Knowledge Extraction from Multilingual and Historical Texts for Advanced Question Answering

Arianna Graciotti[1]

[1]*Department of Modern Languages, Literatures, and Cultures, University of Bologna, 40126 Bologna, Italy*

Abstract

Building knowledge graphs from text is a challenge, especially when working with multilingual and historical documents. This PhD thesis addresses this task with the goal of supporting question-answering applications against repositories of diachronic textual resources. The proposed method combines Semantic Web and Natural Language Processing techniques and addresses specific difficulties inherent to historical texts, such as data deterioration from Optical Character Recognition and challenges in Entity Linking due to bias towards contemporary entities. The framework is evaluated against a corpus of historical documents covering the Musical Heritage domain.

Keywords

Natural Language Processing and Understanding, Knowledge Extraction, Knowledge Graphs, Diachronic and Multilingual Corpora, Question Answering

1. Introduction

Knowledge Extraction (KE) from text for creating Knowledge Graphs (KGs) is pivotal for enabling question-answering (QA) systems. These systems offer access to this knowledge while concurrently enabling an assessment of its quality. State-of-the-art (SotA) KE technologies primarily rely on contemporary text crawled from the web. This research approaches the challenges associated with historical and multilingual documents, often overlooked by Natural Language Processing/Understanding (NLP/U) systems. A diachronic and plurilingual musical heritage (MH) corpus[1] is used as an empirical basis for experimenting and validating our proposal. Through integrating Semantic Web (SW) technologies, we aim to create logically robust knowledge graphs from the extracted knowledge, accessible via ontology-based querying.

1.1. Problem Statement

Text-to-KG pipelines have been advanced by both the NLP and SW communities. NLP has leveraged Machine Learning (ML) and Neural Networks (NN) advancements for improved semantic parsing. While promising results have been achieved with sequence-to-sequence

[1]The Polifonia Textual Corpus (PTC) https://github.com/polifonia-project/Polifonia-Corpus; https://doi.org/10.5281/zenodo.6772330, a large diachronic corpus covering English, Italian, Spanish, French, German and Dutch language documents.

models in single-language scenarios [1] and multi-language settings [2], those models fall short in interoperability due to the formalisms' heterogeneity and lax logic.

SW has used interoperable ontologies for the formal representation of extracted knowledge, which favours knowledge augmentation and alignment with other ontologies. FRED [3] achieves this goal but relies on an obsolete and monolingual-input NLP pipeline. Both communities' SotA tools integrate Named Entity Recognition and Classification (NERC) and Entity Linking (EL) systems trained on contemporary texts whose performance is not satisfactory when applied to historical documents.

Recently, Large Language Models (LLMs, [4]) have shown capabilities of converting text into KGs. However, LLMs are not explicitly trained to prioritize ontologies' and KGs' modelling best practices, such as Ontology Design Patterns (OPDs). Additionally, the ability of LLMs to handle variations in lexical and syntactic structures in the input text, and thus to avoid generating redundant triples, remains unproven. We hypothesise that these issues are especially relevant when dealing with historical documents, which might be under-represented in the models' training data, largely made of web-crawled content.

1.2. Importance

This research aims to overcome SotA KE technologies' limitations with plurilingual and non-contemporary textual sources. It proposes using semantic parsing based on Abstract Meaning Representation (AMR, [5]) to transform multilingual and historical texts into RDF/OWL event-centric Knowledge Graphs (KGs), leveraging AMR's resilience to surface variations in the input text. Pre-processing challenges, like mitigating information loss due to Optical Character Recognition (OCR) and data preparation tasks (such as sentence-splitting, required for semantic parsing), are also targeted. The KGs created, enriched with existing Knowledge Bases (KBs), are ready for SPARQL querying, maintaining links to the original sources.

The RDF/OWL KGs generated through our hybrid pipeline also allow for populating validated datasets and enabling comparative analysis with KGs generated by LLMs. Text-to-KG output obtained through the two approaches can finally be compared with regard to stability amidst lexical and syntactic variations, alignment proficiency with Ontology Design Patterns (ODP), and efficiency in handling structured queries.

While the experimental focus is on the MH domain, our results and methods are generalisable and applicable to any arbitrary collection of texts. This research will enable scholars, e.g. musicologists, to access historical and multilingual sources more widely and easily.

2. Expected Contribution and Methods

My research project aims at addressing the following research questions:
RQ1: What methodologies can be employed to automatically derive RDF/OWL KGs from multilingual and diachronic textual corpora?
RQ1.1: What information should be retained when transforming text to KG, specifically addressing historical documents' characteristics?
RQ1.2: What is the relevant implicit knowledge behind a text that shall be explicit in a KG?
RQ1.3: How can we ensure scalable and consistent quality of the KGs?

RQ1.4: What aspects of the methodology should be tailored to specifically address historical documents' characteristics?

This research proposes a hybrid KE framework combining symbolic and neural architectural components to generate KGs automatically from multilingual and historical documents (RQ1). The proposed methodology encompasses techniques to pre-process the input text, such as co-reference resolution and post-OCR correction (RQ1.1). Post-processing strategies are applied to enrich the output KGs with selected knowledge derived from external Knowledge Bases (KBs) (RQ1.2). A hybrid evaluation approach is proposed, which comprises automatic evaluation metrics and tools for human-driven validation of the graphs (RQ1.3). Focus is placed on testing SotA NERC and EL neural models to identify their limits and improve them to deliver optimal performance when dealing with non-contemporary texts (RQ1.4).

RQ2: Can the automatically generated KGs provide useful answers to both the general public's and experts' questions, and how?

RQ2.1: How can ontology-based structured querying be leveraged to complement KGs quality evaluation?

RQ2.2: Can the automatically generated KGs be used to create silver training data for QA, and how?

This research aims to optimize the text-to-KG output for QA applications. QA allows us to foster access to diachronic resource repositories and enable extrinsic evaluation of the KGs (RQ2). We contemplate methods and tools for translating expert Competency Questions (CQs) into SPARQL queries, and we outline procedures for experts to evaluate the results of these queries (RQ2.1). Alongside this, we design a data augmentation methodology for the synthetic generation of training data in the form of question-answer-provenance tuples for fine-tuning domain-specific QA models (RQ2.2).

3. Related Works

3.1. Text-to-KG

Recent advancements in semantic parsing, such as AMR parsing, offer promising solutions for text-to-KG challenges. AMR outputs an event-centric graph-based sentence meaning representation based on PropBank *Frames*[2] [6]. Neural AMR parsers, such as SPRING [1], utilize transformers' transfer learning capabilities. However, these models have been predominantly tested on contemporary English texts. Furthermore, AMR output is informal as opposed to logic-based KGs. As discussed in [7], the SW community has focused on KE frameworks that encode results in formal SW standards, with FRED [3] being a SotA machine reader able to encode information into KGs for structured queries and alignment with other KBs. However, FRED relies on a cumbersome NLP pipeline that supports only input in English. My research strives to overcome both approaches' limitations by exploiting SotA neural AMR parsers' multilingual capabilities and transforming their output into logically sound frame-based KGs that

---

[2]PropBank Frames are the core lexicon of the PropBank paradigm and consist of predicate-argument structures named "rolesets". A complete list of PropBank frames can be found at http://propbank.github.io/v3.4.0/frames/
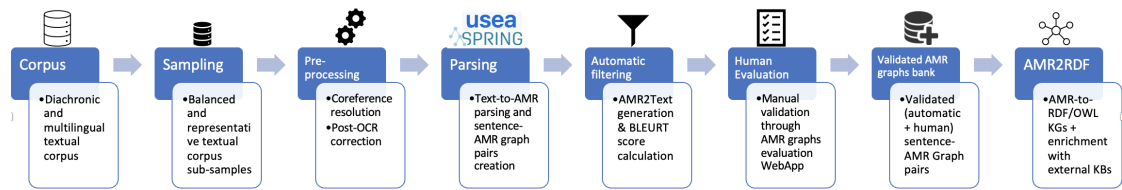
Figure 1: Overview of the hybrid text-to-KG framework proposed in this research.

follow FRED's theoretical paradigm.

3.2. Historical NERC and EL

Historical documents present unique challenges for NERC and EL, such as noisy OCR output, variations in the spelling, sentence structure, and naming conventions [8]. The scarcity of annotated historical text datasets hinders progress in deep learning approaches to this task [9]. SotA neural EL systems fall short in entity disambiguation performance when applied to historical texts, rarely included in popular training datasets [10]. [11] pioneered the research in this field by proposing diaNED, a time-aware entity disambiguation approach for diachronic corpora, which leverages KB-driven temporal information and text-driven temporal expressions. Recent evaluation campaigns [12] have provided datasets of historical and multilingual documents with gold-standard annotations for NERC and EL and novel transformer-based models to approach the tasks. My research mitigates the lack of gold-standard resources containing NERC- and EL-annotated historical documents by contributing with a manually annotated benchmark. It also aims at advancing SotA models' performance proposing a neuro-symbolic EL model fine-tuned on historical data enhanced by knowledge-based heuristics.

3.3. Question Generation and Answering Using KGs

Research has widely investigated QA over unstructured text and over structured knowledge (Knowledge Base Question Answering, KBQA) [13]. LLMs demonstrated remarkable performance in QA over text. In the KBQA realm, recent advancements in text-to-SPARQL technologies facilitate structured querying [14], and data augmentation approaches alleviate the need for domain-specific training data [15]. The text-to-KG pipeline developed within this research enables the development of a domain- and language-agnostic methodology for generating training data for QA and fine-tuning cross-domain and cross-language QA systems. It also allows for comparing QA over text and KBQA performance.

4. Method and Results Achieved So Far

4.1. Text-to-KG

(RQ1.1-1.3) This research contributes to the development of Polifonia Knowledge Extractor (PKE)[3], a pipeline that relies on AMR parsing as an intermediate step for extracting knowl-

---

[3]https://github.com/polifonia-project/Polifonia-Knowledge-Extractor

Table 1

Statistics for the *Musical heritage Historical named Entities Recognition, Classification and Linking - Version 0.1* (MHERCL v0.1) benchmark.

Dataset	Lang	#docs	#sents	#tokens	#named entities mentions				
					all	types	noisy	linked	not linked
MHERCL v0.1	EN	20	2.181	51.301	2.757	65	502 (18%)	1849 (67%)	908 (33%)

edge from diachronic corpora. The text-to-AMR parsing relies on the neural semantic parsers SPRING [1] for English and USeA [16] for languages other than English. The pipeline implements methodologies for minimising the loss of information in the source text by performing co-reference resolution and OCR post-correction[4]. Through PKE, this research contributed to releasing an AMR graphs bank focused on the music domain[5], containing text from contemporary and historical documentary sources. Also, it contributed to developing an enhanced version of AMR2FRED [17], a tool that exploits similarities between AMR and FRED's graphs, such as both being frame-based and event-centric, to transform AMR graphs into RDF/OWL KGs that follow FRED's theoretical model. This transformation allows the enrichment of the resulting KGs through Framester[6][7]. PKE[3] and AMR2FRED tools lay the basis for our text-to-KG framework[8]. Figure 1 depicts the various steps of our framework. Within this research, this framework was applied to a module of PTC[1], the MusicBO pilot's corpus[9], focusing on a selection of documents in English and Italian. The resulting MusicBO Knowledge Graph is available on GitHub[10] and through a public SPARQL endpoint[11]. Data stories extracted from it are published on MELODY[12].

4.2. Historical NERC and EL

(RQ1.4) Striving to advance NERC and EL on historical documents, this research sponsored the building of the *Musical heritage Historical named Entities Recognition, Classification and Linking* (MHERCL) benchmark of manually annotated sentences with NERC and EL information,

---

[4]Our initial approach to OCR post-correction is a rule-based method detailed at https://github.com/polifonia-project/rulebased-postocr-corrector. It targets sentence cohesion issues, particularly those caused by historical periodicals' peculiar formats, like multiple columns leading to incorrect sentence breaks.

[5]Available at https://zenodo.org/record/7025779#.ZDls8OxBy3I

[6]Framester (https://github.com/framester/Framester) is an extensive KG formalising and interlinking diverse lexical and ontological resources.

[7]Framester facilitates KGs enrichment, for example, by enabling the addition of semantics to the generic argument slots of PropBank predicates as found in AMR graphs. Specifically, the "ARG0-PAG" argument of the "play.11" predicate can, through Framester, be formally associated with the URI https://w3id.org/framester/pb/data/role/play-11/performer.

[8]The framework can be invoked to obtain *named graphs* via the Machine Reading suite (available at: https://github.com/anuzzolese/machine-reading) exploiting the Text-to-AMR-to-FRED API (available at: http://framester.istc.cnr.it/txt-amr-fred/api/docs)

[9]MusicBO pilot's corpus collects documents regarding the role of music in the cultural heritage history of Bologna.

[10]https://github.com/polifonia-project/musicbo-knowledge-graph

[11]https://polifonia.disi.unibo.it/musicbo/sparql

[12]https://projects.dharc.unibo.it/melody/musicbo/music_in_bologna_knowledge_graph_overview

extrapolated from historical periodicals in English. Such a resource is intended to support the testing of neural models for NERC and EL and the implementation of neuro-symbolic architectures specialised in historical documents. MHERCL v0.1 is a manually annotated benchmark from the English *Periodicals* module of PTC[1], including historical documents from 1823 to 1900. The sentences were selected based on four criteria: English language, sourced from the *Periodicals* module of the PTC, being part of the AMR Graphs Bank[5], and containing at least one named entity (NE) recognised by the PKE[3]. Two undergraduate Foreign Languages and Literatures interns manually annotated the sentences, recognising NEs and linking them to their corresponding WikiData ID (QID) based on guidelines harmonised from the AMR and the *Impresso*[13] project, which specifically targets historical documents. Statistics of MHERCL v0.1 are summarised in Table 1. It is worth noticing that 18% NE is noisy due to OCR errors. Also, 33% of the NEs' mentions could not be linked to any QID by the annotators. MHERCL v0.1 will be released in tab-separated values (TSV) format (UTF-8 encoded) compliant with the HIPE-2022 data format, to ease future integration.

5. Evaluation

5.1. Text-to-KG

(RQ1.1-1.3) Non-standard texts, like those in historical documents, can affect the accuracy of SotA neural AMR parsers. To assess the quality of AMR graphs derived from the PTC[1] sentences, we blend an automatic filtering strategy using an AMR-to-text back-translation approach and human AMR graphs evaluation. This approach yields BLEURT[14] scores, which reflect the similarity between the original and the back-translated sentences. High BLEURT scores, we hypothesize, correspond to high-quality AMR graphs. In this initial stage, we discard all graphs whose corresponding back-translated sentence yield a negative BLEURT score. To validate the automatic filtering, we have developed an AMR evaluation web tool[15] aimed at supporting human validation of the graphs and strengthening the evaluation methodology. We plan to utilize the RDF/OWL KGs generated by our hybrid pipeline to populate validated datasets and conduct a comparative analysis with KGs generated by LLMs, building upon the approach described in [18]. We will pay attention to a careful definition of the comparison protocol to mitigate the risk of framing a biased comparison task.

5.2. Historical NERC and EL

(RQ1.4) The gold standard benchmark developed (see Paragraph 4.2) allows for evaluating the performance of SotA EL tools on historical documents. Preliminary tests were conducted on a subset of MHERCL v0.1, comprising 1545 unique tuples (sentence, NE mention, QID). NEs mentions that could not be linked to any QID by the annotators were excluded from this test. When we passed these tuples to BLINK, the output QIDs and gold QIDs matched in 1135 instances (73%), but differed in 410 cases (27%), showing room for improvement.

---

[13]https://impresso-project.ch/
[14]https://github.com/google-research/bleurt
[15]http://framester.istc.cnr.it/amr-eval

5.3. Question Generation and Answering Using KGs

(RQ2.1-2.2) We will leverage KBQA as an interface to maximize accessibility to our KE outcomes and to evaluate the quality of our text-generated KGs. This will involve translating CQs crafted by experts into SPARQL queries and allowing experts to assess the results. Data augmentation strategies will be evaluated by leveraging the rich landscape of benchmarks specialised for QA [19]. Being our KG automatically extracted from text, we plan to compare KBQA performance to that of QA over text powered by LLMs.

6. Next Steps and Conclusion

As next steps, we will address the multilingualism challenge by expanding the language scope of the documents within the PTC[1] to be processed through our text-to-KG pipeline.

The project will also continue to build historical NERC and EL benchmarks targeting the Italian language, for which, to the best of my knowledge, there is profound resource scarcity. In collaboration with musicologists (as our current case study is in MH domain), we will validate the NEs that the annotators could not link to WikiData and consider adding them as new entries to that KB. A novel neuro-symbolic framework for NERC and EL will be developed by fine-tuning SotA neural EL using historical documents and employing knowledge-based heuristics.

Further future work will concentrate on strengthening the evaluation methods: we plan to exploit QA applications to assess the generated RDF/OWL KGs capacity to answer domain-specific queries. We will continue the work on data augmentation for QA, with the aim of making our KGs interrogable also through a custom fine-tuned QA model that supports natural language questions.

Finally, it will be key to establish solid comparison pipelines between AMR-driven frame-based RDF/OWL KGs and those derived from LLMs. This endeavour is crucial to enhance our understanding of the capabilities and limitations of LLMs and to identify to which aspects we could bring the most value with a hybrid KE framework at the crossroads of NLP/NLU and SW.

Acknowledgments

References

[1] M. Bevilacqua, et al., One SPRING to Rule Them Both: Symmetric AMR Semantic Parsing and Generation without a Complex Pipeline, Proc. of the AAAI Conference on Artificial Intelligence 35 (2021) 12564–12573. doi:10.1609/aaai.v35i14.17489.

[2] L. Procopio, et al., SGL: Speaking the Graph Languages of Semantic Parsing via Multilingual Translation, in: K. Toutanova, et al. (Eds.), Proc. of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL, Online, 2021, pp. 325–337. doi:10.18653/v1/2021.naacl-main.30.

[3] A. Gangemi, et al., Semantic Web Machine Reading with FRED, Semantic Web 8 (2017) 873–893. doi:10.3233/SW-160240.

[4] W. X. Zhao, et al., A Survey of Large Language Models, 2023. doi:10.48550/arXiv.2303.18223.

[5] L. Banarescu, et al., Abstract Meaning Representation for sembanking, in: Proc. of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, ACL, Sofia, Bulgaria, 2013, pp. 178–186. URL: https://aclanthology.org/W13-2322.

[6] S. Pradhan, et al., PropBank Comes of Age—Larger, Smarter, and more Diverse, in: Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics, Seattle, Washington, 2022, pp. 278–288. doi:10.18653/v1/2022.starsem-1.24.

[7] F. Babič, et al., Review of Tools for Semantics Extraction: Application in Tsunami Research Domain, Information 13 (2022). doi:10.3390/info13010004.

[8] M. Ehrmann, et al., Named Entity Recognition and Classification on Historical Documents: A Survey, 2021. doi:10.48550/arXiv.2109.11406.

[9] Ö. Sevgili, , et al., Neural entity linking: A survey of models based on deep learning, Semantic Web 13 (2022) 527–570. doi:10.3233/SW-222986.

[10] C. Möller, et al., Survey on English Entity Linking on Wikidata: Datasets and approaches, Semantic Web 13 (2022) 925–966. doi:10.3233/SW-212865.

[11] P. Agarwal, et al., diaNED: Time-aware named entity disambiguation for diachronic corpora, in: I. Gurevych, et al. (Eds.), Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 686–693. doi:10.18653/v1/P18-2109.

[12] M. Ehrmann, et al., Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents, in: Proc. of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, CEUR-WS, Bologna, Italy, 2022, pp. 1038 – 1063. doi:10.5281/zenodo.6979577.

[13] R. S. Roy, et al., Introduction, Springer International Publishing, Cham, 2022, pp. 1–5. doi:10.1007/978-3-031-79512-1_1.

[14] D. Banerjee, et al., Modern Baselines for SPARQL Semantic Parsing, in: Proc. of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2260–2265. doi:10.1145/3477495.3531841.

[15] O. Agarwal, et al., Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training, in: K. Toutanova, et al. (Eds.), Proc. of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 3554–3565. doi:10.18653/v1/2021.naacl-main.278.

[16] R. Orlando, et al., Universal Semantic Annotator: the First Unified API for WSD, SRL and Semantic Parsing, in: N. Calzolari, et al. (Eds.), Proc. of LREC 2022, European Language Resources Association, Marseille, France, 2022, pp. 2634–2641. URL: https://aclanthology.org/2022.lrec-1.282.

[17] A. Meloni, et al., AMR2FRED, A Tool for Translating Abstract Meaning Representation to Motif-Based Linguistic Knowledge Graphs, in: E. Blomqvist, et al. (Eds.), The Semantic Web: ESWC 2017 Satellite Events, Springer International Publishing, Cham, 2017, pp. 43–47. doi:10.1007/978-3-319-70407-4_9.

[18] M. Trajanoska, et al., Enhancing Knowledge Graph Construction Using Large Language Models, 2023. doi:10.48550/arXiv.2305.04676.

[19] A. Rogers, et al., QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension, ACM Comput. Surv. 55 (2023). doi:10.1145/3560260.