Named Entity Recognition and Knowledge Extraction from Spanish Golden Age theatre

BVMC.Labs

Digital Humanities Center

CENTRO BIBLIOTECA VIRTUAL MIGUEL CERVANTES

Gustavo Candela [1] Juan Carlos García [1] Pilar Escobar [1] Dolores Sáez [1] David Quintela [1] Manuel Marco-Such [1]

[1]University of Alicante, Spain

Introduction

Libraries and cultural heritage institutions, commonly known as Galleries, Libraries, Archives and Museums (GLAM), are exploring new methods of accessing to the digital collections in innovative ways that have recently emerged in order to facilitate the application of computational methods based on Data Science, Machine Learning (ML) and Artificial Intelligence (AI) [4, 1].

Background

The study of the Spanish Golden Age theatre has traditionally gained attention [5]. However, the scarcity of annotated benchmark datasets for particular approaches and languages, such as Spanish Golden Age theatre, has resulted into relatively less progress in these domains. In addition, many of the approaches are offline, computationally expensive and mainly based on predominant languages. Previous approaches in the cultural heritage domain have addressed the identification of artwork titles using NER techniques [3].

Contributions

1. A method to train a NER model for Spanish Golden Age theatre based on TEI documents
2. The NER model generated
3. The results obtained after the analysis and assessment

Framework

Experiment Introduce a method to train a NER model based on Spanish Golden Age theatre. Figure 1 shows the **framework** proposed:

- **Data preparation and tagging:** events, places and mythological characters
- **Training a NER model:** using spaCy V3.0
- **Evaluation:** assessing the NER results

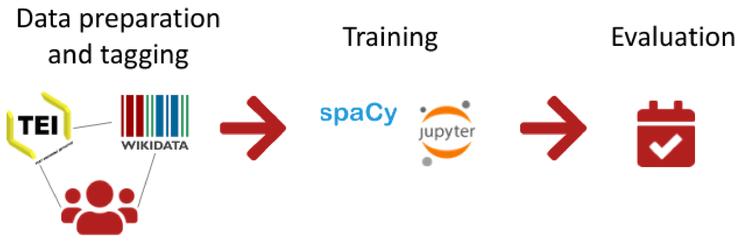

Figure 1. A framework to annotate TEI documents.

TEI tagging

- Dicen que el <wikidata ref="https://www.wikidata.org/wiki/Q530936" type="event">Siglo Dorado</wikidata>
- de generoso en <wikidata ref="https://www.wikidata.org/wiki/Q217196" type="place">Castilla</wikidata>.
- adonde <wikidata ref="https://www.wikidata.org/wiki/Q3954" type="mit">Neptuno</wikidata> reina.</l>

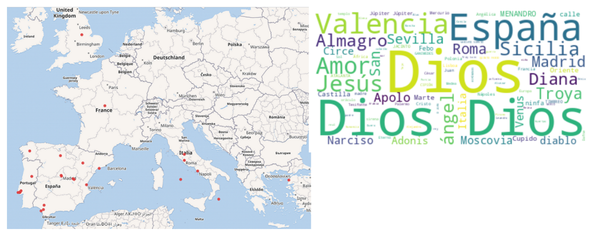

Figure 2. Data visualisations based on the annotations and Wikidata.

Results & Conclusions

The results of this study are publicly available at the BVMC Labs and can be applied to other domains such as historical documents and newspapers [2].

Table 1. TEI works annotated using Wikidata as a repository

Author	Title	Wikidata links
Calderón de la Barca	La vida es sueño: comedia famosa	47
Francisco de Rojas Zorrilla	Abrir el ojo: gran comedia	84
Guillén de Castro	El Conde Alarcos : comedia	21
Lope de Vega	El anzuelo de Fenisa : comedia famosa	197
Lope de Vega	Adonis y Venus : tragedia	214
Lope de Vega	La viuda valenciana : comedia famosa	195
María de Zayas y Sotomayor	La traición en la amistad : comedia famosa	153
Miguel de Cervantes	Tragedia de Numancia	76
Tirso de Molina	El burlador de Sevilla	106

- **Reproducible code** based on a Jupyter Notebook collection
- **Research into production:** the annotation workflow will be enhanced
- **Students engagement** at the University of Alicante
- **Future work** includes the annotation of additional works to increase the corpus and the engagement with additional students to enrich the works

References

[1] Sally Chambers et al. Position Statements -> Collections as Data: State of the field and future directions, May 2023.

[2] Biblioteca Virtual Miguel de Cervantes. DARIAH Annual Event 2023 y la BVMC, May 2023.

[3] Nitisha Jain, Alejandro Sierra Múnera, Jan Ehmueller, and Ralf Krestel. Generation of training data for named entity recognition of artworks. *Semantic Web*, 14(2):239–260, 2023.

[4] Mahendra Mahey et al. *Open a GLAM lab*. International GLAM Labs Community, Book Sprint, Doha, Qatar, 09 2019.

[5] Borja Navarro. A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects. In Anna Feldman, Anna Kazantseva, Stan Szpakowicz, and Corina Koolen, editors, *Proceedings of the Fourth Workshop on Computational Linguistics for Literature, CLfL@NAACL-HLT 2015, June 4, 2015, Denver, Colorado, USA*, pages 105–113. The Association for Computer Linguistics, 2015.

gcandela@ua.es