Using Distant Supervision to Build a Proposition Bank

Peter Exner, Marcus Klang, Pierre Nugues

Department of Computer Science, Lund University, Sweden
{peter.exner, marcus.klang, pierre.nugues}@cs.lth.se

Abstract

Semantic role labeling has become a key module of many language processing applications. To build an unrestricted semantic role labeler, the first step is to develop a comprehensive proposition bank. However, building such a bank is a costly enterprise, which has only been achieved for a handful of languages. In this paper, we describe a technique to build proposition banks for new languages using distant supervision. Starting from PropBank in English and loosely parallel corpora such as versions of Wikipedia in different languages, we carried out a mapping of semantic propositions we extracted from English to syntactic structures in Swedish using named entities. We could identify 2,333 predicate–argument frames in Swedish.

1. Introduction

Semantic role labeling has become a key module of many language processing applications and its importance is growing in fields like question answering (Shen and Lapata, 2007), information extraction (Christensen et al., 2010), sentiment analysis (Johansson and Moschitti, 2011), and machine translation (Liu and Gildea, 2010; Wu et al., 2011). To build an unrestricted semantic role labeler, the first step is to develop a comprehensive proposition bank. However, building proposition banks is a costly enterprise and as a consequence of that, they only exist for a handful of languages such as English, Chinese, German, or Spanish. In this paper, we describe a technique to build proposition banks for new languages using distant supervision.

Distant supervision is an alternative to unsupervised and supervised approaches that was introduced by Craven and Kumlien (1999). They used a knowledge base of existing biological relations, automatically identified sentences containing these relations, and could train a classifier to recognize the relations. Distant supervision has been successfully transferred to other fields. Mintz et al. (2009) describe a method for creating training data and relation classifiers without a hand-labeled corpus. The authors used Freebase and its binary relations between entities, such as (/location/location/contains, Belgium, Nijlen). They extracted entity pairs from the sentences of a text and matched them to those found in Freebase. Using the entity pairs, the relations, and the corresponding sentence text, they could train a relation extractor.

2. Distant Supervision to Extract Semantic Propositions

We designed a method to build a Swedish proposition bank using distant supervision. Starting from an existing proposition bank, ProbBank in English (Palmer et al., 2005), and loosely parallel corpora such as versions of Wikipedia in different languages, we carried out a mapping of the semantic propositions we extracted from English to syntactic structures in the target language. We parsed the English edition of Wikipedia up to the predicate–argument structures using a semantic role labeler (Björkelund et al., 2010) and the Swedish Wikipedia using a dependency parser (Nivre et al., 2006). We extracted all the named entities we found in the propositions and we disambiguated them using the Wikidata nomenclature. Using recurring entities, we aligned sentences in the two languages and we identified 2,333 predicate–argument frames in Swedish.

Similarly to Mintz et al. (2009), we used an external resource of relational facts and we matched the entity pairs in the relations to a Swedish text corpus. However, our approach substantially differs from theirs by the form of the external resource, which is a parsed corpus. To our best knowledge, there is no Swedish repository of relational facts between entities in existence. Instead, we semantically parsed an English corpus, in our case the English edition of Wikipedia, and we matched, article by article, the resulting semantic structures to sentences in the Swedish edition of Wikipedia.

We believe that by only using pairs of corresponding articles in different language editions and, hence, by restraining cross-article supervision using the unique identifiers given by Wikipedia, we can decrease the number of false negatives. We based this conviction on the observation that many Swedish Wikipedia articles are loosely translated from their corresponding English article and therefore express the same facts or relations.

3. Architecture

Our system consists of three parts:

- The first one parses the Swedish Wikipedia up to the syntactic layer and carries out a named entity identification.
- The second part carries out a semantic parsing of the English Wikipedia and applies a named entity identification.
- The third part aligns propositions having identical named entities in both languages using the Wikidata Q number.

To complete these tasks, we used a Hadoop-based architecture, Koshik (Exner and Nugues, 2014), that we ran on a cluster of 12 machines.

Given the sentences:

Cologne is located on both sides of the Rhine River

and

Köln ligger på båda sidorna av floden Rhen,

Figure 1 shows the parsing results in terms of predicate–argument structures for English, and functions for Swedish. We identify the named entities in the two languages, *Cologne* and *Rhine*, respectively, *Köln* and *Rhen*, link them to their Wikidata identifiers, http://www.wikidata.org/wiki/Q365 and http://www.wikidata.org/wiki/Q584, and finally align the predicates and arguments. We obtain the complete argument spans by projecting the yield from the argument token. If the argument token is dominated by a preposition, the preposition token is used as the root token for the projection.

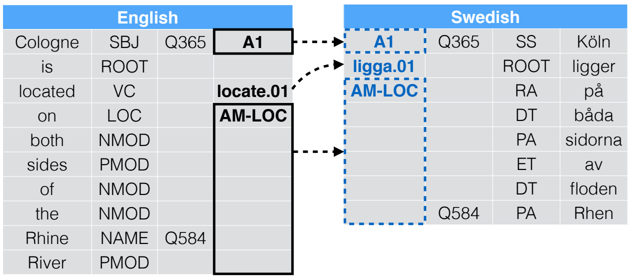

Figure 1: Outline of the distant supervision process.

4. Named Entity Disambiguation

Named entity disambiguation (NED) is the core step to anchor the parallel sentences and propositions with distantly supervised techniques. NED usually consists of two steps: extract the entity mentions, usually noun phrases, and if a mention corresponds to a proper noun – a named entity –, link it to a unique identifier.

For the English part, we used Wikifier (Ratinov et al., 2011). There was no similar disambiguator for Swedish and we implemented one: NEDforia. In addition, as most disambiguators are designed for English and require resources that do not exist for Swedish, we created a specific algorithm.

NEDforia starts from a Wikipedia dump and automatically collects a list of named entities from the corpus. It then extracts the links and contexts of these entities to build disambiguation models. Given an input text, NEDforia recognizes and disambiguates the named entities, and annotates them with their corresponding Wikidata number.

5. Results and Future Work

By aligning 17,115 English sentences with 16,636 Swedish sentences, we managed to generate 19,121 propositions from which we extracted 2,333 Swedish predicate–argument frames[1]. Tables 1 and 2 show, respectively, an overview of the extraction statistics and the predicate names of the ten most frequent Swedish frames.

[1]These predicate–argument frames are available at http://semantica.cs.lth.se

Property	Count
English articles	4,152,283
Swedish articles	2,792,089
Supervising sentences (English)	17,115
Supervised sentences (Swedish)	16,636
Number of supervisions	19,121
Generated frames	2,333

Table 1: An overview of extraction statistics.

Swedish predicate	English predicate	Count
vinna.01	win.01	125
följa.01	follow.01	107
bli.01	become.01	93
spela.01	play.01	67
ligga.01	locate.01	55
flytta.01	move.01	55
förekomma.01	find.01	41
föda.02	bear.02	41
använda.01	use.01	39
släppa.01	release.01	37

Table 2: The ten most frequent Swedish frames.

We aligned the sentences using entities and frequency counts to select the most likely frames. While this relatively simple approach could be considered inadequate for other distant supervision applications, such as relation extraction, it worked surprisingly well in our case. We believe this can be attributed to the named entity disambiguation, which goes beyond a simple surface form comparison and uniquely identifies the entities used in the supervision. Similarly, we go beyond distant supervision that uses infobox relations, and instead form new predicates with different senses. Using infobox relations would have limited us to relations already described by the infobox ontology.

Since our technique builds on repositories of entities extracted from Wikipedia, such as DBpedia (Bizer et al., 2009) and YAGO2, one future improvement could be to exploit the semantic information residing in these repositories. Another possible improvement would be to apply a coreference solver to anaphoric mentions to increase the number of sentences that could be aligned.

Acknowledgements

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

References

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia—a crystallization point for the web of data. *Journal of Web Semantics*, pages 154–165.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstration Volume*, pages 33–36, Beijing, August 23-27. Coling 2010 Organizing Committee.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2010. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, FAM-LbR '10, pages 52–60.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB'99)*, pages 77–86.

Peter Exner and Pierre Nugues. 2014. KOSHIK: A large-scale distributed computing framework for NLP. In *Proceedings of ICPRAM 2014 – The 3rd International Conference on Pattern Recognition Applications and Methods*, pages 464–470, Angers, March 6-8.

Richard Johansson and Alessandro Moschitti. 2011. Extracting opinion expressions and their polarities: exploration of pipelines and joint models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers. Volume 2*, pages 101–106.

Ding Liu and Daniel Gildea. 2010. Semantic role features for machine translation. In *COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics*, pages 716–724, Beijing, June.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 1003–1011.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, pages 2216–2219, Genoa.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: an annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the ACL-HLT 2011 – Volume 1*, pages 1375–1384.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 12–21, Prague, June.

Dekai Wu, Pascale Fung, Marine Carpuat, Chi kiu Lo, Yongsheng Yang, and Zhaojun Wu. 2011. Lexical semantics for statistical machine translation. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, pages 236–252. Springer, Heidelberg.