A Distant Supervision Approach to Semantic Role Labeling

Exner, Peter; Klang, Marcus; Nugues, Pierre

Published in:
Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)

2015

Document Version:
Publisher's PDF, also known as Version of record

A Distant Supervision Approach to Semantic Role Labeling

Peter Exner **Marcus Klang** **Pierre Nugues**

Lund University
Department of Computer Science
Lund, Sweden
{Peter.Exner, Marcus.Klang, Pierre.Nugues}@cs.lth.se

Abstract

Semantic role labeling has become a key module for many language processing applications such as question answering, information extraction, sentiment analysis, and machine translation. To build an unrestricted semantic role labeler, the first step is to develop a comprehensive proposition bank. However, creating 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 trained a semantic parser on the generated Swedish propositions and we report the results we obtained. Using the CoNLL 2009 evaluation script, we could reach the scores of 52.25 for labeled propositions and 62.44 for the unlabeled ones. We believe our approach can be applied to train semantic role labelers for other resource-scarce languages.

1 Introduction

Semantic role labeling has become a key module for 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 create proposition banks for new languages using distant supervision. Our approach builds on the transfer of semantic information through named entities. Starting from an existing proposition bank, PropBank 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., 2010a) 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[1]. Using recurring entities, we aligned sentences in the two languages; we transferred the semantic annotation from English sentences to Swedish sentences; and we could identify 2,333 predicate–argument frames in Swedish.

Finally, we used the resulting corpus to train a semantic role labeler for Swedish that enabled us to evaluate the validity of our approach. Beyond Swedish, we believe it can apply to any resource-

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

*Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 239–248,
Denver, Colorado, June 4–5, 2015.

scarce language.

2 Previous Work

The techniques we applied in this paper are similar to those used in the extraction of relations between entity mentions in a sentence, where relational facts are often expressed in the form of triples, such as: (Seoul, CapitalOf, South Korea). While supervised and unsupervised techniques have been applied to the extraction of such relations, they both suffer from drawbacks. Supervised learning relies on labor-intensive, hand-annotated corpora, while unsupervised approaches have lower precision and recall levels.

Distant supervision is an alternative to these 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 trained 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.

Padó and Lapata (2009) used parallel corpora and constituent-based models to automatically project FrameNet annotations from English to German. Hoffmann et al. (2010) introduced Wikipedia infoboxes in relation extraction, where the authors trained a classifier to predict the infobox schema of an article prior to the extraction step. They used relation-specific lexicons created from a web crawl to train individual extractors for 5,025 relations and, rather than running all these extractors on every article and sentence, they first predicted the schema of an article and then executed the set of corresponding extractors. Early work in distant supervision assumed that an entity pair expresses a unique explicit relation type. Surdeanu et al. (2012) describe an extended model, where each entity pair may link multiple instances to multiple relations. Ritter et al. (2013) used a latent-variable approach to model information gaps present in either the knowledge base or the corresponding text.

As far as we know, all the work on relation extraction focused on the detection of specific semantic relations between entities. In this paper, we describe an extension and a generalization of it that potentially covers all the relations tied to a predicate and results in the systematic extraction of the semantic propositions observed in a corpus.

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. Using the generated Swedish semantic structures, we could train a semantic role labeler.

3 Extending Semantic Role Labeling

In our approach, we employ distantly supervised techniques by combining semantic role labeling (SRL) with entity linking. SRL goes beyond the extraction of n-ary relations and captures a semantic meaning of relations in the form of predicates–argument structures. Since SRL extracts relations between a predicate and its arguments, it can be considered as a form of relation extraction which involves a deeper analysis.

However, the semantic units produced by classical semantic role labeling are still shallow, as they do not resolve coreference or disambiguate named entities. In this work, we selected the propositions, where the arguments corresponded to named entities and we resolved these entities in unique identifiers. This results in a limited set of extended propositions that we think are closer to the spirit of logical forms and can apply in a cross-lingual setting.

Input	Explanation	# cand.	Output
helsingborg c	*Railway station in Helsingborg*	1	wikidata:Q3062731
kärna	*Medieval tower in Helsingborg*	27	wikidata:Q1779457
berga	*District in Helsingborg*	33	wikidata:Q25411

Table 1: Entries in the detection dictionary, all related to the city of Helsingborg in Sweden, with their unique Wikidata Q-number and a short explanation in italics.

4 Named Entity Linking

Named entity linking (or disambiguation) (NED) is the core step of distant supervision to anchor the parallel sentences and propositions. NED usually consists of two steps: first, 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) to disambiguate entities. There was no similar disambiguator for Swedish and those described for English are not directly adaptable because they require resources that do not exist for this language. We created a disambiguator targeted to Swedish: NEDforia. NEDforia uses a Wikipedia dump as input 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.

4.1 Entity Detection

We created a dictionary of entities from Wikipedia using the combination of a POS tagger (Östling, 2013), language-dependent uppercase rules, and two entity databases: Freebase (Bollacker et al., 2008) and YAGO2 (Hoffart et al., 2010). Table 1 shows three dictionary entries, where an entry consists of a normalized form and the output is a list of Wikidata candidates in the form of Q-numbers. The output can be the native Wikipedia page, if a Wikidata mapping could not be found, as for "wikipedia.sv:Processorkärna" ("wikipedia.en:Multi-core_processor" in the English Wikipedia).

The entity detection module identifies the strings in the corpus representing named entities. It tokenizes the text and uses the longest match to find the sequences of tokens that can be associated to a list of entity candidates in the dictionary.

4.2 Disambiguation

We disambiguated the entities in a list of candidates using a binary classifier. We trained this classifier with a set of resolved links that we retrieved from the Swedish Wikipedia articles. As in Bunescu and Paşca (2006), we extracted all the manually created mention–entity pairs, encoded as `[[target|label]]` in the Wikipedia markup, and we marked them as positive instances. We created the negative instances with the other mention–candidate pairs that we generated with our dictionary.

As classifier, we used the L2-regularized logistic regression (dual) from LIBLINEAR (Fan et al., 2008) with three features and we ranked the candidates according to the classifier output. The features are the popularity, commonness (Milne and Witten, 2008), and context. The popularity is the probability that a candidate is linked to an entity. We estimate it through the count of unique inbound links to the candidate article (Table 2). The commonness is the probability the sequence of tokens could be the candidate: $P(\text{candidate}|\text{sequence of tokens})$. We compute it from the target–label pairs (Table 3). The context is the count of unique words extracted from the two sentences before the input string that we intersect with the words found in the candidate's article.

Entity	Occupation	Popularity
Göran Persson	Skåne politician	4
Göran Persson	Musician	5
Göran Persson	Prime minister	257

Table 2: The popularity of some entities.

Entity	Mention	Common.
Scandinavian Airlines	SAS	90.4%
Special Air Service	SAS	5.4%
SAS System	SAS	0.4%
Cable News Network	CNN	99.2%
Cable News Network Int.	CNN	0.8%

Table 3: The commonness of some entities.

5 Distant Supervision to Extract Semantic Propositions

The distant supervision module consists of three parts:

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

5.1 Semantic and Syntactic Parsing

As first step, we parsed the English edition of Wikipedia up to the predicate–argument structures using the Mate-Tools dependency parser and semantic role labeler (Björkelund et al., 2010a) and the Swedish Wikipedia using MaltParser (Nivre et al., 2006). To carry out these parsing tasks, we used a Hadoop-based architecture, Koshik (Exner and Nugues, 2014), that we ran on a cluster of 12 machines.

5.2 Named Entity Disambiguation

The named entity disambiguation links strings to unique Wikidata and is instrumental to the proposition alignment. For the two English-Swedish equivalent sentences:

> Cologne is located on both sides of the Rhine River

and

> Köln ligger på båda sidorna av floden Rhen,

Wikifier, on the English version, identifies *Cologne* and *Rhine river* as named entities and links them respectively to the `en.wikipedia.org/wiki/Cologne` and `en.wikipedia.org/wiki/Rhine` pages, while NEDforia, on the Swedish text, produces a ranked list of entity candidates for the words *Köln* and *Rhen* shown in Table 4. We assign the named entities to the top candidates, Q365 for *Köln* 'Cologne' and Q584 for *Rhen* 'Rhine.' We import the resulting annotated Wikipedia into Koshik, where we map the document titles and anchor targets to Q-numbers.

Words	Entities	English pages
Köln	Q365	Cologne
	Q54096	University of Cologne
	Q104770	1. FC Köln
	Q7927	Cologne (region)
	Q157741	Cologne Bonn Airport
	...	...
Rhen	Q584	Rhine
	Q10650601	No English page

Table 4: The ranked entity candidates matching the words *Köln* 'Cologne' and *Rhen* 'Rhine.' The entities are identified by their Wikidata Q-numbers.

5.3 Alignment of Parallel Sentences

We ran the alignment of loosely parallel sentences using MapReduce (Dean and Ghemawat, 2008) jobs. Both the English and Swedish articles are sequentially read by mappers. For each sentence, the mappers build and emit key-value pairs. The mappers create keys from the entity Q-numbers in each sentence and we use the sentences as values.

The shuffle-and-sort mechanism in Hadoop ensures that, for a given key, each reducer receives all the sentences. In this process, the sentences are aligned by their Q-numbers and given as a group to the reducers with each call. The reducers process each group of aligned sentences and annotate the Swedish sentence by linking the entities by their Q-numbers and by inferring the semantic roles from the aligned English sentences. The annotated Swedish sentences are then emitted from the reducers. For each newly formed Swedish predicate, we

select the most frequent alignments to form the final Swedish predicate–argument frames. Figure 1 shows this alignment process.

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.

5.4 Semantic Annotation Transfer

Figure 2 shows the parsing results for the sentences *Cologne is located on both sides of the Rhine River* and *Köln ligger på båda sidorna av floden Rhen* in terms of predicate–argument structures for English, and functions for Swedish. We identify the named entities in the two languages and we 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.

5.4.1 Forming Swedish Predicates

During the alignment of English and Swedish sentences, we collect token-level mappings between sentences. The mappings keep a record of how many times an English predicate is aligned with a Swedish verb. For each Swedish verb, we then select the most frequent English predicate it is aligned with. We create a new Swedish frame by using the lemmatized form of the verb and attaching the sense of the English predicate. We use the sentences representing the most frequent mappings to generate our final corpus of Swedish propositions. Table 6 shows how two Swedish frames, *vinna.01* and *vinna.03*, are created by selecting the most frequent mappings. Table 7 shows the ten most frequent Swedish frames created using this process.

6 A Swedish Corpus of Propositions

We processed more than 4 million English Wikipedia articles and almost 3 million Swedish Wikipedia pages from which we could align over 17,000 English sentences with over 16,000 Swedish

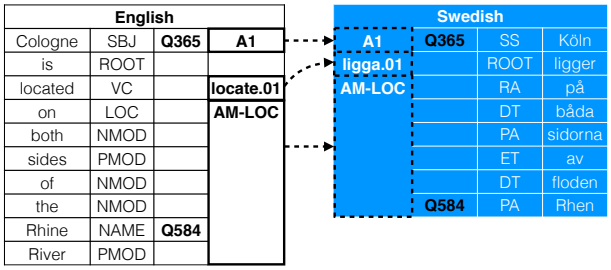

Figure 2: Transfer of the predicate–argument structure from an English sentence to a corresponding Swedish sentence. The sentences are aligned by the two entities that they both share: Cologne (**Q365**) and the Rhine River (**Q584**). The argument roles are transferred using the Q-number entity links. The Swedish predicate is formed using the lemma of the verb token, that is the syntactical parent of the arguments.

Type	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

Table 5: An overview of distant supervision statistics.

sentences. This resulted into 19,000 supervisions and the generation of a corpus of Swedish propositions. Table 5 shows an overview of the statistics of this distant supervision process.

The generated corpus consists of over 4,000 sentences, a subset of the 16,000 Swedish sentences used in the supervision process. These 4,000 sentences participate in the most frequent English to Swedish mappings, as detailed in Sect 5.4.1. Table 8 shows an overview of the corpus statistics.

Table 7 shows the ten most frequent mappings and we can see that all of them form meaningful Swedish frames. We can with caution state that our method of selecting the most frequent mapping works surprisingly well. However, if we examine Table 6, we observe some drawbacks to this approach. Although some unlikely mappings, such as pay.01 are filtered out, defeat.01 and prevail.01 could be used to form new Swedish predicates with different senses of the verb *vinna* 'win'. In addition, the predicates, help.01, take.01, and scoring.01, might participate as auxiliary verbs or otherwise form propositions

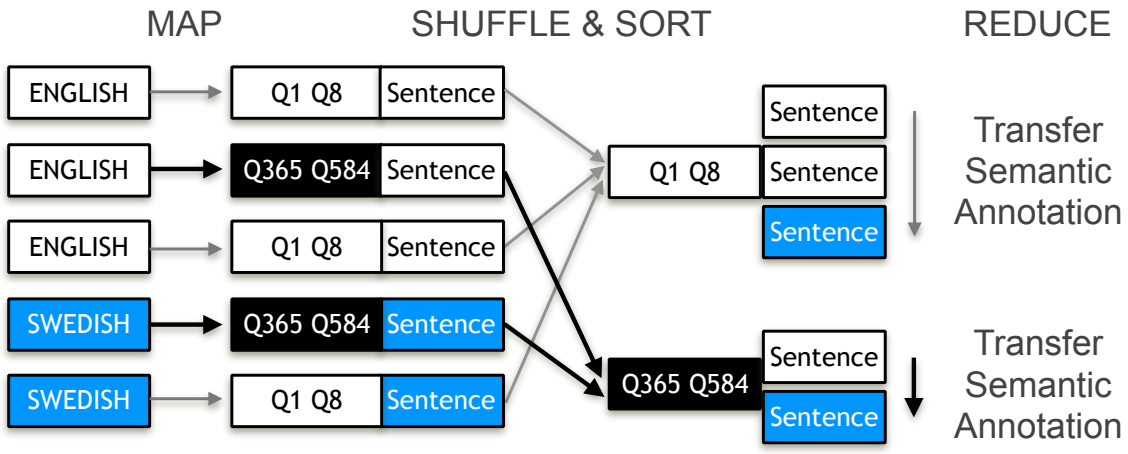

Figure 1: Automatic parallel alignment of sentences through MapReduce. The *Map* phase creates a key-value pair consisting of list of entities and a sentence. The *Shuffle & Sort* mechanism groups the key-value pairs by the list of entities, effectively aligning sentences across the languages. The *Reduce* phase steps through the list of aligned sentences and transfers semantic annotation from a language to another. Figure 2 shows this latter process.

English predicate	Count	Swedish predicate
win.01	**125**	**vinna.01**
defeat.01	24	–
beat.03	**10**	**vinna.03**
help.01	4	–
take.01	4	–
scoring.01	2	–
pay.01	1	–
prevail.01	1	–

Table 6: Selecting the most frequent English to Swedish mapping to form new Swedish predicates for the verb *vinna* 'win'. A bold typeface indicates a newly formed Swedish predicate. A dash indicates that a Swedish predicate for the verb *vinna* was not formed using the corresponding English predicate.

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

Table 7: The ten most frequent Swedish frames and their mappings from English predicates.

having the same meaning as win.01. A more thorough investigation of the roles played by the entities, possibly in combination with the use of additional semantic information from Wikidata, would certainly aid in improving the extraction of Swedish predicates.

7 Semantic Role Labeling

To assess the usefulness of the proposition corpus, we trained a semantic role labeler on it and we compared its performance with that of a baseline parser. Some roles are frequently associated with grammatical functions, such as A0 and the subject in PropBank. We created the baseline using such association rules and we measured the gains brought by the corpus and a statistical training.

We split the generated corpus into a training, development, and test sets with a 60/20/20 ratio. We used the training and development sets for selecting features during training and we carried out a final evaluation on the test set.

7.1 Baseline Parser

The baseline parser creates a Swedish predicate from the lemma of each verbal token and assigns it the sense 01. Any token governed by the verbal to-

ken having a syntactic dependency function is identified as an argument. The Talbanken corpus (Teleman, 1974) serves as training set for the Swedish model of MaltParser. We used four of its grammatical functions: subject (SS), object (OO), temporal adjunct (TA), and location adjunct (RA) to create the roles A0, A1, AM-TMP (temporal), and AM-LOC (locative), respectively.

7.2 Training a Semantic Role Labeler

The SRL pipeline, modified from Björkelund et al. (2010b), consists of four steps: Predicate identification, predicate disambiguation, argument identification, and argument classification.

During predicate identification, a classifier determines if a verb is a predicate and identifies their possible sense. Predicates may have different senses together with a different set of arguments. As an example, the predicate open.01 describes *opening something*, for example, *opening a company branch or a bottle*. This differs from the predicate sense, open.02, having the meaning of *something beginning in a certain state*, such as *a stock opening at a certain price*.

The argument identification and classification steps identify the arguments corresponding to a predicate and label them with their roles.

7.3 Feature Selection

We considered a large number of features and we evaluated them both as single features and in pairs to model interactions. We used the same set as Johansson and Nugues (2008) and Björkelund et al. (2009), who provide a description of them. We used a greedy forward selection and greedy backward elimination procedure to select the features (Björkelund et al., 2010a). We ran the selection process in multiple iterations, until we reached a stable F1 score. Table 10 shows the list of single features we found for the different steps of semantic role labeling: Predicate identification, predicate disambiguation, argument identification, and argument classification.

Interestingly, the amount of features used in argument identification and classification, by far exceeds those used for predicate identification and disambiguation. This hints that, although our generated corpus only considers entities for argument roles, the diverse nature of entities creates a corpus in which arguments hold a wide variety of syntactical and lexical roles.

Property	Unfiltered Count	Filtered Count
Generated frames	2,333	457
Number of propositions	4,369	2,663
Number of sentences	4,152	2,562
Number of tokens	77,015	43,617

Table 8: An overview of corpus statistics.

7.4 The Effect of Singleton Predicate Filtering

We performed a secondary analysis of our generated corpus and we observed that a large number of predicates occurs in only one single sentence. In addition, these predicates were often the result of errors that had propagated through the parsing pipeline.

We filtered out the sentences having mentions of singleton predicates and we built a second corpus to determine what kind of influence it had on the quality of the semantic model. Table 8, right column, shows the statistics of this second corpus. Singleton predicates account for a large part of the corpus and removing them shrinks the number of sentences by almost a half and dramatically reduces the overall number of predicates.

7.5 Validation on the Test Set

Table 9 shows the final evaluation of the baseline parser and the semantic role labeler trained on the generated corpus using distant supervision. The baseline parser reached a labeled F1 score of 22.38%. Clearly, the indiscriminating choice of predicates made by the baseline parser gives a higher recall but a poor precision. The semantic role labeler, trained on our generated corpus, outperforms the baseline parser by a large margin with a labeled F1 score of 39.88%. Filtering the corpus for singleton mention predicates has a dramatic effect on the parsing quality, increasing the labeled F1 score to 52.25%. We especially note a F1 score of 62.44% in unlabeled proposition identification showing the validity of the approach.

Method	Labeled			Unlabeled		
	Precision	Recall	F1	Precision	Recall	F1
Baseline	15.74	38.73	22.38	25.10	**61.78**	35.70
Distant supervision (Unfiltered corpus)	46.99	34.65	39.88	67.06	49.45	56.92
Distant supervision (Filtered corpus)	**58.23**	**47.38**	**52.25**	**69.59**	56.62	**62.44**

Table 9: Summary of semantic role labeling results. The table shows precision, recall, and F1 scores for our baseline and distant supervision methods. Evaluation performed on test set.

Feature	PI	PD	AI	AC
ArgDeprel			•	•
ArgPOS			•	•
ArgWord			•	•
ChildDepSet		•		•
ChildPOSSet		•		•
ChildWordSet		•	•	
DepSubCat	•	•		
DeprelPath			•	•
LeftPOS				•
LeftSiblingPOS				•
LeftSiblingWord				•
LeftWord				•
POSPath			•	•
Position	•		•	•
PredLemma	•		•	•
PredLemmaSense				•
PredPOS		•	•	•
PredParentPOS	•	•	•	•
PredParentWord	•	•		•
PredWord	•			•
RightPOS			•	•
RightSiblingWord			•	
RightWord			•	•

Table 10: List of features used in the four stages of semantic role labeling. PI stands for predicate identification, PD for predicate disambiguations, AI for argument identification, and AC for argument classification.

8 Conclusion

By aligning English and Swedish sentences from two language editions of Wikipedia, we have shown how semantic annotation can be transferred to generate a corpus of Swedish propositions. We trained a semantic role labeler on the generated corpus and showed promising results in proposition identification.

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. In addition, we believe that the implicit entity types that a set of named entities infer, constrain a sentence to a certain predicate and sense. This increases the likelihood that the Swedish aligned sentence contains a predicate which preserves the same semantics as the English verb of the source sentence. Furthermore, we go beyond infobox relations as we infer 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 a repository of entities extracted from Wikipedia, one future improvement could be to exploit the semantic information residing in it, possible from other repositories such as DBpedia (Bizer et al., 2009) or YAGO2. Another possible improvement would be to increase the size of the generated corpus. We envision this being done either by applying a coreference solver to anaphoric mentions to increase the number of sentences that could be aligned or by synthetically generating sentences through the use of a semantic repository. An additional avenue of exploration lies in extending our work to other languages.

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, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of The Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 43–48, Boulder, June 4-5.

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

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010b. A high-performance syntactic and semantic dependency parser. In *COLING (Demos)*, pages 33–36. Demonstrations Volume.

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.

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, Trento, April. Association for Computational Linguistics.

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.

Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

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.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. 2010. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. Research Report MPI-I-2010-5-007, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, November.

Raphael Hoffmann, Congle Zhang, and Daniel S Weld. 2010. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 286–295.

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.

Richard Johansson and Pierre Nugues. 2008. Dependency-based syntactic–semantic analysis with PropBank and NomBank. In *Proceedings of CoNLL-2008: The Twelfth Conference on Computational Natural Language Learning*, pages 183–187, Manchester, August 16-17.

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.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518.

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-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.

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

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

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 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1375–1384.

Alan Ritter, Luke Zettlemoyer, Oren Etzioni, et al. 2013. Modeling missing data in distant supervision for information extraction. *Transactions of the Association for Computational Linguistics*, 1:367–378.

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.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465.

Ulf Teleman. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska*. Studentlitteratur, Lund.

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.