Semantic Enrichment of a Multilingual Archive with Linked Open Data

Max De Wilde
Université libre de Bruxelles
madewild@ulb.ac.be

Simon Hengchen
Université libre de Bruxelles
shengche@ulb.ac.be

The Historische Kranten[1] project involved the digitisation, OCR and online publication of over a million articles from 41 Belgian newspapers published between 1818 and 1972. Articles are written in Dutch, French and English and focus mainly on the city of Ypres and its neighbourhood.

Currently, only full-text indexing has been performed on the collection, which means that search for particular mentions in the corpus suffer from both noise and silence. For instance, a search on the string "Huygens" returns correct results about Christiaan Huygens:

> *Links zien wij Christiaan Huygens die met zijn slingeruurwerk de oplossing bracht voor het meten van de tijd*

But one also gets results that are not relevant in this context (noise):

> *La reconnaissance du cadavre de la veuve Huygens, faite par les hommes de l'art, a fait constater l'existence de neuf blessures sur la tête*

Moreover, interesting results are lost due to variations in spelling (silence):

> *en op het uurwerk toegepast door den Hollander Huyghens (1629-1695).*

We first performed Named-Entity Recognition (NER) on this collection in order to extract meaningful concepts. A second step involved a new approach to Entity Linking with gazetteers (Shen et al., 2014) in order to disambiguate them with DBpedia URIs[2] (Bizer et al., 2009). For instance, http://dbpedia.org/resource/Christiaan_Huygens includes the alternative label "Christian Huyghens" (French spelling) but excludes information about the Belgian painter Léon Huygens (which has his own

[1] http://www.historischekranten.be/

[2] We use DBpedia as an entry point to the Linked Data cloud, enabling access to other resources with the `owl:sameAs` property.

unique URI: http://dbpedia.org/resource/Léon_Huygens) or the crater on Mars named after the Dutch astronomer (http://dbpedia.org/resource/Huygens_(crater)).

We now intend to integrate our findings into the project's web interface in order to improve the search experience of the end-users. We plan to interact with the users to get feedback about the relevance of entities extracted and of automatic related search suggestions based on semantic relatedness, which are currently quite random and of poor quality.

The impact of OCR quality on NER output will also be evaluated. In a similar experiment on Holocaust-related archives, Rodriquez et al. (2012) find, somewhat counter-intuitively, that "manual correction of OCR output does not significantly improve the performance of named-entity extraction". The confirmation of this hypothesis would mean a lot to institutions that lack sufficient funding to perform first-rate OCR on their collections.

References

Bizer, C., Lehman, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia – A crystallization point for the Web of Data. *Web Semantics: science, services and agents on the World Wide* Web, 7(3):154-165.

Rodriquez, K. J., Bryant, M., Blanke, T., and Luszczynska, M. (2012). Comparison of Named Entity Recognition tools for raw OCR text. In Proceedings of KONVENS 2012, pp 410–414. Vienna.

Shen, W., Wang, J., and Han, J. (2014). Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*.