SYSTEMATIC REVIEW
published: 29 October 2020
doi: 10.3389/fcomp.2020.523053

Meta-Analysis of Cross-Language Plagiarism and Self-Plagiarism Detection Methods for Russian-English Language Pair

Alina Tlitova [1]*, *Alexander Toschev* [1], *Max Talanov* [1] *and Vitaliy Kurnosov* [2]

[1] Computer Engineering Department, The Higher Institute of Information Technology and Intelligent Systems, Kazan Federal University, Kazan, Russia, [2] TPPKM Department, Institute of Polymers, Kazan National Research Technological University, Kazan, Russia

Scientists need to publish the results of their work to remain relevant and in demanded. The well-known principle of "publish or perish" often forces scientists to pursue an increase in quantity, not quality. Along with the problems of authorship, paid research, fabrication of results, plagiarism, and self-plagiarism are among the most common violations. Their impact is more subtle but no less destructive for the scientific community. Statistics show that the reuse of texts in different languages is very common in various studies. Identification of translated plagiarism is a complex task, and there are almost no such tools for this purpose on the Russian market now. In this article, we have provided an overview of the existing methods for the identification of cross-language borrowings in the scientific articles of the authors. We analyzed solutions by studying the works on various language pairs and paid the great attention to the Russian-English language pair.

Keywords: cross-language plagiarism, plagiarism detection methods, self-plagiarism detection, text borrowings, multilingual plagiarism

OPEN ACCESS

Edited by:
Tom Crick,
Swansea University, United Kingdom

Reviewed by:
Diego R. Amancio,
University of São Paulo, Brazil
Mohamed Mostafa,
Cardiff Metropolitan University, United Kingdom
Imtiaz Khan,
Cardiff Metropolitan University, United Kingdom

****Correspondence:***
Alina Tlitova
AlETlitova@stud.kpfu.ru

Specialty section:
This article was submitted to Digital Education, a section of the journal Frontiers in Computer Science

Received: *26 December 2019*
Accepted: *20 August 2020*
Published: *29 October 2020*

Citation:
Tlitova A, Toschev A, Talanov M and Kurnosov V (2020) Meta-Analysis of Cross-Language Plagiarism and Self-Plagiarism Detection Methods for Russian-English Language Pair. Front. Comput. Sci. 2:523053. doi: 10.3389/fcomp.2020.523053

1. INTRODUCTION

The number of publications has a great influence on the career and wealth of a scientist. To succeed in the form of increasing the degree and recognition, unscrupulous authors translate an existing scientific work into the required language and pass it off as their own or republish their work in different languages without indicating that this information has been published previously (Amancio, 2015). The whole work can be repeated with minor changes, for example, in the title, abstracts (double publication), and excerpts from the previous ones (salami slicing). Such work cannot be called scientific. A good example is the reprinted text.

IEEE defines plagiarism as the reuse of someone else's previous results or words without explicit recognition of the original author and source (IEEE-Faq, 2019). Plagiarism in any form is unacceptable and is considered a serious violation of professional behavior with ethical and legal consequences.

According to the IEEE policy there are several basic factors that are taken into account when assessing possible plagiarism (IEEE, 2019):

- Number (full article, article section, page, paragraph, sentence, and phrases)
- Using quotes for all copied text
- Proper placement of links to sources of borrowing
- Incorrect rephrasing.

Duplication of information without reference to sources is unacceptable, even if it is the property of the author, as this calls into question the relevance and scientific novelty of the idea.

Plagiarism detection systems are often used to identify self-plagiarism.

Many programs successfully cope with plagiarism and self-plagiarism in the same language (Tlitova and Toschev, 2019). However, their significant drawback is the detection of borrowings from different languages, the so-called cross-language plagiarism.

There are several groups of state-of-the-art methods for detecting this type of plagiarism that we reviewed in this work.

The purpose of this study is to analyze existing methods for identifying textual cross-language borrowings by several characteristics to then identify their features and applicability for the Russian-English pair.

2. METHODS

In this research we performed a meta-analysis of existing works and articles aimed at survey systems for identifying cross-language plagiarism, and their relevance for the Russian-English language pair. It was conducted in adherence to the standards of quality for conducting and reporting meta-analyses detailed in the PRISMA statement (Moher et al., 2009).

We exploited two methods of literature meta-analysis.

The first way is looking for publications in databases such as Scopus and Web Of Science and including all publication types obtained through the World Wide Web: unpublished dissertations, peer-reviewed journal articles, book chapters, and conference proceedings. We omitted articles that were published before January 2004, mostly focusing on studies of the last 5 years. We used the following relevant keywords for the search: cross-language plagiarism, plagiarism detection methods, self-plagiarism detection, multilingual plagiarism, and text borrowings.

The second way consisted of a cross-referencing process that included articles that were used to recognize other appropriate works (Horsley et al., 2011). Using a backward-search process, we read the references at the end of articles to find other research that could potentially be used in the meta-analysis. Then we conducted a forward search via Google Scholar (2004) to identify these studies.

A total of 136 studies were identified. After removing 38 duplicates, 49 studies were excluded after title and abstract review. The full text of the remaining items was examined in detail. Part of them were not suitable for inclusion in our study for the following reasons:

- No analysis of methods in practice
- No appropriate description of the models and methods.

After these steps of the literature search, two articles published online that explored Russian-English pair of languages (Kuznecova et al., 2018; Zubarev and Sochenkov, 2019) and seven articles that analyzed models of cross-language plagiarism (Barrón-Cedeño et al., 2010, 2013; Potthast et al., 2011; Franco-Salvador et al., 2016b; Ferrero et al., 2017b; Thompson and Bowerman, 2017; Ehsan et al., 2019) remained and were included in the meta-analysis.

Figure 1 presents a “PRISMA Flow Diagram” the study selection that depicts the flow of information through the various stages of a systematic review.

As known detection of plagiarism has two stages:

1. Search for sources for selection of candidate documents
2. Text alignment to compare a document with each candidate.

We studied the models of the second stage, which are used for the CLPD task, studied existing works aimed to detect plagiarism in a Russian-English pair, and created tables and bar charts for the visual presentation of the results. There are several approaches that offer a solution to the problem of cross-language plagiarism detection for different pairs of languages: Arabic-English (Hanane et al., 2016), (Alaa et al., 2016), Malay-English (Kent and Salim, 2010), Spanish-English and German-English (Franco-Salvador et al., 2016b), Basque-English (Barrón-Cedeño et al., 2010), and Russian-English (Zubarev and Sochenkov, 2019). In the work (Barrón-Cedeño et al., 2010), the authors noted that the effectiveness of the plagiarism detection algorithms directly depends on the degree of relationship between the considered languages. If languages are not in the same linguistic group, it causes additional difficulties for developing an algorithm for identifying borrowings.

There are methods that use different models for detecting cross-language plagiarism from five following groups (Ferrero et al., 2017b); the purpose is to determine whether two text blocks are identical in terms of information content:

- Syntax-based models (Length model, CL-CnG, and Cognateness)
- Dictionary-based models (CL-VSM and CL-CTS)
- Parallel corpora-based models (CL-ASA, CL-LSI, and CL-KCCA)
- Comparable corpora-based models (CL-KGA and CL-ESA)
- Machine translation-based models (T + MA).

The authors of article (Ferrero et al., 2017b) and (Potthast et al., 2011) analyzed the models:

- CL-CnG (Cross-Language Character N-Gram) is based on McNamee and Mayfield models (McNamee and Mayfield, 2004) and represents documents with character n-grams.
- CL-CTS (Cross-Language Conceptual Thesaurus-based Similarity) is aimed at determining semantic similarity using abstract concepts from words in the text.
- CL-ASA (Cross-Language Alignment-based Similarity Analysis) determines how a text unit is a potential translation of another text unit using a bilingual unigram dictionary that contains pairs of translations (and their probabilities) extracted from parallel corpora.
- CL-ESA (Cross-Language Explicit Semantic Analysis) is based on the explicit semantic analysis model, which represents the value of a document as a vector based on a dictionary derived from Wikipedia.

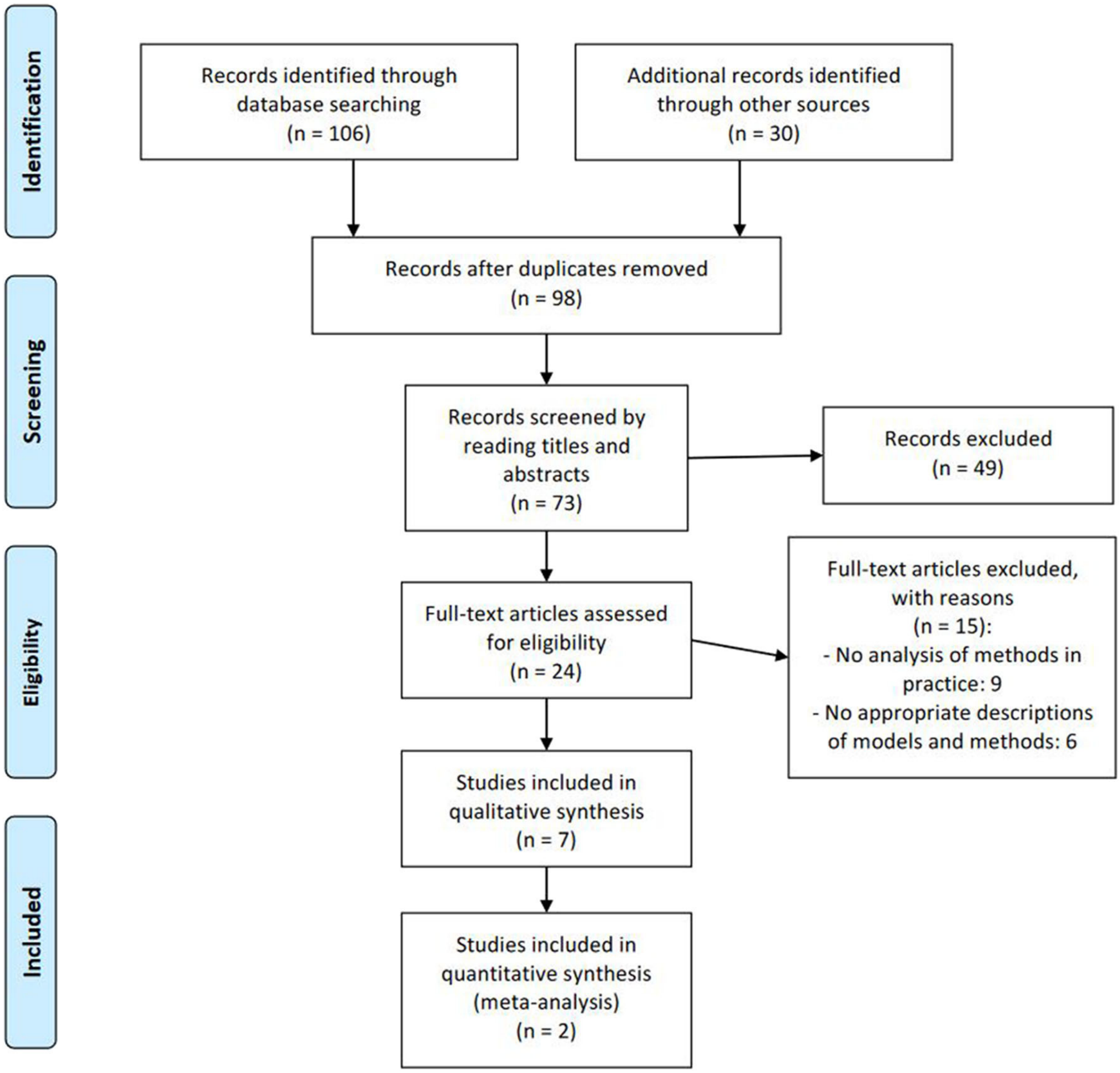
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 106)
Additional records identified through other sources (n = 30)
Records after duplicates removed (n = 98)
Records screened by reading titles and abstracts (n = 73)
Records excluded (n = 49)
Full-text articles assessed for eligibility (n = 24)
Full-text articles excluded, with reasons (n = 15):
- No analysis of methods in practice: 9
- No appropriate descriptions of models and methods: 6
Studies included in qualitative synthesis (n = 7)
Studies included in quantitative synthesis (meta-analysis) (n = 2)

FIGURE 1 | Study selection flow diagram by PRISMA (Moher et al., 2009) template.

- T + MA (Translation + Monolingual Analysis) consists of translating suspicious plagiarism back into the same source language to make a unilingual comparison between them.
- CL-VSM (Cross-Language Vector Space Model) is based on a vector space model.
- CL-LSI (Cross-Language Latent Semantic Indexing) is based on hidden semantic indexing.
- CL-KCCA (Cross-Language Kernel Canonical Correlation Analysis) is based on canonical core correlation analysis.

CL-LSI and CL-KCCA achieve high search quality, but the runtime is very long, which makes them inapplicable to many practical tasks. CL-VSM requires a lot of effort to remove ambiguities, and the availability of dictionaries for translation of manual works depends on the frequency of translations between the respective languages for this model. So that we excluded these models from our comparison. CL-CnG, CL-ESA, and CL-ASA provide good search quality and do not require manual fine-tuning.

The multilingual dataset in Ferrero et al. (2017b) was specially designed for evaluation of cross-language textual similarity detection. It is based on parallel and comparable corpora (mix of Wikipedia, scientific conference papers, amazon product reviews, Europarl, and JRC) including French, English, and Spanish texts.

The authors of Barrón-Cedeño et al. (2013) compared T+MA model with CL-CNG and CL-ASA using the Spanish-English partition of PAN'11 competition.

We also found a comparative analysis of CL-CNG, CL-ESA, and CL-ASA models in Potthast et al. (2011). Authors studied the behavior of the models on 120,000 test documents from the JRC-Acquis parallel corpus and Wikipedia comparable corpus and for each test document highly similar texts were available in English, German, Dutch, Spanish, French, and Polish languages.

We paid attention to the CL-KGA model (Franco-Salvador et al., 2016b), which reflects text parts through knowledge graphs as a language independent content model and which is applicable at the level of CL-ASA, CL-ESA, and CL-CnG. The authors did a comprehensive comparative analysis of CL-CnG, CL-ESA, and CL-ASA with usual CL-KGA and its various models using the largest multilingual semantic network that combines lexicographic information with amplitudinous encyclopedic knowledge BabelNet (Navigli and Ponzetto, 2012) and others. They selected the datasets used for the CL plagiarism detection competition PAN-PC-10 and PAN-PC-11, which consisted of Spanish-English and German-English sections.

In addition, we included in our meta-analysis the less famous models (though they still show good results for this task) that are described in works (Thompson and Bowerman, 2017; Ehsan et al., 2019). The authors of the model used cross-lingual word embeddings (CL-WE) and multilingual translation model (MTM) (Thompson and Bowerman, 2017) used datasets from PAN-PC-11 and PAN-PC-12. PAN-PC-12 is also used by the authors of model proposed in Ehsan et al. (2019).

Despite the variety of described models, the majority of authors use a conventional machine translation (MT) model in methods and algorithms for detecting cross-language borrowings, and the task transforms into identifying monolingual plagiarism. The disadvantage of this approach is that machine translation provides various versions, and authors can change parts of the text that are reused.

Many works use text comparisons based on monolingual or bilingual word vectors (Franco-Salvador et al., 2016a; Ferrero et al., 2017a). However, the authors proposed a method that uses vectors of phrases to detect plagiarism in a Russian-English pair in one recent study (Kuznecova et al., 2018). The paper describes an algorithm that performs monolingual analysis of documents: firstly, the text is completely translated into English, and then not the fragments of the text but the corresponding vectors are compared to reduce instability to the translation ambiguities. This is “proposed” algorithm. Additionally, they took an algorithm based on a shingle as “basic” to reproduce the following steps:

1. Translation of the checked document into English
2. Lemmatization of the obtained text and its division into many overlapping 4-grams
3. Sorting words within each 4-gram to account for the possible permutations of words in translation
4. A set of matching sorted 4-grams is the result of comparing documents.

The work (Zubarev and Sochenkov, 2019) is also devoted to plagiarism analysis for the Russian-English pair. The authors present a dataset for the text alignment task as an alternative to existing datasets. They compare two models to detect translated plagiarism. One of them is based on various similarity indicators for texts that use word embedding and neural machine translation. Moreover, the other is built as an addition to the previous one based on a pre-trained language presentation (Bert). Also, they generated two corpora with various count of negative samples per each positive sentence pair that include the source and plagiarized sentences:

1. Negative-1: One negative example is selected randomly from the most similar sentences. The authors use this dataset for training and tuning models.
2. Negative-4: Four negative examples are selected (one most similar sentence for each used similarity score). They use this dataset for testing purposes to check how models handle a larger amount of negative examples.

The authors (Zubarev and Sochenkov, 2019) conduct computational experiments using various classifiers:

- NMT is a neural machine translation that measure similarity on 1-grams using OpenNMT-py4 library to train a machine translation system as an additional criterion to evaluate the pairwise similarity between the sentences.
- NMT2 is the same neural machine translation but measured similarity on 2-grams.
- LR-1 is a logistic regression classifier with L2 regularization using two similarity scores: one based on sentence embeddings and one calculated after the substitution of all words with the most similar ones in the other language.
- LR-2 is a logistic regression classifier with C = 1.0 using only sentence embeddings similarity scores and the word substitution similarity score.
- LASER [Language-Agnostic SEntence Representations (Artetxe and Schwenk, 2018)] is a method to obtain sentence embeddings that provides an encoder called BiLSTM and trained on 93 languages.
- BERT [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding] Authors used the BERT Multilingual model and considered a simple linear layer for the sentence pair classification on top of the pooled output of Bert.

The authors (Zubarev and Sochenkov, 2019) used Word2Vec for vector representation of words and created a specified dataset consisting of 16,000 sentence pairs from Yandex parallel corpus that did not take part in learning word embeddings and 4,000 sentences written by students by looking for sources in English on the internet and via translation using Yandex and Google Translate services, making adjustments for getting the right Russian text. Small parts of sentences were translated manually without these tools and served as positive instances of plagiarized pairs of sentences. The authors of the Russian language work (Kuznecova et al., 2018) used FastText library for representation of words and chose 18.5 million parallel sentences from Opus corpora that we show in **Table 5**.

3. RESULTS

We chose six commonly used models that could be scaled to work in a real-world setting for CL plagiarism detection: CL-ASA, CL-ESA, CL-CnG, CL-CTS, CL-KGA, T+MA, and two models from (Thompson and Bowerman, 2017; Ehsan et al.,

TABLE 1 | Comparative table of models T+MA, CL-ESA, CL-ASA.

Parameter	T+MA	CL-ESA	CL-ASA
Comparison with	CL-CNG, CL-ESA, CL-CTS, CL-ASA, CL-KGA	CL-CNG, T+MA, CL-CTS, CL-ASA, CL-KGA	CL-CNG, CL-ESA, CL-CTS, T+MA, CL-KGA
Dependence on corpora	Independent	Gives better results on comparable corpora like Wikipedia	Gives better results on parallel corpora like JRC or Europarl
Dependence on kinship of languages	Can be used on language pairs whose alphabet and syntax are not related	Can be used on language pairs whose alphabet and syntax are not related	Can be used on language pairs whose alphabet and syntax are not related
Dependence on machine translator	Depends on the quality of machine translation	No data	Uses statistical machine translation, and depends on its quality
Dependence on the length of document	More favorable F1 in all cases except long. Poorly detects plagiarism in small documents	No data	Its formula tends to minimize the number of false positions in short-length texts
Productivity	High recall (erroneous translations at the stage of normalization of the text can be the cause of low precision). More effective at sentence detalization than CL-ASA	Performance is close to CL-CNG but depends on language pairs. It is based on similarities to a collection of documents and gives a large number of false positives, because it was originally intended for tasks of similarity, not plagiarism	High precision and performance in long documents (from some paragraphs up to entire documents). Shows good results in professional, automatic translations and receives a small number of false positives. Better detects human translations than T+MA

TABLE 2 | Comparative table of models CL-CNG, CL-CTS, CL-KGA.

Parameter	CL-CNG	CL-CTS	CL-KGA
Comparison with	T+MA, CL-ESA, CL-CTS, CL-ASA, CL-KGA	CL-CNG, CL-ESA, T+MA, CL-ASA, CL-KGA	CL-CNG, CL-ESA, CL-CTS, CL-ASA
Dependence on corpora	Gives better results on comparable corpora like Wikipedia	No data	No data
Dependence on kinship of languages	Has low quality for language pairs without lexical and syntactic similarities. More effective than CL-ASA, CL-ESA in cases when languages are syntactically related	No data	No data
Dependence on machine translator	No data	No data	No data
Dependence on the length of document	No data	Its formula tends to minimize the number of false positions in short-length texts	No data
Productivity	High recall. Provides acceptable retrieval quality	The behavior depends on granularity and the level of detail. More effective at sentence detalization than CL-ASA	Provides high results for all indicators through the use of knowledge graphs. Offers better performance with Babel Net's high reach and interconnectivity concept than CL-ESA

2019). The qualitative results of the analysis of them are presented in **Tables 1–3**.

Table 4 introduces our comparative results of the approaches used in the articles for the Russian-English pair. Precision, recall, and F1 characteristics are also represented in **Figures 2–4**.

F1 is a balance of accuracy and completeness of classification that is calculated as the following:

$$F1 = \frac{2PR}{P + R} \tag{1}$$

where P—precision, R—recall (Kuznecova et al., 2018).

As for algorithms that authors used for the Russian-English pair, the Bert language model shows a high performance; however, it is inappropriate during large-scale checking of borrowings because it exhibited the worst results over time, as can be seen in **Table 4**. It shows good results only after retraining for a specific task. Additionally, Bert is quite complex and requires great hardware capacity, and its use is thus limited. The authors of Zubarev and Sochenkov (2019), therefore proposed a classifier with a reduced space for features for effective filtering: only with sentence embeddings and word substitution measures. They considered it reasonable to use both context-free models and context models together in modern plagiarism

TABLE 3 | Comparative table of model used cross-lingual word embeddings (CL-WE) and multilingual translation model (MTM) (Thompson and Bowerman, 2017) and model proposed in Ehsan et al. (2019).

Parameter	Model used cross-lingual word embeddings (CL-WE) and multilingual translation model (MTM) (Thompson and Bowerman, 2017)	Model proposed in Ehsan et al. (2019)
Comparison with	T+MA	CL-CNG, T+MA
Dependence on corpora	Does not require parallel or comparable corpora, but it is necessary to compile a dictionary to teach the model. Not limited to bilingual CLPD tasks	Uses a simple dictionary (no probability of translation) as the only translation resource
Dependence on kinship of languages	Applicable in any pair of languages that have any translation resource	Applicable in any pair of languages that have any translation resource
Dependence on machine translator	Depends on the quality of translation of the dictionary for model training	Does not use a machine translation system and does not depend on the availability or quality of machine translation systems
Dependence on the length of document	No data	No data
Productivity	Preserves the precision of the T + MA model without losing recall	More productive than CL-CNG and T+MA

TABLE 4 | A quantitative analysis of the effectiveness of methods for a Russian-English pair.

Article	Algorithm	Precision	Recall	F1	Computation time (seconds)
Cross-language text alignment for plagiarism detection based on contextual and context-free models (Zubarev and Sochenkov, 2019) Negative-1	Sentence embeddings	0.75	0.77	0.76	2.89
	Words substitution	0.84	0.76	0.80	2.63
	NMT	0.85	0.80	0.82	34.15
	NMT-2	0.83	0.64	0.72	240.13
	LR-1	0.91	0.80	0.85	39.68
	LR-2	0.87	0.80	0.85	5.53
	Laser	0.90	0.89	0.89	7.63
	Bert	0.96	0.93	0.95	91.95
Cross-language text alignment for plagiarism detection based on contextual and context-free models (Zubarev and Sochenkov, 2019) Negative-4	Sentence embeddings	0.45	0.77	0.57	4.02
	Words substitution	0.60	0.76	0.66	3.3
	NMT	0.61	0.80	0.69	34.31
	NMT-2	0.54	0.64	0.58	240.29
	LR-1	0.73	0.80	0.76	41.65
	LR-2	0.64	0.80	0.71	7.34
	Laser	0.70	0.89	0.78	11.04
	Bert	0.88	0.93	0.90	197.45
Detection of translated borrowings in large arrays of scientific documents (Kuznecova et al., 2018)	Basic	0.99	0.15	0.26	-
	Proposed	0.93	0.80	0.85	-

TABLE 5 | Comparative table for the used technologies.

Article	Datasets	Vector representation of words
Cross-language text alignment for plagiarism detection based on contextual and context-free models (Zubarev and Sochenkov, 2019)	44 million sentences (for each language): parallel sentences from Opus corpora + sentences from the Yandex Parallel corpus (English-Russian Parallel Corpora, 2015) (16,000 sentence pairs that were not used for learning word embeddings) + parallel concepts from Wikidata + 4,000 sentences manually written by students	Word2Vec
Detection of translated borrowings in large arrays of scientific documents (Kuznecova et al., 2018)	18.5 million parallel sentences from Opus corpora + 10 million sentences from the English version of Wikipedia + articles from journals included in the Russian Science Citation Index (RSCI)	FastText

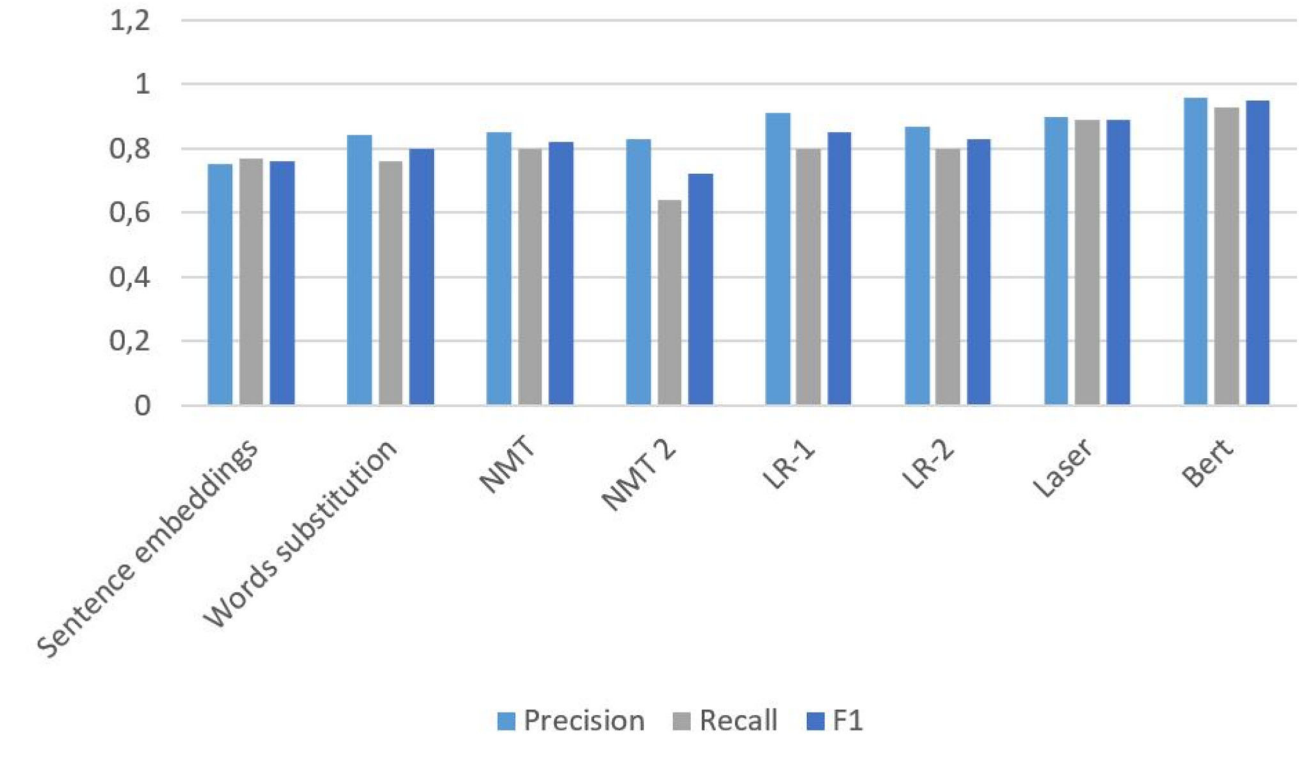

FIGURE 2 | Results of methods from article (Zubarev and Sochenkov, 2019) with Negative-1.

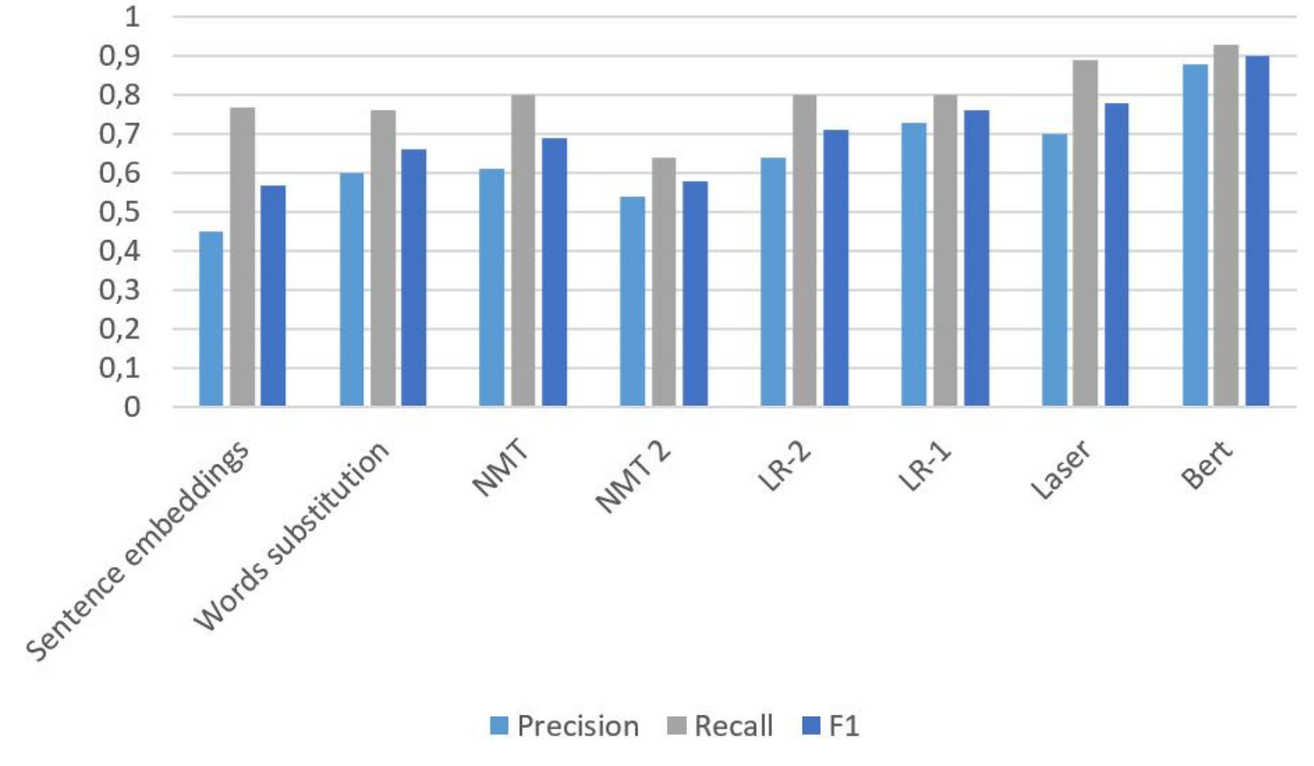

FIGURE 3 | Results of methods from article (Zubarev and Sochenkov, 2019) with Negative-4.

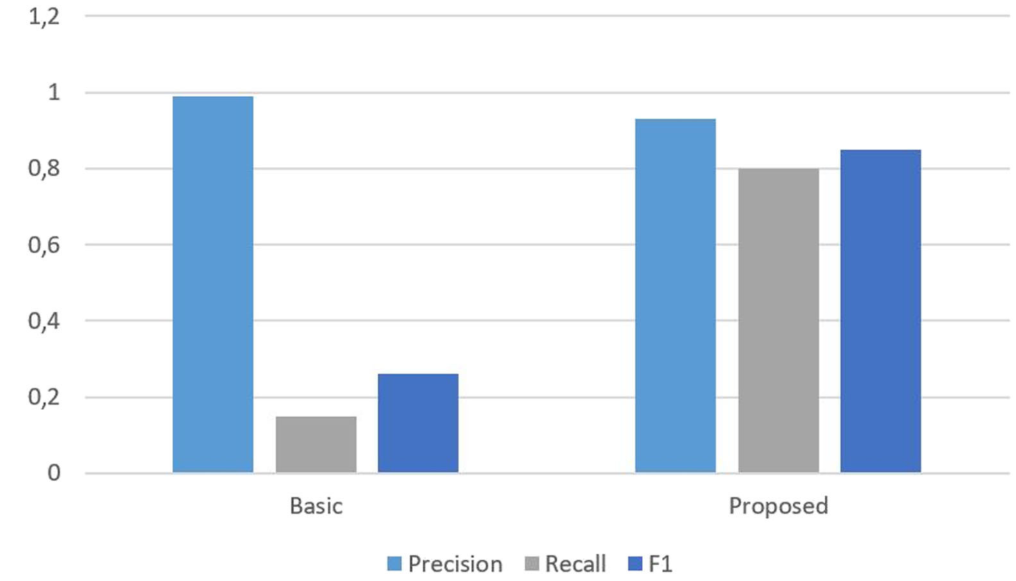

FIGURE 4 | Results of "basic" and "proposed" methods from article (Kuznecova et al., 2018).

detection systems. Cross-language embeddings of words based on large parallel corpora were prepared there to analyze the similarity of two sentences. In turn, these were used to analyze two different similarity ratings: one is based on the sentences embeddings and the other calculated after replacing all the words with the most similar ones in another language. LR-1 classifier showed performance that can be comparable with Bert. This classifier, tuned to maximize recall, can greatly reduce the load on the more sophisticated processing downstream. It is more than two times faster than the Bert. The LASER showed "the golden mean" between F1 and computation time. In addition, its speed can be higher because the authors did not use pre-learned English embeddings. The NMT-2 method had the worst speed, which is even more than the Bert's and the least Recall measure. In all analyzed approaches, a higher number of negative examples (Negative-4) means lesser precision and F1, though recall stayed the same. Computation time is also increased on Negative-4.

With respect to the algorithm proposed in the Russian-language article (Kuznecova et al., 2018), we found out that it showed quite high precision compared to the above methods but is slightly less than the accuracy of the basic algorithm. The high accuracy of the basic algorithm is due to the fact it considered the similarity of only almost-duplicate text. Despite this, the proposed algorithm has better recall and F1 indicators that the basic.

We discovered the most effective methods to evaluate the quantity of plagiarism are Bert and, as proposed in article (Kuznecova et al., 2018), an algorithm used for machine translation. As there is no such characteristic as the time of calculation in Russian language articles, we could not compare them, but the precision, recall, and F1 measure of these approaches showed the best results.

4. CONCLUSION

We conducted a meta-analysis of approaches used for detection of cross-language plagiarism and studied the methods for identifying cross-language plagiarism during the course of this research. Comparison results are shown in **Tables 1–3**.

For the Russian-English pair we perform an in-depth analysis comparing the models by using the following characteristics: precision, recall, F1, computing time, datasets, and vector representation of words. Results are presented in **Tables 4**, **5**, and bar charts in **Figures 2–4**.

We selected machine translation mainly to reduce the task to the analysis of documents in one language. However, the disadvantage of this approach is that repeated sections of the text may not be detected due to the peculiarities of translation and interpretation.

Despite the large number of developed programs for detecting plagiarism, the problem of detecting translated borrowings in weakly related languages is still relevant.

Extending vocabulary can be considered an issue for cross-language plagiarism detection systems. Many available parallel corpora contain a common lexis, but detection of borrowings should also be well applicable and accurate for scientific papers where a multitude of special terms exist. Additionally the next possible solution is to create parallel corpora from comparable corpora with the help of the system for translated plagiarism detection and extend vocabulary with new parallel data.

Based on early research, it is fair to say that CL-CnG model is less effective for the Russian-English pair because these languages are not syntactically related. CL-ESA is more fit for tasks of relatedness than plagiarism detection.

As for other models, not only machine translation but CL-ASA, CL-CTS, and CL-KGA can be used in methods and algorithms to develop a cross-language borrowing verification systems for a Russian-English pair. However, as proposed in articles (Thompson and Bowerman, 2017; Ehsan et al., 2019) models seem to be most suitable for this language pair due to an independence from resources like parallel or comparable corpora, network connectivity, and the availability of online translators throughout the entire text comparison process.

We found that the results of model comparison do not contradict each other in the process of studying the articles and works. In general, the difference in the productivity of models is small, and their application most often depends on the required speed of work, used resources, corpora, dictionaries, as well as the pair of languages analyzed and their relationship.

In future work, it is planned that we apply that which is most suitable for Russian-English pair methods in practice and assess their indicators of Precision, Recall, and F1 to confirm or refute the applicability of the models.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

ATl, ATo, MT, and VK contributed conception and design of the study. ATl wrote the first draft of the manuscript. VK reviewed the first draft and suggested improvements. ATl and MT wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was partly funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities number 2.8303.2017/8.9. The reported research was partly funded by Russian Foundation for Basic Research and the Methods and models for the formation of the digital infrastructure of the scientific and educational cluster of the Republic of Tatarstan, grant N 18-47-160012. The reported study was partly funded by RFBR according to the research project N 19-29-03057.

ACKNOWLEDGMENTS

The work partly was performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

REFERENCES

Alaa, Z., Tiun, S., and Abdulameer, M. (2016). Cross-language plagiarism of Arabic-English documents using linear logistic regression. *J. Theoret. Appl. Inform. Technol.* 83, 20–33.

Amancio, D. R. (2015). Comparing the topological properties of real and artificially generated scientific manuscripts. *Scientometrics* 105, 1763–1779. doi: 10.1007/s11192-015-1637-z

Artetxe, M., and Schwenk, H. (2018). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR abs/1812.10464*. doi: 10.1162/tacl/_a/_00288

Barrón-Cedeño, A., Gupta, P., and Rosso, P. (2013). Methods for cross-language plagiarism detection. *Knowledge Based Syst.* 50, 211–217. doi: 10.1016/j.knosys.2013.06.018

Barrón-Cedeño, A., Rosso, P., Agirre, E., and Labaka, G. (2010). "Plagiarism detection across distant language pairs," in *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference* (Beijing), 37–45.

Ehsan, N., Shakery, A., and Tompa, F. W. (2019). Cross-lingual text alignment for fine-grained plagiarism detection. *J. Inform. Sci.* 45, 443–459. doi: 10.1177/0165551518787696

English-Russian parallel corpora (2015).

Ferrero, J., Agnes, F., Besacier, L., and Schwab, D. (2017a). Using word embedding for cross-language plagiarism detection. doi: 10.18653/v1/W17-2502

Ferrero, J., Besacier, L., Schwab, D., and Agnes, F. (2017b). Deep investigation of cross-language plagiarism detection methods, 6–15. doi: 10.18653/v1/E17-2066

Franco-Salvador, M., Gupta, P., Rosso, P., and Banchs, R. E. (2016a). Cross-language plagiarism detection over continuous-space and knowledge graph-based representations of language. *Knowledge Based Syst.* 111, 87–99. doi: 10.1016/j.knosys.2016.08.004

Franco-Salvador, M., Rosso, P., and y Gamez, M. M. (2016b). A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Inform. Process. Manage.* 52, 550–570. doi: 10.1016/j.ipm.2015.12.004

Google Scholar (2004).

Hanane, E., Erritali, M., and Oukessou, M. (2016). "Semantic similarity/relatedness for cross language plagiarism detection," in *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)* (Morocco). doi: 10.1109/CGiV.2016.78

Horsley, T., Dingwall, O., and Sampson, M. (2011). Checking reference lists to find additional studies for systematic reviews. *Cochrane Datab. System. Rev.* 8:MR000026. doi: 10.1002/14651858.MR000026.pub2

IEEE (2019). *Advanced Technology for Humanity*.

IEEE-Faq (2019). *Advanced Technology for Humanity*.

Kent, C. K., and Salim, N. (2010). "Web based cross language plagiarism detection," in *2010 Second International Conference on Computational Intelligence, Modelling and Simulation* (Tuban), 199–204. doi: 10.1109/CIMSiM.2010.10

Kuznecova, R. V., Bahteev, O. U., and Chekhovich, U. V. (2018). Detection of translated borrowings in large arrays of scientific documents (Detektirovanie perevodnyh zaimstvovanij v bolshih massivah nauchnyh dokumentov).

McNamee, P., and Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. *Inform. Retriev.* 7, 73–97. doi: 10.1023/B:INRT.0000009441.78971.be

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., et al. (2009). Prisma statement.

Navigli, R., and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.* 193, 217–250. doi: 10.1016/j.artint.2012.07.001

Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011). Cross-language plagiarism detection. *Lang. Resour. Eval.* 45, 45–62. doi: 10.1007/s10579-009-9114-z

Thompson, V., and Bowerman, C. (2017). Detecting cross-lingual plagiarism using simulated word embeddings.

Tlitova, A. E., and Toschev, A. S. (2019). Review of existing tools for detecting plagiarism and self-plagiarism (Obzor sushchestvuyushchih instrumentov vyyavleniya plagiata i samoplagiata). *Elektronnye Biblioteki* 22, 143–159. doi: 10.26907/1562-5419-2019-22-3-143-159

Zubarev, D., and Sochenkov, I. (2019). Cross-language text alignment for plagiarism detection based on contextual and context-free models.

Conflict of Interest: 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.