MINING SCIENTIFIC PAPERS: NLP-ENHANCED BIBLIOMETRICS

EDITED BY: Iana Atanassova, Marc Bertin and Philipp Mayr

PUBLISHED IN: Frontiers in Research Metrics and Analytics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88945-964-3
DOI 10.3389/978-2-88945-964-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

MINING SCIENTIFIC PAPERS: NLP-ENHANCED BIBLIOMETRICS

Topic Editors:
Iana Atanassova, Université de Bourgogne Franche-Comté, France
Marc Bertin, Université Claude Bernard Lyon 1, France
Philipp Mayr, Leibniz Institute of Social Sciences (GESIS), Germany

Citation: Atanassova, I., Bertin, M., Mayr, P., eds. (2019). Mining Scientific Papers: NLP-enhanced Bibliometrics. Lausanne: Frontiers Media.
doi: 10.3389/978-2-88945-964-3

Table of Contents

04 ***Editorial: Mining Scientific Papers: NLP-enhanced Bibliometrics***
Iana Atanassova, Marc Bertin and Philipp Mayr

07 ***Is the Abstract a Mere Teaser? Evaluating Generosity of Article Abstracts in the Environmental Sciences***
Liana Ermakova, Frederique Bordignon, Nicolas Turenne and Marianne Noel

24 ***The Termolator: Terminology Recognition Based on Chunking, Statistical and Search-Based Scores***
Adam L. Meyers, Yifan He, Zachary Glass, John Ortega, Shasha Liao, Angus Grieve-Smith, Ralph Grishman and Olga Babko-Malaya

38 ***Deep Reference Mining From Scholarly Literature in the Arts and Humanities***
Danny Rodrigues Alves, Giovanni Colavizza and Frédéric Kaplan

51 ***Temporal Representations of Citations for Understanding the Changing Roles of Scientific Publications***
Jiangen He and Chaomei Chen

60 ***Resolving Citation Links With Neural Networks***
Tadashi Nomoto

73 ***The NLP4NLP Corpus (I): 50 Years of Publication, Collaboration and Citation in Speech and Language Processing***
Joseph Mariani, Gil Francopoulo and Patrick Paroubek

103 ***The NLP4NLP Corpus (II): 50 Years of Research in Speech and Language Processing***
Joseph Mariani, Gil Francopoulo, Patrick Paroubek and Frédéric Vernier

frontiers
in Research Metrics and Analytics

EDITORIAL
published: 30 April 2019
doi: 10.3389/frma.2019.00002

Editorial: Mining Scientific Papers: NLP-enhanced Bibliometrics

Iana Atanassova[1]*, *Marc Bertin*[2] *and Philipp Mayr*[3]

[1] *CRIT, Université de Bourgogne Franche-Comté, Besançon, France,* [2] *ELICO, Universié Claude Bernard Lyon 1, Lyon, France,* [3] *GESIS - Leibniz Institute for the Social Science, Cologne, Germany*

Keywords: text mining, scientific papers, scientometrics, natural language processing, computational linguistics, citation content analysis, academic search

Editorial on the Research Topic

Mining Scientific Papers: NLP-enhanced Bibliometrics

1. INTRODUCTION

The Research Topic on "NLP-enhanced Bibliometrics" aims to promote interdisciplinary research in bibliometrics, Natural Language Processing (NLP) and computational linguistics in order to enhance the ways bibliometrics can benefit from large-scale text analytics and sense mining of papers. The objectives of such research are to provide insights into scientific writing and bring new perspectives to the understanding of both the nature of citations and the nature of scientific papers and their internal structures. The possibility to enrich metadata by the full-text processing of papers offers a new field of investigation, where the major problems arise around the organization and structure of text, the extraction of information and its representation at the level of metadata.

Recently, the ever growing availability of datasets and papers in full text and in machine-readable formats has made possible a change in perspective in the field of bibliometrics. From preprint databases to the Open Access and the Open Science movements, the development of online platforms such as ArXiv, CiteSeer or PLoS and so forth, largely contribute to facilitating the experimentation with datasets of articles, making it possible to perform bibliometric studies not only considering the metadata of papers but also their full text content.

The field of NLP offers methodological frameworks and tools for the full text processing of papers that can enlighten bibliometric studies. Some of the open source tools for text processing that have been recently applied to such tasks include NLTK, Mallet, OpenNLP, CoreNLP, Gate, CiteSpace, AllenNLP, and others. Many datasets are now freely available for the community: e.g., PubMed OA, CiteSeerX, JSTOR, ISTEX, Microsoft Academic Graph, ACL anthology, etc. The further developments in this field of study need producing annotated corpora and shared evaluation protocols in order to enable the comparison between different tools and methods. The development of such resources is an important step to making scientific reproducibility possible.

2. PAPERS IN THIS RESEARCH TOPIC

The seven papers published in this Research Topic were all reviewed by two independent reviewers.

In the paper "Is the Abstract a Mere Teaser? Evaluating Generosity of Article Abstracts in the Environmental Sciences," Ermakova et al. examines the abstracts of scientific papers. In fact, the abstract points out the information that is the most important for the reader and is often used as a proxy for the content of an article. The authors propose the GEM score that measures the representativeness of an abstract or its "generosity." To obtain this score, sections in the papers were weighted according to their importance to the reader and sentences in the abstracts were

OPEN ACCESS

Edited and reviewed by:
Chaomei Chen,
Drexel University, United States

****Correspondence:***
Iana Atanassova
iana.atanassova@univ-fcomte.fr

Received: *28 March 2019*
Accepted: *10 April 2019*
Published: *30 April 2019*

Citation:
Atanassova I, Bertin M and Mayr P (2019) Editorial: Mining Scientific Papers: NLP-enhanced Bibliometrics. Front. Res. Metr. Anal. 4:2. doi: 10.3389/frma.2019.00002

assigned to different sections based on their similarity with the content of the sections. More than 36,000 papers in environmental sciences, retrieved from the ISTEX database, were processed to observe the trends in the GEM score over an 80-year period of time. The results show that abstracts tend to be more generous in recent publications and there seems to be no correlation between the GEM score and the citation rate of the papers.

In the paper "The Termolator: Terminology Recognition Based on Chunking, Statistical and Search- Based Scores," Meyers et al. propose an open-source high-performing terminology extraction system called Termolator which utilizes a combination of knowledge-based and statistical components. The Termolator tool includes chunking that favors chunks containing out-of-vocabulary words, nominalizations, technical adjectives, and other specialized word classes and supports term chunk ranking. The authors analyse the effectiveness of all involved components to the overall system's performance and compare their Termolator system with a terminology extraction system called Termostat. They use a gold standard consisting of manually annotated instances of inline terms (multi-word nominal expressions) of different types of documents (e.g., patent, journal article).

In the paper "Deep Reference Mining From Scholarly Literature in the Arts and Humanities," Rodrigues Alves et al. work on a deep learning architecture for the detection, extraction and classification of references within the full text of scholarly publications. The authors explore word and character-level word embeddings, different prediction layers (Softmax and Conditional Random Fields) and multi-task over single-task learning components. Their experiments are based on a published dataset of annotated references from a corpus of publications on the historiography of Venice (books and journal articles in Italian, English, French, German, Spanish and Latin) published from the nineteenth century to 2014. In the evaluation the authors show the relative positive contribution of their character-level word embeddings. The authors release two implementations of the architecture, in Keras and TensorFlow, along with all the data to train and test. Their results strongly support the adoption of deep learning methods for the general task of reference mining.

In the paper "Temporal Representations of Citations for Understanding the Changing Roles of Scientific Publications," He and Chen propose an analysis of the temporal characteristics of citations in order to represent the dynamic role of scientific publications. For this purpose, they study and compare different types of citation contexts in order to identify articles that play important role in the development of science. The proposed methods can have different applications, such as improving citation-based techniques at the individual or collective level, but also improving recommendation systems dedicated to information retrieval by identifying articles of importance or interest.

In the paper "Resolving Citation Links With Neural Networks," Nomoto presents a novel way to tackle the citation resolution through the application of neural network models and identifying some of the operational factors that influence their behavior. The author introduces the notion *approximately correct targets* which is "an idea that we should treat sentences that occur in the vicinity of true targets as equally correct, whereby we try to identify an area which is likely to include a true target, instead of finding its exact location." Experiments in the paper are conducted using three datasets developed by the CL-SciSumm Shared Task (ACL repository) and a cross validation style setup.

The two papers "The NLP4NLP Corpus (I and II): 50 Years of Publication, Collaboration and Citation in Speech and Language Processing" by Mariani et al. and Mariani et al., present the results of an extensive study of a dataset in the field of Natural Language Processing (NLP) and Spoken Language Processing (SLP) for the period 1956–2015. The authors investigate various trends that can be observed from the publications in this specific research domain. The study is presented in two companion papers that each provides a different perspective of the analysis. The first paper describes the corpus and presents an overall analysis of the number of papers, authors, gender distributions, co-authorship, collaboration patterns and citation patterns. The second paper investigates the research topics and their evolution over time, the key innovative topics and the authors that introduced them, and also the reuse of papers and plagiarism. Together, the two papers provide a survey of the literature in NLP and SLP and the data to understand the trends and the evolution of research in this research community. This study can also be seen as a methodological framework for producing similar surveys for other scientific areas. The authors report on the major obstacles that appear during such processing. The first one are the errors that are due to the automatic processing of the full text of papers and in particular scanned content. The second obstacle is the lack of a consistent and uniform identification of authors, affiliations, conference titles, etc. which all require manual corrections by experts in the research area that is investigated.

3. CONCLUSION

The large number of studies on the use of scientific documents with bibliometric applications shows the growing interest of the bibliometric community in this subject. Since 2016, we have been maintaining the "Bibliometric-enhanced-IR Bibliography[1]" which is a bibliography of all scientific articles (workshops and journals) on this Research Topic. In 2018, two special issues closely related to this Research Topic were published. The first one is the special issue on "Bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL)" in the *International Journal on Digital Libraries* (Mayr et al., 2018). The second one is "Bibliometric-enhanced Information retrieval and Scientometrics" in *Scientometrics* (Cabanac et al., 2018).

The articles published in this Research Topic contribute to the state of the art through theoretical discoveries, practical methods and technologies for the processing of scientific corpora involving full text processing, classification of citations but also their temporal representation, semantic analysis, text mining, and related topics. Taken together, these papers identify some of the new challenges in this area and pave the way for future theoretical frameworks.

[1] https://github.com/PhilippMayr/Bibliometric-enhanced-IR_Bibliography/

The development of deep learning techniques is emerging in this field with approaches based on neural network models and can play a fundamental role in the exploitation of citations and their contexts in the scientific literature. While the development of neural network models requires large resources, the increasing number of datasets that are available today allows the implementation of this type of technology for the analysis of citations. Indeed, two of the articles in this Research Topic deal with the implementation of neural network models for citation analysis (Rodrigues Alves et al. and Nomoto), and other two with the construction and exploitation of a large scale corpus of papers (Mariani et al. and Mariani et al.).

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

The work done by PM was partly funded by the DST-DAAD Project-based Personnel Exchange Programme (PPP) under DAAD project number: 57318047 DST project number: DST/INT/FRG/DAAD/P-28/2017.

REFERENCES

Cabanac, G., Frommholz, I., and Mayr, P. (2018). Bibliometric-enhanced information retrieval: preface. *Scientometrics* 116, 1225–1227. doi: 10.1007/s11192-018-2861-0

Mayr, P., Frommholz, I., Cabanac, G., Chandrasekaran, M. K., Jaidka, K., Kan, M.-Y., et al. (2018). Introduction to the Special Issue on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). *International Journal on Digital Libraries*, 19, 107–111.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Research Metrics and Analytics

ORIGINAL RESEARCH
published: 15 May 2018
doi: 10.3389/frma.2018.00016

Is the Abstract a Mere Teaser? Evaluating Generosity of Article Abstracts in the Environmental Sciences

Liana Ermakova[1,2]*, *Frederique Bordignon*[3], *Nicolas Turenne*[4] *and Marianne Noel*[4]

[1] HCTI–EA 4249, Université de Bretagne Occidentale, Brest, France, [2] Analyse et Traitement Informatique de la Langue Française (ATILF), Université de Lorraine, Nancy, France, [3] Direction de la Documentation, Ecole des Ponts ParisTech, Champs-sur-Marne, France, [4] LISIS, Centre National de la Recherche Scientifique, Université Paris-Est Marne-la-Vallée, Institut National de la Recherche Agronomique, ESIEE Paris, Champs-sur-Marne, France

An abstract is not only a mirror of the full article; it also aims to draw attention to the most important information of the document it summarizes. Many studies have compared abstracts with full texts for their informativeness. In contrast to previous studies, we propose to investigate this relation based not only on the amount of information given by the abstract but also on its importance. The main objective of this paper is to introduce a new metric called GEM to measure the "generosity" or representativeness of an abstract. Schematically speaking, a generous abstract should have the best possible score of similarity for the sections important to the reader. Based on a questionnaire gathering information from 630 researchers, we were able to weight sections according to their importance. In our approach, seven sections were first automatically detected in the full text. The accuracy of this classification into sections was above 80% compared with a dataset of documents where sentences were assigned to sections by experts. Second, each section was weighted according to the questionnaire results. The GEM score was then calculated as a sum of weights of sections in the full text corresponding to sentences in the abstract normalized over the total sum of weights of sections in the full text. The correlation between GEM score and the mean of the scores assigned by annotators was higher than the correlation between scores from different experts. As a case study, the GEM score was calculated for 36,237 articles in environmental sciences (1930–2013) retrieved from the French ISTEX database. The main result was that GEM score has increased over time. Moreover, this trend depends on subject area and publisher. No correlation was found between GEM score and citation rate or open access status of articles. We conclude that abstracts are more generous in recent publications and cannot be considered as mere teasers. This research should be pursued in greater depth, particularly by examining structured abstracts. GEM score could be a valuable indicator for exploring large numbers of abstracts, by guiding the reader in his/her choice of whether or not to obtain and read full texts.

Keywords: abstract, full text, generosity, environmental sciences, measure, metric, scientific articles, text-mining

OPEN ACCESS

Edited by:
*Iana Atanassova,
Université Bourgogne
Franche-Comté, France*

Reviewed by:
*Chengzhi Zhang,
Nanjing University of Science and
Technology, China
Kevin Boyack,
SciTech Strategies, Inc., United States*

***Correspondence:**
*Liana Ermakova
liana.ermakova@univ-brest.fr*

Received: *31 January 2018*
Accepted: *23 April 2018*
Published: *15 May 2018*

Citation:
Ermakova L, Bordignon F, Turenne N and Noel M (2018) Is the Abstract a Mere Teaser? Evaluating Generosity of Article Abstracts in the Environmental Sciences. Front. Res. Metr. Anal. 3:16. doi: 10.3389/frma.2018.00016

INTRODUCTION

Scientific journals use abstracts to succinctly communicate research results. Acting as separate entities with respect to full papers, abstracts are generally a free material with easy access.

Abstracts of published manuscripts were introduced in the 1950s (Zhang and Liu, 2011). The notion of an abstract is part of everyday language, but its definitions are multiple: the term "abstract" is used loosely to refer to almost any brief account of a longer paper. Most definitions refer to ideal abstracts produced by professional summarizers. Orasan (2001) argues that it is very unlikely that an abstract produced by the author(s) of a paper is intended to be used as a replacement for the whole document. Therefore, we suggest using a simple functional definition of an abstract: "a concise representation of a document's contents to enable the reader to determine its relevance to a specific information" (Johnson, 1995). So, the abstract is no longer a "mirror" of the document; instead it is intended to draw attention to the most important information of the document it is supposed to summarize (Orasan, 2001).

The abstract represents the primary point of entry to a scientific article, a "point de passage obligé" (Callon and Latour, 1991; Crosnier, 1993). In the context of a rapid increase in the number of scientific journals, abstracts are useful to capture a large volume of documents. Abstracts are also an answer to external demands: publishers of some periodicals and the ANSI NISO standard [ANSI/NISO Z39.14-1997 (R2009)] require or recommend specific information that represents the content of texts reporting results of experimental work, or descriptive or discursive studies to be present in abstracts.

Scientific articles typically have a number of different audiences: the referees, who help the journal editor decide whether a paper is suitable for publication; the journal readers themselves, who may be more or less knowledgeable about the topic addressed in the paper[1]. Most journals ask for between 150 and 200 words for traditional abstracts (i.e., those without subheadings). Structured abstracts, which are divided into a number of named sections, can be longer than traditional ones (Hartley, 2004).

The abstract has been the subject of many research projects, including attempts to evaluate their quality (Narine et al., 1991; Timmer et al., 2003; Sharma and Harrison, 2006; Prasad et al., 2012; Fontelo et al., 2013). In the past two decades, researchers have carried out a number of studies on structured abstracts from different perspectives, and compared abstracts in biomedical journals with those from social sciences journals (see review of James Hartley's research on structured abstracts; Zhang and Liu, 2011).

What we argue here is that the abstract is based on a series of terminological, syntactical and stylistic choices made by the author(s) (Crosnier, 1993). Through a psycholinguistic analysis and readability tests, Guerini et al. (2012) showed that the linguistic style of abstracts contributes to determining the success and viral capability of a scientific article. Scientific texts allow the construction of knowledge claims (Myers, 1985). The act of writing a paper corresponds to an attempt to claim ownership of a new piece of knowledge, which is to be integrated into the repository of scientific knowledge in the author's field by the process of peer review and publication (Teufel et al., 2009).

In this paper, we look at the issue from the perspective of the researcher, who is both an author and a reader. We introduce cognitive processes, i.e., the intention of the author when writing what we call a "*generous*" or "*non-generous*" abstract. While the journal may issue instructions for the abstract, in the act of writing, the author[2] makes his/her own choices (in terms of terminology, syntax and style) and this is what we aim to catch through our measurement of generosity. Our goal in this paper is to define a set of principles from which the generosity score (of an abstract X to its corresponding full text Y) can be calculated. It differs from previous work in that it weights different sections of the paper by their importance.

In our definition, generosity means more than informativeness (a ratio of Y found in X). Indeed, we could have an abstract that scores excellently compared to the full text it summarizes, but which is not very generous. Schematically speaking, a generous abstract should have the best possible score of similarity with the sections that are important to the reader; sections must therefore be weighted according to their importance in the calculation. Matching sentences from the abstract with those issued from the full text was inspired by the work of (Atanassova et al., 2016), who aimed to compare abstract sentences with sentences issued from a full text.

Our study aims to answer the following research questions:

1) Is the abstract a teaser rather than an exact reflection of the article content? By teaser we mean a promotional device or advert intended to arouse interest or curiosity for what will follow.
2) Are authors who write generous abstracts also generous in providing open access to their work?
3) Has generosity of abstracts evolved over time in the case study field of environmental sciences?

These are the questions addressed in remainder of this paper using text-mining techniques and the voluminous database available from ISTEX, combined with the results of an online questionnaire. In the Related Work section, we clarify the motivation for the work presented and situate the focus of our research. The Materials and Methods section includes the constitution of a dataset of 36,237 articles in the environmental sciences and details the two approaches chosen: on one hand, an online questionnaire on researchers' practices and their relationship with the abstract; on the other hand, the definition of the automatic metric GEM (for GEnerosity Measure) that calculates an abstract's generosity. The Results section presents evaluations of the section classification tool and GEM score. Finally, we conducted an experiment aiming to apply GEM to the defined dataset.

[1] https://www.nature.com/scitable/topicpage/scientific-papers-13815490

[2] In case of a multiple authorship, we make the hypothesis that the authorship is endorsed collectively.

RELATED WORK

Overview of Studies on Scientific Abstract

Our research is relevant to several aspects of the scientific literature, on which we have chosen to focus. First, there is a need to apply text-mining techniques to retrieve information from the ever-increasing number of scientific documents, in order to help researchers identify the most appropriate work to base their future research upon.

Many studies have been conducted to compare scientific texts, particularly between the different contents or versions of a publication: title, abstract, keywords, preprint, and published version. Because of the massive quantities of information produced in biological and medical research within a short period of time and the necessity for researchers to stay up to date, experiments have been carried out in life sciences and medicine to check whether it was worth the effort to mine full texts or whether the title, abstract, and keywords freely available could be sufficient to gain a clear picture of what is relevant and useful. Shah et al. (2003) demonstrated that even though abstracts display many keywords in a small space there is much more relevant information (at least in a ratio of 1:4 regarding gene names, anatomical terms, organism names, etc.) in the rest of the article.

PubMed Central is the most comprehensive index to medical literature and has been pioneering in open access since 1997. It opened the door to the free building of text collections for automatic extraction leading to the first web-based platform in molecular biology, called iHOP (Information Hyperlinked over Proteins)[3]. By using specific genes and proteins as hyperlinks between sentences and abstracts, the information in PubMed can be converted into one navigable resource. Based on named entity recognition, iHOP processed 14 million abstracts to extract 11 million molecular relationships for 2,700 living organisms (Blaschke et al., 1999). In the field of biomedicine, some studies for drug target discovery (Kafkas et al., 2017) integrated full texts and abstracts into a massive database, successfully mining more than 26 million abstracts and about 1.2 million full texts for 1.1 million target-drug discoveries. However, when considering paragraph-sized segments of full text articles, searching performed on abstracts only is shown to be far less efficient.

Using their own technology to compare 23 million PubMed abstracts and 2.5 million full text biology articles, Elsevier (2015) showed that more relevant and interesting facts are retrieved from a full text corpus than one containing abstracts alone. More recently, with a similar corpus and methodology, Westergaard et al. (2017) came to the same conclusion. In fields other than biology, Klein et al. (2016) investigated the textual similarity of scholarly preprints and their final published counterparts (12,202 published versions of articles on physics, mathematics, statistics, and computer sciences) and found no significant difference between preprints and published versions.

Using the TREC-2007 genomics track test collection (162,259 full text articles assembled in 2006), Lin (2009) showed that treating an entire article as an indexing unit is not consistently more effective than an abstract-only search. However, when considering paragraph-sized segments of full text articles, searching performed on abstracts alone was shown to be far less efficient. These findings are consistent with Corney's (Corney et al., 2004) conclusions showing that the density of 'interesting' facts found in the abstract is much higher than the corresponding density in the full text.

Scientific papers are highly discursive since they aim to show a view with demonstrative arguments (or proofs). Discourse analysis can help to capture the organization of discursive elements related to argumentation: alternative views, arguments from authority, pros and cons arguments, etc. (Perelman and Olbrechts-Tyteca, 1958; Toulmin, 2003). Khedri et al. (2013) used what they call meta-discourse markers (such as "firstly" and "in conclusion") that refer explicitly to aspects of the organization of a text. Mann and Thompson (1988) developed a grammar theory called "Rhetorical Structure Theory" (RST) about the recurrent structure of scientific paper content. Teufel and Kan (2011) investigated the potential of weakly-supervised learning for argumentative zoning of scientific abstracts. They chose seven categories of argumentative zone: background, objectives, methods, results, conclusion, related work, and future work. Our work builds upon a method relating to such zoning and introduces weighting of sections from the full text that match content of the abstract.

Automatic Metrics for Summary Evaluation

As far as descriptive statistics are concerned, different notions of "similarity" between texts have been incorporated in text-comparison algorithms. The literature provides many string metrics (also known as a string similarity metrics or string distance functions) that are used for approximate string matching or comparison and in fuzzy string searching, e.g., cosine (Manning et al., 2008), Dice (Sørensen, 1948), or Jaccard similarity (Tanimoto, 1958). Similarity between the full text and an abstract may also be estimated by the number of shared n-grams or longest common subsequence, etc. (Cormen et al., 2009).

Other metrics are more specific to the task of document summarization. The simplest metric is a compression rate, i.e., the proportion of summary length in relation to full text length. This metric is opposed to a retention rate, i.e., the proportion of information retained, which is difficult to formalize (Gholamrezazadeh et al., 2009). Thus, a good summary should have a low compression rate and a high retention rate.

The metrics commonly used in information retrieval, such as recall and precision over the number of terms/sentences appearing in the full text and the abstract (Gholamrezazadeh et al., 2009) could also be applied. The F-measure (Lin, 2004) is less useful in summary analysis than in search engines since it is based on recall, and the results returned by search engines are potentially infinite while a summary is limited.

One of the most commonly used metrics of summary evaluation is the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) family (Lin, 2004): ROUGE-N (n-grams recall), ROUGE-N-MULTI (maximal value of pairwise *n*-gram

[3] http://www.ihop-net.org (Accessed December 2017)

recalls), ROUGE-L (longest common substring shared by two sentences), ROUGE-S (shared bigrams which may be separated by other words), ROUGE-SU (unigram smoothing). ROUGE-BE, DemokritosGR2, catholicasc1, and CLASSY1 significantly outperformed ROUGE-2, which is the best performing of all ROUGE variants at the Automatically Evaluating Summaries of Peers (AESOP) task within the Text Analysis Conference (TAC) (Owczarzak et al., 2012). Normalized pairwise comparison LCS-MEAD (Radev et al., 2002) is similar to ROUGE-L, but LCS-MEAD takes the maximal value of longest common substring (LCS), while ROUGE-L deals with the union of LCSs (Hovy and Tratz, 2008). One of the serious shortcomings of LCS is the fact that it does not consider the distance between words. An attempt was made to overcome this drawback by using weighted LCS, which takes into account the length of consecutive matches. LCS-based algorithms are a special case of edit distance (Bangalore et al., 2000).

Campr and JeŽek (2015) proposed to use the similarity within semantic representation such as LSA, LDA, Word2Vec, and Doc2Vec. However, ROUGE-1 outperformed all these metrics. In (Ng and Abrecht, 2015), the ROUGE metric was modified by word embedding, but this variant showed lower results than the standard one.

A Pyramid score is based on the number of repetitions of information in the gold-standard model summaries (Nenkova et al., 2007), which can be replaced by a full text. Because Pyramid score requires heavy manual annotation of both gold-standard and candidate summaries it is not applicable to large corpora.

In (Owczarzak et al., 2012), a responsiveness metric is proposed. This metric shows how well a summary satisfies the user's information need expressed by a given query and is completely manual. Louis and Nenkova (2013) suggest using the full text instead of a set of reference summaries for summary evaluation. They estimated summary score by Kullback–Leibler divergence, Jensen Shannon divergence, and cosine similarity measure. Although these metrics have some correlation with ROUGE score, ROUGE-1 gave better results. In the INEX Tweet Contextualization Track 2011–2014, summaries were evaluated by the Kullback–Leibler divergence and simple log difference (Bellot et al., 2016). The authors state that the Kullback–Leibler divergence is very sensitive to smoothing in case of small numbers of relevant passages in contrast to the absolute log-diff between frequencies (Bellot et al., 2016). Cabrera-Diego et al. (2016) introduced a trivergent model that outperformed the divergence score.

In this paper, our main task is to provide a measure of the generosity of an abstract of a scientific article with regard to the full text. The use of the full text rather than a set of reference summaries for summary evaluation provides low results (Louis and Nenkova, 2013) since traditional metrics are designed for the comparison with summaries created by humans. Thus, they are not appropriate for comparison of an abstract produced by humans with the full text. All these existing metrics have relative values allowing candidate summaries to be ranked, which has two major consequences. First, these measures are not applicable for comparison of an isolated abstract with the full text, e.g., ROUGE score would depend on the length of the full text. Second, it is not possible to compare metric scores for abstracts of different documents.

Another problem with the existing metrics is their output values. Theoretically, the majority of metrics are normalized, but in practice, the values tend to be quite small (usually <0.2).

Last, but not least, the final drawback is that none of these measures take into account document structure. As demonstrated by Fontelo et al. (2013), "structured abstracts appear to be informative." One of the metrics considering document structure is BM25F (Robertson et al., 2004) which is a field-based extension of Okapi's BM25 widely used in information retrieval. However, it is not suitable for abstract scoring since it also gives a relative score allowing search result ranking.

In contrast to the state-of-the-art measures listed above, the metric proposed in this paper (GEM) has absolute values in the interval [0,1]. It also considers the importance of different sections by introducing weighting of sections in full text that match with sentences in the abstract. These weightings were determined by an online questionnaire of researchers' opinions described in the next section.

MATERIALS AND METHODS

Dataset

Our analysis was based on a corpus of articles in the field of environmental sciences published from 1930 to 2013. This corpus was obtained from the Excellence Initiative for Scientific and Technical Information (ISTEX) database[4]. ISTEX provides the French higher education and research community with online access to scientific archives in all disciplines. At the time of writing of the present article, this archive contains collections of scientific literature from all disciplines, covering journal archives, digital books, databases, text corpora, etc. from the following publishers: Elsevier, Wiley, Springer, Oxford University Press, British Medical Journal, IOP Publishing, Nature, Royal Society of Chemistry, De Gruyter, Ecco Press, Emerald, Brill, and Early English Books Online.

The ISTEX platform provides a set of services via an HTTP-based web Application Programming Interface (API)[5] within a RESTful (REpresentational State Transfer) paradigm, i.e., the platform allows access and manipulation of textual representations of resources using a uniform and predefined set of stateless operations. A Graphical User Interface (GUI) is also available as a form of demonstration[6]. The API enables to search for documents and their metadata. Search results and document metadata in JSON or MODS formats are available on open access, while access to retrieved documents is restricted and requires authentication. Documents are available in the following formats:

- PDF (full text);
- TEI (full text and enrichments);
- XML provided by a publisher;

[4] http://www.istex.fr/

[5] https://api.istex.fr/documentation/

[6] http://demo.istex.fr/

- Different formats (images, videos, sounds, etc.) corresponding to appendices and publication covers.

We retrieved 66,518 articles (tagged as *research-article* or *article* in the ISTEX database) categorized by ISTEX as "*Environmental Studies*" or "*Environmental Science*" (according to the Web of Science classification). We selected articles for which we could retrieve a full text and an abstract from the PDF file. Out of the 59,419 article/abstract pairs thus obtained, we then chose to filter out documents having less than four section classes in the full text: 23,181 articles were therefore considered unsuitable for further analysis. The definitive dataset was composed of 36,237 articles (see published dataset of results in Bordignon and Ermakova, 2018).

Online Questionnaire

An online questionnaire was designed to analyze the way in which a sample of researchers read and write abstracts. The questionnaire was developed on the basis of a broad definition of the abstract, which is divided into seven sections. The following definitions of abstract section classes were provided in the questionnaire:

Introduction—Context

This section describes what is already known about the subject in a way that is understandable to researchers from all fields.

Objectives

The aim here is to describe what is not yet known but which can be discovered or answered by the research or reasoning developed in the article.

Methods—Design

This section informs the reader of the techniques and strategies used to conduct the research and demonstrate its validity (for instance, material used, methodological framework, population being studied, data collection process, sample size, etc.).

Results—Observations—Findings

The main results are presented here, accompanied by the data (possibly quantified) that made it possible to characterize them. These may also be negative results that do not support the initial hypothesis.

Conclusions

This part contains the main message of the article. It shows how the results are interpreted and how the initial question from the objectives is answered.

Limits

If any limitations have been identified, they are presented here.

Perspectives

The aim here is to position the results of the study in a more general context in order to show to what extent there has been progress in understanding and how further studies could lead to new developments.

The questionnaire was strictly anonymous (identities, first and last names, contact details, or e-mail addresses were not asked), and no consent was needed as we retrieved no individual information. The questionnaire had no commercial intent, didn't target individuals and participants were informed of their participation in a research project. It was signed by us and respondents were informed of our status. The link to respond was open to anyone and sent via our professional mailing lists and Twitter accounts. The data did not need to be anonymized and are published (Bordignon and Noël, 2018).

This online questionnaire was completed by 630 individuals between 08/24/2016 and 09/27/2016, to whom these definitions were presented. The large majority of respondents are researchers: 50% are PhD students or postdocs, 43% are professors or permanent researchers. Interviewees were asked to provide a maximum of two disciplinary fields (from a list of 12) that characterize their research.

We asked the respondents to rank the seven sections in their respective fields according to the following scale: essential, important, marginal, optional or unusual, or unknown. We also asked if a good abstract is more like a summary or a teaser. They had the opportunity to send us one or two abstracts that they consider successful, and examples of journal names whose abstracts they consider satisfactory (see published dataset of answers; Bordignon and Noël, 2018).

The last question in the questionnaire was about generosity, a concept that we intentionally did not define in the questionnaire: *"In your opinion, which section must imperatively be present in the abstract so that it can be qualified as 'generous'?"*

Out of the respondents, 32% considered that the Results—Observations—Findings section must be present in the abstract if it is to be considered generous and 27% thought mentioning the Objectives in the abstract to be a sign of generosity. Conclusions (16%) and Methods—Design (12%) were in third and fourth place in terms of interest, respectively. Introduction—Context (5%), Perspectives (5%), and Limits (3%) were the sections considered to be of least interest with regard to generosity (see **Figure 1**). These results were then used to weight the sections detected in the full texts, whose equivalents were either found or not found in the abstract. **Table 1** shows there was

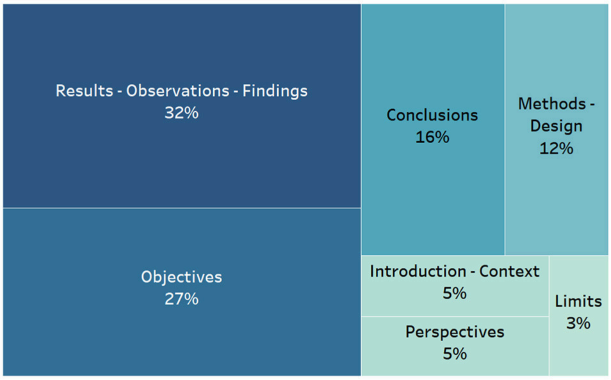


FIGURE 1 | Online questionnaire answers to the question "Which section must imperatively be present in the abstract so that it can be qualified as "generous"?" (630 respondents).

TABLE 1 | Answers distribution according to the disciplines.

	Social sciences (%)	Engineering (%)	Computer science (%)	Physics (%)	Environmental sciences (%)	Life sciences (%)	Chemistry (%)	Economics and finance (%)	Maths (%)	Planets and universe (%)	Cognitive sciences (%)	Statistics (%)
Conclusions	19	16	16	17	20	16	11	22	16	15	11	16
Introduction—Context	7	4	5	6	3	5	2	3	5	8	0	5
Limits	4	2	3	4	6	3	0	3	3	8	0	3
Methods—Design	8	16	12	10	6	12	13	11	12	8	33	12
Objectives	24	24	27	23	29	27	30	30	27	8	22	27
Perspectives	7	8	5	5	1	5	2	0	5	8	22	5
Results—Observations—Findings	31	30	32	34	35	32	43	32	32	46	11	32
Number of respondents	190	139	105	77	70	53	48	37	33	13	9	3

no significant difference among the disciplines the respondents identify themselves with, more particular for fields with more than 30 respondents. There was indeed no need to take various disciplines into account when weighting sections differently.

GEM, A MEASURE OF ABSTRACT GENEROSITY

We introduce here a completely automatic metric for the estimation of abstract generosity called GEM (for GEnerosity Measure), which attributes an absolute score [0,1] to an abstract. GEM relies on the importance of the different sections of a scientific paper according to the researchers' opinions obtained from the questionnaire results described above (**Figure 2**).

First of all, we considered that the score calculated by GEM was reliable only if at least four section classes (out of the seven section classes we listed above and submitted to the respondents of the questionnaire) could be automatically identified in the full text using the GROBID tool for section splitting and our algorithm for sentence classification presented below. Otherwise, we considered the estimated score to be unreliable, as GEM is based on the weighting of the detected sections.

Thus, the main steps were the following:

1. Section detection in the full text (using GROBID to split it into sections);
2. Classification of the sections from the full text (position, section embedding, regular expressions, and quantitative features such as number of tables, references, and figures);
3. Sentence splitting in the abstract by Stanford CoreNLP[7] (Manning et al., 2014) and estimation of similarity between article sections and corresponding abstract sentences (cosine similarity measure between TF-IDF representations);
4. Calculation of the GEM score. The informativeness rate was weighted according to the importance of the sections.

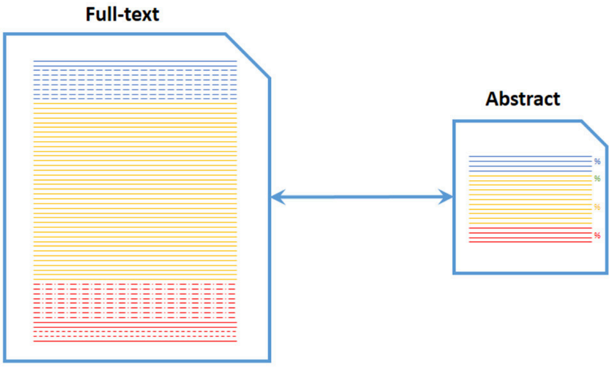


FIGURE 2 | Principle of GEM score as a comparison between the full text and abstract relying on detection of sections.

Figure 3 presents the flow diagram of the algorithm. This model was implemented in Java (Ermakova, 2018).

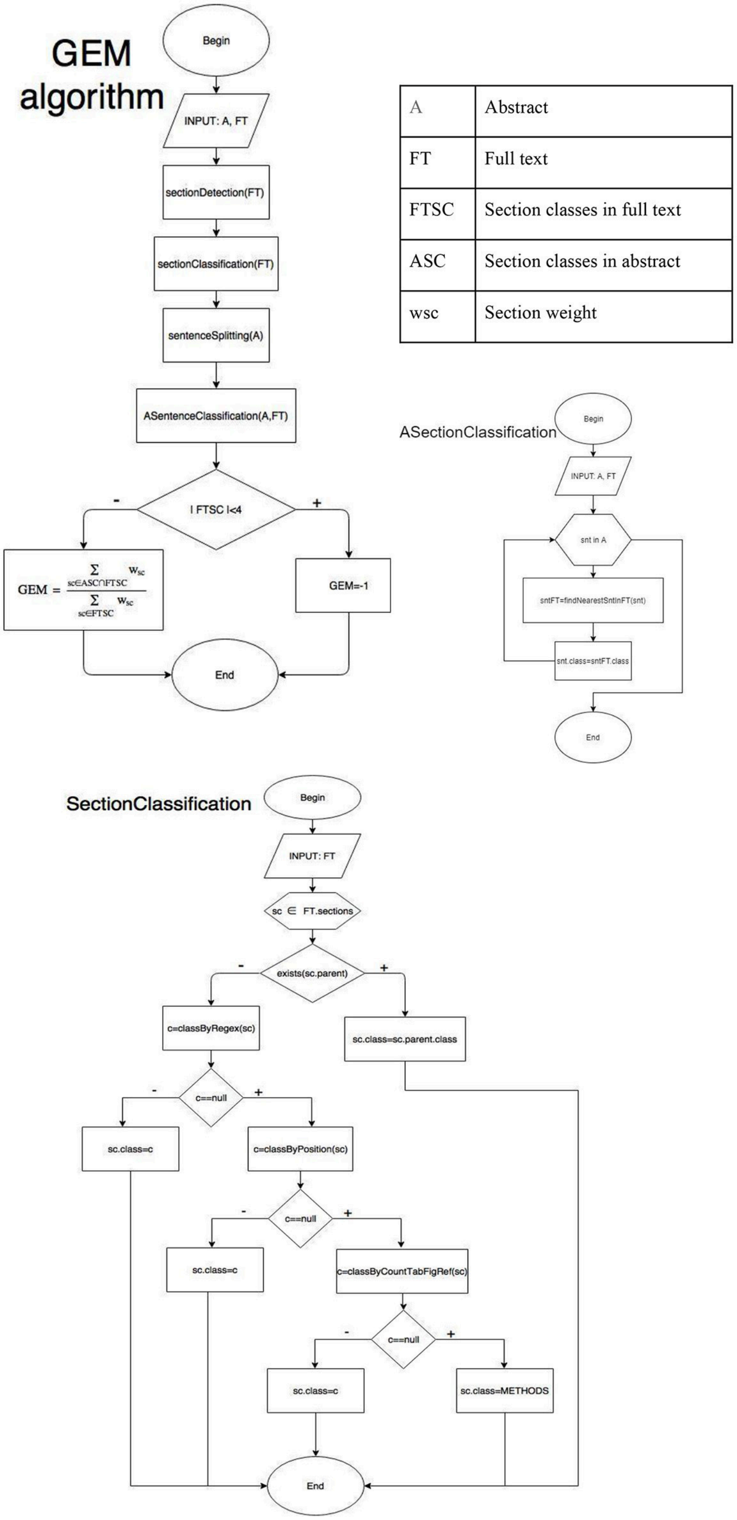
Article Section Detection

The first step of our algorithm is section detection by GROBID software[8]. GeneRation Of BIbliographic Data (GROBID) is a machine-learning library for parsing PDF documents into structured TEI format designed for technical and scientific publications. The tool was conceived in 2008 and became available in open source in 2011. Its applications include ResearchGate, HAL Open Access repository, the European Patent Office, INIST, Mendeley and CERN.

GROBID enables:

[7] https://stanfordnlp.github.io/CoreNLP/

[8] https://github.com/kermitt2/grobid



A	Abstract
FT	Full text
FTSC	Section classes in full text
ASC	Section classes in abstract
wsc	Section weight

FIGURE 3 | GEM calculation algorithm.

- Header extraction and parsing from articles in PDF format (e.g., extraction of title, abstract, authors, affiliations, keywords, address, etc.);
- Reference extraction and parsing from articles in PDF format, including references in footnotes, isolated references, and patent references;
- Parsing of dates;
- Full text extraction from PDF articles with document segmentation.

Extraction and parsing algorithms use the Wapiti CRF (Conditional Random Fields) library[9]. Wapiti is a toolkit for segmenting and labeling sequences with discriminative models based on maximum entropy Markov models and linear-chain CRF. GROBID is available in batch mode, as well as RESTful and JAVA APIs. We integrated GROBID in our tool using JAVA API.

Section Classification

After extracting sections from a PDF article we classified them into the seven classes described below. As a first step, we classified the sections into four classes: INTRO, METHODS, RESULTS, and CONCLUSION, according to the following rules. The rules were applied as it is given in **Figure 3**. Thus, only one rule can be applied (i.e. only one section is assigned) since if a rule is activated the following rules are not evaluated. We looked for section embedding based on section numbers if they were provided by GROBID or analysis of empty sections with titles only; otherwise, we considered that a section was not embedded in another, i.e., that it was not a subsection. If a section was a subsection, it was assigned the class of its parent; otherwise, we tried to apply regular expressions to a section title in order to classify it (see **Table 2**). If the title did not match any regular expression, we analyzed its relative position in the text, e.g., the first section was considered to be the INTRO. If none of the previous rules were applicable, we assigned the class RESULTS if the section contained figures or tables, or the class INTRO if it contained more than five references. The default class was METHODS.

Second, we applied regular expressions for searching for sentences related to OBJECTIVES in sections attributed to INTRO and sentences referring to LIMITS and PERSPECTIVES in sections already assigned the class CONCLUSION. Splitting into sentences was performed by Stanford CoreNLP.

Words in regular expressions were considered as representative, but we are aware that they are not exhaustive.

Abstract Sentence Splitting and Classification

Our approach to abstract segmentation is inspired by the work of Atanassova et al. (2016), which aimed to compare abstract sentences with sentences issued from a full text. At this step, splitting into sentences was performed by Stanford CoreNLP. Then, we searched for the most similar sentence in the full text and assigned its class to the abstract sentence under consideration. Thus, only one class can be assigned the class of the section that contains the sentence the most similar to the sentence from the abstract under consideration.

TABLE 2 | Regular expressions used for section detection.

SECTION CLASS	DESCRIPTION	REGEX
INTRO	Description of the research context, i.e., introduction of the already known information/problem	(?i).*introduction.* (?i).*state.*of.*the.*art.* (?i).*related.*work.*
OBJECTIVES	A new piece of knowledge that is the focus of a given article	(?i).*objective.* (?i).*the purpose of this.* (?i).* aim.* (?i).*in this paper.* (?i).*in this study.* (?i).*in this research.* (?i).*in this work.* (?i).*a new.*is proposed.* (?i).*we.* propose.*
METHODS	Methods used for the research and its validation, e.g., materials, data, methods etc.	(?i).*method.*
RESULTS	Results obtained (usually numerical data with their interpretation)	(?i).*result.*
CONCLUSION	The main contribution of the paper, answers on the research questions	(?i).*conclu.*
LIMITS	Limitations of the presented research	(?i).*limit.* (?i).*only.* (?i).*wrong.* (?i).*drawback.* (?i).*shortcom.*
PERSPECTIVES	Potential applications and future work	(?i).*potential.* (?i).*perspective.* (?i).*in the pursuit.* (?i).*futur.* (?i).* will.* (?i).*further.*

Many researchers consider text content as weighted phrases (Radev and McKeown, 1998; Erkan and Radev, 2004; Seki, 2005). Phrases are often identified by their frequency in a document or collection or by their distribution in a text.

We hypothesized that TF-IDF cosine similarity should be suitable for capturing similarity between sentences. TF-IDF is a short for term frequency–inverse document frequency. It is a numerical statistics that reflects how important a word is to a document in a corpus. A TF-IDF score is achieved with a high term frequency in the document and a low document frequency of the term in the collection. IDF refers to term specificity. As a term appears in more documents, the IDF (and, therefore, TF-IDF) becomes closer to 0. Hence, the weights tend to filter out common terms. We tested the hypothesis that the TF-IDF measure is able to capture keywords by comparison with author-provided keywords and expert analysis. More than 70% of the top words retrieved by the TF-IDF measure coincided with human-provided keyword lists.

[9] http://wapiti.limsi.fr/

Thus, we applied the TF-IDF-based cosine similarity measure between an abstract sentence *Sa* and a sentence from the full text *Si*:

$$\cos(Sa, Si) = \frac{\sum_{j=1}^{|V|} Sa_j \times Si_j}{\sum_{j=1}^{|V|} {Sa_j}^2 \times \sum_{j=1}^{|V|} {Si_j}^2}$$

where Sa_j and Si_j are TF-IDF scores of the term j in *Sa* and *Si*, respectively, and $|V|$ is vocabulary size. Then, we selected the sentence with the maximal cosine similarity and assigned its class to *Sa*.

It should be noticed here that, in contrast to section classification in the full text, classification in the abstract is performed based on the similarity with sentences from the full text only. Thus, we do not directly consider the regular expressions mentioned above. This decision makes impossible to use key phrases to a trigger section score without any

EXAMPLE 1 | GEM score calculation for Piringer and Steinberg (2008).

Abstract sentence	Closest sentence from the full text	Class
Energy budgets for agricultural production can be used as building blocks for life-cycle assessments that include agricultural products, and can also serve as a first step toward identifying crop production processes that benefit most from increased efficiency.	Moreover, identifying the most energy- consuming steps in wheat production helps to focus energy efficiency efforts, which in turn are likely to reduce important environmental burdens of industrial agriculture, such as nutrient leaching and soil erosion.	INTRO
A general trend toward increased energy efficiency in U.S. agriculture has been reported.	For example, the average electricity generation output in the U.S. is 39.6% of input energy and the average transmission and distribution efficiency in the nationwide grid is 92%.	RESULTS
For wheat cultivation, in particular, this study updates cradle-to-gate process analyses produced in the seventies and eighties.	Some of the resulting detailed analyses of energy coefficients are applicable to wheat production as well and may thus assist in a reevaluation of the earlier studies from the seventies.	INTRO
Input quantities were obtained from official U.S. statistics and other sources and multiplied by calculated or recently published energy coefficients.	Averages for input quantities or embodied energy coefficients were not available.	METHOD
The total energy input into the production of a kilogram of average U.S. wheat grain is estimated to range from 3.1 to 4.9 MJ/kg, with a best estimate at 3.9 MJ/kg.	Based on data mostly from the last decade, the average energy input into the production of a kilogram of U.S. wheat grain is estimated to range from 3.1 to 4.9 MJ/kg, with a best estimate at 3.9 MJ/kg.	CONCLUSION
The dominant contribution is energy embodied in nitrogen fertilizer at 47% of the total energy input, followed by diesel fuel (25%), and smaller contributions such as energy embodied in seed grain, gasoline, electricity, and phosphorus fertilizer.	The dominant contribution to energy input into wheat production is nitrogen fertilizer, accounting for almost half the total energy input.	CONCLUSION
This distribution is reflected in the energy carrier mix, with natural gas dominating (57%), followed by diesel fuel (30%).	Not surprisingly, the energy carrier mix mirrors this distribution, with natural gas (the major energy source in nitrogen fertilizer manufacturing) and diesel fuel (the largest direct energy input) as the dominant inputs, at 57 and 30% of the total energy, respectively.	CONCLUSION
High variability in energy coefficients masks potential gains in total energy efficiency as compared to earlier, similar U.S. studies.	Thus, potential gains in total energy efficiency as compared to earlier, similar studies are masked by the range of the current estimate.	CONCLUSION
Estimates from an input-output model for several input processes agree well with process analysis results, but the model 's application can be limited by aggregation issues: Total energy inputs for generic food grain production were lower than wheat fertilizer inputs alone, possibly due to aggregation of diverse products into the food grain sector.	Its main limitation was demonstrated by the fact that an estimate of total energy inputs into generic food grain production was lower than an estimate of fertilizer energy; this apparent inconsistency may be attributable to influences of nonwheat products that are aggregated with wheat into the U.S. food grain sector.	CONCLUSION

INTRO 0.05
METHOD 0.12
RESULTS 0.32
CONCLUSION 0.16
PERSPECTIVES 0
LIMITS 0

$$GEM = \frac{0.05 + 0.12 + 0.32 + 0.16}{0.05 + 0.12 + 0.32 + 0.16 + 0.05 + 0.03} = 0.89$$

EXAMPLE 2 | GEM score calculation for Schmid et al. (2012).

Abstract sentence	Closest sentence from the full text	Class
The aim of this study was to investigate the effectiveness of different shielding materials in protective clothing using dicentric frequency in human peripheral lymphocytes as a marker of radiation-induced damage.	The present experiments indicate different yields of dicentrics in human lymphocytes exposed to the broad spectrum of diagnostic 70 kV x-rays immediately behind commercially available non-lead based shielding materials in radioprotective clothing.	CONCLUSION
Blood samples from a healthy donor were exposed to 70 kV x-rays behind shielding materials lead (Pb), tin/antimony (Sn + Sb) and bismuth barrier/tin/tungsten (Bi + Sn + W) with the same nominal lead equivalent value of 0.35 mm lead.	In four independently performed experiments (I-IV), blood was exposed to x-rays behind three types of shielding material cut from x-ray protective aprons with the same nominal lead equivalent value (LEV) of 0.35 mm lead: shielding materials lead (Pb), tin/antimony (Sn + Sb) and bismuth barrier/tin/tungsten (Bi + Sn + W).	METHOD
Irradiation was performed either in contact (exposure position A, containing secondary radiation) or at a distance of 19 cm behind the shielding materials (exposure position B, containing only the unaffected transmitted photons).	In experiment I, blood was exposed to 217.2 mGy at two different positions of each shielding material but without moving the blood sample position (Figure 1): in contact with the shielding material (exposure position A) or at a distance of 19 cm behind the shielding material (exposure position B).	METHOD
Using shielding material Sn + Sb, a significantly higher dicentric yield was determined at exposure position A relative to position B, whereas no significant differences were found between the exposure positions using shielding materials Pb or Bi + Sn + W. For doses up to 434.4 mGy at exposure position A, the slopes of the linear dose-response curves for dicentrics obtained behind shielding materials Pb and Bi + Sn + W were not significantly different, whereas a significantly higher slope was determined behind Sn + Sb relative to Pb and Bi + Sn + W. Using moderately filtered 220 kV x-rays as a reference, maximum RBE values at low doses (RBE M) of 1.22 ± 0.10, 2.28 ± 0.19 and 1.03 ± 0.12 were estimated immediately behind shielding materials Pb, Sn + Sb and Bi + Sn + W, respectively.	For exposure to 217.2 mGy (experiment I), no significant difference was determined between exposure positions A and B using shielding materials Pb or Bi + Sn + W, whereas a significantly higher dicentric yield was obtained behind shielding material Sn + Sb at position A relative to position B. Using exposure position A, the dicentric yield behind shielding material Sn + Sb was also significantly higher than the corresponding dicentric yields behind shielding materials Pb or Bi + Sn + W. However, using exposure position B, no significantly different dicentric yields were determined behind the three shielding materials.	RESULTS
These findings indicate a significantly higher RBE M of 70 kV x-rays behind shielding material Sn + Sb with respect to Pb or Bi + Sn + W. Using previous dicentric data obtained for exposure of blood from the same donor to x-rays at energies lower than 70 kV, it can be assumed that the increased RBE M of the broad spectrum of 70 kV x-rays (mean energy of 44.1 keV) may be attributed predominately to secondary (mainly fluorescence) radiation generated in the shielding material Sn + Sb that is able to leave the 0952-4746/12/ 03N129 +11 $ 33.00	In fact, taking into account the large uniform data set obtained with blood from the same donor (ICRP, 2003) showing a strong increase in coefficient α with decreasing photon energy, it can be assumed that the increased RBE M of the broad spectrum of 70 kV x-rays obtained in the present investigation in blood from the same donor should be attributed predominately to photon energies lower than the mean energy of 44.1 keV.	RESULTS

INTRO 0
METHOD 0.12
RESULTS 0.32
CONCLUSION 0.16

$$GEM = \frac{0.12 + 0.32 + 0.16}{0.05 + 0.12 + 0.32 + 0.16} = 0.923$$

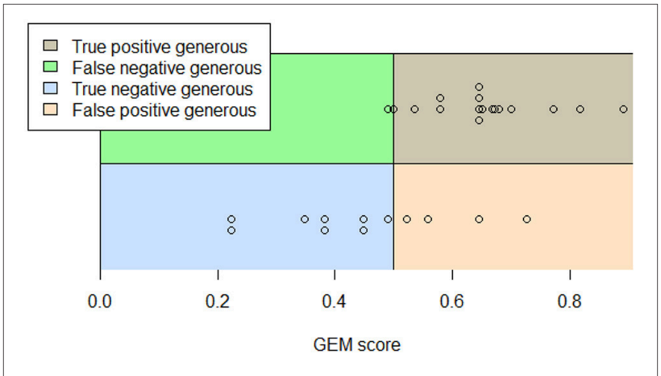


FIGURE 4 | GEM scores according to ground truth.

relation to the full text, e.g., the use of the phrase "we report our results" without actually reporting any results does not necessarily provoke the assignment of the result section score. However, the quality of the full text is out of scope of this research.

GEM Score

The GEM score is an interval [0,1]. If we detected less than four section classes in a full text, we assigned the score −1. This was motivated by the fact that GEM is based on section detection and classification and we believe that our score is more reliable in cases where we detect at least four section classes. The GEM score was calculated as a sum of weights of section classes $w(sc)$ retrieved both in an abstract and a full text normalized over the total sum of weights of section classes in a full text:

$$GEM = \frac{\sum_{sc \in ASC \cap FTSC} wsc}{\sum_{sc \in FTSC} wsc}$$

where *FTSC* denotes section classes in the full text, *ASC* refers to section classes in the abstract, *wsc* corresponds to section weight. Dividing by the sum of all weights of sections from the full text penalizes abstracts that do not reflect sections from the full text, e.g., an abstract representing only result section would have lower score that an abstract of the same length that contains also limits. However, an abstract that presents

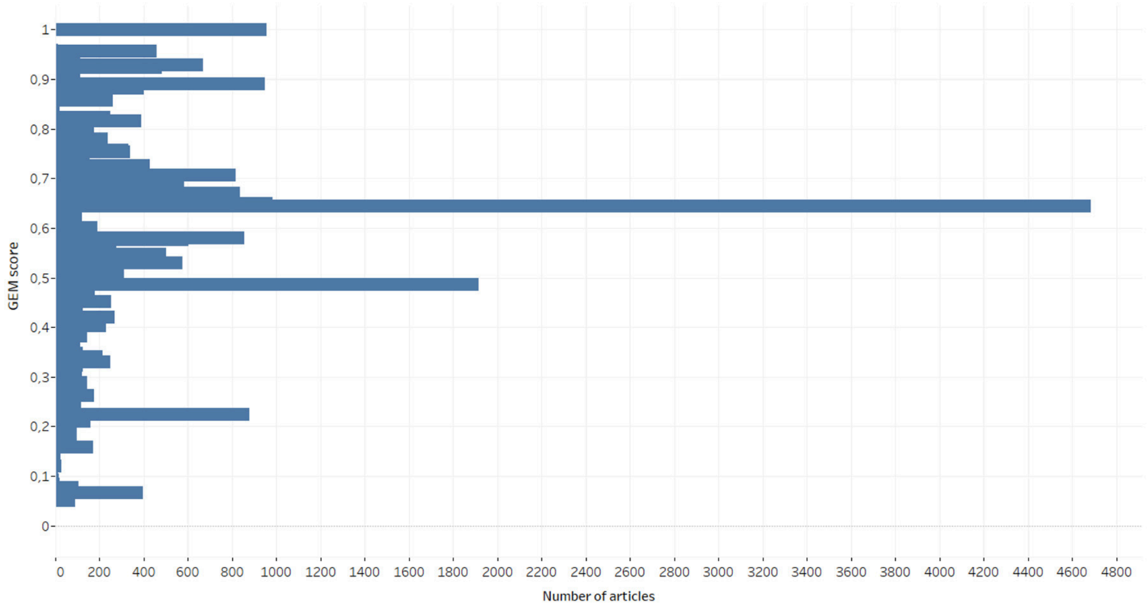


FIGURE 5 | GEM Score distribution for the whole dataset ($n = 36{,}237$).

limits only would be scored lower than an abstract that only details results. GEM does not consider the number nor the length of sentences in the abstract that reflect different full text sections. It measures the presence/absence of the sections in the abstract weighted by their importance according to the scientific community.

Examples of GEM score calculation are given for two articles having different contents and styles above (**Example 1** and **Example 2**).

RESULTS

Section Classification Evaluation

Section classification evaluation was performed over a dataset annotated manually. For manual evaluation, we chose 20 documents at random. For each article, each sentence was tagged by two experts who are both researchers. The first of these experts has expertise in chemistry and the other has experience in economics and environmental sciences. We treated about 4,000 classified sentences. The quality of our classification algorithm was evaluated by a commonly used metric, namely accuracy. Accuracy of our classification was calculated as the number of correctly classified items over the total number of items and was found to be above 80%.

GEM Score Evaluation

We conducted three types of experiment to evaluate the GEM score.

In the first evaluation experiment, we hypothesized that the score assigned to the abstract of a given article should be higher than the score of the abstract coming from another article. Thus, we compared the score assigned to the original abstract with the scores of all other abstracts from the test set. We obtained 25% errors, i.e., in 25% of cases the scores of abstracts corresponding to other articles were higher than the scores of the original ones, while the random score produced 55% of errors on the same dataset. In all cases, the errors of GEM were produced for non-generous original abstracts.

We compared the GEM score with the scores assigned by the experts as in the previous subsection. Forty-two documents were annotated at least by one expert and 20 of these documents were assigned a score by both evaluators. The correlation between GEM scores and the mean of the human assigned scores was 0.59. The correlation between the human annotators was 0.56. We can thus conclude that GEM score reliability is comparable human reliability.

The intuition underlying the third evaluation framework is that a good metric should assign a high score to a generous abstract and a low score to a non-generous one. Rather than calculating the correlation between the scores assigned to abstracts by assessors and metrics, we propose to compare the accuracy, i.e., the percentage of cases where a very generous summary is scored lower than a non-generous one. The motivation is the relative simplicity for a human to distinguish very generous abstracts and abstracts that are not generous at all. We considered only not conflicting assignments as the ground truth. We manually chose 19 generous abstracts and 12 non-generous ones for which we could calculate GEM score. Thus, we had $19 * 12 = 228$ pairs for which we know the preferences. In 90% of cases we obtained a higher score for generous abstracts than for non-generous ones. GEM scores are plotted on **Figure 4**.

TABLE 3 | Numbers and typology of abstracts according to the structure of the full text (sections missing from the abstract appear in red).

Gem score	No. of occurrences	Full text structure	Abstract structure
0.64473684	4683	INTRO OBJECTIVES METHODS RESULTS	INTRO METHODS RESULTS
0.48684211	1916	INTRO OBJECTIVES METHODS RESULTS	INTRO RESULTS
0.65	982	INTRO OBJECTIVES METHODS RESULTS CONCLUSION PERSPECTIVES LIMITS	INTRO METHODS RESULTS CONCLUSION
1	957	All section classes from the full text are presented in the abstract. Different structures can correspond to this value	
0.89041096	948	INTRO METHODS RESULTS CONCLUSION PERSPECTIVES LIMITS	INTRO METHODS RESULTS CONCLUSION
0.22368421	876	INTRO OBJECTIVES METHODS RESULTS	INTRO METHODS
0.57894737	854	INTRO OBJECTIVES METHODS RESULTS	METHODS RESULTS
0.67010309	836	INTRO OBJECTIVES METHODS RESULTS CONCLUSION PERSPECTIVES	INTRO METHODS RESULTS CONCLUSION
0.70652174	816	INTRO OBJECTIVES METHODS RESULTS CONCLUSION	INTRO METHODS RESULTS CONCLUSION
0.92857143	669	INTRO METHODS RESULTS CONCLUSION PERSPECTIVES	INTRO METHODS RESULTS CONCLUSION

Experimental Results

We calculated the GEM score for articles from the definitive dataset ($n = 36{,}237$) (see **Figure 5**).

The most frequent GEM value, 0.6447, occurred 4,683 times. As shown in **Table 3**, this value was attributed to abstracts where three section types (INTRO, METHODS, and RESULTS) were detected in the abstract out of four found in the full text (OBJECTIVES was missing in the abstract). The second largest value (0.4868) corresponds to detection of INTRO and RESULTS in the abstract while four section types are found in the full text (INTRO, OBJECTIVES, METHODS, and RESULTS). INTRO, METHODS, RESULTS, and CONCLUSION are section types that our algorithm looks for at the first stage. They are often organized as well-defined blocks of text in the articles. These results suggest that the sections INTRO, METHODS, and RESULTS are the most frequently presented in the abstract.

As **Figure 6** shows for articles published in the last 40 years, we detected that abstracts tended to become more generous over time. We did not take the period 1930–1975 into account because of the small number of articles.

The fall in the number of articles in 2002 shown in **Figure 6** is inherent to the ISTEX database and more particularly to the end of data acquisition from Elsevier. The number of the remaining articles is still significant because it is above 500 articles a year. This fall in numbers had no effect on the growth of the GEM score over time.

In order to illustrate the GEM score potential, we ambitiously propose additional analyses even if they appear to be premature.

Nine publishers were identified in the definitive dataset (**Table 4**). Half of the dataset articles were published in an Elsevier journal.

We found significant differences between publishers: abstracts from Sage and Springer journals appeared to be less generous than those of other publishers (see **Figure 7**). These results need to be further investigated in order to identify whether the guidelines for authors or even instructions about structured abstracts could have impacted this trend.

The environmental sciences dataset we tested also includes articles from journals categorized in one or more additional subject areas (according to the Elsevier journal classification). **Table 5** shows the distribution among subject areas and **Figure 8** compares the seven most important subject areas excluding environmental sciences that are obviously common to all articles.

This provided an opportunity to compare GEM score between disciplines. No significant differences were found except for the abstracts of articles in the social sciences ($n = 3{,}494$) which were the least generous. In the commentaries collected in our online questionnaire, we also came across views consistent with this conclusion:

"In the field of literary studies, we do not have any abstract of this kind [...]. I tried to answer your questionnaire anyway, but this type of publication is simply not part of our practice (we're talking about articles, codified but not as rigidly)."

These results need to be investigated further, including making a comparison with a social sciences corpus that could also be retrieved from the ISTEX database.

Finally, we used the oaDOI API[10] to look for open access versions of the articles. As far as we know, literature about openness and open access to publications does not deal with abstract content. So we aimed to identify whether authors who wrote generous abstracts were also generous in providing

[10] https://oadoi.org/api

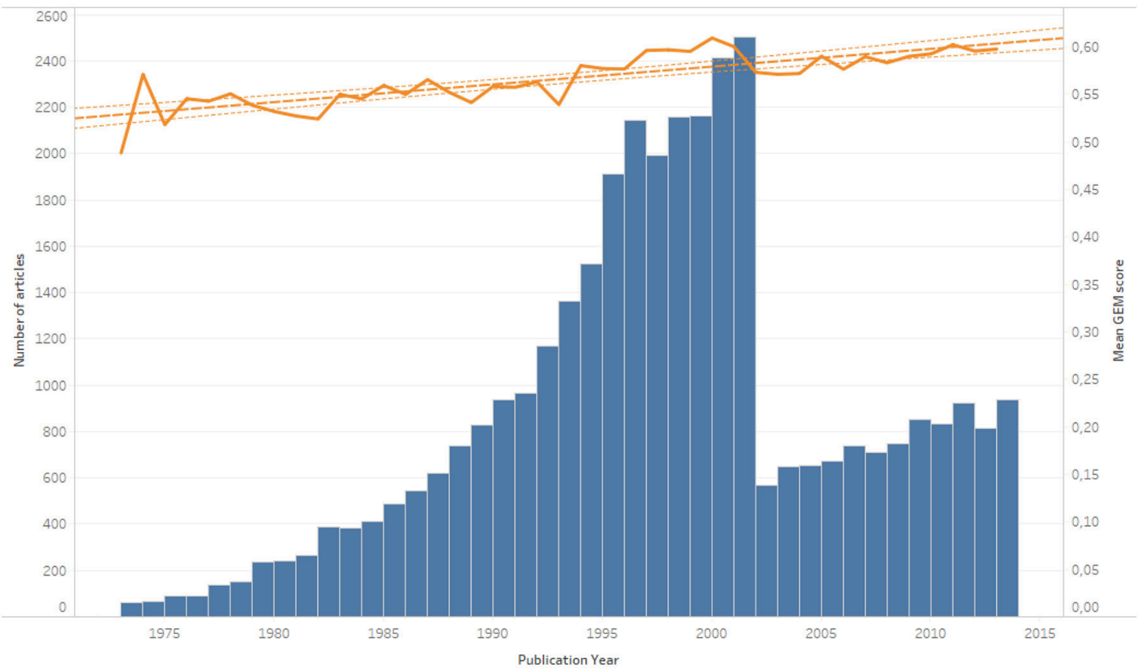


FIGURE 6 | Temporal distribution of the number of articles and mean GEM score (1975–2013).

TABLE 4 | Article distribution across publishers.

Publishers	No. of articles	%
BMJ	912	2.5
De Gruyter Journals	90	0.2
Elsevier	18,236	50.3
Emerald	182	0.5
IOP	398	1.1
RSC	268	0.7
Sage	1,079	3.0
Springer	4,100	11.3
Wiley	10,972	30.3
Total	36,237	100

TABLE 5 | Article distribution across subject areas.

Subject area	Number of articles	%
Medicine	11,342	36.4
Earth and Planetary Sciences	3,575	11.5
Social Sciences	3,494	11.2
Agricultural and Biological Sciences	2,730	8.8
Chemistry	2,619	8.4
Chemical Engineering	2,134	6.8
Pharmacology, Toxicology, and Pharmaceutics	1,983	6.4
Energy	815	2.6
Engineering	773	2.5
Economics, Econometrics and Finance	514	1.6
Business, Management and Accounting	453	1.5
Nursing	433	1.4
Mathematics	117	0.4
Immunology and Microbiology	71	0.2
Arts and Humanities	70	0.2
Psychology	38	0.1
Materials Science	24	0.1
Decision Sciences	7	0.0
Total	31,192	100

open access to their work. There are two routes for achieving open and unrestricted access: the green and the gold routes. The green route is based on the idea of authors making their work publicly accessible by depositing their manuscripts in a repository, or freely-accessible database. Under the gold route, publications are made open access through publishers' websites. We found no significant difference between mean GEM scores for open access articles (0.57) and non-open access articles (0.58), even with the most recently published articles in the dataset (see **Table 6**).

There is clearly not a perfect correlation between the GEM score and the mean citation rate (see **Table 7**), but it should be noted that the lowest citations rates were for the articles with the lowest scores ($\leq$0.4).

CONCLUSION AND PERSPECTIVES

In this paper we introduce the notion of generosity of an abstract in relation to the full text that it is supposed to summarize. We developed this concept with a user study (an online questionnaire) in which we questioned researchers.

We propose a new, completely automatic, measure of abstract generosity with absolute values in the interval [0,1], which

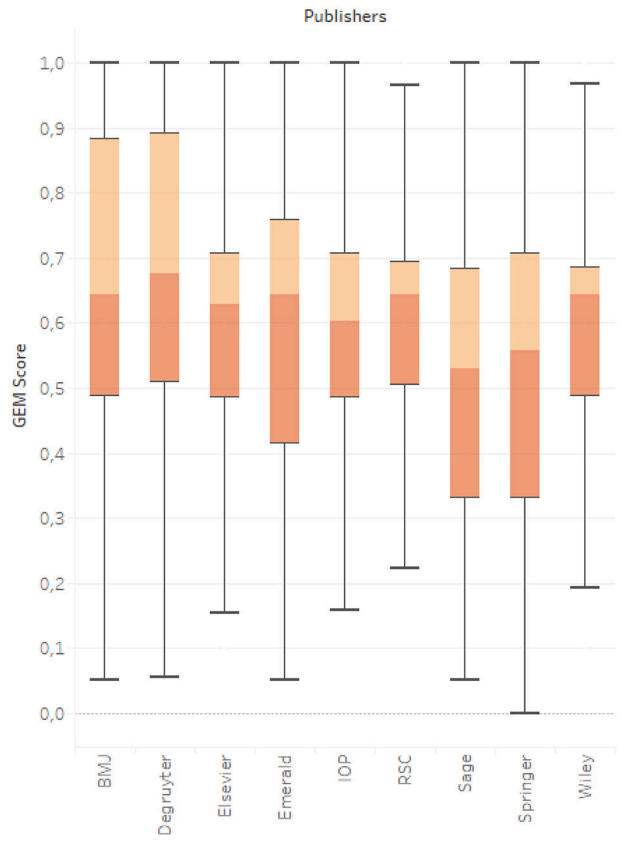


FIGURE 7 | Boxplot for GEM score distribution across publishers.

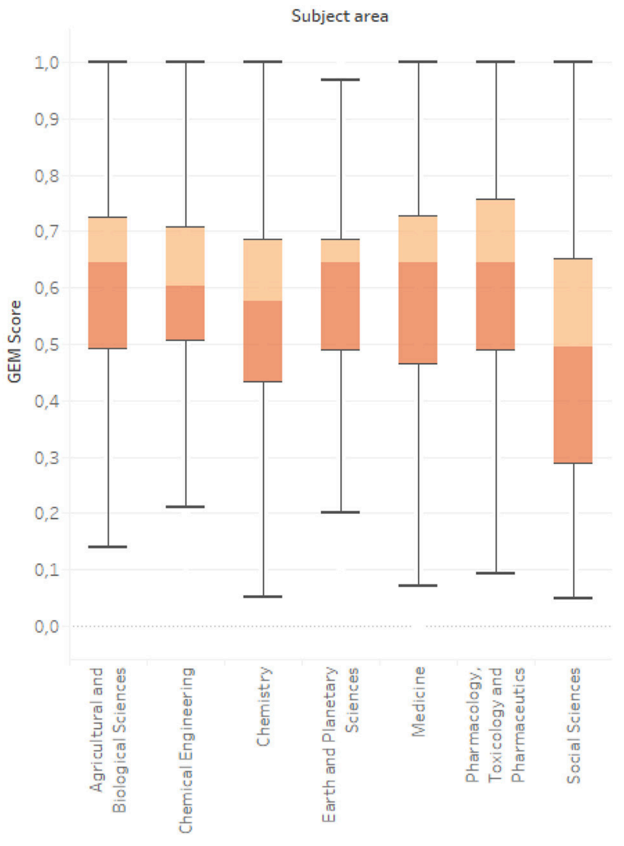


FIGURE 8 | Boxplot for GEM score distribution across subject areas.

differs from the state-of-the-art informativeness metrics. Our score (GEM) considers the importance of different sections by introducing the weighting of sections from the full text that match with sentences in the abstract. The accuracy of section splitting and section classification compared with human judgment is above 80%. The error rate of the GEM score compared with scores assigned by experts is not entirely satisfactory but it could be better with improvements to the GEM formulation.

GEM scores show differences among publishers and subject areas based on the analysis of a large corpus in the environmental sciences.

Our results show that GEM scores have increased over time. The evolution of scores over time is consistent with a codification in the writing of articles. The IMRaD structure, which was widely adopted in health sciences journals in the 1980s (Sollaci and Pereira, 2004), was pioneering in the growing use of standards and reporting guidelines developed in the 1990s through 2010s.

Results suggest that abstracts are more generous in recent publications than earlier ones and cannot be considered as mere teasers. These findings are consistent with those of the questionnaire: when asked about the abstract, 74% of respondents considered it as a summary while only 26% considered it a teaser. The questionnaire results provide also section importance weightings, a unique and very useful information.

One of the possible improvements of the proposed measure is to revise the rules we used for section classification to include regular expressions. We also need to supplement the list of words used in the latter. Another means of improvement would be to learn section weights from an annotated corpus.

This research does start the process of measuring the quality of an abstract. It could be taken further, in particular by exploiting structured abstracts that are included in the dataset. It would be interesting to calculate GEM scores for such abstracts, which have a structure imposed by the journals or publishers, and to compare them with those written without guidelines.

Recommendation systems have emerged recently because document databases enable learning from usage. A user can hence define by their own usage a small pool of interesting documents from which recognition will be made for language modeling (Beel et al., 2016). The proposed measure, based on a series of choices made by author(s) and reader(s), is user-oriented. Following our preliminary results, we suggest that GEM score could be a promising recommendation concept and approach. It could be a valuable indicator in exploring a

TABLE 6 | Mean GEM score and open access status over two time periods.

oaDOI results	All articles (1930–2013)			Most recent articles (2010–2013)		
	Mean GEM Score	Number of articles	%	Mean GEM Score	Number of articles	%
No DOI	0.53	43	0.1	–	0	0
No info	0.58	2,406	6.6	0,64	89	2.5
Not OA	0.58	31,371	86.6	0,6	2,459	70.1
OA	0.57	2,417	6.7	0,58	959	27.3
Total	0.58	36,237	100	0,6	3,507	100

TABLE 7 | GEM score and mean citation rate.

GEM score	Mean citation rate	Number of articles
[0.9;1]	32	2,900
[0.8;0.9]	33	2,562
[0.7;0.8]	34	2,759
[0.6;0.7]	34	8,576
[0.5;0.6]	35	4,825
[0.4;0.5]	33	4,230
[0.3;0.4]	31	1,927
[0.2;0.3]	29	2,443
[0.1;0.2]	27	971
[0;0.1]	28	1,084

large amount of documents by guiding the reader in his/her choices. It could also be a valuable indicator for exploring a large number of abstracts by guiding the reader in his/her choice of whether to obtain the full text to read it or not. Combined with price information, it could also provide useful information for researchers who have very limited access to journal subscriptions from their institutions and who are thus forced to purchase individual articles on a limited budget.

ETHICS STATEMENT

An ethics approval was not required as per the Institution's guidelines and national regulations and the consent of the participants was obtained by virtue of survey completion.

AUTHOR CONTRIBUTIONS

MN, FB, and NT initiated the study and designed the work with LE. LE collected the data and wrote the code. LE and FB made the calculations. FB and MN designed the online questionnaire. LE, FB, and MN participated equally in the analysis of the results, the drawing of conclusions and the writing of most of the manuscript. NT contributed in the state-of-the-art.

FUNDING

The work was supported by the ISTEX project (reference: ANR-10-IDEX-0004, Chantiers d'usage program) under the acronym FULLAB. LE was funded through a postdoctoral scholarship at the Université de Lorraine in partnership with the LISIS.

REFERENCES

Atanassova, I., Bertin, M., and Larivière, V. (2016). On the composition of scientific abstracts. *J. Document.* 72, 636–647. doi: 10.1108/JDOC-09-2015-0111

Bangalore, S., Rambow, O., and Whittaker, S. (2000). "Evaluation metrics for generation," in *Proceedings of the First International Conference On Natural Language generation* (Mitzpe Ramon), 1–8.

Beel, J., Gipp, B., Langer, S., and Breitinger, C. (2016). Research-paper recommender systems: a literature survey. *Int. J. Digit. Libr.* 17, 305–338. doi: 10.1007/s00799-015-0156-0

Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., and Tannier, X. (2016). INEX tweet contextualization task: evaluation, results and lesson learned. *Inform. Process. Manage.* 52, 801–819. doi: 10.1016/j.ipm.2016.03.002

Blaschke, C., Andrade, M. A., Ouzounis, C. A., and Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 7, 60–67.

Bordignon, F., and Ermakova, L. (2018). Data for: 'Is the abstract a mere teaser? Evaluating generosity of article abstracts in the environmental sciences' 1. doi: 10.17632/j39gjcjz5p.1

Bordignon, F., and Noël, M. (2018). Données d'enquête pour la construction d'un indice de générosité des abstracts 1. doi: 10.17632/43trgycgmh.1

Cabrera,-D., Adrián, L., Torres-Moreno, J.-M., and Durette, B. (2016). "Evaluating multiple summaries without human models: a first experiment with a trivergent model," in *Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings*, eds E. Métais, F. Meziane, M. Saraee, V/ Sugumaran, and S. Vadera (Cham: Springer International Publishing), 91–101. doi: 10.1007/978-3-319-41754-7_8

Callon, M., and Latour, B. (1991). *La science Telle Qu'elle se Fait*. Paris: La Découverte.

Campr, M., and JeŽek, K. (2015). "Comparing semantic models for evaluating automatic document summarization," in *Text, Speech, and Dialogue: 18th International Conference, TSD 2015, Pilsen,Czech Republic, September 14-17, 2015, Proceedings*, eds P. Král and V. Matoušek (Cham: Springer International Publishing), 252–260. doi: 10.1007/978-3-319-24033-6_29

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms, 3rd Edn*. Cambridge, MA; London, UK: The MIT Press.

Corney, D. P., Buxton, B. F., Langdon, W. B., and Jones, D. T. (2004). BioRAT: extracting biological information from full-length papers. *Bioinformatics* 20, 3206–3213. doi: 10.1093/bioinformatics/bth386

Crosnier, E. (1993). L'abstract scientifique anglais - français : contraintes et libertés. *ASp. Rev. GERAS* 2, 177–198. doi: 10.4000/asp.4287

Erkan, G., and Radev, D. R. (2004). LexRank: graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.* 22, 457–479.

Ermakova, L. (2018). GEM: measure of the generosity of the abstract comparing to the full text. doi: 10.5281/zenodo.1162951

Elsevier (2015). *Extracting Value from Scientific Literature: The Power of Mining Full-Text Articles for Pathway Analysis Harnessing the Power of Content*. Available online at: https://www.elsevier.com/__data/assets/pdf_file/0016/83005/R_D-Solutions_Harnessing-Power-of-Content_DIGITAL.pdf

Fontelo, P., Gavino, A., and Sarmiento, R. F. (2013). Comparing data accuracy between structured abstracts and full-text journal articles: implications in their use for informing clinical decisions. *Evid. Based Med.* 18, 207–211. doi: 10.1136/eb-2013-101272

Gholamrezazadeh, S., Salehi, M. A., and Gholamzadeh, B. (2009). "A comprehensive survey on text summarization systems," *2nd International Conference on Computer Science and Its Applications* (Jeju), 1–6.

Guerini, M., Pepe, A., and Lepri, B. (2012). "Do linguistic style and readability of scientific abstracts affect their virality?," in *ArXiv:1203.4238 [Cs]. Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM 2012)* (Dublin). Available online at: http://arxiv.org/abs/1203.4238

Hartley, J. (2004). Current findings from research on structured abstracts. *J. Med. Libr. Assoc.* 92, 368–371. doi: 10.3163/1536-5050.102.3.002

Hovy, E., and Tratz, S. (2008). "Summarization evaluation using transformed basic elements," in *Proceedings TAC 2008* (Gaithersburg, MD).

Johnson, F. (1995). Automatic abstracting research. *Libr. Rev.* 44, 28–36. doi: 10.1108/00242539510102574

Kafkas, S., Dunham, I., and McEntyre, J. (2017). Literature evidence in open targets - a target validation platform. *J. Biomed. Seman.* 8:20. doi: 10.1186/s13326-017-0131-3

Khedri, M., Heng, C. S., and Ebrahimi, S. F. (2013). An exploration of interactive metadiscourse markers in academic research article abstracts in two disciplines. *Discour. Stud.* 15, 319–331. doi: 10.1177/1461445613480588

Klein, M., Broadwell, P., Farb, S. E., and Grappone, T. (2016). "Comparing published scientific journal articles to their pre-print versions," in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries* (New Jersey, NJ) 153–162. doi: 10.1145/2910896.2910909

Lin, C.-Y. (2004). "ROUGE: a package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop* (Barcelona).

Lin, J. (2009). Is searching full text more effective than searching abstracts? *BMC Bioinformatics* 10:46. doi: 10.1186/1471-2105-10-46

Louis, A., and Nenkova, A. (2013). Automatically assessing machine summary content without a gold standard. *Comput. Linguist.* 39, 267–300. doi: 10.1162/COLI_a_00123

Mann, W. C., and Thompson, S. A. (1988). Rhetorical structure theory: toward a functional theory of text organization. *Text Interdiscipl. J. Study Disc.* 8, 243–281. doi: 10.1515/text.1.1988.8.3.243

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. New York, NY: Cambridge University Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Baltimore, MD), 55–60. Available online at: http://www.aclweb.org/anthology/P/P14/P14-5010.

Myers, G. (1985). Texts as knowledge claims: the social construction of two biology articles. *Soc. Stud. Sci.* 15, 593–630. doi: 10.1177/030631285015004002

Narine, L., Yee, D. S., Einarson, T. R., and Ilersich, A. L. (1991). Quality of abstracts of original research articles in CMAJ in 1989. *Canad. Med. Assoc. J.* 144:449.

Nenkova, A., Passonneau, R., and McKeown, K. (2007). The pyramid method: incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.* 4:4. doi: 10.1145/1233912.1233913

Ng, J.-P., and Abrecht, V. (2015). "Better summarization evaluation with word embeddings for ROUGE," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1925–1930 (Lisbon: Association for Computational Linguistics). Available online at: http://aclweb.org/anthology/D15-D1222

Orasan, C. (2001). "Patterns in scientific abstracts," in *Proceedings of Corpus Linguistics 2001 Conference*, eds P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja (Lancaster), 433–443.

Owczarzak, K., Conroy, J. M., Dang, H. T., and Nenkova, A. (2012). "An assessment of the accuracy of automatic evaluation in summarization," in *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization* (Stroudsburg, PA: Association for Computational Linguistics), 1–9. Available online at: http://dl.acm.org/citation.cfm?id=2391258.2391259

Perelman, C., and Olbrechts-Tyteca, L. (1958). *Traité de L'argumentation. Logos (Bucureşti. 1996)*, T. *1*. Presses Universitaires de France. Available online at: https://books.google.fr/books?id=CEA6RAAACAAJ

Prasad, S., Lee, D. J., Yuan, C., Barao, V. A., Shyamsunder, N., and Sukotjo, C. (2012). Discrepancies between Abstracts Presented at International Association for Dental Research Annual Sessions from 2004 to 2005 and Full-Text Publication. *Int. J. Dent.* 2012, 1–7. doi: 10.1155/2012/859561

Piringer, G., and Steinberg, L. J. (2008). Reevaluation of energy use in wheat production in the United States. *J. Indus. Ecol. 10*, 149–167. doi: 10.1162/108819806775545420

Radev, D. R., and McKeown, K. R. (1998). Generating natural language summaries from multiple on-line sources. *Comput. Linguist. Spec. Iss. Nat. Lang. Generat.* 24, 470–500.

Radev, D., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Elebi, A., et al. (2002). *Evaluation of Text Summarization in a Cross-Lingual Information Retrieval Framework*. Baltimore, MD: Center for Language and Speech Processing, Johns Hopkins University.

Robertson, S., Zaragoza, H., and Taylor, M. (2004). "Simple BM25 extension to multiple weighted fields," in *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM'04 (New York, NY: ACM), 42–49. doi: 10.1145/1031171.1031181

Schmid, E., Panzer, W., Schlattl, H., and Eder, H. (2012). Emission of fluorescent x-radiation from non-lead based shielding materials of protective clothing: a radiobiological problem? *J. Radiol. Protect. 32*, N129–N139. doi: 10.1088/0952-4746/32/3/N129

Seki, Y. (2005). Automatic summarization focusing on document genre and text structure. *ACM SIGIR Forum* 39, 65–67. doi: 10.1145/1067268.1067294

Shah, P. K., Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2003). Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics* 4:20. doi: 10.1186/1471-2105-4-20

Sharma, S., and Harrison, J. E. (2006). Structured abstracts: do they improve the quality of information in abstracts? *Am. J. Orthodont. Dentofac. Orthoped.* 130, 523–530. doi: 10.1016/j.ajodo.2005.10.023

Sollaci, L. B., and Pereira, M. G. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *J. Med. Libr. Assoc.* 92, 364–367.

Sørensen, T. J. (1948). *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. Biologiske Skrifter. I kommission hos E. Munksgaard. Available online at: https://books.google.fr/books?id=rpS8GAAACAAJ

Tanimoto, T. T. (1958). *An Elementary Mathematical Theory of Classification and Prediction*. International Business Machines Corporation. Available online at: https://books.google.de/books?id=yp34HAAACAAJ

Teufel, S., and Kan, M.-Y. (2011). "Robust argumentative zoning for sensemaking in scholarly documents," in *Advanced Language Technologies for Digital Libraries*, eds R. Bernardi, S. Chambers, B. Gottfried, F. Segond, and I. Zaihrayeu, Vol. 6699 (Berlin; Heidelberg: Springer), 154–170. doi: 10.1007/978-3-642-23160-5_10

Teufel, S., Siddharthan, A., and Batchelor, C. (2009). "Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics," in *EMNLP ′09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Vol. 3* (Stroudsburg, PA), 1493–1502. doi: 10.3115/1699648.1699696

Timmer, A., Sutherland, L. R., and Hilsden, R. J. (2003). Development and evaluation of a quality score for abstracts. *BMC Med. Res. Methodol.* 3:2. doi: 10.1186/1471-2288-3-2

Toulmin, S. E. (2003). *The Uses of Argument*. Cambridge: Cambridge University Press.

Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen, L., J., and Brunak, S. (2017). Text mining of 15 million full-text scientific articles. *BioRxiv*. Available online at: http://www.biorxiv.org/content/early/2017/07/11/162099

Zhang, C., and Liu, X. (2011). Review of James Hartley's research on structured abstracts. *J. Inform. Sci.* 37, 570–576. doi: 10.1177/0165551511420217

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Research Metrics and Analytics

ORIGINAL RESEARCH
published: 15 June 2018
doi: 10.3389/frma.2018.00019

The Termolator: Terminology Recognition Based on Chunking, Statistical and Search-Based Scores

Adam L. Meyers[1]*, *Yifan He*[1], *Zachary Glass*[1], *John Ortega*[1], *Shasha Liao*[2], *Angus Grieve-Smith*[3], *Ralph Grishman*[1] and *Olga Babko-Malaya*[4]

[1] Department of Computer Science, New York University, New York, NY, United States, [2] Google Inc., Mountain View, CA, United States, [3] Department of Information Technology, Columbia University, New York, NY, United States, [4] BAE Systems, Burlington, MA, United States

OPEN ACCESS

Edited by:
Philipp Mayr,
Leibniz Institut für
Sozialwissenschaften (GESIS),
Germany

Reviewed by:
Chengzhi Zhang,
Nanjing University of Science and
Technology, China
Isola Ajiferuke,
University of Western Ontario, Canada

***Correspondence:**
Adam L. Meyers
meyers@cs.nyu.edu

Received: *09 February 2018*
Accepted: *15 May 2018*
Published: *15 June 2018*

Citation:
Meyers AL, He Y, Glass Z, Ortega J, Liao S, Grieve-Smith A, Grishman R and Babko-Malaya O (2018) The Termolator: Terminology Recognition Based on Chunking, Statistical and Search-Based Scores. Front. Res. Metr. Anal. 3:19. doi: 10.3389/frma.2018.00019

The Termolator is an open-source high-performing terminology extraction system, available on Github. The Termolator combines several different approaches to get superior coverage and precision. The in-line term component identifies potential instances of terminology using a chunking procedure, similar to noun group chunking, but favoring chunks that contain out-of-vocabulary words, nominalizations, technical adjectives, and other specialized word classes. The distributional component ranks such term chunks according to several metrics including: (a) a set of metrics that favors term chunks that are relatively more frequent in a "foreground" corpus about a single topic than they are in a "background" or multi-topic corpus; (b) a well-formedness score based on linguistic features; and (c) a relevance score which measures how often terms appear in articles and patents in a Yahoo web search. We analyse the contributions made by each of these components and show that all modules contribute to the system's performance, both in terms of the number and quality of terms identified. This paper expands upon previous publications about this research and includes descriptions of some of the improvements made since its initial release. This study also includes a comparison with another terminology extraction system available on-line, Termostat (Drouin, 2003). We found that the systems get comparable results when applied to small amounts of data: about 50% precision for a single foreground file (*Einstein's Theory of Relativity*). However, when running the system with 500 patent files as foreground, Termolator performed significantly better than Termostat. For 500 refrigeration patents, Termolator got 70% precision vs. Termostat's 52%. For 500 semiconductor patents, Termolator got 79% precision vs. Termostat's 51%.

Keywords: terminology extraction, terminology, technology forecasting, information extraction, multiword expressions

INTRODUCTION

Automatic terminology extraction systems aim to collect word sequences to be used as Information Retrieval key words, terms to be included in domain-specific glossaries or ontologies. Terms are also tracked by technology forecasting applications and are potential arguments of information extraction relations. Terminology extraction systems such as the ones described in Damerau (1993), Drouin (2003), Navigli and Velardi (2004), and others find terminology by comparing the distribution of potential terms in foreground and background corpora, where a foreground corpus consists of text that is about some topic of interest and a background corpus consists of varied documents about all different topics. Potential terms being considered can be single words, bigrams, other n-grams or a constituent type such as a noun groups (Justeson and Katz, 1995).

This paper describes ***the Termolator***, an open source terminology extraction system available on Github[1]. We build on our previous Termolator papers (Meyers et al., 2014a, 2015), adding subsequent improvements (caching information for efficiency, an improved stemming procedure) and additional evaluation experiments, including a comparison to Termostat, another terminology extraction program (Drouin, 2003). The Termolator selects the terms (scientific noun sequences) that are characteristic of a particular technical area. The system identifies all potential instances of terminology in sets of files using a sequential pattern matching process called chunking. Our chunker is similar to the noun group chunkers used in many natural language processing systems, but includes additional constraints so that the selected noun group chunks must contain words belonging to specialized vocabulary classes including: out-of-vocabulary words, nominalizations, technical adjectives, and others. To find chunks that are characteristic of a topic, the system compares the frequencies of particular terms in 2 sets of documents: the foreground corpus (documents about a single topic) and the background corpus (documents about a mixture of topics). It uses several statistical measures to make this determination including Document Relevance Document Consensus or DRDC (Navigli and Velardi, 2004), Term Frequency-Inverse Document Frequency (TFIDF, Spärck Jones, 1972) and Kullback-Leibler Divergence or KLD (Cover and Thomas, 1991; Hisamitsu et al., 1999). For each foreground set of documents, the system produces a list of terms, which is initially ordered based on the distributional means just described. Two other types of scores are factored in to the system's ranking: a well-formedness score based on linguistic constraints, and a relevance score, based on how often a Yahoo (https://search.yahoo.com) web-search results for that term point to patents or articles. The final ranking is used to extract the top terms. We have found that given about 5000 foreground documents and 5,000 background documents, we can generate about 5,000 terms that are approximately 80–85% correct. The system has been tested on US patents, Web of Science abstracts, Open American National Corpus documents (http://www.anc.org/data/oanc/), books from project Gutenberg (https://www.gutenberg.org/) and English journal articles from the PubMed Central corpus (http://www.ncbi.nlm.nih.gov/pmc/). We have implemented some of these components of a Chinese version of the system and are considering developing a system for Spanish for future work. Many other terminology extraction systems, mentioned throughout this paper, also compare the distribution of potential terms in a foreground corpus with a background in order to select characteristic terms. The main things that make Termoloator different are: our particular chunking method for selecting potentential terms (other systems use single words, n-grams or standard noun groups); and our reranking (or filtering methods). Thus Termolator combines the advantages of knowledge-based and statistical techniques to produce superior results.

SYSTEM DESCRIPTION (ENGLISH)

System Overview

As depicted in **Figure 1**, Termolator runs in three stages: (1) terminological chunking and abbreviation; (2) distributional ranking; and (3) filtering (or reordering). The first stage identifies instances of potential terms in text. The second stage orders the terms according to their relative distribution in the foreground and background corpora. The final stage reorders the top N terms from the second stage based on a well-formedness metric and a relevance metric[2]. The so-called filtering criteria sometimes simply rule-out terms completely, and other times they change their ranking in the term list[3]. The assumption behind the ranking is that the higher ranked terms are preferred over lower ranked ones in three respects: (1) higher ranked terms are less likely to be errors (ill-formed as noun groups) and less likely to be "normal" noun sequences, phrases that are part of the general vocabulary, rather than specialized vocabulary (aka terminology); (2) higher ranking terms tend to be more characteristic of a particular field of interest than lower ranking terms; and (3) higher ranking terms tend to have greater relevance than the low ranking ones, i.e., specialists and others are currently more interested in the concepts represented by the high ranking terms.

Stage 1: Terminological Chunking and Abbreviation

In this section, we describe the component of our system designed for identifying terms in sentences, independent of their distribution in sets of documents. Like Justeson and Katz (1995), we assume that most instances of terminology are noun

[1] Termolator's NYU website: http://nlp.cs.nyu.edu/termolator/ English System: https://github.com/AdamMeyers/The_Termolator/ Chinese System: https://github.com/ivanhe/termolator/

[2] There are actually two parameters to determine the cutoff of the terms considered for the third stage. There is a top N parameter (which defaults to 30,000) and a top P percent parameter (which defaults to 30% of the initial term list). P% of the entire list is considered unless it exceeds N terms, in which case we just use N terms. Our defaults assume that the lowest 70% of a ranked list of terms are likely to be of low quality. At the same time, for our purposes we rarely need to look at more than 30 K terms.

[3] For example, a score of zero in any of the metrics will cause the term to simply be ruled out, whereas a higher ranking may cause it to be more preferred or less preferred.

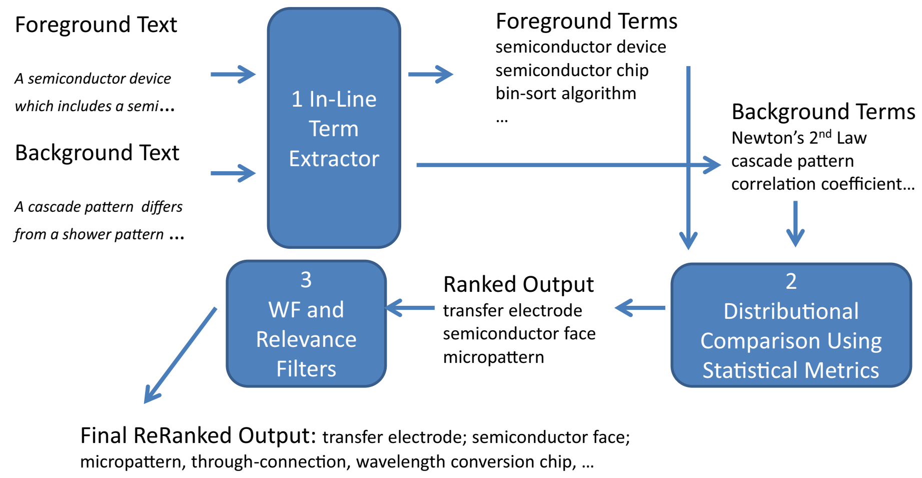

Foreground Text
A semiconductor device which includes a semi…
Background Text
A cascade pattern differs from a shower pattern …
1 In-Line Term Extractor
Foreground Terms
semiconductor device
semiconductor chip
bin-sort algorithm
…
Background Terms
Newton's 2nd Law
cascade pattern
correlation coefficient…
2 Distributional Comparison Using Statistical Metrics
Ranked Output
transfer electrode
semiconductor face
micropattern
3 WF and Relevance Filters
Final ReRanked Output: transfer electrode; semiconductor face; micropattern, through-connection, wavelength conversion chip, …

FIGURE 1 | Termolator system overview.

groups, head nouns and pre-modifiers other than determiners. Consequently, we currently exclude non-noun instances of terminology (verbs like *calcify* or *coactivate*; adjectives like *covalent* or *model-theoretic* and adverbs like *deterministically* or *stochastically*). Unlike previous approaches, we consider only a subset of noun groups as we adapt a more stringent set of chunking rules than used for standard noun group detection. We also identify an additional set of terms by means of rules for identifying abbreviations. We call these terms in-line terms, as this stage is geared toward finding instances of term tokens in documents, rather than identifying classes of terms (types) across a set of documents (the larger task of the full-system)[4].

Terminology Chunking

We incorporate into our chunking rules requirements that constituents contain nominalizations, out of vocabulary words, technical adjectives and other classes of a more fine-grained nature than typical parts of speech used in noun chunking. Nominalizations, such as *amplification* and *radiation* are identified and classified using the NOMLEX_PLUS dictionary (Macleod et al., 1998; Meyers et al., 2004)[5], contributing to the ranking of the terms *optical amplification medium fiber* and *optical radiation*. Out of vocabulary words (e.g., *photoconductor* and *collimate*) are words not found in the lexicon COMLEX Syntax (Macleod et al., 1997), thus selecting terms like *electrophotographic photoconductor* and *optical collimate*[6]. Technical adjectives are adjectives found in COMLEX or classified by a POS tagger that end in *-ic, -cal*, or *–ous,* but are not part of a manually selected out-list (e.g., *public*, *jealous*)[7]. The chunking component is modeled as a finite state machine (FSM) using a fine-grained set of parts of speech (FPOS) to determine transitions between **Beginning**, **Ending**, **Inside,** and **Outside** states in the style of Ramshaw and Marcus (1995). These noun chunks are sequences of these categories. The rules omit preceding determiners, normal adjectives and other words that are not likely to be parts of instances of terminology[8]. The FSM identifies potential terms (PTs). PTs that meet an additional set of constraints are marked as in-line terms. The FSM uses the following FPOS tags:

- **Adjectives**, words with POS tags JJ, JJR or JJS, are subdivided into:
  - ○ **TECH-ADJ**: If an adjective ends in a suffix indicating (-*ic, -cous, -xous*, and several others) it is a technical word, but it is not found in our list of exceptions, it is marked TECH-ADJ.
  - ○ **NAT-ADJ**: An adjective, usually capitalized, that is the adjectival form of a country, state, city or continent, e.g., *European, Indian, Peruvian, …*
  - ○ **CAP-ADJ**: Adjective with the first letter capitalized (but not NAT-ADJ).
  - ○ **ADJ**: Other adjectives

[4] We identify small number of additional term types, specifically chemical formulas and gene sequences, using regular expressions.

[5] NOMLEX-PLUS is described in Meyers et al. (2004). It extends the original Nomlex lexicon described in Macleod et al. (1998).

[6] We have found the word list in COMLEX to be a reasonably good filter for identifying in-vocabulary words. For some domains, we have had to supplement with dictionaries of special in-vocabulary words, words that we treat as out-of-vocabulary, even though they are in COMLEX. For example, we have a dictionary of chemical names, which we always use. We also have a legal dictionary, which we are experimenting with for the legal domain (e.g., court decisions). If extended to social media, we of course would have to add additional dictionaries as well. For the most part, however, mostly words that don't occur in COMLEX tend to be genuine neologisms. The "basic" lexicon of the language actually changes very slowly.

[7] There are 1,445 adjectives in COMLEX with these endings, so it was possible to quickly go through these by eye in a few hours. All but 237 of these adjectives were deemed to be technical.

[8] This set of constraints is based on informal observations of the composition of valid terms in corpora. We validate this set of constraints by showing that results that are constrained this way have higher scores than results that are not so constrained, as discussed below in the Evaluation section.

- **Nouns** are marked NN or NNS by the POS tagger and are the default POS for out of vocabulary (OOV) words. POS tags like NNP, NNPS, and FW (proper nouns and foreign nouns) are not reliable for our POS tagger (trained on news) when applied to patents and technical articles. So NOUN is also assumed for these. Subclasses include:
    - **O-NOUN**: (Singular or plural) nouns not found in any of our dictionaries (COMLEX Syntax plus some person names) or nouns found in lists of specialized vocabulary which currently include chemical names.
    - **PER-NOUN**: Nouns beginning with a capital that are in our dictionary of first and last names.
    - **C-NOUN**: Nouns with POS NN that are not marked O-NOUN or PER-NOUN. A subset of these are nominalizations, a distinction used by constraints applied to the output of the FSM.
    - **PLUR-NOUN**: Nouns with POS NNS nouns that are not marked O-NOUN or PER-NOUN. These include plurals of nominalizations.
- **Verbs that can be modifiers:**
    - **ING-VER**B—verbs marked VBG. These verbs ending in –ing can function as head nouns and can pre-modify nouns.
    - **EN-VERB**—verbs marked VBN and VBD. Past-participles can pre-modify nouns like adjectives. Although these are normally marked VBN, we assume that VBD is a common POS tagging error when past tense and past participles share the same form of a given verb (e.g., *cooked* can be either VBN or VBD).
- **POSS**: Part of speech of the *'s*, separated from a possessive noun by the POS tagger.
- **PREP**: All prepositions (POS IN and TO)
- **ROM-NUM**: Roman numerals (I, II, ..., MMM)
- **Other**: The tag used for all other parts of speech, including verbs hat are neither ING-VERBs not EN-VERBS.

The transitions in the FST are represented in **Table 1** The states are: **B-T** (**Beginning of Term**); **I-T** (**Inside Term), E-T (End of Term)**, **O** (**Outside term), S (Start Sentence), and E (End Sentence)**. This finite state machine recognizes potential terms (PTs). A **PT** is a sequence consisting of 1 **B-T**, followed by 0 or more **I-T** and an optional **E-T**. This can be represented by the following context free phrase structure rule:

$$\textbf{Potential Term} \rightarrow \textbf{B}-\textbf{T}\,\textbf{I}-\textbf{T}^{*}\textbf{E}-\textbf{T?} \qquad (1)$$

where the Kleene star (*) means 0 or more instances and the question mark indicates optionality. As per **Table 1**. each transition to a new state is conditioned on combinations of previous FPOS, current FPOS and the previous state. For example, the table suggests that if (i) the previous word is an out of vocabulary noun (O-noun), a common singular noun (C-NOUN) or plural noun (PLUR-NOUN; (ii) the current FPOS is a roman numeral (ROM-NUM); and (iii) the previous chunk tag is either B-T or I-T, then the new chunk tag should be E-T, a transition which could help identify a term like ***GFP-myosin II***.

The PTs recognized by the FSM are filtered out unless they meet several constraints. To be accepted by the system, an in-line **term** must meet all of the following criteria:

1. It must contain at least one noun.
2. It must be more than one character long, not counting a final period.
3. It must contain at least one word consisting completely of alphabetic characters.
4. It must not end in a common abbreviation from a list (e.g., cf., etc., . . .).
5. It must not contain a word that violates a morphological filter, designed to rule out numeric identifiers (patent numbers), mathematical formulas and other non-words. This rules out tokens beginning with numbers that include letters; tokens including plus signs, ampersands, subscripts, superscripts; and tokens containing no alphanumeric characters at all, etc.
6. It must not contain any word from a list of common patent section headings.

Additionally, each in-line term **T** must satisfy at least one of the following conditions:

1. T contains at least one O-NOUN.
2. T consists of at least 4 words, at least 3 of which are either nominalizations (C-NOUNs found in NOMLEX-PLUS: Meyers et al., 2004; Meyers, 2007) or TECH-ADJs.
3. T is a single word, a nominalization at least 11 characters long.
4. T is a multi-word sequence, ending in a common noun and containing a nominalization.

A final filter aims to distinguish named entities from in-line terms. It turns out that named entities, like jargon terms, include many out of vocabulary words. Thus we look for NEs among those PTs that remain after stage 3 and contain capitalized words (a single capital letter followed by lowercase letters). These NE filters are based on manually collected lists of named entities and nationality adjectives, as well as common NE endings. Dictionary lookup is used to assign GPE (ACE's Geopolitical Entity) to New York or American; LOC(ation) to Aegean Sea and Ural Mountains; and FAC(ility) to Panama Canal and Suez Canal. Plurals of nationality words, e.g., Americans are filtered out as non-terms. Terms are filtered by endings typically associated with non-terms, e.g., et al. signals that a potential term is actually a citation to articles. Honorifics (Esq, PhD, Jr, Snr) indicate that a phrase is probably a PER(son) NE. Finally, if at least one of the words in a multi-word term is a first or last person name, we can further filter them by the last word in the phrase. An ORGanization NE is assumed if the last word is ***agency, association, college*** or 65 other words. The words ***Heights, Township, Park,*** and others indicate GPE named entities. ***Street, Avenue,*** and ***Boulevard*** indicate LOC(ation) named entities. It turns out that 2 word capitalized structures including at least one person name are usually either ORG or GPE in our patent

TABLE 1 | State transition table for terminology chunker.

Previous POS	Current POS	Previous state	New state
Anything	POSS, other	Anything	O
O-NOUN, C-NOUN, PLUR-NOUN	ROM-NUM	B-T or I-T	E-T
Anything	PLUR-NOUN, C-NOUN, PER-NOUN, O-NOUN	B-T or I-T	I-T
Anything	ADJ, CAP-ADJ	I-T	I-T
O-Noun	CAP-ADJ, TECH-ADJ, NAT-ADJ	B-T or I-T	I-T
Anything	CAP-ADJ, TECH-ADJ, NAT-ADJ, ING-VERB, ED-VERB, C-NOUN, O-NOUN, PER-NOUN	E-T, O, Start	B-T
TECH-ADJ, NAT-ADJ, ADJ, CAP-ADJ	TECH-ADJ, NAT-ADJ, ADJ, CAP-ADJ	B-T or I-T	I-T
Everything else			O

corpus, and we maintain this ambiguity, but mark them as non-terms[9].

Identifying Terms by Abbreviations

We extract instances of abbreviations and full forms, using pattern matching similar to Schwartz and Hearst (2003) in contexts where a full form/abbreviation pair are separated by an open parentheses, e.g., *Hypertext Markup Language (HTML).* In the simplest case, the abbreviation consists of the initials for each word of the full form (e.g., *SAS* is an abbreviation for Statistical Analysis System), but we also allow for several more complex cases. Abbreviations can skip stop words like *the, a, in, out, and*, others, e.g., *YHL* abbreviates *Years of Healthy Life* (no initial corresponds to the word *of*). Multiple letters can match a single word, e.g., *Hypertext* corresponds to the *HT* of *HTML*). There can be a correspondence between Greek and Roman letters, e.g., *TGF-β* abbreviates *Transforming Growth Factor Beta*). These and other special cases are all accounted for. After establishing a full-form/abbreviation correspondence, we use keyword-based heuristics and gazetteers to differentiate non-terminology abbreviation cases from terminology ones. For example, *New York University (NYU)* and *Acbel Polytech Inc. (API)*, are ruled out as terminology because the words *Inc.* and *University* indicate organizations; *British Columbia (BC)* is ruled out due to a gazetteer. Each term abbreviation (e.g., *html*) and the associated longer term (e.g., *Hypertext Markup Language*) are classified as instances of a single term (*Hypertext Markup Language*) for purposes of subsequent stages.

Summary of Stage 1

Both the terminology chunker and the abbreviation system identify terms in sentences in each document. These instances are collected and output to be used for stage 2. The chunker uses a FSM with the transitions conditioned on FPOS tags, to identify potential in-line terms. Additional filters based on linguistic features are used to identify the final in-line terms. The abbreviation system uses standard patterns to identify instances in the text where a phrase is linked to its corresponding abbreviation, both of which are likely to be either an in-line term or a NE. We use word lists and heuristics to eliminate the NE instances of abbreviations. Selecting these in-line terms is a major differentiation between our approach and other approaches. We find word sequences that are likely to be instances of terms, sequences containing nouns that are too rare to be included in a general purpose dictionary (O-Nouns) and other words that tend to be technical. Additionally, abbreviations are likely to be terms because authors tend to abbreviate important technical phrases. Together, these methods find good candidates for subsequent stages of Termolator. Arguably, this process of term candidate selection is a major differentiator between Termolator and other systems.

Other Details About Stage 1: Compound Terms and Stemming

Compound Terms

The Stage 1 system can combine instances of two adjacent or nearly adjacent inline terms to form *compound terms*. The two smaller terms are combined when they fall into one of the following 2 patterns:

1. There are 1 or 2 words between the first and second term, such that a preposition from the set {*of, for*} immediately follows the first inline term. The preposition is optionally followed by a determiner from the set {*a, the, an*}, e.g., *alignment algorithms for rna secondary structures* is a combination of the inline terms *alignment algorithms* and *rna secondary structures* (a singular form of this same term could include a determiner in the second short inline term as in *alignment algorithm for an rna secondary structure*).
2. The first and second term are one right after the other, e.g., *Post-HF event medical management* is the combination of the inline terms *Post-HF event* and *medical management*.

Both the initial in-line terms and the longer longer compound inline terms are output by the system as potential terms and are treated separately in Stage 2.

Stemming

In Stage 2, the instances of particular terms derived in stage 1 will be "counted." For purposes of counting, equivalences are established between terms that share the same lemma. Thus, we must make some assumptions about which items are regularized to the same lemma. Plural forms of terms are regularized to their singular counterparts, e.g., *Optical Character Recognition*

[9] We are currently experimenting with a modification to the system that allows the user to provide the output of a named entity tagger (or similar program) to block particular types of phrases from being considered as terms.

Systems ➔ *Optical Character Recognition System*, and thus plural and singular forms count as instances of the same term lemma. Given a noun that is also a verb, the –ing form is regularized to the singular noun, e.g., *network modeling* ➔ *network model*. Abbreviations are regularized to the fully spelled out form, e.g., *OCR* ➔ *Optical character Recognition*. Finally, compound terms with the prepositions *for* or *of* are regularized to prenominal noun modifier equivalents. Given a compound term of the form **NP1 preposition NP2**: (1) the determiner is dropped from NP2 and the final noun, if plural is converted to singular form; (2) NP2 is moved before NP1. For example, *Recognition of Optical Characters* is regularized to *Optical Character Recognition*. Thus for statistical purposes, a single lemma *Optical Character Recognition* will be correspond to instances of: *Optical Character Recognition, Optical Character Recognitions, OCR, OCRs,* and *Recognition of Optical Characters*. The output of lemmatization is included in the output of Stage 1, both as information associated with each recognized term and as a dictionary from lemmas to possible phrases that map to these lemmas. The dictionary is used to augment the final set of ranked terms (lemmas) to include the variants of each form, e.g., if *Optical Character Recognition* is in the output list, it would be associated with any variants of the term that actually occur in the input text, a subset of: {*Optical Character Recognition***,** *Optical Character Recognitions*, *OCR*, *OCRs*, *Recognition of Optical Characters*}.

Applications of Stage 1 Output

As discussed in the introduction, the output of stage 1 is the input to stage 2. However, we have found other applications of inline terms, the output of stage 1. We used them as potential arguments of the Information Extraction relations discussed in Meyers et al. (2014b). Some example relations from the PubMed corpus follow:

1. found *in the* ***IκB protein****, an* ***inhibitor of NF-κB***
   - Relation: **Exemplify**, Arg$_1$: ***IκB protein***, Arg$_2$*:* ***inhibitor of NF-κB***
   - Interpretation: Arg$_1$ is an instance of Arg$_2$
2. *a* ***necrotrophic effector system*** *that is an exciting contrast to the* ***biotrophic effector models*** *that have been intensively studied*
   - Relation: **Contrast**, Arg$_1$: ***necrotrophic effector system***, Arg$_2$: ***biotrophic effector models***
   - Interpretation: Arg$_1$ and Arg$_2$ are in contrast with each other
3. ***Bayesian networks*** *hold a considerable advantage over* ***pairwise association tests***
   - Relation**: Better than**, Arg$_1$: ***Bayesian networks,*** Arg$_2$: ***pairwise association tests***
   - Interpretation: Arg$_1$ is better than Arg$_2$ (in some respect)
4. ***housekeeping gene 36B4*** (***acidic ribosomal phosphoprotein P0***)
   - Relation: **Alias**, Arg$_1$: ***housekeeping gene 36B4***, Arg$_2$: ***acidic ribosomal phosphoprotein P0***
   - Interpretation: Arg$_1$ and Arg$_2$ are alternative names for the same concept, but neither is a shortened form (acronym or abbreviation).

Additionally, we have begun some research that uses in-line terms to improve Machine Translation (MT). It hypothesize that it is useful to treat in-line terms (and other fixed phrases like named entities) differently from other source language input. For phrase-based MT, these words are unlikely to be in the phrase table from (general domain) training data; these words are more likely than other words to be translated as themselves in the target language; these words are likely to be translated as single units (the constituent boundaries of the terms should not be interrupted by other translations) and finally, these phrases may correspond to terminology detected in the target language using terminology extraction. We are looking toward using fuzzy-match repair methods for translation of these units, along the lines of Ortega et al. (2016). More generally, inline terms appear to be good candidate entities that represent technical concepts for possibly a large variety of NLP applications.

While Stage 2 provides a way of selecting the "most important" terms for certain applications. Stage 1 provides a way of finding a large subset of terms useful for a variety of other applications, where finding only the most "important" terms is not sufficient[10].

Stage 2: Distributional Ranking

While stage 1 identifies term instances or tokens, stage 2 groups together these tokens into general types, clustering together variants of terms and representing types their common lemmas, e.g., *Optical Character Recognition* is a type that is realized in the actual texts in a variety of ways, as noted above. The term types are returned by the system in the form of a ranked list, ranking terms by how characteristic the terms are to one set of documents about a single topic (foreground), as compared to another set of documents about a diverse set of topics (background). Essentially, a highly ranked (more characteristic) term occurs much more frequently in the foreground than it does in the background. This methodology is based on many previous systems for identifying terminology (Damerau, 1993; Drouin, 2003; Navigli and Velardi, 2004; etc.) which aim to find nouns or noun sequences (N-grams or noun groups) that are the most characteristic of a topic. The output of systems of this type have been used as Information Retrieval key words (Jacquemin and Bourigault, 2003), terms to be defined in thesauri or glossaries for a particular field (Velardi et al., 2001) and terms tracked over time as part of technology forecasting (Daim et al., 2006; Babko-Malaya et al., 2015)[11].

In Stage 2, we rank our terms using a combination of three metrics: (1) a version of the standard Term Frequency Inverse Document Frequency (TFIDF) metric; (2) the Document

[10] Obtaining the inline terms is a relatively fast process, that is dominated in our implementation (timewise) by POS tagging. The later stages of Termolator are more computationally expensive.

[11] In Technology forecasting applications, systems seek to identify patterns of changing terminology usage in corpora divided by topic and by epoch. In principle, given increased usage of particular terminology over a sequence of epochs, one can predict the increasing prominence of a technology associated with that terminology.

Relevance Document Consensus (DRDC) metric (Navigli and Velardi, 2004); and (3) the Kullback-Leibler Divergence (KLD) metric (Cover and Thomas, 1991; Hisamitsu et al., 1999). The TFIDF metric selects terms specific to a domain by favoring terms that occur more frequently in the foreground (abbreviated as *Fore*) documents than they do in the background (abbreviated as *Back*).[12] The formula is:

$$TFIDF\,(t) = \frac{freqFore(t)}{freqBack(t)} * \log\left(\frac{numBackDocs}{numBackDocContains(t)}\right)$$

where freqFore(t) and freqBack(t) respectively refer to the number of times a term occurs in the foreground and background corpora. The first term is simply a ration of foreground/background frequencies. The second term is the standard inverse document frequency of a term in the background corpus (number of background documents divided by the total number of such documents containing the term). In the DRDC metric, two factors are considered: (i) document relevance (DR), which measures the specificity of a terminological candidate with respect to the foreground via comparative with the background (the same first term as in TFIDF); and (ii) document consensus (DC), which measures the distributed use of a terminological candidate in the target domain (favoring terms that occur in lots of foreground documents). The formula for DRDC is:

$$DRDC\,(t) = \frac{freqFore(t)}{freqBack(t)} * \sum_{d \in Fore} \frac{freq\,(d,t)}{freqFore\,(t)} * \log\left(\frac{freqFore(t)}{freq(t,d)}\right)$$

The KLD metric measures the difference between two probability distributions: the probability that a term will appear in the foreground corpus vs. the background corpus. The formula is[13]:

$$KLD(t) = (\log(freqFore(t)) - \log(freqBack(t))) * freqFore(t)$$

These three metrics are combined together with equal weights, ranking both the terms produced in stage 1 and substrings of those terms, producing an ordered list.

Stage 2 uses some of the same metrics as previous work, but may achieve different results due to the differences between the stage 1 output (technical noun groups or inline terms) that Termolator uses as opposed to the normal noun groups or bigrams used by previous work. In the Experiments and Evaluation section, we compare some results of running the system using different types of input terms and demonstrate that our inline terms provide better results.

Crucially, the terms that the system outputs depend on the choice of both the foreground and the background document sets. For example, a foreground of surgery patents entails that the output may include surgical terms and/or patent terms. Different backgrounds will result in different subsets of terms. Thus given, surgery patents as foreground and a general non-patent (e.g., news) corpus as background, the output would probably include some terms specific to patents in general, even if they were not related specifically to surgery. However, given a varied set of patent documents as the background, the output terms would probably mostly be about surgical matters and not include general patent terms. This corroborates with some of the experiments described in the Experiment and Results section in which we compare Termolator with Termostat, a terminology extraction system that has a distributional component similar to Termolator's, but currently uses a fixed corpus as its background corpus for all foreground corpora.

Stage 3: Well-Formedness Score and Relevance Score

The previous stages produce a ranked list of terms, the ranking derived from the distributional score, which we normalize to ***D***, a percentile score between 0 and 1. We then combine this score with other scores between 0 and 1. We multiply all the 0–1 scores together to produce a new percentile ranking. Weights can be applied as exponents on each of the scores, resulting in one aggregate score that we use for reranking the terms. However, we currently assume all weights to equal 1. We assume 2 scores, in addition to D: *W*, a well-formedness score and ***R***, a relevance score. The aggregate score which we use for reranking purposes is simply: ***D* W* R.*** Like stage 1, the stage 2 components (***W*** and ***R***) can be used separately from the other portions of Termolator, to score or rank terms entered by a user, e.g., terms produced by other terminology extraction systems.[14]

Well-Formedness Score

Our well-formedness (W) score is based on several linguistic rules and subjective evaluations about violations of those rules. Many of these linguistic rules are built into the chunking rules in stage 1 and thus the most common score for W is 1 when used as part of Termolator. However, W does contribute to the ranking and eliminates some potential terms with scores of 0 (a 0 score for D, W or R eliminates a term since these scores are combined by multiplication). We assume that applications of the following rules are reason to give a candidate term a perfect score (1.0):

- **ABBREVIATION_OR_TERM_THAT_IS_ABBREVIATED** – This rule matches terms that are either abbreviations or a full length term that has been abbreviated, e.g., *html, hypertext markup language, OCR, optical character recognition*, ...
- **Out_of_Vocabulary_Word** – This rule matches terms consisting of single words (and their plurals) that are not found in our dictionaries, e.g., *radionuclide, photoconductor*, . . .
- **Hyphenated Word + OOV Noun** – This applies if a word contains one or more hyphen and the part of the word following the last hyphen would matches the conditions described in the previous bullet, e.g., *mono-axial, lens-pixel*, ...

[12] In Meyers et al. (2014a, 2015), we refer to Foreground documents as "Related Document Groups," i.e., a group of documents that are related as they are about the same topic. We also referred to some of the numbers referring to counts as total document counts, even though they actually refer to counts in the background documents.

[13] Our KLD function is a simplified version of KL Divergence.

[14] We have used these components to evaluate sets of terms that were not produced by the Termolator as part of the FUSE project. Our subjective analysis is that they can be used effectively in this way to rate or rerank such terms, but a formal evaluation is outside the scope of this paper.

These rules yield a score of **0.7:**

- **Common_Noun_Nominalization** – This means that the term is a single word, identified as a nominalization using dictionary lookup, e.g., *demagnetization, overexposure*,
- **Hyphenated Word + Nominalization** – This applies if a word contains one or more hyphen and the part of the word following the last hyphen would match the conditions described in the previous bullet, e.g., *de-escalation, cross-fertilization*

This rule gives a score of **0.3**:

- **Normal_Common_Noun_or_Number** – This means that the term consists of a single word that is either a number, a common noun, a name or a combination of numbers and letters (e.g., *ripcord, H1D2*).

The following rules have scores that vary, depending on the type of words found in the phrase:

- **Normal_NP** – This means that the term consists of a word sequence that is part of a noun group according to our chunker, described above. The score can be as high as **1.0** if the term contains an OOV words (e.g., *electrophotographic photoconductor* contains two OOV words). A noun group containing one "unnecessary" element such as a preceding adjective, would have a score of **0.5 (*acceptable organic solvent*)**. Other noun groups or noun phrases would have scores of **0.2 (*wheel drive capacity*)**.

There are several other rules which have scores of 0 associated with them including:[15]

- **Single_Word_Non_Noun** – This means that the word is identified as a non-noun, either by dictionary lookup or by simple morphological rules, e.g., we assume that an out of vocabulary word ending in *-ly* is an adverb, e.g., *downwardly, optical, tightening*
- **Bad_character** – This means that the term contains at least one character that is not either: a) a letter; b) a number; c) a space; d) a hyphen; e) a period; or f) an apostrophe, e.g., box$^{\text{TM}}$, *sum_l*, *slope* Δa
- **Contains_conjunction** – This rule matches sequences including coordinate conjunctions (*and, or, but, nor*), e.g., *or reproducing*, *asic or other integrated*
- **Too many verbs** – This means that the sequence contains multiple verbs, e.g., *insulating film corresponding, emitting diodes disposed*
- **Verbal or Sentential Structure** – This means that some chunking rules found a verbal constituent other than an adjective-like pre-modifier (*broken record*), e.g., *developer containing, photoelectric converting*
- **Unexpected_POS_sequence** – This applies to multi-word terms that do not fit any of the profiles above, e.g., *of the developing roll, beam area of the charged*.

In addition to ranking the output of Stage 1, Stage 2 also ranks highly frequent substrings of stage 1, e.g., if *intravascular balloon catheter* and *cannulated balloon catheter* are frequent terms, the system may also recognize that the common substring *balloon catheter* is a frequent term. So one function of W is to rule-out ill-formed substrings by assigning them a score of 0. For example, the noun *balloon* is a substring of *balloon catheter* (and the superstrings noted above), but is not a valid term by itself–it is just a normal, non-technical common noun. So when applied to our own stage 1 terms, ***W*** usually has a value of 1, but it assigns a score of 0 to some substrings. Intermediate values occur less frequently, but may serve to rank terms containing OOV words more highly than those well-formed terms that do not, e.g., *protective shield* has a low score (0.6) because although it is well-formed (the noun *shield* is arguably a nominalization of the verb *shield*), it does not contain any OOV words or other technical words.

Relevance Score

The relevance score is derived by searching for the term using Yahoo's search engine (powered by Microsoft Bing)[16] and applying some heuristics to the search result. This score is intended to measure the "relevance" of a term to technical literature. The Relevance Score $\boldsymbol{R} = HT^2$ where the two factors $\boldsymbol{H}$ and T are defined as follows and the weight on $\boldsymbol{T}$ was determined experimentally:

- $\boldsymbol{H}$ = the total number of hits for an exact match. The log 10 of this number (up to a maximum of 10) is normalized between 0 and 1.
- $\boldsymbol{T}$ = the percentage of the top 10 hits that are either articles or patents

The following information from a Yahoo search are used to compute this score: (1) the total number of hits; (2) a check to see if this result is based on the search or if a similar search was substituted, i.e., if the result includes the phrase ***including results for*** or the phrase ***showing results for***, then we know that our search was not matched at all and we should assume that there are 0 hits; and (3) the top 10 search results as represented by URLs, titles and summaries. If there are fewer than 10 hits, we assume that there are actually 500 hits, when calculating *H*. For each result, we search the URL, title and summary for key words which indicate that this hit is probably an article or a patent (*patent, article, sciencedirect, proceedings, journal, dissertation, thesis, abstract*). ***T*** is equal to the number of these search results that match, divided by 10. In practice, this heuristic seems to capture the intuition that a good term is likely to be the topic of current scientific articles or patents, i.e., that the term is relevant.

Today's web search programs (Google, Bing, etc.) find documents from a query, using a combination of standard information retrieval metrics like TF-IDF and a metric such as PageRank (Page et al., 1998) that measures how prominent

[15] Some additional patterns also yield a score of 0, e.g., terms consisting of a single character.

[16] In theory, a different search engine could be used instead of Yahoo. While we currently use the free version, pay versions could be substituted. In practice, some additional coding may be necessary to make the output of a new search engine compatible with Termolator.

documents are on the web. By using a web search query with our terms, we are indirectly using that search engine's prominence measure (in the current case Yahoo/Microsoft's prominence measure) and, in principle, ranking prominent terms more highly.

Runtime is a limiting factor for the Relevance scores because it takes about 0.75 s to search for each term. This means that producing Relevance scores for 30 K terms takes about 6 h, a substantial portion of the overall runtime.

EXPERIMENTS AND EVALUATION

Stage 1 Annotation and Evaluation

We evaluated Stage 1's inline terms by manually annotating all the instances of inline terms in a few documents and comparing the inline terms annotated by the human annotators with those selected by the system. For purposes of annotation, we defined an (in-line) term as a word or multi-word nominal expression that is specific to some technical sublanguage. It is conventionalized in one of the following two ways:

1. The term is defined early (possibly by being abbreviated) in the document and used repeatedly (possibly only in its abbreviated form).
2. The term is special to a particular field or subfield (not necessarily the field of the document being annotated).

It is not enough if the document contains a useful description of an object of interest– there must be some conventional, definable term that can be used and reused. Thus multi-word expressions that are defined as terms must be somewhat word-like—mere descriptions that are never reused verbatim are not terms. Justeson and Katz (1995) goes further than we do: they require that terms be reused within the document being annotated, whereas we only require that they be reused (e.g., frequent hits in a web search). Criterion 2 leaves open the question of how specific to a genre an expression must be to be considered a jargon-term. At an intuitive level, we would like to exclude words like *patient*, which occur frequently in medical texts, but are also commonly found in non-expert, everyday language. By contrast, we would like to include words like *tumor* and *chromosome*, which are more intrinsic to technical language insofar as they have specialized definitions and subtypes within medical language. To clarify, we posited that a term must be sufficiently specialized so that a typical naive adult should not be expected to know the meaning of the term. We developed 2 alternative models of a naive adult:

1. Homer Simpson, an animated TV character who caricatures the typical naive adult–the annotators invoke the question: Would Homer Simpson know what this means?
2. The Juvenile Fiction sub-corpus of the COCA: The annotators go to http://corpus.byu.edu/coca/ and search under FIC:Juvenile – a single occurrence of an expression in this corpus suggests that it is probably not a jargon-term.

In addition, several rules limited the span of terms to include the head and left modifiers that collocate with the heads. Decisions about which modifiers to include in a term were difficult. However, as this evaluation task came on the heels of the relation extraction task (Meyers et al., 2014b), we based our extent rules on the definitions and the set of problematic examples that were discussed and cataloged during that project. This essentially formed the annotation equivalent of case-law for extents.

For evaluation purposes, we annotated all the instances of inline-terms in a speech recognition patent (SRP), a sunscreen patent (SUP) and a journal article about a virus vaccine (VVA). For purposes of this task, only the longest strings need be detected, e.g., if *cannulated balloon catheter* is recognized, the substring *balloon catheter* need not be annotated separately, even though it is also a valid term. Each document was annotated by 2 people and then adjudicated by Annotator 2 after discussing controversial cases **Table 2** scores annotator 1, annotator 2 and a few versions of the system by comparing each against the answer key. The table includes number of terms in the answer key, number of matches, precision, recall and F-measure. The "strict" scores are based on exact matches between system terms and answer key terms, whereas the "sloppy" scores count as correct instances where part of a system term matches part of an answer key term (span errors). For example, given an answer key item of *cannulated balloon catheter,* the strings *balloon catheter* and *cannulated balloon* would each count as incorrect for purposes of the strict score and correct for purposes of the sloppy score.

As the SRP document was annotated first, some of specification agreement process took place after annotation and the scores for annotators are somewhat lower than for the other documents. However, Annotator 1's scores for SUP and **VVA** are good approximations of how well a human being should be expected to perform and the system's scores should be compared to Annotator 1 (i.e., accounting for the adjudicator's bias).

There are four system results: two baseline systems the results of running the system and two versions of the Stage 1 system: one admitting all potential terms (PTs) and one that filters out some of the terms with the filters described in the Stage 1 chunking section. Baseline 1 assumes terms derived by removing determiners from noun groups – we used an MEMM chunker using features from the GENIA corpus (Kim et al., 2003). That system has relatively high recall, but overgenerates, yielding a lower precision and F-measure than our full system – it is also inaccurate at determining the extent of terms. Baseline 2 restricts the noun groups from this same chunker to those with O-NOUN heads. This improves the precision at a high cost to recall. Next we ran our finite state machine to derive potential in-line terms, but we did not run the subsequent filters, and the final score is for our full system. Clearly our more complex strategy performs better than these baselines and the linguistic filters increase precision more than they reduce recall, resulting in higher F-measures (though low-precision high-recall output may be better for some applications).

Evaluation of Stages 2 and 3

We ran the complete system with 5000 patents about optical systems and components as the foreground (US patent codes 250, 349, 356, 359, 362, 385, 398, and 399) and 5,000 diverse patents as background. We collected a total of 219 K terms,

TABLE 2 | Evaluation of terminology chunking annotation and system output.

			Strict				Sloppy			
	Doc	**Terms**	**Matches**	**Prec**	**Rec**	**F**	**Terms**	**Prec**	**Rec**	**F**
Ann 1	SRP	1131	798	70.8%	70.6%	70.7%	1041	92.5%	92.0%	92.2%
	SUP	2166	1809	87.5%	83.5%	85.5%	1992	96.3%	92.0%	94.1%
	VVA	919	713	90.9%	77.6%	83.7%	762	97.2%	82.9%	89.5%
Ann 2	SRP	1131	960	98.4%	84.9%	91.1%	968	99.2%	85.6%	91.9%
	SUP	2166	1999	95.5%	92.3%	93.8%	2062	98.5%	95.2%	96.8%
	VVA	919	838	97.4%	91.2%	94.2%	855	99.4%	93.0%	96.1%
BL1	SRP	1131	602	24.3%	53.2%	33.4%	968	44.2%	96.8%	60.7%
	SUP	2166	1367	36.5%	63.1%	46.2%	1897	50.6%	87.6%	64.2%
	VVA	919	576	28.5%	62.7%	39.2%	887	44.0%	96.5%	60.4%
BL 2	SRP	1131	66	24.9%	5.8%	9.5%	151	57.0%	13.4%	21.6%
	SUP	2166	771	52.3%	35.6%	42.4%	1007	68.4%	46.5%	55.3%
	VVA	919	270	45.8%	29.4%	35.8%	392	66.5%	42.6%	51.9%
Sys W/O filter	SRP	1131	932	39.0%	82.4%	53.0%	1121	46.9%	99.1%	63.7%
	SUP	2166	1475	39.7%	68.1%	50.2%	1962	52.8%	90.6%	66.7%
	VVA	919	629	27.8%	68.4%	39.5%	900	39.8%	97.9%	56.6%
Full sys	SRP	1131	669	69.0%	59.2%	63.7%	802	82.8%	70.9%	76.4%
	SUP	2166	1193	64.7%	55.1%	59.5%	1526	82.8%	70.5%	76.1%
	VVA	919	581	62.1%	63.2%	62.7%	722	77.2%	78.6%	77.9%

TABLE 3 | System Output with aggregate scores, component scores and correctness judgements.

Rank	**Term**	**D**	**W**	**R**	**Total**	**Correct**
41	Stimulable phosphor	0.866	1	0.174	0.151	Yes
104	Ion beam profile	0.889	1	0.117	0.126	Yes
346	X-ray receiver	0.906	1	0.099	0.089	Yes
533	Wavelength-variable	0.838	1	0.091	0.076	Yes
556	Irradiation time t	0.460	1	0.163	0.075	No
1275	Quadrupole lens	0.460	1	0.113	0.052	Yes
1502	Evolution	0.439	1	0.109	0.048	No
1581	Proximity correction	0.451	1	0.103	0.046	Yes
1613	Dfb laser	0.943	1	0.049	0.046	Yes
1685	Asymmetric stress	0.493	1	0.067	0.033	Yes
3834	Panoramagram	0.483	1	0.056	0.027	Yes
4203	Crystal adjacent	0.316	1	0.080	0.025	No
4244	Single-mode optical fiber	0.875	1	0.029	0.025	Yes
4467	Total reflection plane	0.988	1	0.024	0.024	Yes
4879	Photosensitive epoxy resin	0.286	1	0.079	0.022	Yes

ranked by the stage 2 system. We selected the top 30 K of these terms and ran the stage 3 processes on these 30 K terms. We ranked these top terms 3 different ways, each time selecting a different top 5,000 terms for evaluation. We selected the top 5,000 terms after ranking these 30 K terms in the following ways: (a) according to stage 2 (Distributional Score); (b) according to the Relevance Score (c) according to the Combined Score (D*R*W). As W primarily was used to remove ill-formed examples, it was not well-suited for this test as a separate factor. For each list of 5,000 terms, we sampled 100 terms, took 20 random terms from each 20% interval, manually inspected the output, and rated each term as correct or incorrect. 71% of the terms ranked according to D only were correct; 82% of the terms ranked according to R were correct and 86% of the terms ranked according to the Combined Score were correct. While we believe that it is significant that the combined score produced the best result, it is unclear whether the fact that R alone did better than the stage 2 ranking because the R score was applied to the 30 K terms out of 219 K terms with the highest D scores. While in principle, we could run R on all 219 K terms, time constraints make it impractical to do this, in general, for all output of our system[17].

Coverage of a term extractor is difficult to measure for terms without having a human being do the task, e.g., reading all 5,000 articles and writing out the list of terms[18]. Informally however, we have observed a significant increase in term output since we adopted the chunking model described above, compared to a previous version of the system that used a standard noun chunker. In other words, we are able to take a larger number of top ranked terms than before without a major decline in accuracy. One of the tasks for future work is to develop a good metric for measuring this.

Example Term Output From These Experiments

Table 3 provides some sample potential terms along with scores ***D, W, R*** and the aggregate score. The table is arranged in descending order by the aggregate score. These terms are excerpts from the best of the three rankings described in the previous section, i.e., the terms ordered by the total score. In the right-most column is an indication of whether or not these are valid terms, as per the judgment of one of the authors. The incorrect examples include: (a) ***irradiation time t***, which is really a variable (a particular irradiation time), not a productively used noun group that should be part of a glossary or a key word; (b) ***evolution***, a common word that is part of the general language and should no longer be relegated to a list of specialized vocabulary; and (c) ***crystal adjacent***, a word sequence that does not form a natural constituent – it is part of longer phrases like ***a one-dimensional photonic crystal adjacent to the magneto-optical metal film.*** In this sequence the word ***crystal***, is modified by a long adjectival modifier beginning with the word ***adjacent*** and it would be an error to consider this pair of words a single constituent.

Comparison With Termostat

Termostat (Drouin, 2003) is a terminology extraction tool that is readily available for public use without installation[19]. To our knowledge, Termostat is the only terminology extraction system that is both available for research purposes and that can perform essentially the same task as Termoloator[20].

- There are a number of key differences between Termolator and Termostat which may explain some of the differences in the results presented below:
- Termostat uses a single foreground document about the topic of interest. This is the only input to the system. In contrast, the Termolator uses a set of foreground documents that are about the same topic, e.g., patents that share a patent code; or other documents that are known to share subject matter
- Termostat uses one general purpose background corpus in common. This is part of the system. It does not change for different foreground corpora. In contrast, Termolator expects the user to supply a set of background documents, the documents that the foreground documents should be compared to.
- Both systems use chunking procedures to find candidate terms. The most significant difference is that Termolator's chunking procedure explicitly favors chunks containing OOV and technical words, whereas Termostat relies on standard Part of Speech tags.
- The two systems use different (but similar) distributional measures to rank terms.
- Termolator adds on additional well-formedness and relevance filters.

Termostat is easy to run. One simply uploads a file to Termostat's website and it creates a list of terms from it. For our first experiment, we attempted to simulate Termostat's use case as closely as possible. We chose a single document as the foreground: a copy of Einstein's Theory of Relativity, downloadable from Project Gutenberg[21]. We removed some initial and final meta-data from Project Gutenberg before using it. We constructed a background corpus that was as close as possible to the one used by Termostat, so Termolator would be running under similar conditions. Specifically, we used the British National Corpus for Termolator's background[22]. After running both Termolator and Termostat on these data, we manually evaluated the results, using the same technique as above. Termolator's stage 2 system generated 673 terms and stage 3 ranked the top 204 of these, since for relatively small lists of terms, the system only keeps the top 30%. Termostat output 1407 terms, of which we only ranked the top 30% or 422 terms. As before, we sampled 100 terms (20 from each fifth) and then manually rated terms as valid or invalid. We rated 53% of the Termolator and 50% of the Termostat terms as being valid terms. Given the difficulty of this annotation task, we believe that it is safe to assume that the systems had roughly the same accuracy.

[17] We evaluated the correctness of terms ourselves. We previously did some experiments in which graduate biology students evaluated our biology terms. We discontinued this practice primarily because we could not afford to have experts in all of the domains for which we had terms. In addition, the domain expertise was rarely accompanied by linguistic expertise. So the process of training domain experts to make consistent determinations about what does and does not constitute a linguistic unit was difficult. In contrast, using one set of annotators resulted in more consistent evaluation. Most unknown terms could be looked up and identified with high accuracy.

[18] There are no established sets of manually encoded data to test the system with. Note that the SemEval keyword extraction task (Kim et al., 2010) while overlapping with terminology extraction, does not capture the task we are doing here. In particular, we are not attempting to find a small number of keywords for a small number of articles, but rather large sets of terms that cover fields of study. We believe that constructing such a shared task manually would be prohibitive.

[19] http://termostat.ling.umontreal.ca/index.php?lang=en_CA

[20] Much of the work that assumes a similar terminology task either precedes Droun 2003 or is not readily available for testing purposes (Justeson and Katz, 1995; Navigli and Velardi, 2004, etc.). Other "terminology extraction" systems assume different tasks, e.g., Defminer (Jin et al., 2013) describes a task of finding terms and their term definitions from computational linguistics research papers. Kim et al. (2010) describes yet another task (key word extraction) which is similar, but not the same as the terminology extraction task described here (i.e., key words are not the same as terminology). Termostat seems to be the only currently available system that frames the terminology detection task the same way as we do.

[21] http://www.gutenberg.org/ebooks/5001.txt.utf-8

[22] The British National Corpus is described here: http://www.natcorp.ox.ac.uk/. Termostat's background corpus includes both the British National Corpus and 13.7K articles from *The Gazelle*, a Montreal newspaper. We only had access to the former, so we could not use it in the background for Termolator.

Another noticeable difference is that there were more 1-word terms in Termostat's output (31%) vs. Termolator's output (20%), especially toward the beginning of the ranking— for the first 1/5 of the terms, 45% of the Termostat terms and 10% of the Termolator terms consisted of single words. In an additional experiment, we ran the filters from Stage 3 (well-formedness and relevance) on the Termostat output and sampled 100 terms in the same manner. These terms were valid 53% of the time, the same as the run with Termolator. This suggests that if the difference in accuracies turns out to be significant, this difference may be due to the Stage 3 filters. 29% of the terms generated from this experiment were single word terms, a similar percentage as with before the application of the filter.

Next we then ran both Termolator and Termostat on some patent data. We downloaded the 2002 US patent applications from the US patent office[23]. We randomly chose a 5,000 file background corpus from these files. We also selected two sets of foreground files based on patent codes for refrigeration (062) and semiconductors (438)[24]. We selected 500 documents randomly about refrigeration and 5,000 randomly about semiconductors. We ran Termolator two times, both using the patent background corpus and once with each of the two foreground corpora. Then we endeavored to run Termostat using these two foreground corpora and Termostat's standard background corpus. Since Termostat requires a single file as input, we needed to merge these files together into two foreground files, one for each domain[25]. It was no problem to run Termostat with the Refrigeration file, but the Semiconductor file (235 mb) proved too large for the web version of Termostat. However, Patrick Drouin, the author of Termostat was kind enough to run it for us on his server. We evaluated the output files in the same manner as before. For the refrigeration topic, Termolator got 70% of the sample correct, whereas Termostat got 52% correct. For the semiconductor topic, Termolator got 79% correct and Termostat got 51% correct. For the refrigeration topic, Termolator detected 37,000 possible terms, of which 30,000 went through Stage 3 and were reranked. Then the 100 being manually scored were selected from the top 5,000 (20 randomly from the first 1,000, 20 randomly from the second 1,000, etc.). Termostat selected 11,675 possible terms, the top 30% or 3,502 were sampled for scoring (we chose the top 5,000 or the top 30%, whichever is less). For the semiconductor topic, Drouin provided us with the 3,073 terms that had at least 300 instances in the input text. We sampled the 100 terms from this group and scored them.

The first use case in which there was a single input file (*Einstein's Theory of Relativity*), Termolator and Termostat produced approximately the same quality output. However, for the second use case, involving a large set of foreground files, Termolator did noticeably better. A number of factors contributed to these differences. First of all, we have found that Termolator tends to produce a larger number of good terms than other systems[26]. We believe that our chunking system provides a larger pool of good candidates, so the distributional metrics have better input and therefore can produce a larger amount of high-quality output. Secondly, this use case fits Termolator's model better than it does Termostat's. Some of Termolator's measures test how many different files contain a term – this is not possible if the foreground and background are both single files. Thirdly, by selecting a background corpus in the patent domain, this means that many of the patent-specific terminology will be ruled out (terms about legal matters and inventions in general)[27]. In contrast, by comparing to a general purpose corpus, patent terms will naturally stand out, just as much as refrigeration or semiconductor terms. Finally, although we have shown that our Stage 3 filters improve the quality of Termolator output, we have yet to prove that they will improve the output of other systems. Our initial attempt to prove this was only suggestive, giving a probably-insignificant 3 percentage point boost to Termostat's output on the Einstein document.

Caching for Efficiency

We include caching options for several parts of Termolator that are reused when the system is run multiple times with similar types of input documents. This can substantially decrease the run time (after the first time the system is run). The following caching options have been implemented:

- ***Background Statistics***: It is common to run different foreground corpora against the same background corpus. For example, we have created foreground corpora, each based on different patent codes and thus covering different specific subject matter for those patents. We then ran these systems against a background corpus consisting of a wide variety of patents. We will choose all the patent documents from the same epoch, e.g., from the same year. It turns out that each of our distributional metrics (TFIDF, DRDC, and KLD) have some components based on the foreground and others based on the background. Specifically, for the background corpus, we only need one opportunity to count the number of times that a term occurs in the background documents and its Inverse Document Frequency or IDF (log of the number of documents containing a term divide by the number of background documents). By storing this information in a file, we can use it to calculate these metrics for terms in any new foreground file.
- **Relevance Scores:** The relevance scores for terms is another example. These scores can take as much as 0.75 s per term as they are based on web searches. However, these results will change very slowly over time. Within a fairly large time window, it is reasonable to store all relevance score calculated. Thus table look up can be used for finding relevance scores

[23] All the zip files from: https://bulkdata.uspto.gov/data/patent/application/redbook/fulltext/2002/

[24] These patent codes are part of a system used for U.S. patents until 2011. It is describe here: https://www.uspto.gov/web/patents/classification/selectnumwithtitle.htm. Starting 2011, the US switched to the world-wide CPC system.

[25] The Termolator can run on XML text, including text in the format of the U.S. patents, whereas Termostat requires plain text files. Thus in creating the input to Termostat, we combined some intermediate .txt files created by Termolator.

[26] We made some informal observations in the past when comparing results. However, until now, it has proven difficult to do a formal comparison.

[27] Additionally, some of our term filters specifically rule out known patent terms, e.g., *embodiment, claim, copyright*.

whenever possible and every newly calculated score is added to the table (and the table is stored in a file).

THE CHINESE SYSTEM

Our current Chinese Termolator implements several components parallel to the English system and we intend to implement additional components in future work. The Chinese Termolator uses an in-house CTB[28] word segmenter and part-of-speech tagger and a rule based noun group chunker, but without additional rules with regard to technical words. Stage 2 is similar to the English system in that we compare word distribution in a given domain with word distribution in a general background set and find topic words of the given domain.

One challenge for the Chinese system is that Chinese word boundaries are implicit, and are automatically induced by the word segmenter, which is prone to errors. We accordingly implemented an accessor-variety (AV) based filter (Feng et al., 2004), which calculates an accessor-variety score for each word based on the number of distinct words that appear before or after it. Character sequences with low AV scores are not independent enough, and usually should not be considered as valid Chinese words (Feng et al., 2004). We therefore filter out words whose accessor-variety scores are less than 3. We evaluated the precision of extracted terms on a set of speech processing patents: the precision was 85% for the top 20 terms and 78% for the top 50 terms. This evaluation was based on 1,100 terms extracted from 2,000 patents related to speech processing.

We developed a well-formedness-based automatic evaluation metric for Chinese terms, which follows the same spirit as the English well-formedness score. This metric penalizes noun phrases that contain non-Chinese characters, contain words that are not nouns or adjectives, contain too many single character words, or are longer than 3 characters. Since this error is exactly the sort of error that would be ruled out by the AV-based filter, we do not use it as part of our own terminology system. Rather, we use it when we are applying our filters to score term lists created externally, just as we are doing with parts of the English system.

We expect to implement a version of the Relevance Score that will work with Chinese language search engines in future work. As with the English, this will be a separable component of the system that can be applied to Chinese term lists created independently from our system.

CONCLUDING REMARKS

We have described a terminology system with state-of-the-art results for English that combines several different methods including linguistically motivated rules, a statistical distribution metric and a web-based relevance metric. We can derive at least 5,000 highly accurate (80–86%) terms from 5,000 documents about a topic. Given fewer input documents, the accuracy scores may be somewhat lower – the experiment on a single file (Einstein's Theory of Relativity) resulted in 54% accuracy and the experiment on 500 refrigeration patents resulted in 70% accuracy and the experiment with semi-conductor patetens resulted in 79% accuracy. More evaluation is necessary to determine if this is a consistent trend or is confounded by other factors, e.g., perhaps some topics are easier than others.

One important characteristic of our system is its combination of knowledge-based and statistical components. The knowledge-based components (dictionaries, manual-rule based chunkers, etc.) improve the results, but slow down the expansion of the system, e.g., the creation of systems for extracting terminology in other languages. Most alternatives involve substituting statistical components, e.g., the results of web searches for the knowledge-based components. However, Termolator already has statistical components and in future work, we would consider adding more such components. We do not see statistical and knowledge-based components to be an either-or question. Rather, we seek to combine the best knowledge-based components with the best statistical ones. For example, we have shown that a knowledge-based chunker produces better input to our distributional component than other types of input.

For future work, we are interested in improving on the one document use-case. Indeed, we imagine that it would be interesting to find the top N terms for all the single documents in a collection—the terms that represent the topic of the document. We have done some preliminary experiments with supreme court decisions and are finding this to be a challenging area.

As reported, the Chinese version of Termolator currently achieves accuracy of 78% accuracy for the first 50 terms, when run on 1100 patents. In future work, we intend to further develop the system for Chinese, possibly to include additional features similar to those currently implemented only in the English system. We are also considering, creating a version of Termolator for Spanish.

AUTHOR CONTRIBUTIONS

AM: Project lead, design, implementation, research and evaluation of all stages of English system. YH: Design, evaluation and implementation of the Chinese system. ZG: Design, implementation and research for stage 2 system. JO: Optimization and evaluation of stage 2 system. SL: Design and implementation of original Stage 2 system. AG-S: Evaluation of Stage 1 system. RG: Design and technical guidance. OB-M: Design and technical guidance, evaluation, and providing use-cases.

FUNDING

Supported, in part, by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20154.

[28] https://catalog.ldc.upenn.edu/LDC2013T21

ACKNOWLEDGMENTS

This paper combines and updates work reported in Meyers et al. (2014a, 2015). Authors of this paper hold the copyrights to these preprints. Copies of the preprints are available at: http://www.aclweb.org/anthology/W/W14/W14-6002.pdf and http://ceur-ws.org/Vol-1384/paper5.pdf.

REFERENCES

Babko-Malaya, O., Seidel, A., Hunter, D., HandUber, J., Torrelli, M., and Barlos, F. (2015). "Forecasting technology emergence from metadata and language of scientific publications and patents," in *15th International Conference on Scientometrics and Informetrics* (Istanbul).

Cover, T., and Thomas, J. A. (1991). *Elements of Information Theory*. New York, NY: Wiley-Interscience.

Daim, T. U., Rueda, G., Martin, H., and Gerdsri, P. (2006). Forecasting emerging technologies: use of bibliometrics and patent analysis. *Technol. Forecast. Soc. Change* 73, 981–1012. doi: 10.1016/j.techfore.2006.04.004

Damerau, F. J. (1993). Generating and evaluating domain-oriented multiword terms from texts. *Inform. Process. Manage.* 29, 433–447.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology* 9, 99–115. doi: 10.1075/term.9.1.06dro

Feng, H., Chen, K., Deng, X., and Zheng, W. (2004). Accessor variety criteria for chinese word extraction. *Comput. Linguist.* 30, 75–93. doi: 10.1162/089120104773633394

Hisamitsu, T., Niwa, Y., Nishioka, S., Sakurai, H., Imaichi, O., Iwayama, M., et al. (1999). "Term extraction using a new measure of term representativeness," in *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition* (Tokyo).

Jacquemin, C., and Bourigault, D. (2003). "Term extraction and automatic indexing," in *Handbook of Computational Linguistics,* ed R. Mitkov (Oxford: Oxford University Press).

Jin, Y., Kan, M., Ng, J., and He, X. (2013). "Mining scientific terms and their definitions: a study of the ACL anthology," in *EMNLP-2013* (Seattle: ACL).

Justeson, J. S., and Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Nat. Lang. Eng.* 1, 9–27.

Kim, S. N., Medelyan, O., Kan, M. Y., and Baldwin, T. (2010). SemEval-2010 task 5: automatic keyphrase extraction from scientific articles. *SemEval* 21–26.

Kim, J.-D., Ohta, T., Tateisi, Y., Tsujii, J. (2003). GENIA corpus-a semantically annotated corpus for bio-textmining. *Bioinformatics* 19(Suppl. 1), i180–i182. doi: 10.1093/bioinformatics/btg1023

Macleod, C., Grishman, R., and Meyers, A. (1997). COMLEX Syntax. *Comp. Human.* 31, 459–481.

Macleod, C., Grishman, R., Meyers, A., Barrett, L., and Reeves, R. (1998). "Nomlex: a lexicon of nominalizations," in *Proceedings of Euralex* (Liège), 98.

Meyers, A. (2007). *Those Other NomBank Dictionaries – Manual for Dictionaries that Come with NomBank*. Available online at: http:nlp.cs.nyu.edu/meyers/nombank/nomdicts.pdf

Meyers, A., Glass, Z., Grieve-Smith, A., He, Y. S. L., and Grishman, R. (2014a). "Jargon-term extraction by chunking," in *COLING Workshop on Synchronic and Diachronic Approaches to Analyzing Technical Language* (Dublin).

Meyers, A., He, Y., Glass, Z., and Babko-Malaya, O. (2015). "The termolator: terminology recognition based on chunking, statistical and search-based scores," in *Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics* (Istanbul).

Meyers, A., Lee, G., Grieve-Smith, A., He, Y., and Taber, H. (2014b). Annotating relations in scientific articles. *LRE*C-2014.

Meyers, A., Reeves, R., Macleod, C., Szekeley, R., Zielinska, V., and Young, B. (2004). "The cross-breeding of dictionaries," in *Proceedings of LREC-2004* (Lisbon).

Navigli, R., and Velardi, P. (2004). Learning domain ontologies from document warehouses and dedicated web sites. *Comput. Linguist.* 30, 151–179. doi: 10.1162/089120104323093276

Ortega, J., Forcada, M., and Sánchez-Martinez, F. (2016). Using any translation source for fuzzy-match repair in a computer-aided translation setting. *Assoc. Mach. Trans. Am.* 1:204.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). "The pagerank citation ranking: bringing order to the web," in *Proceedings of the 7th International World Wide Web Conference* (Brisbane, QLD).

Ramshaw, L. A., and Marcus, M. P. (1995). "Text chunking using transformation-based learning," in *ACL Third Workshop on Very Large Corpora* (Cambridge, MA), 82–94.

Schwartz, A., and Hearst, M. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac. Composium Biocomput.* 451–462.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *J. Documentation* 28, 11–21.

Velardi, P., Missikoff, M., and Basili, R. (2001). "Identification of relevant terms to support the construction of domain ontologies," in *Workshop on Human Language Technology and Knowledge Management* (Toulouse).

Disclaimer: The Termolator has been approved for public release; unlimited distribution.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Research Metrics and Analytics

ORIGINAL RESEARCH
published: 13 July 2018
doi: 10.3389/frma.2018.00021

Deep Reference Mining From Scholarly Literature in the Arts and Humanities

Danny Rodrigues Alves[1], *Giovanni Colavizza*[1,2]* *and Frédéric Kaplan*[1]

[1] Digital Humanities Laboratory, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, [2] The Alan Turing Institute, London, United Kingdom

We consider the task of reference mining: the detection, extraction and classification of references within the full text of scholarly publications. Reference mining brings forward specific challenges, such as the need to capture the morphology of highly abbreviated words and the dependence among the elements of a reference, both following codified reference styles. This task is particularly difficult, and little explored, with respect to the literature in the arts and humanities, where references are mostly given in footnotes. We apply a deep learning architecture for reference mining from the full text of scholarly publications. We explore and discuss three architectural components: word and character-level word embeddings, different prediction layers (Softmax and Conditional Random Fields) and multi-task over single-task learning. Our best model uses both pre-trained word embeddings and characters embeddings, and a BiLSTM-CRF architecture. We test our solution on a dataset of annotated references from the historiography on Venice and, using a linear-chain CRF classifier as a baseline, we show that this deep learning architecture improves by a considerable margin. Furthermore, multi-task learning performs almost on par with a single-task approach. We thus confirm that there are important gains to be had by adopting deep learning for the task of reference mining.

Keywords: reference mining, natural language processing, conditional random fields, deep learning, recurrent neural networks, bibliometrics, arts and humanities, history

OPEN ACCESS

Edited by:
Philipp Mayr,
Leibniz Institut für Sozialwissenschaften (GESIS), Germany

Reviewed by:
Lin Zhang,
KU Leuven, Belgium
Animesh Prasad,
National University of Singapore, Singapore

****Correspondence:***
Giovanni Colavizza
gcolavizza@turing.ac.uk

Received: *26 February 2018*
Accepted: *12 June 2018*
Published: *13 July 2018*

Citation:
Rodrigues Alves D, Colavizza G and Kaplan F (2018) Deep Reference Mining From Scholarly Literature in the Arts and Humanities. Front. Res. Metr. Anal. 3:21. doi: 10.3389/frma.2018.00021

1. INTRODUCTION

Reference mining (or parsing) is a Natural Language Processing (NLP) task focused on the detection, extraction and classification of bibliographic references and their constituent components from scholarly literature. It is a necessary step toward the creation of relational citation data, a task commonly performed in view of building citation indexes (Garfield, 1979). Compared to other NLP tasks, reference mining stands in the broader category of sequence labeling problems, which includes among others Part Of Speech (POS) tagging and Named Entity Recognition (NER). Traditional machine learning methods for sequence labeling tasks, including Hidden Markov Models (HMM) and (linear-chain) Conditional Random Fields (CRF), depend on a considerable amount of external knowledge in the form of hand-engineered features and task-specific resources like gazetteers and lexicons. However, these resources are costly to produce and are not easy to adapt to variations of a given task, especially so because they require expert human knowledge.

In recent years deep learning, or the use of deep neural network models trained on large amounts of data, has been changing the whole field of machine learning, considerably improving on most tasks (LeCun et al., 2015; Schmidhuber, 2015). Yet the openly available non-commercial tools for reference parsing still mostly rely on previous-generation techniques (Tkaczyk et al., 2018). Quite consequently, this paper contribution is to take a deep learning approach by applying current state-of-the-art architectures for sequence labeling to the specific task of reference mining.

A further motivation for the use of deep learning comes from the scholarly domain which we interest ourselves into: the arts and humanities. Where reference mining applications targeting most scientific publications need to focus on relatively uniform reference lists, scholarly publications in the arts and humanities are more varied in this respect (Sula and Miller, 2014; Colavizza et al., 2017). A set of challenges must be considered: references are made to (at least) both primary and secondary sources, and primary sources are by definition more varied than secondary ones. References can happen anywhere in the text of a publication, especially so in footnotes, and not just in reference lists. In this case, references are often given once in full form and abbreviated thereafter. It must also be noted that it is not customary to cite primary sources in reference lists. Lastly, the variety of publication venues, languages, scholarly communities in the arts and humanities is broader, making reference practices and styles less uniform. For these and other reasons, the scholarly literature from the arts and humanities is still not well indexed (Mongeon and Paul-Hus, 2016) nor studied (Ardanuy, 2013) using citation data.

We consider and compare several components of a recurrent neural network architecture for reference mining. In particular, we experiment with different approaches in the input layer, by considering both character and word-level embeddings. We also test a Conditional Random Field instead of the canonical Softmax prediction layer. Finally, we experiment with multi-task learning in order to test whether the learning our best model does is shared across different tasks. All models are built around a single BiLSTM layer, a proven key ingredient in a variety of sequence labeling tasks. We make two implementations available, one using *Keras* (Chollet et al., 2015) (relying on *TensorFlow* as back-end), and another directly in *TensorFlow* (Dean et al., 2015), in order to facilitate the reuse of results and further experimentation.[1] Our experiments are based on a published dataset of annotated references from a corpus of publications on the history of Venice (Colavizza and Romanello, 2017).

This paper is organized as follows. We briefly discuss previous work in section 2, then introduce the task of reference mining and the dataset in section 3. In the same section, a CRF baseline model is discussed. Section 4 describes the general architecture we propose and test in all its components. Section 5 contains our results, as well as the details of the best architecture and model configuration, with its validation. We finally conclude in section 6.

2. RELATED WORK

In a recent survey and evaluation, several non-commercial reference parsing tools, Tkaczyk et al. (2018) found that the best three performing ones all use a CRF approach: GROBID (Lopez, 2009), CERMINE (Tkaczyk et al., 2015) and ParsCit (Councill et al., 2008). All three benefit from task-specific tuning using extra annotated data, with GROBID showing the best off-the-shelf results. Indeed seven out of the total of thirteen surveyed tools use a CRF approach, while the rest mainly adopt regular expressions. To date, all published non-commercial reference mining tools rely on these or rule-based methods[2]. Heckmann et al. (2016) attempted to tackle some of the main challenges to be found in humanities literature, namely: "multilingual citation entries, lack of data redundancy, inconsistencies, and noise from OCR input." Their knowledge-based approach relying on Markov logic networks was found to substantially outperform a CRF baseline. A useful insight for the task at hand also came from Körner et al. (2017), where a CRF is used to classify lines of text containing references in advance to considering their constituent tokens. The proposed method, RefExt, outperformed several above-mentioned state-of-the-art solutions.

As deep learning started to gain momentum in recent years, attention has been given to the use of unsupervised feature extraction techniques in a variety of NLP tasks, mainly in the form of word embeddings, which lead to state-of-the-art results when used to augment, rather than replace, hand-crafted features (Collobert et al., 2011). More recent work on sequence labeling tasks relies instead on deep learning techniques such as convolutional or recurrent neural network models (CNNs LeCun et al., 1989 and RNNs Rumelhart, 1986, respectively), without the need for any hand-crafted features (Kim, 2014; Huang et al., 2015; Zhang et al., 2015; Chiu and Nichols, 2016; Lample et al., 2016; Ma and Hovy, 2016; Yang et al., 2016; Strubell et al., 2017). RNNs in particular, typically rely on a neural network architecture built using one or more Bidirectional Long-Short Term Memory (BiLSTM) layers, as this type of neural cell provides for variable-length memory allowing the model to capture relationships within sequences of proximal words. Such architectures have achieved state-of-the-art performance for both POS and NER tasks on popular datasets (Reimers and Gurevych, 2017b). Current state-of-the-art architectures for sequence labeling include the use of a CRF prediction layer (Huang et al., 2015) and the use of character-level word embeddings to complement word embeddings, trained either with CNNs (Ma and Hovy, 2016) or BiLSTM RNNs (Lample et al., 2016). Character-level word embeddings have indeed been shown to perform well on a variety of NLP tasks (Dos Santos and Gatti de Bayser, 2014; Kim et al., 2015; Zhang et al., 2015). Attention mechanisms have also been proposed for the same tasks (Rei et al., 2016; Shen and Lee, 2016). In this paper we will apply, tune and compare two architectures (Lample et al., 2016; Ma and Hovy, 2016) to the specific task of reference mining.

[1] Keras version `2.1.1` and TensorFlow version `1.4.0`.

[2] An exception is Neural ParsCit https://github.com/opensourceware/Neural-ParsCit, a yet unpublished adaptation of the architecture proposed in Lample et al. (2016) for the task of reference parsing.

3. TASK DEFINITION, DATASET, AND BASELINE MODEL

A bibliographic reference is a contiguous sequence of text where all the necessary information on a citation to any primary or secondary source is contained. Most typically, previous scholarship and primary evidence such as archival documents or works of art and literature can be cited in arts and humanities scholarly literature. What constitutes necessary information is relative: usually, the first citation to a source contains all information necessary for its unambiguous identification, a substantial part of this same information can be dropped or abbreviated in subsequent citations to the same source within the same publication.

A reference is usually composed of several information components, such as the *author*, *title* or *publisher* of a cited publication, encoded in a systematic way following some editorial guidelines specific to the venue and time of publication (e.g., using double quotes or italics for the title). An example of a reference is

```
G. Ostrogorsky, History of the Byzantine
  State, Rutgers University Press, 1986.
```

This reference has four components: the author's name, title, publisher and year of publication. In this example, the components are separated by a comma and the author's name is abbreviated using initials followed by a dot. The same reference might be given elsewhere following a different *reference style*, defined as: "a specific combination of elements in a reference, such as author and title, encoded in a predefined way" (Colavizza et al., 2017, p. 4). For example, it might be given as *"Ostrogorsky, G. (1986). History of the Byzantine State, Rutgers University Press,"* where the combination of elements as well as their encoding has changed.

If we consider a text as a stream of tokens organized into lines (sequences of characters separated by white space), the goal of reference mining is to:

- **Detect** that a token is part of a reference. A token part of a reference can be anywhere, most typically in footnotes.
- **Extract** a reference, i.e., individuate its first and last tokens (begin-end).
- Optionally **classify** a full reference and its constituent components: in our case, a reference might be to a primary or secondary source (this information is useful for further processing steps such reference disambiguation to establish citations, as this step typically relies on existing catalogs look-up), and each reference might contain a variety of components (author, title, archive and record group, etc.).

In this article we consider all three actions, and use the processing unit (sequence) of the line of text. Our motivation to use sequences as lines of text is given by the need to parse the full-text of publications in order to capture footnotes, and the irregular positioning of references therein. The extraction and detection of references is done using *begin-end* token classification to mark, respectively, the beginning and end of a reference within a stream of tokens. With respect to classification, two annotation schemes (tags) are considered: *specific* and *generic*. A specific annotation identifies a component of a reference, such as author or title. A generic annotation refers to the typology of the cited source, distinguishing among primary sources, books and other contributions such as journal articles. More in detail, given the plain text of a publication, our goal is to assign the most likely tag to each token (token by token classification). We define three tasks as follows:

- Task 1: *reference components*. Each token is classified using a taxonomy of 27 specific tags, unevenly represented in the annotated dataset, which include a non-reference tag. The taxonomy is given and discussed in Colavizza and Romanello (2017) and in the accompanying code repository. The reason to have 27 tags is mainly the presence of references to archival documentation, which requires a classification on its own.
- Task 2: *reference typology*. Each token is classified according to the generic annotation scheme. As mentioned above, tags include: *primary* sources (e.g., archival documents), *secondary* sources (books), and *meta*-sources, i.e., publications contained within other publications (e.g., journal articles). Furthermore, *b*egin, *e*nd and *i*n reference tags are prepended to a generic tag, and an *o*ut of reference tag is used too. For example, *b-secondary* marks the first token of a reference to a book-form publication.
- Task 3: *reference span*. Each token is classified simply using the *b*egin, *e*nd, *i*n and *o*ut schema. For example, *e-r* marks the last token of a reference.

The different tasks are illustrated in **Figure 1**, using the example given above.

3.1. Dataset

We use a published dataset containing more than 40,000 annotated references from a corpus of publications on the historiography on Venice. The corpus includes books and journal articles published from the 19th century to 2014. It considers publications in a variety of languages: mostly Italian, followed by English, French, German, Spanish and Latin. The annotated corpus includes references taken from reference lists and footnotes, as a consequence, a considerable variety of referencing styles and referred sources is present. Annotated references for every publication are a representative sample of the total amount. For reasons of copyright, this dataset does not contain the full text of publications, but only the text lines where a reference (or part of it) appears; therefore some lines of text include out-of-reference tokens, preceding or following a reference (these tokens are important to learn to assign begin-end tags). Full details, including corpus acquisition and annotation sampling strategy and procedure, are given in Colavizza and Romanello (2017) [3].

A new export of this dataset is used here, prepared as follows. Initially, every publication with annotated references is randomly

[3] The dataset and accompanying code are available in, respectively GitHub: https://github.com/dhlab-epfl/LinkedBooksReferenceParsing and Zenodo: http://doi.org/10.5281/zenodo.579679

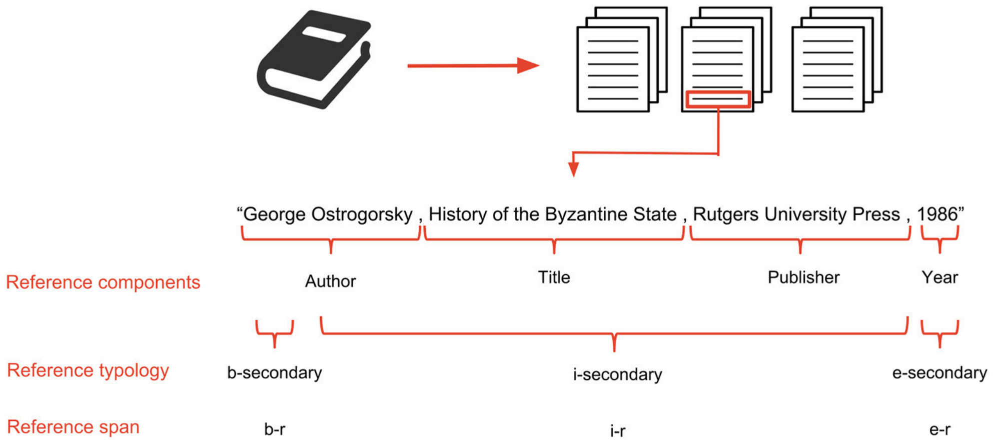


FIGURE 1 | Example of a reference annotated according to tasks 1–3. Task 1 covers reference components, task 2 considers the span of a reference plus its general typology, task 3 instead only considers the span of a reference. We use a clear-cut example for illustration purposes, yet in fact most references given in footnotes have text before and afterwards, often on the same line.

allocated in a train, test or validation set, with an 80/10/10 split respectively[4]. The number of references in each set does not precisely follow the same proportion, as different publications have a varied amount of annotated references. Nevertheless, a publication-level split is important in order to reduce reference style data snooping. Next, the annotated lines of plain text for every publication are considered independently and split into tokens using the NLTK word-punkt tokenizer (Bird et al., 2009), thus considering several punctuation symbols as a separate token. The dataset is at this point composed of a set of lines of text, which will be parsed independently, each including at least part of a reference, split into tokens and associated with the annotation schemes of the different tasks. By all means, a reference can be part of multiple lines of text. The choice of considering lines of text independently reduces the dependency window that the classification method can rely upon, and is to be considered a limitation of this study.

This reprocessed dataset is made available using the CoNLL convention: each line in a file (test, train and validation) corresponds to a token in a sequence (original line of text), and sequences are separated by a blank line. Each token line contains the token surface form followed by the corresponding tags for each task, separated by a white space. To encode the relative position of a token in a reference, the IOBE convention is used, where `i-label` stands for a token inside a reference (not begin or end), `o` outside, `b-label` if the token is the first of a reference and `e-label` the last. The IOBE is a variant of the more common IOB scheme. Using a more expressive tagging scheme like IOBE has been shown to marginally improve model performance (Ratinov and Roth, 2009; Dai et al., 2015) and ease the retrieval of references spanning across several lines.

Our example *"G. Ostrogorsky, History of the Byzantine State, Rutgers University Press, 1986,"* assuming it spans a single line (sequence), is encoded as:

```
G author b-secondary b-r
. author i-secondary i-r
Ostrogorsky author i-secondary i-r
, author i-secondary i-r
History title i-secondary i-r
of title i-secondary i-r
the title i-secondary i-r
Byzantine title i-secondary i-r
State title i-secondary i-r
, title i-secondary i-r
Rutgers publisher i-secondary i-r
University publisher i-secondary i-r
Press publisher i-secondary i-r
, publisher i-secondary i-r
1986 year e-secondary e-r
. year e-secondary e-r
```

3.2. CRF Baseline

We train and test a Conditional Random Field (Lafferty et al., 2001) baseline using the same dataset. The CRF classifier is trained over a rich set of hand-crafted features considering a size-two bi-directional window: the features for a token at position t in a sequence include features extracted for the two preceding and two following tokens too, that is positions $t-2$, $t-1$, $t+1$, $t+2$, following previous work where the specificities of applying CRF to the humanities are amply discussed (Colavizza and Romanello, 2017). This model is trained with Stochastic Gradient Descent applying both L1 and L2 regularization, using the *CRFSuite* package (Okazaki, 2007)[5]. The code and training details are given in this work's accompanying repository. The best cross-validated

[4] Sometimes in the literature what we refer as test dataset, to assess the results of training, is named development dataset, and the validation dataset, what we use at the end to test for generalization, is named test dataset. We will use what we call test dataset for development and what we call validation dataset for final testing.

[5] We used the *CRFsuite* implementation from sklearn-crfsuite, version 0.3.6 available at https://github.com/TeamHG-Memex/sklearn-crfsuite.

configuration of this model yields the following F1 validation scores for each task:[6]

Task 1 gives an F1 score of **`82.63%`**
(precision `82.88%`, recall `82.76%`).

Task 2 gives an F1 score of **`71.04%`**
(precision `71.32%`, recall `71.1%`).

Task 3 gives an F1 score of **`92.50%`**
(precision `92.64%`, recall `92.41%`).

4. MODEL

We consider a recurrent architecture organized into three layers: input (word representations), inner and prediction, following the best performing models for sequence labeling tasks (Lample et al., 2016; Ma and Hovy, 2016). The network firstly receives a sequence of (one-hot encoded) words $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, ..., \mathbf{w}^{(n)}$ as input and transforms it into a sequence of dense vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(n)}$, using a combination of word and character-level word embeddings. Secondly, word representations are passed to a bidirectional LSTM composed of two layers: a forward layer where the word representations are processed starting with input representation $\mathbf{x}^{(1)}$ to $\mathbf{x}^{(n)}$, and a backward layer from $\mathbf{x}^{(n)}$ to $\mathbf{x}^{(1)}$. The outputs of these two layers are concatenated and used in the prediction layer, which outputs a sequence of predictions $\hat{\mathbf{y}}^{(1)}, \hat{\mathbf{y}}^{(2)}, ..., \hat{\mathbf{y}}^{(n)}$ for each initial input word $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, ..., \mathbf{w}^{(n)}$. We remark that we refer to words from now on, to adopt the most common terminology in the literature, where in practice generic tokens are considered.

4.1. Input Layer

The input layer combines word and character-level word embeddings for each input token in a sequence, in order to create a word representation.

4.1.1. Word Embeddings

Word embeddings are a common staple of sequence classification tasks, and are often trained over large corpora like *Wikipedia*[7] or *Reuters*[8], in order to embed richer information than just using task-specific data. Yet considering the dataset used in this project, publicly available embeddings will likely not help. Instead, word embeddings were pre-trained using Word2Vec (Mikolov et al., 2013a,b) on the full contents of all the publications from which our references were extracted. We used the Gensim Word2vec implementation (Řehůřek and Sojka, 2010), a window of 5 words and the skip-gram model. The vectors are trained for all words appearing at least five times in the dataset, while less frequent words have been regrouped under an unknown token `$UNK$` and all digits have been merged into a `$NUM$` token. We tested word embeddings with a dimensionality of 100 and 300. Word embeddings can be randomly initialized and trained with the model, pre-trained and kept fix, or pre-trained and further trained with the model. The pre-trained word embedding vocabulary comprises 727,902 words, of which 51,569 are actually used in the published dataset.

4.1.2. Character-Level Word Embeddings

Tokens part of references contain relevant information at the orthographic and morphological levels, such as prefixes and suffixes and the use of punctuation or abbreviations. Given the relative small amount of annotated data at hand, it is likely the case that these features will not be learned at the word level in a satisfactory way. Conversely, character-level word embeddings can help into learning task-specific features at this level, with fewer examples. These features have in particular found useful application to deal with out-of-vocabulary words and morphologically rich languages (Dos Santos and Zadrozny, 2014). Furthermore, character-level word embeddings can help reduce the impact of OCR errors and help deal with rare words. Character-level word embeddings are a representation of a word from the compounded representation of sequences of characters the word is composed of. They can be learned either via CNNs or BiLSTMs. The character-level word embeddings are trained by first considering randomly initialized character embeddings. In the CNN case, we then feed them to a single 1d convolution layer followed by a max pool layer, using a filter stride of 1 and various widths. Alternatively, we use a BiLSTM and concatenate its outputs.

4.1.3. Word Representation Architecture

Figure 2 describes the architecture to build a word representation input made of the concatenation of a word embedding and a character-level word embedding trained with a BiLSTM. The word embeddings consist of a lookup to the precomputed Word2Vec embeddings, or randomly initialized ones, and the character-level word embeddings are computed through additional neural network layers as described above. The final word representation is a concatenation of its word embedding and character-level word embedding.

To prevent the model from too strongly depending on word and character-level word embeddings, dropout layers are added after the BiLSTM or CNN layers (for character-level word embeddings) and after word and character-level word embeddings are concatenated. More generally, as sketched in **Figure 3**, dropout layers are applied on several components of the final model. Dropout is a regularization technique where randomly selected neurons are turned off during training. It helps to prevent overfitting and to avoid the model to depend to heavily on individual neurons (Srivastava et al., 2014).

4.2. Inner Layer

Long-Short Term Memory cells (LSTM) are part of the Recurrent Neural Networks (RNN) family, designed to account for flexibly long memory dependences (Hochreiter and Schmidhuber, 1997). LSTMs overcome in part the limitations of vanilla RNNs, such as the practically short memory dependence and the tendency to suffer from vanishing or exploding gradients (Bengio et al., 1994).

[6] Here c1 and c2 refer to the model coefficients for L1 and L2 regularization, respectively. For task 1 cross validated parameters were set at `c1=1.3099` and `c2=0.0773`; `c1=0.9298` and `c2=0.0229` for task 2; `c1=2.1334` and `c2=0.0142` for task 3.

[7] https://nlp.stanford.edu/projects/glove/

[8] https://www.cs.umb.edu/~smimarog/textmining/datasets/

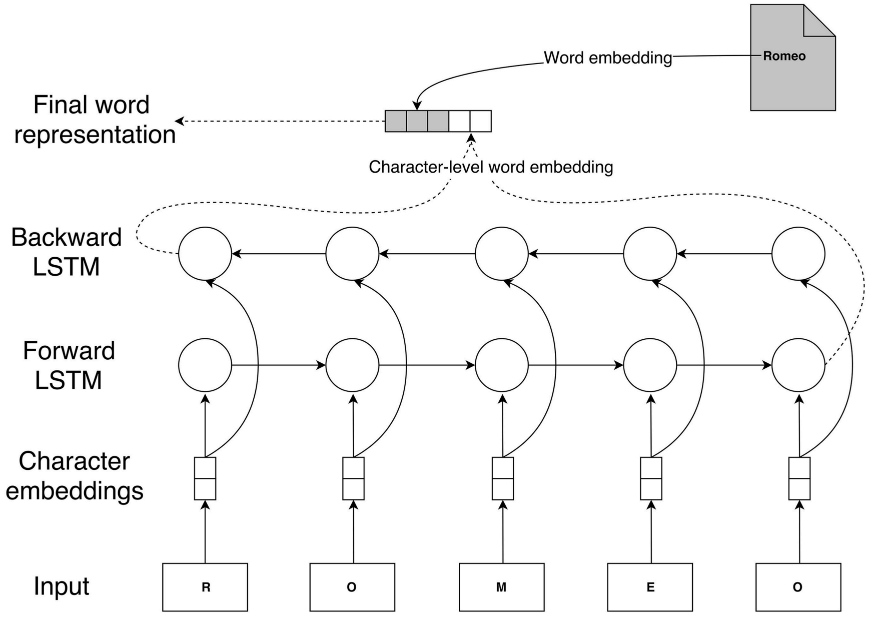


FIGURE 2 | The word representation architecture using both pre-trained word embeddings and BiLSTM character-level word embeddings, used in the example to construct the representation of the word "Romeo." Rectangles are used for inputs, sequences of squares for vectors, rounds for neuron cells and dashed lines for dropout connections.

An RNN cell with sigmoid activation and softmax prediction can be described as follows:

$$\mathbf{h}^{(t)} = \sigma(\mathbf{b} + \mathbf{W}\mathbf{x}^{(t)} + \mathbf{U}\mathbf{h}^{(t-1)})$$
$$\hat{\mathbf{y}}^{(t)} = \mathit{softmax}(\mathbf{c} + \mathbf{V}\mathbf{h}^{(t)})$$

where $\mathbf{x}^{(t)}$ is the input word representation in position t of the current sequence, $\mathbf{h}^{(t)}$ represents the hidden state at the same position, **b** and **c** are bias vectors and **W**, **U** and **V** are parameter matrices to be learned. An LSTM instead introduces three gates to the RNN configuration: an input gate **i**, a forget gate **f**, and an output gate **o**, in order to provide the cell with a means to retain information on previous states more effectively. An LSTM cell with softmax prediction, as implemented in Keras, can be described as follows:

$$\mathbf{i}^{(t)} = \sigma(\mathbf{b}^{i} + \mathbf{W}^{i}\mathbf{x}^{(t)} + \mathbf{U}^{i}\mathbf{h}^{(t-1)})$$
$$\mathbf{f}^{(t)} = \sigma(\mathbf{b}^{f} + \mathbf{W}^{f}\mathbf{x}^{(t)} + \mathbf{U}^{f}\mathbf{h}^{(t-1)})$$
$$\mathbf{c}^{(t)} = \mathbf{f}^{(t)} \odot \mathbf{c}^{(t-1)} + \mathbf{i}^{(t)} \odot \mathit{tanh}(\mathbf{b}^{c} + \mathbf{W}^{c}\mathbf{x}^{(t)} + \mathbf{U}^{c}\mathbf{h}^{(t-1)})$$
$$\mathbf{o}^{(t)} = \sigma(\mathbf{b}^{o} + \mathbf{W}^{o}\mathbf{x}^{(t)} + \mathbf{U}^{o}\mathbf{h}^{(t-1)})$$
$$\mathbf{h}^{(t)} = \mathbf{o}^{(t)} \odot \mathit{tanh}(\mathbf{c}^{(t)})$$
$$\hat{\mathbf{y}}^{(t)} = \mathit{softmax}(\mathbf{c} + \mathbf{V}\mathbf{h}^{(t)})$$

where σ is the element-wise hard-sigmoid function and $\odot$ is the element-wise product. As before, $\mathbf{x}^{(t)}$ is the input word representation in position t of the current sequence and $\mathbf{h}^{(t)}$ represents the hidden state at the same position. $\mathbf{x}^{(t)}$ represent the current cell state, as a function of the forget gate applied to the previous step cell state, and the input gate applied to a non-linear transformation (hyperbolic tangent in this case) of a vanilla RNN internal state. The final hidden state is then given by a product of the output gate with a further non-linear transformation of the cell state. The different bias vectors **b** and **c** and matrices **W**, **U**, and **V** are all learned parameters. A BiLSTM is made of two LSTM layers, one being fed the input in the original order, the other in reversed order. The final hidden layer is the concatenation of the two: $\mathbf{h}^{(t)} = \left[\overrightarrow{\mathbf{h}}^{(t)}; \overleftarrow{\mathbf{h}}^{(t)}\right]$.

Since inputs are processed in temporal order, a possible shortcoming of LSTMs is their inability to make use of subsequent context (Hochreiter et al., 2001). Nevertheless, two LSTMs can be used to process the input in opposite directions, and their results concatenated. This solution, referred to as a Bidirectional LSTM (Schuster and Paliwal, 1997), has shown notable results in a variety of NLP tasks (Graves and Schmidhuber, 2005; Graves et al., 2013; Huang et al., 2015).

4.3. Prediction Layer

A widely adopted prediction layer for multi-class sequence labeling tasks relies on the softmax function. Assuming **z** to be a vector of unnormalized log probabilities from a linear layer, we have:

$$\mathbf{z} = \mathbf{c} + \mathbf{V}\mathbf{h}$$
$$\mathit{softmax}(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1} e^{z_j}}$$

The softmax takes every classification decision independently for every input word, yet sequence labeling tasks seldom present no dependence between proximal tags. For example in our task 3,

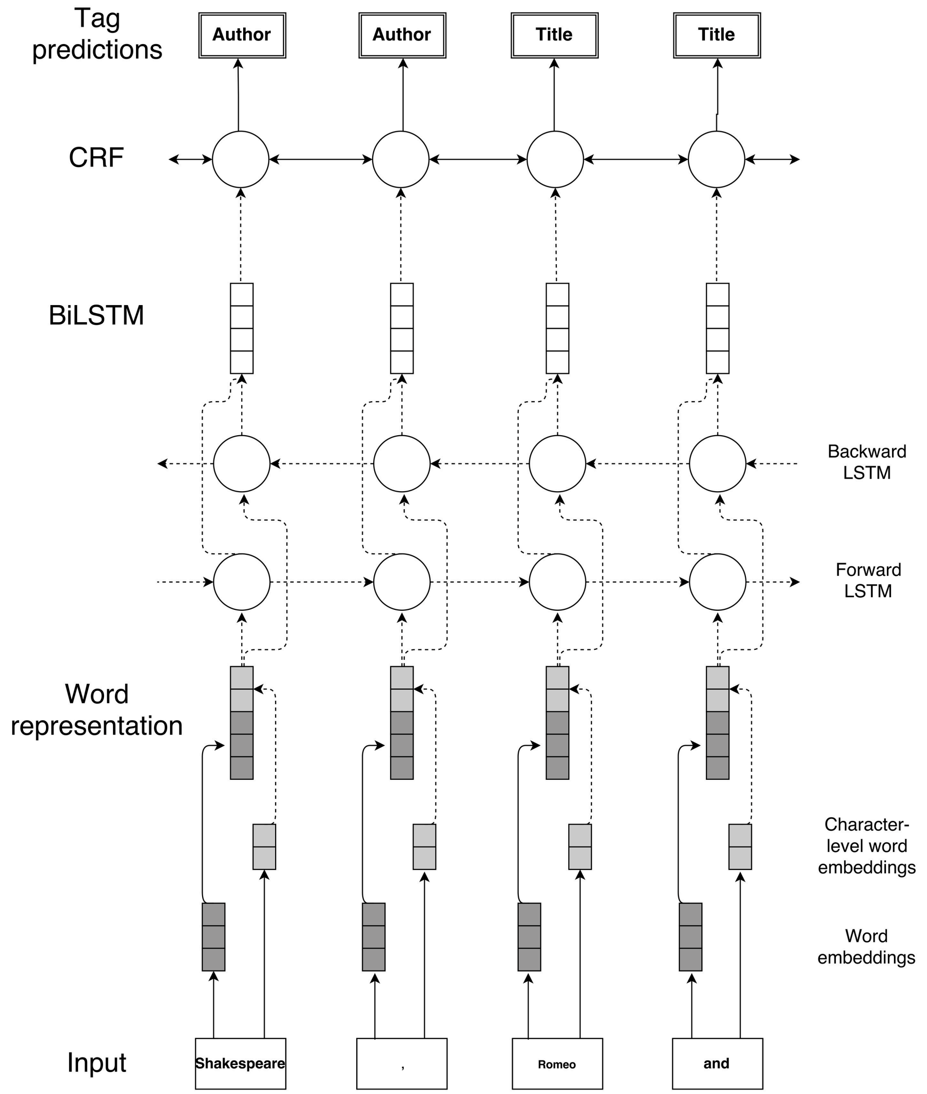


FIGURE 3 | Sketch of the model architecture for a part of the sequence `W. Shakespeare, Romeo and Juliet, Oxford University Press, London, 1914`. Rectangles are used for inputs, double rectangles for outputs, sequences of squares for vectors, rounds for neuron cells and dashed lines for dropout connections.

the tag *i-r* can never be followed by the tag *b-r*. More generally, reference styles entail that few recurring sequences of tags should be learned and predicted.

Using a CRF layer for predictions enables the model to perform classification decisions maximizing the (log) likelihood over the whole sequence of predictions (Lafferty et al., 2001; Sutton and McCallum, 2011). In the context of sequence labeling tasks, a linear-chain CRF is trained to predict a sequence $\mathbf{y} = (\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, ..., \mathbf{y}^{(n)})$ of known tags for a sequence input representation $\mathbf{X} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(n)})$. A linear-chain CRF in this setting uses a combination of unary features for state-observation pairs, and of binary features for each transition (Huang et al., 2015). We consider $\mathbf{Z}$ to be the $n \times k$ matrix of unnormalized scores from the inner BiLSTM layer, where n is the number of words in the sequence, k the number of possible tags (e.g., 27 for task 1). We then consider a square matrix $\mathbf{A}$ of new parameters, such that $\mathbf{A}_{i,j}$ represents the probability of transitioning from tag i to j in a sequence of predictions. In HMM terminology, $\mathbf{Z}$ is referred to as the emission matrix and $\mathbf{A}$ as the transition matrix. The score for the given sequence of tag assignments $\mathbf{y}$ is then calculated, and its probability over the space of possible tag prediction sequences $\hat{\mathbf{Y}}_{\mathbf{X}}$ taken with softmax:

$$score(\mathbf{X}, \mathbf{y}) = \sum_{i=0}^{n} \mathbf{A}_{y_i, y_{i+1}} + \sum_{i=1}^{n} \mathbf{Z}_{i, y_i}$$

$$\mathbb{P}(\mathbf{y}|\mathbf{X}) = \frac{e^{score(\mathbf{x}, \mathbf{y})}}{\sum_{\hat{\mathbf{y}}' \in \hat{\mathbf{Y}}_{\mathbf{X}}} e^{score(\mathbf{x}, \hat{\mathbf{y}}')}}$$

During training, the score of the correct tag sequence is maximized using dynamic programming. The best (maximum a posteriori) tag sequence assignment for a new input sequence can be computed using the Viterbi algorithm.

4.4. Multi-Task Learning

Multi-task learning has been considered to train and predict the three tasks at once, relying on the same architecture. This technique has proved useful to reach results comparable to single-task architectures, at a great reduced computational cost obtained by sharing most of the trained parameters across multiple tasks (Ruder, 2017). In some instances, multi-task learning can even improve single-task results. With respect to reference classification, we expect the inner layers of the network to learn quite similarly across different tasks, therefore it makes sense to attempt a multi-task approach.

Our multi-task architecture is identical to a single-task one up to the hidden layer outputs included. Afterwards, a separate prediction layer is created for each task. The loss function to be optimized is the sum of the losses of each task layer. Considering a softmax prediction layer, and the output $\mathbf{h}^{(t)}$ of the hidden layer at step t, we have:

$$\hat{\mathbf{y}}_1^{(t)} = softmax\big(\mathbf{c}_1 + \mathbf{V}_1\mathbf{h}_1^{(t)}\big)$$
$$\hat{\mathbf{y}}_2^{(t)} = softmax\big(\mathbf{c}_2 + \mathbf{V}_2\mathbf{h}_2^{(t)}\big)$$
$$\hat{\mathbf{y}}_3^{(t)} = softmax\big(\mathbf{c}_3 + \mathbf{V}_3\mathbf{h}_3^{(t)}\big)$$

The model thus has few extra parameters to learn, namely bias vectors **c** and matrices **V**.

5. EXPERIMENTS

In this section we detail the experiments conducted on variants of the neural network architecture under consideration, as well as the fine tuning of our best final model (5.1). We then validate and discuss the results (5.2). For both model selection and fine tuning, task 1 has been considered. Furthermore we used early epoch stopping on the F1 test score with a waiting window of 5 epochs without improvements, and a maximum number of 25 epochs. Both code and dataset are released publicly (see data availability statement).

5.1. Architecture

Three main variants of the architecture were considered in turn: (1) word embeddings (presence or absence, pre-trained or not, further trained or not); (2) character-level word embeddings (presence or absence, BiLSTM or CNN), (3) prediction layer (softmax or CRF). The best components were selected based on the F1 score on testing data[9]. Results reported in **Table 1** indicate that the best architecture uses pre-trained word embeddings which are further trained on the specific task, BiLSMT character-level word embeddings and a CRF prediction layer. The experiments on the architecture of the model always used the configuration given in **Table 2**, following Lample et al. (2016).

TABLE 1 | Results of the experiments on the model architecture.

Word embeddings	Character features	Output	F1 score
Txrain word2vec	BiLSTM	crf	88.36
Train word2vec		crf	87.36
Train word2vec	CNN	crf	87.29
Word2vec	BiLSTM	crf	86.85
Word2vec	CNN	crf	86.16
Train word2vec	BiLSTM	softmax	86.12
Train	BiLSTM	crf	86.10
Word2vec	BiLSTM	softmax	85.96
Train		crf	85.88
Train	CNN	crf	85.56
Train word2vec	CNN	Softmax	85.47
Train word2vec		Softmax	85.41
Word2vec		crf	84.95
word2vec	CNN	Softmax	84.45
Train	BiLSTM	softmax	83.99
Word2vec		Softmax	83.91
Train	CNN	Softmax	83.61
	BiLSTM	crf	83.06
Train		Softmax	83.06
	BiLSTM	Softmax	82.05
	CNN	crf	78.23
	CNN	Softmax	75.28

Configurations are sorted according to the F1 testing score, in decreasing order. A blank cell indicates that the specific component was not included.

TABLE 2 | Configuration for the experiments on model architecture.

Layer	Parameter	Value
Word embeddings	Dimensionality	300
	Min word frequency	5
Character-level word embeddings	Embedding dimensionality	100
	BiLSTM dimensionality	100
BiLSTM	Dimensionality	64
CRF	Metric	Viterbi
Early stopping	Max waiting	5
	Max number of epochs	25
Model	Optimizer—CRF prediction	RMSprop
	Optimizer—Softmax prediction	Adam
	Dropout	0.5
	Learning rate	0.001
	Decay	0
	Batch size	50

Word embeddings can be integrated in a model architecture in three ways:

1. *Train*: Word embeddings initialized at random and trained. This configuration is also known in the literature as random initialization.

[9] The F1 score is the harmonic mean of precision and recall calculated considering every classification action independently.

TABLE 3 | Results of the fine-tuning of the best multi-task architecture, over the batch size, the dimensionality of the inner BiLSTM and the rate of dropout.

Batch size	Dropout	BiLSTM size	Training Task I	Validation Task I	Training Task II	Validation Task II	Training Task III	Validation Task III
100	0.5	200	0.8597	0.8613	0.8010	0.7990	0.9396	0.9316
30	0.5	200	**0.8840**	**0.8952**	**0.8236**	0.8077	0.9073	0.9053
70	0.5	200	0.8804	0.8885	0.8166	0.8005	0.9455	**0.9391**
100	0.7	200	0.8693	0.8837	0.8044	0.8052	**0.9460**	0.9359
30	0.7	200	0.8701	0.8776	0.8051	0.8015	0.9039	0.8998
100	0.5	100	0.8817	0.8882	0.8157	**0.8107**	0.9386	0.9323
100	0.5	30	0.8606	0.8671	0.8021	0.7778	0.9113	0.9076

2. *Word2vec*: pre-trained Word2vec embeddings without further task-specific tuning. Also known as static word embeddings.
3. *Train Word2vec*: Word embeddings initialized with pre-trained Word2vec embeddings and further tuned on the specific task during training. Also known as non-static word embeddings.

Our results strongly support the use of word embeddings, and also indicate that the pre-trained word embeddings carry useful information for the task at hand.

The contribution of character-level word embeddings is instead less impactful, especially as there seems to be a substantial overlap with the contribution of word embeddings: there is only a 1% gain in the best model using both word and BiLSTM character-level word embeddings. Notably, the CNN approach appears to perform less well then the BiLSTM, despite the fact that the gain in speed of a CNN architecture is considerable (3 times faster training, on average). We therefore confirm the relatively low impact of character-level word embeddings, as previously discussed in the literature (Reimers and Gurevych, 2017b), but find that a BiLSTM slightly outperforms a CNN approach for our task.

With respect to the prediction layer, as expected the CRF approach consistently outperforms the softmax, yielding a gain of above 2% when compared with an identical architecture using non-static word embeddings and BiLSTM character-level word embeddings. This result follows from the intuition that tag predictions are not independent in a reference. We eventually tested a multi-task architecture where all layers are shared across the three tasks, besides for the prediction one. Our results are quite encouraging, with performances lowering on average less than 0.5% from the equivalent single-task architecture (**Table 3**). It follows that the input and inner layers learn a set of parameters which are to a large degree shared across tasks.

As discussed in the previous section, the best model is a BiLSTM-CRF network with word embeddings and character-level word embeddings. We fine-tuned this architecture over a set of parameter ranges using grid search, with results presented in **Table 4**.

The results reported in **Table 4** outline the importance of the BiLSTM dimensionality. The best predictions were achieved with a dimensionality of 100 and a medium rate of dropout (0.5), without affecting the running time. The batch size is the parameter with the most influence on the training time: the smaller the batch, the longer the training. A second round of fine-tuning on the best model yielded some further minor improvements, given in **Table 5**. Eventually, **Table 6** reports the final configuration of our best model.

TABLE 4 | Results of the fine-tuning of the best architecture, over the batch size, the dimensionality of the inner BiLSTM and the rate of dropout.

Batch	Dropout	BiLSTM	Testing F1 score
100	0.5	200	89.09
30	0.5	200	88.96
70	0.5	200	88.95
70	0.7	200	88.61
100	0.2	200	88.51
100	0.7	200	88.41
100	0.5	80	88.36
70	0.2	200	88.19
30	0.2	200	88.08
30	0.2	80	88.00
70	0.5	80	87.97
70	0.2	80	87.89
30	0.5	80	87.84
100	0.2	80	87.78
30	0.2	40	87.63
30	0.7	200	87.32
100	0.5	40	87.23
70	0.5	40	87.17
100	0.7	80	86.81
70	0.7	80	86.80
70	0.2	40	86.79
30	0.5	40	86.71
100	0.2	40	86.70
30	0.7	80	86.29
100	0.7	40	84.60
70	0.7	40	83.68
30	0.7	40	83.55

5.2. Evaluation

We report in what follows the validation of the best model, and a discussion of the errors. Some figures and tables are given in the Appendix.

TABLE 5 | Results of the further fine-tuning of the best architecture, over the batch size, the dimensionality of the inner BiLSTM and the rate of dropout.

Batch	Dropout	BiLSTM	Testing F1 score
100	0.5	400	89.56
200	0.5	400	89.24
100	0.5	600	89.13
100	0.5	300	88.99
100	0.5	200	88.89
200	0.5	300	88.61

TABLE 6 | Configuration of the final best model.

Layer	Parameter	Value
Word embeddings	Dimensionality	300
	Min word frequency	5
Character-level word embeddings	Embedding dimensionality	100
	BiLSTM dimensionality	100
BiLSTM	Dimensionality	400
CRF	Metric	Viterbi
Early stopping	Max waiting	5
	Max number of epochs	25
Model	Optimizer	RMSprop
	Dropout	0.5
	Learning rate	0.001
	Decay	0
	Batch size	100

TABLE 7 | Classification report for Task 1.

	Precision	Recall	f1-score	Support
Abbreviation	0.1333	0.0460	0.0684	87
Archivalreference	0.8163	0.4878	0.6107	328
Archive_lib	0.2857	0.8235	0.4242	17
Attachment	0.0000	0.0000	0.0000	0
Author	0.8928	0.9742	0.9317	4581
Box	1.0000	1.0000	1.0000	6
Cartulation	0.0000	0.0000	0.0000	10
Column	1.0000	1.0000	1.0000	6
Conjunction	0.4778	0.7167	0.5733	120
Date	0.6667	0.3158	0.4286	19
Filza	0.8333	0.2143	0.3409	70
Folder	0.0000	0.0000	0.0000	0
Foliation	0.0000	0.0000	0.0000	0
Numbered_ref	0.0000	0.0000	0.0000	87
o	0.8066	0.4445	0.5732	1379
Pagination	0.9504	0.9801	0.9650	1154
Publicationnumber-year	0.8874	0.8767	0.8820	665
Publicationplace	0.9569	0.9421	0.9494	1555
Publicationspecifications	0.4068	0.3982	0.4025	329
Publisher	0.8941	0.8196	0.8552	937
Ref	0.2576	0.4722	0.3333	36
Registry	0.7447	1.0000	0.8537	35
Series	0.7949	0.7209	0.7561	43
Title	0.9390	0.9651	0.9519	13744
Tomo	0.3030	0.3030	0.3030	33
Volume	0.7822	0.5254	0.6286	335
Year	0.9088	0.9582	0.9328	1601
Avg/total	0.9006	0.9022	0.8966	27177

TABLE 8 | Classification report for Task 2.

	Precision	Recall	f1-score	Support
b-meta-annotation	0.7473	0.7500	0.7487	280
b-primary	0.6957	0.3556	0.4706	45
b-secondary	0.7737	0.7022	0.7362	779
e-meta-annotation	0.7970	0.8532	0.8242	879
e-primary	0.4382	0.2335	0.3047	167
e-secondary	0.8399	0.7789	0.8083	1583
i-meta-annotation	0.7594	0.8269	0.7917	8457
i-primary	0.5772	0.8444	0.6857	270
i-secondary	0.8687	0.8475	0.8580	13682
o	0.7950	0.5507	0.6507	1035
Avg/total	0.8181	0.8162	0.8151	27177

- On Task 1 (**Table 7**), the model achieves an F1 score of 89.66% on the validation dataset, outperforming our CRF baseline by +7.03%. The model performs particularly well on the two most represented tags (*title* and *author*): these two tags combined account for more than the 2/3 of the dataset. All tags with 500 or more examples in the validation dataset perform quite well, at the exception of the *o* and *publisher* tags. The *o* tags are probably both not well represented and difficult to grasp (too generic). When compared with the CRF baseline, in Table S1 (Appendix), the neural network approach performs better for the *title* and *author* tags, and the vast majority of the rest, especially so for the *publisher* and *o* tags.
- On Task 2 (**Table 8**), the model achieves an F1 score of 81.51% on the validation dataset, and outperforms the CRF baseline by +10.47% (Table S2 in Appendix). The model performs well overall for the most represented tags in the dataset, such as the *i-* tags, but it shows issues with the begin and end *primary* annotations, that are often difficult to capture. The lower results of the model on this task, if compared with tasks 1 and 3, suggests that distinguishing between primary or secondary references might not be a sequence labeling problem but a classification one, over the entire line/reference.
- On Task 3 (**Table 9**), the model achieves an F1 score of 95.09% on the validation dataset, and outperforms the CRF baseline by +2.59% (Table S3 in Appendix). In particular, the model improves on the *o*, begin and in tags, while lowering its performance on the end tag.

We further discuss the error confusion matrices over the validation dataset, in order to compare the proportion of

TABLE 9 | Classification report for Task 3.

	Precision	Recall	f1-score	Support
b-r	0.8498	0.7636	0.8044	1104
e-r	0.8963	0.7703	0.8286	1145
i-r	0.9633	0.9904	0.9767	23893
o	0.8491	0.5217	0.6463	1035
Avg / total	0.9515	0.9541	0.9509	27177

classifications gone well or wrong, for each tag. Starting with Task 1 (Figure S1 in Appendix), we can see how systematic errors tend to be caused by two reasons: under-represented tags or very similar encoding styles or contents for different styles. Examples are the *date* tag mistaken for a *year*, or an *abbreviation* mistaken for an *author* (initials). The confusion matrix for Task 2 (Figure S2 in Appendix), broadly follows along the same lines, further highlighting how most frequent tags tend to ac as attractors of wrong classification actions. Indeed, the tag *i-secondary* is often misassigned. Interestingly, *i-secondary* tags are sometimes predicted as *i-meta-annotation*, the second most frequent tag in the training dataset: indeed, their contents are often very similar. Quite crucially, when a prediction is wrong it is often assigned to the correct IBOE tag, but the wrong reference type. This would allow to adopt a voting system to refine a classification at a further stage. The confusion matrix for Task 3 (Figure S3 in Appendix), shows that the inside tag is correctly predicted, but reveals a fragility in the *e-r* tag predictions. Indeed, a lot of *e-r* tags are labeled as *i-r* by the model. The model also performs poorly in predicting the out-of-reference *o* tag.

In conclusion, the neural network model substantially outperformed the CRF baseline in all tasks, with minor downgrade of performance on some infrequent tags, but an important gain on most of the rest. All systematic errors can be explained either by the important imbalance in the amount of training examples per tag, or by the similarity in either contents or referencing styles between some tags.

6. CONCLUSION

In this work, we applied a state-of-the-art deep learning architecture to the task of reference mining, with a focus on applications in the arts and humanities. In particular, the model is trained to extract and parse references within the full text of publications, such as in footnotes, yet it can be applied more generally. The final architecture follows previous work in sequence labeling tasks, by integrating word embeddings and character-level word embeddings into word representations as inputs, an inner BiLSTM layer and a CRF prediction layer. As was shown for a variety of similar tasks, important components of the network result to be pre-trained word embeddings, which integrate information on the use of words within a broader textual corpus, and the CRF prediction layer, which accounts for the dependency among tag predictions (Reimers and Gurevych, 2017a). Furthermore, for the specific task at hand, we showed the relative positive contribution of character-level word embeddings. Given the importance of morphological and orthographical features in references, and the lack of large quantities of annotated data to learn word representations from, character-level features proved to be a minor yet positive addition. This model was tested on a dataset of annotated references extracted from a corpus of scholarly literature on the history of Venice, and it improved considerably over a CRF baseline using a rich set of hand-crafted features, with F1 gains going from +2.59% to +10.47% on different tasks. Furthermore, a multi-task architecture was found to perform almost on par on all tasks combined. We released two implementations of the architecture, in Keras and TensorFlow, along with all the data we used to train and test it. These results strongly support the adoption of deep learning methods for the general task of reference mining.

This work used a relatively small dataset with some limitations, reflecting the current situation with respect to reference mining and, more broadly, citation indexing in the arts and humanities. The dataset contains several sources of noise, including OCR errors, referencing errors or inconsistencies, annotation errors. In part for this reason, we consider as the most important next step for future work to explore how active learning or semi-supervised learning techniques might be used in order to maximize the model gain while at the same time minimizing the costly process of manual annotation (Peters et al., 2017; Shen et al., 2018). At the same time, we plan to explore how to align and use existing annotated datasets with coverage in the arts and humanities (Anzaroot and McCallum, 2013). Furthermore, it remains to be tested to what extend reference parsers trained on scientific publications could be adapted for the literature in the arts and humanities.

AUTHOR CONTRIBUTIONS

GC conceived the research project, DR and GC developed code and experiments and wrote the paper, FK contributed with advice.

FUNDING

GC was supported by the Swiss National Fund with grant number 205121_159961 for part of this research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frma.2018.00021/full#supplementary-material

Supplementary Data

The dataset and code for this article is available at the following address on GitHub: https://github.com/dhlab-epfl/LinkedBooksDeepReferenceParsing, the pre-trained word vectors are available on Zenodo at the following address: http://doi.org/10.5281/zenodo.1175212.

REFERENCES

Anzaroot, S. and McCallum, A. (2013). *A New Dataset for Fine-Grained Citation Field Extraction.* Atlanta, GA: JMLR: W&CP.

Ardanuy, J. (2013). Sixty years of citation analysis studies in the humanities (1951-2010). *J. Am. Soc. Inform. Sci. Technol.* 64, 1751–1755. doi: 10.1002/asi.22835

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* 5, 157–166. doi: 10.1109/72.279181

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly.

Chiu, J. P., and Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* 4, 357–370.

Chollet, F. et al. (2015). *Keras*. Availbale online at: https://github.com/keras-team/keras.

Colavizza, G., and Romanello, M., (2017). Annotated References in the Historiography on Venice: 19th-21st centuries. *J. Open Human. Data.* 3:2. doi: 10.5334/johd.9

Colavizza, G., Romanello, M., and Kaplan, F. (2017). The references of references: a method to enrich humanities library catalogs with citation data. *Int. J. Digit. Libr.* 18, 1–11. doi: 10.1007/s00799-017-0210-1

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* 12, 2493–2537.

Councill, I. G., Giles, C. L., and Kan, M.-Y. (2008). "ParsCit: an open-source CRF reference string parsing package," in *Proceedings of the Language Resources and Evaluation Conference (LREC 2008)* (Marrakesh).

Dai, H.-J., Lai, P.-T., Chang, Y.-C., and Tsai, R. T.-H. (2015). Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. *J. Cheminform.* 7:S14. doi: 10.1186/1758-2946-7-S1-S14

Dean, J., Monga, R., and Google Research (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Available online at: https://www.tensorflow.org/.

Dos Santos, C., and Gatti de Bayser, M. (2014). "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers* (Santa Fe), 69–78.

Dos Santos, C., and Zadrozny, B. (2014). "Learning character-level representations for part-of-speech tagging," in *ICML, Vol. 32 of JMLR Workshop and Conference Proceedings* (JMLR.org), 1818–1826.

Garfield, E. (1979). *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. New York, NY: John Wiley & Sons.

Graves, A., Mohamed, A.-R., and Hinton, G. (2013). "Speech recognition with deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (icassp), 2013 IEEE International Conference on IEEE* (Budapest), 6645–6649.

Graves, A., and Schmidhuber, J. (2005). "Framewise phoneme classification with bidirectional lstm networks," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005, Vol. 4*, 2047–2052.

Heckmann, D., Frank, A., Arnold, M., Gietz, P., and Roth, C. (2016). Citation segmentation from sparse and noisy data: a joint inference approach with Markov logic networks. *Digit. Schol. Hum.* 31, 333–356. doi: 10.1093/llc/fqu061

Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J. (2001). "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *A Field Guide to Dynamical Recurrent Neural Networks*, eds, S. C. Kremer and J. F. Kolen (IEEE Press).

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint*.

Kim, Y. (2014). "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha), 1746–1751.

Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2015). "Character-aware neural language models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)* (Arizona), 2741–2749.

Körner, M., Ghavimi, B., Mayr, P., Hartmann, H., and Staab, S. (2017). Evaluating reference string extraction using line-based conditional random fields: a case study with German language publications. *New Trends in Databases and Information Systems*, Vol. 767, eds M. Kirikova, K. Nørvåg, G. A. Papadopoulos, J. Gamper, R. Wrembel, J. Darmont, and S. Rizzi (Cham: Springer International Publishing), 137–145.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 282–289.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *CoRR* abs/1603.01360. doi: 10.18653/v1/N16-1030

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 541–551. doi: 10.1162/neco.1989.1.4.541

Lopez, P. (2009). "GROBID: combining automatic bibliographic data recognition and term extraction for scholarship publications," in *Research and Advanced Technology for Digital Libraries* (Springer), 473–474.

Ma, X., and Hovy, E. H. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *CoRR*, abs/1603.01354. doi: 10.18653/v1/P16-1101

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Mongeon, P., and Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics* 106, 213–228. doi: 10.1007/s11192-015-1765-5

Okazaki, N. (2007). *Crfsuite: A Fast Implementation of Conditional Random Fields*. Available online at: http://www.chokkan.org/software/crfsuite/

Peters, M. E., Ammar, W., Bhagavatula, C., and Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*

Ratinov, L., and Roth, D. (2009). "Design challenges and misconceptions in named entity recognition," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, (Stroudsburg, PA: Association for Computational Linguistics), 147–155.

Řehůřek, R., and Sojka, P. (2010). "Software Framework for Topic Modelling with Large Corpora" in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (ELRA: Valletta), 45–50.

Rei, M., Crichton, G. K., and Pyysalo, S. (2016). Attending to characters in neural sequence labeling models. *arXiv preprint arXiv:1611.04361*.

Reimers, N., and Gurevych, I. (2017a). Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*.

Reimers, N., and Gurevych, I. (2017b). Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. *CoRR*, abs/1707.09861.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098.

Rumelhart, D. E. (1986). Learning internal representations by error propagation. *Nature* 323, 533–536. doi: 10.1038/323533a0

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 2673–2681. doi: 10.1109/78.650093

Shen, S.-S., and Lee, H.-Y. (2016). Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection. *arXiv preprint arXiv:1604.00077*.

Shen, Y., Yun, H., Lipton, Z., Kronrod, Y., and Anandkumar, A. (2018). "Deep active learning for named entity recognition," in *Proceedings of the 2nd Workshop on Representation Learning for NLP* (Vancouver, BC), 252–256.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Strubell, E., Verga, P., Belanger, D., and McCallum, A. (2017). “Fast and accurate entity recognition with iterated dilated convolutions,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen), 2660–2670.

Sula, C. A., and Miller, M. (2014). Citations, contexts, and humanistic discourse: toward automatic extraction and classification. *Liter. Linguist. Comput.* 29, 452–464. doi: 10.1093/llc/fqu019

Sutton, C., and McCallum, A. (2011). An introduction to conditional random fields. *Found. Trends Mach. Learn* 4, 267–373. doi: 10.1561/2200000013

Tkaczyk, D., Collins, A., Sheridan, P., and Beel, J. (2018). Evaluation and comparison of open source bibliographic reference parsers: a business use case. *arXiv preprint arXiv:1802.01168*.

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., and Bolikowski, L. (2015). CERMINE: automatic extraction of structured metadata from scientific literature. *Int. J. Doc. Anal. Recogn.* 18, 317–335. doi: 10.1007/s10032-015-0249-8

Yang, Z., Salakhutdinov, R., and Cohen, W. (2016). Multi-Task Cross-Lingual Sequence Tagging from Scratch. arXiv:1603.06270 [cs]

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. *ArXiv e-prints*.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Research Metrics and Analytics

ORIGINAL RESEARCH
published: 19 September 2018
doi: 10.3389/frma.2018.00027

Temporal Representations of Citations for Understanding the Changing Roles of Scientific Publications

Jiangen He and Chaomei Chen**

Department of Information Science, College of Computing and Informatics, Drexel University, Philadelphia, PA, United States

Researchers may describe different aspects of past scientific publications in their publications and the descriptions may keep changing in the evolution of science. The diverse and changing descriptions (i.e., citation contexts) on a publication characterize the impact and contributions of the past publication. In this article, we aim to provide an approach to understanding the changing roles of a publication characterized by its citation contexts in the full text of publications. We proposed approaches for representing the changing citation contexts of cited publications in different periods as sequences of vectors by training temporal embedding models. We can utilize the temporal representations to quantify how much the roles of publications changed and interpret how they changed. We also evaluated the performance of three ways of constructing citation contexts for representation learning. Our study in the biomedical domain shows that our metric on the changes of publication roles is stable at the group level but it can account for the variation of individual publications.

Keywords: in-text citation, citation context, embedding learning, citation analysis, document representation, full-text literature

Edited by:
Marc Bertin,
Claude Bernard University Lyon 1,
France

Reviewed by:
Nicolás Robinson-Garcia,
Universitat Politècnica de València,
Spain
Benjamin Vargas-Quesada,
Universidad de Granada, Spain
Jiang Wu,
Wuhan University, China

****Correspondence:***
Jiangen He
jiangen.he@drexel.edu
Chaomei Chen
chaomei.chen@drexel.edu

Received: *18 February 2018*
Accepted: *27 August 2018*
Published: *19 September 2018*

Citation:
He J and Chen C (2018) Temporal Representations of Citations for Understanding the Changing Roles of Scientific Publications.
Front. Res. Metr. Anal. 3:27.
doi: 10.3389/frma.2018.00027

1. INTRODUCTION

Although the content of a scientific publication cannot be changed once it was published, how other researchers cite and evaluate the publication may keep changing. The actual scientific contributions and impact of a specific publication are changing within the evolving intellectual spaces constructed by other publications. Besides the role of a publications are ever-changing, the role of may be complex because of the varied contributions made the publication, especially the publications contributed to interdisciplinary or fundamental research topics. The changing and complex roles of cited publications can be characterized by their citations and citation contexts.

Both citation network and citation context (i.e., the sentences containing in-text citations) can be utilized for analyzing scientific publications (Elkiss et al., 2008). A relevant intellectual structure characterized by citation network is commonly used as a foundation for analyzing the role of a publication played in scientific dynamics, such as identifying the place where the analyzed publication is in the intellectual structure (Orosz et al., 2016) or the structural alteration caused by the publication (Chen, 2012). The text of citation contexts were also used to characterize publications for various applications, such as publication summarization (Qazvinian et al., 2010), survey article generation (Mohammad et al., 2009), and information retrieval (Huang et al., 2015). Quantitative metric for quantifying the role changes of publications can be derived from

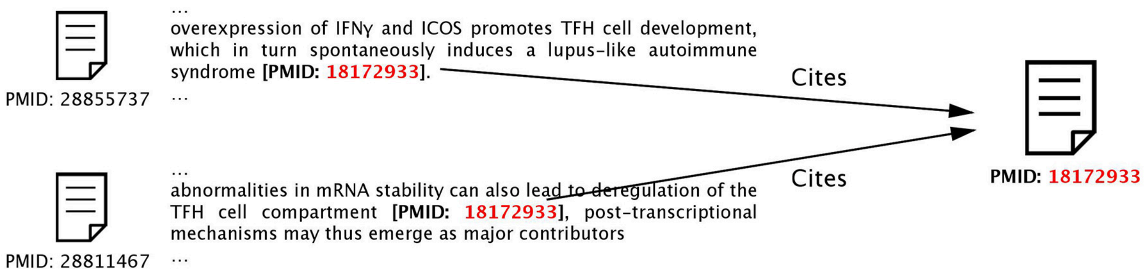

PMID: 28855737
…
overexpression of IFNγ and ICOS promotes TFH cell development, which in turn spontaneously induces a lupus-like autoimmune syndrome [PMID: 18172933].
…
Cites
PMID: 28811467
…
abnormalities in mRNA stability can also lead to deregulation of the TFH cell compartment [PMID: 18172933], post-transcriptional mechanisms may thus emerge as major contributors
…
Cites
PMID: 18172933

FIGURE 1 | Two examples of in-text citations and citation contexts of a PubMed article (PubMed ID: 18172933). This figure is a revised version from a workshop paper that was presented at CLBib-2017 (He and Chen, 2017).

citation network analysis, but interpreting the changes is not straightforward which always relied on techniques of text mining and visual analytics. While analyzing the text of citation contexts naturally has interpretable results but a unified quantitative measurement is challenging to be built on the unstructured textual data.

In this article, we proposed methods for learning temporal representations of in-text citations of publications by word embedding models, which can be used to characterize and analyze the changing roles of the publications. The in-text citations of publications are the citations referred to this publication within the full text of publications cited this publication; The text around the in-text citation is the citation context text (see **Figure 1** for examples). We proposed and compared different ways of constructing the citation contexts for representation learning. Based on the temporal representations of citations, we introduced a simple method to quantify the role changes of publications characterized by their citation contexts. We also analyzed the distribution of change scores and described applications of how to identify and interpret the changes by making use of the embedding representations.

2. RELATED WORK

Due to the availability full-text data of scientific articles, such as PubMed Central, many citation-based studies went beyond the article metadata and citation links. The proximity of citations was combined with co-citation analysis to provide co-citation networks at multiple levels of granularity (Liu and Chen, 2012) or to identify related work (Gipp and Beel, 2009). Citation contexts also has been utilized to improve co-citation network analysis (Callahan et al., 2010; Small and Klavans, 2011; Boyack et al., 2013) and enhance the application of direct citation network (Sugiyama and Kan, 2013). Some studies emphasized the literal features of citation by analyzing the citation context, such author's reason for citing (Teufel et al., 2006) and sentiment of citation (Small, 2011). Besides, various applications based on citation context have been developed, such as information retrieval (Eto, 2013; Liu S. et al., 2014a) and article recommendation (He et al., 2010; Liu X. et al., 2014b).

More recently, embedding learning techniques were employed in representing key elements of scientific knowledge, such as publications (Ganguly and Pudi, 2017), authors (Ganesh et al., 2016), citations (Berger et al., 2017), and research topics (He and Chen, 2018). Paper2vec (Ganguly and Pudi, 2017) leveraged both citation networks and textual information of publications to represent a publication, but the textual information they used is the full text of publications which is the description from authors of publication rather than scientific communities. Another study also named *Paper2vec* (Tian and Zhuo, 2017) focused on utilizing neighbor nodes of publications in citation network to represent the publications. *Cite2vec* (Berger et al., 2017) represented publications by using their citation contexts as ours and provided visualization for exploration, while the temporal feature of citation contexts is ignored. In our study, we emphasize representing the changes of publications characterized by the citation contexts of the publications over time.

3. METHODS

In this section, we describe how we train temporal citation embedding models, which includes data preprocessing, constructing citation contexts, embedding model training over periods, and embedding model alignment. We also present our approach to quantifying the role changes of publications. **Figure 2** describes the overview of our methods.

3.1. Data and Preprocessing

The dataset we use for training is the PubMed Central Open Access Subset (PMC OAS), which is an open access XML formatted full-text document repository from biomedicine and life sciences maintained by the U.S. National Library of Medicine (NLM)[1]. We can parse in-text citations and citation contexts from PMC OAS because of the well-structured XML files. In this study, we trained embedding models by documents published from 2007 to 2016 in PMC OAS, which comprises 1,361,455 full-text scientific publications.

To train the citation embeddings, we need to use a unique identifier to indicate a publication in full text. Many references cited by publications in the PMC OAS have unique publication identifiers such as PubMed ID (PMID), PubMed Central ID (PMCID), and Digital Object Identifier (DOI). However, many cited references don't have unique identifiers. We assign unique

[1] http://www.ncbi.nlm.nih.gov/pmc/

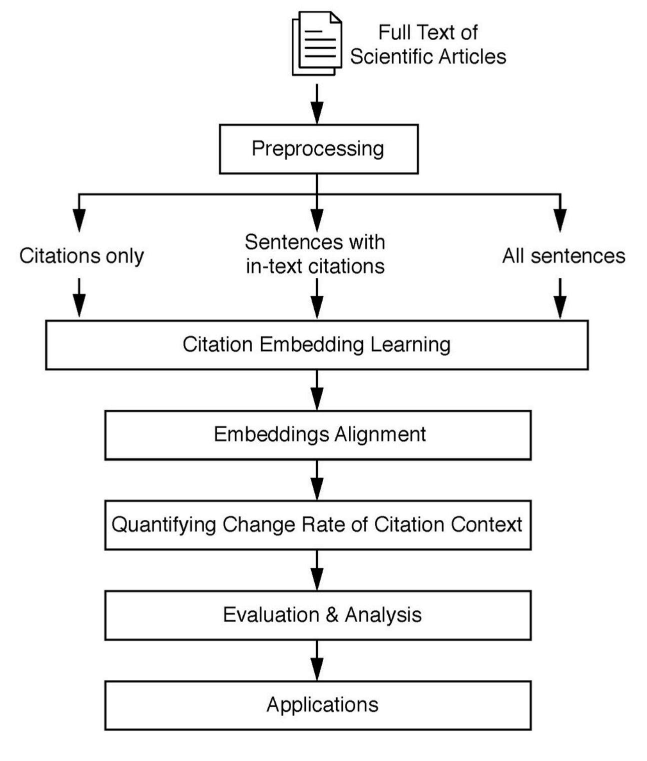

Full Text of Scientific Articles
Preprocessing
Citations only
Sentences with in-text citations
All sentences
Citation Embedding Learning
Embeddings Alignment
Quantifying Change Rate of Citation Context
Evaluation & Analysis
Applications

FIGURE 2 | An overview of methods.

identifiers for these references by using their metadata in the form of 'FA_VE_YR_VO_FP' where FA is the first author's first name and last name, VE is the name of venue (journal or conference proceeding), YR is the year of the publication date, VO is the volume number, and FP is the first page number of the publication. If a cited reference has neither a standard identifier nor identifiable metadata for constructing a unique identifier, the reference will be ignored in this study.

About 5.5% references cannot be identified in the PMC OAS dataset. Excluding these unidentifiable references and their in-text citations may have effects on learning representations of citations, because the cin-text citations are a part of citation contexts for learning. However, the effects may not be significant. First, both words and in-text citations are the citation contexts for learning, but words constitute the major part of the citation contexts. Second, other identified in-text citations and words can work as substitutes to provide effective contextual information to diminish the effects.

To facilitate citation embedding learning, we convert sentences with in-text citation XML tags into plain text. We only retain text and in-text citations by removing XML tags or converting XML tags into text. We replace the in-text citation tag <xref ref-type="bibr"></xref> by using a unique identifier. It is worth noting the various usages of <xref> tag in the XML full text. For example, a single <xref> may refer to a group of citations (see **Figure 3A**) and some in-text citations are not explicit in XML files (the purple citation identifier in **Figure 3B** is the citation omitted in the XML file).

Since our embedding learning method learns the representation of a citation by capturing the context of the citation within its sentence, we need sentence tokenizer to segment the full text into sentences. Then, we conduct a series of preprocessing by using *NLPre*[2], including dash removal, capitalization normalization, and replacing phrase from Medical Subject Headings (MeSH) dictionary.

3.2. Constructing Citation Contexts

We use three methods to construct citation contexts for embedding learning. They retain different context information for learning citation representations. To illustrate the methods, we denote a paragraph p in a scientific publication as two sets of sentences $S = \{s_1, ..., s_i, ..., s_m\}$ and $SC = \{sc_1, ..., sc_j, ..., sc_n\}$, where s_i is a sentence without any in-text citation and sc_j is a sentence with a set of in-text citations $C_j = \{c_1^j, c_2^j, ...c_{k_j}^j | k_j \geq 1\}$.

- **CITATION_ONLY**. We only use citation identifiers for embedding learning. Since a single sentence usually has very few in-text citations, we use a sequence of citation identifiers in a paragraph as an input record. One input sequence that can be derived from paragraph p for **CITATION_ONLY** is $\{c_1^1, ..., c_{k_1}^1, ..., c_1^n, ..., c_{k_n}^n\}$.
- **WITH_CITATION**. We use sentences as input, but only sentences with at least one in-text citation are used. m input sequences that can be derived from paragraph p are $sc_1, sc_2, ..., sc_n$.
- **FULL_SET**. All sentences in the full text of articles are used for embedding learning. $m+n$ input sequences that can be derived from paragraph p are $s_1, ..., s_m, sc_1, ..., sc_n$.

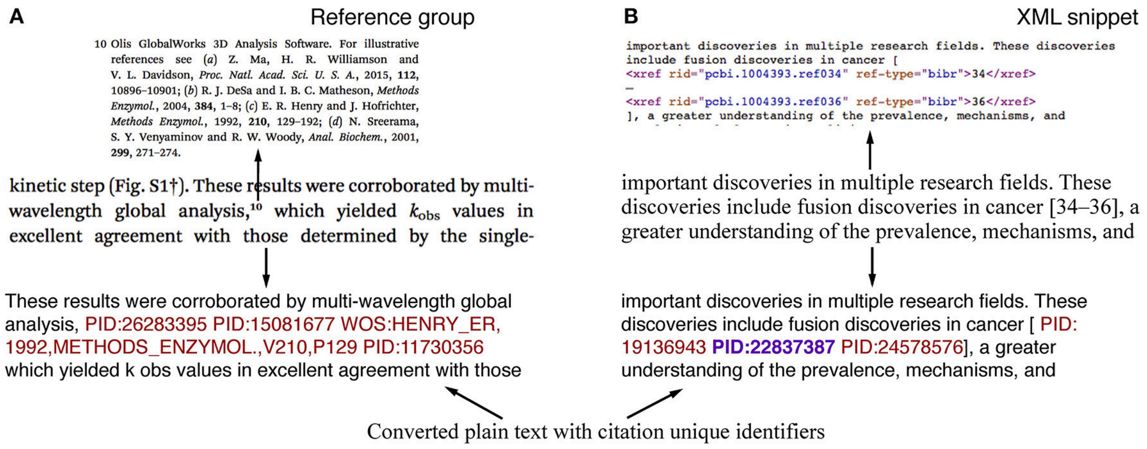
3.3. Embedding Learning Methods

We build citation embeddings for understanding how researchers described cited publications. Word embedding techniques were proved to be able to capture semantic and syntactic effectively (Mikolov et al., 2013). We use skip-gram with negative sampling (SGNS) to learn citation embedding based on the context words of citation in the full text of publications. Given a citation or a phrase w_i in training dataset, skip-gram maps it into a continuous representation vector $\mathbf{w}_i$. $\mathbf{w}_i$ is used to predict the context words of w_i. The objective of skip-gram is to maximize the log probability:

$$\frac{1}{T}\sum_{T}^{i=1}\sum_{i-c\leqslant j\leqslant i+c}\log p(w_j|w_i) \tag{1}$$

where T is the occurrence of each word or citation in the training data, c is the window size of context and w_j is the context of w_i. Negative sampling builds "negative" context words for each w_i to accelerate the training procedure. We separately constructed citation embeddings from publication text data for each period by SGNS algorithm.We used the implementation of word2vec provided by gensim (Řehůřek and Sojka, 2010) for embedding learning. We empirically set embedding length as 100, negative sampling size as 5, and the number of iteration

[2] https://pypi.python.org/pypi/nlpre


A
Reference group
B
XML snippet
Converted plain text with citation unique identifiers

10 Olis GlobalWorks 3D Analysis Software. For illustrative references see (*a*) Z. Ma, H. R. Williamson and V. L. Davidson, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 10896–10901; (*b*) R. J. DeSa and I. B. C. Matheson, *Methods Enzymol.*, 2004, **384**, 1–8; (*c*) E. R. Henry and J. Hofrichter, *Methods Enzymol.*, 1992, **210**, 129–192; (*d*) N. Sreerama, S. Y. Venyaminov and R. W. Woody, *Anal. Biochem.*, 2001, **299**, 271–274.

kinetic step (Fig. S1†). These results were corroborated by multi-wavelength global analysis,[10] which yielded k_{obs} values in excellent agreement with those determined by the single-

These results were corroborated by multi-wavelength global analysis, PID:26283395 PID:15081677 WOS:HENRY_ER, 1992,METHODS_ENZYMOL.,V210,P129 PID:11730356 which yielded k obs values in excellent agreement with those

```
important discoveries in multiple research fields. These discoveries
include fusion discoveries in cancer [
<xref rid="pcbi.1004393.ref034" ref-type="bibr">34</xref>
–
<xref rid="pcbi.1004393.ref036" ref-type="bibr">36</xref>
], a greater understanding of the prevalence, mechanisms, and
```

important discoveries in multiple research fields. These discoveries include fusion discoveries in cancer [34–36], a greater understanding of the prevalence, mechanisms, and

important discoveries in multiple research fields. These discoveries include fusion discoveries in cancer [PID: 19136943 PID:22837387 PID:24578576], a greater understanding of the prevalence, mechanisms, and

FIGURE 3 | Two examples of converting XML into plain text.

as 5. However, we set different window size for each type of citation context. For **ONLY_CITATION**, we set a relatively small window size 3 because of a small length of input sequences and technical practices of training word embeddings of short text. For **WITH_CITATION** and **FULL_SET**, we set a larger window size 10. Many in-text citations were placed at the end of sentences, so learning model with a large window size can capture essential context information.

3.4. Temporal Embedding Alignment

Since our embedding models are constructed separately for different periods, the models are in different vector space because of differences in stochastic initialization of the weights of the neural network in SGNS algorithm. We need to align the models for different periods into the same coordinate axes to compare citation representations overtime and quantify the citation changes of articles. Following the method proposed by Hamilton et al. (2016), we use orthogonal Procrustes to align the learned embeddings. Defining $\mathbf{W}^{(t)} \in \mathbb{R}^{d \times |\nu|}$ as the matrix of word embeddings learn at period t, we align across time periods while preserving cosine similarities by optimizing

$$\mathbf{R}^{(t)} = \arg \min_{\mathbf{Q}^{\mathrm{T}}\mathbf{Q}=\mathbf{I}} \left\| \mathbf{Q}\mathbf{W}^{(t)} - \mathbf{W}^{(t+1)} \right\|_F \tag{2}$$

with $\mathbf{R}^{(t)} \in \mathbb{R}^{d \times d}$. The alignment is performed in an iterative fashion, i.e., $(\mathbf{W}^{(1)}, \mathbf{W}^{(2)}), (\mathbf{W'}^{(2)}, \mathbf{W}^{(3)}), ..., (\mathbf{W'}^{(T-1)}, \mathbf{W}^{(T)})$ where $\mathbf{W'}^{(t)}$ is the aligned matrix of word embeddings at t, an alignment of $(\mathbf{W'}^{(t-1)}, \mathbf{W}^{(t)})$ produces an aligned matrix $\mathbf{W'}^{(t)}$, and T is the last time-period.

3.5. Quantify the Changes of Citation Contexts

The representations of citations can be compared over periods after aligning the citation embeddings over time. The difference between representations of a cited article over periods can be utilized to quantify the change rate of the article's citation contexts. We measure the difference based on commonly used cosine similarity. Therefore, we quantify the citation change rate of a cited article ca occurred at t as

$$Change^t(ca) = 1 - \text{cos\_sim}(\mathbf{w}^t_{ca}, \mathbf{w}^{t-1}_{ca}) \tag{3}$$

where $\mathbf{w}^t_{ca}$ is the vector representation of article ca at t derived from the citation contexts of ca at t.

3.6. Evaluation Metric Based on MeSH

We can train the citation embeddings by using three types of citation context described in section 3.2. To evaluate their ability to quantify the change rates of citations of cited articles, we propose an evaluation metric based on MeSH and evaluated the three types of citation context.

We use MeSH to derive an implicit gold standard concerning the topical changes of a publication's citations. Most of publications in this MEDLINE/PubMed (88.25%[3]) were manually assigned a set of descriptors from MeSH by biomedical experts at the U.S. National Library of Medicine (NLM). Similar to multiple prior studies on measuring document similarity (Zhu et al., 2009; Gipp et al., 2015), We view MeSH indexing as an accurate topical description of biomedical articles. The assigned MeSH descriptors of a article collection can describe the topical information of the article collection, so the change over time of assigned MeSH descriptors of the collection of articles cited a certain article can reflect the topical change of the article's citations. Although our temporal citation representation is not designed for representing topical change, a good representation should reflect the topic change as well. Thus, we build an evaluation metric based on the MeSH indexing.

We create the evaluation metric by following the approaches used by CITREC (Gipp et al., 2015) which is evaluation framework for citation-based and text-based similarity measures of documents. However, the evaluation metric proposed by CITREC is for document similarity rather than citation change

[3] https://www.nlm.nih.gov/bsd/licensee/2017_stats/2017_LO.html

of a cited article. Thus, we modified the approaches of CITREC and proposed a metric for measuring the citation change as our evaluation metric. At first, we measure the similarity of MeSH descriptors by utilizing the tree-like structure of MeSH thesaurus. Then, we measure the topical change over time of citations based on the topical information derived from the assigned MeSH descriptors of the citations.

A MeSH descriptor may have multiple tree numbers, which means a descriptor can occur multiple times within the tree structure of MeSH thesaurus. We view the tree numbers as different concepts. To measure the similarity of descriptors, we need to measure the similarity of the concepts behind the descriptors at first. The basic idea of measuring the similarity of two concepts c and c' is that the similarity reflects the information they have in common (Gipp et al., 2015). We use the assessment of information content (IC) proposed by Resnik (1995) to quantify the common information of concepts. We quantify information content IC of a concept c by a negative log-likelihood function as

$$IC(c) = -\log \frac{1 + s(c)}{N} \tag{4}$$

where $s(c)$ is the number of concepts subsumed to concept c and N is the total number of concepts in the MeSH thesaurus ($N = 58,760$). The common information content of two concepts c and c' can be represented as information content of their closest subsuming concept $c_s(c, c')$. Then, we calculate the similarity of c and c' using Lin's generic similarity measure (Lin et al., 1998) as

$$\mathrm{sim}(c, c') = \frac{2 \times IC(c_s(c, c'))}{IC(c) + IC(c')} \tag{5}$$

To measure the similarity of two MeSH descriptors m and m', we compare the sets of the descriptors' concepts C and C'. We use the average maximum match, a similarity measure proposed by Zhu et al. (2009), to calculate the similarity of two MeSH descriptors m and m' as

$$\mathrm{sim}(m, m') = \frac{\sum_{c \in C} \max_{c' \in C'} \mathrm{sim}(c, c') + \sum_{c' \in C'} \max_{c \in C} \mathrm{sim}(c, c')}{|C| + |C'|} \tag{6}$$

The similarity of citations of two cited articles ca and ca' is determined by the similarity of two sets of articles $D = \{d | d \text{ cited } ca\}$ and $D' = \{d' | d' \text{ cited } ca'\}$. We use the average maximum match between the two sets of MeSH descriptors M and M' assigned to citing articles in D and D' respectively to measure the citation similarity of ca and ca' as

$$\begin{aligned} \text{citation-sim}(ca, ca') &= \mathrm{sim}(D, D') = \mathrm{sim}(M, M') \\ &= \frac{\begin{array}{c}\sum_{m \in M} c(m) \times \max_{m' \in M'} \mathrm{sim}(m, m') \\ + \sum_{m' \in M'} c(m') \times \max_{m \in M} \mathrm{sim}(m, m')\end{array}}{|M| + |M'|} \end{aligned} \tag{7}$$

where $c(m)$ is the frequency of m in the descriptor set M and $c(m')$ is the frequency of m' in the descriptor set M'.

The change of citation of a cited article ca at period t is determined by the similarity of ca's citations at t and $t - 1$. It is computed as

$$\mathrm{Change}^{t}_{topic}(ca) = 1 - \mathrm{sim}(D^{t-1}, D^{t}) \tag{8}$$

where $D^t = \{d | d \text{ cited } ca \text{ and published at } t\}$.

4. RESULTS

In this section, we compared three ways of constructing citation contexts to identify the possibly best practice for representation learning. Based on the choice of constructing citation contexts, we preliminarily investigated the characteristics and the patterns of citation context changes by applied our proposed metric on PMC OAS dataset. At last, we conducted two simple applications to show the practical potential of our proposed representation learning method and metric.

4.1. Data Description

We used full-text scientific articles from PMC OAS in a recent decade for our analysis and divided the decade into five periods for further analysis. The articles without a citation of identifiable articles in the full text were excluded. 1,205,407 publications have at least one effective citing sentence, and they have 31 citing sentences on average. In recent 6 years, much more articles with references have been available in PMC OAS. Each cited article roughly received 3 in-text citations on average within each period. In this study, we aim to represent cited articles by their citation contexts, so we focus on the cited articles (CA) which have enough citation context information for representation learning. We identified cited articles with more 50 in-text citations for further analysis. We show the data descriptions in **Table 1**.

4.2. Comparison Results

To compute the change score of a cited article ca at t, both representation of ca at $t - 1$ and t would be used. For each t, only ca has more than 50 in-text citations at both $t - 1$ and t were used in this evaluation for the robustness. We used 11,628 cited articles within the five periods for evaluation.

We used Spearman's rank correlation and Kendall's tau correlation analysis for evaluation. The correlation analysis allow comparing the similarity of ordered two types of changes scores. The results of correlation analysis is shown in **Table 2**. The results of two analysis are consistent. The change scores derived from three types of citation context are significantly correlated with the topical change score. The results of **WITH_CITATION** have highest correlation coefficients. **WITH_CITATION** can produce citation representations reflecting topical changes best. Therefore, we used citation representations derived from **WITH_CITATION** for further investigation.

4.3. Distribution of Change Scores

We computed the change score for cited articles at each period and observed the average of the scores by five groups (see

TABLE 1 | Publications and cited publications in PMC OAS from 2007 to 2016.

Period	Articles	Articles with references	Cited articles	Times cited	Cited articles[a] with citations > 50
2007–2008	69,394	49,105	1,431,012	3,138,614	425
2009–2010	138,204	92,773	2,523,797	6,265,277	1,568
2011–2012	252,714	225,101	4,283,203	12,288,738	4,841
2013–2014	398,620	360,902	6,367,475	19,653,641	8,994
2015–2016	502,523	477,526	8,019,222	24,905,208	11,809

[a] *In-text citation times of cited articles.*

TABLE 2 | Comparing citation contexts.

	ONLY_CITATION		WITH_CITATION		FULL_SET	
	Coeff.	*P*-value	Coeff.	*P*-value	Coeff.	*P*-value
Spearman's rank	0.082	$9.5*10^{-19}$	**0.305**	$2.1*10^{-249}$	0.285	$7.0*10^{-217}$
Kendall's tau	0.055	$1.2*10^{-18}$	**0.208**	$3.3*10^{-247}$	0.194	$1.1*10^{-216}$

The highest coefficients are in bold.

TABLE 3 | Temporal distribution of the change scores of cited articles from 2007 to 2016.

Period	50 < citations ⩽ 56	56 < citations ⩽ 64	64 < citations ⩽ 77	77 < citations ⩽ 107	107 < citations
2007–2008	0.130 (*SD* = 0.045, *N* = 74)	0.132 (*SD* = 0.047, *N* = 69)	0.132 (*SD* = 0.044, *N* = 68)	0.125 (*SD* = 0.045, *N* = 78)	0.123 (*SD* = 0.050, *N* = 63)
2009–2010	0.142 (*SD* = 0.058, *N* = 263)	0.136 (*SD* = 0.053, *N* = 227)	0.127 (*SD* = 0.043, *N* = 262)	0.132 (*SD* = 0.048, *N* = 280)	0.119 (*SD* = 0.049, *N* = 239)
2011–2012	0.124 (*SD* = 0.045, *N* = 956)	0.120 (*SD* = 0.043, *N* = 860)	0.119 (*SD* = 0.042, *N* = 839)	0.112 (*SD* = 0.038, *N* = 840)	0.104 (*SD* = 0.037, *N* = 867)
2013–2014	0.116 (*SD* = 0.043, *N* = 1,706)	0.113 (*SD* = 0.041, *N* = 1,611)	0.110 (*SD* = 0.040, *N* = 1,583)	0.106 (*SD* = 0.039, *N* = 1,700)	0.095 (*SD* = 0.035, *N* = 1,700)
2015–2016	0.108 (*SD* = 0.042, *N* = 2,234)	0.104 (*SD* = 0.039, *N* = 1,994)	0.102 (*SD* = 0.037, *N* = 2,168)	0.098 (*SD*=0.035, *N* = 2,223)	0.086 (*SD*=0.033, *N* = 2,379)

SD, Standard deviation; N, the Number of observed cited articles.

Table 3). These five groups divided the cited articles from 2007 to 2016 into groups with roughly even number of cited articles. Each group has about 20% cited articles over each period and the cited times of the articles within a group are in the same interval. We use the groups to observe the differences of citation changes over time and citation counts.

The change scores differ slightly between groups. The change scores of the first four groups are relatively consistent, but most of the periods have lower change scores in the fifth group. The lower change scores in the fifth group may be caused by high citations of cited articles in this group, which may indicate that highly cited articles are relatively stable in terms of their roles in science. It is reasonable to expect that the scientific community has a more rigid consensus on the scientific contribution of a more highly cited article.

The change scores show a slightly decreasing trend over time within each group, but but its underlying factors remain unclear. A possible factor is the change of PMC OAS's journal coverage, because the journal coverage has effects on the completeness of semantic information for representation learning. However, the effects of the coverage need to be validated and proved by further evidence.

We also analyzed change score value distribution of the five groups (see **Figure 4**). The group of more frequently cited articles is encoded by more intensely orange. The five groups have similar distributions where most of the cited articles' change scores lie in the range of 0.04 to 0.2 and peak in the range of 0.06 to 0.1. Additionally, the distributions are roughly normal, but the groups with more than 107 citations are less slightly peaked in a lower score than the ones with fewer citations. It is consistent with our observations in **Table 3**.

4.4. Applications

We show two simple application examples by using the change scores and temporal citation representations to identify and understand the citation changes of cited articles.

4.4.1. Identifying Cited Articles With Most Changing Citation Contexts

We listed 5 cited articles with highest average change scores over recent 5 years (2012–2016) in **Table 4**. These articles have greatly changed descriptions within the full text of articles cited these articles in recent 5 years. The high change scores may be indications of various reasons, such as high novelty or controversy. The underlying reasons need a further examination.

4.4.2. Understanding the Changing Citation Contexts

The change scores alone are not informative for us to understand the changes of citation contexts of an article. Based on the citation representations, we can retrieve a series of similar words and articles at each time to interpret the changes. We use the article with the highest average change score over recent 5 years (Olsen et al., 2006) as an example to demonstrate the interpretability of the temporal representations of publications. We listed the most similar articles and a group of most similar words of the publication at each year from 2012 to 2016 in **Table 5**. We can see the words which describe the original content of the article like "phosphosite"; we can also see the words describing the scientific development related to this publication like "kinase specific phosphorylation site prediction" and "UbPred." It is worth noting that most similar items are other articles rather than words. It is quite reasonable because publications naturally share more syntactic and semantic features than with words. The similar articles may also provide a proxy to understand the changes.

TABLE 4 | Top 5 publications with highest average change score over recent 5 years (2012–2016).

PubMed ID	Published year	In-text citations	Actual citations[a]	Avg. change score
17081983	2006	392	2,804	0.186
18171944	2003	423	1,656	0.156
19372393	2005	379	2,194	0.151
12845331	2008	594	1,684	0.143
19608861	2009	375	1,711	0.141

[a] *From Google Scholar.*

5. DISCUSSION

We compared different types of citation context for learning citation representations and offered methods for identifying and understanding the changing roles of cited articles played in scientific dynamics.

We quantified the change rate of citation contexts by citation embeddings and analyzed the distribution of changes scores of a large set of biomedical articles. Both of the average and standard deviations of change scores over article groups differ in a small range. Besides, article groups have a similar distribution of change scores. These observations indicate the stability of the metric at the group level. Meanwhile, from the normal distributions in **Figure 4**, we observed a significant individual variability that can distinguish cited individual articles greatly. The change score we proposed is not only stable but also effective for identifying outstanding individuals.

Citation is a fundamental feature of scholarly communication used by researchers to position their research and lend support for claims they made (Mansourizadeh and Ahmad, 2011). Citation has become a well-established proxy for measuring scholarly impact (Garfield, 1979). Various citation-based techniques have been developed and applied to delineating and analyzing scientific structures and dynamics (Kessler, 1963; Small, 1973). Although cited articles play a dynamic role in the development of science, the dynamic aspect of citations characterized by the full text of articles hasn't been emphasized in citation-based techniques due to the lack computational and interpretable citation representation. The methods proposed for representing citations and the metric for quantifying the changes

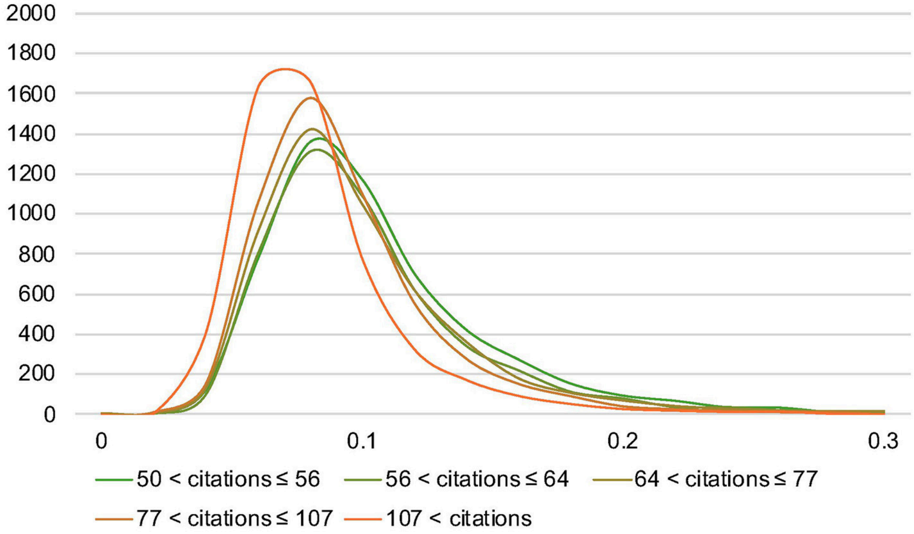


FIGURE 4 | The distribution of change score value over groups. This figure is a revised version from a workshop paper that was presented at CLBib-2017 (He and Chen, 2017).

TABLE 5 | The changes of citation contexts of Olsen et al. (2006).

Year	Change score	The most similar publications[a,b]	Most similar publications[b]
2012	0.168	20068231 (0.84)	PHOSIDA (0.78), PhosPhAt (0.77), PhosphoSite (0.76), MiCroKit (0.75), NetPhos (0.74), ChloroP1.1 (0.74), CisGenome (0.74), Phospho.ELM (0.74)
2013	0.215	21177495 (0.79)	Guittard (0.74), Scansite(0.70), phosphosite, (0.70), Tyr216 (0.69), pY (0.68), AKT (0.68), Sarbassov(0.68), IRAG (0.68), phospho-protein (0.68)
2014	0.198	21183079 (0.83)	phosphopeptide (0.73), Phosida (0.73), NetPhos (0.73),ChIP-seq (0.73), SignalP4.1 (0.72), Scansite (0.72), kinase specific phosphorylation site prediction (0.72), mNgn2 (0.71)
2015	0.145	21183079 (0.83)	phosphosite (0.75), phosphopantetheinylation (0.71), OGlcNAcylation, (0.71), Schwanhausser (0.70), phosphotyrosine-containing (0.70), phosphopeptide (0.69), phosphoamino (0.69), PTM (0.69)
2016	0.201	21081558 (0.84)	Hornbeck,(0.80), UbPred (0.79), PHOSIDA (0.79), NetPhosK (0.78), PhosphoSitePlus, (0.78), NetPhos (0.76), PhosphoSite (0.76), pY (0.75)

[a]PMID. [b]The value in the parentheses is similarity score.

have a variety of practical implications for improving the citation-based techniques at the individual and group levels.

The representation and metric can reveal important dynamics of individual articles in the evolution of science. First, the metric has potential to identify articles of importance or interests in many applications, such as academic article recommendation and information retrieval. Second, the impact dynamics of an article may be interpreted by the metrics and the representations, for example, understanding the sudden attention attracted by a *sleeping beauty* (Van Raan, 2004) in science and identifying underlying changes of an article's impact. Third, the metric may have the ability of serving as an early indication of an article's impact dynamics. Forth, the metric may provide supplementary information for scientific evaluation based on citation.

The representation and metric may also be used to enhance citation-based approaches to science mapping, such as bibliographic coupling (Kessler, 1963) and co-citation analysis (Small, 1973). Integrating the metric with citation-based approaches can reveal scientific dynamics that conveys foresights into emerging trends (Chen, 2016). For example, a cluster of articles where many of the articles cited references in new contexts may be an early sign of a emerging research topic.

6. CONCLUSION

This study has limitations and we plan to improve our methods and further investigate the factors affecting the change of citation context. We didn't use the information of how a citation was mentioned in a sentence in the representation learning. For example, a citation can play an explicit grammatical role within a sentence or play no explicit grammatical role in a sentence usually by being placed within a bracket (Thompson and Tribble, 2001). In the future, we will construct different contexts for citations with different forms. The investigation on the factors affecting the changes of citation contexts in this study is limited. We will investigate more factors and their effects. The mechanism of how the changes of citation contexts affect future impact of cited articles is another interesting question we will study in the future.

In conclusion, we introduced an embedding learning method to represent scientific articles by using the citation context text in other articles. Our method emphasizes the temporal features of citation text to characterize the dynamic role of scientific publications. The temporal representation can be used to quantify how much the role of a publication changed as well as interpret how the role changed over time. Base on the study on a large biomedical full-text literature dataset, we evaluated different citation contexts for representing citation over time and found that using sentences with in-text citation reflect topical change best. We also concluded that the metric for quantifying the changes of articles' roles is stable over time at the population level and there is significant individual variability to distinguish individuals. We hope these insights will facilitate further research into improving citation-based indicators and analysis approaches by modeling citation contexts.

AUTHOR CONTRIBUTIONS

JH designed the study, conducted the experiment, and wrote the first version of the manuscript. CC designed the evaluation method, improved the methods of the study and revised the manuscript.

FUNDING

The work is supported by the National Science Foundation (Award Number: 1633286).

ACKNOWLEDGMENTS

This is an extended and revised version of a workshop paper that was presented at CLBib-2017 (He and Chen, 2017). We thank Zhipeng Zheng at Department of Chemistry, University of Pennsylvania for his help in the example interpretation.

REFERENCES

Berger, M., McDonough, K., and Seversky, L. M. (2017). Cite2vec: citation-driven document exploration via word embeddings. *IEEE Trans. Visual. Comput. Graph.* 23, 691–700. doi: 10.1109/TVCG.2016.2598667

Boyack, K. W., Small, H., and Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *J. Assoc. Inform. Sci. Technol.* 64, 1759–1767. doi: 10.1002/asi.22896

Callahan, A., Hockema, S., and Eysenbach, G. (2010). Contextual cocitation: augmenting cocitation analysis and its applications. *J. Assoc. Inform. Sci. Technol.* 61, 1130–1143. doi: 10.1002/asi.21313

Chen, C. (2012). Predictive effects of structural variation on citation counts. *J. Am. Soc. Inform. Sci. Technol.* 63, 431–449. doi: 10.1002/asi.21694

Chen, C. (2016). Grand challenges in measuring and characterizing scholarly impact. *Front. Res. Metr. Anal.* 1:4. doi: 10.3389/frma.2016.00004

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., and Radev, D. (2008). Blind men and elephants: what do citation summaries tell us about a research article? *J. Assoc. Inform. Sci. Technol.* 59, 51–62. doi: 10.1002/asi.20707

Eto, M. (2013). Evaluations of context-based co-citation searching. *Scientometrics* 94, 651–673. doi: 10.1007/s11192-012-0756-z

Ganesh, J., Ganguly, S., Gupta, M., Varma, V., and Pudi, V. (2016). "Author2vec: Learning author representations by combining content and link information," in *Proceedings of the 25th International Conference Companion on World Wide Web, WWW '16 Companion*. (Geneva: International World Wide Web Conferences Steering Committee), 49–50.

Ganguly, S., and Pudi, V. (2017). "Paper2vec: combining graph and text information for scientific paper representation," in *European Conference on Information Retrieval* (Aberdeen: Springer), 383–395.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics* 1, 359–375. doi: 10.1007/BF02019306

Gipp, B., and Beel, J. (2009). "Citation proximity analysis (CPA): a new approach for identifying related work based on co-citation analysis," in *ISSI-09: 12th International Conference on Scientometrics and Informetrics*, (Rio de Janeiro) 571–575.

Gipp, B., Meuschke, N., and Lipinski, M. (2015). "CITREC: an evaluation framework for citation-based similarity measures based on TREC genomics and pubmed central," in *Proceedings of the iConference 2015* (Newport Beach, CA).

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). "Diachronic word embeddings reveal statistical laws of semantic change," in *Proceedings Association Computational Linguistics* (Berlin: ACL).

He, J., and Chen, C. (2017). "Understanding the changing roles of scientific publications via citation embeddings," in *Proceedings of the Second Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics (CLBib-2017) Co-located with 16th International Conference on Scientometrics and Informetrics (ISSI 2017)* (Wuhan), 42–48.

He, J., and Chen, C. (2018). Predictive effects of novelty measured by temporal embeddings on growth in science. *Front. Res. Metr. Anal.* 3:9. doi: 10.3389/frma.2018.00009

He, Q., Pei, J., Kifer, D., Mitra, P., and Giles, L. (2010). "Context-aware citation recommendation," in *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, NC: ACM), 421–430.

Huang, W., Wu, Z., Liang, C., Mitra, P., and Giles, C. L. (2015). "A neural probabilistic model for context based citation recommendation," in *AAAI 2015: Proceedings of the Twenty-ninth AAAI Conference on Artificial Intelligence*, 2404–2410.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *J. Assoc. Inform. Sci. Technol.* 14, 10–25. doi: 10.1002/asi.5090140103

Lin, D. (1998). "An information-theoretic definition of similarity," in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)* (San Francisco, CA) 296–304.

Liu, S., and Chen, C. (2012). The proximity of co-citation. *Scientometrics* 91, 495–511. doi: 10.1007/s11192-011-0575-7

Liu, S., Chen, C., Ding, K., Wang, B., Xu, K., and Lin, Y. (2014a). Literature retrieval based on citation context. *Scientometrics* 101, 1293–1307. doi: 10.1007/s11192-014-1233-7

Liu, X., Yu, Y., Guo, C., Sun, Y., and Gao, L. (2014b). "Full-text based context-rich heterogeneous network mining approach for citation recommendation," in *IEEE/ACM Joint Conference on Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on IEEE* (London), 361–370.

Mansourizadeh, K. and Ahmad, U. K. (2011). Citation practices among non-native expert and novice scientific writers. *J. Engl. Acad. Purp.* 10, 152–161. doi: 10.1016/j.jeap.2011.03.004

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13 (Curran Associates Inc.), 3111–3119.

Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., et al. (2009). "Using citations to generate surveys of scientific paradigms," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics) (Boulder, CO), 584–592.

Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., et al. (2006). Global, *in vivo*, and site-specific phosphorylation dynamics in signaling networks. *Cell* 127, 635–648. doi: 10.1016/j.cell.2006.09.026

Orosz, K., Farkas, I. J., and Pollner, P. (2016). Quantifying the changing role of past publications. *Scientometrics* 108, 829–853. doi: 10.1007/s11192-016-1971-9

Qazvinian, V., Radev, D. R., and Özgür, A. (2010). "Citation summarization through keyphrase extraction," in *Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10.* (Stroudsburg, PA: Association for Computational Linguistics), 895–903.

Řehůřek, R. and Sojka, P. (2010). "Software framework for topic modelling with large corpora. in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta: ELRA), 45–50.

Resnik, P. (1995). "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95 (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 448–453.

Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *J. Assoc. Inform. Sci. Technol.* 24, 265–269.

Small, H. (2011). Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics* 87, 373–388. doi: 10.1007/s11192-011-0349-2

Small, H., and Klavans, R. (2011). "Identifying scientific breakthroughs by combining co-citation analysis and citation context," in *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics* (Leuven), 783–793.

Sugiyama, K., and Kan, M.-Y. (2013). "Exploiting potential citation papers in scholarly paper recommendation." in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital libraries* (New York, NY: ACM), 153–162.

Teufel, S., Siddharthan, A., and Tidhar, D. (2006). "Automatic classification of citation function," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (Sydney, NSW: Association for Computational Linguistics), 103–110.

Thompson, P., and Tribble, C. (2001). Looking at citations: using Corpora in English for academic purposes. *Lang. Learn. Technol.* 5, 91–105.

Tian, H., and Zhuo, H. H. (2017). Paper2vec: citation-context based document distributed representation for scholar recommendation. *arXiv [preprint] arXiv:1703.06587*.

Van Raan, A. F. (2004). Sleeping beauties in science. *Scientometrics* 59, 467–472. doi: 10.1023/B:SCIE.0000018543.82441.f1

Zhu, S., Zeng, J., and Mamitsuka, H. (2009). Enhancing medline document clustering by incorporating mesh semantic similarity. *Bioinformatics* 25, 1944–1951. doi: 10.1093/bioinformatics/btp338

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 25 October 2018
doi: 10.3389/frma.2018.00031

Resolving Citation Links With Neural Networks

*Tadashi Nomoto**

National Institute of Japanese Literature, Tokyo, Japan

This work demonstrates how neural network models (NNs) can be exploited toward resolving citation links in the scientific literature, which involves locating passages in the source paper the author had intended when citing the paper. We look at two kinds of models: triplet and binary. The triplet network model works by ranking potential candidates, using what is generally known as the triplet loss, while the binary model tackles the issue by turning it into a binary decision problem, i.e., by labeling a candidate as true or false, depending on how likely a target it is. Experiments are conducted using three datasets developed by the CL-SciSumm project from a large repository of scientific papers in the Association for Computational Linguistics (ACL) repository. The results find that NNs are extremely susceptible to how the input is represented: they perform better on inputs expressed in binary format than on those encoded using the TFIDF metric or neural embeddings of specific kinds. Furthermore, in response to a difficulty NNs and baselines faced in predicting the exact location of a target, we introduce the idea of *approximately correct targets* (ACTs) where the goal is to find a region which likely contains a true target rather than its exact location. We show that with the ACTs, NNs consistently outperform Ranking SVM and TFIDF on the aforementioned datasets.

Keywords: neural network model, citation resolution, text similarity, ACL anthology, machine learning, natural language processing

Edited by:
Philipp Mayr,
Leibniz Institut für Sozialwissenschaften (GESIS), Germany

Reviewed by:
Haozhen Zhao,
Navigant, United States
Muthu Kumar Chandrasekaran,
National University of Singapore, Singapore

Correspondence:
Tadashi Nomoto
nomoto@acm.org

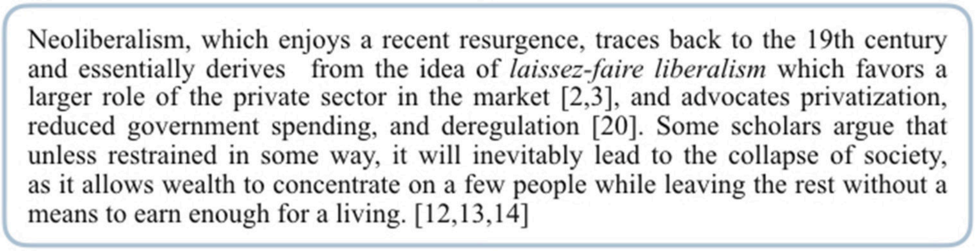
Received: *29 January 2018*
Accepted: *12 September 2018*
Published: *25 October 2018*

Citation:
Nomoto T (2018) Resolving Citation Links With Neural Networks. Front. Res. Metr. Anal. 3:31. doi: 10.3389/frma.2018.00031

1. INTRODUCTION

The work described in this paper owes its birth to recent efforts at CL-SciSumm Shared Task Project (Jaidka et al., 2016) to develop a systematic approach to relating citing snippets to their sources in the paper they refer to. The CL-SciSumm started in 2014 as a part of the NIST sponsored Text Analysis Conference to encourage the development of techniques to facilitate a computer aided understanding of the scholarly documents. The CL-SciSumm in the current format included three related tasks: (1) the citation linkage, where one is asked to find a way to locate passages in another paper which the citation refers to; (2) the facet classification, whose goal is to identify a discourse function of a referred-to passage; and (3) the summarization, which aims at creating a summary using parts that serve as a source of citations. We will explain somewhat in detail what the task (1) is about, as it will be the topic of the current work.

Consider an excerpt in **Figure 1**.



Neoliberalism, which enjoys a recent resurgence, traces back to the 19th century and essentially derives from the idea of *laissez-faire liberalism* which favors a larger role of the private sector in the market [2,3], and advocates privatization, reduced government spending, and deregulation [20]. Some scholars argue that unless restrained in some way, it will inevitably lead to the collapse of society, as it allows wealth to concentrate on a few people while leaving the rest without a means to earn enough for a living. [12,13,14]

FIGURE 1 | An example of citation in a scientific publication. Note that the examples are all fictitious.

Figure 1 has a segment that reads:

> Some scholars argue that unless restrained in some way, it will inevitably lead to the collapse of society, as it allows wealth to concentrate on a few people while leaving the rest without a means to earn enough for a living. [12,13,14]

The goal of (1) is to find out exactly which part of the referred-to papers the author had in mind when putting down the passage. As a further example, consider **Figure 2**, which has a citing instance,

> This is something we call a citing instance or *citance* (Pocus et al., 1980).

How do we get to its target, marked by the red box in **Figure 2**? This is a problem the task (1) challenges us to solve.

The results from the task (1) at CL-SciSumm give an overwhelming sense that supervised approaches, in particular, SVMs, are failing to a degree that is almost indistinguishable in performance from a method as simple as TFIDF. In light of this, we turn our attention to neural networks (NNs), which made a huge stride in recent years, to see whether they have any relevance to solving the problem. One particular (much publicized) feature of NNs is that they are an end-to-end system, meaning that they are designed to learn whatever features they need by themselves, freeing humans of the drudgery of making them up. This is something that has not been explored in the previous CL-SciSumm literature, with an exception of Nomoto (2016), who presented a preliminary attempt to leverage neural network to address resolving citation links, from which the current work descends[1]. One important difference between the earlier and present approach is that the latter takes more seriously how the input is represented, which as we show below, has a huge impact on how well models perform.

Our contribution mainly consists of presenting a novel way to tackle the citation resolution through the application of NNs and identifying some of the operational factors that influence their behavior.

In section 2, we discuss some of the past efforts to capture semantic relatedness among articles, and how they expanded to make use of information that reside outside the text such as citation counts, social relations. Section 3 introduces neural network models. We explain in detail how they actually work to spot potential targets for citations. Sections 5 and 6 will discuss how our approach compares against more conventional baselines, including Ranking-SVM and TFIDF.

2. RELATED WORK

Much of prior work on semantic relatedness among articles focused on exploiting features internal to the text itself such as term frequency, named entity, topical structure, collocation, and burstiness (Lavrenko et al., 2001; Brown, 2002; Chen and Chen, 2002; Chen et al., 2003; Nallapati, 2003; Larkey et al., 2004; Lee and Kageura, 2006; Zhang et al., 2008). Despite a large effort put into research through a project like TDT (Topic Detection and Tracking) (Allen, 2002), a general consensus that emerged out of the experience was that cosine similarity based on TFIDF, simple as it may seem, is the best option, which as it turned out, rivaled or even beat technically more informed approaches. Lavrenko et al. (2001) and Larkey et al. (2004) stand out as an interesting exception with their emphasis on the use of relevance feedback in link detection.

There is another growing trend in the literature, in which people are more concerned about how articles are connected to one another, and try to explain similarity among them through the hyperlink structure (Milne and Witten, 2008; West et al., 2009). Milne and Witten (2008) propose to make use of what they call *context terms*, or terms in a Wikipedia page likely to serve as an outgoing link, as a part of mechanism to disambiguate word senses. West et al. (2009), meanwhile, seek to enrich the hyperlink structure of Wikipedia by automatically adding links that are useful but left out by humans. A basic idea is to encode a given Wikipedia page in terms of connections it has to the rest of Wikipedia and use the principal component analysis to predict links that are missing from the original structure. Compared to Milne and Witten (2008), which mostly relies on the number of shared links to determine the relatedness of terms, an approach by West et al. (2009) achieves a level of sophistication far beyond that of Milne and Witten (2008).

Bethard and Jurafsky (2010) aim at identifying potential papers that an author may cite in his or her work. Besides textual similarity between citing and cited papers, they look at features such as whether authors are citing papers they have

[1] Some of the NNs we develop here are an adaptation of the embedding models proposed by Weston et al. (2010, 2013) and Bordes et al. (2013, 2014).

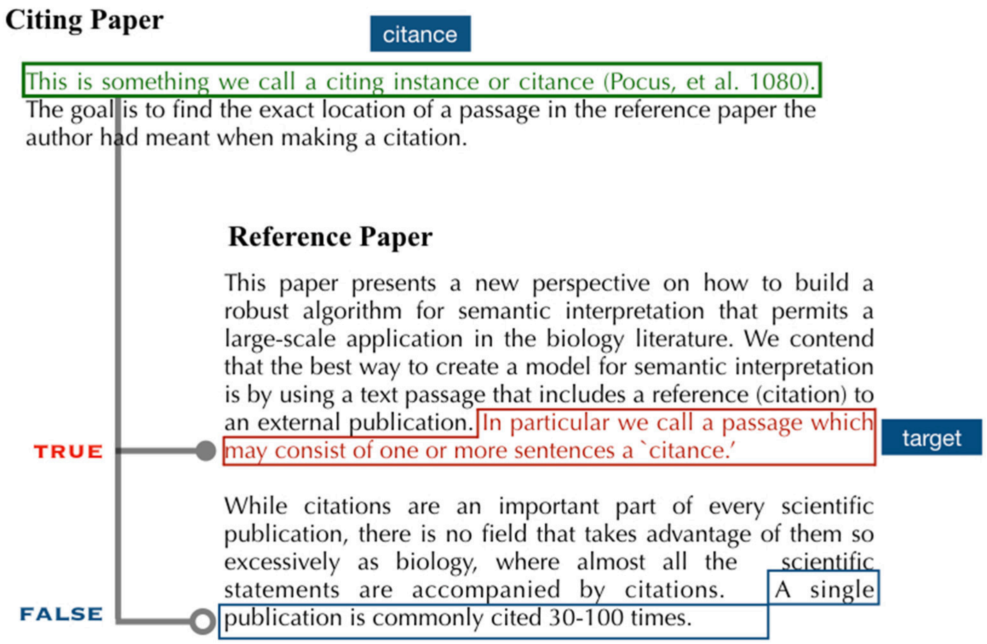

Citing Paper
citance
This is something we call a citing instance or citance (Pocus, et al. 1080).
The goal is to find the exact location of a passage in the reference paper the author had meant when making a citation.
Reference Paper
This paper presents a new perspective on how to build a robust algorithm for semantic interpretation that permits a large-scale application in the biology literature. We contend that the best way to create a model for semantic interpretation is by using a text passage that includes a reference (citation) to an external publication. In particular we call a passage which may consist of one or more sentences a `citance.'
target
TRUE
While citations are an important part of every scientific publication, there is no field that takes advantage of them so excessively as biology, where almost all the scientific statements are accompanied by citations. A single publication is commonly cited 30-100 times.
FALSE

FIGURE 2 | A citance and its target. A sentence boxed in red represents a true target for the citance (one in green box) and one underlined in blue a false target. Note that the examples are all fictitious.

cited in the past, whether they are citing works done by their past co-authors, and how many times papers are cited by other authors. The significance of Bethard and Jurafsky (2010) lies in their finding that much of identifying potential papers is actually driven by factors extraneous to the content of a paper, such as recency, authorship and citation counts. The finding is also consonant with an observation by Meij and de Rijke (2007) that contextual information such as the number of citations has a visible impact on the effectiveness of document retrieval in the scientific literature.

Among the systems that participated in the 2016 CL-SciSumm conference, those from Cao et al. (2016), Li et al. (2016), and Moraes et al. (2016) are most notable. Cao et al. (2016) split the text into n-sentence long segments and used the SVM Ranking to find a stretch of text likely to be a source of the citation. Moraes et al. (2016) found that an approach using TFIDF together with some preprocessing options (stemming, cutting off sentences that exceed a certain limit) outperformed that based on a tree-kernel. Meanwhile, Li et al. (2016) turned to a rule based model, where they combined diverse similarity metrics (Jaccard, word2vec, idf-based similarity, etc.), each weighted with some hand-picked coefficient, to arrive at a prediction. The approach is manually demanding because how much contribution each feature makes to the final outcome has to be decided by humans, and its ability to generalize is unknown because it was tested only on one particular dataset provided by the CL-SciSumm in 2016.

Despite differences in ways people tackled the problem, a curious commonality emerged from the studies: that a simple similarity metric such as TFIDF or Jaccard works better than those that rely on supervision (Li et al., 2016 even found that Word2Vec fell behind Jaccard). Later in the paper, we will examine whether what they found holds true for the current setup, while looking at how NNs fare against Jaccard and TFIDF.

3. RESOLVING CITATION LINKS WITH NEURAL NETWORKS

In this work, we explore two approaches to modeling citation resolution, both based on neural networks: one is what we might call a "triplet model" which aims to rank sentences in terms of how similar they are to the source sentence (citance); and the other is a binary classification model which labels a given sentence as "true" or "false," depending on how likely a target it is.

3.1. Triplet Model

We start with the triplet model. Its objective is to provide a scoring function h that favors a true target[2] over a false one, or more precisely, to build a function that ensures that $h(s, t^{+}) > h(s, t^{-})$, where s denotes a citing snippet, t^{+} denotes a true target (a sentence humans judged as a target) and t^{-} a false target (i.e., a sentence not selected as target). Here and throughout, we assume that both s and t consist of exactly one sentence. If we take **Figure 2** as an example, the green box corresponds to s, the red to t^{+} and the blue to t^{-}

[2] By *target*, we mean one or more sentence in the referred-to paper (RP) that serve as a source for a snippet or text citing the RP: for example, a sentence boxed in red in **Figure 2**; and those not boxed are said to be *false targets* with respect to the citance.

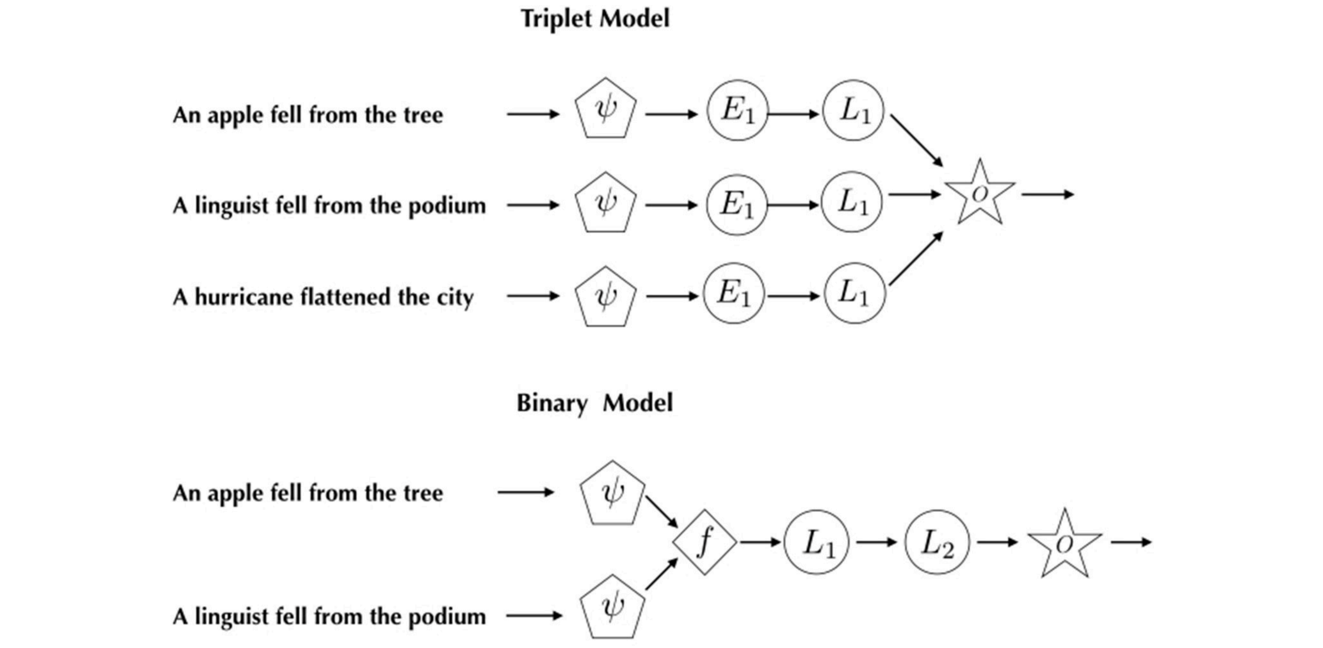

Triplet Model
An apple fell from the tree
A linguist fell from the podium
A hurricane flattened the city
ψ
E_1
L_1
O
Binary Model
An apple fell from the tree
A linguist fell from the podium
ψ
f
L_1
L_2
O

FIGURE 3 | Citation resolution models. ψ is a function to transform the input, mapping it into a matrix of real numbers. f denotes an arithmetic operation on outputs of ψ, $L_{1,2}$ hidden layers. A star marked with an "O" represents an output layer. The number of hidden units in L_1 in the binary model is 100, and that in L_2 is 2. L_1 in the triplet model consists of 60 hidden units. E_1 is an embedding layer containing 10 units.

We define h by:

$$h(s,t) = \mathbf{V}(s)^{\top}\mathbf{V}(t), \tag{1}$$

where $\mathbf{V}(s)$ denotes a vector derived from s through the application of some neural network and similariy for $\mathbf{V}(t)$. One way to ensure that s's similarity with its true target (t^+) ranks higher than that with a false target (t^-) is to require the following constraint to hold for h (Weston et al., 2010; Bordes et al., 2014):

$$\forall_{i,j}\ h(s,t^+) > h(s,t^-). \tag{2}$$

Noting that we need to ensure that $h(s,t^-) - h(s,t^+) < -C$ for some constant C ($\neq 0$), the above formula turns into a loss function:

$$L_1 = \max(0, C - h(s,t^+) + h(s,t^-)). \tag{3}$$

One way to think about t^+ and t^- is to take the former as a sentence labeled by humans as a true target and the latter as one of those sentences that are similar to t^+ (we call it *target-centric supervision* as opposed to *citance-centric supervision*, which we later explain).

Figure 3 gives a general picture of how we move through a neural architecture to $C - h(s,t^+) + h(s,t^-)$. $\psi(\cdot)$ denotes a representation function that maps a sentence into a discrete or continuous multi-dimensional space. While there are a number of ways to define ψ, we focus on the following three. Note that N is the size of the vocabulary.

$$\psi_e(s) = \begin{pmatrix} v_{11} & v_{12} & v_{13} & \dots & v_{1N} \\ v_{21} & v_{22} & v_{23} & \dots & v_{2N} \\ \dots & \dots & \dots & \dots & \dots \\ v_{I1} & v_{I2} & v_{I3} & \dots & v_{IN} \end{pmatrix} \tag{4}$$

$$\psi_b(s) = \{0,1\}^N \tag{5}$$

$$\psi_t(s) = \begin{pmatrix} w_1 & w_2 & w_3 & \dots & w_N \end{pmatrix} \tag{6}$$

(4) represents what is generally known as "word embedding," where each word in a sentence is assigned to a vector of randomly generated real numbers, whose length I is also arbitrarily chosen (we set I to 10 in the experiments later described). (5) produces a representation that consists of binary values, with 0 indicating the absence and 1 the presence of a particular word in the sentence. (6) works like (5), except that it associates each word with its tfidf value.

We project $\psi(s)$ and $\psi(t)$ into a hidden layer l via a matrix $\mathbf{W}$ ($\in \mathbb{R}^{N\times K}$)[3]. K represents the number of neural units in l.

[3] Note that the shape of $\mathbf{W}$ will become I times N by K, when working with word embedding. As a further note, $\psi_b(s)$ is of shape $S \times N$, where S is the length of sentence s (the number of words) and N the size of the vocabulary, with each word represented as a "one-hot" vector, meaning it consists of a single vector of size N, with all cells set to 0, except for one that corresponds to the relevant word, which is set to 1. $\psi_t(s)$ works the same way, except that each word vector has a tfidf value where $\psi_b(s)$ has 1.

Intuitively, one could think of an element of $\mathbf{W}$ as indicating the strength of relationship between a word and a corresponding hidden unit. How to determine it is a primary concern of the neural model.

Now we define a layer $\mathbf{G_1}$ by

$$\mathbf{G_1}(s) = g(\psi(s)\,\mathbf{W_1} + b_1) \tag{7}$$

b_1 is a parameter for the bias and g an activation function which we take to be a rectifier: i.e., $g(x) = \max(0, x)$. For each $(s_i, t_i^+, t_i^-) \in D$, we run the stochastic gradient descent (SGD) to minimize:

$$\max(0, C - \mathbf{G_1}(s)\mathbf{G_1}(t^+)^\top + \mathbf{G_1}(s)\mathbf{G_1}(t^-)^\top).$$

It is important to note that minimizing has the effect of increasing the chance that the similarity of the citance to its true target is larger than that to a false target. For SGD, we use an optimizer known as ADAM, which makes use of bias corrected moments to adaptively change learning rates (Kingma and Ba, 2015; Ruder, 2016). We set C to 1.0 in the experiments below, following Weston et al. (2013).

3.2. Binary Classification Model

The binary model takes as input a vector of features of the from $(f(s_{i1}, t_{i1}), \ldots, f(s_{iN}, t_{iN}))$[4], which we feed into the layer $\mathbf{G_1}$, whose output is further fed to the following:

$$\mathbf{G_2}(u) = m(\mathbf{G_1}(s)\mathbf{W_2} + b_2). \tag{8}$$

The loss function is given by:

$$L_2 = -\mathbf{y}^* \log(\mathbf{G_2}(u)). \tag{9}$$

where $\mathbf{W_2}$ is of the shape $K \times 2$, $\mathbf{y}$ is a true label for a given sentence, m is a softmax function. $\mathbf{y}$ is assigned to $(1, 0)$ if t is a true target of s and $(0, 1)$, otherwise. We define $x \cdot y$ as an inner product of x and y. We assume f to be either an element-wise multiplication or a squared distance.

4. DATA SETS

We created training data from three sources: (1) the "Development-Set-Apr8" dataset (henceforth, DSA2016) (Jaidka et al., 2016); (2) a pilot study corpus which was created as a part of the Text Analysis Conference (TAC2014), prior to DSA2016, and (3) the data made available for the shared task conference at BIRNDL2016 (hereafter, SRD2016). Regardless of where it originates, each dataset contains a number of folders representing a *topic*, which is composed of one reference paper (RP) and a number of papers that make reference to it (or CPs) [5].

TABLE 1 | Corpus profiles.

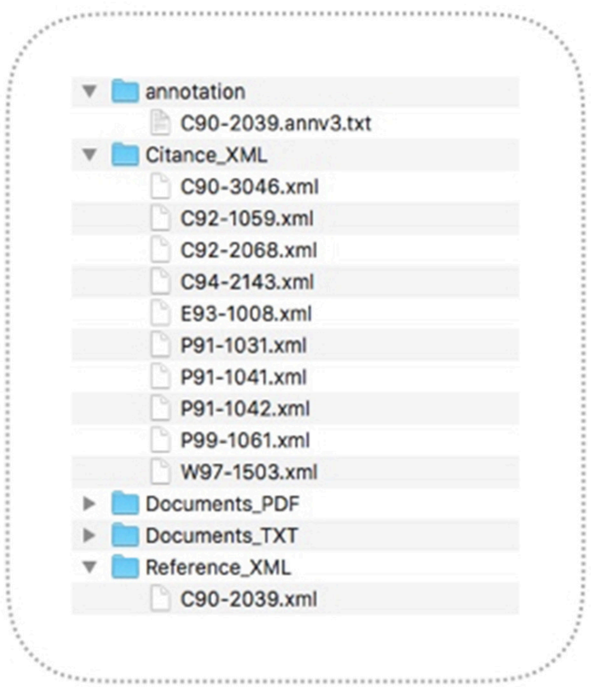
RP	|RP|	#CPs	|*T*|	#Citances
TAC2014				
C90-2039	211	10	33	16
C94-2154	118	5	12	5
E03-1020	99	9	19	15
H05-1115	190	8	19	12
H89-2014	152	8	19	11
J00-3003	586	9	24	10
J98-2005	105	9	26	21
N01-1011	195	8	16	8
P98-1081	164	9	60	25
X96-1048	363	9	21	12
DSA2016				
C02-1025	205	18	31	23
C08-1098	226	22	37	29
C10-1045	321	13	42	33
D10-1083	248	11	21	18
E09-2008	63	10	8	8
N04-1038	258	20	44	24
P06-2124	247	12	38	18
W04-0213	161	13	28	18
W08-2222	165	9	13	9
W95-0104	338	25	68	39
SRD2016				
C00-2123	204	16	24	20
C04-1089	177	16	19	17
I05-5011	213	19	33	23
J96-3004	473	47	109	69
N06-2049	156	16	35	22
P05-1004	235	12	14	14
P05-1053	219	34	90	71
P98-1046	177	26	34	31
P98-2143	157	43	93	59
W03-0410	275	10	29	24

Table 1 give some statistical profiles of TAC2014, DSA2016, and SRD2016. $|T|$ is the number of sentences in RP which CPs' citations are pointing to (*target sentences*). |RP| represents the number of sentences that comprise the RP, #CPs the number of citing papers, and #Citances the number of citing instances (or *citances*) in CPs that make reference to RP. $|T|$ tends to be greater than #Citances, as most of target sentences appear in more than one citance.

In what follows, we mean by $\mathbb{CP}$ a set of sentences that comprise a citing paper and by $\mathbb{RP}$ those that comprise its reference paper. We build a training set by creating a set D of

[4] s_{ij} denotes the j-th word in the vocabulary that appears in s_i. The same applies to t_{ij}.

[5] **Figure 4** shows one such cluster: what appears under Citance_XML is a group of papers that contain passages (*citances*) that refer to paper C90-2039, which is placed in a directory called Reference_XML. The former corresponds to CPs and the latter to RP. Associated with each topic cluster is a file that contains human created annotations (e.g., C90-2039-annv3.txt in **Figure 4**) that indicate which part of the RP a given citing passage relates to, an example of which is found in **Figure 5**: the area shaded in green contains information on a citing passage and the one in yellow indicates a target sentence.



```
▼ annotation
    C90-2039.annv3.txt
▼ Citance_XML
    C90-3046.xml
    C92-1059.xml
    C92-2068.xml
    C94-2143.xml
    E93-1008.xml
    P91-1031.xml
    P91-1041.xml
    P91-1042.xml
    P99-1061.xml
    W97-1503.xml
▶ Documents_PDF
▶ Documents_TXT
▼ Reference_XML
    C90-2039.xml
```

FIGURE 4 | Directory plot of topic cluster C90-2039.

triplets such that: for a given $s \in \mathbb{CP}$ and $t \in \mathbb{RP}$,

$$D = \{(s, t, u)\}, \exists u \in R, u \neq t, \tag{10}$$

where $R \subset \mathbb{RP}$. Thus, if we have a reference paper with four sentences $\{a, b, c, d\}$ a citing instance s from $\mathbb{CP}$, and a target sentence b, we will have $\{(s, b, a), (s, b, c), (s, b, d)\}$ as the training data. The size of the training data for topic cluster C roughly sums to:

$$\sum_{v \in \mathbb{I}(C)} r(v)|R \cap \overline{T}|, \tag{11}$$

where $r(v)$ the number of target sentences associated with a citing instance v, and $|R|$ the size of R. $\mathbb{I}(C)$ stands for a set of citing instances in C, $\overline{T}$ a complement of T (a set of target sentences). We set $|R|$ to 10 in the experiments described below.

5. EVALUATION

To evaluate, we followed a cross validation style setup where we set aside one cluster for testing, using the rest for training, and report an average performance we get from the validation test on each of the clusters contained in a dataset.

TAC2014, DSA2016, and SRD2016 each came with ten topic clusters, consisting of one reference paper and a number of papers which cite that paper. We gauged performance by a metric known as MRR (mean reciprocal rank), which produces the average of the inverted ranks of first true targets retrieved by the models. The closer an MRR is to 1, the better the performance is. Formally, MRR is defined as

$$R(C, M) = \frac{1}{|C|} \sum_{t[i] \,:\, i \in C} \frac{1}{\mathrm{rank}_M(t[i])} \tag{12}$$

C is a set of citances (see **Figure 5** for an example) and M a model[6]. In case a citance or a target involves multiple sentences, we split them into pairs of sentences in such a way that each pair will consist of exactly one sentence from s and one from t. This makes it easier for NNs to handle inputs, as they require that the length of input to stay fixed. $\mathrm{rank}_M(t[i])$ represents the rank of the *first* true target returned by M for the i-th citance. Note that MRR is meant to measure performance not in terms of how similar the output is to target sentences as was done in the past SciSumm events, but in terms of how accurately we locate a true target. We believe this is more in line with the goal of the CL-SciSumm[7], [8].

To find how neural models compare to some of the more conventional methods, we also included Ranking SVM (SVR, henceforth) (Joachims, 2006)[9] in a roster of models we put to the test. As mentioned earlier, given the relative paucity of positive instances available, it may be hard to get a meaningful insight by running a discriminatory version of SVM, as it requires a sizable amount of training data for each of the labels we are interested in. Since we have on the average, positive instances accounting for only about 10% of the corpus amenable to use in training, the discrete classification with SVM is a non starter. This is the reason we turn to SVR.

Preliminary tests we ran on NNs found that NNs trained on triplets created with the target-centric supervision (section 3.1) were not able to produce performance on a par with baselines. Which promoted us to come up with an alternative setup, where we regard sentences in $\mathbb{RP}$ that are most similar to those in $\mathbb{CP}$ as targets, entirely dismissing the annotations supplied by humans (which we call *citance-centric supervision* or CCS)[10]. As it turned out, the adoption of CCS brought a clear gain to NNs, propelling them over baselines by a comfortable margin.

Under the new setup, we have a triplet of the form:

$$D_1 = \{(s, t, u)\}_n, \tag{13}$$

where t is a sentence in $\mathbb{RP}$ that ranks between the 1st and 10th in terms of the similarity to s ($\in \mathbb{CP}$) and u ($\in \mathbb{RP}$) is a sentence

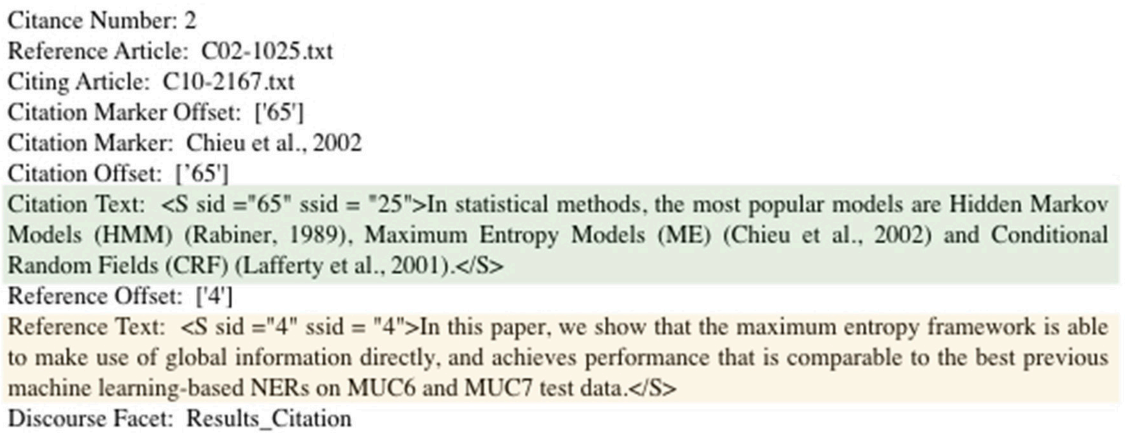
[6] It is safe to assume that a ground truth citance consists of two types of information: the site of a citing instance in $\mathbb{CP}$ and that of its target in $\mathbb{RP}$.

[7] We were told that they abandoned an accuracy based metric because it was not able to distinguish participating systems, with their performance crawling around 0.

[8] While a particular way we set up the evaluation makes it infeasible to make a direct comparison to the systems at the CL-SciSumm, we included in the evaluation, our analogs of baselines that were noted at the conference for their strong performance over other competing methods, in particular, TFIDF, Jaccard, and SVM-Ranking, which would provide some sense of how the current setup compares to the previous approaches.

[9] We used the sklearn library to implement the SVR (http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html).

[10] This is equivalent to what is generally known as *distant supervision* in the machine learning literature, where training labels are created artificially with heuristics or rules (cf. Mintz et al., 2009).



Citance Number: 2
Reference Article: C02-1025.txt
Citing Article: C10-2167.txt
Citation Marker Offset: ['65']
Citation Marker: Chieu et al., 2002
Citation Offset: ['65']
Citation Text: <S sid ="65" ssid = "25">In statistical methods, the most popular models are Hidden Markov Models (HMM) (Rabiner, 1989), Maximum Entropy Models (ME) (Chieu et al., 2002) and Conditional Random Fields (CRF) (Lafferty et al., 2001).</S>
Reference Offset: ['4']
Reference Text: <S sid ="4" ssid = "4">In this paper, we show that the maximum entropy framework is able to make use of global information directly, and achieves performance that is comparable to the best previous machine learning-based NERs on MUC6 and MUC7 test data.</S>
Discourse Facet: Results_Citation

FIGURE 5 | Citation and target. A 65th sentence in C10-2167 is seen as referring to a fourth sentence in C02-1025.

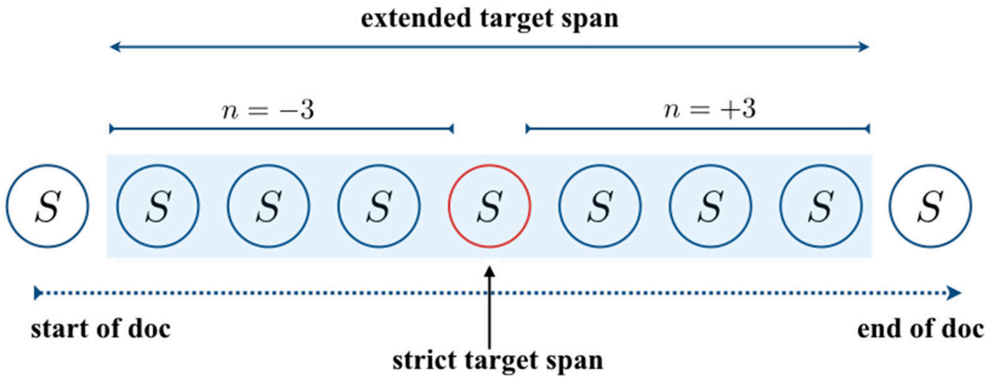


FIGURE 6 | Strict vs. extended target span.

that ranks between the 11th and 20th (the similarity was given in the dot-product). n is the number of training instances. We also applied the idea to the binary classification model: we treated top 10 most similar sentences as true targets and those that did not make it to the top 10 but ranked above 21 as corrupt or false. The training and test data for the binary classifier look like:

$$D_2 = \{(f(\psi(s), \psi(t)), y)\}_n. \tag{14}$$

ψ is either ψ_b or ψ_t. f is an element-wise multiplication or a squared distance. $y \in \{1, 0\}$. $s \in \mathbb{CP}$ and $t \in \mathbb{RP}$. SVR was trained on D_2 with f set to the element-wise multiplication.

In addition to SVR, we consider two other baselines, TFIDF and BDOT Both make use of the inner product to determine whether t is a target for s. They differ only in what they take as input: TFIDF operates on vectors of values expressed in TFIDF, while BDOT on those that consist of binary values. Formally, they will come to:

$$\textsc{tfidf}(s, t) = \psi_t(s)\psi_t(t)^\top \tag{15}$$

$$\textsc{bdot}(s, t) = \psi_b(s)\psi_b(t)^\top \tag{16}$$

Notice that BDOT is in practice equivalent to a familiar Jaccard coefficient:

$$\psi_b(s)\psi_b(t)^\top \approx \frac{|s \cap t|}{|s \cup t|}, \tag{17}$$

which the previous literature found to be singularly effective for identifying the target passage (Li et al., 2016). In BDOT, we can safely ignore the denominator as it remains invariant: $|\psi_b(s) \cup \psi_b(t)| = N$ (the size of the vocabulary, i.e., the number of unique tokens that comprise the dataset).

Finding MRR involves ranking for a given $s \in \mathbb{CP}$, a sentence t in $\mathbb{RP}$ in accordance to how similar s is to t. For SVR, TFIDF, and BDOT, this is straightforward. For NNs, however, things get somewhat tricky, as their outputs represent the loss, not the similarity. So we use $\mathbf{G}_1$ instead [see (7)] as an indicator of the strength of the relationship between s and t: in other words, we quantify the similarity by $\mathbf{G}_1(s)\mathbf{G}_1(t)^\top$.

TABLE 2 | Results in MRR for CCS.

Model	±0	±1	±2	±3	±4	±5
TAC2014						
NN: in = binary,loss = entropy	**0.0940**	**0.1742**	0.2151	0.2404	**0.2993**	0.3151
NN: in = binary, loss = triplet	0.0900	0.1601	**0.2600**	**0.2476**	0.2620	**0.3198**
NN: in = tfidf, loss = triplet	0.0710	0.1546	0.2212	0.1998	0.2305	0.2101
NN: in = embedding, loss = triplet	0.0720	0.1308	0.1793	0.2175	0.2465	0.2621
NN: in = word2vec_1, loss = triplet	0.0028	0.0215	0.0340	0.0468	0.0715	0.0785
NN: in = word2vec_2, loss = triplet	0.0054	0.0249	0.0430	0.0472	0.0575	0.0776
SVR	0.0694	0.1351	0.1652	0.2182	0.2618	0.2983
TFIDF	0.0711	0.1297	0.1677	0.2060	0.2387	0.2552
BDOT	0.0777	0.1440	0.1865	0.2291	0.2556	0.2852
DSA2016						
NN: in = binary, loss = entropy	0.0864	**0.1724**	0.2166	**0.2533**	**0.3020**	0.3340
NN: in = binary, loss = triplet	**0.0918**	0.1400	**0.2198**	0.1871	0.2026	**0.3439**
NN: in = tfidf, loss = triplet	0.0643	0.1627	0.2107	0.2363	0.2721	0.3016
NN: in = embedding, loss = triplet	0.0799	0.1492	0.2013	0.2414	0.2696	0.3082
NN: in = word2vec_1, loss = triplet	0.0025	0.0255	0.0348	0.0401	0.0486	0.0613
NN: in = word2vec_2, loss = triplet	0.0104	0.0196	0.0290	0.0484	0.0591	0.0678
SVR	0.0687	0.1500	0.1847	0.2290	0.2725	0.3060
TFIDF	0.0828	0.1383	0.1804	0.2068	0.2425	0.2758
BDOT	0.0909	0.1557	0.2003	0.2317	0.2685	0.3009
SRD2016						
NN: in = binary, loss = entropy	0.0771	0.1373	**0.1781**	**0.2137**	0.2287	0.2476
NN: in = binary, loss = triplet	0.0942	**0.1435**	0.1642	0.1641	0.1922	0.2535
NN: in = tfidf, loss = triplet	0.0746	0.1144	0.1438	0.1724	0.1882	0.2039
NN: in = embedding, loss = triplet	0.0808	0.1134	0.1379	0.1533	0.1723	0.1853
NN: in = word2vec_1, loss = triplet	0.0043	0.0174	0.0342	0.0545	0.0643	0.0728
NN: in = word2vec_2, loss = triplet	0.0067	0.0083	0.0138	0.0801	0.0806	0.0951
SVR	0.0747	0.1183	0.1534	0.1729	0.1932	0.2238
TFIDF	0.0688	0.1081	0.1428	0.1828	0.2061	0.2315
BDOT	**0.0819**	0.1196	0.1604	0.1969	**0.2319**	**0.2560**

The highest mark achieved for each span radius is highlighted in bold.

Moreover, to get a broad picture of how the models perform, we introduce an idea we call *approximately correct targets* (ACTs), where we are not only interested in finding out whether they pick up exact sentences humans labeled as true targets, but also finding out how close predictions are to the true targets.

Consider **Figure 6**. A target sentence of interest is one circled in red. We mean by *approximately correct targets*, those sentences that appear n sentences away (both forward and backward) from the target, where n is arbitrarily chosen (which is set at 3 in the example) and take any sentence that occurs within the region to be as correct as the true target. In **Figure 6**, ACTs are found in the area shaded in light blue. A motivation for this idea comes from our curiosity to find out whether it is possible to achieve meaningful performance by making the citation resolution less hard (the preliminary experiments suggest it will not happen if we stick to the strict target span).

6. RESULTS AND DISCUSSION

Tables 2, **3** show the outcome of running NNs in four different setups, along with baselines[11]. "NN: in = binary,loss = entropy" refers to a binary classification model using ψ_b for the input transformation, and the L_2 loss function. "NN: in = binary, loss = triplet" denotes a triplet model with ψ_b for the input and the loss measured in L_1. "NN: in = tfidf, loss = triplet" and "NN: in = embedding, loss = triplet" are like the previous model except that the former uses ψ_t and the latter, ψ_e in place of ψ_b. The top row indicates the length of the target span. For instance, $n = \pm 3$ means that any of the three sentences that either precede or follow the target sentence is considered ACTs; and $n = \pm 0$ means that there is no ACT other than the target sentence itself.

[11] We performed stemming and removed stop words, using the NLTK package (Bird et al., 2009). All the NNs described here were created using the tensorflow package (https://www.tensorflow.org).

TABLE 3 | Results in MRR for TCS.

Model	±0	±1	±2	±3	±4	±5
TAC2014						
NN: in = binary, loss = entropy	0.0557	0.1131	0.1484	0.1700	0.1946	0.2240
NN: in = binary, loss = triplet	0.0552	**0.1485**	**0.2020**	**0.2394**	**0.2760**	**0.3024**
NN: in = tfidf, loss = triplet	0.0537	0.1382	0.1804	0.2008	0.2301	0.2412
NN: in = embedding, loss = triplet	**0.0784**	0.1324	0.1551	0.1922	0.2178	0.2335
NN: in = word2vec_1, loss = triplet	0.0041	0.0204	0.0389	0.0555	0.0842	0.0906
NN: in = word2vec_2, loss = triplet	0.0109	0.0282	0.0446	0.0502	0.0587	0.0775
SVR	0.0542	0.1121	0.1483	0.1890	0.2293	0.2559
TFIDF	0.0711	0.1297	0.1677	0.2060	0.2387	0.2552
BDOT	0.0777	0.1440	0.1865	0.2291	0.2556	0.2852
DSA2016						
NN: in = binary, loss = entropy	0.0870	**0.1700**	**0.2124**	**0.2542**	**0.2898**	**0.3204**
NN: in = binary, loss = triplet	0.0699	0.1273	0.1951	0.2126	0.2571	0.2892
NN: in = tfidf, loss = triplet	0.0781	0.1450	0.1813	0.2054	0.2317	0.2713
NN: in = embedding, loss = triplet	0.0781	0.1306	0.1788	0.2276	0.2399	0.2585
NN: in = word2vec_1, loss = triplet	0.0171	0.0457	0.0590	0.0650	0.0724	0.0792
NN: in = word2vec_2, loss = triplet	0.0126	0.0220	0.0325	0.0515	0.0634	0.0714
SVR	0.0872	0.1512	0.1866	0.2119	0.2559	0.2811
TFIDF	0.0828	0.1383	0.1804	0.2068	0.2425	0.2758
BDOT	**0.0909**	0.1557	0.2003	0.2317	0.2685	0.3009
SRD2016						
NN: in = binary, loss = entropy	0.0728	**0.1352**	**0.1706**	**0.2014**	0.2189	0.2322
NN: in = binary, loss = triplet	0.0714	0.1291	0.1685	0.1993	0.2168	0.2362
NN: in = tfidf, loss = triplet	0.0682	0.1161	0.1488	0.1668	0.1995	0.2100
NN: in = embedding, loss = triplet	0.0794	0.1137	0.1303	0.1437	0.1600	0.1733
NN: in = word2vec_1, loss = triplet	0.0054	0.0142	0.0316	0.0536	0.0623	0.0689
NN: in = word2vec_2, loss = triplet	0.0139	0.0188	0.0254	0.0858	0.0926	0.1054
SVR	0.0593	0.1092	0.1382	0.1696	0.1883	0.2200
TFIDF	0.0688	0.1081	0.1428	0.1828	0.2061	0.2315
BDOT	**0.0819**	0.1196	0.1604	0.1969	**0.2319**	**0.2560**

The highest mark achieved for each span radius is highlighted in bold.

TABLE 4 | Effects of multiplication vs. squared distance on performance.

Model	±0	±1	±2	±3	±4	±5
TAC2014						
NN: in = binary/mul, loss = entropy	0.0940	0.1742	0.2151	0.2404	0.2993	0.3151
NN: in = tfidf/mul, loss = entropy	0.0741	0.1425	0.1866	0.2334	0.2644	0.2817
NN: in = binary/sqrd, loss = entropy	0.0622	0.1764	0.2225	0.2443	0.2784	0.2457
NN: in = tfidf/sqrd, loss = entropy	0.0471	0.0854	0.1403	0.1739	0.2080	0.2558
DSA2016						
NN: in = binary/mul, loss = entropy	0.0864	0.1724	0.2166	0.2533	0.3020	0.3340
NN: in = tfidf/mul, loss = entropy	0.0790	0.1581	0.2048	0.2416	0.3045	0.3412
NN: in = binary/sqrd, loss = entropy	0.0462	0.1032	0.1621	0.1774	0.2015	0.2293
NN: in = tfidf/sqrd, loss = entropy	0.0392	0.0929	0.1374	0.1675	0.1946	0.2132
SRD2016						
NN: in = binary/mul, loss = entropy	0.0771	0.1373	0.1781	0.2137	0.2287	0.2476
NN: in = tfidf/mul, loss = entropy	0.0998	0.1470	0.1819	0.2174	0.2401	0.2596
NN: in = binary/sqrd, loss = entropy	0.0338	0.0671	0.0835	0.0978	0.1215	0.1414
NN: in = tfidf/sqrd, loss = entropy	0.0505	0.0744	0.0991	0.1249	0.1499	0.1677

The numbers in the tables show MRRs averaged over 10 topic clusters (An MRR for each cluster was produced via a by-topic cross validation, where we set aside one topic cluster for testing and use the rest for training).

6.1. CCS vs. TCS

The results in CCS or *citance-centric supervision* (**Table 2**) show a clear tendency for the NNs to score higher than the baselines, which include SVR, TFIDF, and BDOT, across varying lengths of the span. Of a particular note is the performance of models which take the input in binary format, against those that employ TFIDF or the embedding[12]. We see the former consistently outperforming the latter. It is safe to say that representing the input in binary format led to the superior performance, which parallels BDOT outperforming TFIDF across the datasets.

By contrast, in TCS (*target-centric supervision*), NNs suffer an across-the-board decline in performance, with some of the top performers under CCS dipping below baselines. It is interesting that they seem to suffer more in TAC than in DSA and SRD, which could be attributable to the fact that TAC contains *less* citing papers than the other two (cf. **Table 1**). Yet the triplet model still performs better on inputs represented in binary format than on those in tfidf, which echoes what we found in CCS.

The finding that CCS yields a better performance happen than TCS regardless of models we choose, is signficant. Because what it implies is that hand created annotations are no better than those created automatically by using a simple similarity metric (or L1-norm in our case), in terms of quality they permit as training data. To further investigate the matter, we examined whether the performance we see under CCS is significantly different from that we have under TCS. The results are shown in **Table 5**. We see “NN: in = binary, loss = entropy” having 0.0006* for *p*-value, meaning that the performance it had in CCS was significantly different from results it produced in TCS. So there appear to be some grounds for arguing that whether one works with CCS or TCS does have consequences for neural models, though how much the choice affects them varies from model to model. “NN: in = binary, loss = entropy” tends to be more sensitive to the choice. The same appears to be the case with models involving an embedding of one sort or another. In constrast, “NN: in = binary, loss = triplet” is totally blind to whatever differences there might be between the two modes of annotation. Yet the fact that CCS in general leads to better results, some of which we found to be statistically different from those under TCS, does cast some doubt over the rationale of using human created annotations as gold standards. Whether this is due to a particular way the datasets were created or to difficulties inherent to annotating papers for citation links, remains to be seen.

TABLE 5 | Significance Test of CCS against TCS (with paired *t*-test).

Model	***p*-value**
TAC2014	
NN: in = binary, loss = entropy	0.0006*
NN: in = binary, loss = triplet	0.1125
NN: in = tfidf, loss = triplet	0.5003
NN: in = embedding, loss = triplet	0.0533
NN: in = word2vec_1, loss = triplet	0.0394**
NN: in = word2vec_2, loss = triplet	0.0292**
DSA2016	
NN: in = binary, loss = entropy	0.1021
NN: in = binary, loss = triplet	0.7375
NN: in = tfidf, loss = triplet	0.0350**
NN: in = embedding, loss = triplet	0.0186**
NN: in = word2vec_1, loss = triplet	0.0001*
NN: in = word2vec_2, loss = triplet	0.0002*
SRD2016	
NN: in = binary, loss = entropy	0.0082*
NN: in = binary, loss = triplet	0.8765
NN: in = tfidf, loss = triplet	0.5083
NN: in = embedding, loss = triplet	0.0232**
NN: in = word2vec_1, loss = triplet	0.0477**
NN: in = word2vec_2, loss = triplet	0.0002*

Single-starred and double starred numbers break 1 and 5% significance levels, respectively.

6.2. Word2Vec Models

Meanwhile, being somewhat struck by a unremarkable performance of the embedding model, we decided to explore the use of a pre-trained embedding model based on Word2Vec, which recently has proven its utility across a wide range of NLP tasks. Word2Vec is a single layer neural network whose primary goal is to predict a given word using its left and right contexts (CBOW) or words that occur closely to what is given as an input (Skip-Gram). Its importance lies not in its ability to predict missing words *per se*, but in the implication that a hidden structure built while training the model can be used as a latent “semantic” representation of word. To determine its utility in the current context, we did experiments with two versions of Word2Vec, one based on the Google News corpus (word2vec_1) [13], and the other built from the training data (word2vec_2). In either case, we made use of the Skip-Gram variant of Word2Vec (the latter model was trained with GENSIM[14]). We set the length of a hidden layer to 300, both for the version that employs the Google News and the one we created locally. The results are found in **Tables 2**, **3**.

Faced with the way they turned out, which was as underwhelming as the non-Word2Vec embedding model we saw previously, we came to a conclusion that the semantics may have little relevance to predicting target passages. The unimpressive performance of the embedding models stands in sharp contrast to the binary models, which rely only on the superficial overlap between source and target sentences to produce a more decent performance. The fact that BDOT—which looks at the amount of

[12] The statement here is not meant to suggest that *any form* of embedding will meet the same fate.

[13] GoogleNews-vectors-negative300.bin.gz (https://github.com/mmihaltz/word2vec-GoogleNews-vectors.git).

[14] https://radimrehurek.com/gensim/models/word2vec.html

tokens shared among the source and target to determine where the citation comes from—is closing in on the binary models lends a further support for the idea that a superficial match is a more reliable indicator of the target than the semantics served by the embedding models. Remarkably, the finding is consonant with what Li et al. (2016) found in their ablation study, who observed that Jaccard coefficient was by far the most effective feature among those they considered, including Word2Vec.

6.3. Factors Influencing MRRs

In light of the discussion so far, it would be interesting to ask whether there is any feature of the dataset that affects the models' performance. **Table 6** gives some insights, which lists the result of regressing MRR on features that we used previously to describe each dataset, namely, |RP|, CP, |*T*|, and |*C*|, where statistically significant features are noted with usual markings. It shows clearly that |RP|, the length of the reference paper is a single most significant predictor of the model's performance. In other words, the shorter the reference, the better the outcome. It turns out that neither the number of CPs nor that of targets is nearly as predictive as the length of RP.

In the meantime, **Table 4** looks at whether the choice for *f* has any effect on performance of the binary model. "mul" indicates the element-wise multiplication and "sqrd" the squared distance. The results clearly indicate that "mul" is a better choice, regardless of how the input is represented. Why this is so, is a curious question. One hypothesis is that the multiplication over binary inputs has the effect of eliminating all the words which occur only in *s* or *t*, which somehow caused some useful patterns to emerge, which NNs exploited. Though it is not clear at this time where the truth lies, a general lesson one might draw from the experiments is that how the inputs are encoded is at least as important as how the models are configured and some careful thinking must go into designing how to express what one feeds into NNs.

TABLE 6 | An analysis on how much individual cluster features affect the task performance, based on a multiple linear regression (where MRR is regressed on |RP|, CP, |*T*| and |*C*| at each span radius).

Predictor	Estimate	Std. error	*t*-statistic	*p*-value
±0				
$\lvert \mathrm{RP} \rvert$	−2.157e-04	8.763e-05	−2.461	0.0211**
CP	1.033e-04	2.183e-03	0.047	0.9626
$\lvert T \rvert$	−1.073e-03	1.156e-03	−0.928	0.3622
$\lvert C \rvert$	1.005e-03	2.071e-03	0.485	0.6317
±1				
$\lvert \mathrm{RP} \rvert$	−0.0002979	0.0001267	−2.351	0.0269**
CP	−0.0022792	0.0031563	−0.722	0.4769
$\lvert T \rvert$	−0.0013701	0.0016719	−0.820	0.4202
$\lvert C \rvert$	0.0022852	0.0029942	0.763	0.4525
±2				
$\lvert \mathrm{RP} \rvert$	−0.0004514	0.0001430	−3.157	0.00413*
CP	−0.0008764	0.0035613	−0.246	0.80763
$\lvert T \rvert$	−0.0017530	0.0018864	−0.929	0.36163
$\lvert C \rvert$	0.0021410	0.0033784	0.634	0.53201
±3				
$\lvert \mathrm{RP} \rvert$	−0.0005469	0.0001642	−3.330	0.0027*
CP	−0.0007249	0.0040910	−0.177	0.8608
$\lvert T \rvert$	−0.0015880	0.0021670	−0.733	0.4705
$\lvert C \rvert$	0.0019830	0.0038810	0.511	0.6139
±4				
$\lvert \mathrm{RP} \rvert$	−0.0006571	0.0002097	−3.133	0.00437*
CP	−0.0004621	0.0052238	−0.088	0.93022
$\lvert T \rvert$	−0.0018476	0.0027670	−0.668	0.51044
$\lvert C \rvert$	0.0015511	0.0049556	0.313	0.75688
±5				
$\lvert \mathrm{RP} \rvert$	−0.0007519	0.0002249	−3.344	0.00261*
CP	−0.0008374	0.0056014	−0.149	0.88237
$\lvert T \rvert$	−0.0007575	0.0029670	−0.255	0.80057
$\lvert C \rvert$	0.0006840	0.0053138	0.129	0.89860

** and ** indicate 1 and 5% significance levels, respectively.*

6.4. Summary of Findings

We conclude the section by highlighting some of the key findings from the experiments.

- Semantic representation: the token-wise overlap (BDOT/Jaccard) is a stronger indicator of a target than the implicit semantic representation induced via the embedding or Word2Vec (Note that by the embedding, we mean a *random* projection of a word into a continuous space, which obviously is distinct from a Word2Vec induced representation, as the latter is constructed using weights (associated with a particular layer) that Word2Vec learned during the training).
- CCS (automatic annotation) vs. TCS (manual annotation): we found statistically significant differences between CCS and TCS in terms of how they impact the performance, though how much susceptible models are to the differences varies from one model to another. "NN: in = binary, loss = triplet" benefitted more from a shift from TCS to CCS, compared to other models.
- Cross-entropy vs. triplet loss: there is a tendency for the former to result in a superior performance over the latter.

Before leaving the section, as a way to assist the reader with an intuitive understanding of what differences and similarities lie among the topic clusters, we provide in **Figure 7** plots of by-topic performance (in CCS) of "NN: in = binary, loss = entropy" on each of the relevant datasets.

7. CONCLUSIONS

We have presented approaches to linking citation and reference that draw upon neural networks (NNs), and described in detail what machinery is involved and what we found in experiments with the three datasets, TAC2014, DSA2016, and SRD2016. We introduced the notion of *approximately correct targets*, an idea that we should treat sentences that occur in the vicinity of true targets as equally correct, whereby we try to identify

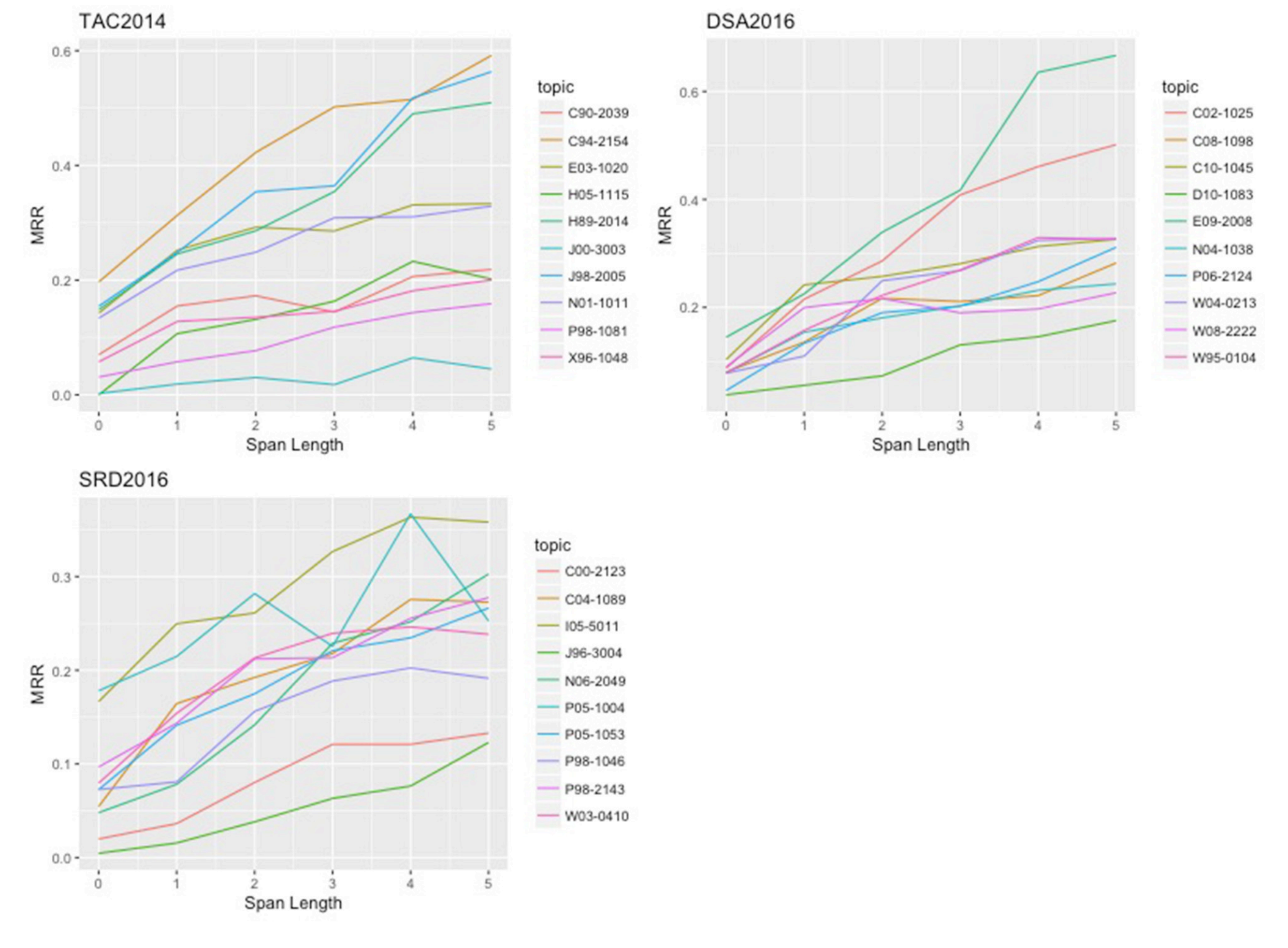


FIGURE 7 | Plot of by-topic performance of NN:binary+entropy.

an *area* which is likely to include a true target, instead of finding its exact location. The experiments found that expanding the target region by 5 sentences in radius led to a four fold increase in MRR across the models. Another curious fact the experiments brought to light was the significance of the way the input is expressed: it turned out that NNs worked visibly better with the binary representation than with either TFIDF or embeddings of the sort we considered in this paper. Also worthy of some attention is a finding that dispensing human created labels altogether led to an improvement (recall discussion on target- vs. citance-centric labeling). How it is so is an interesting question we have yet to answer. The paucity of human annotations, and the lack of consistent patterns in human labelings are some of the possible causes that immediately come to mind. To fully answer the question, however, may require finding out how well humans agree on their judgments as well as collecting additional data, topics we will leave to the future research.

AUTHOR CONTRIBUTIONS

The author confirms that he is the sole contributor of the present work and responsible for any error and misrepresentation thereof.

REFERENCES

Allen, J. (ed.) .(2002). *Topic Detection and Tracking: Event-Based Information Organization*. Dordrecht: Kluwer Academic Publishers.

Bethard, S., and Jurafsky, D. (2010). "Who should i cite: learning literature search models from citation behavior," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (New York, NY: ACM), 609–618. doi: 10.1145/1871437.1871517

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media Inc.

Bordes, A., Usunier, N., Weston, J., and Yakhnenko, O. (2013). "Translating embeddings for modeling multi-relational data," in *Advances in Neural Information Processing Systems* (Lake Tahoe, CA), 1–9.

Bordes, A., Weston, J., and Usunier, N. (2014). "Open question answering with weakly supervised embedding models," in *ECML PKDD 2014* (Nancy).

Brown, R. (2002). "Dynamic stopwording for story link detection," in *Proceedings of HLT 2002: Second International Conference on Human Language Technology* (San Diego, CA), 190–193.

Cao, Z., Li, W., and Wu, D. (2016). "Polyu at cl-scisumm 2016," in *BIRNDL 2016 Joint Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries* (Newark, NJ), 132–138.

Chen, F., Farahat, A., and Brants, T. (2003). "Story link detection and new event detection are asymmetric," in *Proceedings of HLT-NACCL 2003* (Edmonton), 13–15.

Chen, Y.-J., and Chen, H.-H. (2002). "NLP and IR approaches to monolingual and multilingual link dectection," in *The 19th International Conference on Computational Linguistics (COLING-2002)* (Taipei).

Jaidka, K., Chandrasekaran, M. K., Rustagi, S., and Kan, M.-Y. (2016). "Overview of the 2nd computational linguistics scientific document summarization shared task (cl-scisumm 2016)," in *The Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016)* (Newark, NJ).

Joachims, T. (2006). "Training linear SVMs in linear time," in *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)* (Philadelphia, PA).

Kingma, D., and Ba, J. L. (2015). "Adam: a method for stochastic optimization," in *Proceedings of the International Conference on Learning Representation* (San Diego, CA), 1–13.

Larkey, L. S., Feng, F., Connell, M., and Lavrenko, V. (2004). "Language-specific models in multilingual topic tracking," in *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY: ACM), 402–409.

Lavrenko, V., DeGuzman, J. A. E., LaFallme, D., Pollard, V., and Thomas, S. (2001). "Relevance models for topic detection and tracking," in *Proceedings of the Conference on Human Language Technology* (Stroudsburg, PA), 102–110.

Lee, K.-S., and Kageura, K. (2006). Korean-Japanese story link detection based on distributional and contrastive properties of event terms. *Inform. Process. Manage.* 42, 538–550. doi: 10.1016/j.ipm.2005.02.005

Li, L., Mao, L., Zhang, Y., Chi, J., Huang, T., and Heng Peng, X. C. (2016). "Cist system for cl-scisumm 2016 shared task," in *BIRNDL 2016 Joint Workshop on Bibliometric-Enhanced Information Retrieval and NLP for Digital Libraries* (Newark, NJ), 156–166.

Meij, E., and de Rijke, M. (2007). "Using prior information from citataions in literature search," in *Proceedings of RIAO2007* (Pittsburgh, PA).

Milne, D., and Witten, I. H. (2008). "Learning to link with wikipedia," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (New York, NY: ACM), 509–518. doi: 10.1145/1458082.1458150

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). "Distant supervision for relation extraction without labeled data," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2* (Stroudsburg, PA: Association for Computational Linguistics), 1003–1011.

Moraes, L., Baki, S., Verma, R., and Lee, D. (2016). "University of houston at cl-scisumm 2016: Svms with tree kernels and sentence similarity," in *BIRNDL 2016 Joint Workshop on Bibliometric-Enhanced Information Retrieval and NLP for Digital Libraries* (Newark, NJ), 113–121.

Nallapati, R. (2003). "Semantic language models for topic detection and tracking," in *Proceedings of the HLT-NAACL 2003 Student Research Workshop* (Edmonton), 1–6.

Nomoto, T. (2016). "NEAL:a neurally enhanced approach to linking citation and reference," in *The Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries* (Newark, NJ).

Ruder, S. (2016). An overview of gradient descent optimisation algorithms. *arxiv preprint arxiv:1609.04747*.

West, R., Precup, D., and Pineau, J. (2009). "Completing wikipedia's hyperlink structure through dimensionality reduction," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (New York, NY: ACM), 1097–1106. doi: 10.1145/1645953.1646093

Weston, J., Bengio, S., and Usunier, N. (2010). Large scale image annotation: Learning to rank with joint word-image embeddings. *Mach. Learn.* 81, 21–35. doi: 10.1007/s10994-010-5198-3

Weston, J., Bordes, A., Yakhnenko, O., and Usunier, N. (2013). "Connecting language and knowledge bases with embedding models for relation extraction," in *Empirical Methods in Natural Language Processing* (Seattle, WA), 1366–1371.

Zhang, X., Wang, T., and Chen, H. (2008). "Story link detection based on dynamic information extending," in *Proceedings of the Third International Join Conference on Natural Language Processing* (Hyderabad), 40–47.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Research Metrics and Analytics

ORIGINAL RESEARCH
published: 07 February 2019
doi: 10.3389/frma.2018.00036

The NLP4NLP Corpus (I): 50 Years of Publication, Collaboration and Citation in Speech and Language Processing

Joseph Mariani[1]*, *Gil Francopoulo*[2] *and Patrick Paroubek*[1]

[1] LIMSI-CNRS, Université Paris-Saclay, Orsay, France, [2] Tagmatica, Paris, France

This paper introduces the NLP4NLP corpus, which contains articles published in 34 major conferences and journals in the field of speech and natural language processing over a period of 50 years (1965–2015), comprising 65,000 documents, gathering 50,000 authors, including 325,000 references and representing ~270 million words. Most of these publications are in English, some are in French, German, or Russian. Some are open access, others have been provided by the publishers. In order to constitute and analyze this corpus several tools have been used or developed. Many of them use Natural Language Processing methods that have been published in the corpus, hence its name. The paper presents the corpus and some findings regarding its content (evolution over time of the number of articles and authors, collaborations between authors, citations between papers and authors), in the context of a global or comparative analysis between sources. Numerous manual corrections were necessary, which demonstrated the importance of establishing standards for uniquely identifying authors, articles, or publications.

Keywords: speech processing, natural language processing, text analytics, bibliometrics, scientometrics, informetrics

OPEN ACCESS

Edited by:
Iana Atanassova,
Université Bourgogne
Franche-Comté, France

Reviewed by:
Gérald Kembellec,
Conservatoire National des Arts et
Métiers (CNAM), France
Desfriches Doria Orelie,
Université de Lyon, France

****Correspondence:***
Joseph Mariani
joseph.mariani@limsi.fr

Received: *30 January 2018*
Accepted: *07 December 2018*
Published: *07 February 2019*

Citation:
Mariani J, Francopoulo G and Paroubek P (2019) The NLP4NLP Corpus (I): 50 Years of Publication, Collaboration and Citation in Speech and Language Processing. Front. Res. Metr. Anal. 3:36. doi: 10.3389/frma.2018.00036

This work is composed of two parts, of which this is part I. Please read also part II (Mariani et al., 2018).

INTRODUCTION

Preliminary Remarks

The aim of this study was to investigate a specific research area, namely Natural Language Processing (NLP), through the related scientific publications, with a large amount of data and a set of tools, and to report various findings resulting from those investigations. The study was initiated by an invitation of the Interspeech 2013 conference organizers to look back at the conference content on the occasion of its twenty-fifth anniversary. It was then followed by similar invitations at other conferences, by adding new types of analyses and finally by extending the data to many conferences and journals over a long time period. We would like to provide elements that may help answering questions such as: What are the most innovative conferences and journals? What are the most pioneering and influential ones? How large is their scope? How are structured the corresponding communities? What is the effect of the language of a publication? Which paradigms appeared and disappeared over time? Were there any epistemological ruptures? Is there a way to identify weak signals of an emerging research trend? Can we guess what will come next? What were the merits of authors in terms of paper production and citation, collaboration activities and innovation? What is the use of Language Resources in research? Do authors plagiarize each other? Do they publish similar papers in the same or in different conferences and journals? The results of this study are presented in two companion papers. The present one introduces the corpus with various analyses: evolution over time of the number of papers and authors, including their distribution by gender, as well as collaboration among authors and citation patterns among authors and papers. In the second paper (Mariani et al., 2018), we will consider the evolution of research topics over time and identify the authors who introduced and mainly contributed to key innovative topics, the use of Language Resources over time and the reuse of papers and plagiarism within and across publications. We provide both global figures corresponding to the whole data and comparisons of the various conferences and journals among those various dimensions. The study uses Natural Language Processing methods that have been published in the corpus considered in the study, hence the name of the corpus. In addition to providing a revealing characterization of the speech and language processing community, the study also demonstrates the need for establishing a framework for unique identification of authors, papers and sources in order to facilitate this type of analysis, which presently requires a heavy manual checking.

Text Analytics of Scientific Papers

The application of text analytics to bodies of scientific papers has become an active area of research in recent years (see for example Li et al., 2006; Tang et al., 2008; Dunne et al., 2012; Osborne et al., 2013; Ding et al., 2014; Gollapalli and Li, 2015; Jha et al., 2016). For example, the Stanford Large Network Dataset Collection (SNAP)[1] is a recently launched effort to study research networks by providing social networks and collaboration and citation graphs for conferences in Astrophysics, High Energy Physics, General Relativity and Condensed Matter. Studies of research publication data mine conference and workshop proceedings to determine trends in publications within a given area or field on various aspects, such as various kinds of collaboration networks, authors and papers citation graphs, author/topic pairings, topic shifts over time, authors and participants demographics, with the goal of better understanding research trends, collaborations, participation and publication data, etc. In the field of Speech and Natural Language Processing (SNLP), several studies of this type have recently been conducted, including the following:

- ACL Anthology[2] (Bird et al., 2008) analysis (Radev et al., 2013), presented in several papers at the Association for Computational Linguistics (ACL) workshop entitled "Rediscovering 50 Years of Discoveries in Natural Language Processing" on the occasion of ACL's fiftieth anniversary in 2012[3]. The workshop included the contributions of 23 authors through 13 papers (Banchs, 2012).
- Analysis of 25 years of research contained in the International Speech Communication Association (ISCA) Archive[4] (assembled by Wolfgang Hess) published in proceedings of various conferences in the ISCA series [e.g., European Conference on Speech Technology (ECST), Eurospeech, International Conference on Spoken Language Processing (ICSLP), Interspeech] between 1987 and 2012 (Mariani et al., 2013).
- Analysis of the proceedings of the TALN conference organized yearly by the French ATALA (*Association pour le Traitement Automatique des Langues*) (Boudin, 2013)[5].
- Results from the Saffron[6] project, which performs automatic analysis of proceedings in the areas of Natural Language Processing [LREC, the ACL Anthology (ACL Annual Conferences, COLING, EACL, HLT, ANLP)], Information Retrieval [CLEF (Cross Language Evaluation Forum)], and the Semantic Web (Semantic Web Dog Food) and publishes its results as linked data (Bordea et al., 2014).
- Analysis of 15 years of research contained in the Language Resources and Evaluation Conference (LREC) proceedings between 1998 and 2012 (Mariani et al., 2014a) then 15 + 2 years, adding LREC 2014 (Mariani et al., 2016).
- Analysis of 20 years of research in Language Technology as published in the Language and Technology Conference (L&TC) from 1995 to 2015 (Mariani et al., 2015).

Studies of this kind can reveal patterns and shifts that may otherwise go unnoticed, and which can ultimately affect perceptions and practices in a given field. For example, an analysis conducted on publications from the IEEE ICASSP conference series between 1976 and 1990 (Mariani, 1990) showed that the percentage of papers on speech decreased over time, from about 50% in 1976 to 30% in 1990. Further analysis showed that the US produced most of the papers on speech (> 50%) within the conference, including on those years when the ICASSP conference took place outside the US; however at these conferences, the total participation increased, including a virtually undiminished level of US participation together with a dramatic increase in the number of European and Asian participants. As a result of this analysis, the speech community decided to begin organizing fully international conferences specifically devoted to spoken language processing, namely Eurospeech in Europe, starting in 1989 (Mariani, 2013), and ICSLP in Asia, starting in 1990 (Fujisaki, 2013).

The NLP4NLP Speech and Natural Language Processing Analysis

In order to conduct this study, we produced a corpus containing research papers on spoken and written language processing, called the NLP4NLP corpus, a name chosen to reflect the fact that the study uses NLP methods that are presented in papers contained in the corpus content itself (Francopoulo et al.,

[1] http://snap.stanford.edu/data/

[2] https://aclanthology.coli.uni-saarland.de/

[3] Results of these analyses together with corresponding data and tools are available on-line at the University of Michigan http://clair.eecs.umich.edu/aan/index.php

[4] http://www.isca-speech.org/iscaweb/index.php/archive/online-archive

[5] Available online at: http://talnarchives.atala.org/TALN/TALN-2013/taln-2013-court-001.pdf

[6] http://saffron.insight-centre.org/

2015a,b). The NLP4NLP corpus contains papers from thirty-four conferences and journals on natural language processing (NLP) and spoken language processing (SLP) published over 50 years (1965–2015) (**Table 1**), thereby providing a good picture of research within the international SNLP community. However, we should stress the fact that many papers, including important papers, related to this field may have been published in other publications than those. We included material from conferences and journals only, as workshops may have widely varying ways of reviewing papers. For the conferences, we will call *venue* the event constituted by holding the conference. Conferences may have different frequencies. They may have annual venues, appear every

TABLE 1 | The NLP4NLP Corpus of Conferences (24) and Journals (10).

Short name	# Docs	Format	Long name	Language	Access to content	Period	# Venues
acl	4,264	Conference	Association for Computational Linguistics Conference	English	Open*	1979–2015	37
acmtslp	82	Journal	ACM Transactions on Speech and Language Processing	English	Private	2004–2013	10
alta	262	Conference	Australasian Language Technology Association	English	Open*	2003–2014	12
anlp	278	Conference	Applied Natural Language Processing	English	Open*	1983–2000	6
cath	932	Journal	Computers and the Humanities	English	Private	1966–2004	39
cl	776	Journal	American Journal of Computational Linguistics	English	Open*	1980–2014	35
coling	3,813	Conference	Conference on Computational Linguistics	English	Open*	1965–2014	21
conll	842	Conference	Computational Natural Language Learning	English	Open*	1997–2015	18
csal	762	Journal	Computer Speech and Language	English	Private	1986–2015	29
eacl	900	Conference	European Chapter of the ACL	English	Open*	1983–2014	14
emnlp	2,020	Conference	Empirical methods in natural language processing	English	Open*	1996–2015	20
hlt	2,219	Conference	Human Language Technology	English	Open*	1986–2015	19
icassps	9,819	Conference	IEEE International Conference on Acoustics, Speech and Signal Processing—Speech Track	English	Private	1990–2015	26
ijcnlp	1,188	Conference	International Joint Conference on NLP	English	Open*	2005–2015	6
inlg	227	Conference	International Conference on Natural Language Generation	English	Open*	1996–2014	7
isca	18,369	Conference	International Speech Communication Association	English	Open	1987–2015	28
jep	507	Conference	Journées d'Etudes sur la Parole	French	Open*	2002–2014	5
lre	308	Journal	Language Resources and Evaluation	English	Private	2005–2015	11
lrec	4,552	Conference	Language Resources and Evaluation Conference	English	Open*	1998–2014	9
ltc	656	Conference	Language and Technology Conference	English	Private	1995–2015	7
modulad	232	Journal	Le Monde des Utilisateurs de L'Analyse des Données	French	Open	1988–2010	23
mts	796	Conference	Machine Translation Summit	English	Open	1987–2015	15
muc	149	Conference	Message Understanding Conference	English	Open*	1991–1998	5
naacl	1,186	Conference	North American Chapter of the ACL	English	Open*	2000–2015	11
paclic	1,040	Conference	Pacific Asia Conference on Language, Information and Computation	English	Open*	1995–2014	19
ranlp	363	Conference	Recent Advances in Natural Language Processing	English	Open*	2009–2013	3
sem	950	Conference	Lexical and Computational Semantics/Semantic Evaluation	English	Open*	2001–2015	8
speechc	593	Journal	Speech Communication	English	Private	1982–2015	34
tacl	92	Journal	Transactions of the Association for Computational Linguistics	English	Open*	2013–2015	3
tal	177	Journal	Revue Traitement Automatique du Langage	French	Open	2006–2015	10
taln	1,019	Conference	Traitement Automatique du Langage Naturel	French	Open*	1997–2015	19
taslp	6,612	Journal	IEEE/ACM Transactions on Audio, Speech and Language Processing	English	Private	1975–2015	41
tipster	105	Conference	Tipster DARPA text program	English	Open*	1993–1998	3
trec	1,847	Conference	Text Retrieval Conference	English	Open	1992–2015	24
Total incl. duplicates	67,937					1965–2015	577
Total excl. duplicates	65,003					1965–2015	558

Joint conferences and the corresponding papers are counted once in the total number of venues and documents.
**Included in the ACL Anthology.*

TABLE 2 | Sources attached to each of the three research areas.

Research area	Sources	# Docs
NLP oriented	acl, alta, anlp, cath, cl, coling, conll, eacl, emnlp, hlt, ijcnlp, inlg, lre, lrec, ltc, mts, muc, naacl, paclic, ranlp, sem, tacl, tal, taln, tipster, trec	28,027
Speech oriented	acmtslp, csal, icassps, isca, jep, lre, lrec, ltc, mts, speechc, taslp	43,056
IR oriented	modulad, muc, tipster, trec	2,333

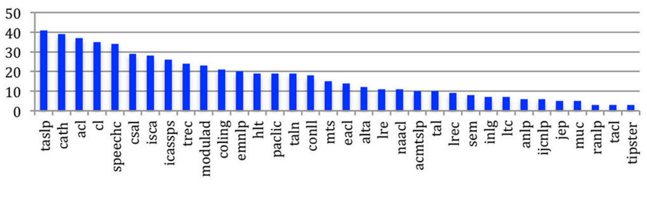


FIGURE 1 | Number of venues or issues for each source.

2 years on even years (this is the case usually for COLING, EACL, JEP, LREC) or on odd years (IJCNLP, L&TC, RANLP). They may also be organized jointly in the same year. For the journals, we will call *issue* a set of papers corresponding to a volume or to a year.

In the present paper, we used the entire corpus to study collaboration among authors and citations of authors and papers in general, but also within each source, and from and to each source, as it gives an analysis on how the community related to each source considers and is being considered by its general scientific environment. A study of reuse and plagiarism within each source but also across sources has also been conducted and is presented in a companion paper.

In order to study the possible differences across different communities, we considered 3 different research areas, Speech, NLP, and Information Retrieval (IR), and we attached the sources to each of those areas (**Table 2**), given that some sources (LREC, LRE, L&TC, MTS) may be attached to several research domains. We see that the number of documents related to Speech is larger than the one related to NLP, and much larger than the one related to IR. We only considered the papers related to Speech processing (named ICASSPS) in the IEEE ICASSP conference, which also includes a large number of papers on Acoustics and Signal Processing in general.

The number of venues, for the conferences, or issues, for the journals, may strongly vary (**Figure 1**), from 41 venues for the *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, which changed its name over the years (initially *Transactions on Acoustics, Speech and Signal Processing* from 1974 to 1990, then *Signal Processing* until 1993, then *Speech and Audio processing* until 2006, then *Audio, Speech, and Language Processing* before merging in 2013 with the *ACM Transactions on Speech and Language Processing*) to 3 venues for Tipster, RANLP or the recently created *Transactions of the ACL (TACL)*. The time span is also different, from 50 years for COLING to 3 years for the *Transactions of the ACL* (**Figure 2**).

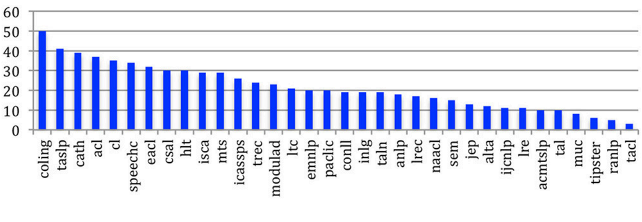


FIGURE 2 | Time span for each source (years).

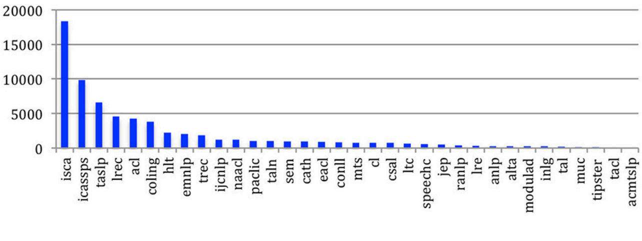


FIGURE 3 | Number of documents for each source.

The number of papers across sources may therefore also strongly vary, from 18,369 for the ISCA conference series to 82 in the case of the *ACM Transactions on Speech and Language Processing* (ACMTLSP) (**Figure 3**).

GLOBAL ANALYSIS OF THE CONFERENCES AND JOURNALS

As a convention, we refer to each conference or journal as a *source* and the conference or journal publication as a *document*. A *paper* or *article* corresponds to a *document* that may have been published in one or several conference series when presented at a joint conference. We refer to individual *authors* and mention their *authorships*, *contributions*, or *signatures* to a publication where they act as *contributors*. The same author may sign several papers at a given conference, as a single author or together with one or several *co-authors*.

Number of Sources Over the Years

As it appears in **Table A1**, the number of sources, including conferences and journals, globally increased over the year but seems now to be stabilizing at 34 (**Figure 4**).

However some conferences are biennial and other only occur from time to time. Some conferences as well as some journals also stopped. Therefore, the number of sources may fluctuate over the years (**Figure 5**), even if the total number globally increases. We took into account the sources we have access to. For example, ACL was founded in 1963 and the first ACL conference took place in 1965. However, we only had access to the content of the ACL conference, through the ACL Anthology, starting in 1979. The number of sources decreases on the last year that we take into account (2015), as some biennial conferences didn't take place on that year (e.g., Coling, LREC, EACL) and because some of the data was only available later in 2016.

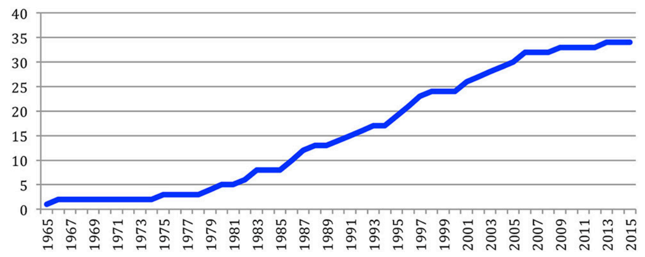


FIGURE 4 | Cumulated number of different sources (conferences and journals) over the years.

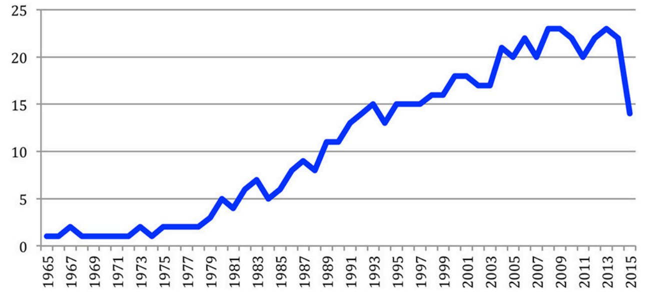


FIGURE 5 | Number of sources (conferences and journals) considered each year.

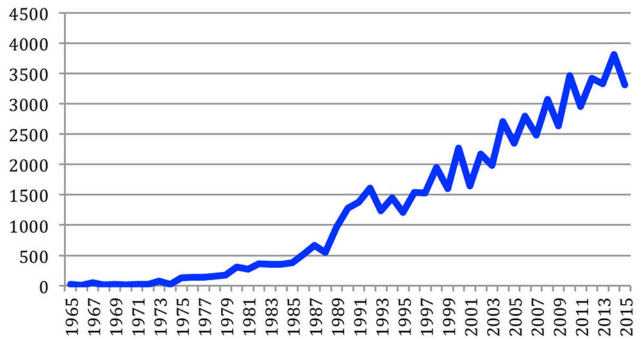


FIGURE 6 | Number of papers each year.

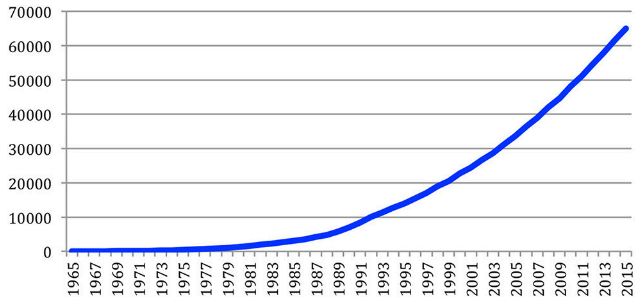


FIGURE 7 | Cumulated number of papers over the years.

Journals

The following journals have been considered: *Computer and the Humanities* (since 1966), IEEE *Transactions on Acoustics, Speech and Signal Processing* and the following titles (since 1975), *Computational Linguistics* (since 1980), *Speech Communication* (since 1982), *Computer Speech and Language* (since 1986), *Modulad* (since 1988), the ACM *Transactions on Speech and Language Processing* (since 2004), *Language Resources and Evaluation* and *TAL* (since 2006) and the *Transactions of the ACL* (since 2013). Most of those publications are in English, except *TAL* and *Modulad* that are mainly in French.

Conferences

The following conferences have been considered: Coling (since 1965), Conference of the ACL (since 1979), ANLP and EACL (since 1983), HLT (since 1986), the "ISCA" conference series (ECST, Eurospeech, Interspeech, ICSLP) and the MT Summit (since 1987), the part devoted to speech and language processing in the IEEE ICASSP conferences (since 1990), MUC (since 1991), TREC (since 1992), and TIPSTER (since 1993), L&TC and PACLIC (since 1995), EMNLP and INLG (since 1996), CONLL and TALN (since 1997), LREC (since 1998), NAACL and Semeval (since 2001), JEP (since 2002), ALTA (since 2003), IJCNLP (since 2005) and RANLP (since 2009). Most of those conferences are in English, except JEP and TALN that are mainly in French.

Documents

Over the years, 67,937 documents have been published in the 34 sources. However, this number comprises papers that were published at joint conferences. The total number of different papers thus reduces to 65,003 (**Table 1**), with a steady increase over time from 24 papers in 1965 to 3,314 in 2015 (**Figure 6**). The number of documents fluctuates over the years, mainly due to the biennial frequency of some conferences. The largest number of papers has been published in 2014 (3,817 papers).

The total number of papers itself still increases steadily at a high rate, reaching 65,003 different documents as of 2015 (**Figure 7**).

Data and Tools

Origin of Data

Most of the proceedings are freely available online on the ACL Anthology website, others are freely available in the ISCA Archive. The corresponding websites include metadata (list of authors and sessions, content of the sessions and, for each article, title, authors, affiliations, abstract, and bibliographic references) as well as the full content of the articles. IEEE ICASSP and TASLP have been obtained through the IEEE, and LRE through Springer, while their website also includes metadata (for each article, title, authors, affiliations, abstract, and bibliographic references). For this study, we only considered the papers written in English and French, but it should be stressed that the papers may contain examples in many different languages.

Extraction and Quality of Data

Most of the documents are available in PDF. Those that are only available as scanned images had to be transferred in a PDF format. In order to do so, a preprocessing was applied in a first step, to extract the textual content by means of PDFBox (Litchfield, 2005) and when the document consisted in a sequence of images, the Optical Character Recognizer (OCR) system Tesseract-OCR[7] was called to produce a textual content.

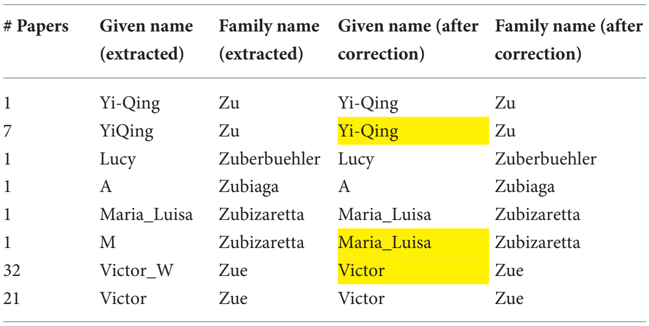
A benchmark to estimate the error rate of the extracted content was established based on a simple heuristics, which is that "rubbish" character strings are not entries in lexicons. This estimation is computed as the number of unknown words divided by the number of words. The number of errors was computed from the result of the morphological module of TagParser (Francopoulo, 2008), a deep industrial parser based on a broad English lexicon and Global Atlas (a knowledge base containing more than one million words from 18 Wikipedias) (Francopoulo et al., 2013). Variations in performance quality measures were used to control the parameterization of the content preprocessing tools.

Following this content extraction, another step in our preprocessing was dedicated to split the content into abstract, body and references sections. Initially, we attempted to use ParsCit (Councill et al., 2008), which had been used to extract citations from the ACL Anthology; however, it was not suited for Slavic, German, extended Latin, and phonetic alphabets included in our data, and retraining the program would have required too much time. We therefore created a small set of rules in Java to extract the abstract and body of the papers and compute their quality, which yielded a 2.5% higher performance than ParsCit.

The result of the preprocessing is summarized in **Table A2**, and it can be noticed that the corpus contains close to 270 million words. We see that the overall quality improved over time. We extracted from those papers the sections related to the abstract and to the references, which didn't exist or could not be extracted in some cases.

Manual Checking and Correction

The study of authors is problematic due to variations of the same name (family name and given name, initials, middle initials, ordering, married name, etc.). It therefore required a tedious semi-automatic cleaning process (Mariani et al., 2014b). On the first survey we conducted on the ISCA archive, about two thirds of the raw family names or given names had to be corrected or harmonized: starting from an initial list of 51,145 authors' names, it resulted in a list of 16,540 different authors. Given the tedious nature of this manual checking process, a cost-benefit perspective suggests that we focus on the data that have the greatest influence on survey goals. Normalizing the names of authors who published only one or two papers over 50 years has only a small effect compared with the required effort. This is especially important given that more than half of the authors (26,870 upon 48,894) published only one paper. In contrast, resolving the different names of an active author is important, because otherwise this person will not appear with the correct ranking. **Figure 8** provides an example of this cleaning process, which focuses on the most prolific authors according to the number of papers they published, as merging variant wordings may drastically change their ranking (see the case of Victor Zue/Victor W. Zue, with 53 papers in total). This suggests a need to determine ways to uniquely identify researchers, which has been proposed (Joerg et al., 2012), and may also be solved through organisms, such as ORCID[8].



# Papers	Given name (extracted)	Family name (extracted)	Given name (after correction)	Family name (after correction)
1	Yi-Qing	Zu	Yi-Qing	Zu
7	YiQing	Zu	Yi-Qing	Zu
1	Lucy	Zuberbuehler	Lucy	Zuberbuehler
1	A	Zubiaga	A	Zubiaga
1	Maria_Luisa	Zubizaretta	Maria_Luisa	Zubizaretta
1	M	Zubizaretta	Maria_Luisa	Zubizaretta
32	Victor_W	Zue	Victor	Zue
21	Victor	Zue	Victor	Zue

FIGURE 8 | Example of cleaning authors' given names and family names. Values colored in yellow indicate manual corrections.

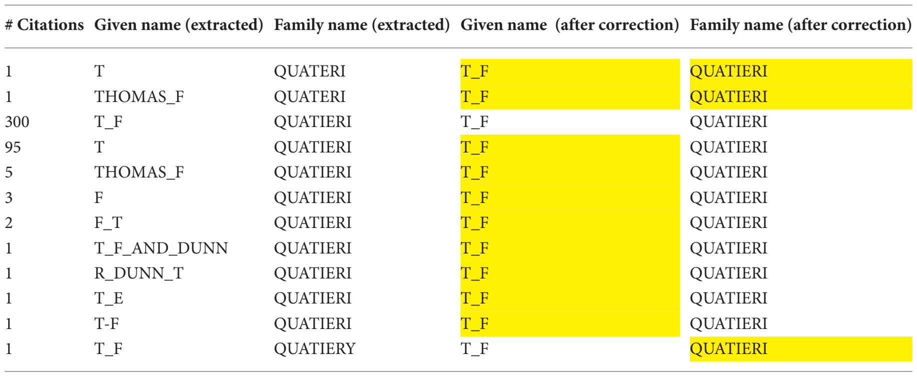
The same process was applied to the analysis of the authors cited in papers. The problem is even more difficult, as the data is extracted from the paper content and may therefore contain segmentation errors. Also the number of cited papers' authors is much larger than the number of papers' authors. We first automatically cleaned the data by using the results of the former process on the authors' names, before conducting a manual cleaning. Here also the focus is put on the most cited authors. In the example of **Figure 9**, the number of citations appears in the first column. Merging variant wordings may drastically change the ranking (from 300 to 412 citations for T.F. Quatieri, for example).

Similarly, we also had to clean the sources of the citations, which may belong to several categories: conferences and workshops, journals or books. The cleaning was first conducted on a single year. The resulting filter was then used for all the years, and the full data received a final review. Here also, the focus is put on the most cited sources, as merging variant wordings change their ranking, and only the most cited sources were considered (more than five citations). **Figure 10** provides an example for IEEE-ICASSP, where the number of mentions appears on the first column.

The analysis of the acknowledgments of the Funding bodies in the papers also necessitated a manual cleaning. The nationality of each funding agency was introduced, and the spelling variants were harmonized in order to estimate the agencies and countries that are the most active in funding research on SNLP. **Figure 11** provides an example for the French National Research Agency (ANR), including cases where several Funding Agencies are mentioned. The nationality of the Funding Agency is also included.

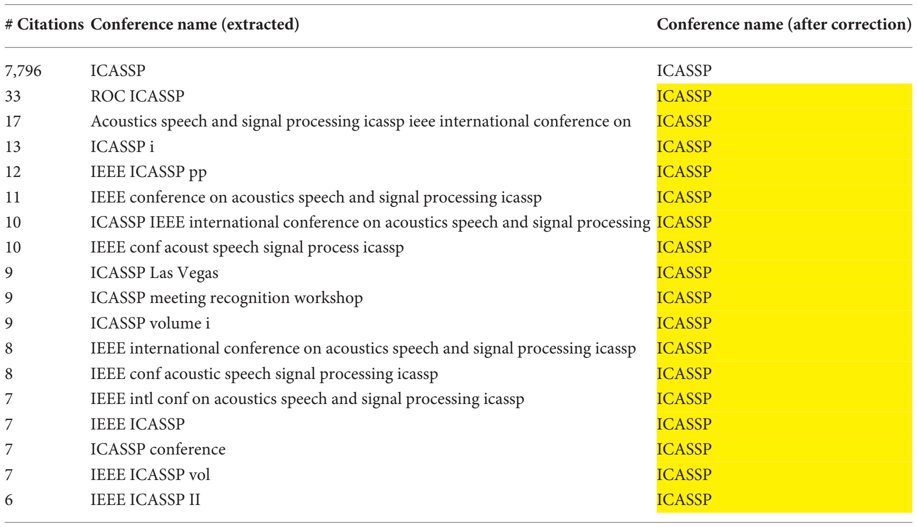
[7] https://code.google.com/p/tesseract-ocr/

[8] Open Researcher and Contributor ID.



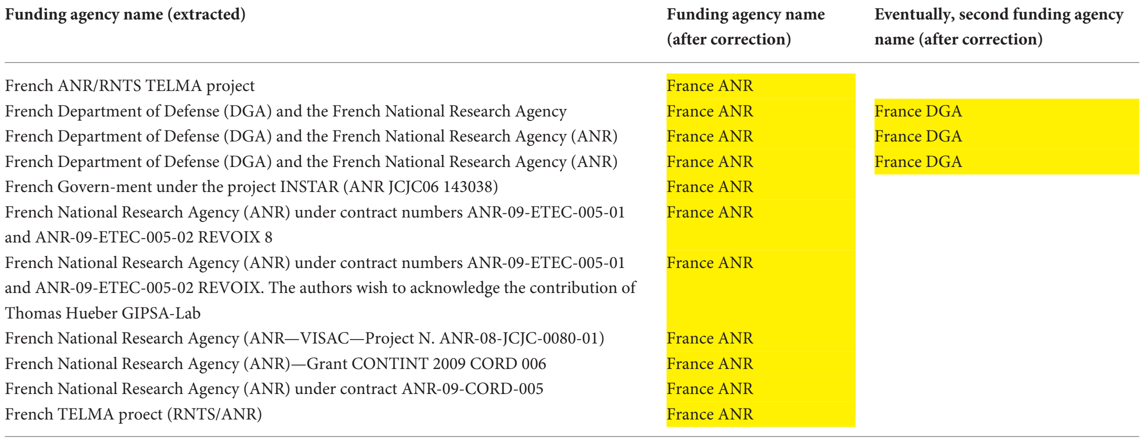
# Citations	Given name (extracted)	Family name (extracted)	Given name (after correction)	Family name (after correction)
1	T	QUATERI	T_F	QUATIERI
1	THOMAS_F	QUATERI	T_F	QUATIERI
300	T_F	QUATIERI	T_F	QUATIERI
95	T	QUATIERI	T_F	QUATIERI
5	THOMAS_F	QUATIERI	T_F	QUATIERI
3	F	QUATIERI	T_F	QUATIERI
2	F_T	QUATIERI	T_F	QUATIERI
1	T_F_AND_DUNN	QUATIERI	T_F	QUATIERI
1	R_DUNN_T	QUATIERI	T_F	QUATIERI
1	T_E	QUATIERI	T_F	QUATIERI
1	T-F	QUATIERI	T_F	QUATIERI
1	T_F	QUATIERY	T_F	QUATIERI

FIGURE 9 | Example of cleaning cited authors' given names and family names: the case of T.F. Quatieri.



# Citations	Conference name (extracted)	Conference name (after correction)
7,796	ICASSP	ICASSP
33	ROC ICASSP	ICASSP
17	Acoustics speech and signal processing icassp ieee international conference on	ICASSP
13	ICASSP i	ICASSP
12	IEEE ICASSP pp	ICASSP
11	IEEE conference on acoustics speech and signal processing icassp	ICASSP
10	ICASSP IEEE international conference on acoustics speech and signal processing	ICASSP
10	IEEE conf acoust speech signal process icassp	ICASSP
9	ICASSP Las Vegas	ICASSP
9	ICASSP meeting recognition workshop	ICASSP
9	ICASSP volume i	ICASSP
8	IEEE international conference on acoustics speech and signal processing icassp	ICASSP
8	IEEE conf acoustic speech signal processing icassp	ICASSP
7	IEEE intl conf on acoustics speech and signal processing icassp	ICASSP
7	IEEE ICASSP	ICASSP
7	ICASSP conference	ICASSP
7	IEEE ICASSP vol	ICASSP
6	IEEE ICASSP II	ICASSP

FIGURE 10 | Example of cleaning cited conferences: the case of IEEE ICASSP.



Funding agency name (extracted)	Funding agency name (after correction)	Eventually, second funding agency name (after correction)
French ANR/RNTS TELMA project	France ANR	
French Department of Defense (DGA) and the French National Research Agency	France ANR	France DGA
French Department of Defense (DGA) and the French National Research Agency (ANR)	France ANR	France DGA
French Department of Defense (DGA) and the French National Research Agency (ANR)	France ANR	France DGA
French Govern-ment under the project INSTAR (ANR JCJC06 143038)	France ANR	
French National Research Agency (ANR) under contract numbers ANR-09-ETEC-005-01 and ANR-09-ETEC-005-02 REVOIX 8	France ANR	
French National Research Agency (ANR) under contract numbers ANR-09-ETEC-005-01 and ANR-09-ETEC-005-02 REVOIX. The authors wish to acknowledge the contribution of Thomas Hueber GIPSA-Lab	France ANR	
French National Research Agency (ANR—VISAC—Project N. ANR-08-JCJC-0080-01)	France ANR	
French National Research Agency (ANR)—Grant CONTINT 2009 CORD 006	France ANR	
French National Research Agency (ANR) under contract ANR-09-CORD-005	France ANR	
French TELMA proect (RNTS/ANR)	France ANR	

FIGURE 11 | Example of cleaning cited Funding Agencies: the case of the French ANR.

Tools

After this preprocessing phase, the metadata and contents are ready to be processed by higher level tools based on the R statistical suite (The R Journal, 2012), iGraph (Csárdi and Nepusz, 2006), the search engine swish-e[9], RankChart, Tulip (Auber et al., 2012) and a series of Java programs that we wrote (Francopoulo et al., 2015a,b, 2016).

Overall Analysis

Papers and Authors

The number of authors varies across the sources, from 16,540 different authors who published in the ISCA conference series to 156 different authors at Tipster (**Figure 12**).

The number of documents per venue or per issue may also vary across the sources (**Figure 13**). The ISCA conferences are the conferences that publish the largest number of papers in a single event (656 papers on average), followed by LREC (506), ICASSP-Speech (378), IJCNLP (198) and Coling (182). The *ACM Transactions on Speech and Language Processing* only had 8 papers on average at each issue.

Accordingly, the number of authorships also rose steadily, from 32 in 1965 to 11,457 in 2015 (**Figure 14**).

Co-authorship

The number of co-authors per paper is most often two to three (**Figure 15**). The largest number of co-authors for a paper is 44, in a paper published by the META-NET[10] EC project partners at LREC 2014. The average number of co-authors per paper increased over time, from 1.33 in 1965 up to 3.45 in 2015 (i.e., two more authors on average) (**Figure 16**). It is interesting to notice that the number of papers with a single author was 75% in 1965 and decreased to 5% in 2015. This clearly demonstrates the change in the way research is being conducted, going progressively from individual research investigations to large projects conducted within teams or in collaboration within consortia, often in international projects and programs.

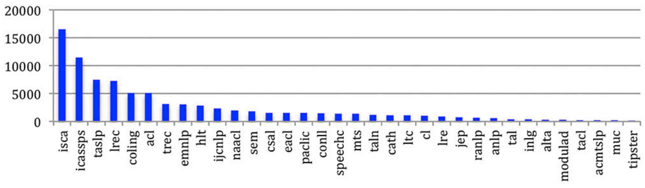


FIGURE 12 | Number of different authors having published at each source.

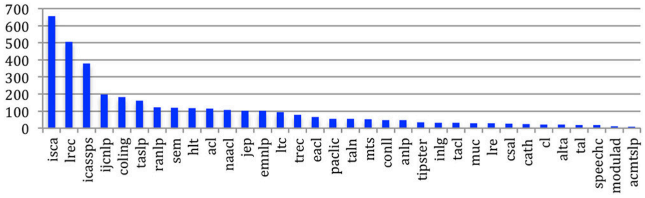


FIGURE 13 | Average number of documents at each venue (conferences) or issue (journals).

The average number of co-authors per paper also varies across the sources (**Figure 17**). TREC, MUC, Semeval and the LREC conference, as well as the LRE Journal, show the largest number of co-authors per paper, while journals, such as *Computer*

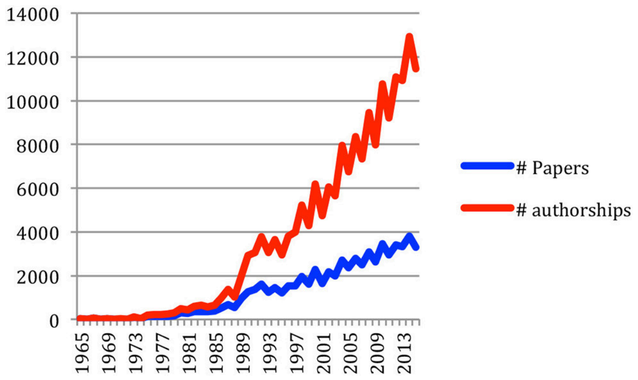


FIGURE 14 | Number of papers and authorships over time.

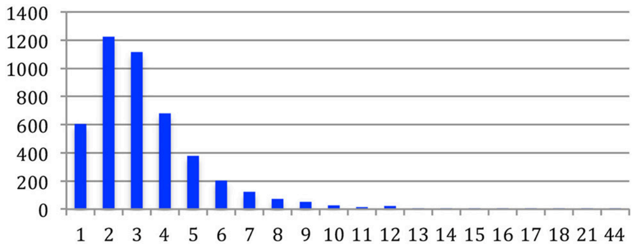


FIGURE 15 | Number of papers according to the number of co-authors.

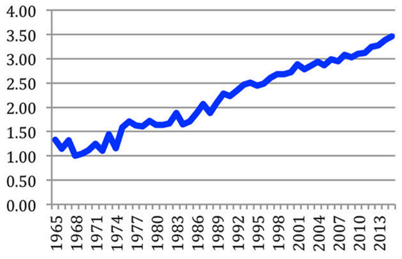


FIGURE 16 | Average number of authors per paper.

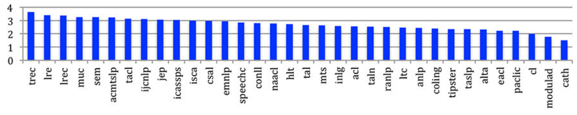


FIGURE 17 | Average number of authors per paper across the sources.

[9] http://www.searchtools.com/tools/swish.html

[10] Multilingual Europe Technology Alliance Network.

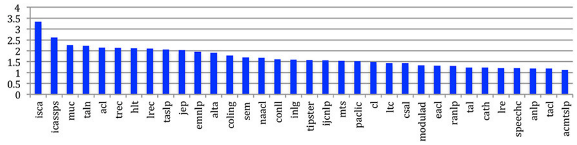


FIGURE 18 | Average number of papers published by each different author across the sources.

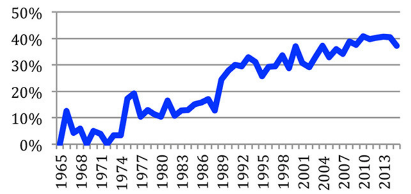


FIGURE 19 | Author redundancy over time.

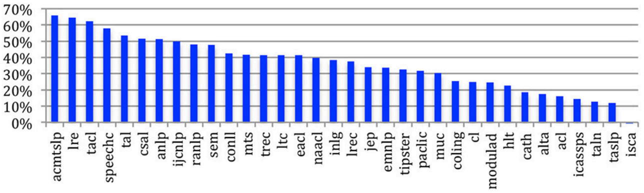


FIGURE 20 | Author redundancy across the sources.

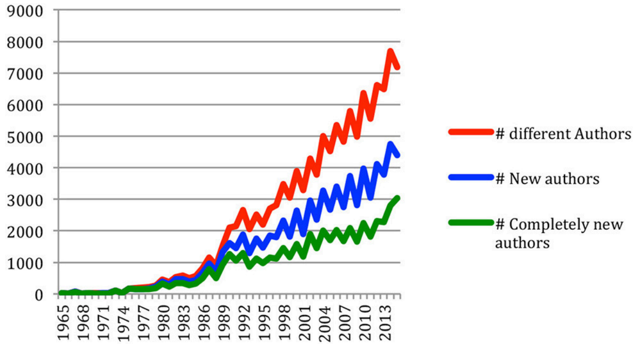


FIGURE 21 | Number of different authors, new authors and completely new authors over time.

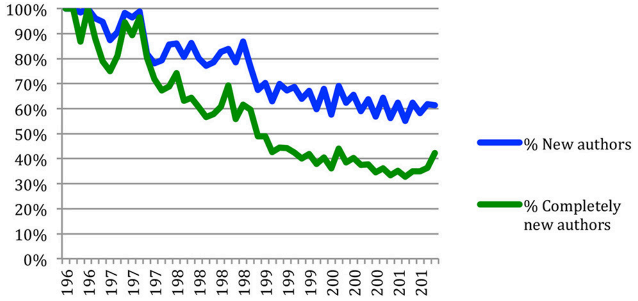


FIGURE 22 | Percentage of new authors and completely new authors over time.

and the Humanities, *Modulad* and *Computational Linguistics* and conferences, such as PACLIC, EACL, and ALTA show the smallest number of co-authorship on average.

Authors' Renewal and Redundancy

We studied the number of repeated authors at successive conferences (**Table A3**). For each conference, we identified the authors who did not publish at the previous conference (*new authors*). We also studied those who had not published at any previous conference (*completely new authors*).

The ratio of the total number of papers (65,003) to the overall number of different authors (48,894) represents the global productivity of the community: each author published on average 1.33 papers over 50 years. The ratio of the total number of authorships (184,050) to the overall number of different authors (48,894) represents the individual productivity of each author: each author contributed on average in 3.76 papers over 50 years.

If we consider the situation across the sources (**Figure 18**), we see that ISCA and ICASSPS authors are very productive, with an average of more than 2.5 papers per author, while the productivity in journals is naturally much lower (about one paper per author on average).

The ratio of the number of different authors to the number of authorships at each conference reflects the *variety* of authors. This ratio would be 100% if each author's name appears on a single paper. We define *author redundancy* as 100%-*author variety*. It appears that this redundancy increased over time and has now stabilized at about 40% (**Figure 19**).

If we consider this measure across the sources (**Figure 20**), we see that this redundancy is of course very large in journals while it is very low in the ISCA conference series, where the number of authors is even larger than the number of papers.

We then studied the authors' renewal. It clearly showed (**Figure 21**) that the number of different authors globally increased over time. The number of new authors from one conference to the next similarly increased over time. The same trend applies to the number of completely new authors, which still increased in 2015 with 3,033 new authors who never published at any of the NLP4NLP conferences and journals before!

This same trend applies to percentages of different authors from 1 year to the next (**Figure 22**), which decreased from 100% in 1966 to 61% in 2015, while the number of completely new authors decreased from 100% in 1966 to about 42% in 2015. This suggests a stabilization of the research community over time, but it also still reflects the existence of "new blood" in the field.

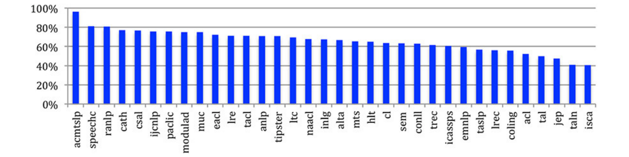


FIGURE 23 | Percentage of completely new authors in the last venue/issue across the sources.

If we consider the percentage of completely new authors at the last venue of conferences or the last issue of journals (**Figure 23**), we see that this percentage ranges from 40 to 80%, and even to 96% in the case of the ACM *Transactions on Speech and Language Processing*. The large conferences show the lowest percentages (from 41% for ISCA to 52% for ACL, 56% for COLING and LREC and 61% for IEEE ICASSPS).

Authors' Gender

An author gender study was performed with the help of a lexicon of 27,509 given names with gender information (66% male, 31% female, 3% epicene[11]). As noted above, variations due to different cultural habits for naming people (single vs. multiple given names, family vs. clan names, inclusion of honorific particles, ordering of the components etc.) (Fu et al., 2010), and changes in editorial practices and sharing of the same name by large groups of individuals contribute to make identification by name a real issue (Vogel and Jurafsky, 2012). In some cases, we only had an initial for the first name, which made gender guessing impossible unless the same person appears with his/her first name in full in another publication. Although the result of the automatic processing was hand-checked by an expert of the domain for the most frequent names, the results presented here should therefore be considered with caution, allowing for an error margin.

The analysis over the 34 sources shows that 49% of the authors are male, while 14% of the authors are female and 37% are of unknown gender, either because their given name is epicene, or because we only have the initials of the given name. If we assume that the authors of unknown gender have the same gender distribution as the ones that are categorized, male authors account for 77% and female authors for 23%. If we now consider the authorships, which take into account the authors' productivity, we see that 61% of the signatures are male, while 13% are female and 26% are of unknown gender (**Figure 24**). If we assume that the authors of unknown gender have the same gender distribution as the ones that are categorized, male authors account for 82% and female authors for 18% of the published papers (**Figure 25**).

If we consider the situation across the various sources (**Figure 26**), we see that the IEEE *Transactions on Speech and Language Processing* and ICASSPS have the largest participation of male authors (respectively 90 and 88%), while the French conferences and journals, together with LRE and LREC have the smallest (from 63 to 70%).

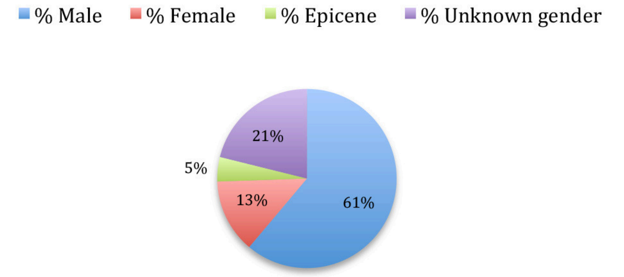


FIGURE 24 | Authorships' gender.

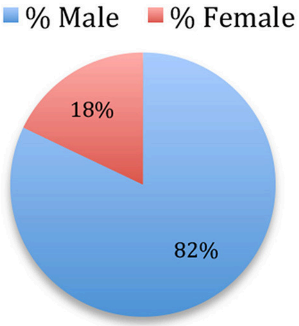


FIGURE 25 | Extrapolated authorships' gender.

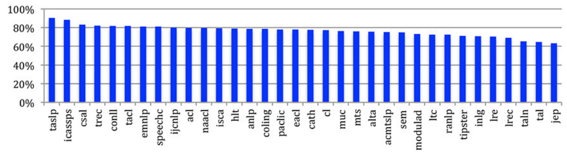


FIGURE 26 | Percentage of male authors across the sources.

The analysis of the authors' gender over time (**Figure 27**) shows that the ratio of female authorship slowly increased over time from 10% to about 20%.

Authors' Production and Co-production

The most productive author published 358 papers, while 26,870 authors (about 55% of the 48,894 authors) published only one paper (**Figure 28**). **Table 3** gives the list of the 10 most productive authors, accompanied by the number of papers they published as a single author. **Table 4** gives the number of authors who published papers as single authors. 42,471 authors (87% of the authors) never published a paper as single author[12].

[11] "Epicene" means that the given name is gender ambiguous.

[12] Keynote papers are not always taken into account if they were not included in the conference programs or proceedings.

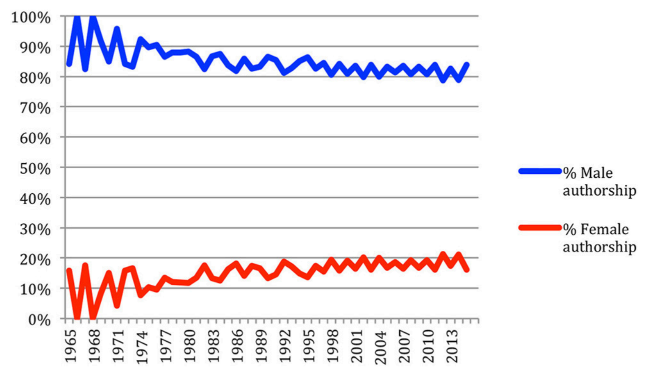


FIGURE 27 | Gender of the authors' contributions over time.

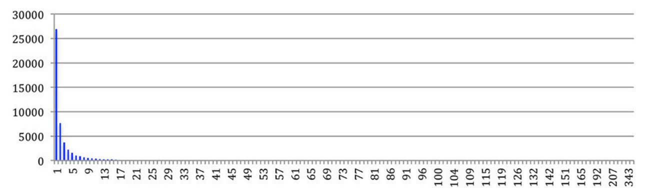


FIGURE 28 | Number of papers per number of authors.

TABLE 3 | Ten most productive authors, including the number of papers published as single author.

Name	Number of papers (= number of authorships)	Number of papers as single author
Shrikanth S. Narayanan	358	0
Hermann Ney	343	10
John H. L. Hansen	299	3
Haizhou Li	257	1
Chin-Hui P. Lee	218	5
Alex Waibel	207	2
Satoshi Nakamura	205	1
Mark J. F. Gales	195	9
Lin-Shan Lee	193	0
Li Deng	192	6
Keikichi Hirose	187	1
Kiyohiro Shikano	184	0

Collaborations

Authors' Collaborations

The most collaborating author published with 299 different co-authors, while 2,401 authors always published alone (**Figure 29**). On average, an author collaborated with 6.6 other authors. 108 authors published with 100 or more different co-authors (**Table 5**).

We may also consider the number of collaborations, possibly with the same co-authors. **Table 6** gives the list of the 12 authors who have the largest number of collaborations.

TABLE 4 | Number of single author papers.

# Papers	# Authors	Author name
0	42,471	...
1	4,402	...
2	1,038	...
3	416	...
4	211	...
5	131	...
6	76	...
7	49	...
8	27	...
9	24	...
10	10	Aravind K. Joshi, Eckhard Bick, Hermann Ney, Hugo Van Hamme, Joshua T. Goodman, Karen Spärck Jones, Kuldip K. Paliwal, Mark Hepple, Raymond S. Tomlinson, Roger K. Moore
11	10	Dekang Lin, Eduard H. Hovy, Jörg Tiedemann, Marius A. Pasca, Michael Schiehlen, Olov Engwall, Patrick Saint-Dizier, Philippe Blache, Stephanie Seneff, Tomek Strzalkowski
12	9	David S. Pallett, Harvey F. Silverman, Jen-Tzung Chien, Kenneth Ward Church, Lynette Hirschman, Martin Kay, Reinhard Rapp, Ted Pedersen, Yorick Wilks
13	4	John Makhoul, Paul S. Jacobs, Rens Bod, Robert C. Moore
14	2	Dominique Desbois, Sadaoki Furui
15	2	Donna Harman, Takayuki Arai
16	2	Jerry R. Hobbs, Steven M. Kay
17	2	Beth M. Sundheim, Kenneth C. Litkowski
18	3	Douglas B. Paul, Mark A. Johnson, Rathinavelu Chengalvarayan
20	1	Olivier Ferret
21	1	Ralph Grishman
25	1	Ellen M. Voorhees
26	1	Jerome R. Bellegarda
27	1	W. Nick Campbell

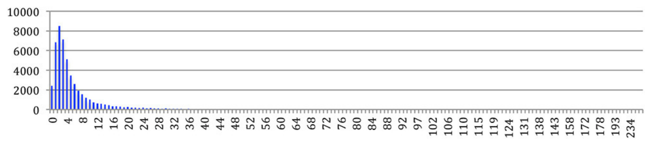


FIGURE 29 | Number of authors as a function of the number of different co-authors.

Collaboration Graph

A *collaboration graph*[13] (CollG) is a model of a social network where the *nodes* (or vertices) represent participants of that network (usually individual people) and where two distinct participants are joined by an *edge* whenever there is a collaborative relationship between them. As opposed to a citation graph, a CollG is undirected. It contains no *loop-edge* (an author does not collaborate with himself/herself) and no *multiple edges* (there is a single edge between two authors, whatever the number of papers they published together).

[13] http://en.wikipedia.org/wiki/Collaboration_graph

TABLE 5 | The 12 authors with the largest number of co-authors.

Name	# Co-authors
Shrikanth S. Narayanan	299
Hermann Ney	254
Haizhou Li	252
Satoshi Nakamura	234
Alex Waibel	212
Mari Ostendorf	199
Chin-Hui P. Lee	194
Sanjeev Khudanpur	193
Frank K. Soong	188
Lori Lamel	185
Hynek Hermansky	179
Yang Liu	178

TABLE 6 | The 12 authors with the largest number of collaborations.

Name	# Collaborations
Shrikanth S. Narayanan	1,035
Haizhou Li	899
Hermann Ney	890
Satoshi Nakamura	672
Alex Waibel	580
Chin-Hui P. Lee	544
Richard M. Schwartz	534
John H. L. Hansen	520
Lori Lamel	513
Bin Ma	503
Li Deng	498
Andreas Stolcke	491

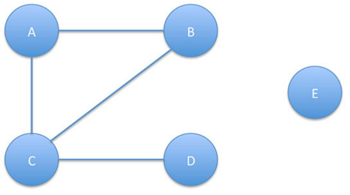


FIGURE 30 | Collaboration graph.

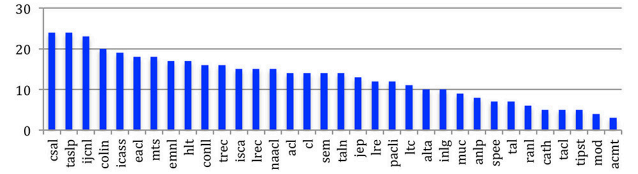


FIGURE 31 | Diameter of the CollG for the 34 sources.

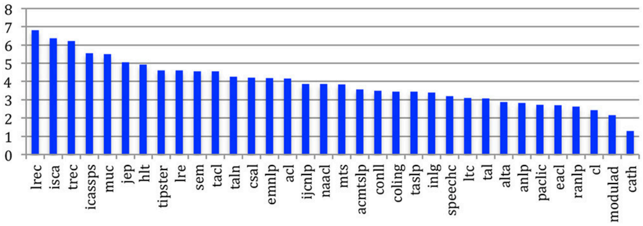


FIGURE 32 | Mean degree of the CollG for the 34 sources.

As it appears in **Figure 30**, the CollG nodes need not be fully connected, i.e., people who never co-authored a joint paper are represented by isolated nodes (E). Those who are connected constitute a *connected component* (this is the case for A, B, C, D). When a connected component gathers a majority of the nodes, it may be called a *giant component*. *Cliques* are fully connected components where all authors published with one another. The *collaboration distance* is the geodesic distance, or path-length, between two nodes in a CollG, which is equal to the smallest number of edges in an edge-path, or *collaboration path*, connecting them. The *diameter* of the CollG is the longest collaboration path in that graph. If no path connecting two nodes in a CollG exists, the collaboration distance between them is considered to be infinite. The *degree* of a node (number of edges attached to the node) reflects the number of co-authors associated with each author, as an absolute measure of his/her collaboration activity. The *clustering coefficient* of a node is a measure of the degree to which its neighboring nodes tend to cluster together: i.e., how close they are to form a clique. The *density* of a graph is the fraction of all possible edges that actually exists in the CollG, thus providing a measure of the density of collaboration: if all authors have published at least one paper with all the other authors, the density of collaboration of the graph would be equal to 1.

The NLP4NLP CollG contains 48,894 nodes corresponding to the 48,894 different authors. There are 162,497 edges. The global diameter is 17. Five pairs have this distance. The sources with the largest diameter are *Computer Speech and Language* and the *IEEE Transactions on Audio, Speech and Language Processing* (24), which reflects the cohesion of the related communities (**Figure 31**).

The mean degree (average number of co-authors for each author) is 6.6. It goes from over 6 for LREC, ISCA and TREC to close to 1 for *Computer and the Humanities*, given that this journal starts being considered very early in the 60s, a period when authors did not collaborate as much as today (**Figure 32**). The max degree (corresponding to the author who collaborated with the largest number of different co-authors) is 299 (as already mentioned in **Table 5**).

The density of the complete CollG is 0.0001. If we consider the difference across the sources, we see that this density goes from 0.03 for Tipster and 0.025 for MUC, which corresponds to evaluation campaigns where there is a strong collaboration

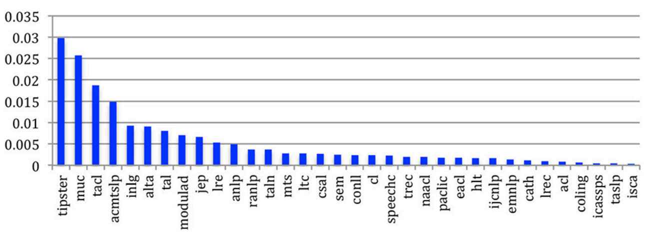


FIGURE 33 | Density of the CollG for the 34 sources.

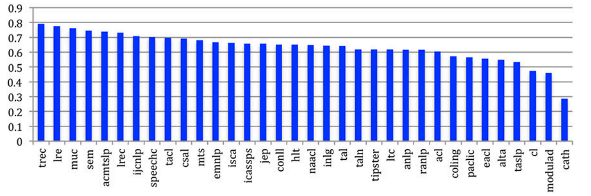


FIGURE 34 | Average clustering coefficient of the CollG for the 34 sources.

among all the authors, to 0.0004 (almost 100 times less) for the ISCA conference series (**Figure 33**).

The average clustering coefficient is 0.6. It goes from more than 0.7 for conferences related to evaluation campaigns (TREC, MUC and Semeval), where the collaboration is strong, to <0.3 for *Computer and the Humanities* (**Figure 34**).

Connected Components

As shown in **Table 7**, the CollG contains 4,585 connected components. The largest one groups 39,744 authors, which means that 81% of the 48,894 authors are connected through a collaboration path. The authors of the largest connected component published 58,208 papers (89% of the total number of papers), and the average path length is 5.5. The second connected component groups 29 authors, who published together but never with any of the 39,744 previous ones. The remaining connected components contain far fewer authors, each of whom has never published with any of the authors of the largest connected component; these components tend to represent small communities often related to the study of a specific topic or a specific language. As already mentioned, 5% of the authors (2,401) have never published jointly with any other author. As it turned out, in our corpus the largest clique could be identified by simply looking at the paper with the largest number of co-authors [44 co-authors in the LREC 2014 paper related to the *Multilingual Europe Technology Alliance Network* (META-NET)].

Figure 35 gives the percentages of authors in the largest Connected Component for the 34 sources. We see that some conferences, either international (ISCA, LREC, ICASSPS, EMNLP, HLT) or national (jep, taln), are more focused than others where the collaboration is sparser. For twelve sources, the largest Connected Component gathers more than 50% of the nodes and may therefore be considered as a Giant Component.

TABLE 7 | Connected components in the collaboration graph.

Connected component size	# Of connected components	# Of authors	% Of authors in the connected components	% Of connected components
39,744	1	39,744	81	0
29	1	29	0	0
27	1	27	0	0
21	1	21	0	0
18	3	54	0	0
17	1	17	0	0
15	1	15	0	0
14	1	14	0	0
12	2	24	0	0
11	9	99	0	0
10	5	50	0	0
9	14	126	0	0
8	26	208	0	1
7	38	266	1	1
6	60	360	1	1
5	120	600	1	3
4	252	1,008	2	5
3	535	1,605	3	12
2	1,113	2,226	5	24
1	2,401	2,401	5	52
39,963	4,585	48,894	100	100

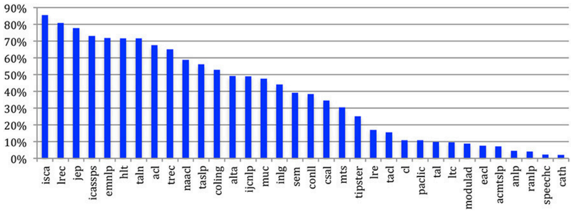


FIGURE 35 | Percentage of authors in the largest connected component of the CollG for the 34 sources.

Measures of Centrality

We explored the role of each author in the CollG in order to assess his/her centrality. In graph theory, there exist several types of centrality measures (Freeman, 1978). The *Closeness distance* has been introduced in Human Sciences to measure the efficiency of a Communication Network (Bavelas, 1948, 1950). It is based on the shortest geodesic distance between two authors regardless of the number of collaborations between the two authors. The *Closeness centrality* is computed as the average closeness distance of an author with all other authors belonging to the same connected component. More precisely, we use the *harmonic centrality* which is a refinement introduced recently by Rochat (2009) of the original formula to take into account the whole graph in one step instead of each connected component separately. The *degree centrality* is simply the number of different co-authors of each author, i.e., the number of edges attached to the corresponding node. The *betweenness centrality* is based on the number

TABLE 8 | Computation and comparison of the closeness centrality, degree centrality and betweenness centrality for the 10 most central authors.

Closeness centrality			Degree centrality		Betweenness centrality		
Author's name	**Harmonic centrality**	**Norm on first**	**Author's name**	**Index and norm on first**	**Author's name**	**Index**	**Norm on first**
Mari Ostendorf	11,958	1	Shrikanth S. Narayanan	1	Shrikanth S. Narayanan	23,492,104	1
Shrikanth S. Narayanan	11,890	0.994	Hermann Ney	0.854	Haizhou Li	21,312,971	0.907
Chin Hui P. Lee	11,869	0.993	Haizhou Li	0.854	Satoshi Nakamura	20,451,472	0.871
Hermann Ney	11,824	0.989	Satoshi Nakamura	0.784	Chin Hui P. Lee	18,488,513	0.787
Haizhou Li	11,803	0.987	Alex Waibel	0.714	Hermann Ney	16,131,472	0.687
Julia B. Hirschberg	11,756	0.983	Mari Ostendorf	0.671	Frank K. Soong	15,473,696	0.659
Nelson Morgan	11,700	0.978	Sanjeev Khudanpur	0.648	Alex Waibel	14,639,035	0.623
Sanjeev Khudanpur	11,659	0.975	Chin Hui P. Lee	0.645	Yang Liu	13,433,061	0.572
Satoshi Nakamura	11,657	0.975	Frank K. Soong	0.635	Lori Lamel	13,160,473	0.56
Alex Waibel	11,655	0.975	Lori Lamel	0.625	Khalid Choukri	13,150,169	0.56

of paths crossing a node and reflects the importance of an author as a bridge across different sets of authors (or sub-communities).

Looking at **Table 8**, we see that some authors who appear in the Top 10 according to the Closeness Centrality also appear in the other two types of centrality, eventually with a different ranking, while others do not.

Citations

Papers' Citations

We studied citations in papers that are accessible in digital form. 58,204 papers contain a list of references, and the number of missing references decreases over time as the quality of the source data increases (see **Table A2**).

If we consider the average number of references in papers, we see that it increased over time from close to 0 in 1965 to 8.5 in 2015 (**Figure 36**). Even if we only consider here the NLP4NLP data, its seems that it is a general trend that goes together with the citing habits and the increase of the number of published papers in the literature.

If we now consider the average number of citations per NLP4NLP paper over the years (**Figure 37**), the trend is less clear. Obviously the most recent papers are less cited than the older ones, with an average number of more than seven citations for the papers of the most cited year (2003) and 0.4 citations on average for the papers published in 2015, given that they have only been cited by the papers published on the same year, but the eldest papers before 1974 are also cited less than once on average.

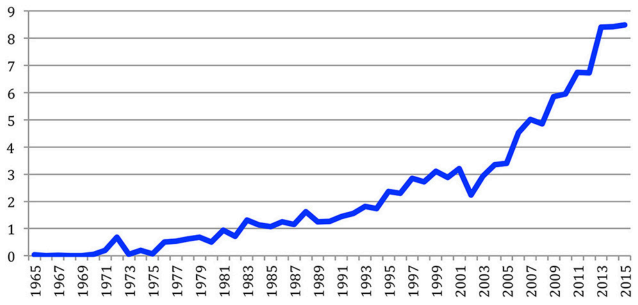


FIGURE 36 | Average number of references per paper over the years.

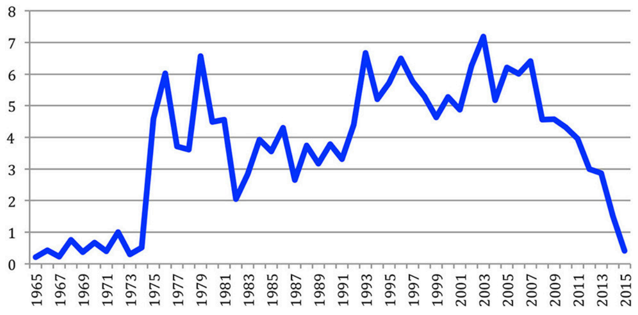


FIGURE 37 | Average number of citations per paper over the years.

The comparative study of the number of references and of the number of citations over the years for the 34 sources is difficult to handle. If we limit this study to the eight most important conferences (ACL, COLING, EACL, EMNLP, ICASSP, ISCA, LREC, NAACL), we see that the number of references strongly increased over time in the ISCA conference series (**Figure 38**). This is directly in agreement with the ISCA Board policy which decided in 2005 to enlarge the number of pages in the yearly conference papers from 6 to 7, with the rule that the allowed extra page should only consist of references, in order to encourage authors to better cite the work of the other authors. The saw tooth aspect of LREC, EACL, and NAACL is due to the fact that those conferences are biennial.

Similarly, it is difficult to analyze the variation of cited papers over time (**Figure 39**). Here also the saw tooth aspect of LREC, EACL, and NAACL is due to the fact that those conferences are biennial.

In order to solve this problem mostly due to the conference frequency, we may integrate the number of papers being cited **up**

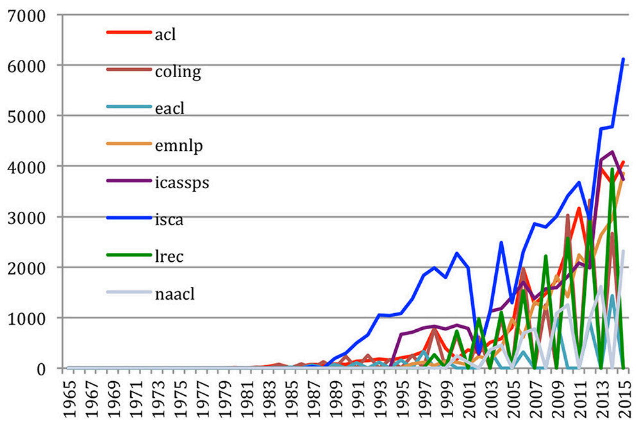


FIGURE 38 | Number of references in papers over the years for the eight most important conferences.

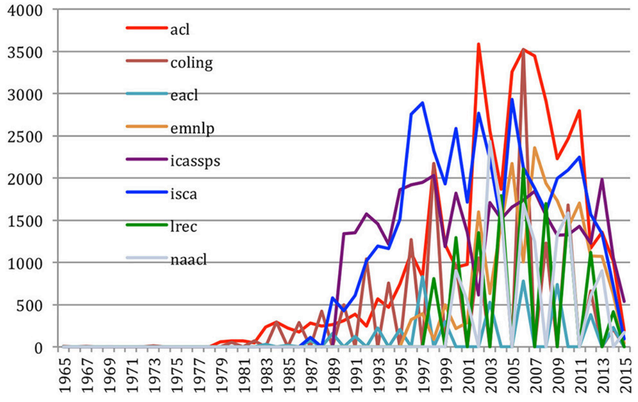


FIGURE 39 | Number of papers being cited over the years for the eight most important conferences.

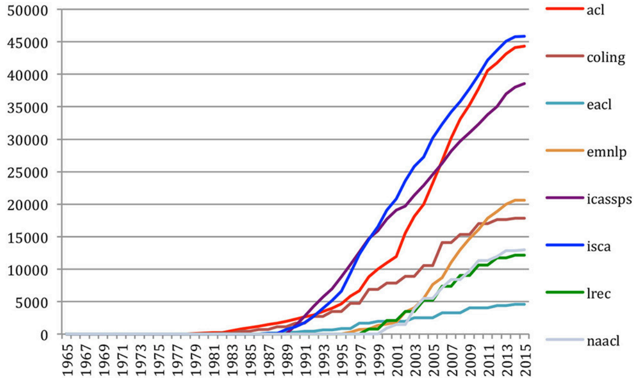


FIGURE 40 | Number of papers that have been cited over the years for the eight most important conferences.

to the given year. In this case, we see (**Figure 40**) that the number of ISCA papers being cited grows at a high rate over time. The same appears for ACL with some delay, which is now caught up.

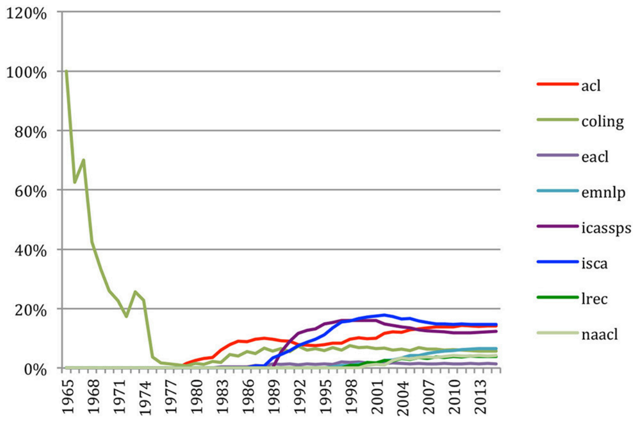


FIGURE 41 | Percentage of the papers that have been cited over the years for the eight most important conferences.

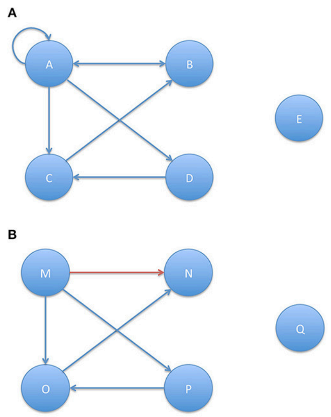


FIGURE 42 | (A) Authors' citation graph. **(B)** Papers' citation graph.

ICASSPS comes in the third position. We then find a group of two with COLING and EMNLP, followed by LREC and NAACL. Then comes EACL.

Finally, we studied the same in terms of percentage over time for each of the 8 conferences (**Figure 41**). We find the same group of 3 (ISCA, ACL, and ICASSPS) at the first rank in 2015 with 12–15% of the citations. COLING, which was alone in 1965 is now at 6% close to EMNLP (7%), while LREC and NAACL represent 4% each and EACL 1% of the citations.

Citation Graph

Unlike the CollG, a *citation graph* (CitG) is directed. In an *authors citation graph* (ACG), nodes (or vertices) represent individual authors (**Figure 42A**). We may consider the *citing authors graph (CgAG)*, in which a citing author is linked to all the authors of the papers that he/she cites by an edge directed toward those authors, and the *cited authors graph (CdAG)*, where each cited author is linked to the authors who cite him/her by an edge directed toward this author. These graphs may have *loop-edges*, as an author may cite and be cited by him/herself, but they have no *multiple edges*:

there is only one edge between two authors, whatever the number of times an author cites or is being cited by another author.

In a *papers citation graph* (PCG), nodes represent individual papers (**Figure 42B**). Here also, we may consider the *citing papers graph (CgPG)*, in which a paper is linked to all the papers it cites by an edge directed toward those papers, and the *cited papers graph (CdPG)*, where each paper is linked to all the papers that cite it by an edge directed toward those papers. These graphs contain *no loop-edge*, as a paper does not cite itself, and no *multiple edges*: there is only one edge between two papers, whatever the number of times a paper cite or is being cited by another paper. Bi-directional arrows are common in ACGs (as Author A may cite Author B while Author B cites Author A), but uncommon in PCGs (if Paper M cites Paper N, it is very unlikely that Paper N will cite Paper M, as papers typically reference papers that have been already published. It may however happen in case of simultaneous publications).

The citation graphs need not be connected, as an author may not cite any author and may not be cited by any author, not even him/herself (E), or a paper may not cite any paper and may not be cited by any other paper (Q); in these cases, corresponding authors or papers appear as isolated nodes in the citation graphs. The nodes that are connected through a directed path (as it is the case for A, B, C, D in **Figure 42A** where Author A cites Authors B, C, and D, and himself/herself, Author B cites Author A, Author C cites Author B and Author D cites Author C), constitute a *strongly connected component*. If the nodes are connected in both directions, they constitute a *symmetric strongly connected component* (**Figure 43**).

The *citation distance* between two nodes is the smallest number of directed edges in an edge-path connecting them. The *diameter* of a citation graph is the longest path in the graph, which is identical in both the citing and cited graphs. If no path connecting two nodes in a citation graph exists, the citation distance between them is said to be infinite. In a citing graph, the degree of a node (the number of directed edges issued from that node) reflects the absolute number of authors (or papers) cited by each author (or paper). In a cited graph, the degree of a node reflects the absolute number of authors (or papers) citing each author (or paper). As in the CollG, the *clustering coefficient* of a node is a measure of the degree to which its neighbors tend to cluster together. The *density* of a citation graph, which is the fraction of possible edges that exist in the graph, provides a measure of the density of citation: if all authors (or papers) cite at least once each other author (or paper), the density of citation of the graph would be equal to 1.

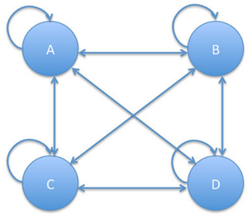


FIGURE 43 | Authors' citation graph symmetric connected component.

We studied the four Citing and Cited/Authors and Papers Graphs for each of the 34 sources, either internally or in the context of the NLP4NLP corpus, which also includes the individual source and represents the general Speech and Natural Language Processing scientific community (SNLP).

We thus studied:

- the citation in the source papers of papers of the same source (*Internal Papers Citations*: the citations within the source) (**Figure 44A**),
- the citation in the source papers of NLP4NLP papers, including those from the same source (*Outgoing Global Papers Citations:* how the source cites its scientific environment, which also includes the source) (**Figure 44B**),
- the citation in NLP4NLP papers of the source papers (*Ingoing Global Papers Citations:* how the source is being cited by its scientific environment, which also includes the source) (**Figure 44C**).

Similarly, we also studied:

- the citation by the source authors of the source authors (*Internal Authors Citations*),
- the citation by the source authors of SNLP authors (*Outgoing Global Authors Citations*),
- the citation by SNLP authors of the source authors (*Ingoing Global Authors Citations*).

where the "source authors" means the authors for the papers they have published in the source, while they may also have published elsewhere.

We give some elements of comparison across sources, keeping in mind that the time scales are different, as well as the frequency and number of venues for conferences (9 venues over 17 years for LREC, to be compared with 28 venues over 27 years for ISCA or 36 venues over 35 years for ACL, for example), or the number of publications for journals.

We considered the 67,937 papers we have in NLP4NLP, which include 324,422 references (**Table A2**).

Authors' Citations

Internal authors' citations

We first consider *internal authors citations*: the citation by authors, in the source papers, of authors for their source papers.

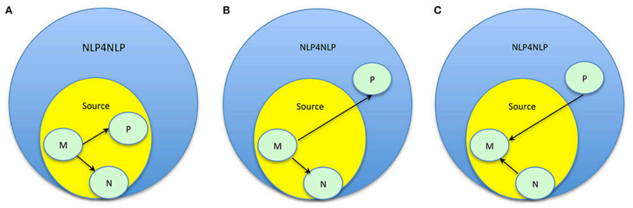


FIGURE 44 | (A) Example of internal citing papers graph: source paper M cites source papers N and P. **(B)** Example of outgoing global citing papers graph: Source paper M cites NLP4NLP papers N and P. **(C)** Example of ingoing global citing papers graph: NLP4NLP papers N and P cite source paper M.

If we consider for the 34 sources the average number of authors (*mean degree*) from the source being cited by the authors of papers of the same source (**Figure 45**) in the CgAG, we see that some communities, such as ACL and EMNLP are used to cite each other. Let's mention that the Mean Degree of the internal Citing Authors Graph (CgAG) is equal to the Mean Degree of the internal Cited Authors Graph (CdAG).

The density reaches 0.008 for MUC, 0.006 for Tipster and 0.005 for Semeval, which correspond to evaluation campaigns where there are many cross-citations among all the authors (**Figure 46**).

For ten sources, the largest Strongly Connected Component gathers more than 50% of the nodes and may be considered as Giant Components. The *Computational Linguistics* journal has the largest Strongly Connected Component, which contains 72% of the authors. It is followed by several ACL related sources (EMNLP, CONLL, HLT, NAACL, ACL, TACL) that illustrates the way authors highly cite each other in this community (**Figure 47**).

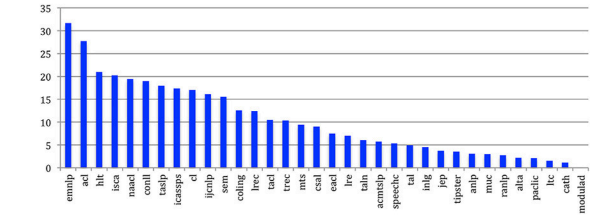


FIGURE 45 | Mean degree of authors citing and being cited within their community for the 34 sources.

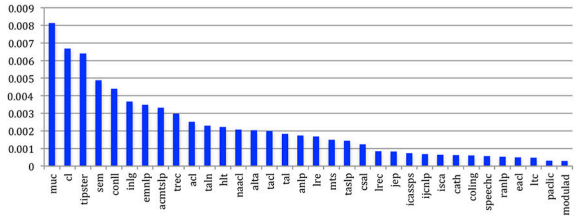


FIGURE 46 | Density of the internal authors citation graph.

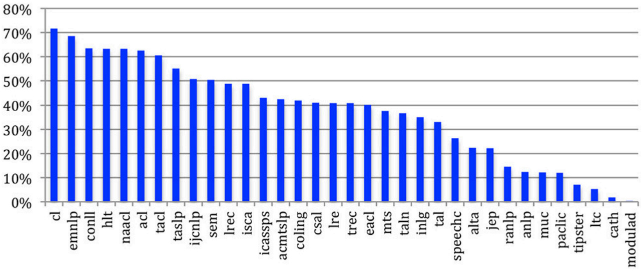


FIGURE 47 | Percentage of authors in the largest strongly connected component.

We compared LREC, ACL, and ISCA (**Table 9**). The largest strongly connected component for LREC has 3,581 nodes among the 7,282 LREC authors (49% of the authors). This is comparable to ISCA (49%), but less than ACL (63%) and illustrates a less focused network of citations than ACL.

In LREC, the number of strongly connected components with symmetric links is 4,798 (**Table 9**). The largest strongly connected component with symmetric links includes 43 authors who all cite each other and correspond to partners in the French Quaero project. It attains 99 authors in ISCA (**Figure 48**).

Global authors' citations

We now consider *global authors citations*: the citation by authors, in papers published in each source, of SNLP authors.

If we now consider the general habit of **citing** other authors (**Figure 49**), we also see that the NLP community (TACL,

TABLE 9 | Comparison of LREC, ACL, and ISCA internal Cg/CdAG strongly connected components, without or with symmetric links.

Internal citing/cited authors graphs (Cd/CgACGs)	lrec	acl	isca
# Of strongly connected components	3,581	1,912	8,102
Size of the largest strongly connected component	3,626	3,140	8,322
% Of authors in the largest strongly connected component	49%	63%	49%
# Of strongly connected components with symmetric links	4,798	3,254	11,252
Size of the largest strongly connected component with symmetric links	43	51	99

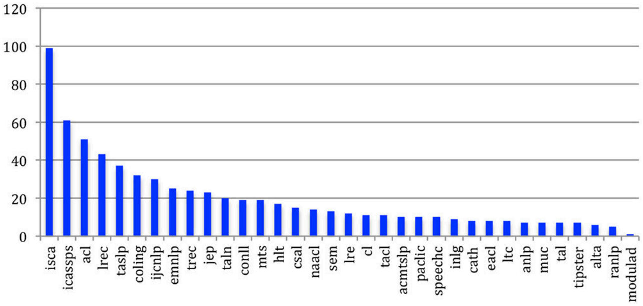


FIGURE 48 | Number of authors in the largest strongly connected component with symmetric links.

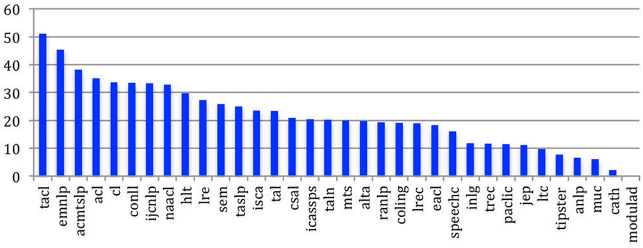


FIGURE 49 | Mean degree of authors citing authors in general for the 34 sources.

EMNLP, ACL, CL, CONLL, IJCNLP) has in general a larger habit of citation than the Speech one (TASLP, ISCA, CSAL, ICASSPS).

If we now consider the authors **being cited** in each of the 34 sources (**Figure 50**) through the CdAG, we see that authors who publish in *Computational Linguistics* are the most cited. It is followed by HLT and ACL, then EMNLP and NAACL. Speech conferences and journals show lower scores. This is in agreement also with the citation habits of the corresponding communities. Authors are obviously less cited for the papers they publish in languages other than English (e.g., JEP and *Modulad*).

Most cited authors

Table 10 gives the list of the 20 most cited authors, with the number of references for each author, and the number of papers written by the author. We see that this ratio may largely vary, some people having few papers but a large audience for this limited set of papers. We also provide the ratio of self-citation (citation of the author in a paper written by the author).

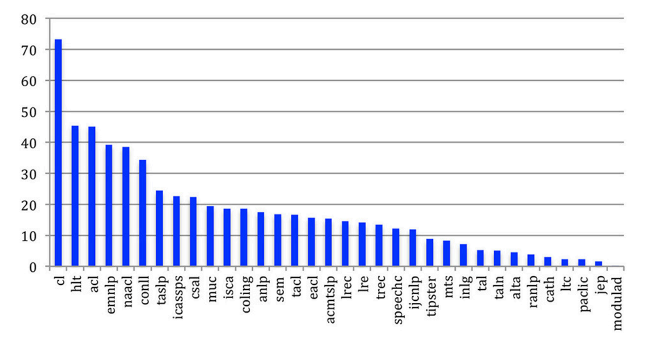


FIGURE 50 | Mean degree of authors being cited for the 34 sources.

We provide in **Table 11** the number of citations, either by themselves (self) or by others (extra), for the most productive authors already mentioned in **Table 3**. We notice that the most productive authors rather sign as last author.

Authors' h-index

We finally computed the h-index for each author. **Table 12** provides the list of the 20 authors with the largest h-index. We see that Christopher Manning has the largest h-index: he published 32 papers which were cited at least 32 times.

Papers Citations

Internal papers citations

Here also, we first consider *internal papers citations*: the citation in a source paper of papers published in the same source.

If we first consider the average number of papers being cited by papers of the same source for the 34 sources (**Figure 51**), we see that some communities, such as ACL and EMNLP, and the papers published in journals, such as TASLP or *Computational Linguistics* are used to cite each other, with an average of two papers from the same source or more being cited in each paper. Let's mention that, just as for authors, the Mean Degree of the internal Citing Papers Graph is equal to the Mean Degree of the internal Cited Papers Graph.

If we compare LREC, ACL and ISCA, we see that an LREC paper is internally cited less than once on average (0.9) in LREC papers, which is less than ACL (2.5) but comparable to ISCA (1.2).

TABLE 10 | Twenty most cited authors.

Name	# References	Nb of papers written by the author	Ratio # references/nb of papers written by the author	Percentage of self-citations
Hermann Ney	5,200	343	15.160	17.538
Franz Josef Och	4,098	42	97.571	2.221
Christopher D. Manning	3,972	116	34.241	5.060
Philipp Koehn	3,121	39	80.026	2.435
Dan Klein	3,080	99	31.111	7.532
Michael John Collins	3,077	53	58.057	3.640
Andreas Stolcke	3,053	130	23.485	7.141
Mark J. F. Gales	2,540	195	13.026	18.858
Salim Roukos	2,505	67	37.388	2.236
Chin-Hui P. Lee	2,450	218	11.239	18.245
Daniel Marcu	2,210	53	41.698	2.715
Philip Charles Woodland	2,154	145	14.855	14.624
Alejandro Acero	2,141	165	12.976	9.715
Vincent J. Della Pietra	2,138	16	133.625	0.655
Fernando C. N. Pereira	2,107	56	37.625	2.421
Li Deng	2,059	192	10.724	23.021
Robert L. Mercer	2,012	29	69.379	0.895
Daniel Jurafsky	1,995	86	23.198	3.609
Jean-Luc Gauvain	1,875	143	13.112	16.907
Keiichi Tokuda	1,864	133	14.015	18.509

TABLE 11 | Number of citations for the 20 most productive authors.

Number of written papers	Name	# As first author	% As first author	# As last author	% As last author	# As sole author	% As sole author	# Self-citations	Ratio of # self-citations/ number of written papers	# Extra-citations	Ratio of # extra-citations/ number of written papers
358	Shrikanth S. Narayanan	13	4	304	85	0	0	506	1.413	921	2.573
343	Hermann Ney	27	8	279	81	10	3	912	2.659	4,288	12.501
299	John H. L. Hansen	24	8	241	81	3	1	580	1.940	552	1.846
257	Haizhou Li	13	5	174	68	1	0	304	1.183	878	3.416
218	Chin-Hui P. Lee	13	6	167	77	5	2	447	2.050	2,003	9.188
207	Alex Waibel	13	6	175	85	2	1	175	0.845	1,183	5.715
205	Satoshi Nakamura	17	8	139	68	1	0	99	0.483	276	1.346
195	Mark J. F. Gales	32	16	87	45	9	5	479	2.456	2,061	10.569
193	Lin-Shan Lee	9	5	179	93	0	0	304	1.575	370	1.917
192	Li Deng	57	30	68	35	6	3	474	2.469	1,585	8.255
187	Keikichi Hirose	28	15	94	50	1	1	121	0.647	216	1.155
184	Kiyohiro Shikano	1	1	141	77	0	0	270	1.467	780	4.239
176	Mari Ostendorf	29	16	89	51	5	3	254	1.443	1,573	8.938
165	Alejandro Acero	12	7	121	73	3	2	208	1.261	1,933	11.715
161	Frank K. Soong	9	6	70	43	0	0	172	1.068	724	4.497
160	Hervé Bourlard	9	6	107	67	2	1	192	1.200	675	4.219
152	Tatsuya Kawahara	31	20	77	51	0	0	188	1.237	513	3.375
151	Douglas O'Shaughnessy	11	7	127	84	9	6	76	0.503	222	1.470
148	Sadaoki Furui	24	16	121	82	14	9	122	0.824	846	5.716
148	Yang Liu	33	22	67	45	3	2	179	1.209	781	5.277

TABLE 12 | List of the 20 authors with the largest h-index.

Name	H-index
Christopher D. Manning	32
Hermann Ney	29
Andreas Stolcke	28
Dan Klein	25
Michael John Collins	24
Alejandro Acero	23
Mari Ostendorf	23
Elizabeth E. Shriberg	23
Douglas A. Reynolds	23
Stephen J. Young	22
Franz Josef Och	22
Noah A. Smith	22
Daniel Jurafsky	22
Li Deng	22
Mirella Lapata	21
Keiichi Tokuda	21
Joakim Nivre	21
Jean-Luc Gauvain	21
Daniel Marcu	21
Philip Charles Woodland	21

The density reaches 0.00045 for Tipster, 0.00025 for MUC and 0.0015 for Semeval, which correspond to evaluation campaigns where there are many cross-citations among all the papers (**Figure 52**).

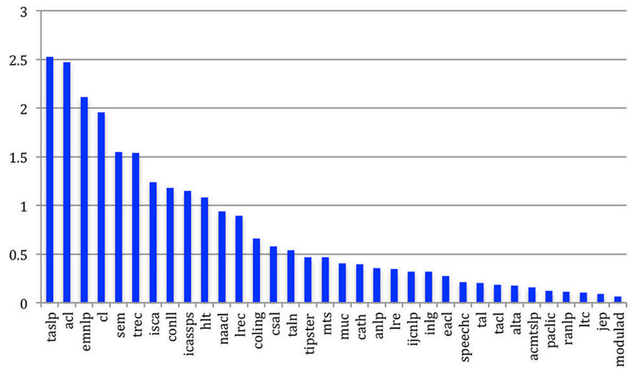


FIGURE 51 | Mean degree of citing and cited papers within the same source for the 34 sources.

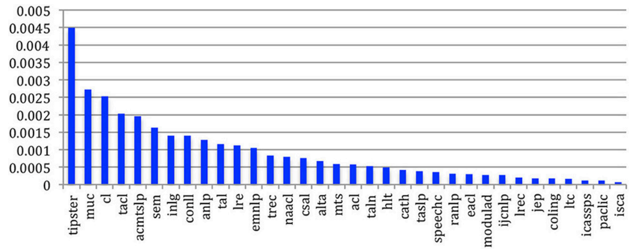


FIGURE 52 | Density of the internal papers citation graph.

Global papers citations

We now consider *global papers citations*: citation in papers published in each source of NLP4NLP papers in general.

If we now consider the general habit of **citing** other papers (**Figure 53**), we also see, just as when we considered the authors, that the NLP community (TACL, EMNLP, CL, CONLL, IJCNLP, NAACL, ACL) has in general a bigger habit of citation than the Speech one (CSAL, *Speech Communication*, TASLP, ICASSPS, ISCA). The average number of references in TACL papers is especially impressive (more than 18).

If we consider the papers **being cited** from each of the 34 sources (**Figure 54**), we see that papers published in *Computational Linguistics* are by far the most cited (more than 20 times on average). It is followed by NAACL, ACL and EMNLP, then HLT and CONLL, and is in agreement with the citing habits in those sources. Speech journals (CSAL, TASLP, *Speech Communication*) and especially speech conferences show lower scores. Papers are obviously less cited if they are published in languages other than English (e.g., TAL, TALN, JEP, *Modulad*).

If we compare LREC, ACL, and ISCA, we see that an LREC paper is cited 2.7 times on average, which is comparable to ISCA (2.5) but much less than ACL (10.4).

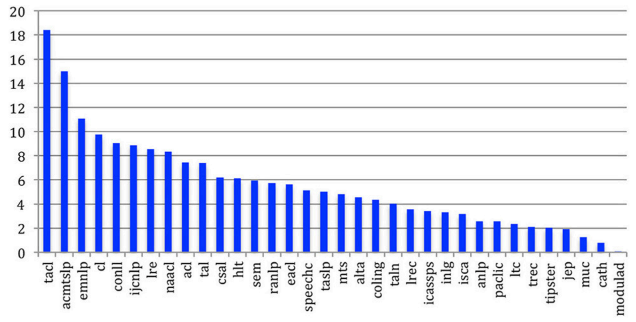


FIGURE 53 | Mean degree of papers citing papers in general for the 34 sources.

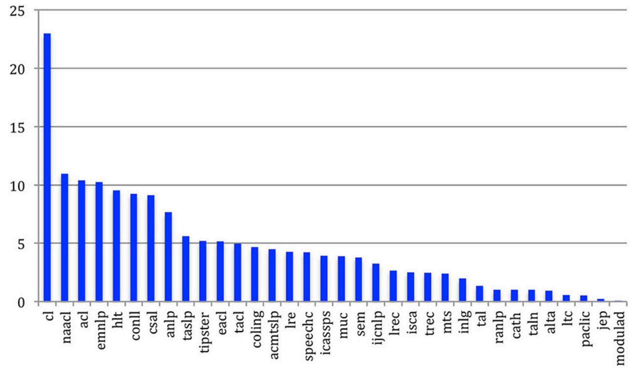


FIGURE 54 | Mean degree of papers being cited for the 34 sources.

Most cited papers

Table 13 gives the list of the 20 most cited papers. We see that the most cited papers are related to an evaluation metrics (Bleu), a Language Resource (Penn Treebank), a tool (Moses, SRILM) or a survey (Statistical alignment, Statistical translation). The largest number of papers comes from the *Computational Linguistics* journal (6), the ACL conference (4), and the IEEE *Transactions on Acoustics, Speech and Language* (3).

Among the 48,894 authors, 20,387 (42%) are never cited, and even 21,670 (44%) if we exclude self-citations (**Table 14**). However, after checking Google Scholar, it appears that many of those never cited authors come from neighboring research domains (machine learning, medical engineering, phonetics, general linguistics), where they may be largely cited. Among the 65,003 papers, 28,283 (44%) are never cited, and even 35,229 (54%) if we exclude self-citations.

Sources' h-index

Figure 55 gives the internal (papers being cited by papers of the same source) h-index for the 34 sources. The largest h-index is obtained by the IEEE TASLP, where 36 papers are cited in other IEEE TASLP papers 36 times or more. It is followed by ACL (34), ISCA (32), ICASSPS (27), EMNLP (22), and LREC (16).

If we now consider the general h-index (**Figure 56**) for the 34 sources, we see that the largest h-index is obtained by ACL, where 75 papers are cited 75 times or more in the NLP4NLP papers. It is followed by TASLP (66), *Computational Linguistics* (58), HLT (56), EMNLP (55), ICASSPS (54), and ISCA (51).

We also compared here LREC to ACL and ISCA. The internal h-index of LREC is 16: i.e., 16 papers published at LREC are cited 16 times or more in LREC papers (to be compared with 34 for ACL and 32 for ISCA). The h-index of LREC according to the NLP4NLP set of 34 conferences and journals is 36: i.e., 36 papers published at LREC are cited 36 times or more in NLP4NLP papers (75 for ACL and 51 for ISCA). However, it should be stressed once again that both ACL and ISCA conferences are annual and cover a much longer time period than LREC.

As of March 2016, Google Scholar[14] (**Table 15**) places ACL first in the ranking of computational linguistics conferences and journals with an h-index of 65 within the last 5 years (therefore on the same citation time period) and an h5-median mean of 99, followed by EMNLP (56), NAACL (48), LREC (38), COLING (38), CSAL (32), *Computational Linguistics* (31), CONLL (24), LRE (23), Semeval (23), EACL (21), and IJCNLP (20). In the Signal Processing category, we find IEEE ICASSP (54), IEEE TASLP (51), Interspeech (39), CSAL (32), and *Speech Communication* (32). Let's stress the point that this ranking covers the last 5 years and therefore reflects the recent trends compared with our own results, which concern a smaller number of sources and a closer scope but a larger time period. Therefore, the ranking may be different. For example, the new ISCA policy of opening the ISCA Archive to all, not only to members, has significantly increased the number of references to ISCA-Interspeech papers. Here also, LREC gets a lower h-index

[14] http://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computationallinguistics

TABLE 13 | Twenty most cited papers.

Title	Corpus	Year	Authors	# Citations
Bleu: a Method for Automatic Evaluation of Machine Translation	acl	2002	Kishore A. Papineni, Salim Roukos, Todd R. Ward, Wei-Jing Zhu	1,514
Building a Large Annotated Corpus of English: The Penn Treebank	cl	1993	Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz	1,146
Moses: Open Source Toolkit for Statistical Machine Translation	acl	2007	Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst	860
A Systematic Comparison of Various Statistical Alignment Models	cl	2003	Franz Josef Och, Hermann Ney	855
SRILM—an extensible language modeling toolkit	isca	2002	Andreas Stolcke	831
Statistical Phrase-Based Translation	hlt, naacl	2003	Philipp Koehn, Franz Josef Och, Daniel Marcu	829
The Mathematics of Statistical Machine Translation: Parameter Estimation	cl	1993	Peter E. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer	820
Minimum Error Rate Training in Statistical Machine Translation	acl	2003	Franz Josef Och	726
Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models	csal	1995	Chris Leggetter, Philip Charles Woodland	566
Suppression of acoustic noise in speech using spectral subtraction	taslp	1979	Steven F. Boll	566
Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains	taslp	1994	Jean-Luc Gauvain, Chin-Hui P. Lee	514
Accurate Unlexicalized Parsing	acl	2003	Dan Klein, Christopher D. Manning	513
Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator	taslp	1984	Yariv Ephraim, David Malah	488
Maximum likelihood linear transformations for HMM-based speech recognition	csal	1998	Mark J. F. Gales	483
Europarl: A Parallel Corpus for Statistical Machine Translation	mts	2005	Philipp Koehn	472
Head-Driven Statistical Models for Natural Language Parsing	cl	2003	Michael John Collins	470
Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms	emnlp	2002	Michael John Collins	465
A Maximum Entropy Approach to Natural Language Processing	cl	1996	Adam L. Berger, Vincent J. Della Pietra, Stephen A. Della Pietra	443
A Maximum-Entropy-Inspired Parser	naacl	2000	Eugene Charniak	437
Class-Based n-gram Models of Natural Language	cl	1992	Peter F. Brown, Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra, Jennifer C. Lai	432

TABLE 14 | Absence of citations of authors and papers within NLP4NLP.

	Number	%
Never cited articles (incl. self-citations)	28,283	44
Never cited articles (excl. self-citations)	35,229	54
Never cited authors (incl. self-citations)	20,387	42
Never cited authors (excl. self-citations)	21,670	44

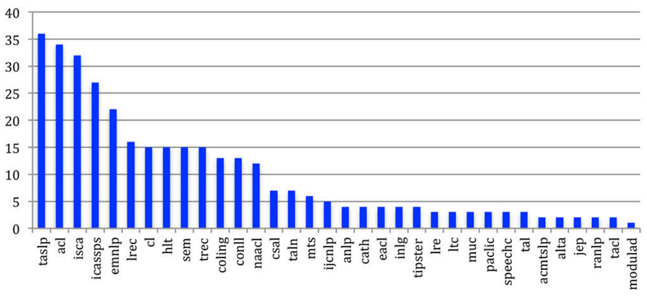


FIGURE 55 | Internal h-index of the 34 sources.

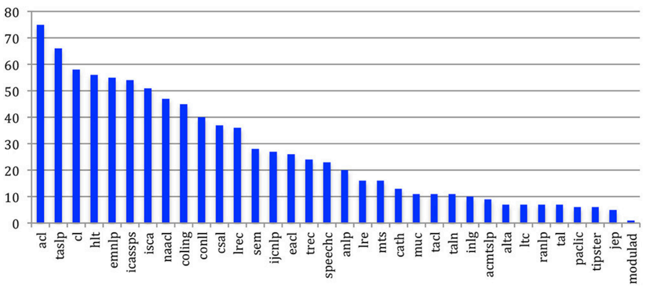


FIGURE 56 | General h-index of the 34 sources.

than ACL, but is similar to ISCA-Interspeech. It shows that the h-index reflects both the quality of a conference or journal, but also the number of papers that are published, which may therefore cite and be cited by other papers of the same conference or journal and also by other ones. The biennial conferences are under-scored with the h5-index as it takes into account either the two or the three previous conferences depending on the year, both in terms of possibly citing and cited papers. The h-index is

TABLE 15 | Ranking of 20 top sources according to Google Scholar h5-index over the 5 last years (2011–2015).

Rank	Source	h-5 index	h-5 Median
1	Meeting of the Association for Computational Linguistics (ACL)	65	99
2	Conference on Empirical Methods in Natural Language Processing (EMNLP)	56	81
3	IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)	54	73
4	IEEE Transactions on Audio, Speech, and Language Processing (TASLP)	51	78
5	North American Chapter of the Association for Computational Linguistics (NAACL)	48	71
6	International Conference on Spoken Language Processing (INTERSPEECH)	39	70
7	International Conference on Language Resources and Evaluation (LREC)	38	64
8	International Conference on Computational Linguistics (COLING)	38	59
9	arXiv Computation and Language (cs.CL)	37	70
10	Computer Speech & Language (CSL)	32	51
11	Speech Communication (SpeCom)	32	49
12	Computational Linguistics (CL)	31	40
13	Conference on Computational Natural Language Learning (CONLL)	24	36
14	Language Resources and Evaluation (LRE)	23	42
15	International Workshop on Semantic Evaluation (SEMEVAL)	23	41
16	Conference of the European Chapter of the Association for Computational Linguistics (EACL)	21	34
17	International Joint Conference on Natural Language Processing (IJCNLP)	20	27
18	IEEE Spoken Language Technology Workshop (SLT)	18	28
19	Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)	18	27
20	Workshop on Statistical Machine Translation	18	24

h5-index is the h-index for articles published in the last 5 complete years. It is the largest number h such that h articles published in 2010–2014 have at least h citations each. h5-median for a publication is the median number of citations for the articles that make up its h5-index.

a different measure of the quality of a conference or journal than the rejection rate, and in our opinion less biased, as it appears as an a-posteriori, not a-priori, quality evaluation. Interestingly, even if all submitted papers were accepted, it would not change the h-index, which only considers the most cited papers.

CONCLUSIONS

The production of the NLP4NLP corpus showed the importance of having an open access to data. In this analysis, we benefited from the fact that most of the source data are freely available on-line. Dealing with proprietary data needed a larger effort in communicating with the data owners, and raises the problems of distributing the data, replicating the results and updating the corpus.

The eldest data was not available in a text format and therefore had to be scanned, which introduced some errors. Additionally, we struggled with the lack of a consistent and uniform identification of entities (authors names, gender, affiliations, paper language, conference, and journal titles, funding agencies, etc.), which required a tedious manual correction process only made possible because we knew the main components of the field. In those conditions, it would have been impossible to conduct a comparable analysis on another research field unknown to us, with the same level of reliability. We already faced that problem when considering neighboring domains. Establishing standards for such domain-independent identification will demand an international effort in order to ensure that the identifiers are unique, which appears as a challenge for the scientific community.

PERSPECTIVES

We plan to produce an RDF version of the corpus and make the results available over the web as Linked Open Data. We would like to improve automatic information (names, references, terms) extraction by taking into account the context, in order to make the distinction between real and false occurrences of the information. It would avoid the tedious manual checking that we presently conduct and would improve the overall process.

In the next paper (Mariani et al., 2018), we will present an analysis of the evolution of the research topics, with the identification of the authors who introduced them and of the publication where they were first presented, and the detection of epistemological ruptures. Linking the metadata, the paper content and the references allowed us to propose a measure of innovation for the research topics, the authors and the publications. In addition, it allowed us to study the use of language resources, in the framework of the paradigm shift between knowledge-based approaches and content-based approaches, and the reuse of articles and plagiarism between sources over time.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication. JM launched the initiative following an invitation to give a keynote talk at Interspeech 2013 to celebrate the 25th anniversary of this major conference in spoken language processing and coordinated the following related and extended works from 2013 to 2018. GF produced the NLP4NLP corpus and developed all the tools that were used for analyzing the corpus. PP participated in the research group and provided advices on the use of NLP tools.

ACKNOWLEDGMENTS

The authors wish to thank the ACL colleagues, Ken Church, Sanjeev Khudanpur, Amjbad Abu Jbara, Dragomir Radev and Simone Teufel, who helped them in the starting phase, Isabel Trancoso, who gave her ISCA Archive analysis on the use of assessment and corpora, Wolfgang Hess, who produced and provided a 14 GBytes ISCA Archive, Emmanuelle Foxonet who provided a list of authors given names with genre, Florian Boudin, who made available the TALN Anthology, Helen van der Stelt, and Jolanda Voogd (Springer) who provided the LRE data and Douglas O'Shaughnessy, Denise Hurley, Rebecca Wollman, and Casey Schwartz (IEEE) who provided the IEEE ICASSP and TASLP data, Nancy Ide and Christopher Cieri who largely improved the readability of parts of this paper. They also thank Khalid Choukri, Alexandre Sicard, and Nicoletta Calzolari, who provided information about the past LREC conferences, Victoria Arranz, Ioanna Giannopoulou, Johann Gorlier, Jérémy Leixa, Valérie Mapelli, and Hélène Mazo, who helped in recovering the metadata for LREC 1998, and all the editors, organizers, reviewers, and authors over those 50 years without whom this analysis could not have been conducted!.

APOLOGIES

This survey has been made on textual data, which cover a 50-years period, including scanned content. The analysis uses tools that automatically process the content of the scientific papers and may make errors. Therefore, the results should be regarded as reflecting a large margin of error. The authors wish to apologize for any errors the reader may detect, and they will gladly rectify any such errors in future releases of the survey results.

RELATIONSHIP WITH OTHER PAPERS AND REUSE OF PREVIOUS MATERIAL

The present paper is accompanied by a second paper "Mariani, Joseph, Paroubek, Patrick, Francopoulo, Gil and Vernier, Frédéric (2018). The NLP4NLP Corpus (II): 50 Years of Research in Speech and Language Processing," in the same special issue of *Frontiers in Research Metrics and Analytics* on "Mining Scientific Papers: NLP-enhanced Bibliometrics" edited by Iana Atanassova, Marc Bertin and Philipp Mayr, which describes various analysis which were conducted on this corpus. A summary of the joint two papers has been presented as a keynote talk at the Oriental-Cocosda conference in Seoul ("Joseph Mariani, Gil Francopoulo, Patrick Paroubek, Frédéric Vernier, Rediscovering 50 Years of Discoveries in Speech and Language Processing: A Survey. Oriental Cocosda conference, Seoul, 1–3 November 2017") (Mariani et al., 2017).

This paper assembles the content of several former papers, which described various facets of the NLP4NLP corpus (http://www.nlp4nlp.org).

This corpus was first introduced in 2015 in two different conferences: "Francopoulo, Gil, Mariani, Joseph and Paroubek, Patrick (2015a). NLP4NLP: The Cobbler's Children Won't Go Unshod, 4th International Workshop on Mining Scientific Publications (WOSP2015), Joint Conference on Digital Libraries 2015 (JCDL 2015), Knoxville (USA), June 24, 2015." and "Francopoulo, Gil, Mariani, Joseph and Paroubek, Patrick (2015b). NLP4NLP: Applying NLP to written and spoken scientific NLP corpora, Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics, 15th International Society of Scientometrics and Informetrics Conference (ISSI 2015), Istanbul (Turkey), June 29, 2015."

Material from previously published sources, listed below, is re-used within permission, implicit or explicit open-license rights, as follows:

(1) "Mariani, Joseph, Paroubek, Patrick, Francopoulo, Gil and Hamon, Olivier (2014). Rediscovering 15 Years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis, LREC 2014, 26–31 May 2014, Reykjavik, Iceland", published within the Proceedings of LREC Conference 2014, http://www.lrec-conf.org/proceedings/lrec2014/index.html.
This paper analyzes the Language Resources and Evaluation Conference (LREC), which is one of the 34 publications contained in NLP4NLP, over 15 years (1998–2014).
The reused material concerns **Tables A1**, **A2**, **3**, **4**, **Figures 9–11**, section *Global Analysis of the Conferences and Journals* (mainly sub section *Manual Checking and Correction*).
(2) "Mariani, Joseph, Paroubek, Patrick, Francopoulo, Gil and Hamon, Olivier (2016). Rediscovering 15 + 2 Years of Discoveries in Language Resources and Evaluation, *Language Resources and Evaluation* Journal, 2016, pp. 1–56, ISSN: 1574-0218, doi: 10.1007/s10579-016-9352-9."
This paper has been selected among the LREC 2014 papers to be published in a special issue of the *Language Resources and Evaluation* Journal. It is an extended version of the previous paper, in the following dimensions: extension of the LREC content with the LREC 2014 conference itself (hence the change in the title of the paper: "15 + 2 Years" instead of "15 Years"), and comparison with two other conferences among those contained in NLP4NLP (namely ACL and Interspeech).
The reused material concerns section *Introduction* (mainly sub section *Preliminary Remarks*), section *Global Analysis of the Conferences and Journals* (mainly sub sections *Origin of Data, Extraction and Quality of Data*), section *Conclusions*, section *Perspectives* and subsection *Citation Graph*.

REFERENCES

Auber, D., Archambault, D., Bourqui, R., Lambert, A., Mathiaut, M., Mary, P., et al. (2012). *The Tulip 3 Framework: A Scalable Software Library for Information Visualization Applications Based on Relational Data*. Research Report, RR-7860. Available online at: http://hal.archives-ouvertes.fr/hal-00659880

Banchs, R. E. (2012). *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries Association for Computational Linguistics 2012 Jeju, Korea*. Available online at: https://aclanthology.coli.uni-saarland.de/papers/W12-3200/w12-3200

Bavelas, A. (1948). A mathematical model for small group structures. *Hum. Organ.* 7, 16–30.

Bavelas, A. (1950). Communication patterns in task oriented groups. *J. Acoust. Soc. Am.* 22, 271–282.

Bird, S., Dale, R., Dorr, B. J., Gibson, B., Joseph, M. T., Kan, M.-Y., et al. (2008). "The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics," in *Proceedings of LREC 2008* (Marrakesh), May 2008.

Bordea, G., Buitelaar, P., and Coughlan, B. (2014). "Hot topics and schisms in NLP: community and trend analysis with saffron on ACL and LREC proceedings," in *Proceedings of LREC 2014* (Reykjavik), May 2014.

Boudin, F. (2013). "TALN archives: une archive numérique francophone des articles de recherche en traitement automatique de la langue," in *TALN-RÉCITAL 2013* (Les Sables d'Olonne).

Councill, I. G., Giles, C., and Kan, M.-Y. (2008). "ParsCit: an open-source CRF reference string parsing package," in *Proceedings of LREC 2008* (Marrakesh).

Csárdi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal 2006. *Complex Syst.* 1695, 1–9. Available online at: http://igraph.org

Ding, Y., Rousseau, R., and Wolfram, D. (eds.). (2014). *Measuring Scholarly Impact*. Springer. doi: 10.1007/978-3-319-10377-8

Dunne, C., Shneiderman, B., Gove, R., Klavans, J., and Dorr, B. (2012). Rapid understanding of scientific paper collections: integrating statistics, text analytics, and visualization. *J. Am. Soc. Inf. Sci. Technol.* 63, 2351–2369. doi: 10.1002/asi.22652

Francopoulo, G. (2008). "TagParser: well on the way to ISO-TC37 conformance," in *ICGL (International Conference on Global Interoperability for Language Resources)* (Hong Kong).

Francopoulo, G., Marcoul, F., Causse, D., and Piparo, G. (2013). "Global atlas: proper nouns, from wikipedia to LMF," in *LMF-Lexical Markup Framework*, ed G. Francopoulo (ISTE/Wiley), 227–241.

Francopoulo, G., Mariani, J., and Paroubek, P. (2015a). "NLP4NLP: the cobbler's children won't go unshod," in *4th International Workshop on Mining Scientific Publications (WOSP2015), Joint Conference on Digital Libraries 2015 (JCDL 2015)* (Knoxville).

Francopoulo, G., Mariani, J., and Paroubek, P. (2015b). "NLP4NLP: applying NLP to written and spoken scientific NLP corpora," in *Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics, 15th International Society of Scientometrics and Informetrics Conference (ISSI 2015)* (Istanbul).

Francopoulo, G., Mariani, J., and Paroubek, P. (2016). "Text mining for notability computation," in *Cross-Platform Text Mining and Natural Language Processing Interoperability Workshop, LREC 2016, Tenth International Conference on Language Resources and Evaluation* (PortoroŽ).

Freeman, L. C. (1978). Centrality in social networks, conceptual clarifications. *Soc. Netw.* 1, 215–239. doi: 10.1016/0378-8733(78)90021-7

Fu, Y., Xu, F., and Uszkoreit, H. (2010). "Determining the origin and structure of person names," in *Proceedings of LREC 2010* (Valletta).

Fujisaki, H. (2013). *History of ICSP and PC-ICSLP, ISCA Web site – About ISCA – History*. Available online at: http://www.isca-speech.org/iscaweb/index.php/about-isca/history

Gollapalli, S. D., and Li, X.-L. (2015). "EMNLP versus ACL: analyzing NLP research over time," in *EMNLP 2015* (Lisbon), (September 17–21, 2015).

Jha, R., Jbara, A.-A., Qazvinian, V., and Radev, D. R. (2016). NLP-driven citation analysis for scientometrics. *Nat. Lang. Eng.* 23, 93–130. doi: 10.1017/S1351324915000443

Joerg, B., Höllrigl, T., and Sicilia, M.-A. (2012). "Entities and identities in research information systems," in *11th International Conference on Current Research Information Systems (CRIS2012): "e-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production"* (Prague).

Li, H., Councill, I. G., Lee, W. C., and Giles, C. (2006). "CiteSeerx: an architecture and web service design for an academic document search engine," in *Proceedings of the 15th Int. Conference on the World Wide Web* (Edinburgh). (May 23–26, 2006).

Litchfield, B. (2005). *Making PDFs Portable: Integrating PDF and Java Technology*, March 24, 2005. Java Developers Journal. Available online at: http://java.sys-con.com/node/48543 (PDFBox is available at: http://pdfbox.apache.org/).

Mariani, J. (1990). *La Conférence IEEE-ICASSP de 1976 à 1990: 15 ans de recherches en Traitement Automatique de la Parole*, Notes et Documents LIMSI 90-8.

Mariani, J. (2013). *The ESCA Enterprise, ISCA Web site – About ISCA – History*. Available online at: http://www.isca-speech.org/iscaweb/index.php/about-isca/history

Mariani, J., Cieri, C., Francopoulo, G., Paroubek, P., and Delaborde, M. (2014b). "Facing the identification problem in language-related scientific data analysis," in *Proceedings of LREC 2014* (Reykjavik).

Mariani, J., Francopoulo, G., Paroubek, P., and Vernier, F. (2017). "Rediscovering 50 years of discoveries in speech and language processing: a survey," in *Oriental Cocosda Conference* (Seoul: IEEE XPlore).

Mariani, J., Francopoulo, G., Paroubek, P., and Vetulani, Z. (2015). "Rediscovering 10 to 20 years of discoveries in language & technology," in *Proceedings of L&TC 2015* (Poznan).

Mariani, J., Francopoulo, G., Paroubek, P., and Vernier, F. (2018). The NLP4NLP Corpus (II): 50 Years of Research in Speech and Language Processing. *Front. Res. Metr. Anal.* 3:37. doi: 10.3389/frma.2018.00037

Mariani, J., Paroubek, P., Francopoulo, G., and Delaborde, M. (2013). "Rediscovering 25 years of discoveries in spoken language processing: a preliminary ISCA archive analysis," in *Proceedings of Interspeech 2013* (Lyon).

Mariani, J., Paroubek, P., Francopoulo, G., and Hamon, O. (2014a). "Rediscovering 15 years of discoveries in language resources and evaluation: the LREC anthology analysis," in *Proceedings of LREC 2014* (Reykjavik).

Mariani, J., Paroubek, P., Francopoulo, G., and Hamon, O. (2016). Rediscovering 15 + 2 years of discoveries in language resources and evaluation. *Lang. Resour. Eval. J.* 50, 1–56. doi: 10.1007/s10579-016-9352-9

Osborne, F., Motta, E., and Mulholland, P. (2013). "Exploring scholarly data with rexplore," in *International Semantic Web Conference* (Sydney, NSW).

Radev, D. R., Muthukrishnan, P., Qazvinian, V., and Abu-Jbara, A. (2013). The ACL anthology network corpus. *Lang. Resour. Eval.* 47, 919–944. doi: 10.1007/s10579-012-9211-2

Rochat, Y. (2009). "Closeness centrality extended to unconnected graphs: the harmonic centrality index," in *Applications of Social Network Analysis (ASNA)* (Zurich).

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008). "ArnetMiner: extraction and mining of academic social networks," in *Proceeding of the 14th Int. Conference on Knowledge Discovery and Data Mining* (Las Vegas, NV) (August 24-27, 2008)..

The R Journal (2012). Available online at: http://journal.r-project.org/

Vogel, A., and Jurafsky, D. (2012). "He said, she said: gender in the ACL anthology," in *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries (ACL'12)* (Stroudsburg, PA: Association for Computational Linguistics), 33–41.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

TABLE A1 | List of sources with number of papers.

Year	# Sources	Cumulated # sources	Sources	# Documents	Cumulated # documents
1965	1	1	**coling**	24	24
1966	1	2	**cath**	7	31
1967	2	2	cath coling	54	85
1968	1	2	cath	17	102
1969	1	2	cath	24	126
1970	1	2	cath	18	144
1971	1	2	cath	20	164
1972	1	2	cath	19	183
1973	2	2	cath coling	80	263
1974	1	2	cath	25	288
1975	2	3	cath **taslp**	131	419
1976	2	3	cath taslp	136	555
1977	2	3	cath taslp	141	696
1978	2	3	cath taslp	155	851
1979	3	4	**acl** cath taslp	179	1,030
1980	5	5	acl cath **cl** coling taslp	307	1,337
1981	4	5	acl cath cl taslp	274	1,611
1982	6	6	acl cath cl coling **speechc** taslp	364	1,975
1983	7	8	acl **anlp** cath cl **eacl** speechc taslp	352	2,327
1984	5	8	acl cath cl speechc taslp	353	2,680
1985	6	8	acl cath cl eacl speechc taslp	384	3,064
1986	8	10	acl cath cl coling **csal hlt** speechc taslp	518	3,582
1987	9	12	acl cath cl csal eacl **isca mts** speechc taslp	669	4,251
1988	8	13	acl anlp cath cl coling **modulad** speechc taslp	546	4,797
1989	11	13	acl cath cl csal eacl hlt isca modulad mts speechc taslp	965	5,762
1990	11	14	acl cath cl coling csal hlt **icassps** isca modulad speechc taslp	1,277	7,039
1991	13	15	acl cath cl csal eacl hlt icassps isca modulad mts **muc** speechc taslp	1,378	8,417
1992	14	16	acl anlp cath cl coling csal hlt icassps isca modulad muc speechc taslp **trec**	1,611	10,028
1993	15	17	acl cath cl csal eacl hlt icassps isca modulad mts muc speechc taslp **tipster** trec	1,239	11,267

(Continued)

TABLE A1 | Continued

Year	# Sources	Cumulated # sources	Sources	# Documents	Cumulated # documents
1994	13	17	acl anlp cath cl coling csal hlt icassps isca modulad speechc taslp trec	1,454	12,721
1995	15	19	acl cath cl csal eacl icassps isca **ltc** modulad mts muc **paclic** speechc taslp trec	1,209	13,930
1996	15	21	acl cath cl coling csal **emnlp** icassps **inlg** isca modulad paclic speechc taslp tipster trec	1,536	15,466
1997	15	23	acl anlp cath cl **conll** csal emnlp icassps isca modulad mts speechc **taln** taslp trec	1,530	16,996
1998	16	24	acl cath cl csal emnlp icassps isca **lrec** modulad muc paclic speechc taln taslp tipster trec	1,953	18,949
1999	16	24	acl cath cl conll csal eacl emnlp icassps isca modulad mts paclic speechc taln taslp trec	1,603	20,552
2000	18	24	acl anlp cath cl coling conll csal emnlp icassps inlg isca lrec modulad naacl paclic speechc taln taslp trec	2,271	22,823
2001	18	26	acl cath cl conll csal emnlp hlt icassps isca modulad mts **naacl** paclic **sem** speechc taln taslp trec	1,644	24,467
2002	17	27	acl cath cl coling conll csal emnlp icassps isca **jep** lrec modulad paclic speechc taln taslp trec	2,174	26,641
2003	17	28	acl **alta** cath cl conll csal eacl emnlp hlt icassps isca modulad mts paclic speechc taln taslp trec	1,984	28,625
2004	21	29	acl **acmtslp** alta cath cl coling conll csal emnlp hlt icassps isca jep lrec modulad paclic sem speechc taln taslp trec	2,712	31,337
2005	20	30	acl acmtslp alta cl conll csal emnlp icassps **ijcnlp** isca lre ltc modulad mts paclic speechc taln taslp trec	2,355	33,692
2006	22	32	acl acmtslp alta cl conll csal eacl emnlp hlt icassps inlg isca **lre** lrec modulad paclic speechc **tal** taln taslp trec	2,794	36,486
2007	20	32	acl acmtslp alta cl conll csal hlt icassps isca lre ltc modulad mts paclic sem speechc tal taln taslp trec	2,489	38,975
2008	23	32	acl acmtslp alta cl coling conll csal emnlp icassps ijcnlp inlg isca jep lre lrec modulad paclic speechc tal taln taslp trec	3,078	42,053
2009	23	33	acl acmtslp alta cl conll csal eacl emnlp hlt icassps isca lre ltc modulad mts paclic **ranlp** speechc tal taln taslp trec	2,637	44,690
2010	22	33	acl acmtslp alta cl coling conll csal emnlp hlt icassps inlg isca lre lrec modulad paclic sem speechc tal taln taslp trec	3,470	48,160
2011	20	33	acl acmtslp alta cl conll csal emnlp icassps ijcnlp isca lre ltc mts paclic ranlp speechc tal taln taslp trec	2,957	51,117
2012	22	33	acl acmtslp alta cl coling conll csal eacl hlt icassps inlg isca jep lre lrec paclic sem speechc tal taln taslp trec	3,419	54,536
2013	23	34	acl acmtslp alta cl conll csal emnlp hlt icassps ijcnlp isca lre ltc mts paclic ranlp sem speechc **tacl** tal taln taslp trec	3,336	57,872
2014	22	34	acl alta cl coling conll csal eacl emnlp icassps inlg isca jep lre lrec paclic sem speechc tacl tal taln taslp trec	3,817	61,689
2015	14	34	acl conll csal emnlp hlt icassps isca lre ltc mts sem speechc tacl tal taln taslp trec	3,314	65,003

Sources are marked in bold characters on the year they are considered for the first time in NLP4NLP.

TABLE A2 | Quantity and quality of data.

Year	# Papers from metadata	# Papers in PDF	# Papers in XML (= output PDFBox)	# Non-empty papers as extraction result	# Papers with an abstract (from extraction)	# Papers with references (from extraction)	# Unknown words	# Known words	# Words of the content	Evaluation of noise = % (known words/words of the content)	Evaluation of silence = % non-empty papers as extraction result/PDF docs	Combined evaluation of noise and silence	# English papers	# French papers
1965	24	24	21	20	8	11	4,743	99,196	103,939	95.437	83.333	88.975	20	0
1966	7	7	7	7	0	1	321	22,967	23,288	98.622	100.000	99.306	7	0
1967	54	54	40	39	0	11	5,681	127,686	133,367	95.740	72.222	82.335	39	0
1968	17	17	15	15	0	2	800	47,771	48,571	98.353	88.235	93.020	15	0
1969	24	24	24	24	1	3	1,024	80,549	81,573	98.745	100.000	99.368	24	0
1970	18	18	18	17	0	8	1,527	57,516	59,043	97.414	94.444	95.906	17	0
1971	20	20	20	20	0	9	1,637	67,172	68,809	97.621	100.000	98.796	20	0
1972	19	19	19	19	0	10	1,913	83,060	84,973	97.749	100.000	98.862	19	0
1973	80	80	68	64	0	23	9,304	249,180	258,484	96.401	80.000	87.438	64	0
1974	25	25	23	23	0	11	2,719	120,591	123,310	97.795	92.000	94.809	23	0
1975	131	131	130	121	28	91	6,493	376,123	382,616	98.303	92.366	95.242	121	0
1976	136	136	135	128	18	104	7,370	430,888	438,258	98.318	94.118	96.172	128	0
1977	141	141	141	130	26	104	8,515	429,818	438,333	98.057	92.199	95.038	130	0
1978	155	155	152	142	44	104	10,395	495,240	505,635	97.944	91.613	94.673	142	0
1979	179	179	175	168	44	127	14,484	594,051	608,535	97.620	93.855	95.700	168	0
1980	307	307	296	287	67	230	36,304	1,159,270	1,195,574	96.963	93.485	95.193	287	0
1981	274	274	273	251	67	201	27,343	1,066,050	1,093,393	97.499	91.606	94.461	251	0
1982	364	364	341	326	58	230	32,237	1,096,602	1,128,839	97.144	89.560	93.198	326	0
1983	352	352	346	333	106	271	40,954	1,493,712	1,534,666	97.331	94.602	95.947	333	0
1984	469	469	458	334	119	250	37,397	1,459,874	1,497,271	97.502	71.215	82.311	334	0
1985	384	384	373	356	143	297	47,222	1,741,094	1,788,316	97.359	92.708	94.977	356	0
1986	518	518	487	476	226	390	69,861	2,414,873	2,484,734	97.188	91.892	94.466	476	0
1987	669	669	665	652	372	561	52,137	2,520,348	2,572,485	97.973	97.459	97.715	652	0
1988	546	546	515	507	252	436	73,909	2,790,102	2,864,011	97.419	92.857	95.084	503	4
1989	965	965	965	925	517	783	76,521	3,858,131	3,934,652	98.055	95.855	96.943	916	9
1990	1,277	1,277	1,257	1,235	837	877	94,150	4,347,841	4,441,991	97.880	96.711	97.292	1,232	3
1991	1,378	1,378	1,365	1,330	900	927	99,674	4,697,932	4,797,606	97.922	96.517	97.214	1,323	7
1992	1,611	1,611	1,568	1,550	1,034	1,142	153,962	6,079,721	6,233,683	97.530	96.214	96.867	1,545	5
1993	1,239	1,239	1,237	1,232	913	873	83,071	3,942,426	4,025,497	97.936	99.435	98.680	1,222	10
1994	1,454	1,454	1,369	1,360	1,036	975	110,706	4,372,238	4,482,944	97.531	93.535	95.491	1,356	4
1995	1,209	1,209	1,203	1,200	913	1,061	83,192	4,203,754	4,286,946	98.059	99.256	98.654	1,195	5

(Continued)

TABLE A2 | Continued

Year	# Papers from metadata	# Papers in PDF	# Papers in XML (= output PDFBox)	# Non-empty papers as extraction result	# Papers with an abstract (from extraction)	# Papers with references (from extraction)	# Unknown words	# Known words	# Words of the content	Evaluation of noise = % (known words/words of the content)	Evaluation of silence = % non-empty papers as extraction result/PDF docs	Combined evaluation of noise and silence	# English papers	# French papers
1996	1,536	1,536	1,502	1,497	1,122	1,390	131,537	5,282,115	5,413,652	97.570	97.461	97.516	1,493	4
1997	1,603	1,603	1,591	1,516	1,242	1,366	93,589	5,016,182	5,109,771	98.168	94.573	96.337	1,504	12
1998	2,198	2,198	2,168	1,924	1,525	1,797	138,419	6,845,119	6,983,538	98.018	87.534	92.480	1,909	15
1999	1,603	1,603	1,558	1,544	1,297	1,455	105,646	5,651,838	5,757,484	98.165	96.319	97.233	1,502	42
2000	2,271	2,270	2,108	2,067	1,789	1,946	139,268	7,525,857	7,665,125	98.183	91.057	94.486	2,052	15
2001	1,644	1,644	1,536	1,503	1,341	1,395	95,264	5,597,958	5,693,222	98.327	91.423	94.749	1,469	34
2002	2,174	2,174	2,070	2,041	1,838	1,887	188,194	7,854,498	8,042,692	97.660	93.882	95.734	1,909	132
2003	2,059	2,059	2,029	1,950	1,765	1,825	180,750	7,463,774	7,644,524	97.636	94.706	96.149	1,907	43
2004	2,794	2,794	2,736	2,640	2,435	2,515	171,027	9,744,993	9,916,020	98.275	94.488	96.345	2,468	172
2005	2,482	2,482	2,457	2,324	2,080	2,180	225,916	9,533,092	9,759,008	97.685	93.634	95.617	2,257	67
2006	3,179	3,179	3,162	2,761	2,585	2,672	237,231	11,872,296	12,109,527	98.041	86.851	92.107	2,669	92
2007	2,747	2,747	2,726	2,443	2,292	2,396	195,425	10,767,056	10,962,481	98.217	88.933	93.345	2,356	87
2008	3,265	3,265	3,251	3,058	2,883	2,986	239,887	12,952,216	13,192,103	98.182	93.660	95.868	2,869	189
2009	2,997	2,997	2,988	2,616	2,482	2,514	210,649	11,652,127	11,862,776	98.224	87.287	92.433	2,512	104
2010	3,616	3,616	3,607	3,444	3,247	3,388	293,256	15,603,676	15,896,932	98.155	95.243	96.677	3,351	93
2011	2,957	2,957	2,951	2,938	2,825	2,908	263,107	13,852,593	14,115,700	98.136	99.357	98.743	2,843	95
2012	3,655	3,655	3,645	3,396	3,263	3,307	309,524	16,017,843	16,327,367	98.104	92.914	95.439	3,234	162
2013	3,476	3,476	3,475	3,311	3,168	3,249	329,259	16,264,021	16,593,280	98.016	95.253	96.615	3,250	61
2014	3,817	3,817	3,811	3,803	3,599	3,708	393,472	18,602,173	18,995,645	97.929	99.633	98.774	3,649	154
2015	3,818	3,818	3,806	3,290	3,144	3,197	344,629	15,432,402	15,777,031	97.816	86.171	91.625	3,214	76
Total	67,937	67,936	66,883	63,357	53,651	58,204	5,481,615	264,057,605	269,539,220	97.966	93.260	95.555	61,661	1,696

TABLE A3 | Authors' renewal and redundancy.

Year	# Papers	# Authorships	# Authorships/ Paper	# Papers written alone	% Papers written alone	# Different Authors	Author redundancy (%)	# New authors	% New authors	# Completely new authors	% Completely new authors
1965	24	32	1.333	18	75	32	0	32	100	32	100
1966	7	8	1.143	6	86	7	13	7	100	7	100
1967	54	71	1.315	42	78	68	4	67	99	59	87
1968	17	17	1.000	17	100	16	6	16	100	16	100
1969	24	25	1.042	23	96	25	0	24	96	22	88
1970	18	20	1.111	17	94	19	5	18	95	15	79
1971	20	25	1.250	16	80	24	4	21	88	18	75
1972	19	21	1.105	17	89	21	0	19	91	17	81
1973	80	115	1.438	55	69	111	3	109	98	105	95
1974	25	29	1.160	21	84	28	3	27	96	25	89
1975	131	208	1.588	75	57	172	17	170	99	166	97
1976	136	233	1.713	64	47	188	19	154	82	151	80
1977	141	230	1.631	75	53	206	10	161	78	148	72
1978	155	249	1.606	82	53	217	13	172	79	146	67
1979	179	307	1.715	91	51	272	11	233	86	187	69
1980	307	502	1.635	178	58	450	10	387	86	334	74
1981	274	449	1.639	146	53	375	16	303	81	237	63
1982	364	606	1.665	194	53	541	11	467	86	348	64
1983	352	662	1.881	154	44	578	13	463	80	350	61
1984	353	582	1.649	182	52	507	13	391	77	287	57
1985	384	657	1.711	186	48	558	15	438	79	323	58
1986	518	973	1.878	215	42	819	16	677	83	498	61
1987	669	1,380	2.063	242	36	1,144	17	959	84	792	69
1988	546	1,028	1.883	223	41	896	13	703	79	501	56
1989	965	2,010	2.083	357	37	1,517	25	1,319	87	935	62
1990	1,277	2,916	2.283	396	31	2,105	28	1,614	77	1,255	60
1991	1,378	3,070	2.228	420	30	2,146	30	1,449	68	1,052	49
1992	1,611	3,777	2.345	465	29	2,661	30	1,872	70	1,301	49
1993	1,239	3,056	2.467	357	29	2,048	33	1,290	63	873	43
1994	1,454	3,650	2.510	367	25	2,512	31	1,753	70	1,118	45
1995	1,209	2,952	2.442	324	27	2,192	26	1,473	67	970	44
1996	1,536	3,818	2.486	379	25	2,697	29	1,849	69	1,146	43
1997	1,530	3,993	2.610	291	19	2,814	30	1,799	64	1,125	40

(Continued)

TABLE A3 | Continued

Year	# Papers	# Authorships	# Authorships/ Paper	# Papers written alone	% Papers written alone	# Different Authors	Author redundancy (%)	# New authors	% New authors	# Completely new authors	% Completely new authors
1998	1,953	5,236	2.681	398	20	3,472	34	2,325	67	1,453	42
1999	1,603	4,294	2.679	322	20	3,056	29	1,824	60	1,160	38
2000	2,271	6,185	2.723	420	18	3,898	37	2,644	68	1,579	41
2001	1,644	4,740	2.883	261	16	3,283	31	1,892	58	1,191	36
2002	2,174	6,048	2.782	385	18	4,284	29	2,959	69	1,889	44
2003	1,984	5,661	2.853	293	15	3,779	33	2,356	62	1,452	38
2004	2,712	7,954	2.933	414	15	4,996	37	3,276	66	2,006	40
2005	2,355	6,749	2.866	342	15	4,524	33	2,670	59	1,698	38
2006	2,794	8,347	2.987	323	12	5,343	36	3,403	64	2,020	38
2007	2,489	7,334	2.947	279	11	4,832	34	2,744	57	1,668	35
2008	3,078	9,467	3.076	285	9	5,791	39	3,730	64	2,090	36
2009	2,637	7,990	3.030	269	10	4,988	38	2,809	56	1,662	33
2010	3,470	10,761	3.101	310	9	6,364	41	3,968	62	2,242	35
2011	2,957	9,224	3.119	210	7	5,555	40	3,059	55	1,820	33
2012	3,419	11,077	3.240	253	7	6,612	40	4,121	62	2,310	35
2013	3,336	10,930	3.276	245	7	6,485	41	3,780	58	2,269	35
2014	3,817	12,925	3.386	261	7	7,700	40	4,750	62	2,793	36
2015	3,314	11,457	3.457	158	5	7,181	37	4,404	61	3,033	42
Total	65,003	184,050	11,123	2,831						48,894	

frontiers
in Research Metrics and Analytics

ORIGINAL RESEARCH
published: 07 February 2019
doi: 10.3389/frma.2018.00037

The NLP4NLP Corpus (II): 50 Years of Research in Speech and Language Processing

Joseph Mariani[1]*, *Gil Francopoulo*[2], *Patrick Paroubek*[1] *and Frédéric Vernier*[1]

[1] LIMSI-CNRS, Université Paris-Saclay, Orsay, France, [2] Tagmatica, Paris, France

The NLP4NLP corpus contains articles published in 34 major conferences and journals in the field of speech and natural language processing over a period of 50 years (1965–2015), comprising 65,000 documents, gathering 50,000 authors, including 325,000 references and representing ~270 million words. This paper presents an analysis of this corpus regarding the evolution of the research topics, with the identification of the authors who introduced them and of the publication where they were first presented, and the detection of epistemological ruptures. Linking the metadata, the paper content and the references allowed us to propose a measure of innovation for the research topics, the authors and the publications. In addition, it allowed us to study the use of language resources, in the framework of the paradigm shift between knowledge-based approaches and content-based approaches, and the reuse of articles and plagiarism between sources over time. Numerous manual corrections were necessary, which demonstrated the importance of establishing standards for uniquely identifying authors, articles, resources or publications.

Keywords: speech processing, natural language processing, text analytics, bibliometrics, scientometrics, informetrics

OPEN ACCESS

Edited by:
Iana Atanassova,
Université Bourgogne
Franche-Comté, France

Reviewed by:
Qi Yu,
Shanxi Medical University, China
Mariannig Le Béchec,
Claude Bernard University Lyon 1,
France

***Correspondence:**
Joseph Mariani
joseph.mariani@limsi.fr

Received: *30 January 2018*
Accepted: *07 December 2018*
Published: *07 February 2019*

Citation:
Mariani J, Francopoulo G, Paroubek P and Vernier F (2019) The NLP4NLP Corpus (II): 50 Years of Research in Speech and Language Processing. Front. Res. Metr. Anal. 3:37. doi: 10.3389/frma.2018.00037

This work is composed of two parts, of which this is part II. Please read also part I (Mariani et al., 2018b).

INTRODUCTION

Preliminary Remarks

The aim of this study was to investigate a specific research area, namely Natural Language Processing (NLP), through the related scientific publications, with a large amount of data and a set of tools, and to report various findings resulting from those investigations. The study was initiated by an invitation of the Interspeech 2013 conference organizers to look back at the conference content on the occasion of its 25th anniversary. It was then followed by similar invitations at other conferences, by adding new types of analyses and finally by extending the data to many conferences and journals over a long time period. We would like to provide elements that may help answering questions such as: What are the most innovative conferences and journals? What are the most pioneering and influential ones? How large is their scope? How are structured the corresponding communities? What is the effect of the language of a publication? Which paradigms appeared and disappeared over time? Were there any epistemological ruptures? Is there a way to identify weak signals of an emerging research trend? Can we guess what will come next? What were the merits of authors in terms of paper production and citation, collaboration activities and innovation?

What is the use of Language Resources in research? Do authors plagiarize each other? Do they publish similar papers in the same or in different conferences and journals? The results of this study are presented in two companion papers. The former one (Mariani et al., 2018b) introduces the corpus with various analyses: evolution over time of the number of papers and authors, including their distribution by gender, as well as collaboration among authors and citation patterns among authors and papers. In the present paper, we will consider the evolution of research topics over time and identify the authors who introduced and mainly contributed to key innovative topics, the use of Language Resources over time and the reuse of papers and plagiarism within and across publications. We provide both global figures corresponding to the whole data and comparisons of the various conferences and journals among those various dimensions. The study uses NLP methods that have been published in the corpus considered in the study, hence the name of the corpus. In addition to providing a revealing characterization of the speech and language processing community, the study also demonstrates the need for establishing a framework for unique identification of authors, papers and sources in order to facilitate this type of analysis, which presently requires a heavy manual checking.

The NLP4NLP Corpus

In the previous paper (Mariani et al., 2018b), we introduced the NLP4NLP corpus. This corpus contains articles published in 34 major conferences and journals in the field of speech and natural language processing over a period of 50 years (1965–2015), comprising 65,000 documents, gathering 50,000 authors, including 325,000 references and representing ∼270 million words. Most of these publications are in English, some are in French, German or Russian. Some are open access, others have been provided by the publishers.

This paper establishes the link between the different types of information that were introduced in the previous paper and that are contained in NLP4NLP. It presents an analysis of the evolution of the research topics with the identification of the authors who introduced them and of the publication where they were first presented and the detection of epistemological ruptures. Linking the metadata, the paper content and the references allowed us to propose a measure of innovation for the research topics, the authors and the publications. In addition, it allowed us to study the use of language resources, in the framework of the paradigm shift between knowledge-based approaches and content-based approaches, and the reuse of articles and plagiarism between sources over time. Numerous manual corrections were necessary, which demonstrated the importance of establishing standards for uniquely identifying authors, articles, resources or publications.

ANALYSIS OF THE NLP4NLP CORPUS

Topics

Archive Analysis

Modeling the topics of a research field is a challenge in NLP (see e.g., Hall et al., 2008; Paul and Girju, 2009). Here, our objectives were two-fold: (i) to compute the most frequent terms used in the domain, (ii) to study their variation over time. Like the study of citations, our initial input is the textual content of the papers available in a digital format or that had been scanned. Over these 50 years, the archives contain a grand total of 269,539,220 words, mostly in English.

Because our aim is to study the terms of the NLP domain, it was necessary to avoid noise from phrases that are used in other senses in the English language. We therefore adopted a contrastive approach, using the same strategy implemented in TermoStat (Drouin, 2004). For this purpose, as a first step, we processed a vast number of English texts that were not research papers in order to compute a statistical language profile. To accomplish this, we applied a deep syntactic parser called TagParser[1] to produce the noun phrases in each text. For each sentence, we kept only the noun phrases with a regular noun as a head, thus excluding the situations where a pronoun, date, or number is the head. We retained the various combinations of sequence of adjectives, prepositions and nouns excluding initial determiners using unigrams, bigrams and trigrams sequences and stored the resulting statistical language model. This process was applied on a corpus containing the British National Corpus (aka BNC)[2], the Open American National Corpus (aka OANC[3]) (Ide et al., 2010), the Suzanne corpus release-5[4], the English EuroParl archives (Koehn, 2005) (years 1999 until 2009)[5], plus a small collection of newspapers in the domain of sports, politics and economy, comprising a total of 200 M words. It should be noted that, in selecting this corpus, we took care to avoid any texts dealing with NLP.

Terms Frequency and Presence

In a second step, we parsed the NLP4NLP corpus with the same filters and used our language model to distinguish SNLP-specific terms from common ones. We worked from the hypothesis that when a sequence of words is *inside* the NLP4NLP corpus and *not inside* the general language profile, the term is specific to the field of SNLP. The 67,937 documents reduce to 61,661 documents when considering only the papers written in English. They include 3,314,671 different terms (unigrams, bigrams and trigrams) and 23,802,889 term occurrences, provided that this number counts all the occurrences of all the sizes and does not restrict to the longest terms, thus counting a great number of overlapping situations between fragments of texts.

The 500 most frequent terms in the field of SNLP were computed over the period of 50 years, according to the following strategy. First, the most frequent terms were computed in a raw manner, and secondly the synonyms sets (aka synsets) for all most 200 frequent terms of each year (which are frequently the same from 1 year to another) were manually declared in the lexicon of TagParser. Around the term synset, we gathered the variation in upper/lower case, singular/plural

[1] www.tagmatica.com

[2] Version 3 (BNC XML Edition), 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

[3] http://www.anc.org/

[4] www.grsampson.net/Resources.html

[5] www.statmt.org/europarl

TABLE 1 | Twenty most frequent terms overall, with number of occurrences and existences, frequency and presence.

Rank	Term	Variants of all sorts	# Occurrences	Frequency	# existences	Presence	Occurrences/ existences
1	HMM	HMMs, Hidden Markov Model, Hidden Markov Models, Hidden Markov model, Hidden Markov models, hidden Markov Model, hidden Markov Models, hidden Markov model, hidden Markov models	134,060	0.00609	14353	0.22671	9.34
2	SR	ASR, ASRs, Automatic Speech Recognition, SRs, Speech Recognition, automatic speech recognition, speech recognition	128,590	0.00584	20324	0.32102	6.33
3	LM	LMs, Language Model, Language Models, language model, language models	111,582	0.00507	12809	0.20232	8.71
4	Annotation	Annotations	111,142	0.00505	11992	0.18942	9.27
5	POS	POSs, Part Of Speech, Part of Speech, Part-Of-Speech, Part-of-Speech, Parts Of Speech, Parts of Speech, Pos, part of speech, part-of-speech, parts of speech, parts-of-speech	101,333	0.0046	13803	0.21802	7.34
6	classifier	classifiers	98,092	0.00446	11513	0.18185	8.52
7	NP	NPs, noun phrase, noun phrases	94,808	0.00431	9584	0.15138	9.89
8	Parser	Parsers	86,901	0.00395	9636	0.1522	9.02
9	Segmentation	Segmentations	76,232	0.00346	10850	0.17138	7.03
10	SNR	SNRs, Signal Noise Ratio, Signal Noise Ratios, signal noise ratio, signal noise ratios	68,722	0.00312	6848	0.10817	10.04
11	Dataset	Data-set, data-sets, datasets	65,310	0.00297	9941	0.15702	6.57
12	Semantic		61,737	0.0028	12906	0.20385	4.78
13	Parsing	Parsings	58,750	0.00267	9390	0.14832	6.26
14	GMM	GMMs, Gaussian Mixture Model, Gaussian Mixture Models, Gaussian mixture model, Gaussian mixture models	58,297	0.00265	5829	0.09207	10.00
15	MT	MTs, Machine Translation, Machine Translations, machine translation, machine translations	56,703	0.00258	8242	0.13018	6.88
16	Iteration	Iterations	52,772	0.0024	11664	0.18424	4.52
17	Neural network	ANN, ANNs, Artificial Neural Network, Artificial Neural Networks, NN, NNs, Neural Network, Neural Networks, NeuralNet, NeuralNets, neural networks	51,584	0.00234	8473	0.13383	6.09
18	Metric	Metrics	50,690	0.0023	11318	0.17877	4.48
19	SVM	SVMs, Support Vector Machine, Support Vector Machines, support vector machine, support vector machines	50,301	0.00228	5974	0.09436	8.42
20	WER	WERs, Wer, word error rate, word error rates	47,812	0.00217	6381	0.10079	7.49

number, US/UK difference, abbreviation/expanded form and absence/presence of a semantically neutral adjective, like "artificial" in "artificial neural network." Thirdly, the most frequent terms were recomputed with the amended lexicon. We will call "*existence*"[6] the fact that a term exists in a document and "*presence*" the percentage of documents where the term exists. We computed in that way the occurrences, frequencies, existences and presences of the terms globally and over time (1965–2015), and the average number of occurrences of the terms in the documents where they exist (**Table 1**).

The ranking of the terms slightly differs whether we consider the frequency or the presence. The most frequent term overall is "HMM" (*Hidden Markov Models*), while the most present term is "*Speech Recognition,*" which is present in 32% of the papers.

The average number of occurrences of the terms in the documents where they exist varies a lot (from 10 for "*Signal/Noise ratio*" or "*Gaussian Mixture Models*" to 4.5 for "*metric*").

Change in Topics

We studied the evolution over the years among the 200 yearly most popular terms (mixing unigrams, bigrams, and trigrams) representing the corresponding topics of interest, according to their ranking, based on their frequency or presence. We developed for this a visualization tool[7] that allows to play with various parameters related to data selection [use of frequency or presence, type of ranking (raw or proportional to frequency or to presence), use and importance of smoothing, covered time period, number of topics per year (from 10 to 200)] and data visualization (size and colors of the boxes and links, selection of topics, etc.) (Perin et al.,

[6] Sometimes called "Boolean frequency" or "binary frequency."

[7] Gapchart: https://rankvis.limsi.fr/

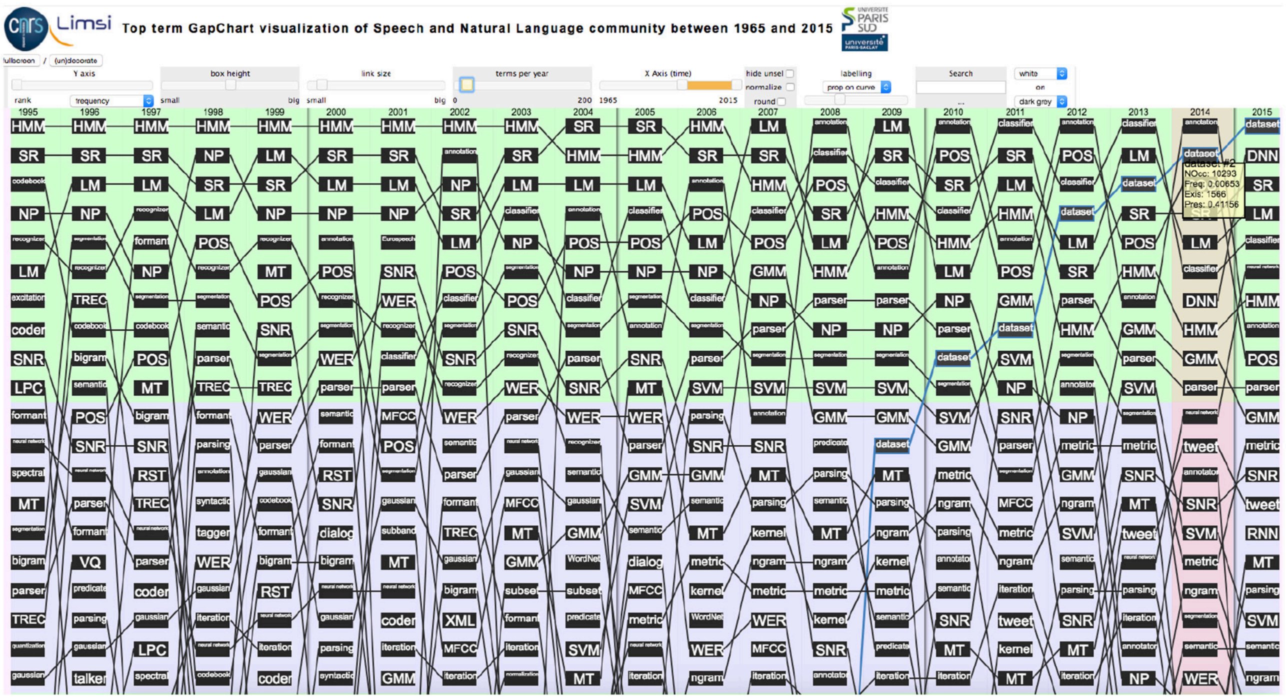


FIGURE 1 | Evolution of the top 20 terms over 20 years (1996–2015) according to their frequency (raw ranking without smoothing. The yellow box indicates the number of Occurrences, Frequency, Number of Existences and Presence of the term "Dataset" ranked 2nd in 2014).

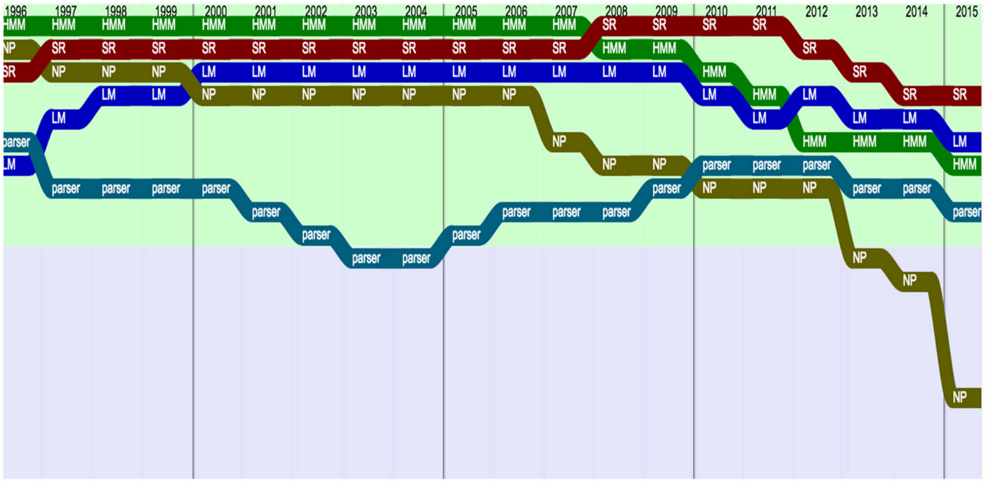


FIGURE 2 | Topics remaining popular (raw ranking, according to Frequency with smoothing).

2016) (**Figure 1**). The raw figure is poorly readable, but focusing on specific terms depicts clear trends as it appears in **Figures 2–6**.

We see that some terms remained popular, such as "*HMM,*" "*Speech recognition,*" "*Language Model,*" "*Noun Phrase*" or "*Parser,*" which stayed in the top 20 terms over 20 years from 1996 to 2015 (**Figure 2**).

We also studied several terms that became more popular over time, such as "*Annotation*" and "*Wordnet,*" which gained a lot of popularity in 1998 when the first LREC was organized, "*Gaussian Mixture Models (GMM)*" and "*Support Vector Machines (SVM),*" "*Wikipedia,*" and, recently, "*Dataset,*" "*Deep Neural Networks (DNN)*" blooming in the top 40 terms in 2013 and "*Tweet*" blooming in the top 20 in 2011 (**Figure 3**).

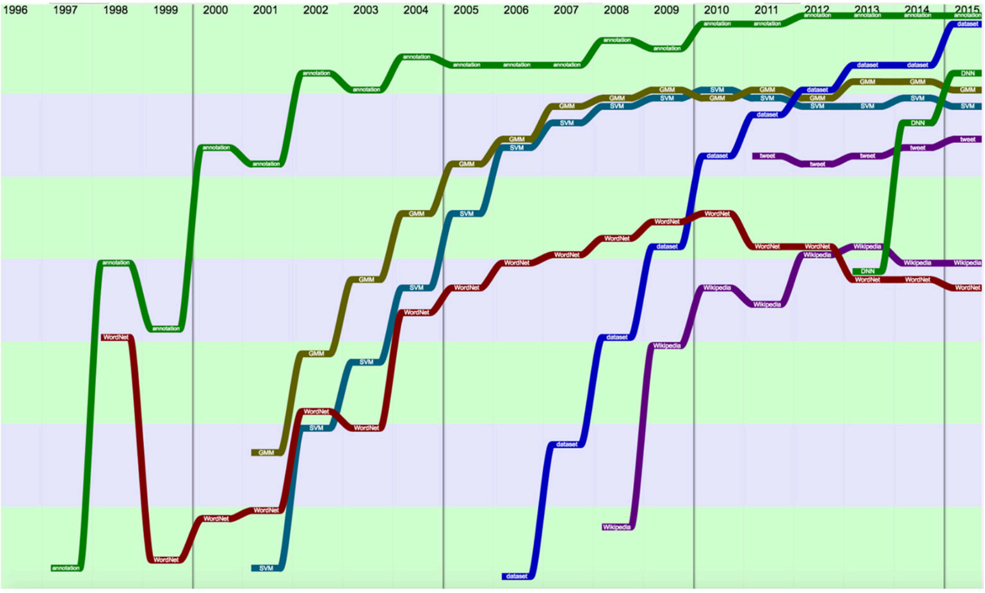


FIGURE 3 | Topics becoming popular (raw ranking, according to Frequency with smoothing).

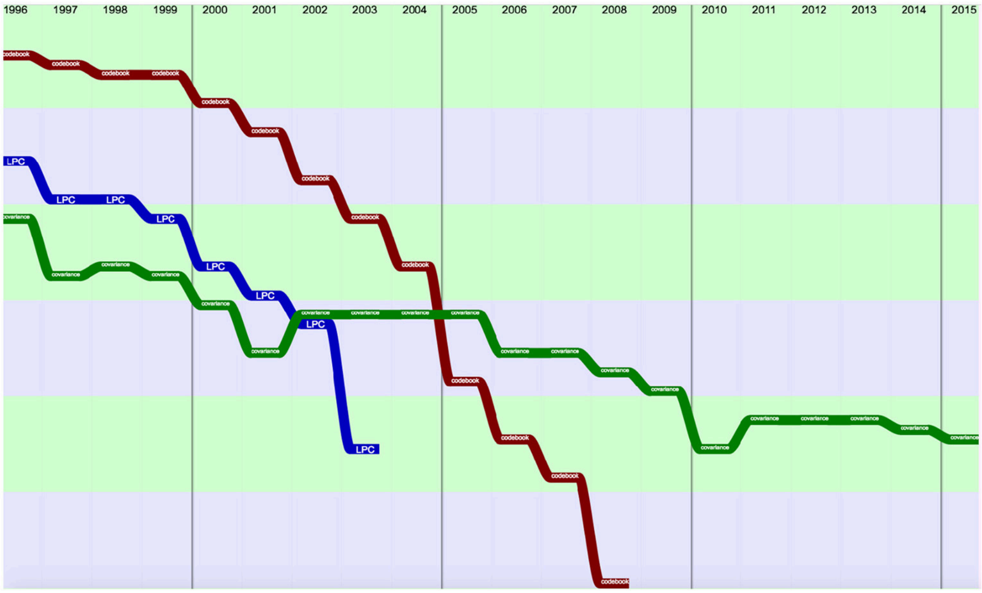


FIGURE 4 | Topics losing popularity (raw ranking, according to Frequency with smoothing).

Among terms losing popularity, we may find "*Codebook,*" "*Covariance,*" and "*Linear Prediction Coding (LPC),*" which disappeared from the top 50 terms in 2005 (**Figure 4**).

We also studied the changes in the use of some related terms, such as "*bigram*" and "*trigram*" that were clearly replaced by "*Ngram*" (**Figure 5**).

We compared the evolution of HMM and Neural Networks over 20 years, in terms of presence (% of papers containing the term) (**Figure 6**). We see a spectacular return of interest for "*Neural Networks*" starting in 2012.

Tag Clouds for Frequent Terms

The aim of Tag Clouds is to provide a global estimation of the main terms used in over the years as well as an indication of the stability of the terms over the years. For this purpose, we use TagCrowd[8] to generate Tag Clouds and we only considered the papers' abstracts.

[8] www.tagcrowd.com. Our thanks to Daniel Steinbock for providing access to this web service.

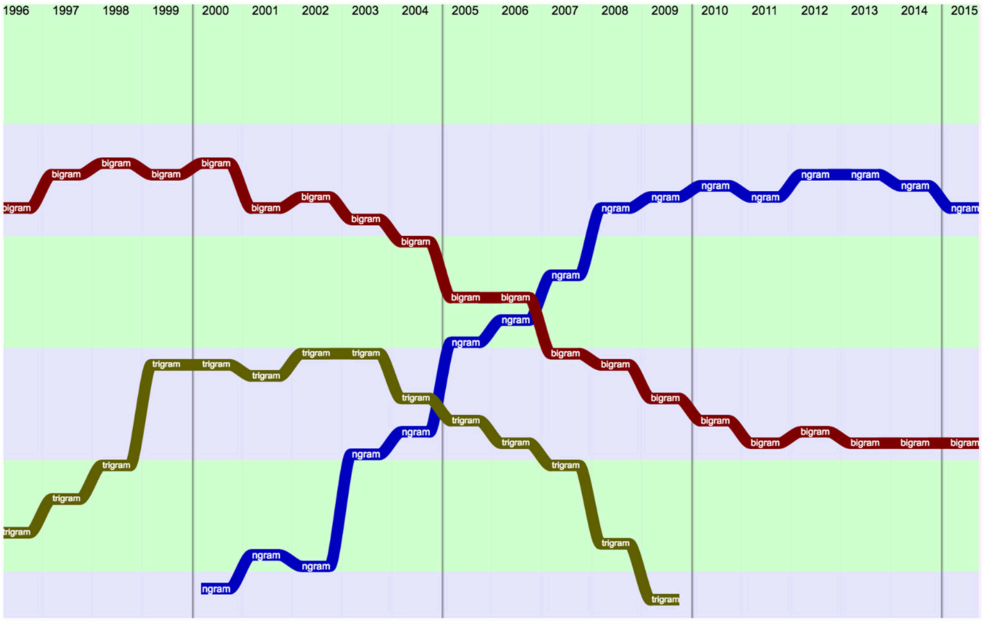


FIGURE 5 | Comparison of bigram, trigram, and Ngram over 20 years (raw ranking, according to Frequency with smoothing).

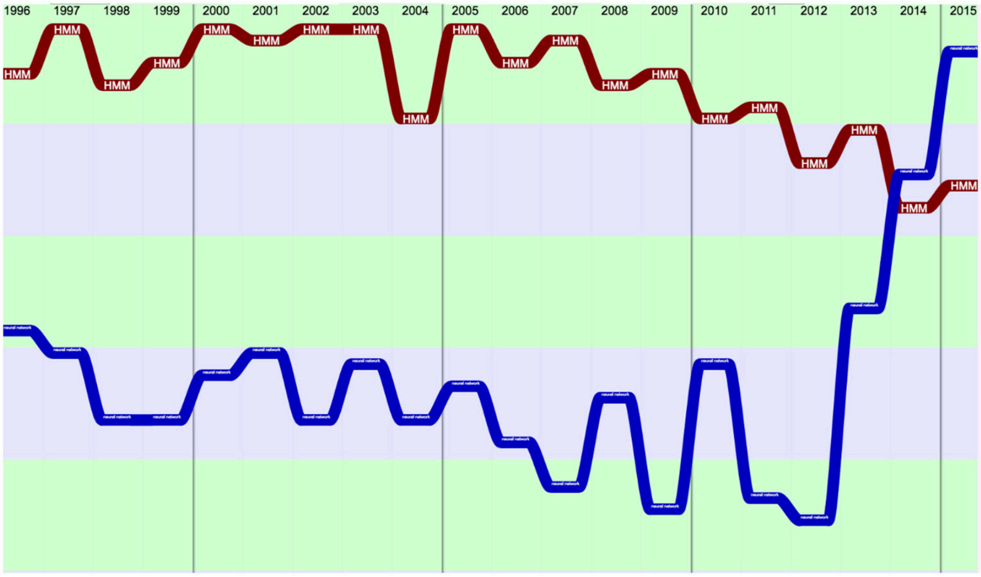


FIGURE 6 | Comparison of HMM and neural networks over 20 years (raw ranking, according to presence).

Figure 7 shows the tag clouds in 10 years intervals from 1965 to 2015. Globally, it appears that the most frequent terms changed over the years. In 1965, only COLING is considered. Most of the terms concerned computation. In 1975, only *Computer and the Humanities* and the *IEEE Transactions on Acoustics, Speech and Signal Processing* are considered. The Tag Cloud still shows a large presence of generic terms, but also of terms attached to audio processing. In 1985, the number of sources is larger and more diversified. The interest for parsing is clear. HMM, and especially discrete models, appear neatly in 1995 together with speech recognition and quantization, while in NLP, *TEI (Text Encoding Initiative), SGML (Standard Generalized Markup Language),* and *MT* are mentioned. The year 2005 shows the growing interest for Language Resources (*Annotation)* and for evaluation *(metric, WER)*, while *MT* is increasing and *GMM* stands next to *HMM*. 2015 is the year of *neural networks [DNN*

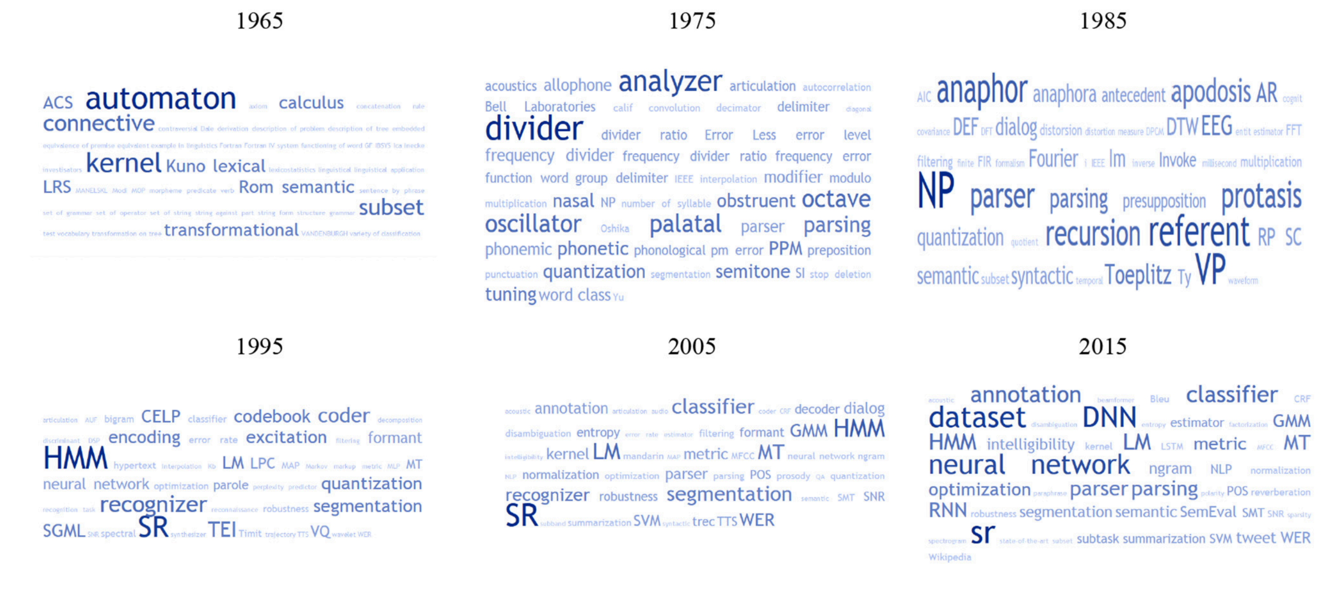

1965
ACS automaton calculus
connective
kernel Kuno lexical
LRS Rom semantic
subset
transformational
1975
acoustics allophone analyzer articulation autocorrelation
Bell Laboratories calif convolution decimator delimiter
divider divider ratio Error Less error level
frequency divider frequency divider ratio frequency error
function word group delimiter IEEE interpolation modifier modulo
multiplication nasal NP number of syllable obstruent octave
oscillator Oshika palatal parser parsing
phonemic phonetic phonological pm error PPM preposition
punctuation quantization segmentation semitone SI stop deletion
tuning word class Yu
1985
AIC anaphor anaphora antecedent apodosis AR
covariance DEF DFT dialog distorsion DTW EEG FFT
filtering FIR formalism Fourier IEEE Im Invoke multiplication
NP parser parsing presupposition protasis
quantization recursion referent RP SC
semantic subset syntactic Toeplitz Ty VP waveform
1995
articulation bigram CELP classifier codebook coder decomposition
encoding error rate excitation filtering formant
HMM hypertext LM LPC MAP MT
neural network optimization parole quantization
recognizer robustness segmentation
SGML spectral SR TEI Timit TTS VQ WER
2005
acoustic annotation audio classifier decoder dialog
disambiguation entropy filtering formant GMM HMM
intelligibility kernel LM mandarin metric MFCC MT neural network ngram
normalization optimization parser parsing POS prosody QA quantization
recognizer robustness segmentation SMT SNR
SR summarization SVM trec TTS WER
2015
annotation Bleu classifier CRF
dataset DNN estimator GMM
HMM intelligibility kernel LM LSTM metric MT
neural network ngram NLP normalization
optimization parser parsing POS reverberation
RNN robustness segmentation semantic SemEval SMT SNR
spectrogram sr subtask summarization SVM tweet WER
Wikipedia

FIGURE 7 | Tag cloud based on the abstracts from 1965 to 2015.

TABLE 2 | Research topics prediction using the Weka software environment.

Observed in 2013	Observed in 2014	Predicted for 2015	Observed in 2015	Rank
Classifier (0.00576)	Annotation (0.00792)	Dataset (0.00653)	Dataset (0.00886)	1
LM (0.00565)	Dataset (0.00639)	Annotation (0.00626)	DNN (0.00613)	2
Dataset (0.00548)	POS (0.00600)	POS (0.00549)	Classifier (0.00491)	3
POS (0.00536)	LM (0.00513)	LM (0.00479)	POS (0.00485)	4
Annotation (0.00509)	Classifier (0.00507)	classifier (0.00466)	Neural network (0.00455)	5
SR (0.00507)	SR (0.00449)	DNN (0.00437)	LM (0.00454)	6
HMM (0.00478)	Parser (0.00388)	SR (0.00429)	SR (0.00439)	7
Parser (0.00404)	DNN (0.00369)	HMM (0.00365)	Parser (0.00436)	8
GMM (0.00367)	HMM (0.00352)	Neural network (0.00345)	Annotation (0.00414)	9
Segmentation (0.00298)	Neural network (0.00326)	Tweet (0.00312)	HMM (0.00384)	10

(Deep Neural Networks), RNN (Recurrent Neural Networks)] together with data *(Dataset)*. *Speech Recognition (SR)* stayed popular since 1995, while *Parsing* comes back to the forefront.

Research Topic Prediction

Machine Learning for Time Series Prediction

We also explored the feasibility of predicting the research topics for the coming years based on the past (Francopoulo et al., 2016a). We used for this the Weka[9] machine learning software package (Witten et al., 2011). We applied each of the 21 algorithms contained in Weka to the time series of terms up to 2014 ordered according to their frequency and retained the one which provided the best results with the corresponding set of optimal parameters (especially the past history time length), after a-posteriori verification on the observed 2015 data. We then applied this software to the full set of the NLP4NLP corpus, year by year.

Table 2 gives the ranking of the most frequent terms in 2013 and 2014 with their frequency, the topic predicted by the selected Weka algorithm for 2015 on the basis of the past rankings and the ranking actually observed in 2015. We see that the prediction is correct for the top term ("*dataset*"). The next predicted term was "*annotation*" which only appears at the 9th rank, probably due to the fact that LREC didn't take place in 2015. It is followed by "*POS,*" which actually appears at the 4th rank with a frequency close to the predicted one.

Prediction Reliability

As we have the information on the actual observations in the annual rankings, it is possible to measure the reliability of the predictions by measuring the distance between the predicted frequencies and the observed frequencies. **Figure 8** gives this distance for the predictions in year 2011 to 2015 based on time series until 2010. We see that the distance largely increases in 2013, i.e., 3 years after the year of prediction. We may therefore think that it is not unreasonable to predict the future

[9] www.cs.waikato.ac.nz/ml/weka

of a research domain within a 2-year horizon (unless a major discovery happens in the meanwhile...).

Scientific Paradigms Ruptures

It is also possible to measure the difference between the prediction and the observation in each year. It provides a measure of the "surprise" between what we were expecting and what actually occurred. The years where this "surprise" is the largest may correspond to epistemological ruptures. **Figure 9** gives the evolution of this distance between 2011 and 2015. We see that 2012 was a year of big changes.

We may also compute this distance for a specific topic, in order to analyze the way this term evolves compared with what was expected. **Figure 10** shows the evolution of the *"Deep Neural Network" (DNN)* topic. We see that up to 2014, we didn't expect the success of this approach in the next year, while, starting in 2014, it became part of the usual set of tools for automatic language processing.

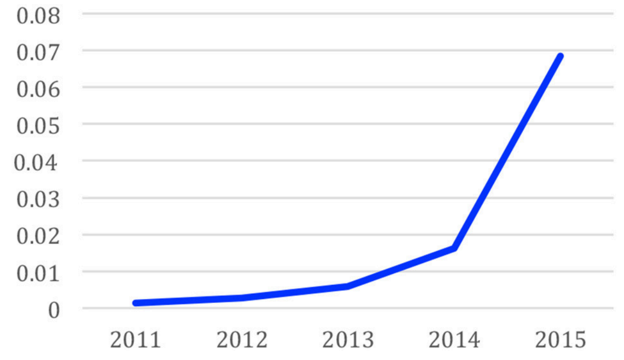


FIGURE 8 | Reliability of the predictions: prediction error over the years from 2011.

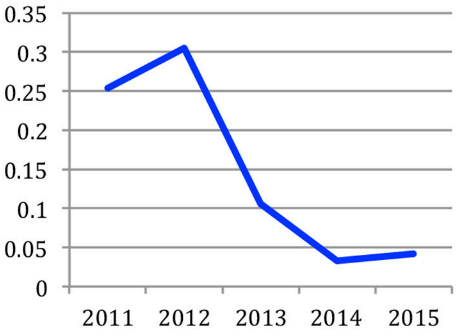


FIGURE 9 | Evolution of the distance between prediction and observation over the years.

Predictions for the Next 5 Years

Table 3 provides the predictions for the next 5 years starting in 2016: not surprisingly, it is expected that *Neural Networks*, more or less *deep* and more or less *recurrent*, will keep on attracting the researchers' attention.

Innovation

New Terms Introduced by the Authors

We then studied when and who introduced new terms, as a mark of the innovative ability of various authors, which may also provide an estimate of their contribution to the advances of the scientific domain (Mariani et al., 2018a). We make the hypothesis that an innovation is induced by the introduction of a term which was previously unused in the community and then became popular. We consider the 61,661

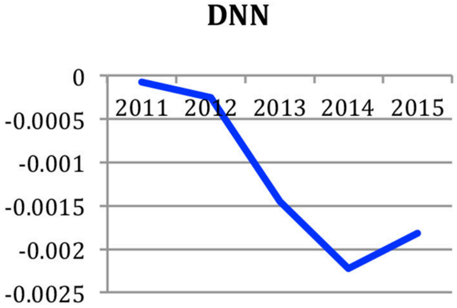


FIGURE 10 | Measure of the expectation of an emerging research topic: Deep Neural Networks (DNN).

TABLE 3 | Predictions for the next 5 years 2016–2020.

Observed 2014	Observed 2015	Prediction 2016	Prediction 2017	Prediction 2018	Prediction 2019	Prediction 2020	Rank
Annotation	Dataset	Dataset	Dataset	Dataset	Dataset	Dataset	1
Dataset	DNN	DNN	DNN	DNN	DNN	DNN	2
POS	Classifier	Annotation	Neural network	Neural network	Neural network	Neural network	3
LM	POS	POS	SR	RNN	RNN	RNN	4
Classifier	Neural network	Neural network	Classifier	POS	Parser	Parser	5
SR	LM	Classifier	LM	Parser	SR	SR	6
Parser	SR	Parser	POS	Annotation	LM	Metric	7
DNN	Parser	SR	RNN	Classifier	Classifier	POS	8
HMM	Annotation	LM	Parser	SR	Metric	Parsing	9
Neural network	HMM	HMM	HMM	Metric	POS	Classifier	10

documents written in English and the 42,278 authors who used the 3,314,671 terms contained in those documents. Two thousand and fifty-four of those terms are present in the 20 documents of the first year (1965), which we consider as the starting point for the introduction of new terms, while we find 333,616 of those terms in the 3,214 documents published in 2015.

We then take into account the terms that are of scientific interest (excluding author's names, unless they correspond to a specific algorithm or method, city names, laboratory names, etc.). For each of these terms, starting from 1965, we determine the author(s) who introduced the term, referred to as the "inventor(s)" of the term. This may yield several names, as the papers could be co-authored or the term could be mentioned in more than one paper on a given year.

Table A1 provides the ranked list of the 10 most popular terms according to their presence in 2015. The ranking of the terms slightly differs if we consider the frequency or the presence. The most frequent term in the archive according to **Table 1**, *Hidden Markov Models (HMM),* doesn't appear on **Table A1** as it is ranked 16th in 2015. The most present term is *Dataset*, which appeared first in 1966, when it was mentioned in a single paper authored by L. Urdang[10], while it was mentioned 14,039 times in 1,472 papers in 2015, and 65,250 times in 9,940 papers overall (i.e., in 16% of the papers!). From its first mention in the introduction of a panel session by Bonnie Lynn Webber at ACL[11] in 1980 to 2015, the number of papers mentioning *Neural Networks* increased from 1 to 1037, and the number of occurrences reached 8,024 in 2015. *Metric, Subset, Classifier, Speech Recognition, Optimization, Annotation, Part-of-Speech,* and *Language Model* are other examples of terms that are presently most popular.

Measuring the Importance of Topics

We then considered the way to measure the importance of a term. **Figure 11A** gives an example of the annual presence (percentage of papers containing the term) for the term "*cross validation,*" which was encountered for the first time in 2 papers in 2000. In order to measure the success of the term over time, we may consider all papers or only those ("external papers" marked in red) that are written by authors who are different than those who introduced the term (marked in blue).

We propose to compute as the annual innovation score of the term the presence of the term on that year (in this example, it went from 0.75% of the papers in 2000 to 4% of the papers in 2014) and to compute as the global innovation score of the term the corresponding surface, taking also into account the inventors' papers in the year of introduction and all the papers in the subsequent years (**Figure 11B**).

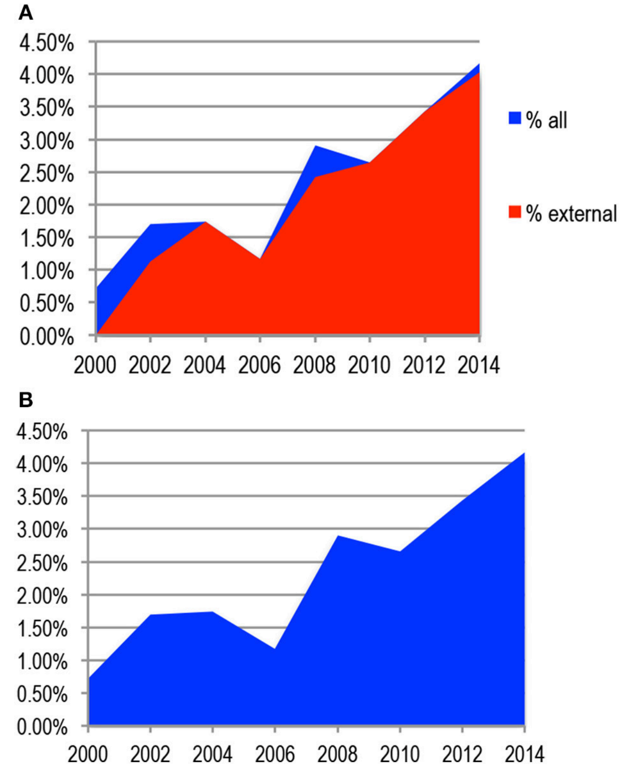


FIGURE 11 | (A) Presence of the term "cross validation" over the years. **(B)** Innovation Score of the term "cross validation".

In this way, it takes into account the years when the term gains popularity (2000 to 2004, 2006 to 2008, and 2010 to 2014 in the case of "*cross validation*"), as well as those when it loses popularity (2004 to 2006 and 2008 to 2010). The innovation score for the term is the sum of the yearly presences of the term and amounts to 0.17 (17%). This approach emphasizes the importance of the term in the first years when it is mentioned, as the total number of papers is then lower. Some non-scientific terms may not have been filtered out, but their influence will be small as their presence is limited and random, while terms that became popular at some point in the past but lost popularity afterwards will remain in consideration.

We considered the 1,000 most frequent terms over the 50-year period, as we believe they contain most of the important scientific advances in the field of SNLP. Given the poor quality and low number of different sources and papers in the first years, we decided to only consider the period from 1975 to 2015. This innovation measure provides an overall ranking of the terms. We also computed separate rankings for NLP and for Speech (**Table 4**), based on the categorization of the sources.

We studied the evolution of the presence of the terms over the years, in order to check the changes in paradigm. However, the fact that some conferences are annual, while others are biennial brings noise, as we already observed when studying

[10] Laurence Urdang (1966), The Systems Designs and Devices Used to Process The Random House Dictionary of the English Language. Computer and the Humanities. Interestingly, the author writes: *"Each unit of information-regardless of length-was called a dataset, a name which we coined at the time. (For various reasons, this word does not happen to be an entry in The Random House Dictionary of the English Language, our new book, which I shall refer to as the RHD)."* a statement which witnesses her authorship of the term.

[11] Interestingly, she mentions the Arthur Clarke's "2001, Space Odyssey" movie: *"Barring Clarke's reliance on the triumph of automatic neural network generation, what are the major hurdles that still need to be overcome before Natural Language Interactive Systems become practical?"* which may appear as a premonition in 1980!

TABLE 4 | Global ranking of the importance of the terms overall and separately for Speech and NLP.

Rank	Terms		
	Overall	**NLP**	**Speech**
1	Speech recognition	Semantic	Speech recognition
2	Subset	Syntactic	Spectral
3	Semantic	NP	Acoustics
4	Filtering	POS	Gaussian
5	HMM	Parser	HMM
6	Spectral	Parsing	Filtering
7	Linear	Subset	Linear
8	Iteration	Lexical	Fourier
9	Language model	Machine translation	Subset
10	POS	predicate	Acoustic

TABLE 5 | Global ranking of authors overall and separately for Speech and NLP.

Rank	Authors		
	Overall	**NLP**	**Speech**
1	Lawrence R. Rabiner	Ralph Grishman	Lawrence R. Rabiner
2	Hermann Ney	Kathleen R. Mckeown	John H. L. Hansen
3	John H. L. Hansen	Jun'Ichi Tsujii	Shrikanth S. Narayanan
4	Shrikanth S. Narayanan	Aravind K. Joshi	Hermann Ney
5	Chin Hui P. Lee	Jaime G. Carbonell	Chin Hui P. Lee
6	Li Deng	Ralph M. Weischedel	Li Deng
7	Mari Ostendorf	Mark A. Johnson	Mark J. F. Gales
8	Alex Waibel	Fernando C. N. Pereira	Frank K. Soong
9	Haizhou Li	Christopher D. Manning	Haizhou Li
10	John Makhoul	Ted Briscoe	Thomas Kailath

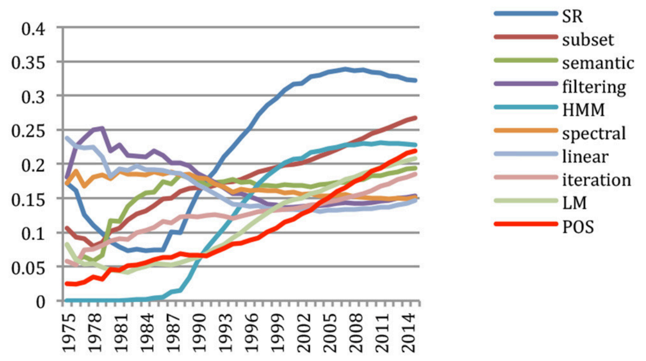


FIGURE 12 | Cumulative presence of the 10 most important terms over time (% of all papers).

citations. Instead of considering the annual presence of the terms (percentage of papers containing a given term **on** a given year), we therefore considered the cumulative presence of the terms (percentage of papers containing a given term **up to** a given year) (**Figure 12**).

We see that *Speech Recognition* has been a very popular topic over the years, reaching a presence in close to 35% of the papers published up to 2008. Its shape coincides with *Hidden Markov Models* that accompanied the effort on *Speech Recognition* as the most successful method over a long period and had then been mentioned in close to 25% of the papers by that time. *Semantic* processing was a hot topic of research by the end of the 80's, and regained interest recently. *Language Models* and *Part-of-Speech* received continuing marks of interest over the years.

Measuring Authors' Innovation

We also computed in a similar way an *innovation score* for each author, illustrating his or her contribution in the introduction and early use of new terms that subsequently became popular. The score is computed as the sum over the years of the annual presence of the terms in papers published by the authors (percentage of papers containing the term and signed by the author on a given year). This innovation measure provided an overall ranking of the authors. We also computed separate rankings for NLP and for Speech Processing (**Table 5**).

We should stress that this measure doesn't place on the forefront uniquely the "inventors" of a new topic, as it is difficult to identify them given that we only consider a subset of the scientific literature over a limited period. It rather helps identifying the early adopters who published a lot when or after the topic was initially introduced. We studied several cases where renowned authors don't appear within the 10 top authors contributing to those terms, such as F. Jelinek regarding *Hidden Markov Models*. The reason is that they initially published in a different research field than SNLP (the *IEEE Transactions on Information Theory* in the case of F. Jelinek, for example) that we don't consider in our corpus. This measure also reflects the size of the production of papers from the authors on emerging topics, with an emphasis on the pioneering most ancient authors, such as L. Rabiner and J. Makhoul, at a time when the total number of papers was low. The overall ranking also favors those who published both in Speech and Language Processing, such as H. Ney or A. Waibel.

We may study the domains where the authors brought their main contributions, and how it evolves over time. We faced the same problem due to the noise brought by the different frequency of the conferences as we did when studying the evolution of the terms, and we rather considered the cumulative contribution of the author specific to that term [percentage of papers signed by the author among the papers containing a given term (that we will call "*topical papers*") **up to** a given year]. We see for example that L. Rabiner brought important early contributions to the fields of *Acoustics*, *Signal Processing* and *Speech Recognition* in general, and specifically to *Linear Prediction Coding (LPC)* and *filtering* (**Figure 13**). He even authored 30% of the papers dealing with *LPC* which were published up to 1976 and the only paper mentioning *endpoint detection* in 1975.

H. Ney brought important contributions to the study of *perplexity* (authoring 10% of the papers which were published on that topic up to 1988) and in *Language Models* (LM) using trigrams and bigrams (**Figure 14**).

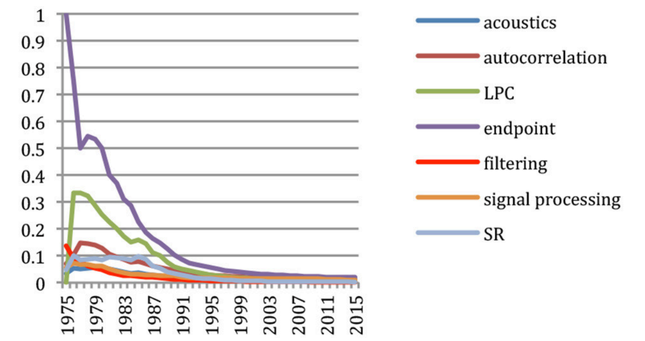


FIGURE 13 | Main contributions areas for L. Rabiner (% of topical papers).

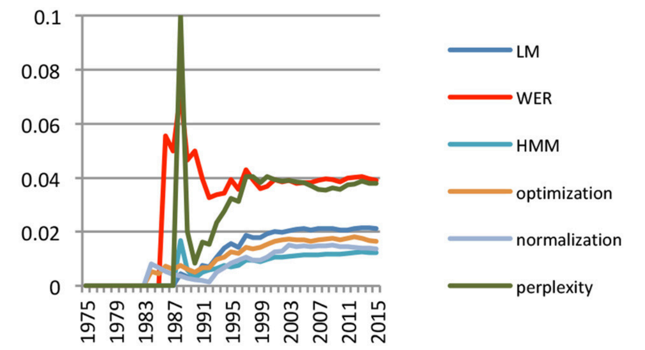


FIGURE 14 | Main contribution areas for H. Ney (% of topical papers).

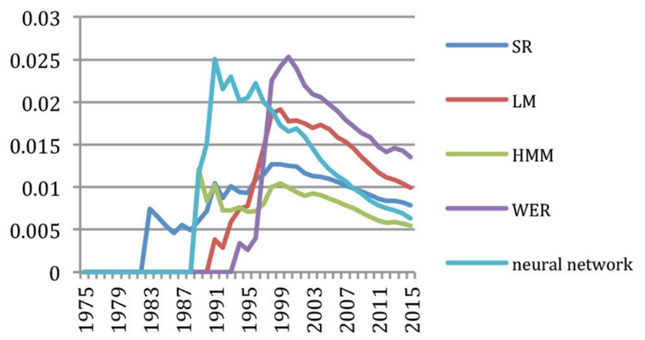


FIGURE 15 | Main contribution areas for A. Waibel (% of topical papers).

A. Waibel brought important contributions in the use of *HMM* and even more of *Neural Networks* for speech and language processing already in the early 90s (**Figure 15**).

We may also wish to study the contributions of authors on a specific topic, using the same cumulative score. **Figure 16** provides the cumulative percentage of papers containing the term *HMM* published up to a given year by the 10 most contributing authors. We also added F. Jelinek as a well-known pioneer in that field and S. Levinson as the author of the first article containing that term in our corpus, which represented 0.4% of the papers published in 1982. We see the contributions of pioneers such as F. Soong, of important contributors in an early stage such as C. H. Lee, S. Furui, or K. Shikano or a later stage such as M. Gales.

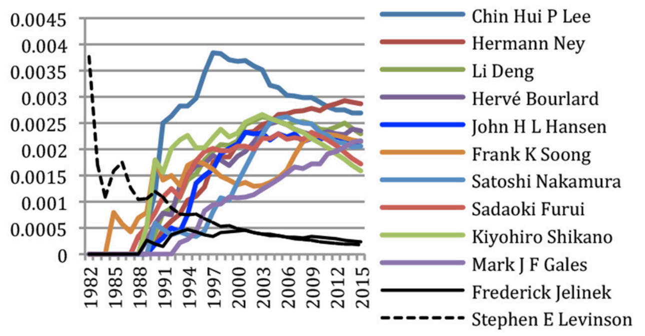


FIGURE 16 | Authors' contributions to HMM in SNLP (% of all papers).

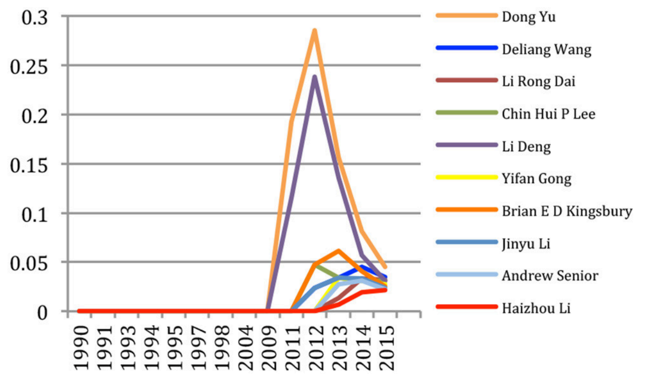


FIGURE 17 | Authors' contributions to the study of DNN in speech and language processing (% of topical papers).

Similarly, we studied the authors' contributions to *Deep Neural Networks (DNN)* which recently gained a large audience (**Figure 17**). We see the strong contribution of Asian authors on this topic, with the pioneering contributions of Dong Yu and Li Deng up to 2012 where they represented altogether about 50% of the papers mentioning DNN since 2009, while Deliang Wang published later but with a large productivity which finally places him at the second rank globally.

Measuring the Innovation in Publications

We finally computed with the same approach an *innovation score* for each publication. The score is similarly computed as the sum over the years of the annual presence of the terms in papers published in the source, conference or journal (percentage of papers containing the term which were published in the publication on a given year). This innovation measure provided an overall ranking of the publication. We also computed separate rankings for NLP and for Speech Processing (**Table 6**).

Just as in the case of authors, the measure also reflects here the productivity, which favors the Speech Processing field where more papers have been published, and the pioneering activities, as reflected by the ranking of *IEEE TASLP*. In the overall ranking,

TABLE 6 | Global ranking of the importance of the sources overall and separately for Speech and NLP.

Rank	Sources		
	Overall	**NLP**	**Speech**
1	taslp	acl	taslp
2	isca	coling	isca
3	icassps	cath	icassps
4	acl	lrec	lrec
5	coling	cl	csal
6	lrec	hlt	speechc
7	hlt	eacl	mts
8	emnlp	emnlp	ltc
9	cl	trec	lre
10	cath	mts	acmtslp

publications that concern both Speech and Language Processing (LREC, HLT) also get a bonus here.

We may study the domains where the publications brought their main contributions, and how it evolves over time. We faced the same problem due to the noise brought by the different frequency of the conferences as we did when studying the evolution of the terms and authors, and we rather considered the cumulative contribution of the publication specific to that term (percentage of papers published in the source among the papers containing the term **up to** a given year). We see for example (**Figure 18**) that ACL showed a strong activity and represented 40% of papers published about *parsing*, 35% of papers published about *semantic, syntactic,* and *lexical* and 25% of papers published about *Machine Translation* up to 1985. Its share in those areas then globally decreases to about 15% of the total number of publications in 2015, due to the launching of new conferences and journals, while the share of publications on *Machine Translation* within ACL recently increased.

We may also wish to study the contributions of publications to a specific term, using the same cumulative score. **Figure 19** provides the cumulative percentage of papers containing the term *HMM* published up to a given year by the 10 most contributing publications. We see that all papers were initially published in the *IEEE Transactions on Speech and Audio Processing*. Other publications took a share of those contributions when they were created (*Computer Speech and Language* starting in 1986, *ISCA Conference series* starting in 1987) or when we start having access to them (*IEEE-ICASSP*, starting in 1990). We see that *ISCA Conference series* represents 45% of the papers published on HMM up to 2015, while *IEEE-ICASSP* represents 25%. We also see that HMMs were first used in speech processing related publications, then in NLP publications as well (ACL, EMNLP), while publications that are placed in both (CSL, HLT, LREC) helped spreading the approach from speech to NLP.

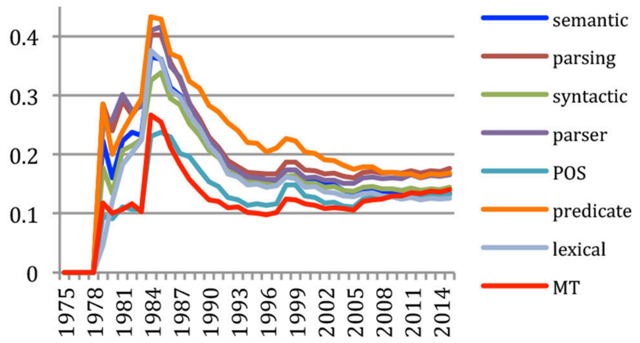


FIGURE 18 | Main domains within the ACL conference series (% of topical papers).

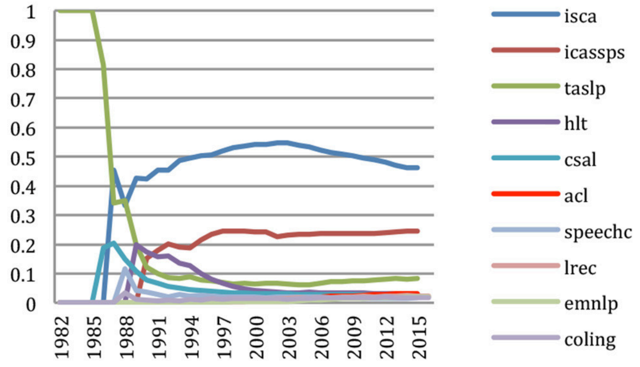


FIGURE 19 | Sources' contributions to the study of HMM (% of topical papers).

The measure of innovation we propose for terms, authors and sources gives an image of the scientific community that seems acceptable. However, it emphasizes the eldest contributions and the productivity, and should be refined. In this analysis, we faced the problem of the lack of quality of the most ancient data that was obtained through OCR from the paper version of the proceedings, which sometimes even contain handwritten comments! For that reason, we focused the study on the period starting in 1975 and we still had to carry out some manual corrections. An automatic term extraction process taking into account the context in which the term is identified would allow making the distinction between real and false occurrences of the terms, especially when they have acronyms as variants. It would avoid the tedious manual checking that we presently conduct and would improve the overall process.

Use of Language Resources

The LRE Map

We have similarly conducted an analysis of the mentions of Language Resources (LR) in the papers of the corpus. Language Resources are bricks that are being used by researchers to conduct their research investigations and develop their system (Francopoulo et al., 2016b). We consider here Language Resources in the broad sense embracing data (e.g., corpus, lexicons, dictionaries, terminological databases, etc.), tools (e.g., morpho-syntactic taggers, prosodic analyzers, annotation tools, etc.), system evaluation resources (e.g., metrics, software, training, dry run or test corpus, evaluation package, etc.), and

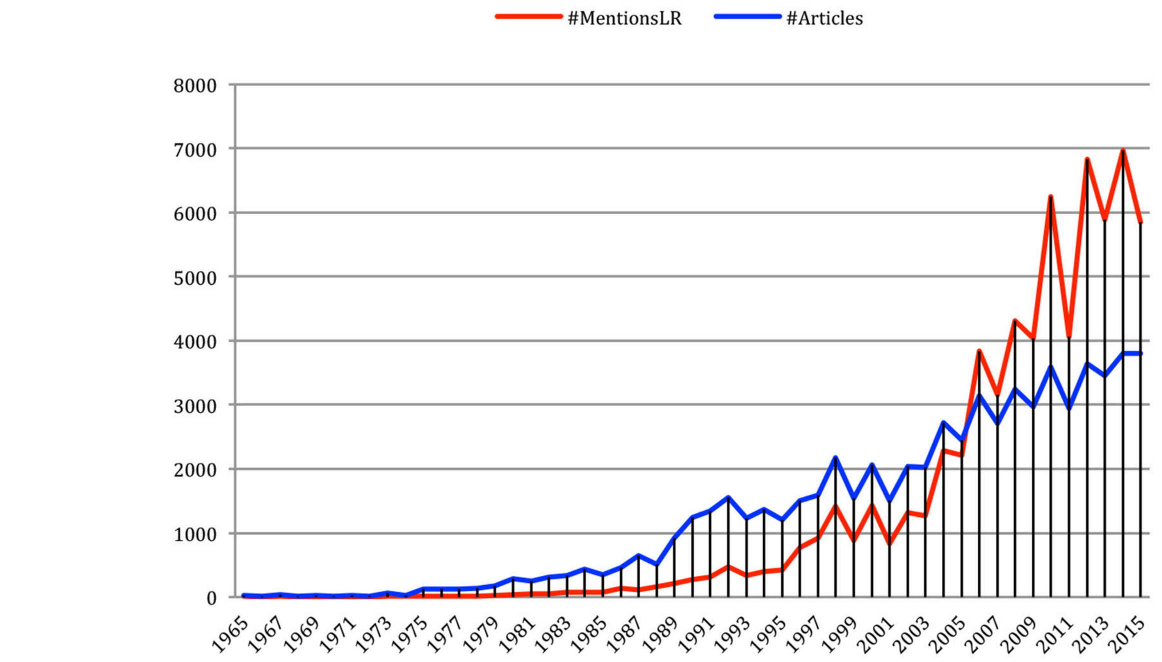


FIGURE 20 | Evolution of the number of mentions of Language Resources in papers over the years.

meta-resources (e.g., best practices, guidelines, norms, standards, etc.).

We considered the Language Resources that are mentioned in the LRE Map (Calzolari et al., 2012). This database was produced in the FlaReNet European project and is constituted by the authors of papers at various conferences of the domain who are invited when submitting their paper to fill in a questionnaire which provides the main characteristics of the Language Resources produced or used in the research investigations that they report in their paper. The LRE Map that we used contains information harvested in 10 conferences from 2010 to 2012, for a total of 4,396 resources. After cleaning those entries (correcting the name of the resources, eliminating the duplicates, regrouping the various versions of resources from the same family, etc.), we ended up with 1,301 different resources that we searched in the NLP4NLP corpus.

Evolution of the Use of Language Resources

Table A2 provides the number of mentions (that we will call "*existences*") of different Language Resources from the LRE Map together with the number of documents that were published each year from 1965 to 2015, with the list of the 10 most cited Language Resources every year. We studied the evolution of the number of different resources mentioned in the papers compared with the evolution of the number of papers over the years (**Figure 20**). It appears that the corresponding curves cross in 2005, date since which more than one Language Resource is mentioned on average in a paper. This may reflect the shift from *Knowledge-based* approaches to *Data-driven* approaches in the history of NLP research.

Table 7 provides the ranking of Language Resources according to the number of papers where they are mentioned ("*existences*"). It also gives for each resource its type (corpus, lexicon, tool, etc.), the number of mentions in the papers ("*occurrences*"), the first authors who mentioned it as well as the first publications, and the first and final year when it was mentioned. We see that "WordNet" comes first, followed by "Timit," "Wikipedia," "Penn Treebank" and the "Praat" speech analysis tool.

One may also track the propagation of a Language Resource in the corpus. **Figure 21** gives the propagation of the "WordNet" resource, which initially appeared in the HLT conference in 1991, and then propagated on the following years, first in computational linguistics conferences, then also in speech processing conferences. **Figure 22** provides another view of the same propagation, which includes the number of mentions in each of the sources.

Language Resources Impact Factor

We may attribute an Impact Factor to Language Resources according to the number of articles that mention the resource as it appears in **Table 7**. **Table 8** provides the Impact Factors for the LR of the "Data" and "Tools" types. It exemplifies the importance of the corresponding LR for conducting research in NLP and aims at recognizing the contribution of the researchers who provided those LR, just like a citation index.

Text Reuse and Plagiarism

Here we study the reuse of NLP4NLP papers in other NLP4NLP papers (Mariani et al., 2016, 2017a).

TABLE 7 | Presence of the LRE Map Language Resources in the NLP4NLP articles.

Rank	Resource	Type	# exist.	# occur.	First authors mentioning the LR	First corpora mentioning the LR	First Year	Last year
1	WordNet	NLPLexicon	4,203	29,079	Daniel A. Teibel, George A. Miller	hlt	1991	2015
2	Timit	NLPCorpus	3,005	11,853	Andrej Ljolje, Benjamin Chigier, David Goodine, David S. Pallett, Erik Urdang, Francine R. Chen, George R. Doddington, H-W Hon, Hong C. Leung, Hsiao-Wuen Hon, James R. Glass, Jan Robin Rohlicek, Jeff Shrager, Jeffrey N. Marcus, John Dowding, John F. Pitrelli, John S. Garofolo, Joseph H. Polifroni, Judith R. Spitz, Julia B. Hirschberg, Kai-Fu Lee, L. G. Miller, Mari Ostendorf, Mark Liberman, Mei-Yuh Hwang, Michael D. Riley, Michael S. Phillips, Robert Weide, Stephanie Seneff, Stephen E. Levinson, Vassilios V. Digalakis, Victor W. Zue	hlt, isca, taslp	1989	2015
3	Wikipedia	NLPCorpus	2,824	20,110	Ana Licuanan, J. H. Xu, Ralph M. Weischedel	trec	2003	2015
4	Penn Treebank	NLPCorpus	1,993	6,982	Beatrice Santorini, David M. Magerman, Eric Brill, Mitchell P. Marcus	hlt	1990	2015
5	Praat	NLPTool	1,245	2,544	Carlos Gussenhoven, Toni C. M. Rietveld	isca	1997	2015
6	SRI Language Modeling Toolkit	NLPTool	1,029	1,520	Dilek Z. Hakkani-Tür, Gökhan Tür, Kemal Oflazer	coling	2000	2015
7	Weka	NLPTool	957	1,609	Douglas A. Jones, Gregory M. Rusk	coling	2000	2015
8	Europarl	NLPCorpus	855	3,119	Daniel Marcu, Franz Josef Och, Grzegorz Kondrak, Kevin Knight, Philipp Koehn	acl, eacl, hlt, naacl	2003	2015
9	FrameNet	NLPLexicon	824	5,554	Beryl T. Sue Atkins, Charles J. Fillmore, Collin F. Baker, John B. Lowe, Susanne Gahl	acl, coling, lrec	1998	2015
10	GIZA++	NLPTool	758	1,582	David Yarowsky, Grace Ngai, Richard Wicentowski	hlt	2001	2015

Data

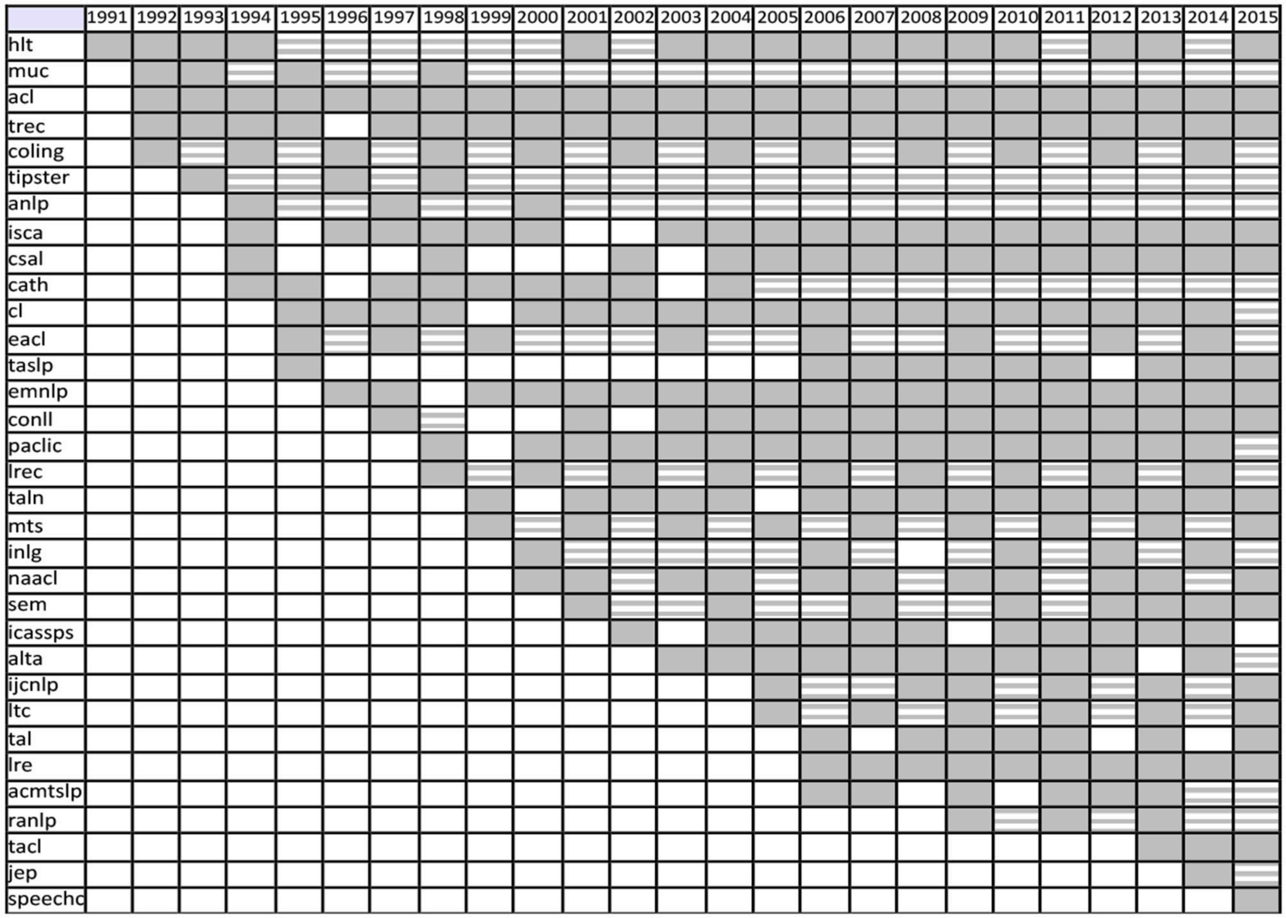
We should remind that we consider here the 67,937 documents coming from various conferences and journals which constitute a large part of the existing published articles in the field, apart from the workshop proceedings and the published books. Some documents are identical as they were published in joint conferences, but we must take them into account individually in order to study the flow of reuse across conferences and journals. The corpus follows the organization of the ACL Anthology with two parts in parallel. For each document, on one side, the metadata is recorded with the author names and the title under the form of a BibTex file. On the other side, the PDF document is recorded on disk in its original form. Each document is labeled with a unique identifier, for instance paper identified as number 1 at the LREC 2000 conference is named "lrec2000_1" and is reified as two files: "lrec2000_1.bib" and "lrec2000_1.pdf." Figures are not extracted because we are unable to compare images. See Francopoulo et al. (2015) for more details about the extraction process as well as the solutions for some tricky problems like joint conferences management or abstract/body/reference sections detection. The majority (90%) of the documents come from conferences, the rest coming from journals. The overall number of words is roughly 270M. The texts are in four languages: English, French, German, and Russian. The number of texts in German and Russian is <0.5%. They are detected automatically and are ignored. The texts in French are a little bit more numerous (3%), so they are kept with the same status as the English ones. This is not a problem as our tool is able to process English and French. The corpus is a collection of documents of a single technical domain which is NLP in the broad sense, and of course, some conferences are specialized in certain topics like written language processing, spoken language processing, including signal processing, information retrieval or machine translation. We also considered here the list of 48,894 authors.

Definitions

As the terminology is fuzzy and contradictory among the scientific literature, we needed first to define four important terms in order to avoid any misunderstanding (**Table 9**):

- The term "**self-reuse**" is used for a copy & paste when the source of the copy has an author who belongs to the group of authors of the text of the paste and when the source is cited.
- The term "**self-plagiarism**" is used for a copy & paste when the source of the copy has similarly an author who belongs to the group of authors of the text of the paste, but when the source is not cited.
- The term "**reuse**" is used for a copy & paste when the source of the copy has no author in the group of authors of the paste and when the source is cited.
- The term "**plagiarism**" is used for a copy & paste when the source of the copy has no author in the group of the paste and when the source is not cited.

Said in other words, the terms "self-reuse" and "reuse" qualify a situation with a proper source citation, on the contrary of "self-plagiarism" and "plagiarism." Let's note that in spite of the



	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
hlt	■	■	■	■	▤	▤	▤	▤	▤	▤	■	▤	■	■	■	■	■	■	■	■	▤	■	■	▤	■
muc		■	■	▤	■	▤	▤	■	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤
acl		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
trec		■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
coling		■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤
tipster			■	▤	▤	■	▤	■	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤
anlp				■	▤	▤	■	▤	▤	■	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤
isca				■		■	■	■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■
csal				■				■				■		■	■	■	■	■	■	■	■	■	■	■	■
cath				■	■		■	■	■	■	■	■		■	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤	▤
cl					■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▤
eacl					■	▤	■	▤	■	▤	▤	▤	■	▤	▤	■	▤	▤	■	▤	▤	■	▤	■	▤
taslp					■											■	■	■	■	■	■		■	■	■
emnlp						■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
conll							■	▤			■		■	■	■	■	■	■	■	■	■	■	■	■	■
paclic								■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▤
lrec								■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤
taln									■		■	■	■	■		■	■	■	■	■	■	■	■	■	■
mts									■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■	▤	■
inlg										■	▤	▤	▤	▤	▤	■	▤		▤	■	▤	■	▤	■	▤
naacl										■	■	▤	■	■	▤	■	■	▤	■	■	▤	■	■	▤	■
sem											■	▤	▤	■	▤	▤	■	▤	▤	■	▤	■	■	■	■
icassps												■		■	■	■	■	■		■	■	■	■	■	
alta													■	■	■	■	■	■	■	■	■	■		■	▤
ijcnlp															■	▤	▤	■	■	▤	■	▤	■	▤	■
ltc															■	▤	■	▤	■	▤	■	▤	■	▤	■
tal																■		■	■	■	■		■		■
lre																■	■	■	■	■	■	■	■	■	■
acmtslp																■	■		■		■	■	■	▤	▤
ranlp																			■	▤	■	▤	■	▤	▤
tacl																							■	■	■
jep																								■	▤
speechc																									■

FIGURE 21 | Propagation of the mention of the "Wordnet" resource in NLP4NLP[12] conferences and journals.

fact that the term "self-plagiarism" seems to be contradictory, we use this term because it is the usual habit within the community of the plagiarism detection. Some authors also use the term "recycling," for instance (HaCohen-Kerner et al., 2010).

Another point to clarify concerns the expression "source papers." As a convention, we call "focus" the corpus corresponding to the source which is studied. The whole NL4NLP collection is the "search space." We examine the copy & paste operations in both directions: we study the configuration with a source paper borrowing fragments of text from other papers of the NLP4NLP collection, in other words, a backward study, and we also study in the reverse direction the fragments of the source paper being borrowed by papers of the NLP4NLP collection, in other words, a forward study.

Algorithm for Computing Papers Similarity

Comparison of word sequences has proven to be an effective method for detection of copy & paste (Clough et al., 2002a) and in several occasions, this method won the PAN contest (Barron-Cedeno et al., 2010), so we will adopt this strategy. In our case, the corpus is first processed with the deep NLP parser TagParser (Francopoulo, 2008) to produce a Passage format (Vilnat et al., 2010) with lemma and part-of-speech (POS) indications.

The algorithm is as follows:

- For each document of the focus (the source corpus), all the sliding windows[13] of 7 lemmas (excluding punctuations) are built and recorded under the form of a character string key in an index locally to a document.
- An index gathering all these local indexes is built and is called the "focus index."
- For each document apart from the focus (i.e., outside the source corpus), all the sliding windows are built and **only the windows** contained in the focus index are recorded in an index locally to this document. This filtering operation is done to optimize the comparison phase, as there is no need to compare the windows out of the focus index.

[12] Hatched slots correspond to years where the conference didn't occur or the journal wasn't published.

[13] Also called "n-grams" in some NLP publications.

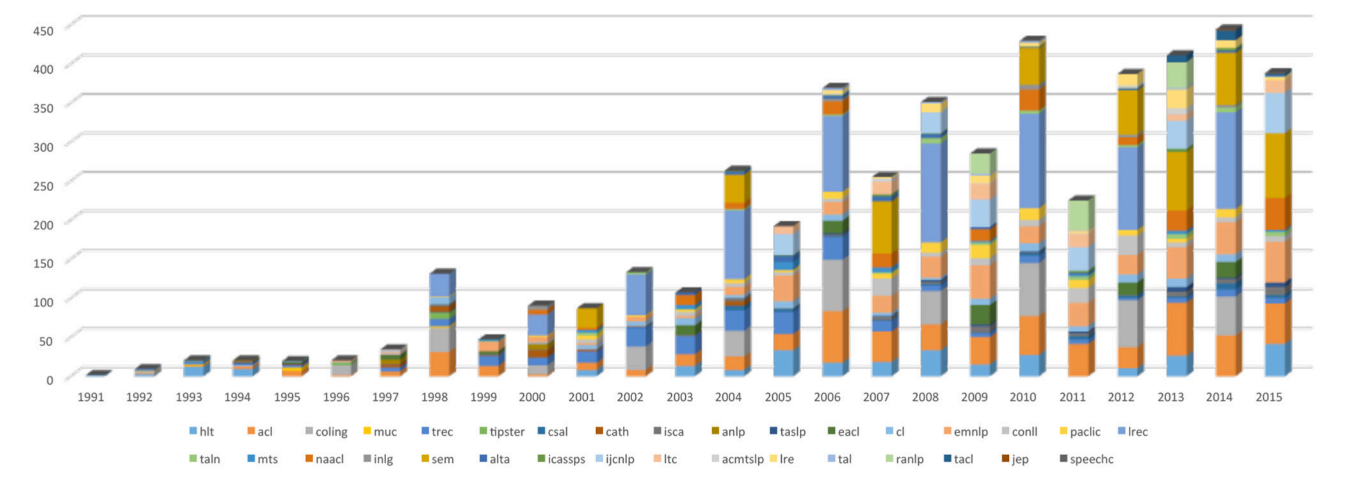


FIGURE 22 | Propagation of the mention of the "Wordnet" resource in NLP4NLP conferences and journals, including the number of mentions.

TABLE 8 | Language resources impact factor (data and tools).

Data	Impact factor	Tools	Impact factor
Wordnet	4203	Praat	1254
Timit	3005	SRI Language Modeling Toolkit	1029
Wikipedia	2824	Weka	957
Penn Treebank	1993	GIZA++	758
Europarl	855		
FrameNet	824		

- Then, the keys are compared to compute a similarity overlapping score (Lyon et al., 2001) between documents D1 and D2, with the Jaccard distance:

$$\mathbf{score(D1, D2)} = \mathbf{sharedwindows\#/union\#} \\ \mathbf{(D1windows, D2windows)}$$

- The pairs of documents D1/D2 are then filtered according to a threshold of 0.04 to retain only significant scoring situations.

In a first implementation, we compared the raw character strings with a segmentation based on space and punctuation. But, due to the fact that the input is the result of PDF formatting, the texts may contain variable caesura for line endings or some little textual variations. Our objective is to compare at a higher level than hyphen variation (there are different sorts of hyphens), caesura (the sequence X/-/endOfLine/Y needs to match an entry XY in the lexicon to distinguish from an hyphen binding a composition), upper/lower case variation, plural, orthographic variation ("normalise" vs. "normalize"), spellchecking (particularly useful when the PDF is an image and when the extraction is of low quality) and abbreviation ("NP" vs. "Noun Phrase" or "HMM" vs. "Hidden Markov Model"). Some rubbish sequence of characters (e.g., a series of hyphens) were also detected and cleaned.

TABLE 9 | Definition of terms.

	Source is quoted	Source is not quoted
At least one author in both papers	Self-reuse	Self-plagiarism
No author in common	Reuse	Plagiarism

Given that a parser takes all these variations and cleanings into account, we decided to apply a full linguistic parsing, as a second strategy. The syntactic structures and relations are ignored. Then a module for entity linking is called in order to bind different names referring to the same entity, a process often labeled as "entity linking" in the literature (Guo et al., 2011; Moro et al., 2014). This process is based on a Knowledge Base called "Global Atlas" (Francopoulo et al., 2013) which comprises the LRE Map (Calzolari et al., 2012). Thus, "British National Corpus" is considered as possibly abbreviated to "BNC," as well as less regular names like "ItalWordNet" possibly abbreviated to "IWN." Each entry of the Knowledge Base has a canonical form, possibly associated with different variants: the aim is to normalize into a canonical form to neutralize proper noun obfuscations based on variant substitutions. After this processing, only the sentences with at least a verb are considered.

We examined the differences between those two strategies concerning all types of copy & paste situations above the threshold, choosing the LREC source as the focus. The results are presented in **Table 10**, with the last column adding the two other columns without the duplicates produced by the couples of the same year.

The strategy based on linguistic processing provides more pairs (+158) and we examined these differences. Among these pairs, the vast majority (80%) concerns caesura: this is normal because most conferences demand a double column format, so the authors frequently use caesura to save place[14]. The other

[14] Concerning this specific problem, for instance, PACLIC and COLING which are one column formatted give much better extraction quality than LREC and ACL which are two columns formatted.

TABLE 10 | Comparison of the two strategies on the LREC corpus.

Strategy	Backward study document pairs#	Forward study document pairs#	Backward + forward document pairs# after duplicate pruning
1. Raw text	438	373	578
2. Linguistic processing (LP)	559	454	736
Difference (LP-raw)	121	81	158

differences (20%) are mainly caused by lexical variations and spellchecking. Thus, the results show that using raw texts gives a more "silent" system. The drawback is that the computation is much longer[15], but we think that it is worth the value. There are three parameters that had to be tuned: the window size, the distance function and the threshold. The main problem we had was that we did not have any gold standard to evaluate the quality specifically on our corpus and the burden to annotate a corpus is too heavy. We therefore decided to start from the parameters presented in the articles related to the PAN contest. We then computed the results, picked a random selection of pairs that we examined and tuned the parameters accordingly. All experiments were conducted with LREC as the focus and NLP4NLP as the search space.

In the PAN related articles, different **window** sizes are used. A window of five tokens is the most frequent one (Kasprzak and Brandejs, 2010), but our results shows that a lot of common sequences like "the linguistic unit is the" overload the pairwise score. After some trials, we decided to select a size of seven tokens.

Concerning the **distance** function, the Jaccard distance is frequently used but let's note that other formulas are applicable and documented in the literature. For instance, some authors use an approximation with the following formula: score (D1, D2) = shared windows# / min(D1 windows#, D2 windows#) (Clough and Stevenson, 2011), which is faster to compute, because there is no need to compute the union. Given that computation time is not a problem for us, we kept the most used function, which is the Jaccard distance.

Concerning the **threshold**, we tried thresholds of 0.03 and 0.04 and we compared the results. The last value gave more significant results, as it reduced noise, while still allowing to detect meaningful pairs of similar papers. We therefore considered as potential reused or plagiarized couples of papers all couples with a similarity score of 4% or more.

Categorization of the Results

After running the first trials, we discovered that using the Jaccard distance resulted in considering as similar a set of two papers, one of them being of small content. This may be the case for invited talks, for example, when the author only provides a short abstract. In this case, a simple acknowledgment to the same institution may produce a similarity score higher than the threshold. The same happens for some eldest papers when the OCR produced a truncated document. In order to solve this problem, we added a second threshold on the minimum number of shared windows that we set at 50 after considering the corresponding erroneous cases. We also found after those first trials erroneous results of the OCR for some eldest papers which resulted in files containing several papers, in full or in fragments, or where blanks were inserted after each individual character. We excluded those papers from the corpus being considered. Checking those results, we also mentioned several cases where the author was the same, but with a different spelling, or where references were properly quoted, but with a different wording, a different spelling (US vs. British English, for example) or an improper reference to the source. We had to manually correct those cases, and move the corresponding couples of papers in the right category (from reuse or plagiarism to self-reuse or self-plagiarism in the case of authors names, from plagiarism to reuse, in the case of references).

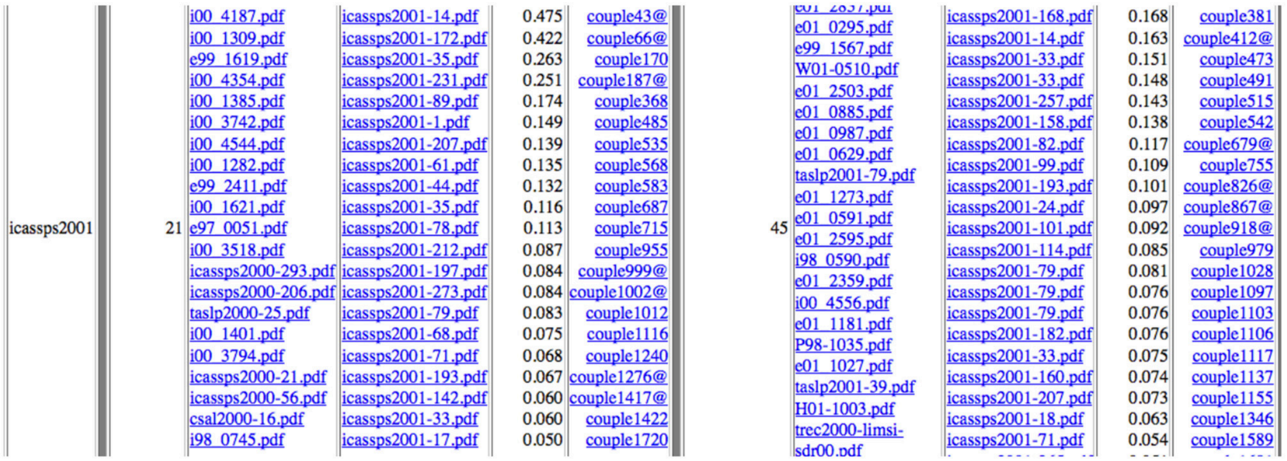
Our aim is to distinguish a copy & paste fragment associated with a citation compared to a fragment without any citation. To this end, we proceed with an approximation: we do not bind exactly the anchor in the text, but we parse the reference section and consider that, globally to the text, the document cites (or not) the other document. Due to the fact, that we have proper author identification for each document, the corpus forms a complex web of citations. We are thus able to distinguish self-reuse vs. self-plagiarism and reuse vs. plagiarism. We are in a situation slightly different from METER where the references are not linked. Let's recall that METER is the corpus usually involved in plagiarism detection competitions (Gaizauskas et al., 2001; Clough et al., 2002b).

Given the fact that some papers and drafts of papers can circulate among researchers before the official published date, it is impossible to verify exactly when a document is issued; moreover we do not have any more detailed time indication than the year, as we don't know the date of submission. This is why we also consider the same year within the comparisons. In this case, it is difficult to determine which are the borrowing and borrowed papers, and in some cases they may even have been written simultaneously. However, if one paper cites the second one, while it is not cited by the second one, it may serve as a sign to consider it as the borrowing paper.

The program computes a detailed result for each individual publication as an HTML page where all similar pairs of documents are listed with their similarity score, with the common fragments displayed as red highlighted snippets and HTML links back to the original 67,937 documents[16]. For each of the 4 categories (Self-reuse, Self-Plagiarism, Reuse and Plagiarism), the program produces the list of couples of "similar" papers according to our criteria, with their similarity score, identification of the common parts and indication of the same authors list or title (**Figures 23–25**).

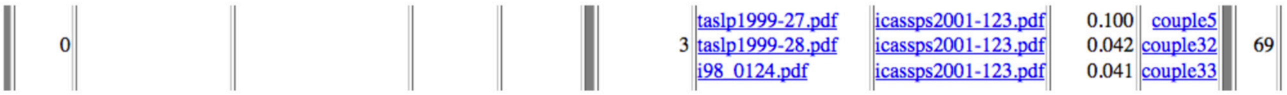
[15] It takes 25 h instead of 3 h on a mid-range mono-processor Xeon E3-1270 V2 with 32G of RAM.

[16] But the space limitations do not allow to present these results in lengthy details. Furthermore, we do not want to display personal results.



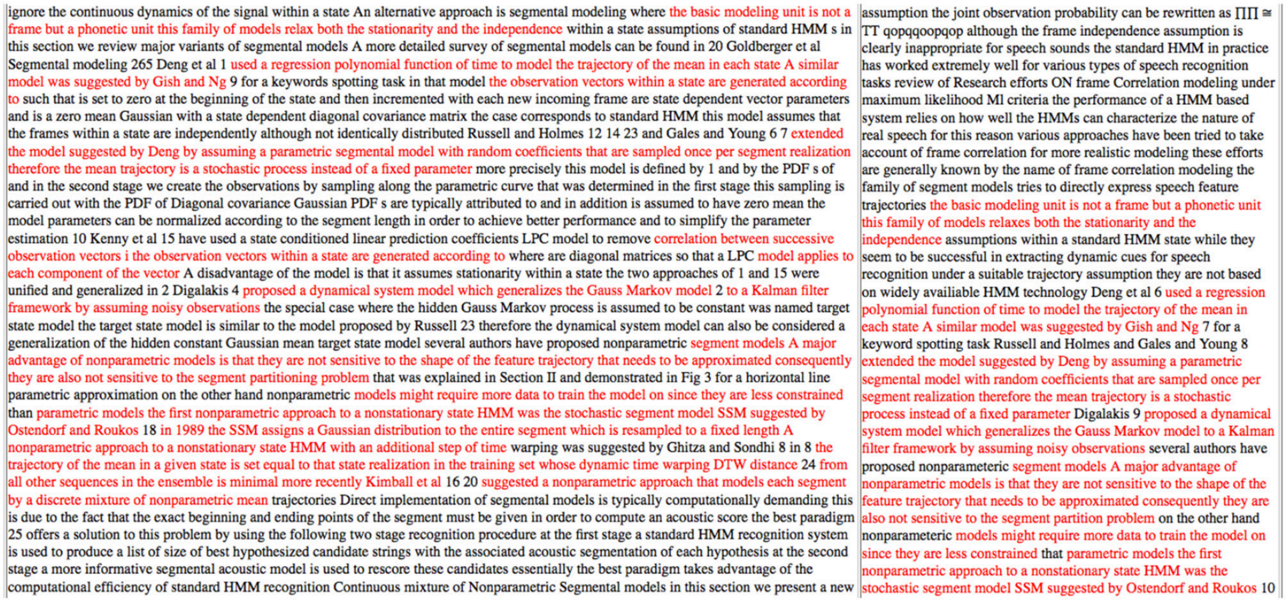
		i00 4187.pdf	icassps2001-14.pdf	0.475	couple43@		e01 2837.pdf	icassps2001-168.pdf	0.168	couple381
		i00 1309.pdf	icassps2001-172.pdf	0.422	couple66@		e01 0295.pdf	icassps2001-14.pdf	0.163	couple412@
		e99 1619.pdf	icassps2001-35.pdf	0.263	couple170		e99 1567.pdf	icassps2001-33.pdf	0.151	couple473
		i00 4354.pdf	icassps2001-231.pdf	0.251	couple187@		W01-0510.pdf	icassps2001-33.pdf	0.148	couple491
		i00 1385.pdf	icassps2001-89.pdf	0.174	couple368		e01 2503.pdf	icassps2001-257.pdf	0.143	couple515
		i00 3742.pdf	icassps2001-1.pdf	0.149	couple485		e01 0885.pdf	icassps2001-158.pdf	0.138	couple542
		i00 4544.pdf	icassps2001-207.pdf	0.139	couple535		e01 0987.pdf	icassps2001-82.pdf	0.117	couple679@
		i00 1282.pdf	icassps2001-61.pdf	0.135	couple568		e01 0629.pdf	icassps2001-99.pdf	0.109	couple755
		e99 2411.pdf	icassps2001-44.pdf	0.132	couple583		taslp2001-79.pdf	icassps2001-193.pdf	0.101	couple826@
		i00 1621.pdf	icassps2001-35.pdf	0.116	couple687		e01 1273.pdf	icassps2001-24.pdf	0.097	couple867@
icassps2001	21	e97 0051.pdf	icassps2001-78.pdf	0.113	couple715	45	e01 0591.pdf	icassps2001-101.pdf	0.092	couple918@
		i00 3518.pdf	icassps2001-212.pdf	0.087	couple955		e01 2595.pdf	icassps2001-114.pdf	0.085	couple979
		icassps2000-293.pdf	icassps2001-197.pdf	0.084	couple999@		i98 0590.pdf	icassps2001-79.pdf	0.081	couple1028
		icassps2000-206.pdf	icassps2001-273.pdf	0.084	couple1002@		e01 2359.pdf	icassps2001-79.pdf	0.076	couple1097
		taslp2000-25.pdf	icassps2001-79.pdf	0.083	couple1012		i00 4556.pdf	icassps2001-79.pdf	0.076	couple1103
		i00 1401.pdf	icassps2001-68.pdf	0.075	couple1116		e01 1181.pdf	icassps2001-182.pdf	0.076	couple1106
		i00 3794.pdf	icassps2001-71.pdf	0.068	couple1240		P98-1035.pdf	icassps2001-33.pdf	0.075	couple1117
		icassps2000-21.pdf	icassps2001-193.pdf	0.067	couple1276@		e01 1027.pdf	icassps2001-160.pdf	0.074	couple1137
		icassps2000-56.pdf	icassps2001-142.pdf	0.060	couple1417@		taslp2001-39.pdf	icassps2001-207.pdf	0.073	couple1155
		csal2000-16.pdf	icassps2001-33.pdf	0.060	couple1422		H01-1003.pdf	icassps2001-18.pdf	0.063	couple1346
		i98 0745.pdf	icassps2001-17.pdf	0.050	couple1720		trec2000-limsi-sdr00.pdf	icassps2001-71.pdf	0.054	couple1589

FIGURE 23 | Example of ICASSP 2001 Speech papers self-reusing (**left**: 21 cases identified) and self-plagiarizing (**right**: 45 cases identified) other papers with similarity scores (@ following the couple number indicates that the two papers have the same full list of authors).



0					3	taslp1999-27.pdf	icassps2001-123.pdf	0.100	couple5	69
						taslp1999-28.pdf	icassps2001-123.pdf	0.042	couple32	
						i98 0124.pdf	icassps2001-123.pdf	0.041	couple33	

FIGURE 24 | Example of ICASSP 2001 Speech papers reusing (**left**: no case identified) and plagiarizing (**right**: 3 cases identified) other papers with similarity scores.



ignore the continuous dynamics of the signal within a state An alternative approach is segmental modeling where the basic modeling unit is not a frame but a phonetic unit this family of models relax both the stationarity and the independence within a state assumptions of standard HMM s in this section we review major variants of segmental models A more detailed survey of segmental models can be found in 20 Goldberger et al Segmental modeling 265 Deng et al 1 used a regression polynomial function of time to model the trajectory of the mean in each state A similar model was suggested by Gish and Ng 9 for a keywords spotting task in that model the observation vectors within a state are generated according to such that is set to zero at the beginning of the state and then incremented with each new incoming frame are state dependent vector parameters and is a zero mean Gaussian with a state dependent diagonal covariance matrix the case corresponds to standard HMM this model assumes that the frames within a state are independently although not identically distributed Russell and Holmes 12 14 23 and Gales and Young 6 7 extended the model suggested by Deng by assuming a parametric segmental model with random coefficients that are sampled once per segment realization therefore the mean trajectory is a stochastic process instead of a fixed parameter more precisely this model is defined by 1 and by the PDF s of and in the second stage we create the observations by sampling along the parametric curve that was determined in the first stage this sampling is carried out with the PDF of Diagonal covariance Gaussian PDF s are typically attributed to and in addition is assumed to have zero mean the model parameters can be normalized according to the segment length in order to achieve better performance and to simplify the parameter estimation 10 Kenny et al 15 have used a state conditioned linear prediction coefficients LPC model to remove correlation between successive observation vectors i the observation vectors within a state are generated according to where are diagonal matrices so that a LPC model applies to each component of the vector A disadvantage of the model is that it assumes stationarity within a state the two approaches of 1 and 15 were unified and generalized in 2 Digalakis 4 proposed a dynamical system model which generalizes the Gauss Markov model 2 to a Kalman filter framework by assuming noisy observations the special case where the hidden Gauss Markov process is assumed to be constant was named target state model the target state model is similar to the model proposed by Russell 23 therefore the dynamical system model can also be considered a generalization of the hidden constant Gaussian mean target state model several authors have proposed nonparametric segment models A major advantage of nonparametric models is that they are not sensitive to the shape of the feature trajectory that needs to be approximated consequently they are also not sensitive to the segment partitioning problem that was explained in Section II and demonstrated in Fig 3 for a horizontal line parametric approximation on the other hand nonparametric models might require more data to train the model on since they are less constrained than parametric models the first nonparametric approach to a nonstationary state HMM was the stochastic segment model SSM suggested by Ostendorf and Roukos 18 in 1989 the SSM assigns a Gaussian distribution to the entire segment which is resampled to a fixed length A nonparametric approach to a nonstationary state HMM with an additional step of time warping was suggested by Ghitza and Sondhi 8 in 8 the trajectory of the mean in a given state is set equal to that state realization in the training set whose dynamic time warping DTW distance 24 from all other sequences in the ensemble is minimal more recently Kimball et al 16 20 suggested a nonparametric approach that models each segment by a discrete mixture of nonparametric mean trajectories Direct implementation of segmental models is typically computationally demanding this is due to the fact that the exact beginning and ending points of the segment must be given in order to compute an acoustic score the best paradigm 25 offers a solution to this problem by using the following two stage recognition procedure at the first stage a standard HMM recognition system is used to produce a list of size of best hypothesized candidate strings with the associated acoustic segmentation of each hypothesis at the second stage a more informative segmental acoustic model is used to rescore these candidates essentially the best paradigm takes advantage of the computational efficiency of standard HMM recognition Continuous mixture of Nonparametric Segmental models in this section we present a new

assumption the joint observation probability can be rewritten as ∏∏ ≅ TT qopqqoopqop although the frame independence assumption is clearly inappropriate for speech sounds the standard HMM in practice has worked extremely well for various types of speech recognition tasks review of Research efforts ON frame Correlation modeling under maximum likelihood Ml criteria the performance of a HMM based system relies on how well the HMMs can characterize the nature of real speech for this reason various approaches have been tried to take account of frame correlation for more realistic modeling these efforts are generally known by the name of frame correlation modeling the family of segment models tries to directly express speech feature trajectories the basic modeling unit is not a frame but a phonetic unit this family of models relaxes both the stationarity and the independence assumptions within a standard HMM state while they seem to be successful in extracting dynamic cues for speech recognition under a suitable trajectory assumption they are not based on widely availiable HMM technology Deng et al 6 used a regression polynomial function of time to model the trajectory of the mean in each state A similar model was suggested by Gish and Ng 7 for a keyword spotting task Russell and Holmes and Gales and Young 8 extended the model suggested by Deng by assuming a parametric segmental model with random coefficients that are sampled once per segment realization therefore the mean trajectory is a stochastic process instead of a fixed parameter Digalakis 9 proposed a dynamical system model which generalizes the Gauss Markov model to a Kalman filter framework by assuming noisy observations several authors have proposed nonparameteric segment models A major advantage of nonparametric models is that they are not sensitive to the shape of the feature trajectory that needs to be approximated consequently they are also not sensitive to the segment partition problem on the other hand nonparameteric models might require more data to train the model on since they are less constrained that parametric models the first nonparametric approach to a nonstationary state HMM was the stochastic segment model SSM suggested by Ostendorf and Roukos 10

FIGURE 25 | Example in ICASSP 2001 of common fragments (marked in red) for couple 5 articles showing a global similarity score of 0.10 (10%).

The program produces also global results in the form of matrices (**Tables 11**, **12**) for each of the four categories (Self-reuse, Self-Plagiarism, Reuse, and Plagiarism) displaying the number of papers that are similar in each couple of the 34 sources, in the forward and backward directions (using sources are on the X axis, while used sources are on the Y axis. The total of used and using papers, and the difference between those totals, are also presented, while the 5 top using or used sources are indicated.

TABLE 11 | Self-reuse and Self-Plagiarism Matrix, with indication in green of the 7 most using and used sources, and of the ones with significant differences between used and using.

Using Used	acl	acmtslp	alta	anlp	cath	cl	coling	conll	csal	eacl	emnlp	hlt	icassps	ijcnlp	inlg	isca	jep	lre	lrec	ltc	modulad	mts	muc	naacl	paclic	ranlp	sem	speechc	tacl	tal	taln	taslp	tipster	trec	Total used	Total using	Difference	
acl	22	8	1	4	8	136	78	25	31	22	83	85	29	31	7	48	0	20	71	4	0	19	1	51	8	5	26	1	2	0	0	24	4	9	863	625	238	acl
acmtslp	1	0	0	0	0	0	0	0	2	0	0	2	3	2	0	6	0	1	1	0	0	0	0	2	0	0	1	0	1	0	0	2	0	0	24	93	−69	acmtslp
alta	3	0	2	0	0	1	5	0	1	2	5	0	0	1	0	4	0	0	4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	4	33	14	19	alta
anlp	7	0	0	1	3	5	8	1	1	2	1	4	0	0	0	1	0	0	5	0	0	1	0	2	1	0	0	0	0	0	0	0	2	5	50	50	0	anlp
cath	1	0	0	1	7	2	0	0	0	1	0	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	18	50	−32	cath
cl	9	0	0	4	3	0	4	0	2	4	3	1	0	0	0	0	0	2	5	0	0	0	0	0	1	0	4	0	0	0	0	0	0	0	42	433	−391	cl
coling	74	10	3	8	7	62	19	24	17	15	43	49	8	24	7	42	0	14	90	4	0	9	2	33	12	5	25	3	0	0	0	12	6	5	632	500	132	coling
conll	26	1	1	1	1	20	18	8	5	6	16	11	2	14	2	2	0	2	10	1	0	3	0	7	0	5	13	0	1	0	0	3	0	0	179	151	28	conll
csal	3	0	0	0	0	4	4	2	7	0	3	2	20	1	0	35	0	2	7	0	0	0	0	0	0	0	2	6	0	0	0	13	0	0	111	643	−532	csal
eacl	16	2	0	2	5	31	12	6	3	1	8	13	3	1	2	9	0	0	21	1	0	1	0	13	1	1	4	0	0	0	0	5	0	1	162	130	32	eacl
emnlp	103	2	2	1	2	44	52	26	18	9	16	30	14	47	1	27	0	5	29	0	0	7	0	22	2	1	19	0	3	0	0	20	1	5	508	355	153	emnlp
hlt	83	12	0	5	3	48	48	11	42	14	33	22	29	30	2	104	0	4	26	1	0	13	2	6	1	0	9	8	0	0	0	25	7	19	607	476	131	hlt
icassps	16	5	0	0	0	3	4	1	130	4	7	21	262	2	0	1005	0	0	19	0	0	2	0	14	2	0	0	65	0	0	0	746	0	3	2311	2160	151	icassps
ijcnlp	27	6	1	0	0	3	29	10	7	2	34	18	2	4	3	7	0	5	19	3	0	9	0	13	4	8	3	0	0	0	0	4	0	1	222	237	−15	ijcnlp
inlg	7	0	0	1	1	6	5	2	0	3	1	3	0	1	2	4	0	1	6	0	0	1	0	4	0	0	0	0	0	0	0	1	0	0	49	35	14	inlg
isca	56	23	0	2	0	13	45	0	317	10	25	116	1531	10	4	879	0	10	133	19	0	12	0	38	6	0	1	233	0	0	0	669	0	5	4157	2460	1697	isca
jep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	16	18	−2	jep
lre	2	1	0	0	0	2	3	0	0	0	0	1	0	0	0	2	0	2	6	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	22	146	−124	lre
lrec	58	3	0	2	6	16	80	6	13	15	16	17	16	10	2	72	0	52	67	12	0	6	0	11	11	4	12	5	2	0	0	6	1	3	524	660	−136	lrec
ltc	4	0	0	0	0	0	0	0	0	0	0	0	0	2	0	15	0	1	35	10	0	2	0	0	6	6	1	4	0	0	0	0	0	0	86	71	15	ltc
modulad	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	modulad
mts	13	0	0	0	0	2	9	2	0	2	9	10	3	9	0	9	0	2	20	2	0	8	0	8	5	2	1	1	0	0	0	2	0	0	119	109	10	mts
muc	2	0	0	2	0	2	3	0	0	1	0	7	0	0	0	0	0	0	0	0	0	0	10	1	0	0	0	0	0	0	0	0	18	1	47	28	19	muc
naacl	46	10	0	2	1	24	30	7	12	11	22	5	15	22	3	30	0	3	16	1	0	9	0	3	0	0	9	1	0	0	0	8	0	3	293	251	42	naacl
paclic	4	0	0	0	1	0	12	1	1	1	1	0	2	8	0	3	0	5	18	7	0	3	0	0	21	7	1	0	0	0	0	1	0	0	97	85	12	paclic
ranlp	3	2	0	0	0	0	2	4	4	2	2	1	0	7	0	0	0	2	19	5	0	2	0	1	2	4	2	1	0	0	0	0	0	1	66	54	12	ranlp
sem	25	2	0	0	0	7	16	14	4	1	12	12	0	8	0	0	0	13	12	1	0	1	0	8	1	4	53	0	0	0	0	0	0	1	195	188	7	sem
speechc	0	0	0	0	0	1	1	0	11	0	0	4	17	0	0	48	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	17	0	0	102	344	−242	speechc
tacl	1	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	7	9	−2	tacl
tal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	18	59	−41	tal
taln	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	53	9	0	0	0	65	22	43	taln
taslp	0	5	0	0	0	0	1	1	13	0	1	4	197	0	0	103	0	0	2	0	0	1	0	2	0	0	0	15	0	0	0	49	0	0	394	1610	−1216	taslp
tipster	3	0	0	3	0	0	6	0	0	0	1	5	0	0	0	0	0	0	2	0	0	0	13	1	0	0	0	0	0	0	0	0	2	7	43	65	−22	tipster
trec	10	0	4	11	2	1	6	0	2	2	11	32	7	3	0	5	0	0	10	0	0	0	0	10	0	1	1	0	0	0	0	2	24	287	431	362	69	trec
Total using	625	93	14	50	50	433	500	151	643	130	355	476	2160	237	35	2460	18	146	660	71	0	109	28	251	85	54	188	344	9	59	22	1610	65	362	12493	12493	0	

TABLE 12 | Reuse and Plagiarism Matrix, with indication in green of the 7 most using and used sources, and of the ones with significant differences between used and using.

Using Used	acl	acmtslp	alta	anlp	cath	cl	coling	conll	csal	eacl	emnlp	hlt	icassps	ijcnlp	inlg	isca	jep	lre	lrec	ltc	modulad	mts	muc	naacl	paclic	ranlp	sem	speechc	tacl	tal	taln	taslp	tipster	trec	Total used	Total using	Difference	
acl	1	0	0	0	1	1	2	2	0	0	4	3	0	3	0	2	0	0	1	1	0	0	1	1	1	1	3	0	0	0	0	0	0	0	28	7	21	acl
acmtslp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	acmtslp
alta	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	alta
anlp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	anlp
cath	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	−2	cath
cl	0	0	0	0	0	0	1	0	0	0	1	1	0	1	0	0	0	0	4	0	0	1	0	1	2	0	0	0	0	0	0	0	0	0	12	5	7	cl
coling	0	0	0	0	1	0	0	0	0	0	0	2	1	1	0	2	0	0	2	0	0	0	1	1	1	0	2	0	0	0	0	1	0	0	15	7	8	coling
conll	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	5	−2	conll
csal	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	3	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	7	6	1	csal
eacl	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	eacl
emnlp	0	0	0	0	0	2	0	2	0	1	1	2	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	2	13	15	−2	emnlp
hlt	2	0	0	0	0	1	0	1	1	0	2	1	1	1	0	2	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	2	17	17	0	hlt
icassps	0	0	0	0	0	0	0	0	1	0	1	2	3	0	0	32	0	0	0	0	0	0	0	2	0	0	0	2	0	0	0	5	0	0	48	37	11	icassps
ijcnlp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	2	9	−7	ijcnlp
inlg	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	inlg
isca	0	0	0	0	0	1	1	0	1	0	0	1	18	1	0	7	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	3	0	0	36	70	−34	isca
jep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	jep
lre	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	−1	lre
lrec	0	0	0	0	0	0	0	0	1	0	2	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	8	8	0	lrec
ltc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	−4	ltc
modulad	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	modulad
mts	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	4	3	1	mts
muc	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3	3	0	muc
naacl	1	0	0	0	0	0	0	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	9	10	−1	naacl
paclic	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	10	−8	paclic
ranlp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	−3	ranlp
sem	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	7	−4	sem
speechc	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	4	5	−1	speechc
tacl	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	tacl
tal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	tal
taln	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	taln
taslp	0	0	0	0	0	0	0	0	1	1	0	0	10	0	0	16	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	30	10	20	taslp
tipster	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	2	0	tipster
trec	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	13	13	0	trec
Total using	7	0	0	0	2	5	7	5	6	2	15	17	37	9	0	70	0	1	8	4	0	3	3	10	10	3	7	5	0	0	0	10	2	13	261	261	0	

Self-Reuse and Self-Plagiarism

Table 11 provides the results for self-reuse (authors reusing their own text while quoting the source paper) and self-plagiarism (authors reusing their own text without quoting the source paper). As we see, it is a rather frequent phenomenon, with a total of 12,493 documents, i.e., 18% of the 67,937 documents! In 61% of the cases (7,650 self-plagiarisms over 12,493), the authors even do not quote the source paper. We found that 205 papers have exactly the same title, and that 130 papers have both the same title and the same list of authors! Also 3,560 papers have exactly the same list of authors.

We see that the most used sources are the large conferences: ISCA, IEEE-ICASSP, ACL, COLING, HLT, EMNLP, and LREC. The most using sources are not only those large conferences, but also the journals: IEEE-*Transactions on Acoustics, Speech and Language Processing* (and its various avatars) (TASLP), *Computer Speech and Language* (CSAL), *Computational Linguistics* (CL), and Speech Com. If we consider the balance between the using and the used sources, we see the flow of papers from conferences to journals. The largest flows of self-reuse and self-plagiarism concern ISCA and ICASSP (in both directions, but especially from ISCA to ICASSP), ICASSP and ISCA to TASLP (also in the reverse direction) and to CSAL, ISCA to Speech Com, ACL to *Computational Linguistics*, ISCA to LREC and EMNLP to ACL.

If we want to study the influence a given conference (or journal) has on another, we must however recall that these figures are raw figures in terms of number of documents, and we must not forget that some conferences (or journals) are much bigger than others, for instance ISCA is a conference with more than 18K documents compared to LRE which is a journal with only 308 documents. If we relate the number of published papers that reuse another paper to the total number of published papers, we may see that 17% of the LRE papers (52 over 308) use content coming from the LREC conferences, without quoting them in 66% of the cases. Also the frequency of the conferences (annual or biennial) and the calendar (date of the conference and of the submission deadline) may influence the flow of papers between the sources.

The similarity scores range from 4 to 100% (**Figure 26**). If we consider the 65,003 different documents, we see that 11,372 couples of documents (18% of the total number of documents) have a similarity score superior or equal to 4%, about 4,560 couples (1.3% of the total) have a similarity score equal or superior to 10% and about 860 (6.6% of the total number) a similarity score superior or equal to 30%. The ones with the largest similarity score correspond to the same paper published by the same author at two successive TREC conferences. The next two couples both correspond to very similar papers published by the same authors first at an ISCA conference, then at ICASSP on the following year. We also found cases of republishing the corrigendum of a previously published paper or of republishing a paper with a small difference in the title and one missing author in the authors' list. In one case, the same research center is described by the same author in two different conferences with an overlapping of 90%. In another case, the difference of the two papers is primarily in the name of the systems being presented, funded by the same project agency in two different contracts, while the description has a 45% overlap!

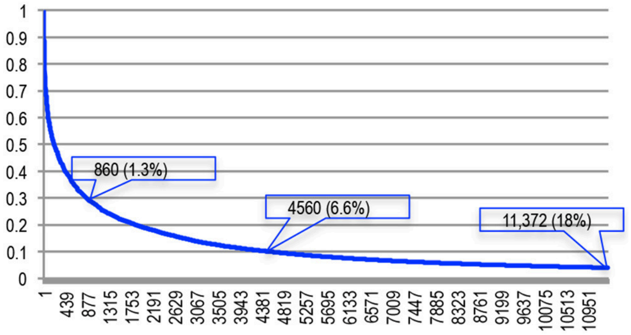


FIGURE 26 | Similarity scores of the couples detected as self-reuse/self-plagiarism.

Reuse and Plagiarism

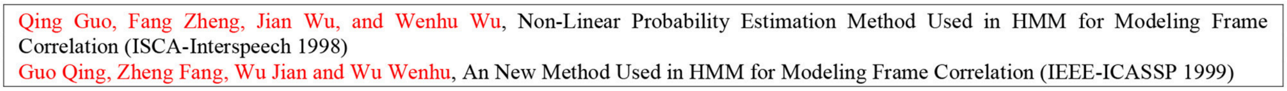
Table 12 provides the results for reuse (authors reusing fragments of the texts of other authors while quoting the source paper) and plagiarism (authors reusing fragments of the texts of other authors without quoting the source paper). As we see, there are very few cases altogether. Only 261 papers (i.e., <0.4% of the 67,937 documents) reuse a fragment of papers written by other authors. In 60% of the cases (146 over 261), the authors do not quote the source paper, but these possible cases of plagiarism only represent 0.2% of the total number of papers. Given those small numbers, we were able to conduct a complete manual checking of those couples.

Among the couple papers placed in the "Reuse" category, it appeared that several have a least one author in common, but with a somehow different spelling and should therefore be placed in the "Self-reuse" category. Among the couples of papers placed in the "Plagiarism" category, some have a least one author in common, but with a somehow different spelling (see **Figure 27**) and should therefore be placed in the "Self-plagiarism" category.

Others correctly quote the source paper, but with variants in the spelling of the authors' names (**Figure 28**), of the paper's title (**Figure 29**) or of the conference or journal. Those variants may also be due to the style guidelines of the conference or journal. We also find the cases of mentioning but forgetting to place the source paper in the references. Those papers should therefore be placed in the "Reuse" category.

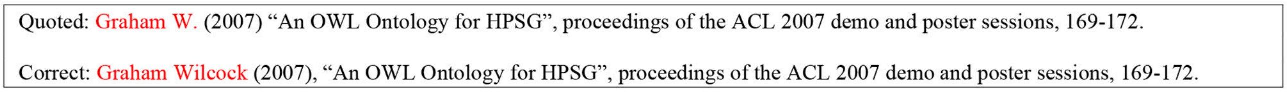
It therefore finally resulted in 104 cases of "reuse" and 116 possible cases of plagiarism (0.17% of the papers) that we studied more closely. We found the following explanations:

- The paper cites another reference from the same authors of the source paper (typically a previous reference, or a paper published in a Journal) (45 cases).
- Both papers use extracts of a third paper that they both cite (31 cases).



Qing Guo, Fang Zheng, Jian Wu, and Wenhu Wu, Non-Linear Probability Estimation Method Used in HMM for Modeling Frame Correlation (ISCA-Interspeech 1998)
Guo Qing, Zheng Fang, Wu Jian and Wu Wenhu, An New Method Used in HMM for Modeling Frame Correlation (IEEE-ICASSP 1999)

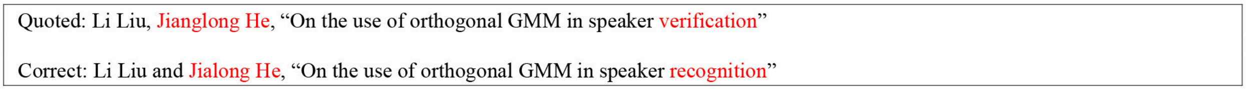
FIGURE 27 | Variants in spelling authors' names.



Quoted: Graham W. (2007) "An OWL Ontology for HPSG", proceedings of the ACL 2007 demo and poster sessions, 169-172.

Correct: Graham Wilcock (2007), "An OWL Ontology for HPSG", proceedings of the ACL 2007 demo and poster sessions, 169-172.

FIGURE 28 | Variants in spelling authors' names in reference.



Quoted: Li Liu, Jianglong He, "On the use of orthogonal GMM in speaker verification"

Correct: Li Liu and Jialong He, "On the use of orthogonal GMM in speaker recognition"

FIGURE 29 | Variants in spelling authors' names and papers titles in reference.

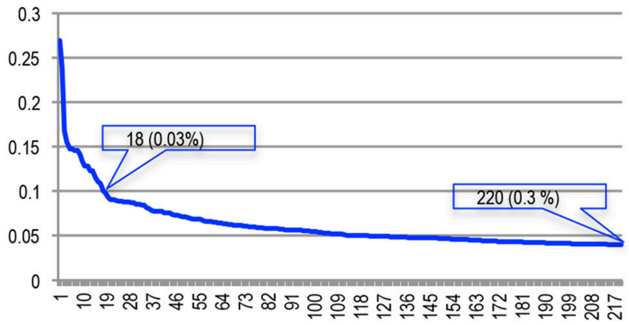


FIGURE 30 | Similarity scores of the couples detected as reuse/plagiarism.

- The authors of the two papers are different, but from the same laboratory (typically in industrial laboratories or funding agencies) (11 cases).
- The authors previously co-authored papers (typically as supervisor and Ph.D. student or postdoc) but are now in a different laboratory (11 cases).
- The authors of the papers are different, but collaborated in the same project which is presented in the two papers (2 cases).
- The two papers present the same short example, result, or definition coming from another event (13 cases).

If we exclude those 113 cases, only 3 cases of possible plagiarism remain that correspond to the same paper which appears as a patchwork of 3 other papers, while sharing several references with them, the highest similarity score being only 10%, with a shared window of 200 tokens (see **Figures 24**, **25**).

Here, the similarity scores range from 4 to 27% (**Figure 30**). If we consider the 65,003 different documents, we see that 220 couples of documents (0.3% of the total number of documents) have a similarity score superior or equal to 4%, and only 18 couples (0.03% of the total number) have a similarity score equal or higher than 10%. For example, the couple showing the highest similarity score comprises a paper published at Interspeech in 2013 and a paper published at ICASSP in 2015 which both describe the *Kaldi system* using the words of the initial paper published at the IEEE ASRU workshop in 2011, that they both properly quote.

Time Delay Between Publication and Reuse

We now consider the duration between the publication of a paper and its reuse (in all 4 categories) in another publication (**Table 13**). It appears that 38% of the similar papers were published on the same year, 71% within the next year, 83% over 2 years, and 93% over 3 years (**Figures 31**, **32**). Only 7% reuse material from an earlier period. The average duration is 1.22 years. Thirty percent of the similar papers published on the same year concern the couple of conferences ISCA-ICASSP.

If we consider the reuse of conference papers in journal papers (**Figures 33**, **34**), we observe a similar time schedule, with a delay of one year: 12% of the reused papers were published on the same year, 41% within the next year, 68% over 2 years, 85% over 3 years and 93% over 4 years. Only 7% reuse material from an earlier period. The average duration is 2.07 years.

Legal and Ethical Limits

The first obvious ascertainment is that self-reusing is much more frequent than reusing the content of others. With a comparable threshold of 0.04, when we consider the total of the two directions, there are 11,372 self-reuse and self-plagiarism detected pairs, compared with 104 reuse and 116 plagiarism detected pairs. Globally, the source papers are quoted only in 40%

TABLE 13 | Number of papers reusing and number of papers being reused over the years (1965–2015), with indication in green of the years with the largest number of reused and reusing papers.

Year	1965	1967	1973	1975	1976	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Total
1965	0	0	0	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
1967	0	0	0	0	0	0	0	1	0	1	2	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
1973	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1975	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1976	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1979	0	0	0	0	0	0	0	1	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
1980	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
1981	0	0	0	0	0	0	0	0	5	3	0	1	4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
1982	0	0	0	0	0	0	0	0	3	9	10	4	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39
1983	0	0	0	0	0	0	0	0	0	4	5	2	16	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28
1984	0	0	0	0	0	0	0	0	0	0	6	4	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20
1985	0	0	0	0	0	0	0	0	0	0	0	6	6	3	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17
1986	0	0	0	0	0	0	0	0	0	0	0	0	17	6	5	2	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32
1987	0	0	0	0	0	0	0	0	0	0	0	0	0	6	9	6	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25
1988	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	10	5	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27
1989	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	43	9	7	7	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	94
1990	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	42	27	17	11	3	0	0	1	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	104
1991	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	35	35	23	5	0	4	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	104
1992	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	62	46	30	10	6	3	1	2	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	163
1993	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	82	55	22	15	8	2	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	187
1994	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	70	45	33	11	7	3	2	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	174
1995	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	70	93	33	16	6	2	1	1	0	0	1	0	1	2	0	0	0	0	0	0	0	226
1996	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	103	101	37	31	11	5	2	3	2	0	1	1	1	0	0	0	0	0	0	0	298
1997	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	78	96	50	29	8	11	4	0	0	0	1	0	0	0	0	0	0	0	0	277
1998	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	189	126	82	40	20	12	9	5	0	0	1	0	0	0	1	0	0	0	485
1999	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	77	129	57	30	10	10	4	2	2	0	1	3	0	0	0	1	0	326
2000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	169	130	70	35	19	6	8	2	3	1	3	2	1	1	0	0	450
2001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	119	90	53	30	14	11	4	4	1	1	0	0	0	0	0	327
2002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	91	98	50	24	17	10	4	3	1	0	0	0	1	1	300
2003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	223	159	59	55	25	4	5	0	1	1	0	0	1	533
2004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	281	200	121	76	18	17	8	1	1	0	2	1	726
2005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	128	210	100	33	28	9	9	7	3	1	1	529
2006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	217	195	87	57	33	14	8	6	2	1	620
2007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	210	206	86	44	25	20	14	4	2	611
2008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	281	201	103	63	35	21	12	1	717
2009	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	209	247	113	60	33	20	5	687
2010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	302	236	122	93	39	12	804
2011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	291	237	129	70	21	748
2012	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	192	210	105	37	544
2013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	265	278	98	641
2014	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	327	230	557
2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	220	220
Total	0	0	0	0	1	0	0	2	11	19	25	17	73	15	23	46	93	74	123	171	163	148	256	236	351	297	426	362	318	438	562	441	642	627	644	610	754	755	685	775	862	631	11676

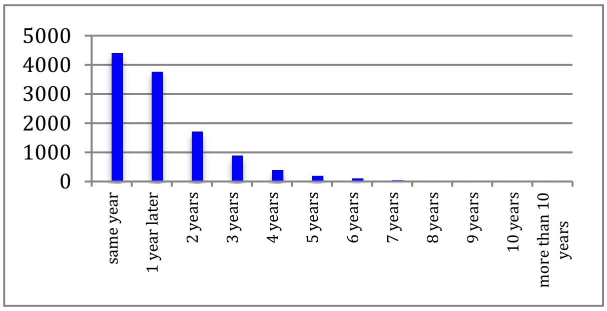


FIGURE 31 | Time delay between publication and reuse.

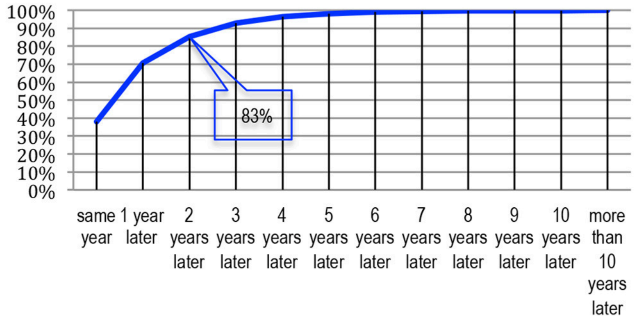


FIGURE 32 | Time delay between publication and reuse (in %).

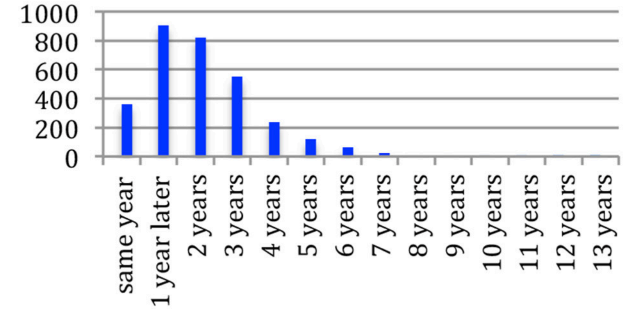


FIGURE 33 | Time delay between publication in conferences and reuse in journals.

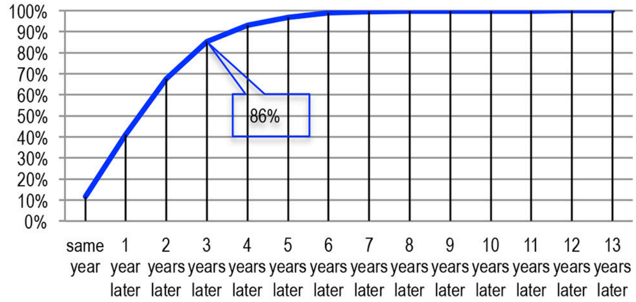


FIGURE 34 | Time delay between publication in conferences and reuse in journals (in %).

of the cases on average, a percentage which falls down from 40 to 25% if the papers are published on the same year.

Plagiarism may raise **legal issues** if it violates copyright, but the *right to quote*[17] exists in certain conditions, considering the Berne convention for the Protection of Literary and Artistic Works[18]: *"National legislations usually embody the Berne convention limits in one or more of the following requirements:*

- *The cited paragraphs are within a reasonable limit,*
- *Clearly marked as quotations and fully referenced,*
- *"The resulting new work is not just a collection of quotations, but constitutes a fully original work in itself,"*
- *"We could also add that the cited paragraph must have a function in the goal of the citing paper."*

Obviously, most of the cases reported in this paper comply with the right to quote. The *limits of the cited paragraph* vary from country to country. In France and Canada, for example, a limit of 10% of both the copying and copied texts seems to be acceptable. As we've seen, it appears that we stay within those limits in all cases in NLP4NLP.

Self-reuse and self-plagiarism are of a different nature and are related to the **ethics and deontology** of a community. Let's recall that they concern papers that have at least one author in common. Of course, a copy & paste operation is easy and frequent but there is another phenomena to take into account which is difficult to distinguish from copy & paste: this is the style of the author. All the authors have habits to formulate their ideas, and, even on a long period, most authors seem to keep the same chunks of prepared words. As we've seen, almost 40% of the cases concern papers that are published on the same year: authors submit two similar papers at two different conferences on the same year, and publish the two papers in both conferences if both are accepted, and they may be unable to properly cite the other paper if it is not yet published or even accepted. It is very difficult for a reviewer to detect and prevent those cases as none of the papers are published when the other one is submitted.

Another frequent case is the publication of a paper in a journal after its publication in a conference. Here also, it is a natural and usual process, sometimes even encouraged by the journal editors after a pre-selection of the best papers in a conference.

As a tentative to moderate these figures and to justify self-reuse and self-plagiarism of previously published material, it is worth quoting Pamela Samuelson (Samuelson, 1994):

- *The previous work must be restated to lay the groundwork for a new contribution in the second work,*
- *Portions of the previous work must be repeated to deal with new evidence or arguments,*
- *The audience for each work is so different that publishing the same work in different places is necessary to get the message out,*

[17] en.wikipedia.org/wiki/Right_to_quote

[18] Berne Convention for the Protection of Literary and Artistic Works (as amended on Sept. 28, 1979). http://www.wipo.int/wipolex/en/treaties/text.jsp?file_id=283693

- *The authors think they said it so well the first time that it makes no sense to say it differently a second time.*

She considers that 30% is an upper limit in the reuse of parts of a paper previously published by the same authors. As we've seen in **Figure 26**, only 1.3% of the documents would fall in this category in NLP4NLP.

We believe that following these two sets of principles regarding (self) reuse and plagiarism will help maintaining an ethical behavior in our community.

CONCLUSIONS

The present paper and its companion one offer a survey of the literature attached to NLP for the last 50 years, and provide examples of the numerous analyses that can be conducted using available tools, some of them resulting from the research conducted in NLP.

As it appears in the various findings, research in NLP for spoken, written and signed languages has made major advances over the past 50 years through constant and steady scientific effort that was fostered thanks to the availability of a necessary infrastructure made up of publicly funded programs, largely available language resources, and regularly organized evaluation campaigns. It keeps on progressing at a high pace, with a very active and coordinated research community. The ethical issues are properly addressed and bridges between the spoken, written and sign language processing communities are being reinforced, through the use of comparable methodologies.

As already mentioned, the lack of a consistent and uniform identification of entities (authors names, gender, affiliations, paper language, conference and journal titles, funding agencies, etc.) required a tedious manual correction process only made possible because we knew the main components of the field. The same applies for Language Resources, where we find initiatives for identifying resources in a persistent and unique way such as the ISLRN (*International Standard Language Resource Number*) (Choukri et al., 2012). Researchers in other disciplines, e.g., biology (Bravo et al., 2015), face the same problems. Establishing standards for such domain-independent identification demands an international effort in order to ensure that the identifiers are unique and appears as a challenge for the scientific community. Therefore, different scientific communities could benefit from mutual experience and methodologies.

PERSPECTIVES

We now plan to investigate more deeply the structure of the research community corresponding to the NLP4NLP corpus. We aim at identifying factions of people who publish together or cite each other. We also plan to refine the study of the polarity of the citations, and deepen the potential detection of weak signals and emerging trends. Establishing links among authors, citations and topics will allow us to study the changes in the topics of interest for authors or factions.

We would like to improve automatic information (names, references, terms) extraction by taking into account the context, in order to make the distinction between real and false occurrences of the information. It would avoid the tedious manual checking that we presently conduct and would improve the overall process.

It should also be noticed that the raw data we gathered and the information we extracted after substantial cleaning could provide data for evaluation campaigns (such as automatic Name Extraction, or Multimedia Gender Detection).

We finally hope that the reader will find interest in the reported results, and may also find inspiration for further interpretation of the reported measures or for conducting other measures on the available data.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication. JM launched the initiative following an invitation to give a keynote talk at Interspeech 2013 to celebrate the 25th anniversary of this major conference in spoken language processing and coordinated the following related and extended works from 2013 to 2018. GF produced the NLP4NLP corpus and developed all the tools that were used for analyzing the corpus. PP participated in the research group and provided advices on the use of NLP tools. FV specifically developed the GapChart visualization tool.

ACKNOWLEDGMENTS

We wish to thank the ACL colleagues, Ken Church, Sanjeev Khudanpur, Amjbad Abu Jbara, Dragomir Radev, and Simone Teufel, who helped them in the starting phase, Isabel Trancoso, who gave her ISCA Archive analysis on the use of assessment and corpora, Wolfgang Hess, who produced and provided a 14 GBytes ISCA Archive, Emmanuelle Foxonet who provided a list of authors given names with genre, Florian Boudin, who made available the TALN Anthology, Helen van der Stelt and Jolanda Voogd (Springer) who provided the LRE data and Douglas O'Shaughnessy, Denise Hurley, Rebecca Wollman and Casey Schwartz (IEEE) who provided the IEEE ICASSP and TASLP data, Nancy Ide and Christopher Cieri who largely improved the readability of parts of this paper. They also thank Khalid Choukri, Alexandre Sicard, and Nicoletta Calzolari, who provided information about the past LREC conferences, Victoria Arranz, Ioanna Giannopoulou, Johann Gorlier, Jérémy Leixa, Valérie Mapelli, and Hélène Mazo, who helped in recovering the metadata for LREC 1998, and all the editors, organizers, reviewers and authors over those 50 years without whom this analysis could not have been conducted!

APOLOGIES

This survey has been made on textual data, which cover a 50-year period, including scanned content. The analysis uses tools that automatically process the content of the scientific papers and may make errors. Therefore, the results should be regarded as reflecting a large margin of error. The authors wish to apologize for any errors the reader may detect, and they will gladly rectify any such errors in future releases of the survey results.

RELATIONSHIP WITH OTHER PAPERS AND REUSE OF PREVIOUS MATERIAL

The present paper is accompanied by another paper "Mariani, Joseph, Paroubek, Patrick, Francopoulo, Gil and Vernier, Frédéric (2018). The NLP4NLP Corpus (II): 50 Years of Research in Speech and Language Processing," in the same special issue of *Frontiers in Research Metrics and Analytics* on "Mining Scientific Papers: NLP-enhanced Bibliometrics" edited by Iana Atanassova, Marc Bertin and Philipp Mayr, which describes the content of this corpus. A summary of the joint two papers has been presented as a keynote talk at the Oriental-Cocosda conference in Seoul ("Joseph Mariani, Gil Francopoulo, Patrick Paroubek, Frédéric Vernier, Rediscovering 50 Years of Discoveries in Speech and Language Processing: A Survey. Oriental Cocosda conference, Seoul, 1-3 November 2017") (Mariani et al., 2017b).

This paper assembles the content of several former papers which described various results of experiments conducted on the NLP4NLP corpus (http://www.nlp4nlp.org). Material from the corresponding previously published sources, listed below, is re-used within permission, implicit or explicit open-licence rights, as follows:

1. Francopoulo, Gil, Mariani, Joseph and Paroubek Patrick (2016). Linking Language Resources and NLP Papers, Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language, LREC 2016, Tenth International Conference on Language Resources and Evaluation, Portorož, Slovenia, May 24, 2016

   This paper analyzes the mention of the Language Resources contained in the LRE Map in the NLP4NLP papers.

   The reused material concerns **Tables 1**, **2** and **Figure 2**.
2. Mariani, Joseph, Paroubek, Patrick, Francopoulo, Gil and Hamon, Olivier (2014). Rediscovering 15 Years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis, LREC 2014, 26-31 May 2014, Reykjavik, Iceland, published within the Proceedings of LREC Conference 2014, http://www.lrec-conf.org/proceedings/lrec2014/index.html

   This paper analyzes the Language Resources and Evaluation Conference (LREC) over 15 years (1998-2014).

   The reused material concerns section *Research Topic Prediction*.
3. Mariani, Joseph, Paroubek, Patrick, Francopoulo, Gil and Hamon, Olivier (2016). Rediscovering 15 + 2 Years of Discoveries in Language Resources and Evaluation, *Language Resources and Evaluation* Journal, 2016, pp. 1-56, ISSN: 1574-0218, doi: 10.1007/s10579-016-9352-9

   This paper has been selected among the LREC 2014 papers to be published in a special issue of the *Language Resources and Evaluation Journal*. It is an extended version of the previous paper, in the following dimensions: extension of the LREC content with the proceedings of the LREC 2014 conference (hence the change in the title of the paper ("15 +2 Years" instead of "15 Years"), and comparison with two other conferences among those contained in NLP4NLP (namely ACL and Interspeech).

   The reused material concerns section *Research Topic Prediction* (mainly subsections *Archive Analysis*, *Terms Frequency and Presence* and *Tag Clouds for Frequent Terms*).
4. Francopoulo, Gil, Mariani, Joseph and Paroubek, Patrick (2016). Predictive Modeling: Guessing the NLP Terms of Tomorrow. LREC 2016, Tenth International Conference on Language Resources and Evaluation Proceedings, Portorož, Slovenia, May 23-28, 2016

   This paper analyzes the possibility to predict the future research topics.

   The reused material concerns section *Research Topic Prediction*.
5. Mariani, Joseph, Francopoulo, Gil and Paroubek, Patrick (2018). Measuring Innovation in Speech and Language Processing Publications, LREC 2018, 9-11 May 2018, Miyazaki, Japan.

   This paper analyzes the innovations brought in the various research topics by the various authors and the various publications within NLP4NLP.

   The reused material concerns section *Innovation*.
6. Mariani, Joseph, Francopoulo, Gil and Paroubek, Patrick (2016). A Study of Reuse and Plagiarism in Speech and Natural Language Processing papers. Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016). 4th Bibliometric-enhanced Information Retrieval (BIR) and 2nd Workshop on text and citation analysis for scholarly digital libraries (NLPIR4DL), Joint Conference on Digital Libraries (JCDL'16), Newark, New Jersey, USA, 23 June 2016.

This paper analyzes the reuse and plagiarism of papers in the NLP4NLP corpus.

The reused material concerns section *Text Reuse and Plagiarism* (mainly subsections *Data*, *Definitions*, *Algorithm for Computing Papers Similarity*, *Categorization of the Results*, and *Time Delay Between Publication and Reuse*).

7. Mariani, Joseph, Francopoulo, Gil and Paroubek, Patrick (2017). Reuse and Plagiarism in Speech and Natural Language Processing Publications, Proc. *International Journal of Digital Libraries*. (2017), doi: 10.1007/s00799-017-0211-0

   This paper has been selected among the BIRNDL 2016 papers to be published in a special issue of the *International Journal of Digital Libraries*. It is an extended version of the previous paper, with a detailed analysis of the findings and a study on the timing of the reuses.

   The reused material concerns section *Text Reuse and Plagiarism* (mainly subsections *Self-Reuse and Self-Plagiarism*, *Reuse and Plagiarism*, and *Legal and Ethical Limits*).

REFERENCES

Barron-Cedeno, A., Potthast, M., Rosso, P., Stein, B., and Eiselt, A. (2010). "Corpus and evaluation measures for automatic plagiarism detection," in *Proceedings of LREC* (Valletta).

Bravo, E., Calzolari, A., De Castro, P., Mabile, L., Napolitani, F., Rossi, A. M., et al. (2015). Developing a guideline to standardize the citation of bioresources in journal articles (CoBRA). *BMC Med*. 13:33. doi: 10.1186/s12916-015-0266-y

Calzolari, N., Del Gratta, R., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., et al. (2012). "The LRE map. harmonising community descriptions of resources," in *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)* (Istanbul).

Choukri, K., Arranz, V., Hamon, O., and Park, J. (2012). "Using the international standard language resource number: practical and technical aspects," in *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)* (Istanbul).

Clough, P., Gaizauskas, R., and Piao, S. S. L. (2002b). "Building and annotating a corpus for the study of journalistic text reuse," in *Proceedings of LREC* (Las Palmas).

Clough, P., Gaizauskas, R., Piao, S. S. L., and Wilks, Y. (2002a). "Measuring text reuse," in *Proceedings of ACL'02* (Philadelphia, PA).

Clough, P., and Stevenson, M. (2011). Developing a corpus of plagiarised short answers. *Lang. Resour. Eval. J.* 45, 5–24. doi: 10.1007/s10579-009-9112-1

Drouin, P. (2004). "Detection of domain specific terminology using corpora comparison," in *Proceedings of the Language Resources and Evaluation Conference (LREC 2004)* (Lisbon).

Francopoulo, G. (2008). "TagParser: well on the way to ISO-TC37 conformance," in *ICGL (International Conference on Global Interoperability for Language Resources)* (Hong Kong).

Francopoulo, G., Marcoul, F., Causse, D., and Piparo, G. (2013). "Global atlas: proper nouns, from Wikipedia to LMF," in *LMF-Lexical Markup Framework,* ed G. Francopoulo (ISTE/Wiley), 227–241.

Francopoulo, G., Mariani, J., and Paroubek, P. (2015). *NLP4NLP: The Cobbler's Children Won't Go Unshod*. Available online at: www.dlib.org/dlib/november15/francopoulo/11francopoulo.html

Francopoulo, G., Mariani, J., and Paroubek, P. (2016a). "Predictive modeling: guessing the NLP terms of tomorrow," in *LREC 2016, Tenth International Conference on Language Resources and Evaluation Proceedings* (Portorož).

Francopoulo, G., Mariani, J., and Paroubek, P. (2016b). "Linking language resources and NLP papers," in *Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language, LREC 2016, Tenth International Conference on Language Resources and Evaluation* (Portorož).

Gaizauskas, R., Foster, J., Wilks, Y., Arundel, J., Clough, P., and Piao, S. S. L. (2001). "The METER corpus: a corpus for analysing journalistic text reuse," in *Proceedings of the Corpus Linguistics Conference* (Lancaster).

Guo, Y., Che, W., Liu, T., and Li, S. (2011). "A graph-based method for entity linking," in *International Joint Conference on NLP* (Chiang Mai).

HaCohen-Kerner, Y., Tayeb, A., and Ben-Dror, N. (2010). "Detection of simple plagiarism in computer science papers," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)* (Beijing).

Hall, D. L.W., Jurafsky, D., and Manning, C. (2008). "Studying the history of ideas using topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)* (Honolulu, HI), 363–371.

Ide, N., Suderman, K., and Simms, B. (2010). "ANC2Go: a web application for customized corpus creation," in *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)* (Valletta: European Language Resources Association).

Kasprzak, J., and Brandejs, M. (2010). "Improving the reliability of the plagiarism detection system lab," in *Proceedings of the Uncovering Plagiarism, Authorship and Social Software Misuse (PAN)* (Padua).

Koehn, P. (2005). "Europarl: a parallel corpus for statistical machine translation," in *Conference Proceedings: The Tenth Machine Translation Summit* (Phuket), 79–86.

Lyon, C., Malcolm, J., and Dickerson, B. (2001). "Detecting short passages of similar text in large document collections," in *Proc. of the Empirical Methods in Natural Language Processing Conference* (Pittsburgh, PA).

Mariani, J., Francopoulo, G., and Paroubek, P. (2016). "A study of reuse and plagiarism in speech and natural language processing papers," in *Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016). 4th Bibliometric-enhanced Information Retrieval (BIR) and 2nd Workshop on text and citation analysis for scholarly digital libraries (NLPIR4DL), Joint Conference on Digital Libraries (JCDL '16)* (Newark, NJ).

Mariani, J., Francopoulo, G., and Paroubek, P. (2017a). Reuse and plagiarism in speech and natural language processing publications. *P. Int. J. Digit Libr*. 19, 113–126. doi: 10.1007/s00799-017-0211-0

Mariani, J., Francopoulo, G., and Paroubek, P. (2018a). "Measuring innovation in speech and language processing publications," in *LREC 2018* (Miyazaki).

Mariani, J., Francopoulo, G., and Paroubek, P. (2018b). The NLP4NLP corpus (I): 50 years of publication, collaboration and citation in speech and language processing. *Front. Res. Metr. Anal.* 3:36. doi: 10.3389/frma.2018.00036

Mariani, J., Francopoulo, G., Paroubek, P., and Vernier, F. (2017b). "Rediscovering 50 years of discoveries in speech and language processing: a survey," in *Oriental Cocosda Conference* (Seoul: IEEE XPlore). doi: 10.1109/ICSDA.2017.8384413

Moro, A., Raganato, A., and Navigli, R. (2014). Entity linking meets word sense disambiguation: a unified approach. *Trans. Assoc. Comput. Linguist.* 2, 231–244.

Paul, M., and Girju, R. (2009). "Topic modeling of research fields: an interdisciplinary perspective," in *Recent Advances in Natural Language Processing (RANLP 2009)* (Borovets).

Perin, C., Boy, J., and Vernier, F. (2016). "GapChart: a gap strategy to visualize the temporal evolution of both ranks and scores," in *IEEE Computer Graphics and Applications, Special Issue on Sports Data Visualization,* Vol. 36.

Samuelson, P. (1994). Self-plagiarism or fair use? *Commun. ACM* 37, 21–25.

Vilnat, A., Paroubek, P., de la Clergerie, E., Francopoulo, G., and Guénot, M. L. (2010). "PASSAGE syntactic representation: a minimal common ground for evaluation," in *Proceedings of LREC 2010* (Valletta).

Witten, I. H., Eibe, F., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, *3rd Edn*. Burlington, VT: Morgan Kaufmann.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

TABLE A1 | Ten most present terms in 2015, with variants, date, authors and publications where they were first introduced, number of occurrences and existences in 2015, number of occurrences, frequency, number of existences and presence in the 50 year archive, with ranking and average number of occurrences of the terms in the documents where they appear.

Rank	Term	Variants of all sorts	Date when the term appeared	Authors who introduced the term	Documents	Archive #Occurrences	Archive frequency	Archive #Existences	Archive Presence	Archive Rank Occurrence	Archive Rank Presence	Archive Ratio occurrences / existences	# occurrences in the last year	# existences in the last year	Frequency in the last year	Presence in the last year
1	Dataset	data-set, data-sets, datasets	1966	Laurence Urdang	cath1966-3	65250	0.003	9940	0.16	11	18	6.6	14039	1472	0.0092	0.44
2	Metric	metrics	1965	A Andreyewsky	C65-1002	50679	0.002	11335	0.18	19	10	4.5	5425	1108	0.0036	0.34
3	Subset	sub set, sub sets, sub-set, sub-sets, subsets	1965	Denis M Manelski, E D Pendergraft, Gilbert K Krulee, Itiroo Sakai, N Dale, Wojciech Skalmowski	C65-1006 C65-1018 C65-1021 C65-1025	45616	0.002	16939	0.27	22	2	2.7	3463	1095	0.0023	0.33
4	Neural network	ANN, ANNs, Artificial Neural Network, Artificial Neural Networks, NN, NNs, Neural Network, Neural Networks, NeuralNet, NeuralNets, neural net, neural nets, neural networks	1980	Bonnie Lynn Webber	P80-1032	54790	0.002	8885	0.14	16	27	6.2	8024	1037	0.0053	0.31
5	Classifier	classifiers	1967	Aravind K Joshi, Danuta Hiz	C67-1007	98229	0.004	11546	0.18	7	9	8.5	8202	1000	0.0054	0.30
6	SR	ASR, ASRs, Automatic Speech Recognition, SRs, Speech Recognition, automatic speech recognition, speech recognition	1970	Josse De Kock	cath1970-9	129979	0.006	20382	0.32	2	1	6.4	8524	1000	0.0056	0.30
7	Optimization	optimisation, optimisations, optimizations	1967	Ellis B Page	C67-1032	35257	0.002	10196	0.16	35	16	3.5	3331	903	0.0022	0.27
8	Annotation	annotations	1967	Kenneth Janda, Martin Kay	cath1967-12 cath1967-8	111084	0.005	11975	0.19	4	7	9.3	7515	896	0.0049	0.27
9	POS	POSs, Part Of Speech, Part of Speech, Part-Of-Speech, Part-of-Speech, Parts Of Speech, Parts of Speech, Pos, part of speech, part-of-speech, parts of speech, parts-of-speech	1965	Denis M Manelski, Dániel Várga, Gilbert K Krulee, Makoto Nagao, Toshiyuki Sakai	C65-1018 C65-1022 C65-1029	102057	0.005	13823	0.22	5	4	7.4	7489	860	0.0049	0.26
10	LM	LMs, Language Model, Language Models, language model, language models	1965	Sheldon Klein	C65-1014	116684	0.005	13117	0.21	3	5	8.9	8522	851	0.0056	0.26

TABLE A2 | Ranked top 10 mentioned LRE map language resources per year (1965–2015).

Year	# existences of LR	# documents	Top10 cited resources (ranked)
1965	7	24	C-3, LLL, LTH, OAL, Turin University Treebank
1966	0	7	
1967	6	54	General Inquirer, LTH, Roget's Thesaurus, TFB, TPE
1968	3	17	General Inquirer, Medical Subject Headings
1969	4	24	General Inquirer, Grammatical Framework GF
1970	2	18	FAU, General Inquirer
1971	0	20	
1972	2	19	Brown Corpus, General Inquirer
1973	7	80	ANC Manually Annotated Sub-corpus, Grammatical Framework GF, ILF, Index Thomisticus, Kontrast, LTH, PUNKT
1974	8	25	General Inquirer, Brown Corpus, COW, GG, LTH
1975	15	131	C-3, LTH, Domain Adaptive Relation Extraction, ILF, Acl Anthology Network, BREF, LLL, Syntax in Elements of Text, Unsupervised incremental parser
1976	13	136	Grammatical Framework GF, LTH, C-3, DAD, Digital Replay System, Domain Adaptive Relation Extraction, General Inquirer, Perugia Corpus, Syntax in Elements of Text, Talbanken
1977	8	141	Grammatical Framework GF, Corpus de Referencia del Español Actual, Domain Adaptive Relation Extraction, GG, LTH, Stockholm-Umeå corpus
1978	16	155	Grammatical Framework GF, C-3, General Inquirer, Digital Replay System, ILF, LLL, Stockholm-Umeå corpus, TDT
1979	23	179	Grammatical Framework GF, LLL, LTH, C-3, C99, COW, CTL, ILF, ItalWordNet, NED
1980	38	307	Grammatical Framework GF, C-3, LLL, LTH, ANC Manually Annotated Sub-corpus, Acl Anthology Network, Automatic Statistical SEmantic Role Tagger, Brown Corpus, COW, CSJ
1981	33	274	C-3, Grammatical Framework GF, LTH, Index Thomisticus, CTL, JWI, Automatic Statistical SEmantic Role Tagger, Brown Corpus, Glossa, ILF
1982	40	364	C-3, LLL, LTH, Brown Corpus, GG, ILF, Index Thomisticus, Arabic Gigaword, Arabic Penn Treebank, Automatic Statistical SEmantic Role Tagger
1983	59	352	Grammatical Framework GF, C-3, LTH, GG, LLL, Unsupervised incremental parser, LOB Corpus, OAL, A2ST, Arabic Penn Treebank
1984	55	353	LTH, Grammatical Framework GF, PET, LLL, C-3, CLEF, TLF, Arabic Penn Treebank, Automatic Statistical SEmantic Role Tagger, COW
1985	53	384	Grammatical Framework GF, LTH, C-3, LOB Corpus, Brown Corpus, Corpus de Referencia del Español Actual, LLL, DCR, MMAX, American National Corpus
1986	92	518	LTH, C-3, LLL, Digital Replay System, Grammatical Framework GF, DCR, JRC Acquis, Nordisk Språkteknologi, Unsupervised incremental parser, OAL
1987	63	669	LTH, C-3, Grammatical Framework GF, DCR, Digital Replay System, LOB Corpus, CQP, EDR, American National Corpus, Arabic Penn Treebank
1988	105	546	C-3, LTH, Grammatical Framework GF, Digital Replay System, DCR, Brown Corpus, FSR, ISOcat Data Category Registry, LOB Corpus, CTL
1989	145	965	Grammatical Framework GF, Timit, LTH, LLL, C-3, Brown Corpus, Digital Replay System, LTP, DCR, EDR
1990	175	1277	Timit, Grammatical Framework GF, LTH, C-3, LLL, Brown Corpus, GG, LTP, ItalWordNet, JRC Acquis
1991	240	1378	Timit, LLL, C-3, LTH, Grammatical Framework GF, Brown Corpus, Digital Replay System, LTP, GG, Penn Treebank
1992	361	1611	Timit, LLL, LTH, Grammatical Framework GF, Brown Corpus, C-3, Penn Treebank, WordNet, GG, ILF
1993	243	1239	Timit, WordNet, Penn Treebank, Brown Corpus, EDR, LTP, User-Extensible Morphological Analyzer for Japanese, BREF, Digital Replay System, James Pustejovsky
1994	292	1454	Timit, LLL, WordNet, Brown Corpus, Penn Treebank, C-3, Digital Replay System, JRC Acquis, LTH, Wall Street Journal Corpus
1995	290	1209	Timit, LTP, WordNet, Brown Corpus, Digital Replay System, LLL, Penn Treebank, Grammatical Framework GF, TEI, Ntimit
1996	394	1536	Timit, LLL, WordNet, Brown Corpus, Digital Replay System, Penn Treebank, Centre for Spoken Language Understanding Names, LTH, EDR, Ntimit
1997	428	1530	Timit, WordNet, Penn Treebank, Brown Corpus, LTP, HCRC, Ntimit, BREF, LTH, British National Corpus
1998	883	1953	Timit, WordNet, Penn Treebank, Brown Corpus, EuroWordNet, British National Corpus, Multext, EDR, LLL, PAROLE
1999	481	1603	Timit, WordNet, Penn Treebank, TDT, Maximum Likelihood Linear Regression, EDR, Brown Corpus, TEI, LTH, LLL
2000	842	2271	Timit, WordNet, Penn Treebank, British National Corpus, PAROLE, Multext, EuroWordNet, Maximum Likelihood Linear Regression, TDT, Brown Corpus

(Continued)

TABLE A2 | Continued

Year	# existences of LR	# documents	Top10 cited resources (ranked)
2001	648	1644	WordNet, Timit, Penn Treebank, Maximum Likelihood Linear Regression, TDT, Brown Corpus, CMU Sphinx, Praat, LTH, British National Corpus
2002	1105	2174	WordNet, Timit, Penn Treebank, Praat, EuroWordNet, British National Corpus, PAROLE, NEGRA, TDT, Grammatical Framework GF
2003	1067	1984	Timit, WordNet, Penn Treebank, AQUAINT, British National Corpus, AURORA, FrameNet, Praat, SRI Language Modeling Toolkit, OAL
2004	2066	2712	WordNet, Timit, Penn Treebank, FrameNet, AQUAINT, British National Corpus, EuroWordNet, Praat, PropBank, SemCor
2005	2006	2355	WordNet, Timit, Penn Treebank, Praat, AQUAINT, PropBank, British National Corpus, SRI Language Modeling Toolkit, MeSH, TDT
2006	3532	2794	WordNet, Timit, Penn Treebank, Praat, PropBank, AQUAINT, FrameNet, GALE, EuroWordNet, British National Corpus
2007	2937	2489	WordNet, Timit, Penn Treebank, Praat, SRI Language Modeling Toolkit, Wikipedia, GALE, GIZA++, SemEval, AQUAINT
2008	4007	3078	WordNet, Wikipedia, Timit, Penn Treebank, GALE, PropBank, Praat, FrameNet, SRI Language Modeling Toolkit, Weka
2009	3729	2637	WordNet, Wikipedia, Timit, Penn Treebank, Praat, SRI Language Modeling Toolkit, GALE, Europarl, Weka, GIZA++
2010	5930	3470	WordNet, Wikipedia, Penn Treebank, Timit, Europarl, Praat, FrameNet, SRI Language Modeling Toolkit, GALE, GIZA++
2011	3859	2957	Wikipedia, WordNet, Timit, Penn Treebank, Praat, SRI Language Modeling Toolkit, Weka, GIZA++, Europarl, GALE
2012	6564	3419	Wikipedia, WordNet, Timit, Penn Treebank, Europarl, Weka, Praat, SRI Language Modeling Toolkit, GIZA++, FrameNet
2013	5669	3336	Wikipedia, WordNet, Timit, Penn Treebank, Weka, SRI Language Modeling Toolkit, Praat, GIZA++, Europarl, SemEval
2014	6700	3817	Wikipedia, WordNet, Timit, Penn Treebank, Praat, Weka, SRI Language Modeling Toolkit, SemEval, Europarl, FrameNet
2015	5597	3314	Wikipedia, WordNet, Timit, SemEval, Penn Treebank, Praat, Europarl, Weka, SRI Language Modeling Toolkit, FrameNet

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership