frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 31 October 2024
DOI 10.3389/fcomp.2024.1472512

OPEN ACCESS

EDITED BY
Stavros Asimakopoulos,
Lancaster University, United Kingdom

REVIEWED BY
Maha Khemaja,
University of Sousse, Tunisia
John Samuel,
École Supérieure de Chimie Physique Electronique de Lyon, France

*CORRESPONDENCE
Cristian Santini
✉ c.santini12@unimc.it

RECEIVED 29 July 2024
ACCEPTED 16 October 2024
PUBLISHED 31 October 2024

CITATION
Santini C (2024) Combining language models for knowledge extraction from Italian TEI editions. *Front. Comput. Sci.* 6:1472512. doi: 10.3389/fcomp.2024.1472512

Combining language models for knowledge extraction from Italian TEI editions

Cristian Santini*

Department of Humanities, University of Macerata, Macerata, Italy

This study investigates the integration of language models for knowledge extraction (KE) from Italian TEI/XML encoded texts, focusing on Giacomo Leopardi's works. The objective is to create structured, machine-readable knowledge graphs (KGs) from unstructured texts for better exploration and linkage to external resources. The research introduces a methodology that combines large language models (LLMs) with traditional relation extraction (RE) algorithms to overcome the limitations of current models with Italian literary documents. The process adopts a multilingual LLM, that is, ChatGPT, to extract natural language triples from the text. These are then converted into RDF/XML format using the REBEL model, which maps natural language relations to Wikidata properties. A similarity-based filtering mechanism using SBERT is applied to keep semantic consistency. The final RDF graph integrates these filtered triples with document metadata, utilizing established ontologies and controlled vocabularies. The research uses a dataset of 41 TEI/XML files from a semi-diplomatic edition of Leopardi's letters as case study. The proposed KE pipeline significantly outperformed the baseline model, that is, mREBEL, with remarkable improvements in semantic accuracy and consistency. An ablation study demonstrated that combining LLMs with traditional RE models enhances the quality of KGs extracted from complex texts. The resulting KG had fewer, but semantically richer, relations, predominantly related to Leopardi's literary activities and health, highlighting the extracted knowledge's relevance to understanding his life and work.

KEYWORDS

large language models (LLMs), knowledge extraction, Semantic Web, Wikidata, TEI/XML, Giacomo Leopardi, entity linking, relation extraction

1 Introduction

In the field of Digital Humanities (DH), extracting knowledge from digitized texts is critical for advancing research and enhancing understanding. This task becomes especially important when handling large collections of cultural heritage materials, such as the works of Giacomo Leopardi (1798–1837). Leopardi, born in Recanati, a small town in central Italy, is widely regarded as one of the most important authors in Italian literature. He was a philologist and prose writer, though he is best known for his poetry, which has been translated into more than twenty languages. Currently, over 15,000 digitized facsimiles of manuscripts handwritten by Leopardi are available across various online platforms.

Knowledge extraction (KE) methods are essential when working with extensive collections of historical texts from literary authors. This is because general-purpose

knowledge graphs (KGs), such as Wikidata[1] or DBpedia,[2] often do not include all the entities and relationships referenced within a specific corpus. For instance, information related to a writer present in Wikidata may omit significant details about their private life, such as personal relationships or secondary occupations, or references to lesser-known works. In the case of a historical author such as Giacomo Leopardi, his writings may refer to entities that are entirely absent from existing knowledge graphs. Consequently, KE methods are pivotal for discovering new entities and facts absent from the current Linked Open Data (LOD) ecosystem and for revealing new relationships between existing graph nodes in the Semantic Web.

KE involves a series of techniques used to convert unstructured textual information into KGs, which can be queried and explored. A KG is automatically extracted using two primary techniques: entity linking (EL) and relation extraction (RE). EL identifies references to entities and concepts in a text and determines the appropriate entry in a knowledge base to which the reference should be linked. RE algorithms establish whether two entities are connected by specific relations, typically defined using a controlled vocabulary or ontology. The introduction of language models has significantly transformed the RE field, particularly with the advent of pre-trained language models (PLMs), which enable direct extraction of `[subject, predicate, object]` triples from texts in an end-to-end manner. Recently, Trajanoska et al. (2023) leveraged the instruction-following abilities of large language models (LLMs) to create KGs from text prompts, providing a flexible methodology adaptable to various domains.

In recent years, substantial progress has been made in employing PLMs for KE. Huguet Cabot and Navigli (2021) introduced a sequence-to-sequence (seq2seq) model for transforming English texts into `[subject, predicate, object]` triples using Wikidata properties. This model, known as REBEL, set a benchmark for semantic KE from unstructured texts. Building on this work, Huguet Cabot et al. (2023) developed mREBEL, a multilingual version of the model, making it applicable to different languages and enhancing its versatility. The development of LLMs has further advanced KE. Xu et al. (2023) demonstrated the potential for few-shot relation extraction using LLMs, underscoring their capability to perform KE with minimal training data. Ma et al. (2023) showed how LLMs could rank the RE output of smaller PLMs, optimizing the extraction process. In addition, Li et al. (2024) improved KE outputs through step-by-step prompting techniques, illustrating how LLMs can iteratively refine results.

Despite the increased effectiveness of LLMs over traditional KE methods in understanding complex texts—such as those written in different historical periods or belonging to specific genres—a common limitation remains. LLMs still produce textual representations, rather than a *de facto* knowledge graph. To transform a set of `[subject, predicate, object]` triples generated by a language model into a KG, the entities and relations must be linked to a knowledge base. This requires the application of effective RE methods to accurately link semantically similar predicates to appropriate properties in external ontologies.

This research aims to bridge the gap between LLM-based KE studies and domain-specific approaches for EL and RE in the Digital Humanities (DH) field. It proposes a system that integrates KE techniques applied to TEI/XML transcriptions of Italian literary texts, with the objective of extracting formal, machine-readable representations of these documents that can be queried, explored, and linked to external resources. The novelty of this approach lies in addressing the limitations of existing relation extraction models, such as REBEL, when applied to Italian literary documents. This is achieved by employing ChatGPT (Openai, 2023) to transform unstructured texts into semi-structured formats that can be more easily interpreted by general-purpose pre-trained models. This research originates from a digitization project focusing on the works of Giacomo Leopardi, preserved in various Italian and international institutions (Melosi and Marozzi, 2021). The objectives of this study are three-fold:

1. To apply instruction-tuned LLMs for extracting knowledge from literary texts, specifically in the context of TEI-encoded documents.
2. To propose a novel method that combines the strengths of LLMs with traditional RE algorithms for generating knowledge graphs using Wikidata properties from TEI-encoded literary editions.
3. To perform an initial evaluation of the proposed approach on a selected collection of Leopardi's letters, assessing the effectiveness and accuracy of the knowledge extraction process.

Section 2 provides an overview of related work concerning KE for texts in the DH field. Section 3 describes the proposed approach and the dataset used as a case study to evaluate the quality of the resulting knowledge graph based on a selection of Leopardi's works. Section 4 presents the results obtained, comparing our approach with a simple baseline and analyzing the extracted content. Finally, Section 5 offers insights on the results, discusses the advantages and limitations of the proposed method, and briefly outlines future research directions.

2 Related work

KE has garnered significant attention in the DH domain due to the wide range of applications it enables. By converting unstructured texts into structured KGs that capture entities and their relationships, researchers are equipped with powerful tools to access and explore extensive humanities corpora. KGs based on OWL/RDF ontologies also support automated reasoning, allowing scholars to uncover new insights from large, interconnected datasets. Sporleder (2010) provides a comprehensive overview of KE tasks within the cultural heritage domain, discussing the challenges of applying NLP techniques for data processing, knowledge extraction, and metadata extraction. A more recent survey by Santini et al. (2022) focuses on KE for Renaissance art-historical texts, highlighting issues related to tasks such as named entity recognition (NER), entity linking (EL), and artwork recognition.

The process of transforming texts into RDF graphs typically relies on two core NLP techniques: EL and RE. EL plays a critical

1 https://www.wikidata.org/

2 https://www.dbpedia.org/

role in identifying entity mentions in a text and linking them to a corresponding knowledge base, resolving ambiguities effectively (Sevgili et al., 2022). RE, in contrast, determines the relationships between identified entities based on a predefined ontology (Zhao et al., 2024). Auxiliary tasks such as NER, coreference resolution (CR), and temporal resolution enhance the accuracy and granularity of the extracted knowledge. NER categorizes entities into types such as person, organization, or location; CR ensures consistent representation of entities throughout a text by clustering names and pronouns; and temporal resolution provides a chronological context for interpreting events and facts.

Several NLP libraries, such as StanfordNLP (Manning et al., 2014), SpaCy (Vasiliev, 2020), and GATE (Cunningham, 2002), support many of these tasks. However, they often face challenges when applied to historical texts and specialized domains. Historical documents introduce additional complexities, including language variation, optical character recognition (OCR) errors, and noisy input data, as noted by Ehrmann et al. (2021). These factors necessitate the development of tailored models designed to address the specific challenges posed by historical texts, particularly for tasks such as NER and EL, where generic models tend to underperform.

In the field of Digital Humanities, EL has become a crucial task as accurate entity disambiguation is essential for constructing reliable KGs. Several studies have utilized off-the-shelf tools for this purpose. For instance, Ruiz and Poibeau (2019) applied DBpedia Spotlight to the Bentham Corpus, while van Hooland et al. (2015) evaluated entity extraction tools on the descriptive fields of the Smithsonian Cooper-Hewitt National Design Museum. Despite their contributions, these studies share common limitations, such as focusing exclusively on English texts and the absence of domain-specific datasets. Brando et al. (2015) addressed EL in French literary texts encoded in TEI, and Linhares Pontes et al. (2022) tackled EL in multilingual historical press articles, offering more domain-specific solutions. However, none of these studies fully addressed the task of extracting relationships between entities to construct comprehensive KGs.

RE has also been explored in several works aimed at improving the accessibility of literary and historical corpora. Reinanda et al. (2013) proposed a hybrid approach combining association finding and RE to construct entity networks, primarily for historical and political documents. Their approach effectively captures explicit and implicit relationships using both statistical co-occurrence measures and machine learning models. However, it encounters difficulties with domain-specific complexities, particularly in multilingual humanities corpora. Similarly, Santini et al. (2024) applied measures from information theory, i.e. that is, pointwise mutual information, to extract social networks from Giorgio Vasari's *Lives of the Artists*, linking artists by means of textual references. Although successful in building KGs, these works' reliance on co-occurrence methods limits their ability to capture more nuanced, event-driven relations.

Other studies have ventured into Open Information Extraction (Open IE) to extract more complex relationships without relying on a predefined vocabulary or ontology. Graham et al. (2020) employed Open IE techniques to extract financial and employment ties, though their approach has limited applicability to the humanities due to the complexity of historical language. Jain et al. (2022) applied Open IE to art history using pre-trained models from StanfordNLP and SpaCy to generate flexible KGs without relying on predefined ontologies. However, their method suffers from noise and lacks precision, particularly in canonicalizing entities and relations. Graciotti (2023) combined Open IE with Abstract Meaning Representation (AMR) to handle multilingual and historical texts in the musical heritage domain. Their use of the *Text2AMR2Fred* pipeline (Gangemi et al., 2023) provides more detailed relations based on an event-centric graph modeled with PropBank *frames*. Nonetheless, their reliance on neural parsers trained on contemporary data introduces limitations in processing historical documents.

Despite the progress in these studies, several challenges persist in KE for humanities texts. Notably, the lack of domain-specific RE models capable of handling the linguistic and syntactic complexities of multilingual and historical corpora remains a major limitation. The absence of standardized benchmarks and resources further complicates model development, resulting in inconsistencies in how relations are extracted and represented across different projects. In addition, while some efforts have been made to harmonize KG outputs with external ontologies such as Wikidata, challenges in entity linking and ontology mapping continue across various works. Finally, although LLMs show promise in improving RE, their effectiveness in extracting RDF graphs from diachronic, historical texts has yet to be fully explored.

3 Materials and methods

3.1 Case study

Platforms such as LiberLiber,[3] Wikisource,[4] and Biblioteca Italiana[5] contain nearly every digital transcription of Leopardi's texts. These transcriptions present a vast amount of information, which can be difficult to navigate without the assistance of domain experts. A particular focus should be placed on Leopardi's correspondence, which includes private anecdotes and factual knowledge often absent from Wikipedia and the Wikidata KG. KE algorithms to uncover the entire network of entities and relations referenced in such documents represent a fascinating challenge. Among the publicly available collections of Leopardi's digitized works, the one hosted by the Cambridge University Digital Library (CUDL)[6] is particularly noteworthy. This portal includes 41 manuscripts: 36 letters sent by Giacomo Leopardi to various correspondents, two brief "translation essays" from classical languages authored by Leopardi, and three additional letters-one by Monaldo Leopardi and two by Paolina Leopardi, addressed to different recipients.

The CUDL collection covers a wide range of topics. Some letters provide information about Leopardi's works, while others discuss his health and his role as a "copy editor" for his friend Giovanni Rosini. Notably, many of the letters addressed to Antonio Fortunato Stella, a publisher based in Milan who printed numerous

3 https://liberliber.it/

4 https://it.wikisource.org/

5 http://www.bibliotecaitaliana.it/

6 https://cudl.lib.cam.ac.uk/collections/leopardi/1

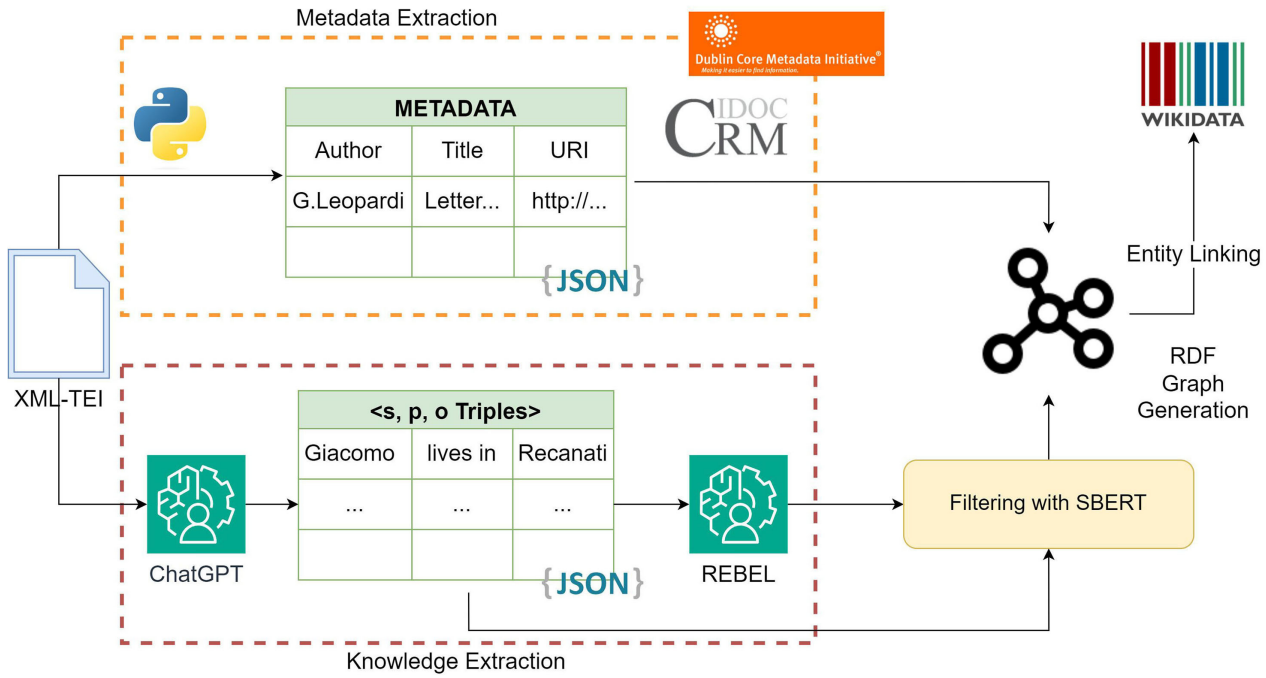

FIGURE 1
Schema of our approach: the process begins with the preprocessing of transcriptions into a structured JSON format using ChatGPT, which identifies entities and relations. The output is then fed into REBEL, which maps these elements to RDF statements using Wikidata properties. The final KG is enriched with metadata and linked to Wikidata.

works by Leopardi, are part of this collection. In addition, the two translation pieces exemplify Leopardi's work as a philologist. This edition is the result of a project led by Gioele Marozzi in collaboration with Cambridge Digital Humanities. Each document is accompanied by a high-resolution image and supplemented with a TEI/XML file that includes a diplomatic transcription of the autograph. Furthermore, each item is annotated with metadata such as title, author, creation date, and letter recipient. In this study, we used these 41 TEI/XML files as the reference dataset to apply our KE pipeline and generate the first version of a KG automatically extracted from Leopardi's texts.

3.2 Methodology

The aim of this study is to extract structured knowledge in the form of a KG from a TEI/XML edition of Leopardi's autographs in Italian, as hosted in CUDL. To achieve this, we structured our pipeline into a series of extraction steps. Initially, a TEI/XML file is processed by a script that parses the XML to extract metadata and text from the edition. Specifically, the metadata considered include the document identifier, repository, title, language support (e.g., "paper with watermark"), extent, place of origin, date of origin, sender, and receiver. Text and metadata are extracted using the `lxml` Python library, which retrieves the text contained within the corresponding XML elements. In addition, during this extraction step, annotations for persons and places found in the `tei:listPerson` and `tei:listPlace` fields are extracted. References to these entities are then identified within the text to capture their surface forms. Finally, the portions of text in which these entities appear are mapped to Wikidata by utilizing the VIAF and GeoNames identifiers provided in the TEI/XML file.

After the extraction step, various language models are employed to generate RDF/XML triples from the unstructured text of the digital edition. This step involves the following components:

1. Zero-shot Triple Extraction with LLM: The proposed methodology involves generating textual triples (e.g., `["Giacomo" "livesIn" "Recanati"]`) in English from the Italian text using a multilingual instruction-tuned LLM such as ChatGPT-4.
2. Relation Extraction with seq2seq model: In this step, the properties in the generated triples are mapped to the Wikidata schema using REBEL (Huguet Cabot and Navigli, 2021), a seq2seq model that extracts Wikidata relations from natural language triples.
3. Filtering with SBERT: A filtering algorithm calculates the similarity between the triples generated by the two previous components by encoding both into an embedding space using SBERT (Reimers and Gurevych, 2019), discarding results that fall below a specified threshold.

The entire pipeline is depicted in Figure 1. The following subsections provide a detailed explanation of how each component processes the data sequentially.

3.2.1 Zero-shot triple extraction with LLM

In this step, an instruction-tuned LLM is employed to extract a preliminary set of `[subject, predicate, object]` triples from the text of a TEI file, without defining a specific

```
prompt = """Extract RDF triples from the following text in Italian:{title}.
Use expressions in English for properties, such as ":dateOfBirth" or
 ":placeOfDeath".
Write triples which have as subject or object people, places, organizations
 and works referenced.
Write the output triples in the following JSON format: [subject, property,
 object].
Check the JSON output to control syntax errors.
Input: {text}"""
```

FIGURE 2
Prompt used to extract `[subject, predicate, object]` triples using ChatGPT.

schema of relations. To achieve this, ChatGPT is queried using the prompt presented in Figure 2 and is provided with the title and text extracted from the TEI/XML file. The LLM then produces a list of triples `[[subject, predicate, object]...]`, which is subsequently converted into a JSON file for further processing. A sample of the output from this step is displayed in Figure 3.

It is important to highlight that the input text is converted into a list of triples in English. This conversion simplifies the extraction process for REBEL, a model trained to extract semantic triples from the English Wikipedia, which cannot be applied directly to texts in Italian. We opted to use REBEL, a monolingual RE model trained on English, instead of its multilingual variant, mREBEL (Huguet Cabot et al., 2023), due to the superior performance observed in our experiments, as demonstrated in Section 4.

A key limitation of this step, however, is that the generated `[subject, predicate, object]` triples do not yet form a complete KG. This is due to the fact that the predicates are freely generated and do not conform to an ontology or metadata schema. To transform the triples into statements that can be serialized in an RDF graph, each property must be linked to its corresponding alias in an external ontology or metadata schema, such as Wikidata. To accomplish this, a relation extraction model is required to classify the relationships between entities in accordance with an external vocabulary or ontology.

3.2.2 Relation extraction with seq2seq model

In this step, the output of ChatGPT is processed into a string by converting the JSON representation into plain text. For example, the triple in Figure 3, `["Paolina Leopardi", ":locationOfWriting", "Recanati"]` is automatically converted into *"Paolina Leopardi location of writing Recanati"* by concatenating the items in the triple with white spaces and by normalizing the predicate into its corresponding literal form (e.g., `locationOfWriting` becomes *location of writing*). This conversion is performed since the RE model extracts relations from plain text and therefore needs to be applied to natural language strings.

More specifically, each triple generated by ChatGPT is rejoined into a string and independently fed into REBEL, a seq2seq model that processes natural language texts and outputs a dictionary of the form `{"head":str, "type":str, "tail":str}`, where


```
[
    [
        "Paolina Leopardi",
        ":wroteLetterTo",
        "Filippo Raffaelli"
    ],
    [
        "Paolina Leopardi",
        ":dateOfLetter",
        "8 Dicembre 1852"
    ],
    [
        "Paolina Leopardi",
        ":locationOfWriting",
        "Recanati"
    ],
    [
        "Paolina Leopardi",
        ":expressesInabilityToProvide",
        "details on Canonico Vogel"
    ]
]
```


FIGURE 3
Sample of the output of ChatGPT4 for the prompt in Figure 2.

"head" and "tail" correspond to the subject and object entities of a triple, and "type" represents the relation between them, based on a semantic property from Wikidata. For instance, when the input *"Paolina Leopardi location of writing Recanati"* is given to REBEL, it outputs the dictionary `{"head":"Paolina Leopardi", "type":"work location", "tail":"Recanati"}`, where "work location" corresponds to the Wikidata property https://www.wikidata.org/entity/P937.

In the proposed pipeline, each triple produced by ChatGPT is independently fed into the `rebel-large`[7] model, which is parameterized to produce one dictionary (i.e., one triple) for each input string. The main advantage of this step is that the relations in the newly generated triples come from a predefined set of Wikidata properties, which may include logical constraints such as symmetry or asymmetry, enabling reasoning (Shenoy et al., 2022). For example, the property "sibling" (https://www.wikidata.org/entity/P3373) is defined in Wikidata as a symmetric property. Consequently, if the triple `["Giacomo Leopardi", "sibling", "Paolina Leopardi"]` is extracted, its symmetric form `["Paolina Leopardi", "sibling", "Giacomo Leopardi"]` can also be included in the KG. Another key advantage is that each triple can now be serialized into an RDF statement using Wikidata URIs for relations. This is critical as it allows the creation of an interoperable KG, in which specific relations or graph patterns can be queried using SPARQL (Hogan et al., 2021).

Finally, by utilizing the links between person names and VIAF identifiers, as well as place names and GeoNames identifiers present in the TEI/XML file, it is possible to map some of the "head" and "tail" entities of the triples to the corresponding Wikidata items. This facilitates the inclusion of new statements in Wikidata. For example, in the dictionary where Paolina Leopardi is the "head" and Recanati is the "tail" of the triple, Paolina is linked to the Wikidata item https://www.wikidata.org/entity/Q3893652, which is associated with VIAF identifier 29611067, while Recanati is linked to https://www.wikidata.org/entity/Q83362, corresponding to the GeoNames identifier 6541846. In conclusion, as the generation of triples in this step is automatic and performed using a neural seq2seq model, a third component is necessary to minimize the risk of errors during this process. This component filters the generated triples to ensure accuracy and coherence before integration into the KG.

3.2.3 Filtering with SBERT

REBEL is a seq2seq BERT model that generates triples based on Wikidata from natural language texts. However, this model is trained on Wikipedia and is not optimized to process texts containing complex factual knowledge, such as Leopardi's letters. As a result, the model can hallucinate and produce false triples that are syntactically correct but semantically inaccurate. For example, the triple `["Giacomo Leopardi", ":sentLetterTo", "Antonio Fortunato Stella"]` can be incorrectly translated by REBEL into `["Giacomo Leopardi", "relative", "Antonio Fortunato Stella"]`, even though the input text does not support this relationship. To mitigate such semantic inaccuracies, we integrated a third component to filter out incorrect triples based on semantic similarity.

Our filtering approach uses SBERT, a Sentence Transformer that encodes long texts such as sentences and paragraphs into embeddings (vectors), representing the meaning of the texts in a linear space (Reimers and Gurevych, 2019). This process ensures that triples with different meanings can be discarded from the output KG as they are located farther apart in the vector space. To apply SBERT to both the output of the triple extraction step (ChatGPT) and the output of the relation extraction step (REBEL), both triples are converted into strings. ChatGPT's triples are converted in the same way described above, while REBEL triples are transformed into strings by simply concatenating the elements of the `[subject, predicate, object]` triple with white spaces. By applying a threshold over the cosine similarity of the embeddings obtained from the two strings, it becomes possible to filter out results that are not semantically aligned.

In our experiments, we used the model `all-mpnet-base-v2`[8] as Sentence Transformer and applied a threshold of 0.9 on the cosine similarity. This high threshold allows us to keep in the KG triples that are very similar to the output of first step, reducing the presence of inaccurate triples generated by REBEL and preserving the correctness of the KG.

3.2.4 RDF Graph generation and entity linking

Finally, the triples extracted by our pipeline are combined with the metadata from the TEI/XML edition to form a KG. To generate the final RDF graph, we reused ontologies and controlled vocabularies that are well-established in the LOD domain. The document is represented using the `E31_Document` class from the CIDOC-CRM ontology. The same ontology is applied to model the relationship with the URI of the organization owning the original text, through the property `P52_has_current_owner`, and to link the document to its display URI with the property `P138i_has_representation`. The Dublin Core vocabulary is used to model document attributes such as title, date, extent, and language.

The triples extracted by our pipeline are represented using RDF reification. In the KG, each triple is represented by a node of type `rdf:Statement`, which includes a label as a natural language representation of the triple and four additional properties: `rdf:subject`, `rdf:predicate`, `rdf:object`, and `dcterms:source`. These properties, respectively, express the relationship between the subject, predicate, object, and source document of each triple. An example of the RDF graph to be obtained is shown in Figure 4. As illustrated, each extracted entity and property can be linked to its corresponding item in Wikidata through the `rdfs:seeAlso` property, facilitating disambiguation and enabling reasoning. A Turtle serialization of the KG extracted from the manuscripts in CUDL has been made publicly available in our GitHub repository.[9]

4 Results

To evaluate the quality of the extracted KG, we compared our pipeline with a simple baseline for end-to-end multilingual RE proposed in Huguet Cabot et al. (2023), namely, mREBEL. This baseline was chosen because, to the best of our knowledge, it is the only RE model trained on Wikidata properties that

7 https://huggingface.co/Babelscape/rebel-large

8 https://huggingface.co/sentence-transformers/all-mpnet-base-v2

9 https://github.com/sntcristian/leopardi_kg

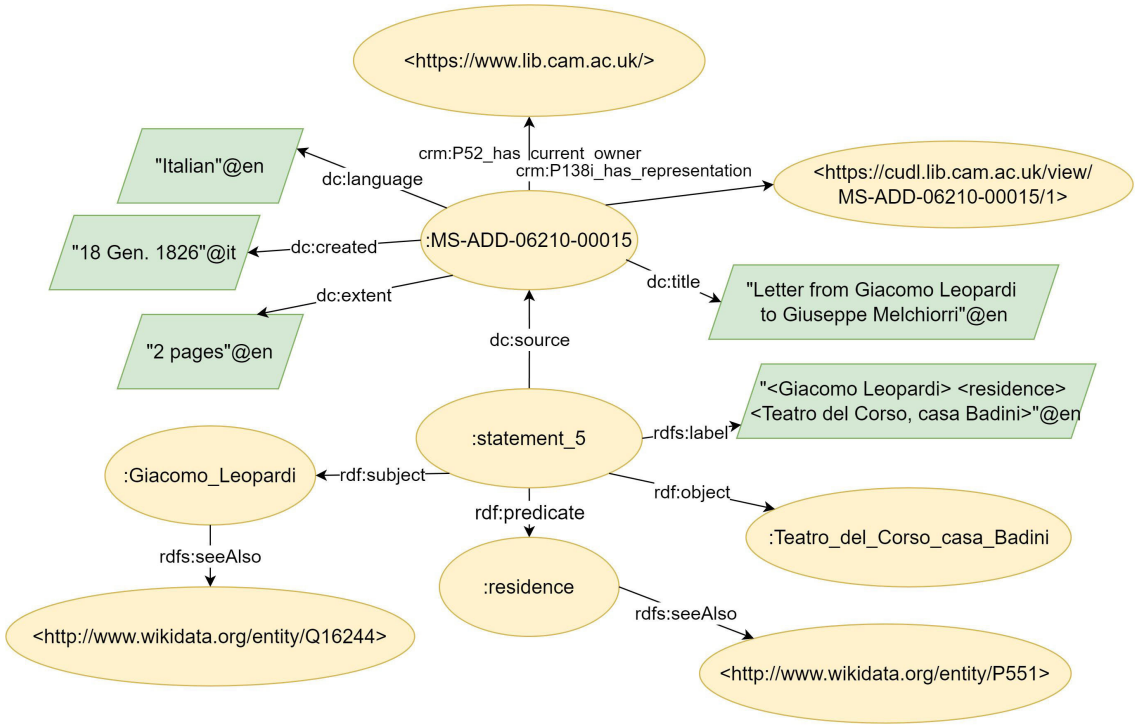

FIGURE 4
Sample of the RDF Graph for a document written by Leopardi contained in the CUDL collection. While document metadata are represented with Dublin Core and CIDOC-CRM, triples automatically extracted from the text are modeled with RDF reification.

is available for Italian. Due to the absence of a benchmark for evaluating RE models on historical Italian literary texts, we adopted different KG quality control metrics to estimate the consistency and accuracy of the knowledge representation produced by various methods. We referenced Wang et al. (2021) to identify the dimensions of KG evaluation (such as semantic accuracy, interlinking, and redundancy) and the specific metrics to be used. Among the metrics discussed in the study, two were considered suitable for our evaluation:

- Semantic accuracy: calculated as the ratio between the triples that represent real-world facts and the total number of triples in the KG.
- Consistency: measured as the ratio of non-contradicting statements to the total number of statements.

Other measures of KG quality, such as completeness and relevancy, were not applied to this analysis due to the challenges of realizing a ground truth containing all the real-world facts which have to be extracted from the corpus and the impossibility of estimating the degree of relevance of the information extracted without a differentiated team of domain experts and users. Moreover, in the assessment of semantic accuracy, we considered a triple to be correct only if all the following three criteria are satisfied:

1. the entities used as subjects or objects of a triple should be unambiguous (e.g., no expression such as "his spouse");
2. the semantics of the property used should exactly match the factual relation between two entities;
3. the fact expressed by a triple should not only be true but also explicitly mentioned in the text.

Metrics on the semantic accuracy and consistency of the KG obtained with our method are reported in Table 1.

As demonstrated, the application of REBEL on a semi-structured text generated by ChatGPT, combined with a filtering approach, significantly strengthens the results in terms of both compliance with real-world facts and consistency. The lower performance of mREBEL can be explained by two factors. First, this model was trained to extract `[subject, predicate, object]` triples from small excerpts of Wikipedia. As a result, it underperforms when applied to longer and more complex texts such as Leopardi's letters. Second, mREBEL is a multilingual model trained on contemporary languages and is not equipped to handle the intricacies and challenges of texts written in nineteenth-century Italian, including the peculiarities of the abbreviations used by Leopardi and the complex intertextual references found in his correspondence. For instance, what Leopardi refers to as *medaglia* (in English, *medal*) in letter `MS-ADD-06210-00010`[10] actually denotes a coin, not a decoration.

An ablation study was conducted to compute intermediate statistics on the extracted graphs at each step of the pipeline. For each sequence of KE processes in Section 3.2, three MultiDiGraphs were created from the textual triples, and the total number of

10 https://cudl.lib.cam.ac.uk/view/MS-ADD-06210-00010/1

TABLE 1 Comparison between the baseline model and our pipeline for number of triples, ratio of semantically accurate triples, and ratio of consistent triples.

Approach	No. of triples	Semantic accuracy	Consistency
mREBEL (baseline)	40	0.1	0.75
Combined LMs (ours)	58	0.67	0.93

TABLE 2 Number of triples, number of unique entities, and number of unique relations at each stage of the pipeline.

Components	Triples	Unique entities	Unique relations	Semantic accuracy
Triple extraction	662	764	400	0.92
Triple extraction + relation extraction	431	435	66	0.24
Triple extraction + relation extraction + filtering	58	98	18	0.67

TABLE 3 First 10 most frequent entities in our KG, number of statements sorted in decreasing order, and Wikidata entity to which it was linked (if possible).

Entity label	No. of statements	Wikidata ID
Giacomo Leopardi	19	Q172599
Roma	14	Q15119
Giuseppe Melchiorri	8	Q88781669
Carlo Emmanuele Muzzarelli	4	Q5041499
Giovanni Rosini	4	Q4396614
Francesco Cancellieri	3	Q3612445
Lettere sopra la condotta di Bonaparte a Sant'Elena	2	Not available
Commentario Storico de Ecclesiis Recanatensi et Lauretana	2	Not available
Luisa Strozzi	2	Not available
flussion d'occhi	2	Not available

triples, unique entities, and unique relations was counted for each one. As previously mentioned, the obtained graph is not *de facto* a knowledge graph until relation extraction with REBEL is applied. Only when Wikidata relations are extracted from ChatGPT's textual triples can a knowledge graph be created, where relations between entities are mapped to properties in the Wikidata schema. The metrics for the ablation study are reported in Table 2, which also includes semantic accuracy ratios for each KE step.

The ablation study highlighted the challenges of applying relation extraction to domain-specific texts. In fact, when using REBEL, the semantic accuracy of the graph drops significantly compared to the semantic accuracy achieved by ChatGPT during triple extraction. This is because, as mentioned earlier, both REBEL and mREBEL are trained on Wikipedia and are not well-suited for complex texts, such as Leopardi's correspondence. However, by applying the proposed filtering approach, it is possible to drastically improve the accuracy of the relation extraction process, demonstrating the effectiveness of our method, even with texts that are outside the domain in which the RE model was trained.

Moreover, a clear effect of our knowledge extraction strategy is that the graph of entities and relations becomes denser at each stage, with fewer relations that nonetheless convey more meaningful semantics. The loss of information caused by performing relation extraction is, in fact, beneficial for the output knowledge representation as the triples extracted by ChatGPT generate relations freely, without adhering to the Wikidata schema. Extracting relations with REBEL allows for semantic queries via SPARQL on the extracted KG by utilizing a more restricted set of relations and enables reasoning by leveraging logical properties, such as the symmetry or transitivity of Wikidata predicates, as discussed in Section 3.2.

The KG obtained with our pipeline overall contains 98 entities, out of which 10 were linked to Wikidata, and 58 relations based on the Wikidata schema. In addition, the obtained KG was queried using SPARQL to identify the types of entities and relations that appear most frequently in the graph. Tables 3, 4 present, respectively, the top 10 most frequent entities and Wikidata relations in the final graph obtained through triple extraction, relation extraction, and filtering. It is noteworthy that relations related to the author's activity as a writer are the most frequent, followed closely by those concerning Leopardi's health, a central theme in the author's life. Table 3 indicates the corresponding Wikidata ID to which each entity in our KG is linked (where applicable). Table 4 provides an example of a triple in the extracted KG for each Wikidata property. For readability purposes, the triples in the table are not serialized in RDF. An RDF serialization of the extracted graph is available online.[11]

5 Discussion

Combining language models to perform KE on a dataset of Italian literary texts can be considered a successful strategy for enhancing the quality of the extracted KG. The results of this study demonstrate how the combination of a multilingual instruction-tuned LLM with REBEL and a filtering approach enables the extraction of RDF statements that are predominantly accurate, with a semantic accuracy of *0.67* and a consistency of *0.93*, outperforming a multilingual RE baseline trained on Wikipedia/Wikidata.

The most significant advantage of our approach lies in its ability to combine the natural language understanding capabilities of LLMs with the use of Wikidata properties in the extracted statements. This feature of our KE pipeline enables semantic queries via SPARQL on the extracted KG and supports reasoning by leveraging logical properties such as the symmetry or transitivity of Wikidata predicates. Another noteworthy benefit of our system is its capacity to enhance the explainability and reliability of the KE process. Seq2seq RE models such as mREBEL extract multiple

11 https://sntcristian.github.io/leopardi_kg/results/leopardi_kg_v1.ttl

TABLE 4 First 10 most frequent Wikidata properties in our KG accompanied by their number of statements sorted in decreasing order and example of triple in our KG.

Property label	No. of statements	Example
Work location	14	["Giuseppe Melchiorri", "work location", "Rome"]
Author	12	["Osservazioni Eusebiane", "author", "Giacomo Leopardi"]
Field of word	4	["Gasparo Mazzi", "field of work", "naturalist"]
Notable work	4	["Giacomo Leopardi", "notable work", "Greek inscription judged by Giacomo Leopardi"]
Medical condition treated	3	["Giacomo Leopardi", "medical condition treated", "eye and nerve disease"]
Part of	3	["Epigramma sopra Amore", "part of", "Libretto del Mosco"]
Position held	3	["Luigi Giambene", "position held", "General Secretary of the Pontifical Posts"]
Instance of	2	["Commentario Storico de Ecclesiis Recanatensi et Lauretana", "instance of", "manuscript"]
Place of death	2	["Giovanni Rosini", "place of death", "location unknown"]
Relative	2	["Francesco Galvani", "relative", "Mario Valdrighi"]

statements from natural language texts in a predictive manner, which can sometimes make it difficult to discern the reasoning behind a specific output. In contrast, our pipeline produces entities and relations sequentially, with increasing semantic complexity: for each RDF statement generated by REBEL, there is a corresponding natural language triple extracted from the text by ChatGPT, which serves as input for generating that statement. In addition, the statements generated by the seq2seq model are verified to be as consistent as possible with the synthetic data generated by the LLM.

However, it is important to acknowledge the limitations of our approach. Since the methodology relies on synthetic data generated by ChatGPT, the pipeline is vulnerable to hallucinations from the LLM, potentially causing a cascade of errors in the output. This issue is also related to the inherent risk of using a combination of tools as it increases the likelihood of errors at each step of the pipeline. Therefore, introducing human supervision in one of the extraction steps could be a successful strategy to reduce errors and make the pipeline's application more responsible.

In conclusion, this work presents the first version of an automatically extracted KG from a small corpus of manuscripts related to Leopardi. This work will be further extended by tackling a series of points. First of all, it is crucial to propose a methodology to discover new statements based on the implicit knowledge present in the KG. This methodology should consider the different attributes and constraints of properties in Wikidata and their equivalent in other ontologies such as FRBR to define a series of rules to discover new information implicit in the extracted graph. Another important task related to this work is the creation of a benchmark from the corpus of Leopardi that can be used to evaluate different KE tasks. Inspired by the work of Graciotti et al. (2024), the creation of a domain-specific benchmark for natural language understanding tasks is crucial to assess how LLMs and other techniques can be used to answer competency questions and retrieve information from texts. Finally, it will be of interest to fine-tune an LLM to perform KE from Leopardi's texts given a reference schema or ontology to assess the applicability of instruction-tuned models for end-to-end KE.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/sntcristian/leopardi_kg.

Author contributions

CS: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

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.

Publisher's note

References

Brando, C., Frontini, F., and Ganascia, J.-G. (2015). "Disambiguation of named entities in cultural heritage texts using linked data sets," in *New Trends in Databases and Information Systems, Communications in Computer and Information Science*, eds. T. Morzy, P. Valduriez, and L. Bellatreche (Cham: Springer International Publishing), 505–514. doi: 10.1007/978-3-319-23201-0_51

Cunningham, H. (2002). "GATE: a framework and graphical development environment for robust NLP tools and applications," in *Proc. 40th annual meeting of the association for computational linguistics (ACL 2002)* (Philadelphia, PA: ACL), 168–175. doi: 10.3115/1073083.1073112

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., and Doucet, A. (2021). Named entity recognition and classification on historical documents: a survey. *arXiv*. [Preprint]. arXiv:2109.11406. doi: 10.48550/arXiv.2109.11406

Gangemi, A., Graciotti, A., Meloni, A., Nuzzolese, A. G., Presutti, V., Recupero, D. R., et al. (2023). "Text2amr2fred, a tool for transforming text into rdf/owl knowledge graphs via abstract meaning representation," in *ISWC (Posters/Demos/Industry)* (Athens).

Graciotti, A. (2023). "Knowledge extraction from multilingual and historical texts for advanced question answering," in *Proceedings of the Doctoral Consortium at ISWC 2023 co-located with 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 7, 2023, volume 3678 of CEUR Workshop Proceedings*, eds. C. d'Amato, and J. Z. Pan. Available at: https://CEUR-WS.org (accessed August 24, 2024).

Graciotti, A., Presutti, V., and Tripodi, R. (2024). "Latent vs explicit knowledge representation: how ChatGPT answers questions about low-frequency entities," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, eds. N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue (Torino: ELRA and ICCL), 10172–10185.

Graham, S. S., Majdik, Z. P., and Clark, D. (2020). Methods for extracting relational data from unstructured texts prior to network visualization in humanities research. *J. Open Humanit. Data* 6:8. doi: 10.5334/johd.21

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., et al. (2021). Knowledge graphs. *ACM Comput. Surv*. 54, 71:1–71:37. doi: 10.1145/3447772

Huguet Cabot, P.-L., and Navigli, R. (2021). "REBEL: relation extraction by end-to-end language generation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, eds. M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih (Punta Cana: Association for Computational Linguistics), 2370–2381. doi: 10.18653/v1/2021.findings-emnlp.204

Huguet Cabot, P.-L., Tedeschi, S., Ngonga Ngomo, A.-C., and Navigli, R. (2023). "RED^{fm}: a filtered and multilingual relation extraction dataset," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. A. Rogers, J. Boyd-Graber, and N. Okazaki (Toronto, ON: Association for Computational Linguistics), 4326–4343. doi: 10.18653/v1/2023.acl-long.237

Jain, N., Múnera, A. S., Lomaeva, M., Streit, J., Thormeyer, S., Schmidt, P., et al. (2022). "Generating domain-specific knowledge graphs: challenges with open information extraction," in *TEXT2KG/MK@ ESWC* (Hersonissos), 52–69.

Li, Y., Ramprasad, R., and Zhang, C. (2024). A simple but effective approach to improve structured language model output for information extraction. *arXiv*. [Preprint]. arXiv:2402.13364. doi: 10.48550/arXiv:2402.13364

Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Boros, E., Hamdi, A., Doucet, A., et al. (2022). MELHISSA: a multilingual entity linking architecture for historical press articles. *Int. J. Digit. Libr*. 23, 133–160. doi: 10.1007/s00799-021-00319-6

Ma, Y., Cao, Y., Hong, Y., and Sun, A. (2023). "Large language model is not a good few-shot information extractor, but a good Reranker for Hard Samples!" in *Findings of the Association for Computational Linguistics: EMNLP 2023* (Sentosa), 10572–10601. doi: 10.18653/v1/2023.findings-emnlp.710

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., McClosky, D., et al. (2014). "The stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (Baltimore, MD: ACL), 55–60. doi: 10.3115/v1/P14-5010

Melosi, L., and Marozzi, G. (2021). Il progetto biblioteca digitale leopardiana: per una catalogazione e digitalizzazione dei manoscritti autografi di Giacomo Leopardi. *DigItalia* 16, 65–81. doi: 10.36181/digitalia-00026

OpenAI (2023). *ChatGPT: Optimizing Language Models for Dialogue*. archive.ph. Available at: https://archive.ph/4snnY (accessed August 24, 2024).

Reimers, N., and Gurevych, I. (2019). Sentence-BERT: sentence embeddings using siamese BERT-networks. *arXiv*. [Preprint]. arXiv:1908.10084. doi: 10.48550/arXiv.1908.10084

Reinanda, R., Utama, M., Steijlen, F., and de Rijke, M. (2013). "Entity network extraction based on association finding and relation extraction," in *Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. Proceedings 3* (Cham: Springer), 156–167. doi: 10.1007/978-3-642-40501-3_16

Ruiz, P., and Poibeau, T. (2019). Mapping the Bentham corpus: concept-based navigation. *J. Data Min. Digit. Humanit*. doi: 10.46298/jdmdh.5044

Santini, C., Garay, N., Posthumus, E., and Sack, H. (2024). "The art of relations," in *Book of Abstracts DHd 2024*. Passau: Zenodo.

Santini, C., Tan, M. A., Tietz, T., Bruns, O., Posthumus, E., Sack, H., et al. (2022). "Knowledge extraction for art history: the case of Vasari's the lives of the artists (1568)," in *Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022) Berlin, Germany, Sept. 19th-23rd, 2022, volume 3234 of CEUR Workshop Proceedings*, ed. A. Paschke . Available at: https://ceur-ws.org/ (accessed August 24, 2024).

Sevgili, Ö., Shelmanov, A., Arkhipov, M., Panchenko, A., and Biemann, C. (2022). Neural entity linking: a survey of models based on deep learning. *Semant. Web* 13, 527–570. doi: 10.3233/SW-222986

Shenoy, K., Ilievski, F., Garijo, D., Schwabe, D., and Szekely, P. (2022). A study of the quality of Wikidata. *J. Web Semant*. 72:100679. doi: 10.1016/j.websem.2021.100679

Sporleder, C. (2010). Natural language processing for cultural heritage domains. *Lang. Linguist. Compass* 4, 750–768. doi: 10.1111/j.1749-818X.2010.00230.x

Trajanoska, M., Stojanov, R., and Trajanov, D. (2023). Enhancing knowledge graph construction using large language models. *arXiv*. [Preprint]. arXiv:2305.04676. doi: 10.4855/arXiv.2305.04676

van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., and Van de Walle, R. (2015). Exploring entity recognition and disambiguation for cultural heritage collections. *Digit. Scholarsh. Humanit*. 30, 262–279. doi: 10.1093/llc/fqt067

Vasiliev, Y. (2020). *Natural language processing with Python and spaCy: A practical introduction*. San Francisco, CA: No Starch Press.

Wang, X., Chen, L., Ban, T., Usman, M., Guan, Y., Liu, S., et al. (2021). Knowledge graph quality control: a survey. *Fundam. Res*. 1, 607–626. doi: 10.1016/j.fmre.2021.09.003

Xu, X., Zhu, Y., Wang, X., and Zhang, N. (2023). How to unleash the power of large language models for few-shot relation extraction? *arXiv* [Preprint]. arXiv:2305.01555. doi: 10.48550/arXiv.2305.01555

Zhao, X., Deng, Y., Yang, M., Wang, L., Zhang, R., Cheng, H., et al. (2024). A comprehensive survey on relation extraction: recent advances and new frontiers. *ACM Comput. Surv*. 56:293. doi: 10.1145/3674501