Named Entity Disambiguation in a Question Answering System

Marcus Klang, Pierre Nugues

Lund University, Department of Computer Science
S-221 00 Lund, Sweden
marcus.klang@cs.lth.se, pierre.nugues@cs.lth.se

Abstract

In this paper, we describe how we use a named entity disambiguation module to merge entities in a question answering system. The question answering system uses a baseline passage retrieval component that extracts paragraphs from the Swedish version of Wikipedia. The passages are indexed and ranked using the Lucene platform. Prior to the indexing, we carried out a recognition and disambiguation of the named entities. We used the Stagger part-of-speech tagger to tag the documents and we implemented a disambiguation module. We extracted and merged the answer candidates using their wikidata Q-number. This enables the question-answering system to benefit from the synonymy facility of Wikipedia as well as an extended set of properties.

1. Introduction

Factoid question answering systems aim at answering short questions, where answers often consist of a single concept or a named entity. The typical architecture of most such systems features a question analysis step, a passage retrieval step, where documents containing the answer are extracted from a collection of texts, and an extraction and ranking step of the candidate answers. See IBM Watson (Ferrucci, 2012) for an example of such an architecture applied to the *Jeopardy!* quiz show.

When the answers correspond to named entities, their identification in the texts that serve as knowledge source and their disambiguation enables the answer extraction step to merge more accurately the candidates coming from different passages. In addition, it makes it possible to carry out inferencing over external structured data sources that can be associated to these entities. While there exist available named entity disambiguators targeted to English, such as AIDA (Hoffart et al., 2011) or SMAPH (Cornolti et al., 2014), to the best of knowledge, there is nothing equivalent for Swedish. In this paper, we describe a named entity disambiguator for Swedish and its integration to the Hajen question-answering system.

2. Named Entity Disambiguation

Named entity disambiguation or named entity linking consists in associating a sequence of words, typically a proper noun, to a unique identifier. As source of identifiers, we can use entity repositories, such as Freebase (Bollacker et al., 2008) or Wikidata[1], that define popular nomenclatures.

In the Hajen system, we use the wikidata identifiers that gather properties collected from the Wikipedia online encyclopedia and the infobox tabulated information that is associated to some of its articles. The city of Lund, Sweden, for example, has Q2167 as wikidata identifier, while Lund University has number Q218506. The associated properties are then accessible from the URL: `http://www.wikidata.org/wiki/Qxxx`.

We carry out the named entity linking in three steps: We first use a part-of-speech tagger, Stagger (Östling, 2013), to recognize the named entities; we link the strings to possible wikidata identifiers; and we finally disambiguate them using their popularity, commonness, and a Boolean context method. In a sentence like:

> Göran Persson var statsminister mellan åren 1996 och 2006 samt var partiledare för Socialdemokraterna
> 'Göran Persson was prime minister between 1996 and 2006 and was leader of the Social Democrats',

we identify *Göran Persson* and *Socialdemokraterna* as proper nouns. We then identify the entities matching these strings. The Swedish Wikipedia lists four *Göran Persson* with four different wikidata Q-numbers:

1. Göran Persson (född 1949), socialdemokratisk partiledare och svensk statsminister 1996–2006 (The Swedish Prime Minister): Q53747;

2. Göran Persson (född 1960), socialdemokratisk politiker från Skåne (A Swedish Member of Parliament): Q5626648;

3. Göran Persson (musiker), svensk proggmusiker (A Swedish composer): Q6042900;

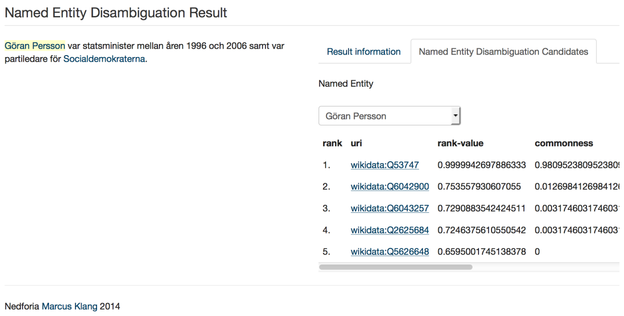
4. Jöran Persson, svensk ämbetsman på 1500-talet (A Swedish state councillor from the 16th century, whose first name can be spelled either Jöran or Göran): Q2625684.

We finally rank these candidates using their popularity and context (Klang and Nugues, 2014). Figure 1 shows the output produced by the disambiguation module.

3. Entity Linking and Question Answering

We integrated our named entity linker in a baseline question answering system. Following IBM Watson (Chu-Carroll et al., 2012), we used the Swedish version of Wikipedia as textual knowledge source. We processed the whole corpus with the linker so that we associated all the entities we could recognize to their Q-number. For the named entities

[1] `http://www.wikidata.org/`



Named Entity Disambiguation Result

Göran Persson var statsminister mellan åren 1996 och 2006 samt var partiledare för Socialdemokraterna.

Result information | Named Entity Disambiguation Candidates

Named Entity

Göran Persson

rank	uri	rank-value	commonness
1.	wikidata:Q53747	0.9999942697886333	0.980952380952380
2.	wikidata:Q6042900	0.753557930607055	0.012698412698412
3.	wikidata:Q6043257	0.7290883542424511	0.003174603174603
4.	wikidata:Q2625684	0.7246375610550542	0.003174603174603
5.	wikidata:Q5626648	0.6595001745138378	0

Nedforia Marcus Klang 2014

Figure 1: The output of the entity disambiguation module

we identified with the POS tagger that had no Q-number, we used the Wikipedia page name as identifier instead, e.g. "Göran_Persson" for the Prime Minister. We segmented the articles into paragraphs and we indexed them using Lucene (Apache Software Foundation, 2014).

We used a corpus of questions and answers transcribed from the SVT *Kvitt eller Dubbelt – Tiotusenkronorsfrågan* game (Thorsvad and Thorsvad, 2005). Given a question, we retrieve the set of Wikipedia passages having the highest similarity using Lucene's $TF.IDF$ implementation.

3.1 Merging the Candidates.

We applied a POS tagger to the passages and we extracted the common nouns, proper nouns, and named entities. We merged all the strings that could be linked to a unique identifier and we created a list of synonyms with the resulting set. When the strings have no identifier, we merge them either by lemma or surface forms. These strings usually consist of a single token: a noun. However, as the POS tagger identifies multiword named entities, a candidate may consist of multiple tokens. In such a case, it is merged in the same way as single tokens.

3.2 Ranking the Candidates.

Ranking the candidate answers to a question is done by frequency, i.e. the number of candidate occurrences after merging. Figure 2 shows the output of the system to the question:

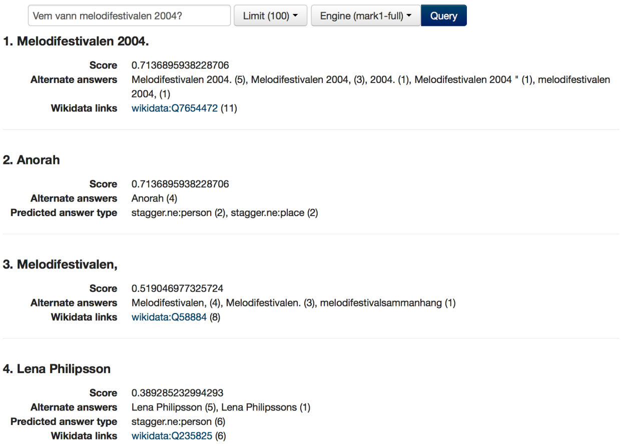
> Vem vann melodifestivalen 2004?
> 'who won the Swedish Melody Festival in 2004?'.

4. Results and Evaluation

We evaluated the question answering system using four metrics: the median rank, mean rank, number of answered questions (recall), and the mean reciprocal rank (MRR), where MRR $= \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}$, Q is the set of answered questions, and rank$_i$ is the rank of question i. The rank metrics only consider answered questions and we set the limit of retrieved paragraphs to 100 for all the systems.

We considered a question answered if the correct answer could be found in the list of candidate answers. As comparison criterion, we used a lowercase exact string match between the corpus answer and the answers provided by the system in the form of lemma and surface form.

We tested three system configurations:



Vem vann melodifestivalen 2004? | Limit (100) | Engine (mark1-full) | Query

1. Melodifestivalen 2004.

Score	0.7136895938228706
Alternate answers	Melodifestivalen 2004. (5), Melodifestivalen 2004, (3), 2004. (1), Melodifestivalen 2004 " (1), melodifestivalen 2004, (1)
Wikidata links	wikidata:Q7654472 (11)

2. Anorah

Score	0.7136895938228706
Alternate answers	Anorah (4)
Predicted answer type	stagger.ne:person (2), stagger.ne:place (2)

3. Melodifestivalen,

Score	0.519046977325724
Alternate answers	Melodifestivalen, (4), Melodifestivalen. (3), melodifestivalsammanhang (1)
Wikidata links	wikidata:Q58884 (8)

4. Lena Philipsson

Score	0.389285232994293
Alternate answers	Lena Philipsson (5), Lena Philipssons (1)
Predicted answer type	stagger.ne:person (6)
Wikidata links	wikidata:Q235825 (6)

Figure 2: The output of the question answering system

System	MRR	Median	Mean	Recall
Nouns only	0.138	29	109.0	0.52
Disambig. only	0.381	4.5	22.8	0.25
Full	0.172	27.0	121.2	0.58

Table 1: The results

Noun only: Only nouns and proper nouns, always single tokens;

Disambiguated only: Named entities that were successfully connected to a wikidata or Wikipedia entity; and

Full: Nouns, named entity candidates, disambiguations, paragraph titles, and numerals.

Table 1 shows the results we obtained. The disambiguation step provide a much better precision but a poorer recall. The opposite applies for nouns, where the recall is higher but precision is lower. The full system with disambiguation increases MRR by nearly 25% compared to only nouns, and the recall by 5.6%.

5. Conclusion

A candidate merging step using a named entity linking module produces high precision results, although due to the entity coverage of Wikidata, it misses a significant part of the answers. A baseline merging method has a much lower precision, but a better recall. When combining both methods, we can observe an increase in both precision and recall over the baseline. More importantly, named entity disambiguation provides entry points to structured data, which would allow questions to be answered using a deeper analysis such as inferencing over structured data.

Acknowledgements

This research was supported by Vetenskapsrådet under grant 621-2010-4800 and the *Det digitaliserade samhället* program.

References

Apache Software Foundation. 2014. Apache Lucene Core. http://lucene.apache.org/core/.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, Vancouver, Canada.

Jennifer Chu-Carroll, James Fan, Nico Schlaefer, and Wlodek Zadrozny. 2012. Textual resource acquisition and engineering. *IBM Journal of Research and Development*, 56(3-4):4:1–4:11.

Marco Cornolti, Paolo Ferragina, Massimiliano Ciaramita, Hinrich Schütze, and Stefan Rüd. 2014. The SMAPH system for query entity recognition and disambiguation. In *Proceedings of Entity Recognition and Disambiguation, ERD'14*, Gold Coast.

David Angelo Ferrucci. 2012. Introduction to "This is Watson". *IBM Journal of Research and Development*, 56(3.4):1:1 –1:15, May-June.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh.

Marcus Klang and Pierre Nugues. 2014. A platform for named entity disambiguation. In *Proceedings of the workshop on semantic technologies for research in the humanities and social sciences (STRiX)*, Gothenburg. To appear.

Robert Östling. 2013. Stagger: an open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology*, 3:1–18.

Karin Thorsvad and Hasse Thorsvad. 2005. *Kvitt eller Dubbelt*. SVT Tactic, Stockholm.