Towards Semantic Alignment of Heterogeneous Structures and Its Application to Digital Humanities

Renata Vieira[1]* and Cassia Trojahn[2]

[1] CIDEHUS, University of Évora, Portugal
renatav@uevora.pt
[2] IRIT, UMR 5505, 1118 Route de Narbonne, F-31062 Toulouse, France
firstname.lastname@irit.fr

1 Introduction

The field of Digital Humanities comprises the use of technology within arts, heritage and humanities research. This brings new methods of inquiry, new means of dissemination, but also constitute a new core of investigation in itself. Not only creation and access to collections of interest for these areas have improved with digitalization of research material, but further use of computing technology is being proposed and discovered [7]. The primary source of information for humanities researchers comes from free, unstructured sources in written language, that is ambiguous and context-dependent. Also the humanities might face difficulties due to the particularities of the source of information, that might be available in ancient forms of registration. For instance, there is a need for identifying specific vocabulary of a historical period and also align non uniform spelling which was usual in old publications [6]. In this perspective, the ability to establish a relationship between different forms of expression of knowledge (from structured and unstructured sources) and its meaning or intent is crucial [5]. This scenario reflects a unifying framework of a wide range of solutions from a variety of domains, including NLP and semantic web.

Different variants of the notion of 'alignment' have been adopted in a range of areas, focusing on homogeneous structures (e.g., text alignment [8], database alignment [1] or ontology alignment [4]) or heterogeneous structures (e.g., annotation of text with ontologies [3], alignment of dictionaries and ontologies [2], alignments between relational databases and ontologies [9]). These alignment approaches, however, take little account of the alignment of multiple structures. This type of approach is becoming increasingly necessary to manage the growing volume of unstructured information sources available on the Web (encyclopedias such as Wikipedia, social media data, etc.) and LOD knowledge bases. In addition, the approaches are mostly developed for the English language. These needs have to be addressed through a global vision of alignment that takes into account a multiplicity of structures in which knowledge can be expressed. This paper seeks a holistic approach to semantic computing and alignment, when considering heterogeneous structures in which knowledge is represented.

2 Proposal

The approach consists of two main steps. First, knowledge extraction approaches will be applied to extract the terminology of the relevant corpora. We plan to specialise general language models, since the corpora present distinctive language characteristics due to scope and time. We also plan to make use of techniques for the recognition of named entities which might help finding important relations and events. On the basis of the models and recognised entities we plan to extract other information with the help of semantic alignment methods. Second, the extracted terminology will be aligned to existing sources of knowledge (available dictionaries, lexicons, corpora and ontologies). In particular, there are basic ontological concepts describing fundamental elements such as persons, places, periods, and that have to be anchored to what is extracted. Ontologies will be the central focus for semantic alignment of textual occurrences of concepts, and its relations with other semantic sources. The alignment may consider previous semantic knowledge, or might be inferred trough semantic similarity analysis.

We plan to apply our approach on current projects such as the Curvo Semedo's works [6]. This is a corpus integrated by six works published between 1707 and 1727, authored by Alentejo doctor João Curvo Semedo (1635-1719), containing medical and pharmacological knowledge constituted and published in Portuguese. The focus reader of his works, at the time they were recorded, was a less educated person, little affected by the materials available only in Latin. The six works gathered include a collection of about 2,150 pages, which are treated and offered in the form of transcripts, in different formats, in original spelling and reproduced, accompanied by descriptions of their terminologies and representations of the content of each one, generated with the support of computational tools. The evaluation phase will be carried out with he help of humanities expert. The proposed methodology has potential utility for other projects with a variety of history and linguistic inquiries.

References

1. J. Cole, Q. Wang, et al. The ribosomal database project: improved alignments and new tools for rrna analysis. *Nucleic acids research*, 37:D141–D145, 2009.
2. B. Dalvi, E. Minkov, P. P. Talukdar, and W. W. Cohen. Automatic gloss finding for a knowledge base using ontological constraints. In *8th Conf. WSDM*, pages 369–378, 2015.
3. M. Erdmann, A. Maedche, H.-P. Schnurr, and S. Staab. From manual to semi-automatic semantic annotation. In *COLING Workshop on Semantic Annotation*, pages 79–85, 2000.
4. J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg, Germany, 2007.
5. M. Matuschek and I. Gurevych. Dijkstra-wsa: A graph-based approach to word sense alignment. *TACL*, 1:151–164, 2013.
6. P. Quaresma and M. J. B. Finatto. Information extraction from historical texts: a case study. In *DHandNLP@PROPOR*, pages 49–56, 2020.
7. S. Schreibman, R. Siemens, and J. Unsworth. *A new companion to digital humanities*. John Wiley & Sons, 2015.
8. D. Tufiş, A. M. Barbu, and R. Ion. Extracting multilingual lexicons from parallel corpora. *Computers and the Humanities*, 38(2):163–189, 2004.
9. D. Uña, N. Rümmele, G. Gange, et al. Machine learning and constraint programming for relational-to-ontology schema mapping. In *27th IJCAI*, pages 1277–1283, 2018.