Bootstrapping a Swedish knowledge base from Wikipedia

Isak Lyckberg, Antonina Tokarchuk
Dept. of Computer Science, LTH,
P.O. Box 118,
221 00 Lund, Sweden
lyckberg@gmail.com, antonina.tt@gmail.com

Abstract

In this paper we identify a need for Swedish knowledge base resources and propose extraction of facts from the first sentence of Wikipedia articles as a way of bootstrapping the creation of such a knowledge base. Our method relies on heuristics exploiting the wiki markup and style of writing in Wikipedia articles, and we show that utilizing a part-of-speech tagger to reformulate and filter out extracted facts is a way of improving the quality. We also conclude that there are limitations to our method, mainly due to errors and inconsistencies in the formulation of Wikipedia articles.

1 Introduction

The structuring of open information into a machine readable knowledge base has been a challenging task in AI research for years (Kasneci et al., 2009). Wu and Weld (2010) lists question answering, ontology learning, and summarization as NLP tasks that would benefit from such repositories. A starting point for the task of automatical extraction of knowledge is Wikipedia: an evolving, collaborative encyclopedia (Wu and Weld, 2010), covering information that the community has valued as knowledge.

Large scale attempts at creating English knowledge bases consisting of unary and binary relations from Wikipedia have been made by e.g. DBpedia and YAGO (Lehmann et al., 2013; Kasneci et al., 2009). These projects have already been used in numerous applications, such as Watson, the question answering system developed by IBM, which builds, inter alia, upon both systems (Ferrucci et al., 2010). DBpedia is also known to be used for e.g. named entity recognition and disambiguation, automatic tag generation and knowledge exploration (Lehmann et al., 2013).

Even though DBpedia and YAGO are mainly focused on the English Wikipedia it is possible to apply these frameworks to its editions in other languages. However, a lot of work has to be put into adapting the frameworks to the lexical and regional standards, as in the case of the Greek chapter of DBpedia (Kontokostas et al., 2012). An alternative to creating a knowledge base from Swedish resources would be to translate the existing English articles. However, as of writing there are 1 602 304 of articles in the Swedish version of Wikipedia, of which only 500 233 have a corresponding article in English. Thus, all information present in the Swedish Wikipedia would not be covered.

YAGO and DBpedia focus heavily on extracting relations from Wikipedia's internal structure, most prevalently by exploiting its categorization of articles, article redirections and predefined relations from infoboxes (Lehmann et al., 2013; Kasneci et al., 2009). While such an approach could be applied, at least as many as 58% of the Swedish articles, particularly the ones small enough not to have an Enlish counterpart, don't have infoboxes to extract knowledge from. An alternative approach is to exploit the encyclopedic style of writing that Wikipedia editors generally strive towards. In particular, the beginning of an article is of great interest since it tends to give an overview of the subject (Lange et al., 2010).

In this paper we describe a method to bootstrap the creation of a Swedish knowledge base by extracting *subject-predicate-object* triples from the first sentence of each article in the Swedish edition of Wikipedia. The method is based on exploitation of the encyclopedia's internal text markup (wiki markup) using an unsupervised heuristic model.

2 Method

Figure 1 shows the data flow in our method. Below, the steps are described in greater detail.

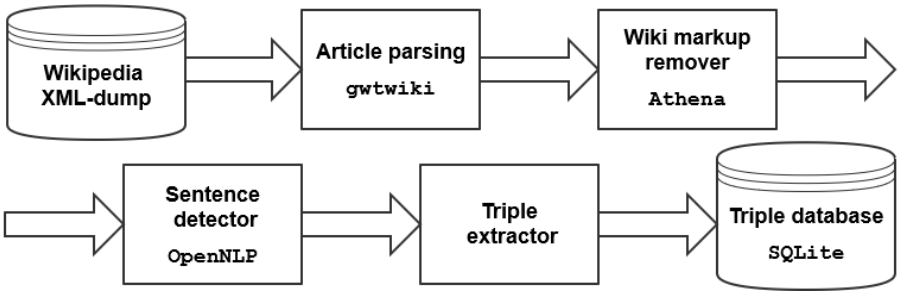

Figure 1: The data flow of our method.

2.1 Preprocessing

To process all of the articles in the Swedish Wikipedia, a compressed XML-dump of the site was downloaded. This file was processed using `gwtwiki`[1], a Java Wikipedia API. Wikipedia disambiguation and redirection pages were filtered away using regex patterns since they lie out of scope of the chosen method.

To remove Wikipedia's markup that generates e.g. infoboxes and article categorization, regex patterns developed for the Athena framework (Exner and Nugues, 2012) were used. Internal Wikipedia links (wiki links) were kept, as these contain information used later when extracting triples.

As only the first sentence of the article was needed, a sentence detector from the Apache OpenNLP library[2] was used, improved by some simple heuristics such as replacing problematic abbreviations with the full words.

2.2 Defining an ontology

The quality of an ontology is crucial for the usability of a knowledge base. An important task in this case is to be able to distinguish between different entities with the same name. Since Wikipedia article names uniquely identify the entities of the articles, they are used to define the ontology of the entity names in order to avoid the issue of named entity disambiguation.

When extracting subject-predicate-object triples from a sentence, the article name is taken to be the subject of a triple. This is combined with the wiki links as objects, as these refer to existing unique Wikipedia articles. Table 1 shows the subject, predicates and objects extracted from a sentence annotated with wiki markup.

[1] `http://code.google.com/p/gwtwiki/`.
[2] `https://opennlp.apache.org/`.

As for the ontology of the predicates, this is defined by the method of triple extraction.

Ystad (SUBJ) is the (PRED 1) [[administrative center]] (OBJ 1) in (PRED 2) [[Ystad municipality]] (OBJ 2) in (PRED 3) [[Skåne county]]. (OBJ 3)

SUBJ	PRED	OBJ
Ystad	is the	administrative center
Ystad	in	Ystad municipality
Ystad	in	Skåne county

Table 1: Example of the first sentence of a Wikipedia article with wiki markup annotation, translated from the Swedish wikipedia article for *Ystad*.

2.3 Naive extraction of triples

First, a naive extraction of triples was implemented. The subject and object of the triples were taken as described above. The predicates in the triples were simply chosen to be the words between the previously extracted object in the sentence (or the mention of the subject, in the case of the first triple), and the next wiki link. In this way, the ontology of the predicates was implicitly defined by the style of writing in the Swedish edition of Wikipedia, as seen in Table 1.

In addition to the described extraction of triples, some heuristics were developed to remove large amounts of bad facts, such as removing parentheses not containing any links and treating birth/death date and year as one object rather than two separate ones.

This naive method yielded about 3.7 million extracted facts for 1.5 million distinct entities, roughly 2.4 triples per subject. Table 2 shows the

most frequently occurring predicates.

2.4 Refinement of the extraction

To find ways to improve the naive method, an evaluation was carried out. Similar to Etzioni et al. (2011), we identify two different kinds of bad predicates: *incoherent extractions* where the facts don't have a meaning, and *uninformative extractions* where the facts make sense but exclude vital information.

To further evaluate the ontology of the predicates, extractions were classified as *obscure* if the triple was correct but the predicate was deemed rare, as e.g. the predicate "*earlier chief of information at*".

In order to improve the coherency and informativeness of the triples the following heuristics were developed.

If the predicate "*i*" (in) follows the predicate "*född*" (born) or "*död*" (dead), "*i*" is replaced with "*född i*" / "*död i*" (born/dead in), since the predicate "*i*" is uninformative in this context.

The predicate "*och*" (and) is replaced with the previous predicate, as in the sentence "*Strindberg (...) was a [[playwright]] and [[painter]]*". Here, the object "*painter*" will be connected to the predicate "*was a*".

Dates in the objects are converted into the ISO 8061 format supported by SQLite ("*YYYY-MM-DD*"), to simplify queries involving temporal constraints.

Reducing the obscurity of the predicates is done mostly with the help of a part-of-speech (POS) tagger for Swedish, Stagger (Östling, 2012). In order to make the ontology of the predicates more homogenous, words POS tagged as determiners, wh-pronouns and adverbs are removed in the predicates.

The Swedish words for "and", "*och*" and "*samt*", are removed if they occur as the first word in a predicate.

Lastly, predicates consisting of more than five words are removed since there is a high risk of them being obscure or incoherent.

The improved method yielded 3.5 million extracted triples for 1.5 million distinct entities. Table 2 shows the most frequently occurring predicates.

3 Evaluation

Evaluation of the extracted triples was carried out in a similar manner to Hoffart et al. (2013) for YAGO2. As no preexisting resource for evaluating the triples was at hand, human judgement was trusted to perform the evaluation. Because of this, facts as such were not evaluated but rather the correctness of the triples with respect to the article from which they were extracted. For each of the two methods, 100 random Wikipedia articles were considered. The extracted triples adhering to the article entity were examined with the first sentence presented next to the facts, and the results are shown in Table 3.

Category	**Naive**	**Improved**
Wiki links	271	268
Extracted triples	256	241
Correct	164 (64%)	171 (71%)
Obscure	35 (14%)	19 (8%)
Incoherent	67 (26%)	56 (23%)
Uninformative	25 (10%)	14 (6%)

Table 3: Results of the evaluation for the two methods.

The improved method results in roughly half the number of distinct predicates in comparison to the naive method and fewer predicates make up for a larger proportion of the triples, as seen in Table 4.

Proportion of triples	**Amount of distinct predicates**	
	Naive	Improved
70%	8	5
75%	13	6
80%	39	11
85%	300	41
90%	11041	534
95%	127641	19104
100%	311671	160495

Table 4: Proportion of total amount of triples and amount of distinct predicates for the two methods.

4 Related work

As mentioned in the introduction, YAGO and DBpedia are two related systems utilizing the structured information in Wikipedia to create knowledge bases. iPopulator (Lange et al., 2010) is a system for automatic generation and refining of

Wikipedia infoboxes through analysis of text in articles. For the English language, ReVerb (Etzioni et al., 2011) is a system for extraction of binary relationships from sentences.

Turning to Swedish NLP resources, Språkbanken[3] (The Swedish Language Bank) at the University of Gothenburg is a hub for various projects, especially corpora and lexicon.

5 Discussion

In this paper we chose to focus on utilizing wiki markup to extract triples, as this yields a consistent ontology. This method relies on a coherent use of wiki links and limits the knowledge base construction to what is expressed in the first sentence of an article. It may also be prone to spelling errors and bad triples due to obscure wiki links. E.g. if an Olympic athlete is of German nationality, the wiki link to the athlete's nationality may refer to the Wikipedia article "Germany at the Olympics". Here, one would rather want simply "Germany".

The rules using the POS tags are fairly simple since they don't consider more advanced syntactical constructions such as combinations of several POS tags. Still, these result in a significant reduction of incoherent and uninformative extractions, and the obscure predicates are significantly reduced as seen in Table 3. We can also see in Table 4 that the overall weight of the predicates is shifted to fewer predicates making up for the bigger part of the total amount.

The method of evaluation relies on human judgement, and is thus inherently fuzzy. It was problematic to distinguish between uninformative and incoherent facts, and between obscure and non-obscure facts, as these categories may be seen as overlapping in some sense.

6 Conclusion

We have pointed to the need of more Swedish knowledge base resources and proposed fact extraction from the first sentence of Wikipedia articles as one mean of bootstrapping the creation of such resources. We proposed a naive method for such extraction of facts which yielded 3.7 million triples for 1.5 million entities. This method was improved by both treatment of specific predicates and more general heuristics utilizing POS tagging. The improved method yielded 3.5 million triples for 1.5 million entities. The size of the ontology was halved and according to manual evaluation the amount of correct facts increased from 64% to 71%.

7 Future work

Further development of the method proposed in this paper may include expanding the use of the POS tags, considering more complex grammatical constructions. The use of a semantic parser such as MaltParser[4] would also be of help when removing e.g. sub-clauses.

Merging of the knowledge base with other Swedish resources such as the DBpedia knowledge base generated from Swedish Wikipedia and lexicons from Språkbanken would also be beneficial.

[3] http://spraakbanken.gu.se/

[4] http://www.maltparser.org/

Naive method		Improved method	
Predicate	**Occurr.**	**Predicate**	**Occurr.**
är en (is a)	1145373	*är* (is)	1271228
i (in)	460920	*som beskrevs av* (as described by)	613269
som beskrevs av (as described by)	438537	*i* (in)	330372
som först beskrevs av (as first described by)	147523	*född* (born)	105694
och (and)	108989	*född i* (born in)	102466
är ett (is a)	95726	*av* (of)	97508
av (of)	90201	*var* (was)	84570
född (born)	86425	*död* (dead)	59711
var en (was a)	63303	*död i* (dead in)	37955
död (dead)	47279	*i regionen* (in the region)	36913
i regionen (in the region)	36919	*från* (from)	20198
från (from)	17742	*var svensk* (was Swedish)	12493
född den (born on the)	15871	*som tillhör* (as belongs to)	12025
och ingår i Sjön har en area på kvadratkilometer och ligger (and is part of the lake has an area of square kilometers and is in)	15095	*en* (one)	11952
som tillhör (as belongs to)	12022	*på* (on)	11249
-	11361	*med* (with)	10052
som ingår i Sjön har en area på kvadratkilometer och ligger (as part of the lake has an area of square kilometers and is in)	10126	*i norra* (in the north of)	9910
på (on)	10021	*i familjen* (in the family)	9875
är (is)	9915	*är svensk* (is Swedish)	8589
i familjen (in the family)	9841	*inom ordningen* (in the order)	8036
i norra (in the north of)	9783	*i delstaten* (in the state)	6222
inom ordningen (in the order)	8034	*i nordöstra* (in the north east of)	5754
var en svensk (was a Swedish)	7971	*mellan* (between)	5690
med (with)	6875	*vid* (with)	5326
är en svensk (is a Swedish)	6766	*i södra* (in the south of)	5190
död den (dead on the)	6220	*för* (for)	4961
i delstaten (in the state)	6218	*den* (that)	4739
och ingår i Sjön är meter djup har en yta på kvadratkilometer och befinner sig (and is part of the lake is meter deep has an area of square kilometers and is located)	6146	*i östra* (in the east of)	4647
vid (at)	5713	*inom* (within)	4461
i nordöstra (in the north east of)	5707	*till* (to)	4420

Table 2: The 30 most frequently occurring predicates extracted by the two methods. English translation shown in parentheses.

References

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. 2011. Open information extraction: The second generation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume One*, IJCAI'11, pages 3–10.

Peter Exner and Pierre Nugues. 2012. Constructing large proposition databases. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, (LREC'12), pages 3836–3840. European Language Resources Association.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. 2010. Building Watson: An overview of the DeepQA project. *AI MAGAZINE*, 31(3):59–79.

Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61.

Gjergji Kasneci, Maya Ramanath, Fabian Suchanek, and Gerhard Weikum. 2009. The YAGO-NAGA approach to knowledge discovery. *SIGMOD Rec.*, 37(4):41–47, March.

Dimitris Kontokostas, Charalampos Bratsas, Sören Auer, Sebastian Hellmann, Ioannis Antoniou, and George Metakides. 2012. Internationalization of linked data: The case of the Greek DBpedia edition. *Web Semantics: Science, Services and Agents on the World Wide Web*, 15:51–61.

Dustin Lange, Christoph Böhm, and Felix Naumann. 2010. Extracting structured information from Wikipedia articles to populate infoboxes. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1661–1664, New York, NY, USA. ACM.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, et al. 2013. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*.

Robert Östling. 2012. Stagger: A modern POS tagger for swedish. In *Proceedings of the Swedish Language Technology Conference (SLTC)*.

Fei Wu and Daniel S. Weld. 2010. Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 118–127, Stroudsburg, PA, USA. ACL.