MusicBO, an application of Text2AMR2FRED to the Musical Heritage domain

Aldo Gangemi[1], Arianna Graciotti[1], Antonello Meloni[2], Eleonora Marzi[1], Andrea Nuzzolese[3], Valentina Presutti[1], Diego Reforgiato Recupero[2,3], Alessandro Russo[3], and Rocco Tripodi[4]

[1] University of Bologna, 40126 Bologna, Italy
[2] University of Cagliari. Via Ospedale 72, 09124 Cagliari, Italy
[3] CNR, via San Martino della Battaglia 44, 00185, Rome, Italy
[4] Ca' Foscari University of Venice, Sestiere Dorsoduro, 3246, 30123 Venezia VE

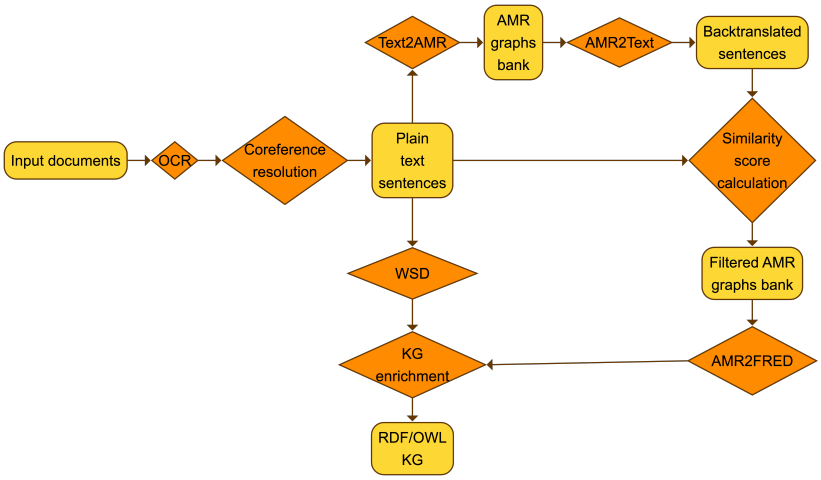
Abstract. Converting textual data into Knowledge Graphs (KGs) poses a significant challenge, particularly when dealing with multilingual and historical documents. In this paper, we describe the application of Text2-AMR2FRED to MusicBO corpus, the former being a tool for transforming text into RDF/OWL KGs via Abstract Meaning Representation (AMR), the latter being a diachronic collection of Musical Heritage (MH) texts.

Keywords: Abstract Meaning Representation · Natural Language Processing · Knowledge Graphs · Semantic Frames

1 Introduction

This paper describes the methods and tools applied for automatically transforming MusicBO, a multilingual and diachronic textual corpus about the role of Musical Heritage (MH) in the city of Bologna, into an OWL-compliant RDF Knowledge Graph (KG). The KG obtained is publicly accessible through a SPARQL endpoint[5], enabling the creation of visual data stories[6] using MELODY[7]. The resulting KG aims to enable scholars with different backgrounds to conduct large-scale qualitative analysis.

Fig. 1. The MusicBO KG creation pipeline schema.



2 The MusicBO Knowledge Graph

MusicBO corpus[8] contains 137 texts in 4 languages (Italian, English, French, and Spanish) published between 1700 to the current era[10].

Table 1. Statistics describing the MusicBO corpus' documents included in the scope of this study.

Language	**#docs**	total length (#tokens)	average length (#tokens)	median length (#tokens)
EN	47	1.873.030	40.718	9.964.5
ITA	51	2.329.054	44.789, 5	14.334

Our approach to transform plain text into a KG, depicted in Figure 1, is based on Text2AMR2FRED[11][3], an enhanced version of FRED [5,4]. We include in the

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

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

[7] MELODY (Make mE a Linked Open Data storY) is a web portal that allows users to query Linked Open Data and create web-ready interactive data stories.

[8] MusicBO corpus is part of the wider Polifonia Textual Corpus[9], a large-scale, multilingual and multigenre diachronic textual corpus.

[10] Due to copyright reasons, the documents of MusicBO corpus cannot be entirely disclosed. Still, we released metadata that allows the reproduction of the corpus `https://doi.org/10.5281/zenodo.6672165`.

[11] `https://arco.istc.cnr.it/txt-amr-fred/`

Table 2. Statistics describing the KG resulting from the application of Text2AMR2FRED to MusicBO corpus.

Language	#(sent, AMR graph) pairs (Text2AMR)	#(filtered sent, AMR graph) pairs (Automatic metrics evaluation)	#triples
EN	51.814	5.798	412.911
ITA	10.563	1.759	118.162
Overall	62.377	7.557	531.073

scope of this study only the MusicBO corpus' documents in English and in Italian (respectively 47 and 51). Statistics[12] of the documents processed in this study are reported in Table 1. The initial formats of these documents encompassed *.pdf*, *images*, or *.docx*. We extract plain text from them through customized Optical Character Recognition (OCR) technologies[13]. Subsequently, we carry out co-reference resolution[14], rule-based minimal post-OCR corrections[15], and sentence splitting on the extracted plain texts. Following this pre-processing stage, we submit the processed sentences to neural models (SPRING for English [1] and USeA for Italian [7]) for text2AMR parsing. AMR graphs, anchored to PropBank *Frames*[16], function as an event-centric representation of the MusicBO corpus' sentences, suited for extracting 'who-did-what-to-whom' information from a text. Through the application of the AMR2FRED tool[17] [6], accessible via the Machine Reading suite[18], we transform AMR graphs into full-fledged RDF/OWL KGs aligned with FRED's theoretical framework. The outcome is a series of *named graphs*, enabling the tracking of each triple to its originating sentence in the corpus. We enrich the resulting KGs through Framester [2], which allows the alignment with external Knowledge Bases (KBs) such as DBPedia[19], Wikidata[2] and Verbatlas[2]. For instance, consider the following triples[20]:

[12] Statistics have been calculated using SpaCy (`https://spacy.io`) NLP library, employing the models `en_core_web_trf` for documents in English language and `it_core_news_lg` for documents in Italian language.

[13] `https://github.com/polifonia-project/textual-corpus-population`

[14] For English language documents, we implemented a co-reference resolution pipeline based on Spacy's *neuralcoref* (`https://spacy.io/universe/project/neuralcoref`). We are currently evaluating tools for Italian.

[15] `https://github.com/polifonia-project/rulebased-postocr-corrector`

[16] PropBank Frames are the core lexicon of the PropBank paradigm and consist of predicate-argument structures named "rolesets".

[17] `https://github.com/polifonia-project/amr2Fred`

[18] `https://github.com/polifonia-project/machine-reading`

[19] `https://www.dbpedia.org/`, `https://www.wikidata.org/`, `https://verbatlas.org/`

[20] Extrapolated from the KG originating from the sentence *"In the year 1814, Barbaja went to Bologna and offered Rossini a better engagement than be-*

```
fred:Barbaja a amr:Person ;
    owl:sameAs dbpedia:Domenico_Barbaia,
        wd:Q908235 .

fred:offer_1 a pbdata:offer-01 ;
    pblr:benefactive_or_entity_offered_to fred:Rossini ;
    pblr:commodity fred:engagement_1 ;
    pblr:entity_offering fred:Barbaja ;
    fschema:subsumedUnder va:0229f,
        fnframe:Offering .

fred:Rossini a amr:Person ;
    owl:sameAs dbpedia:Gioachino_Rossini,
        wd:Q9726 .
```

The reported triples encode the event of an engagement offer delivered from Domenico Barbaja, an opera manager, to the composer Gioachino Rossini[21]. Such knowledge is what scholars who supported the corpus collection aimed to disclose, at scale, from MusicBO documents automatically. Independent scholars can leverage such knowledge encoded in the KG and create data stories through MELODY, such as the one created by University of Bologna students [22], focusing on the representation of Russian composers and classical music in the MusicBo corpus.

Processing non-standard texts may lead to potential inaccuracies of text2AMR parsers, as such data is scarce in their training sets. Manual validation is time-consuming, and no standard benchmarks exist for semantic parsing of historic and OCRed text. We followed a back-translation [8] methodology to address these challenges. We converted the AMR graphs back to sentences using SPRING for English and m-AMR2Text for Italian[23] , followed by similarity score computations using BLEURT [24] for English and cosine similarity for Italian. We posit that high-quality graphs are associated with generated sentences exhibiting high BLEURT or cosine similarity scores. All AMR graphs paired with AMR2Text-generated sentences with a negative BLEURT score or a cosine similarity below 0.90 were discarded. We provide in Table 2 the statistics regarding the KG resulting from the application of Text2AMR2FRED to the MusicBO corpus, in-

fore.", taken from the MusicBO corpus document *The Life of Rossini (Edwards, 1869)*, available at: `https://freeditorial.com/en/books/filter-author/henry-sutherland-edwards`

[21] The named entities are automatically linked to their entry in Wikipedia by BLINK [9], the entity linker used by SPRING, and aligned to Wikidata and DBPedia in the AMR2RDF step of our pipeline

[22] `https://melody-data.github.io/stories/published_stories/story_1687714706.423208.html`

[23] `https://github.com/UKPLab/m-AMR2Text`

[24] `https://github.com/google-research/bleurt`

cluding insights regarding the automatic filtering. Raw data to recreate the KG are stored in a dedicated repository[25].

3 Acknowledgements

The authors acknowledge the support of the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101004746.

References

1. Bevilacqua, M., Blloshmi, R., Navigli, R.: One SPRING to Rule Them Both: Symmetric AMR Semantic Parsing and Generation without a Complex Pipeline. Proc. of the AAAI Conf. on Artificial Intelligence **35**(14), 12564–12573 (May 2021), https://ojs.aaai.org/index.php/AAAI/article/view/17489
2. Gangemi, A., Alam, M., Asprino, L., Presutti, V., Recupero, D.R.: Framester: A Wide Coverage Linguistic Linked Data Hub. In: EKAW 2016. pp. 239–254. Springer International Publishing, Bologna, Italy (2016)
3. Gangemi, A., Graciotti, A., Meloni, A., Nuzzolese, A., Presutti, V., Reforgiato Recupero, D., Russo, A., Tripodi, R.: Text2AMR2FRED, a tool for transforming text into RDF/OWL Knowledge Graphs via Abstract Meaning Representation. In: 22nd ISWC. CEUR Workshop Proc., Athens, Greece (November 2023)
4. Gangemi, A., Hassan, E., Presutti, V., Recupero, D.R.: FRED as an event extraction tool. In: van Erp, M., Hollink, L., Troncy, R., van Hage, W.R., van de Laar, P., Shamma, D.A., Gao, L. (eds.) Proceedings of DeRiVE 2013, co-located with the 12th ISWC 2013, Sydney, Australia, October 21, 2013. CEUR Workshop Proceedings, vol. 1123, pp. 14–17. CEUR-WS.org (2013)
5. Gangemi, A., Presutti, V., Recupero, D.R., Nuzzolese, A.G., Draicchio, F., Mongiovì, M.: Semantic Web Machine Reading with FRED. Semantic Web **8**(6), 873–893 (2017). https://doi.org/10.3233/SW-160240, https://doi.org/10.3233/SW-160240
6. Meloni, A., Reforgiato Recupero, D., Gangemi, A.: AMR2FRED, A Tool for Translating Abstract Meaning Representation to Motif-Based Linguistic Knowledge Graphs. In: The Semantic Web: ESWC 2017 Satellite Events. pp. 43–47. Springer International Publishing, Portorož, Slovenia (2017)
7. Orlando, R., Conia, S., Faralli, S., Navigli, R.: Universal Semantic Annotator: the First Unified API for WSD, SRL and Semantic Parsing. In: Proc. of LREC 2022. pp. 2634–2641. European Language Resources Association, Marseille, France (June 2022), https://aclanthology.org/2022.lrec-1.282
8. Sennrich, R., Haddow, B., Birch, A.: Improving Neural Machine Translation Models with Monolingual Data. In: Proc. of the 54th Annual Meeting of the ACL (Volume 1: Long Papers). pp. 86–96. ACL, Berlin, Germany (August 2016). https://doi.org/10.18653/v1/P16-1009, https://aclanthology.org/P16-1009
9. Wu, L., Petroni, F., Josifoski, M., Riedel, S., Zettlemoyer, L.: Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). pp. 6397–6407. ACL, Online (November 2020). https://doi.org/10.18653/v1/2020.emnlp-main.519, https://aclanthology.org/2020.emnlp-main.519

[25] https://github.com/polifonia-project/musicbo-knowledge-graph/tree/main