Linking, Searching, and Visualizing Entities for the Swedish Wikipedia

Anton Södergren, Marcus Klang, and Pierre Nugues

Department of computer science
Lund University, Lund
karl.a.j.sodergren@gmail.com, marcus.klang@cs.lu.se, pierre.nugues@cs.lth.se

Abstract

In this paper, we describe a new system to extract, index, search, and visualize entities on Wikipedia. To carry out the extraction, we designed a high-performance entity linker and we used a document model to store the resulting linguistic annotations. The entity linker, HERD, extracts the mentions from text using a string matching engine and links them to entities with a combination of rules, PageRank, and feature vectors based on the Wikipedia categories. The document model, Docforia, consists of layers, where each layer is a sequence of ranges describing a specific annotation, here the entities. We evaluated HERD with the ERD'14 protocol (Carmel et al., 2014) and we reached the competitive F1-score of 0.746 on the English development set. We applied HERD to the whole collection of Swedish articles of Wikipedia and we used Lucene to index the layers and a search module to interactively retrieve articles and metadata given a title, a phrase, or a property. The user can then select an entity and visualize concordance in articles or paragraphs. A demonstration of the entity search and visualization is available for Swedish at this address: http://vilde.cs.lth.se:9001/sv-herd/.

1. Introduction

Wikipedia has become a popular NLP resource used in many projects such as text categorization or translation. In addition to its size and diversity, Wikipedia, through its links, also enables to create a graph that associates concepts, entities, and their mentions in text. However, according to the edition rules of Wikipedia, only the first mention of an entity should be linked in an article. The automatic linking of the subsequent mentions is called a wikification (Mihalcea and Csomai, 2007) and most reported works in this field apply to English.

In this paper, we describe a new multilingual system to process, index, search, and visualize entities on Wikipedia. The entity linker, HERD, extracts the mentions from text using a string matching engine and links them to entities with a combination of rules, PageRank, and feature vectors based on the Wikipedia categories. To carry out the indexing, we used a document model to store the linguistic annotations and a high-performance entity linker. The document model consists of layers, where each layer is a sequence of ranges describing a specific annotation, here the entities.

We evaluated HERD with the ERD'14 protocol (Carmel et al., 2014) and we reached the competitive F1-score of 0.746 on the English development set. We applied HERD to the whole collection of Swedish articles of Wikipedia and we used Lucene to index the layers and a search module to interactively retrieve articles and metadata given a title, a phrase, or a property. The user can then select the articles or paragraphs s/he wants to visualize.

2. Entity Linking

2.1 Extraction of Wikipedia Structure

Before we apply the linker to Wikipedia, we convert the HTML dumps into a multilayer document model; see Sect. 4. This preprocessing step parses the HTML documents into DOM trees and extracts the original page structure, text styles, links, lists, and tables. We then resolve all the Wikipedia links to unique Wikidata identifiers, where Wikidata is an entity database, which assigns unique identifiers across all the language editions of Wikipedia. Berlin, for instance, has the unique id: Q64 that enables to retrieve the article pages in English, French, Swedish, or Russian. Figure 1 shows examples of these ids, where we have replaced the manually set Wikipedia anchors (the wikilinks) with their Wikidata numbers: Q183 for *Tyskland* 'Germany' and Q5119 for *capital* 'huvudstad.'

2.2 Entities

Once we have extracted and structured the text, we apply the entity linking module that finds mentions of entities in text and link these mentions to a unique identifiers. We used Wikipedia as a knowledge source for both names and concepts, and Wikidata for unique identifiers. The system collects the links on Wikipedia articles to count and analyze them. The link is seen as a mention, that consists of a label and an address, that the system uses as a name and an identifier. The address is translated into a Wikidata Q-number. When the system parses a new document, each recognized name is linked to a unique identifier.

Following Lipczak et al. (2014), we applied the Solr Text Tagger (Smiley, 2013) to spot the mentions. This tagger uses finite-state transducers and is efficient in terms of memory usage and execution time. The results are then indexed using Lucene. We applied logistic regression, PageRank, and feature vectors based on the Wikipedia categories to improve the name recognition, and select the best candidate for each name. We evaluated the system with the same method as used in the ERD'14 competition (Carmel et al., 2014) and we reached the competitive F1-score of 0.746 on English. We applied our linker to Swedish without any language adaptation.

3. Processing Wikipedia

We deployed the entity linker on a dump of the Swedish Wikipedia. Since the dataset is quite large, we needed to speed up the tagging by parallelizing the process. To do so, we used Apache Spark and the HDFS distributed storage

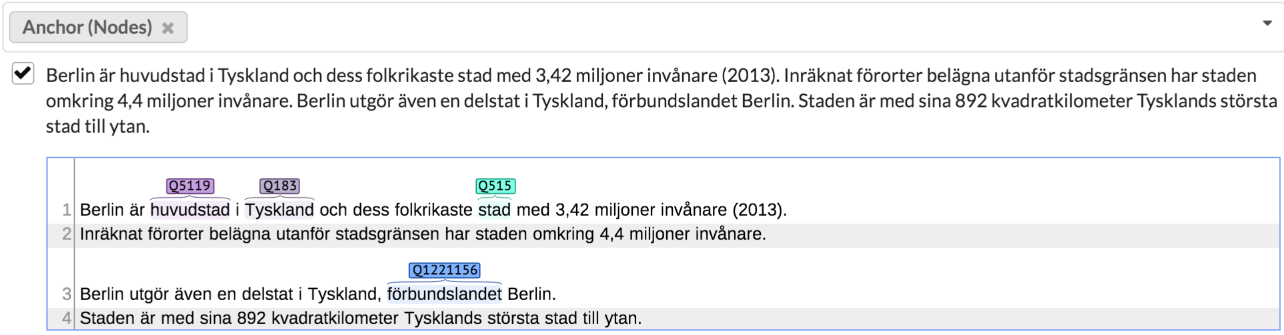


Figure 1: Visualization of anchors with Wikidata Q-numbers

system on a cluster of 12 computers.

In order to run the algorithm efficiently on our cluster, all parts of it has to support concurrent execution. The Solr Text Tagger introduced library conflicts with the Spark environment. To remove them, we modified the Solr Text Tagger to skip its dependencies on Solr and use its underlying indexing engine Lucene directly. We thus improved its performance while reducing the number of dependencies.

We deployed the entity linker on our cluster and we used HDFS to spread the Wikipedia dump across the nodes as well as to save the final result. With the setup we used, the dump of the Swedish Wikipedia took around three hours to process.

4. The Document Model

We represented and stored Wikipedia using the Docforia document model[1] (Klang and Nugues, 2016b; Klang and Nugues, 2016a). Docforia is designed it so that we could store the original markup, as well as the subsequent linguistic annotations. It consists of multiple layers, where each layer is dedicated to a specific type of annotation. The annotations are encoded in the form of graph nodes, where a node represents a piece of data: a token, a sentence, a named entity, etc., delimited by ranges. These nodes are possibly connected by edges as in dependency graphs.

5. Indexing

We created an indexing tool that is based on Lucene that we called Panforia. Lucene is a search and indexing library that is easy to embed. We saved the output of the annotation pipeline as Parquet files that serve as input to the Panforia indexer. Each Docforia record is converted into a Lucene document by mapping record properties and documents to Lucene fields. In addition, a binary copy of the Docforia record is embedded with the indexed fields, which provides the ranges and relationships between nodes needed for the visualization.

Building directly on the Lucene library, instead of existing packages such as Solr or ElasticSearch, allows us to optimize the insertion into an index. One key advantage of the Panforia indexer is that it can read the output from the Wikipedia pipeline, Parquet files, without a conversion step.

[1] `https://github.com/marcusklang/docforia/`

6. Visualization and Demonstration

The front-end of Panforia is a web server that embeds the Docforia library, Lucene, and a client-side web application.

Figure 2 shows an example of results we obtained when we searched the entity Göran Persson, the former Swedish Prime Minister, using its Q-number: Q53747. This mention (*Göran Persson*) is ambiguous and the Swedish Wikipedia lists as many as four different entities with this name: The former Swedish Prime Minister, a progressive musician (Q6042900), a Swedish social democratic politician, former member of the Riksdag (Q5626648), and a Swedish statesman from the 16th century (Q2625684). The latter is also being spelled Jöran Person.

Searching the mention *Göran Persson* would return articles or concordances with any of these entities, while searching the entity through its Q-number only returns the intended person, either with her/his name or with other mentions such as *Persson*. The results are given in the forms of concordances with a left and right contexts (Fig. 2). The column to the left is the Q-number of the source document in the Swedish Wikipedia and the left column is the offset from the beginning of this document.

The demonstration is available at: `http://vilde.cs.lth.se:9001/sv-herd/`.

Acknowledgements

This research was supported by Vetenskapsrådet, the Swedish research council, under the *Det digitaliserade samhället* program.

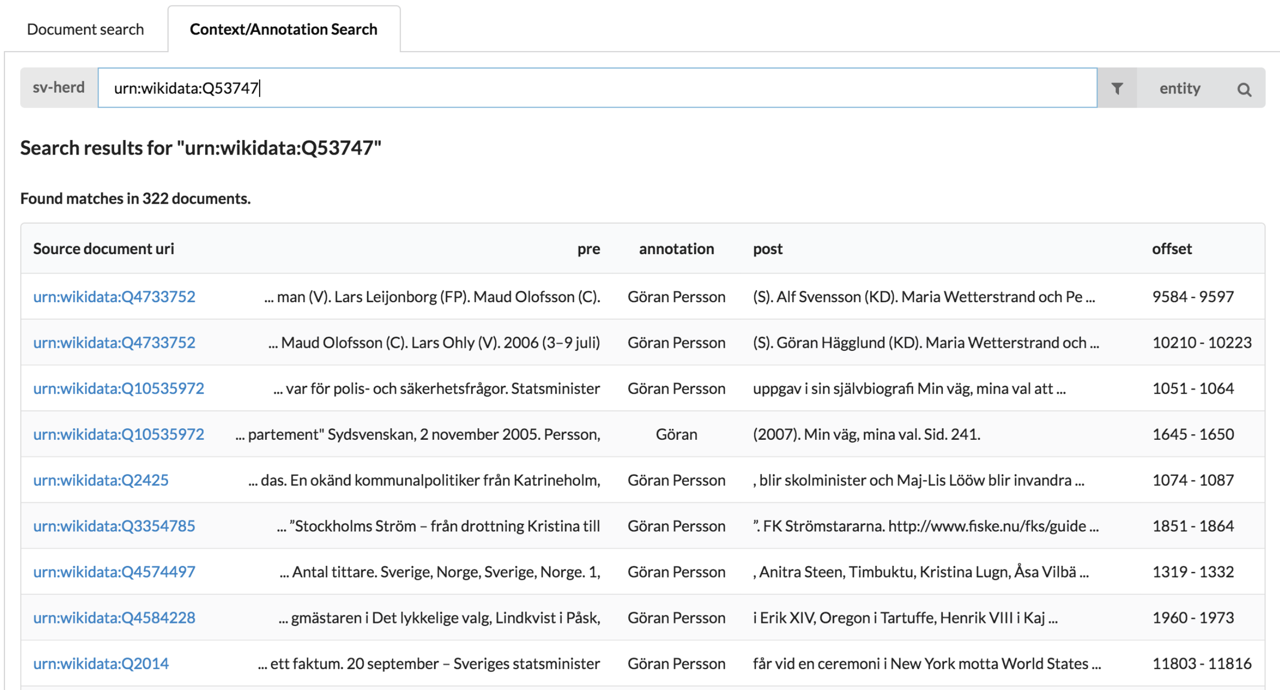
References

David Carmel, Ming-Wei Chang, Evgeniy Gabrilovich, Bo-June Paul Hsu, and Kuansan Wang. 2014. ERD'14: Entity recognition and disambiguation challenge. In *ACM SIGIR Forum*, volume 48, pages 63–77. ACM.

Marcus Klang and Pierre Nugues. 2016a. Docforia: A multilayer document model. In *Proceedings of SLTC 2016*, Umeå, November.

Marcus Klang and Pierre Nugues. 2016b. Wikiparq: A tabulated Wikipedia resource using the Parquet format. In *Proceedings of LREC 2016*, pages 4141–4148, Portorož, Slovenia.

Marek Lipczak, Arash Koushkestani, and Evangelos Milios. 2014. Tulip: Lightweight entity recognition and disambiguation using wikipedia-based topic centroids. In



Source document uri	pre	annotation	post	offset
urn:wikidata:Q4733752	... man (V). Lars Leijonborg (FP). Maud Olofsson (C).	Göran Persson	(S). Alf Svensson (KD). Maria Wetterstrand och Pe ...	9584 - 9597
urn:wikidata:Q4733752	... Maud Olofsson (C). Lars Ohly (V). 2006 (3–9 juli)	Göran Persson	(S). Göran Hägglund (KD). Maria Wetterstrand och ...	10210 - 10223
urn:wikidata:Q10535972	... var för polis- och säkerhetsfrågor. Statsminister	Göran Persson	uppgav i sin självbiografi Min väg, mina val att ...	1051 - 1064
urn:wikidata:Q10535972	... partement" Sydsvenskan, 2 november 2005. Persson,	Göran	(2007). Min väg, mina val. Sid. 241.	1645 - 1650
urn:wikidata:Q2425	... das. En okänd kommunalpolitiker från Katrineholm,	Göran Persson	, blir skolminister och Maj-Lis Lööw blir invandra ...	1074 - 1087
urn:wikidata:Q3354785	... "Stockholms Ström – från drottning Kristina till	Göran Persson	". FK Strömstararna. http://www.fiske.nu/fks/guide ...	1851 - 1864
urn:wikidata:Q4574497	... Antal tittare. Sverige, Norge, Sverige, Norge. 1,	Göran Persson	, Anitra Steen, Timbuktu, Kristina Lugn, Åsa Vilbä ...	1319 - 1332
urn:wikidata:Q4584228	... gmästaren i Det lykkelige valg, Lindkvist i Påsk,	Göran Persson	i Erik XIV, Oregon i Tartuffe, Henrik VIII i Kaj ...	1960 - 1973
urn:wikidata:Q2014	... ett faktum. 20 september – Sveriges statsminister	Göran Persson	får vid en ceremoni i New York motta World States ...	11803 - 11816

Figure 2: Concordances of the entity *Göran Person*, Q53747. The results are given in the forms of concordances with a left and right contexts. The column to the left is the Q-number of the source document in the Swedish Wikipedia and the left column is the offset from the beginning of this document.

Proceedings of the First International Workshop on Entity Recognition & Disambiguation, ERD '14, pages 31–36, New York, NY, USA. ACM.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on CIKM*, CIKM '07, pages 233–242, Lisbon, Portugal.

David Smiley. 2013. Solr text tagger, text tagging with finite state transducers. `https://github.com/OpenSextant/SolrTextTagger`.