Cross-lingual Bootstrapping for Semantic Role Labeling

Mikhail Kozhevnikov
University of Saarland
66123 Saarbrücken, Germany
mkozhevn@mmci.uni-saarland.de

Ivan Titov
University of Saarland
66123 Saarbrücken, Germany
titov@mmci.uni-saarland.de

Abstract

The approach we present uses semantic similarity between parallel sentences to bootstrap semantic role labeling (SRL) models for a pair of languages. The setting is similar to co-training, except for the intermediate model required to convert the SRL structure between the two annotation schemes used for different languages. This approach can facilitate the construction of SRL models for a resource-poor language, while preserving the annotation schemes designed for it and leveraging the resources available for this language. It can also be extended to benefit from the use of the resources in multiple languages simultaneously. We evaluate the model on four language pairs, English vs German, Spanish, Czech and Chinese, against a supervised baseline, and discuss the improvements observed, as well as the factors that affect the performance of the model.

1 Introduction

Annotated resources that have been developed over the course of the last decade for a variety of natural language processing (NLP) tasks are crucial to the success of statistical methods in this area. However, even for such standard problems, as part-of-speech tagging, syntactic parsing, or semantic role labeling (SRL), resources are scarce for many of the world's languages. Fortunately, similarities between syntactic and semantic constructions in different languages suggest that existing resources may significantly reduce the human effort required to produce such resources for a language that lacks them.

This idea has led to the development of cross-lingual annotation projection approaches, such as [1, 2, 3], as well as attempts to adapt models directly to other languages, e.g. [4, 5, 6, 7], most of which make use of parallel data to link the languages together. The success of such methods in transferring various forms of syntactic information suggests that it may prove even more effective when applied to shallow semantic parsing, as semantic structure is, intuitively, more likely to be preserved in translation [8].

Most cross-lingual annotation projection approaches transfer the source language annotation scheme without modification, which makes it hard to combine their output with existing target language resources, as annotation schemes may vary significantly. In the present work, we instead address the problem of information transfer between two *existing* annotation schemes for two different languages.

The ability to bootstrap an SRL model from parallel data may prove useful in a setup similar to annotation projection, where only a small amount of data is available for the target language, as well as when datasets of comparable size exist for both languages, where we can benefit from the fact that certain structures are less ambiguous in one language than in another or from the difference in the coverage of the datasets. We will henceforth refer to these as the *projection setup* and the *symmetric setup*.

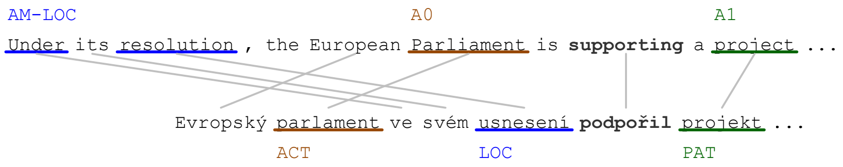
AM-LOC
A0
A1
Under its resolution , the European Parliament is supporting a project ...
Evropský parlament ve svém usnesení podpořil projekt ...
ACT
LOC
PAT

Figure 1: An example of role correspondence in an English-Czech sentence pair.

The intuition behind this approach is that a translated sentence should realize more or less the same set of predicates as the original one, and the corresponding predicates should have similar arguments. We extract the pairs of corresponding arguments for each pair of corresponding predicated based on the word-alignment information. The arguments in such pairs may be expected to bear the same meaning, even if labeled differently in the two annotation schemes [9], see figure 1. We will refer to this underlying meaning as the semantic function of an argument, as opposed to its semantic role. The schemes we consider have a relatively small inventory of semantic roles, and at least some of these roles have a consistent interpretation (a mapping to a certain semantic function) only in the context of a particular predicate.

Defining the semantic function mentioned above and modeling it explicitly is a rather complex problem. Fortunately, we can avoid that by modeling the mapping between the semantic roles of the two annotation schemes in their respective contexts directly. Even in this case, accurate estimation of the model parameters would require a significant amount of parallel annotated data, which is rarely available, so instead we assume that the predictions of our initial models are correct more often than not and make sure that the mapping of semantic role labels is consistent across all occurrences of a given predicate pair. The consistency in enforced by means of a *projection model*, which predicts the semantic role of the target argument based on that of the source one plus the source and target predicates. We perform joint inference to find a compromise between the projection model and the monolingual models (figure 2).

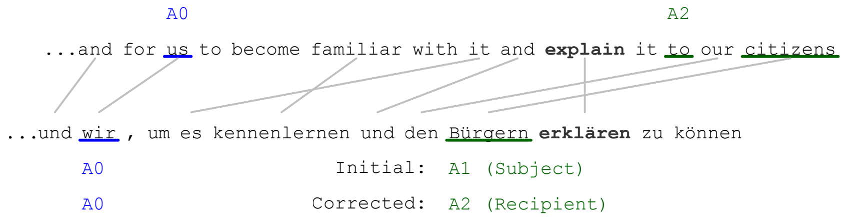
A0
A2
...and for us to become familiar with it and explain it to our citizens
...und wir , um es kennenlernen und den Bürgern erklären zu können
A0
Initial: A1 (Subject)
A0
Corrected: A2 (Recipient)

Figure 2: An erroneously classified argument in the German sentence is corrected by the projection model.

The induced annotations are added to the initial training set and the model trained on the augmented data is evaluated on a held-out test set. To our knowledge, this is the first approach for this setting, so we only compare with *supervised* and *self-training* baselines. The former is the model trained on the initial training set, the latter – on the initial training set plus the output or our model without the projection component.

We show that given a small initial training set our approach yields a moderate improvement for certain language pairs in both symmetric and projection settings and investigate the possible issues with the present approach and ways of resolving them.

2 Approach

Let us consider a pair of languages, α and β, and their corresponding sets T_α and T_β of labeled training instances $\langle S, p, \bar{a}, \bar{r} \rangle$, where p is a predicate in the sentence S, $\bar{a}$ is the argument tuple, containing positions of words, identified as arguments, and $\bar{r}$ is the tuple of semantic role labels for these arguments. We will also need a set aligned predicate instances $\left\langle S^\alpha, S^\beta, p_\alpha, p_\beta, \bar{a}_\alpha, \bar{a}_\beta, l \right\rangle$ extracted from parallel data, where l is a set of aligned argument pairs. Here only the aligned arguments are considered, the rest are removed during preprocessing.

We start by training the monolingual SRL models on T_α and T_β respectively and applying them to the source and target predicate in each pair. The obtained annotations are used to train the projection model, and then the joint inference step is run, using both monolingual models and the projection model. The monolingual models are then retrained on the initial training data augmented with that obtained in the joint inference step.

The above procedure can be run iteratively, as is typical for bootstrapping approaches, but we found it to yield no noticeable improvement in our experiments. It can also be seen as a form of co-training [10] of a pair of monolingual SRL models, with the addition of the statistical agreement model (projection model).

2.1 Joint Inference

In the projection setup we assume that the model for one of the languages, which we will henceforth refer to as *source*, is much better informed than the one for the other language, referred to as *target*, so we only have to propagate the information one way. Let $\bar{r}^*_\alpha$ denote the initial prediction of the source language model $M_\alpha(S^\alpha, \bar{a}_\alpha, p_\alpha) \rightarrow \bar{r}_\alpha$. We assume we also have the initial model $M_\beta(S^\beta, \bar{a}_\beta, p_\beta) \rightarrow \bar{r}_\beta$ for the target language and a projection model $M_{\alpha\beta}(\bar{r}_\alpha, p_\alpha, p_\beta) \rightarrow \bar{r}_\beta$ capable of predicting semantic roles of arguments of the target predicate given those of the corresponding arguments of the source one. The task is then to identify the role assignment $\bar{r}_\beta$ that would maximize the objective $L(\bar{r}_\beta) = \lambda_\beta f_\beta(\bar{r}_\beta, S^\beta, \bar{a}_\beta, p_\beta) + \lambda_{\alpha\beta} f_{\alpha\beta}(\bar{r}_\beta, \bar{r}^*_\alpha, p_\alpha, p_\beta)$, where $\bar{r}^*_\alpha = argmax_{\bar{r}} f_\alpha(\bar{r}_\alpha, S^\alpha, \bar{a}_\alpha, p_\alpha)$ and f_α, f_β are $f_{\alpha\beta}$ are the scoring functions associated with the models.

To maximize this objective, we employ the dual decomposition method, as it fits the structure of the problem well and allows a wide range of monolingual models to be used in this setup. It is implemented using the subgradient descent algorithm, following [11]. The only requirement is that the monolingual model can incorporate a bias towards or away from a certain prediction. While the method is not guaranteed to converge, we observe that it does so within the first 500 iterations approximately 99% of the time in our experiments. Whenever it does converge, the result is guaranteed to represent the global maximum of the sum of the objectives.

In the case of the symmetric setup, where we would like to transfer the information both ways simultaneously, the structure of the problem changes somewhat. It is possible to generalize the approach presented above to this latter case directly, by performing the above procedure independently for the two directions. This means, however, that we fix the initial predictions of the monolingual models and the projection model acts based on those. It would intuitively be more desirable to have the projection model make its predictions on the basis of the updated predictions of the monolingual models, which we achieve by interleaving the subgradient descent algorithm steps for the two problems.

The resulting iterative procedure does not, unfortunately, fit the dual decomposition framework and therefore the optimality certificate does not apply, i.e. even if the algorithm converges, there is no guarantee that the solution is the a global maximum of the sum of the objectives, but we found it to perform better in practice.

2.2 Implementation

The annotation schemes for all languages we consider are those used in the CoNLL Shared Task 2009 [12]. The SRL parser is based on that of [13]. It is a pipeline system of linear classifiers, trained using Liblinear [14]. As we are working with small amounts of data, we have not used the reranker the system provides, but introduced a uniqueness constraint on certain semantic roles, which proves more useful is such a setting.

Projection models are realized by a single linear classifier applied to each argument pair independently. A projection model $M_{\alpha\beta}$ relies on the features derived from the source semantic role, as well as source and target predicates, and produces the semantic role for the argument in the target sentence and vice versa.

3 Results

We evaluate our approach on four language pairs, namely English vs German, Spanish, Czech and Chinese, which we will denote `en-de`, `en-es`, `en-cz` and `en-cn` respectively. The parallel data is drawn from Europarl v6 [15] and MultiUN [16].

In the evaluation we consider small subsets of the training data in order to emulate the scenario with a resource-poor language. We have found that due to the different sources of data for the shared task, sentences contain different proportions of annotated predicates, so we measure the initial training set sizes in the number of argument labels they contain, or *instances*, rather than in the number of sentences.

We evaluate on a held-out test set with predicted syntax provided, as well as an out-of-domain test set, where available, as the parallel data comes from a different domain than the annotated training data. We used 500 instances for each language in the symmetric setup, as well as for the target language in the projection setup. We have also conducted an additional experiment to demonstrate the importance of the projection component in the present approach. Here the projection models are trained on the output of the full models (20000 training instances) and then used in the same projection scenario as above. Thus a more informed, *oracle* projection model is emulated. In practice this can be achieved by e.g. incorporating prior knowledge, including certain language-specific aspects, or using external sources of information.

The results are summarized in 1. The columns correspond to supervised baseline (SUP), self-training baseline (SELF), projection setup (PROJ), symmetric setup (SYM) and projection setup with oracle projection model (ORACLE). The value in parentheses denotes the improvement over the supervised baseline, with the value highlighted in bold if it is statistically significant with $p < 0.005$ according to the permutation test [17]. Note that also for the symmetric setup we only present the accuracies for the target language, as no statistically significant improvement was observed for English there. Note also that self-training has a mostly negative effect on the model's performance [18, 19, 20], which has to be overcome by the joint model.

	SUP	**SELF**	**PROJ**	**SYM**	**ORACLE**
en-cn	76.5	75.1	76.5 (+0.0)	76.4 (−0.0)	76.5 (+0.0)
en-cz	55.8	55.7	56.1 (+**0.4**)	56.3 (+**0.5**)	56.7 (+**1.0**)
en-cz (ood)	57.0	57.0	57.5 (+0.5)	56.6 (−0.4)	57.5 (+0.5)
en-de	61.0	58.6	60.5 (−0.6)	60.6 (−0.5)	61.1 (+0.1)
en-de (ood)	64.3	59.7	68.0 (+**3.7**)	67.8 (+**3.5**)	69.1 (+**4.8**)
en-es	62.3	61.8	63.9 (+1.5)	64.4 (+**2.1**)	65.2 (+**2.9**)

Table 1: Projection setup results. ”ood” indicated evaluation on an out-of-domain test set.

Conclusions

We have presented an approach to information transfer between SRL systems for different language pairs using parallel data. The task proves challenging due to non-trivial mapping between the role labels used in different SRL annotation schemes and the nature of parallel data – the difference in domains and the limited accuracy of the preprocessing tools. Nevertheless, in most experiments, the model achieves a moderate improvement, and we show that a more sophisticated projection model, e.g. one incorporating language- and annotation scheme-specific prior knowledge, can further boost the performance of this approach.

Interesting directions for future work include extending the method to multiple source languages, either by merging together the data produced by running it for each source language separately, or by extending the model to account for multiple source languages directly.

Acknowledgments

The authors acknowledge the support of the MMCI Cluster of Excellence and thank Alexandre Klementiev for valuable suggestions.

References

[1] Lonneke van der Plas, Paola Merlo, and James Henderson. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 299–304, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[2] Paolo Annesi and Roberto Basili. Cross-lingual alignment of framenet annotations through hidden markov models. In *Proceedings of the 11th international conference on Computational Linguistics and Intelligent Text Processing*, CICLing'10, pages 12–25, Berlin, Heidelberg, 2010. Springer-Verlag.

[3] Tahira Naseem, Regina Barzilay, and Amir Globerson. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[4] Greg Durrett, Adam Pauls, and Dan Klein. Syntactic transfer using a bilingual lexicon. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1–11, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[5] Anders Søgaard. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 682–686, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[6] Adam Lopez, Daniel Zeman, Michael Nossal, Philip Resnik, and Rebecca Hwa. Cross-Language Parser Adaptation between Related Languages. In *IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42, Hyderabad, India, January 2008.

[7] Ryan McDonald, Slav Petrov, and Keith Hall. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 62–72, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[8] Sebastian Padó and Mirella Lapata. Cross-linguistic projection of role-semantic information. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 859–866, Vancouver, British Columbia, Canada, 2005.

[9] Tao Zhuang and Chengqing Zong. Joint inference for bilingual semantic role labeling. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 304–314, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[10] Avrim Blum and Tom M. Mitchell. Combining labeled and unlabeled sata with co-training. In Peter L. Bartlett and Yishay Mansour, editors, *COLT*, pages 92–100. ACM, 1998.

[11] David Sontag, Amir Globerson, and Tommi Jaakkola. Introduction to dual decomposition for inference. In Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors, *Optimization for Machine Learning*. MIT Press, 2011.

[12] Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado, June 2009. Association for Computational Linguistics.

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

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

[15] Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT.

[16] Andreas Eisele and Yu Chen. MultiUN: A multilingual corpus from united nation documents. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[17] P. Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer, 2000.

[18] Shan He and Daniel Gildea. Self-training and co-training for semantic role labeling: Primary report. Technical report, University of Rochester, 2006.

[19] Hagen Fürstenau and Mirella Lapata. Graph alignment for semi-supervised semantic role labeling. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 11–20, Singapore, 2009.

[20] D. Goldwasser, R. Reichart, J. Clarke, and D. Roth. Confidence driven unsupervised semantic parsing. In *ACL*, 2011.