NER as a gateway drug to the Linked Data cloud: Application of Named-Entity Recognition on cultural heritage metadata

Simon Hengchen, Max De Wilde and Seth van Hooland

Proposal for the 2014 DHBenelux Conference – The Hague, June 2014

Abstract

Due to massive digitization efforts by libraries and archives, new approaches are needed to make sense of digitized corpora. The notion of *distant reading*, as defined by the literary scholar Moretti[1], has been gaining considerable attention. Instead of traditional methods of *close reading*, consisting of manually reading and interpreting a very limited corpus, cultural heritage institutions are increasingly experimenting with natural language processing to allow *distant reading* practices by end-users. Named-Entity Recognition (NER) is one of these methods, which can help end-users to navigate through large volumes of data, facilitating social network analysis, etc. This method has been mainly applied on full-text documents, but in this presentation we want to demonstrate the possibilities of NER upon descriptive metadata from a cultural heritage corpus.

Different initiatives such as NERD[2] and FreeYourMetadata[3] have proposed a low-cost approach to enable researchers and cultural institutions to use and enrich the available research corpora. Named-Entity Recognition is a technique that allows researchers and cultural institutions to enrich their datasets easily. Nonetheless, we should be wary not to take the outcome of NER at face value. Although the creation of a Gold-Standard Corpus (GSC) allows to objectivate their output in terms of recall, precision and F-score, it also raises other issues. First of all, the creation of a GSC implies the manual tagging of Named Entities in categories. However, experience has shown that NER services propose entities which fall beyond the scope of a GSC: words such as *linocut* or *paleontology* hold potential value and are recognized, but fail to be taken into account in traditional GSC which only include broad categories such as Persons, Organizations and Locations. Another problem raised by the use of NER techniques relates to a core principle of Linked Data: the use of URLs. While NER services extract entities, some of them also disambiguate them – they distinguish the difference between *Washington* as a historical figure, a capital city, a state, a lake, and so on. The four entities are represented by the same chain of characters, yet, as per Linked Data guideline, have different URLs. This feature, while being very useful, also is challenging: what exactly does a URL mean? Since URLs define real-world resources, how should those resources be represented? Which is, between the URL to the Wikipedia article for George Washington and the Wikipedia article itself, the representation of the first president of the USA? Is the URL the identifier for the person, or for the documentation[4] about that person? The question of information resources (that can be electronically represented) and non-information resources (that cannot be electronically represented) remains to be tackled.

In order to demonstrate the relevance of these research questions, a case study using metadata from the historical archives of the city of Québec will be developed. The dataset spans over 22,500 records in French and contains more than 60,000 entities – most of which are persons and locations, but also a few organizations. It covers descriptions of photographs, engravings and slide films of Québec. Based on this concrete case study, future developments for the use and evaluation of NER will be proposed.

*Affiliation: Simon Hengchen, Max De Wilde, Seth van Hooland: *Université Libre de Bruxelles* Corresponding author: shengche@ulb.ac.be

[1] Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London & New York: Verso, 2005. Print.

[2] http://nerd.eurecom.fr

[3] http://freeyourmetadata.org

[4] Note the difference between http://dbpedia.org/resource/George_Washington and http://dbpedia.org/page/George_Washington, the first URL being the resource, the second the documentation. In DBpedia, the *resource* URL automatically redirects to the *page* URL