Generating a Swedish Semantic Role Labeler

Peter Exner, Marcus Klang, Pierre Nugues

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

1. Introduction

Semantic role labeling is a form of shallow extraction method with an increasingly important role in complex tasks as information extraction (Christensen et al., 2010), question answering (Shen and Lapata, 2007), and text summarization (Khan et al., 2015). The development of resources such as FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005), and Swedish FrameNet (Borin et al., 2010) have enabled supervised approaches for training semantic role labelers (SRL) (Surdeanu et al., 2008; Johansson et al., 2012). However, the manual effort behind their development is considerable and, additionally, they face challenges regarding domain and language dependence.

As an alternative to supervised techniques, annotation projection transfers linguistic knowledge between multilingual corpora. Yarowsky et al. (2001) introduced the concept by transferring part-of-speech tags across parallel corpora. Padó and Lapata (2009) and Basili et al. (2009) performed cross-lingual frame-semantic annotation projection. In Exner et al. (2015), we introduced the concept of using entities to align loosely parallel corpora. We extend this approach by: (1) including linguistic units exhibiting entity-like characteristics in a local setting, (2) showing the effectiveness of our approach in a practical setting by evaluation on the Swedish FrameNet corpus, and (3) releasing the source code used and thus providing the generated Swedish PropBank to the scientific community[1].

2. Method

Starting with loosely parallel corpora, the entire English and Swedish editions of Wikipedia, our method uses entities to align sentences expressing the same semantic content. Our goal is to transfer semantic annotation, as described by PropBank, from English to Swedish, generate a Swedish PropBank, and use it to train a SRL. The sentence alignment process consists of the following steps:

1. We extract all the mentions of named entities, e.g. people, places, and organizations, from all the sentences, and we assign them a unique and language-independent identifier provided by Wikidata (Vrandečić and Krötzsch, 2014).

2. We annotate English sentences to a semantic level and Swedish sentences to a syntactic level.

3. For each sentence, we create sets of entities, including entity-like linguistic units and pronouns generalized by case, gender, and number. Entity-like units correspond to sequences of tokens that are only uniquely identifiable in the scope of a sentence pair.

4. Finally, using the unique identifier of each entity, we align and form English-Swedish sentence pairs.

Our baseline method extracts all the entities as a set from a sentence. In addition, we use a projection method, that extracts sets of entities projected either by arguments in English sentences or by a verb in Swedish sentences. From each English-Swedish sentence pair, we extract alignments between English predicates and Swedish verbs, and we record the most frequent alignments.
We then perform the annotation projection using the following steps:

1. For each English-Swedish sentence pair, we transfer the semantic annotation from a predicate to a verb using our record of the most frequent predicate→verb alignments.

2. We transfer argument roles by using the aligned entities between the sentence pair.

3. We assign the argument role to the governing token in the token span covered by each entity. If the argument token is dominated by a preposition, we search for a corresponding preposition in the Swedish sentence and assign it the argument role.

3. Results and Evaluation

We evaluated our method, seeking the answers to how different parameters and approaches affect the results. We also assessed the performance we can expect from the generated Swedish PropBank in a practical setting.

3.1 Predicate-Verb Alignment

In this section, we evaluate how our predicate→verb alignment method performs. We observe the quality of alignments under different settings including the method used, the frequency of alignments, and by varying the number of entities. We group each predicate→verb alignment into three frequency bands, high, medium, and low, and we evaluate by taking a random sample from each band, in total 100 alignments.

[1] http://semantica.cs.lth.se

In Figure 1, we show the precision and number of alignments using our baseline and projected methods. We observe that the precision grows with the number of entities used at a cost of decreasing the number of alignments created. Using three projected entities, our method produces 1,000 alignments with a precision of 80%.

In Figure 2, we use our projected method and show the breakdown of precision curves by three frequency bands. We observe that high to medium occurring alignments produce high precision, when using three projected entities. We believe these results show our method of selecting the most frequent alignments can generate valid alignments, with a precision that scales with the size of the corpora used. In the following evaluations, we use the optimal setting of the projected entities.

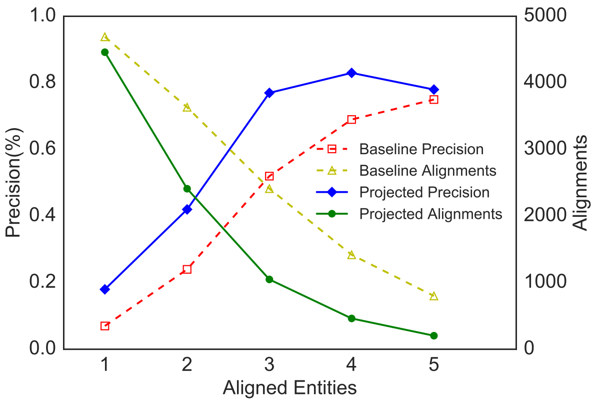

Figure 1: Graph of predicate→verb alignment precision and count under different parameter settings.

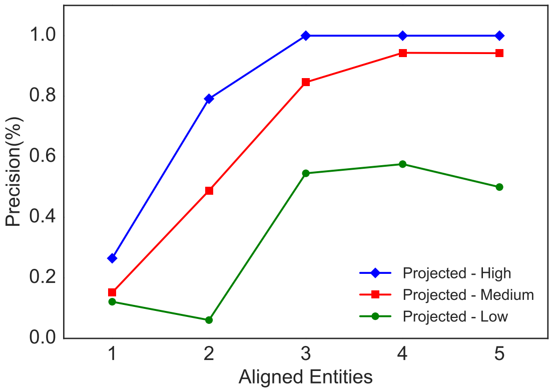

Figure 2: Graph of projected predicate→verb alignment precision, breakdown by three frequency bands: high, medium, and low.

3.2 Experimental Results

We now carry out an evaluation of our generated Swedish PropBank by training an SRL and evaluating it on a random sample of the Swedish FrameNet corpus. To enable this evaluation, we converted the sampled sentences from frame semantics to the semantics used in PropBank. The characteristics of our generated Swedish PropBank and the sampled Swedish FrameNet used in our evaluations are shown in Table 1.

We first train an off-the-shelf SRL (Björkelund et al., 2010), splitting our generated corpora into 60:20:20 training, development, and testing sets, and running a feature selection process using a greedy forward selection and greedy backward elimination (Johansson and Nugues, 2008; Björkelund et al., 2009). Using our trained SRL models, we then automatically parse the sampled Swedish FrameNet. Table 2 shows the performance breakdown of the semantic role labeler under different settings. We observe that our method ranks favorably with the approach described in Padó and Lapata (2009). We especially note that by using entity-like linguistic units, we observe an improvement of the labeled F1-score by 10%.

4. Conclusion and Future Work

In this paper, we have described a method for generating a Swedish PropBank by aligning the Swedish and English editions of Wikipedia using entities and entity-like linguistic units. We have shown two alignment methods that produce predicate→verb alignments of high precision while scaling with the input corpora. In addition, we have evaluated our generated PropBank in a practical setting by training a SRL and automatically parsing and evaluating on a sample of the Swedish FrameNet corpus. Our results show that our method performs favorably in comparison to previous approaches on parallel corpora, and promises an alternative way to creating training data from a growing resource of loosely parallel corpora. In the future, one research direction could investigate the alignment method by including entity taxonomy from an external ontology to reduce the specificity of entities. In addition, we believe our approach could be applied to generate PropBanks from similar resources, e.g. news articles describing the same events.

Acknowledgements

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

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Roberto Basili, Diego De Cao, Danilo Croce, Bonaventura Coppola, and Alessandro Moschitti. 2009. Cross-language frame semantics transfer in bilingual corpora. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 332–345. Springer.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 43–48. Association for Computational Linguistics.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and

DATASET	TOKENS	SENTENCES	PREDICATES	ARGUMENTS
Generated-Swedish	198,008	13,767	14,552	32,659
SweFN++ (TEST)	1,258	101	101	265

Table 1: Characteristics of the generated Swedish PropBank used for training SRL models and the SweFN++ FrameNet used for evaluating the trained model

	LABELED			UNLABELED		
LINGUISTIC UNITS	P	R	F1	P	R	F1
Entities (Baseline)	79.88	36.89	50.47	93.49	43.17	59.07
Entities + Unique Tokens	84.82	44.26	58.17	92.67	48.36	63.55
Entities + Unique Tokens + Pronouns	72.18	52.46	**60.76**	81.58	59.29	**68.67**

Table 2: Evaluation of semantic role labeling on the SweFN++ FrameNet.

semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 33–36. Association for Computational Linguistics.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2010. The past meets the present in swedish framenet+. In *14th EURALEX international congress*.

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.

Peter Exner, Marcus Klang, and Pierre Nugues. 2015. A distant supervision approach to semantic role labeling. In *Fourth Joint Conference on Lexical and Computational Semantics (* SEM 2015)*.

Richard Johansson and Pierre Nugues. 2008. Dependency-based syntactic-semantic analysis with propbank and nombank. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 183–187. Association for Computational Linguistics.

Richard Johansson, Karin Friberg Heppin, and Dimitrios Kokkinakis. 2012. Semantic role labeling with the swedish framenet. In *LREC*, pages 3697–3700.

Atif Khan, Naomie Salim, and Yogan Jaya Kumar. 2015. A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, 30:737–747.

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):71–106.

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, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.