Linking Named Entities in Dutch Historical Newspapers

Theo van Veen, Juliette Lonij and Willem Jan Faber

Koninklijke Bibliotheek, National Library of the Netherlands
{theo.vanveen,juliette.lonij,willemjan.faber}@kb.nl

Abstract. We improved access to the collection of Dutch historical newspapers of the Koninklijke Bibliotheek by linking named entities in the newspaper articles to corresponding Wikidata descriptions by means of machine learning techniques and crowdsourcing. Indexing the Wikidata identifiers for named entities together with the newspaper articles opens up new possibilities for retrieving articles that mention these resources and searching the newspaper collection using semantic relations from Wikidata. In this paper we describe our steps so far in setting up this combination of entity linking, machine learning and crowdsourcing in our research environment as well as our planned activities aimed at improving the quality of the links and extending the semantic search capabilities.

Keywords: named entities, linked data, entity linking, semantic enrichment, semantic search, machine learning, classification, crowdsourcing

1 Introduction

One of the strengths of the semantic web [1] is the possibility it offers to identify resources and link the mentions of these resources to relevant descriptions from external data sources. In the research environment of the Koninklijke Bibliotheek (KB) we started to enrich Dutch historical newspaper articles with named entities (i.e. names of persons, locations, organizations and others) linked to their resource descriptions in various knowledge bases, such as DBpedia [2], Wikidata [3] and VIAF [4]. We combine all relevant links for an entity into a single enrichment record, which is stored in a dedicated enrichment database. We discussed the data model for these enrichment records and the architecture of our enrichment infrastructure in a previous article [5]. This paper will focus on the process of entity linking and the application of the results in semantic search, as currently available in the KB research environment.

Entity linking has received much attention in recent years and often DBpedia plays a central role [6, 7]. In paragraph 2 we will describe the automatic process we developed for generating links to DBpedia based on machine learning techniques and entity context information. As we cannot train the machine-learning algorithm to give results with 100% accuracy, we will consider crowdsourcing as an option to correct false and missing links in paragraph 3. Indexing generated links makes it possible to use them in various forms of semantic search, some of which we have implemented in a purpose-built research portal that we will present in paragraph 4. We will draw

some provisional conclusions and point out the next steps we have planned for improving the quality of the links and extending the semantic search capabilities in paragraphs 5 and 6 respectively.

2 The automatic linking process

The development of the automatic entity linking process has gone through a number of different phases.

Our first attempt of searching all titles from DBpedia descriptions in our newspapers was not very successful. Thus, in the next phase, we did it the other way around: we first performed named entity recognition [8] on the newspaper articles and subsequently looked for matching descriptions in DBpedia. For this purpose we constructed a Solr [9] index out of DBpedia dumps, combining the relevant data, such as label, abstract, VIAF and Wikidata identifiers, for each resource into records. The most appropriate link candidate was selected with a simple, rule-based approach looking at various forms of string matching (e.g. exact match, partial match, last-part match) and a few additional features such as the number of inlinks. This stage is illustrated in figure 1.

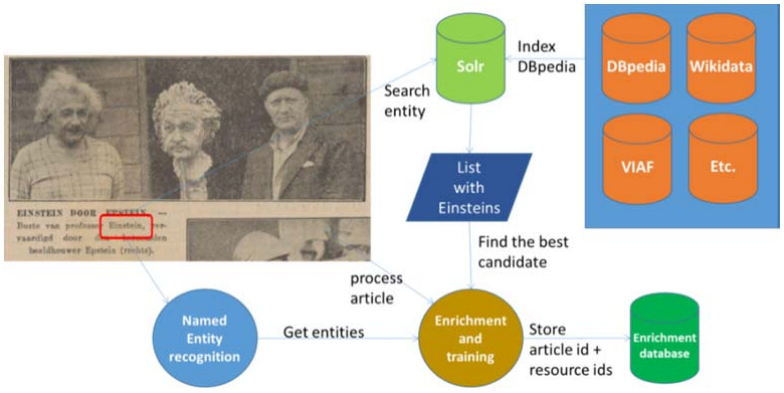

Fig. 1. Schematic overview of the process for entity linking

The third phase was speeding up the process by using the processing capacity of the HPC cloud infrastructure at SURFsara [10]. This has increased the throughput of the number of articles by more than an order of magnitude. Because of the large number of articles and the time it still takes to process them all, we keep the linking process running continuously, with each improved version starting from the point in the collection where the previous version left off. Once all articles have been processed we start at the beginning again. In this way we are able to enrich the entire collection of articles as quickly possible, improving the quality level with each iteration.

In the fourth phase we started applying machine-learning techniques to the linking process, training a Support Vector Machine classifier [11] on a labeled example set of several thousands of potential links. To the string matching features we added a number of features based on contextual information. These include the type and subtype match, e.g. “person” or “politician”, between the named entity and the candidate description, the occurrence of other named entities from the article in the DBpedia abstract, and the compatibility of the publication year of the article and any known year of birth.

During this last phase the fact that DBpedia is language dependent became increasingly problematic. We wanted to include data from both Dutch and English descriptions in the index, because the English DBpedia dumps contain more names and the English descriptions often provide additional data about a resource. To avoid having to deal with different identifiers for the same entity in different languages, we decided to switch to the Wikidata identifiers as the main identifier. From Wikidata we can still obtain all the links to DBpedia, as well as many other databases.

We measured the quality of the results from the second phase onward by means of an accuracy score for a manually linked evaluation set of 349 named entities. With the rule-based approach we were able to obtain an accuracy of 74,50% after the last iteration. For the machine-learning approach, while improving the features, the accuracy gradually increased to 83,09%. Based on manual inspection of the results we estimate that approximately 5% of the examples cannot be automatically linked because of serious OCR errors or because a significant amount of human knowledge about the entity is required. Thus, we expect 95% to be the maximum accuracy possible and we hope to get closer to that number using a neural network approach.

3 Crowdsourcing

Since the accuracy of the results of the automatic linking process will never reach 100% we need user feedback to make corrections and add missing links. This crowdsourced data can be used to extend the training set for the machine-learning algorithm and user feedback is also useful for preventing the suggestion to the end user that the links are 100% reliable.

Users can currently provide feedback through an enrichment page displaying a newspaper article with the linked named entities marked in the text. Clicking on a name shows the linked resources for that name and the option to remove incorrect links. Selecting the text of a name appearing in the article will enable the option to add a new link to that name, either by entering a URL or by choosing the most appropriate entry in DBpedia.

At the moment the enrichment page is only available to KB-employees. We expect a significant number of other users to be intrinsically motivated to contribute links in areas of their interest or expertise, but before we can offer this functionality to the general public we will need to take measures to minimize the chances of abuse and the introduction of errors. Options such as crowdsourced moderation are being considered, but the exact measures have yet to be decided upon.

4 Semantic search

The links resulting from the automatic linking process and from crowdsourcing can be used to improve accessibility and usability of the newspaper collection, which is the actual goal of our project.

The links function as identifiers and, by indexing them along with the newspaper articles, knowledge bases can play the role of thesaurus for the named entities. We believe that Wikidata is the most promising candidate for this purpose, since Wikidata is not restricted to a specific domain, like VIAF, and it is language independent, as opposed to DBpedia. It already has quite a large number of descriptions (over 19 million at the moment [12]) and can easily be extended with entries for entities that are not yet included.

Using the indexed Wikidata identifiers it also becomes possible to search the newspaper articles for resources based on semantic relations present in Wikidata. This concept, as shown in Figure 2, is very simple: a list of Wikidata identifiers resulting from a SPARQL [13] query in Wikidata is used as input for a conventional SRU [14] query in the enriched newspaper article index. As Wikidata resource descriptions contain many other existing resource identifiers, this approach can even be applied to library catalogues and other databases by replacing the obtained Wikidata identifiers with the local identifiers listed in Wikidata.

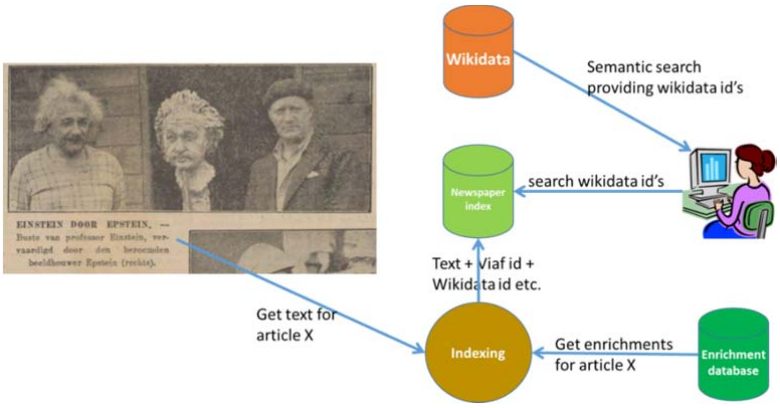

Fig. 2. Overview of the usage of the Wikidata identifier

For demonstration purposes we have developed a research portal [15] for searching and viewing the KB collections and their various enrichments, a screenshot of which is shown in Figure 3. The following semantic query functionality is available for the enriched newspaper index ("Newspapers +"), the last three options for the library catalogue ("KB Catalogue") as well:

- Each resource identified in an article is provided with an infobox that has the option to search articles mentioning this specific resource using its Wikidata identifier.
- The infobox also contains contextual information about the resource in the form of properties from Wikidata and allows searching for all resources in the newspaper articles with the same value for a particular property by clicking on that value. The actual query that is generated looks like [property=value] using the Wikidata identifiers of both property and value.
- Using square brackets in a query, e.g. “[Beatles]”, the application tries a few SPARQL queries using different properties as a first guess, thus avoiding the need for user knowledge on Wikidata property names. In this example the property that can have “Beatles” as its value turns out to be “is member of” and the query results are the articles enriched with the Wikidata identifier of any of five the members of the Beatles (Pete Best being the fifth Beatle).
- For advanced users it is possible to enter a property-value combination between square brackets, such as “[P737=Q1203]”, which will search for articles mentioning resources that are influenced (P737) by John Lennon (Q1203). Here, the user must know the Wikidata identifiers of both the property and the value but in a future version we might give the user some help in finding these identifiers.
- The even more skilled user might enter a very complex SPARQL query on the Wikidata website directly and use the resulting Wikidata identifiers in a query on the research portal. In this case the query is treated as a conventional query, but still the added value of our approach is that the Wikidata identifiers are available in the index.

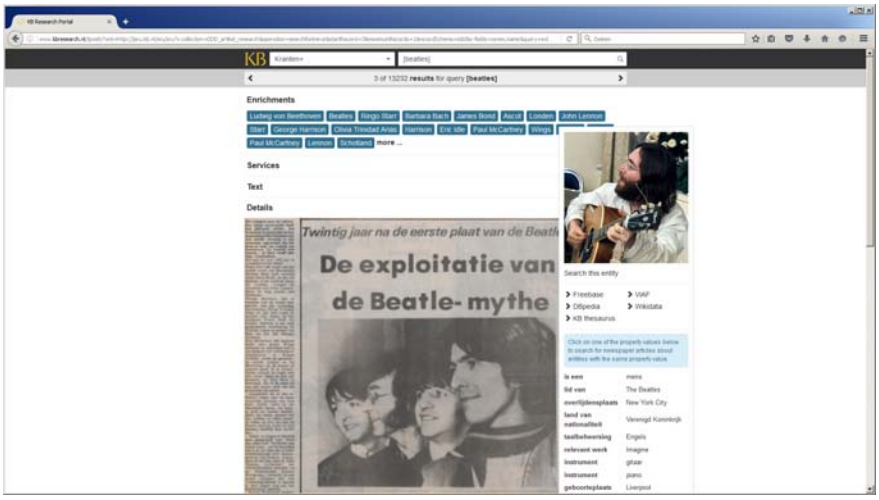

Fig. 3. Screenshot of the research portal

5 Conclusion

In this paper we have described the combination of building blocks making up the entity linking process for Dutch historical newspapers currently running in the KB research environment, as well as the ways in which the resulting links can be used to improve access and usability. The linking quality achieved so far is quite promising, as we expect to be able to further increase it. Our approach of using Wikidata as both thesaurus and data source for semantic search is conceptually simple and makes available new search functionality. Adopting a "release early, release often" strategy has ensured that we can already use and assess these results in our research portal. New links and functionality are added there as soon as they become available.

6 Next steps

We have several next steps planned for further improving the quality of the links and the possibilities for semantic search. The first step is replacing the conventional classification algorithm with a neural network approach. Moreover, we want to start using the data obtained from crowdsourcing as additional input for training the network. Another, somewhat more distant goal is to extract new relations between named entities occurring in the newspaper collection that are not (yet) part of knowledge bases. We hope to be able to present some results of these upcoming steps at the conference.

References

1. Semantic Web, http://www.w3.org/standards/semanticweb/
2. DBpedia, http://dbpedia.org/
3. Wikidata, https://www.wikidata.org/
4. VIAF, Virtual International Authority File, http://viaf.org/
5. Van Veen, T., Lonij, J. & Koppelaar, H. (2015). Semantic Enrichment: a Low-barrier Infrastructure and Proposal for Alignment, D-Lib Magazine, DOI: 10.1045/july2015-vanveen.
6. Odijk, D., Meij, E., & de Rijke, M. (2013). Feeding the Second Screen: Semantic Linking based on Subtitles. In Open research Areas in Information Retrieval (OAIR 2013). Lisbon.
7. Sil, A., Croning, E., et al. (2012). Linking named entities in any database. In EMNLP-CoNLL '12 Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea.
8. Stanford Named Entity Recognizer, http://nlp.stanford.edu/software/CRF-NER.shtml
9. Apache Solr, http://lucene.apache.org/solr/
10. SURFsara, https://www.surf.nl/en/services-and-products/hpc-cloud/
11. mySVM, http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html
12. Wikidata Statistics, https://www.wikidata.org/wiki/Wikidata:Statistics
13. SPARQL, Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/
14. SRU, Search and Retrieval via URL's, http://www.loc.gov/standards/sru/
15. KB Research Portal, http://www.kbresearch.nl/xportal/