Bootstrapping Semantic Role Labelers from Parallel Data

Kozhevnikov, M.; Titov, I.

Publication date
2013
Document Version
Author accepted manuscript
Published in
Second Joint Conference on Lexical and Computational Semantics : *SEM. - Volume 1

Bootstrapping Semantic Role Labelers from Parallel Data

Mikhail Kozhevnikov **Ivan Titov**
Saarland University, Postfach 15 11 50
66041 Saarbrücken, Germany
{mkozhevn|titov}@mmci.uni-saarland.de

Abstract

We present an approach which uses the similarity in semantic structure of bilingual parallel sentences to bootstrap a pair of semantic role labeling (SRL) models. 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. Our approach can facilitate the construction of SRL models for resource-poor languages, while preserving the annotation schemes designed for the target language and making use of the limited resources available for it. We evaluate the model on four language pairs, English vs German, Spanish, Czech and Chinese. Consistent improvements are observed over the self-training baseline.

1 Introduction

The success of statistical modeling methods in a variety of natural language processing (NLP) tasks in the last decade depended crucially on the availability of annotated resources for their training. And while sizable resources for most standard tasks are only available for a few languages, the human effort required to achieve reasonable performance on such tasks for other languages may be significantly reduced by leveraging existing resources and the similarities between languages.

This idea has lead to the development of cross-lingual annotation projection approaches, which make use of parallel corpora (Padó and Lapata, 2009), as well as attempts to adapt models directly to other languages (McDonald et al., 2011). In this paper we consider correspondences between SRL structures in translated sentences from a different perspective. Most cross-lingual annotation projection approaches transfer the source language annotation scheme to the target language without modification, which makes it hard to combine their output with existing target language resources, as annotation schemes may vary significantly. We instead address the problem of information transfer between two *existing* annotation schemes (figure 1) for a pair of languages using an intermediate model of role correspondence (RCM). An RCM models the probability of a pair of corresponding arguments being assigned a certain pair of roles. We then use it to guide a pair of monolingual models toward compatible predictions on parallel data in order to extend the coverage and/or accuracy of one or both models.

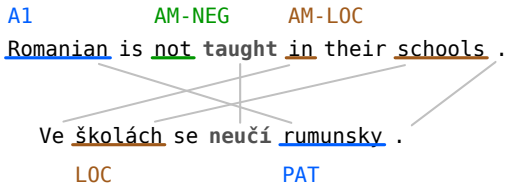

Figure 1: Role correspondence in parallel sentences, an example.

The notion of compatibility here is highly non-trivial, even for sentences translated as close to the original as possible. Zhuang and Zong (2010), for example, observe that in the English-Chinese parallel PropBank (Palmer et al., 2005b) corresponding arguments often bear different labels, even though the same inventory of semantic roles is used for both

languages and the annotation guidelines are similar. When different annotation schemes are considered, the problem is further complicated by the difference in the granularity of semantic roles used and varying notions of what is an argument and what is not.

Manually annotated training data for such a model is hard to come by. Instead, we propose an iterative procedure similar to bootstrapping, where the parameters of the RCM are initially estimated from a parallel corpus automatically annotated with semantic roles using the monolingual models independently, and then the RCM is used to refine these annotations via a joint inference procedure, serving to enforce consistency on the predictions of monolingual models on parallel sentences. The obtained annotations on the parallel corpus are expected to be of higher quality than the independent predictions of the models, so they can be used to improve the SRL models' performance and/or coverage. We evaluate this approach by augmenting the original training data with the annotations obtained on parallel data and observing the change in the model's performance. This is especially useful if one of the languages is relatively poor in resources, in which case the proposed procedure will help propagate information from the stronger model to the weaker one. Even if the two models are comparable in their predictive power, we may be able to benefit from the fact that certain semantic roles are realized less ambiguously in one language than in another. We will henceforth refer to these two alternatives as the *projection* and *symmetric* setups.

The paper is structured as follows. In the next section we present our approach and discuss the issues of role correspondence modeling, then describe the implementation and datasets used in evaluation in section 3, present the evaluation and results in section 4 and conclude with the discussion of related work in section 5.

2 Approach

We consider bootstrapping a pair of SRL models on a parallel corpus, using the correspondence between their predictions on parallel sentences to guide the learning. The models are forced toward compatible predictions, where the notion of compatibility is defined by a (statistical) role correspondence model.

Let us consider a pair of languages, α and β, and their corresponding datasets T^0_α and T^0_β, annotated with semantic roles (the upper indices here denote the iteration number). We will refer to these as the *initial* training sets. We also assume that a word-aligned parallel corpus is available for the pair of languages, which we denote P, with the predicates and their respective arguments identified on both sides.

The procedure is then as follows: we train monolingual models M^0_α and M^0_β on T^0_α and T^0_β, respectively, apply them to the two sides of the parallel corpus, resulting in a labeling P^0. We collect the semantic role co-occurrence information and train the role correspondence model C^0 on it, then proceed to the joint inference step involving M^0_α, M^0_β and C^0, resulting in a refined labeling P^1 of the parallel corpus. The two sides of the P^1 are then used to augment the initial training sets, yielding T^1_α and T^1_β, and new models M^1_β and M^1_β are trained on these. The process can then be repeated using M^1_α and M^1_β instead of the initial models.

We report the model's performance on a held-out test set, drawn from the same corpus as the corresponding initial training set.

The procedure can be seen as a form of co-training (Blum and Mitchell, 1998) of a pair of monolingual SRL models. In our case, however, the question of the models' agreement is not as trivial as in most applications of co-training, requiring a statistical model of its own (C^i).

In the low-resource (*projection*) setup our approach is also similar to self-training with weak supervision coming from the stronger model.

Note that although the approach is iterative, we have observed no significant improvements from repeating the procedure, possibly owing to the noise introduced by the errors in preprocessing. In the evaluation we run only one iteration. In the notation introduced above, the self-training baseline model (SELF) is trained on P^0_β, the joint model (JOINT) – on P^1_β and the combined model (COMB) – on T^1_β.

2.1 Modeling Role Correspondence

It is necessary to distinguish between semantic roles and their interpretation in a particular context. The former can be defined in a variety of

ways, depending on the formalism used. In case of FrameNet (Baker et al., 1998), for example, the interpretation of a semantic role (frame element) is explicitly provided for each separate frame, so a frame and a frame element label together describe the semantics of an argument. PropBank (Palmer et al., 2005a) follows a mixed strategy – the labels for a relatively small set of *core roles* are numbered and their interpretations are provided separately for each predicate (although those of the first two roles, `A0` and `A1`, consistently denote what is known as Proto-Agent and Proto-Patient), while *modifiers* (Merlo and Leybold, 2001) bear labels that are interpreted consistently across all predicates. Other resources, such as Prague Dependency Treebank (Hajič et al., 2006), use a single set of semantic roles (functors), which are interpretable across different predicates.

From the standpoint of defining the semantic similarity of parallel sentences, the important implication is that we cannot assume that the corresponding arguments should bear the same label, even if the annotation schemes used are compatible (Zhuang and Zong, 2010). Nor can we write down a single mapping between the roles that will be valid across different predicates (figure 2), which motivates the need for a statistical model of semantic role correspondence.

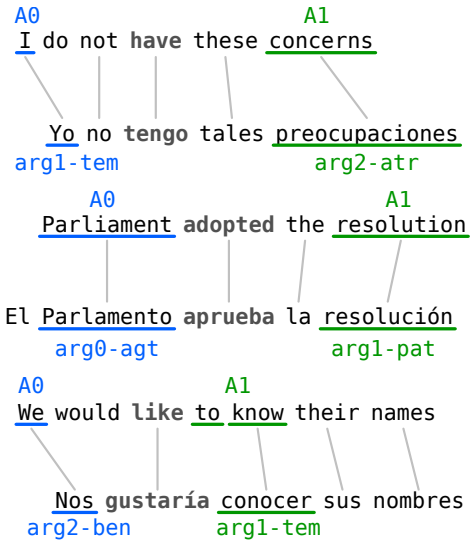

Figure 2: Predicate-specific role mapping. Note that `A0` corresponds to `art0-agt`, `art1-tem` or `art2-ben`, depending on the predicate.

We assume the existence of a one-to-one mapping between semantic roles for a given predicate pair. As the mappings are not completely independent – at least some roles have the same interpretation across different predicate pairs, – we choose to build a single model, which relies on features derived from the pair of predicates in question, rather than create a separate model for each predicate pair. The model can then make decisions specific to particular predicates or predicate pairs, where sufficient data has been observed or back off to a generic mapping where there is not enough data.

For the purpose of this study, we choose to separately model the probability of a target role, given the source one and the necessary contextual information and vice versa. These two components are referred to as *projection models* and realized as a pair of linear classifiers.

Training such a model in a conventional fashion would require a rather specific kind of dataset, namely a parallel corpus annotated with semantic roles, and assuming the availability of such data would severely limit the applicability of the approach proposed, as, to our knowledge, it is currently only available for two language pairs, namely English-Chinese (Palmer et al., 2005b) and English-Czech (Hajič et al., 2012). We instead use the automatically produced annotations on a parallel corpus, effectively enforcing consistency on the role correspondence in the monolingual models' predictions.

2.2 Joint Inference

The joint inference would have been simplest if the arguments were classified independently. This assumption is too restrictive, though, since the interdependencies between the arguments can be used to improve the accuracy of semantic role labeling (Roth and Yih, 2005).

2.2.1 Projection Setup

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. The scoring functions of these two models will be denoted f_s and f_t, respectively, and that of the projection model from source to target – f_{st}. Source and target sentences are denoted S_s and S_t,

and aligned predicates in these sentences – p_s and p_t. The task is then to identify the target language role assignment r_t that would maximize the objective $L(r_t) = \lambda_t f_t(r_t, S_t, p_t) + \lambda_{st} f_{st}(r_t, r_s, p_s, p_t)$, where $r_s = argmax_r f_s(r_s, S_s, p_s)$ is the role assignment of the source-side arguments as predicted by the monolingual model and λ are the weights associated with the models.

The exact maximization of this objective is computationally expensive, so we resort to an approximation. We chose to use the *dual decomposition* method primarily because it fits the structure of the objective particularly well (in that it is a sum of the objectives of two independent models) and since it allows a wide range of monolingual models to be used in this setup. The only requirement here is that the monolingual model must be able to incorporate a bias toward or away from a certain prediction.

To apply this approximation, we decouple the r_t variables into r_t and r_{st} and get $L_1(r_t, r_{st}) = \lambda_t f_t(r_t, S_t, p_t) + \lambda_{st} f_{st}(r_{st}, r_s, p_s, p_t)$ under the condition that $r_t = r_{st}$. Applying the Lagrangian relaxation, we replace the hard equality constraint on r_t and r_{st} with a soft one, using slack variables δ, which results in the following objective:

$$\begin{aligned} &min_\delta max_{r_t, r_{st}} L_1'(r_t, r_{st}, \delta) = \\ &\lambda_t f_t(r_t, S_t, p_t) + \lambda_{st} f_{st}(r_{st}, r_s, p_s, p_t) + \\ &\sum_i \sum_{r \in R_t} \delta^{i,r} \left(I(r_t^i = r) - I(r_{st}^i = r) \right), \end{aligned} \tag{1}$$

where i indexes aligned argument pairs and I is an indicator function. This is equivalent to

$$\begin{aligned} &min_\delta max_{r_t, r_{st}} L_1'(r_t, r_{st}, \delta) = \\ &min_\delta \Big(max_{r_t} g_t(r_t, S_t, p_t, \delta) + \\ &max_{r_{st}} g_{st}(r_{st}, r_s, p_s, p_t, \delta) \Big), \end{aligned} \tag{2}$$

where

$$\begin{aligned} &g_t(r_t, S_t, p_t, \delta) = \\ &\lambda_t f_t(r_t, S_t, p_t) + \sum_i \sum_{r \in R_t} \delta^{i,r} I(r_t^i = r) \\ &g_{st} p(r_{st}, r_s, p_s, p_t, \delta) = \\ &\lambda_{st} f_{st}(r_{st}, r_s, p_s, p_t) - \sum_i \sum_{r \in R_t} \delta^{i,r} I(r_{st}^i = r) \end{aligned} \tag{3}$$

are the augmented objectives of the two component models, incorporating bias factors on various possible predictions.

The minimization with respect to δ is performed using a subgradient descent algorithm following Sontag et al. (2011). Whenever the method converges, it converges to the global maximum of the sum of the objectives. We found that in our case it reaches a solution within the first 1000 iterations over 99% of the time.

2.2.2 Symmetric Setup

If the models have comparable accuracy, the above inference procedure can be extended to perform projection both ways. Formulating this as a dual decomposition problem would require using three separate components, two for the monolingual models and one for the RCM, which would have to make its own predictions for the semantic roles on both sides without conditioning on the predictions of the monolingual models. This calls for a different kind of model than the one we use – a model that will rely on a (possibly simplified) feature representation of the source and target arguments to jointly predict their labels. Instead, we perform the projection setup inference procedure in both directions simultaneously, interleaving gradient descent steps and allowing the projection models to access the updated predictions of the monolingual models. This results in a block gradient descent algorithm with the following updates:

$$\begin{aligned} &r_t^{n+1} = argmax_{r_t} g_t(r_t, S_t, p_t, \delta_t^n) \\ &r_s^{n+1} = argmax_{r_t} g_s(r_s, S_s, p_s, \delta_s^n) \\ &r_{st}^{n+1} = argmax_{r_{st}} g_{st}(r_{st}, r_s^n, p_s, p_t, \delta_t^n) \\ &r_{ts}^{n+1} = argmax_{r_{ts}} g_{ts}(r_{ts}, r_t^n, p_t, p_s, \delta_s^n) \\ &\forall_i \forall_{r \in R_s} \delta_s^{n+1,i,r} = \delta_s^{n,i,r} + \\ &\quad \gamma_s(n)(I(r_{ts}^{n,i} = r) - I(r_s^{n,i} = r)) \\ &\forall_i \forall_{r \in R_t} \delta_t^{n+1,i,r} = \delta_t^{n,i,r} + \\ &\quad \gamma_t(n)(I(r_{st}^{n,i} = r) - I(r_t^{n,i} = r)), \end{aligned} \tag{4}$$

where $\gamma_s(n) = \gamma_t(n) = \frac{\gamma_0}{n+1}$ is the update rate function for step n, and g_s and g_{ts} are defined as in (3), but with the direction reversed.

This procedure allows us to use the same RCM implementation as in the projection setup. Moreover, the inference procedure for projection setup is

a special case of this one with $\gamma_s(n)$ set to 0. The algorithm also demonstrates convergence similar to that of the projection version, although it lacks the optimality guarantees.

3 Experimental Setup

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-zh` respectively.

3.1 Parallel Data

The parallel data for the first three language pairs is drawn from Europarl v6 (Koehn, 2005) and from MultiUN (Eisele and Chen, 2010) for English-Chinese. We applied Stanford Tokenizer for English, tokenizer scripts (Koehn, 2005) provided with the Europarl corpus to German, Spanish and Czech, and Stanford Chinese Segmenter (Chang et al., 2008) to Chinese, then performed POS-tagging, morphology tagging (where applicable) and dependency parsing using MATE-tools (Bohnet, 2010).

Word alignments were acquired using GIZA++ (Och and Ney, 2003) with its standard settings. Predicate identification on the parallel data was done using the supervised classifiers of the monolingual SRL systems, except for German, where a simple heuristic had to be used instead, as only some of the predicates are marked in the training data, which makes it hard to train a supervised classifier. Following van der Plas et al. (2011), we then retain only those sentences where all identified predicates were aligned.

In the experiments we used 50 thousand predicate pairs in each case, as increasing the amount further did not yield noticeable benefits, while increasing the running time.

3.2 Annotated Data

The CoNLL'09 (Hajič et al., 2009) datasets were used as a source of annotated data for all languages. Only verbal predicates were considered and predicted syntax was used in evaluation.

We consider subsets of the training data in order to emulate the scenario with a resource-poor language. Due to the different sources the datasets were derived from, sentences contain different proportions of annotated predicates depending on the language. The German corpus, for example, contains about 6 times fewer argument labels per sentence than the English one. We will therefore indicate the sizes of the datasets used in the number of argument labels they contain, referred to as *instances*, rather than the number of predicates or sentences. The corpus for English, for example, contains 6.2 such instances per sentence on average.

We use the 20 thousand instances of the available data as the training corpus for each language and split the rest equally between the development and the test set. The secondary ("out-of-domain") test sets are preserved as they are.

In dependency-based SRL, only heads of syntactic constituents are marked with semantic roles. The heads of corresponding arguments may or may not align, however, even if the arguments are lexically very similar, because their syntactic structure may differ. In general, one would have to identify the whole phrase for each argument and take into account the links between constituents, rather than single words (Padó and Lapata, 2005). As reconstructing the constituents from the dependency tree is non-trivial (Hwang et al., 2010), we are using a heuristic to address the most common version of this problem, i.e. a preposition or an auxiliary verb being an argument head. In such a case we also take into account any alignment links involving the head's immediate descendants.

3.3 Implementation

Our system is based on that of Björkelund et al. (2009). It is a pipeline system comprised of a set of binary or multiclass linear classifiers. Both here and in the projection model, the classifiers are trained using Liblinear (Fan et al., 2008).

We employed a uniqueness constraint on role labels (Chang et al., 2007), preventing some of them from being assigned to more than one argument in the same predicate, which appears to be more reliable in a low-resource setting we consider than the reranker the original system employed. The constraint is enforced in the monolingual model inference using a beam-search approximation with the beam size of 10. The label uniqueness information was derived from the training sets.

3.4 The Projection Model

Each projection model is realized by a single linear classifier applied to each argument pair independently. It relies on features derived from the source semantic role and source and target predicates, and predicts the semantic role for the argument in the target sentence.

The features include the source semantic role and its conjunctions with (lowercased) forms and lemmata of the source and target predicates. For example, assuming the source semantic role is `A3` and the source and target predicates are `went` and `ging` (past tense of "gehen", German), the features would be as shown in figure 3.

```
FORMPAIR=A3-went-ging
LEMMAPAIR=A3-go-gehen
FORMSRC=A3-went
FORMTGT=A3-ging
LEMMASRC=A3-go
LEMMATGT=A3-gehen
LABEL=A3
```

Figure 3: Projection model features example.

3.5 Parameters

In case of projection there are two parameters, λ_{st} and λ_t, – the weights of the component models in the objective. Only their relative values matter (except in the choice of γ_0), so we set λ_t to 1 and only tune the weight of the role correspondence model.

In the symmetric setup, the objective takes the form $L(r_t, r_s) = \lambda_t f_t(r_t, S_t, p_t) + \lambda_{st} f_{st}(r_t, r_s, p_s, p_t) + \lambda_s f_s(r_s, S_s, p_s) + \lambda_{ts} f_{ts}(r_s, r_t, p_t, p_s)$. Since we assume that the two monolingual models here have comparable performance, we do not tune their relative weights, setting both λ_s and λ_t to 1.

We also use the same weight for both projection models, $\lambda_{st} = \lambda_{ts}$, and this value plays an important role – it basically indicates how strongly we insist on the role correspondence models' correctness. If this weight is set to 0, the RCM will accept the initial predictions the monolingual models make, and if it is set to a sufficiently large value, the predictions of the monolingual models will be biased until they match the mapping suggested by the RCM. The optimal weight will therefore depend on the language pair, the sizes of the initial training sets and the RCM used. We use the value of 0.7 in all projection experiments and 0.5 in the symmetric setup, however, as excessive tuning may be undesirable in the low-resource setting.

3.6 Domains

One important factor in the understanding of the evaluation figures presented is the fact that sources of annotated and parallel data belong to different domains. The former usually contains some sort of newswire text – Wall Street Journal in case of English, Xinhua newswire, Hong Kong news and Sinorama news magazine for Chinese, etc. Parallel data, on the other hand, comes from the proceedings of European Parliament and United Nations, which are quite different. For example, the sentences in the latter domain often start with someone being addressed, either by name or by title, which can hardly be expected to occur as often in a newspaper or a magazine article.

As is well-known, the performance of many statistical tools drops significantly outside the domain they were trained on (Pradhan et al., 2008), and the preprocessing and SRL models used here are no exception, which results in relatively low quality of the initial predictions on the parallel text. The low argument identification performance, in particular, is presumably due to inaccurate dependency parses, on which it heavily relies. Several approached have been proposed to improve the accuracy of dependency parsers and other tools on out-of-domain data, but this is beyond the scope of this paper. In some cases (though seldom), sources of parallel data belonging to the same domain as the annotated training data can be obtained.

Another concern is that the performance of a model trained on automatically labeled parallel data as measured on a test set we use may not reflect the quality of these annotations. To assess the resulting model's coverage, it would be interesting to evaluate it on data outside the original domain, so we consider the out-of-domain (OOD) test sets as provided for the CoNLL Shared Task 2009 where available.

Perhaps the most interesting one of these is the German OOD test set, which is drawn from Europarl (as is the parallel data we use). It was originally annotated with syntactic dependency trees and se-

mantic structure in the SALSA format (Burchardt et al., 2006) for Padó and Lapata (2005), and then converted into a PropBank-like form for the CoNLL Shared Task 2009 (Hajič et al., 2009). The OOD test set for English is drawn from the Brown corpus (Francis and Kucera, 1967) and the one for Czech – from a Czech translation of Wall Street Journal articles (Hajič et al., 2012).

4 Evaluation

The first question we are interested in is how the joint inference affects the quality of the automatically obtained annotations on the parallel data. To answer this, we will run the monolingual models independently and jointly, then train models on the output of these two procedures and compare their performance on a test set. Note that we do not add the initial training data at this point, so the initial model scores are provided for reference, rather than as a baseline.

4.1 Projection Setup

A small initial training set of 600 instances was used here for the target language here and the full training set (20000 instances) for the source one. λ_{st} was set to 0.7 in all experiments in this section.

	INIT	SELF	JOINT	Δ_{SELF}
en-cz*	61.11	60.68	63.01	**2.33**
en-cz	62.45	62.15	63.11	**0.96**
en-de*	66.81	63.96	67.64	**3.69**
en-de	70.40	68.34	70.13	**1.79**
en-es	64.20	64.51	66.01	**1.50**
en-zh	75.80	73.52	74.87	**1.35**
cz-en*	66.82	63.95	64.97	**1.02**
cz-en	74.92	71.60	71.90	**0.29**
de-en*	66.82	63.58	63.21	-0.37
de-en	74.93	71.31	70.72	-0.59
es-en*	66.82	63.95	64.18	0.23
es-en	74.93	71.47	72.09	**0.62**
zh-en*	66.82	64.51	63.67	-0.83
zh-en	74.93	72.26	71.24	-1.01

Table 1: Projection setup results: self-training baseline, refined model and the difference in their performance. Asterisk indicates out-of-domain test set, statistically significant improvements are highlighted in bold.

In table 1, we present the accuracy of the model trained on the output of the joint inference (JOINT) against that of the self-training baseline (SELF). The Δ_{SELF} column contains the difference between the two. Note that the SELF model is trained on the parallel data automatically annotated using monolingual SRL models (not mixed with the initial training set), since we are interested in the effect of joint inference on the quality of the annotation obtained. Where the improvement is positive and statistically significant with $p < 0.005$ according to the permutation test (Good, 2000), they are highlighted in bold.

We can see that the refined model (JOINT) outperforms the self-training baseline in most cases by a moderate, but statistically significant margin, which indicates that the joint inference does improve the quality of annotations on the parallel corpus.

The slightly higher improvement on the German OOD test set supports our hypothesis that the procedure enhances the performance of the model on parallel data, as the data for this test set is also drawn from the Europarl corpus. The improvement over the initial model (Δ_{INIT}) in this case is statistically significant with $p < 0.05$. Higher p-value may be attributed to the smaller test set size.

Figure 4 shows how the performance of the JOINT model changes with the size of the initial training set. The improvements are smaller for en-cz, en-de and en-zh, but they are also statistically significant for initial training sets of up to 2000 instances. Projection to English from other languages performs worse.

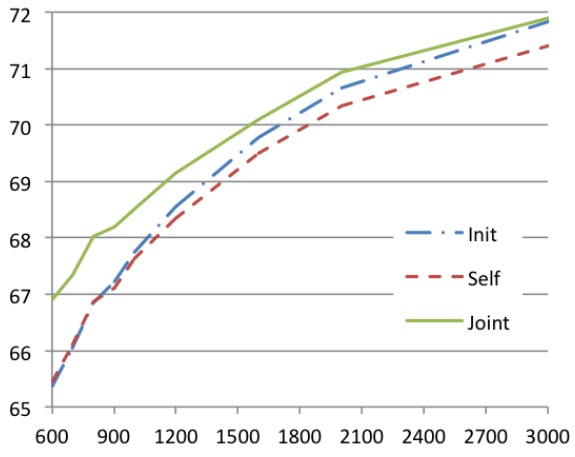

Figure 4: Projection setup, English-Spanish, model performance as a function of the size of the initial training set.

4.2 Combining

In practice, automatically obtained annotations are usually combined with the existing labeled data. For this purpose, the initial training set is replicated so as to constitute 0.3 (an empirically chosen value that appears to work well in most experiments) of the size of the automatically labeled dataset. We compare the performance of the model trained on the resulting dataset (COMB) with that of the JOINT model and the initial models. The results are presented in table 2. We omit projection from other languages to English, since the JOINT model there fails to outperform the initial model and we do not expect to benefit from adding the automatically annotated data to the initial training set in this case.

	INIT	JOINT	COMB	Δ_{JOINT}	Δ_{INIT}
en-cz*	61.11	63.01	62.98	-0.03	**1.87**
en-cz	62.45	63.11	63.30	**0.19**	**0.85**
en-de*	66.81	67.64	67.64	0.00	0.84
en-de	70.39	70.19	70.53	**0.34**	0.15
en-es	64.20	66.01	66.01	0.00	**1.81**
en-zh	75.80	74.87	75.03	0.16	-0.77

Table 2: The effect of adding automatically obtained annotation to the initial training set. Asterisk indicates out-of-domain test set, statistically significant improvements are highlighted in bold.

4.3 Symmetric Setup

In the symmetric setup evaluation, we use a slightly larger initial training set of 1400 instances for both source and target language. The projection model weight is set to 0.5. Table 3 shows the accuracy of the JOINT model and the SELF baseline.

Note that here, unlike section 4.1, the joint inference is run once and then a model is trained for each language and evaluated on the corresponding test set(s).

The results support our intuition that joint inference helps improve the quality of the resulting annotations, at least in some cases.

4.4 Oracle RCM

It would be useful to know to what extent the performance of the role correspondence model affects the quality of the output (and thus the performance of the resulting model). The RCM we use is rather simplistic, and we believe it can be substantially improved for any given language pair by incorporating prior knowledge and/or using external sources of information. In order to estimate the potential impact of such improvements, we simulate a better informed projection model, giving it access to the predictions of more accurate monolingual models on the parallel data – those trained on the full training set, rather than the initial training set used in this particular experiment. We refer to the resulting RCM as *oracle* and assess the difference it makes, compared to a regular one (table 4).

	INIT	SELF	JOINT	Δ_{SELF}
en-cz*	67.07	66.15	68.18	**2.02**
en-cz	67.56	66.42	66.72	**0.30**
en-de*	67.64	66.72	68.57	1.84
en-de	75.13	71.97	73.57	**1.60**
en-es	68.14	67.80	69.04	**1.24**
en-zh	76.28	72.96	75.22	**2.26**
cz-en*	69.37	66.45	66.22	-0.23
cz-en	77.32	74.72	75.02	**0.31**
de-en*	69.37	66.45	66.68	0.23
de-en	77.32	73.56	73.72	0.17
es-en*	69.37	66.64	66.40	-0.23
es-en	77.32	74.05	74.89	**0.84**
zh-en*	69.37	66.08	65.53	-0.56
zh-en	77.32	74.48	74.25	-0.24

Table 3: Comparing JOINT model against the self-training baseline in symmetric setup. Asterisk indicates out-of-domain test set, statistically significant improvements are highlighted in bold.

5 Related Work

There is a number of approaches to semi-supervised semantic role labeling, and most suggest that some external supervision is required for such approaches to work (He and Gildea, 2006), such as measures of syntactic and semantic similarity (Fürstenau and Lapata, 2009) or external confidence measures (Goldwasser et al., 2011). The alternative we propose is primarily motivated by the research on annotation projection (Padó and Lapata, 2009; van der Plas et al., 2011; Annesi and Basili, 2010; Naseem et al., 2012) and direct transfer (Durrett et al., 2012; Søgaard, 2011; Lopez et al., 2008; McDonald et al., 2011). The key difference of the present approach compared to annotation projection is that we assume

	INIT	SELF	JOINT	Δ_{SELF}	Δ_{INIT}
en-cz*	61.11	60.68	72.49	**11.81**	**11.38**
en-cz	62.45	62.15	70.19	**8.04**	**7.74**
en-de*	66.81	63.96	76.78	**12.82**	**9.97**
en-de	70.39	68.34	79.22	**10.88**	**8.84**
en-es	64.20	64.51	75.43	**10.92**	**11.23**
en-zh	75.80	73.52	76.75	**3.22**	**0.94**
cz-en*	66.82	63.95	70.75	**6.80**	**3.93**
cz-en	74.93	71.60	79.70	**8.10**	**4.76**
de-en*	66.82	63.58	69.46	**5.88**	**2.64**
de-en	74.93	71.31	77.34	**6.03**	**2.41**
es-en*	66.82	63.95	69.92	**5.97**	**3.10**
es-en	74.93	71.47	79.55	**8.08**	**4.62**
zh-en*	66.82	64.51	67.19	**2.68**	0.37
zh-en	74.93	72.26	76.51	**4.26**	**1.58**

Table 4: Oracle RCM performance, projection setup: initial model, self-training baseline, refined model and its improvement over the other two. Asterisk indicates out-of-domain test set, statistically significant improvements are highlighted in bold.

the availability of some amount of training data for the target language, possibly using a different inventory of semantic roles.

As mentioned previously, from the training point of view this approach can be seen as similar to co-training (Blum and Mitchell, 1998), other applications of which to NLP are too numerous to list here.

Most closely related is the joint inference in Zhuang and Zong (2010), the main difference being that it relies on a manually annotated parallel corpus, aligned on the argument level, and evaluates only the inference procedure and only on in-domain data.

Other related approaches include Kim et al. (2010), where a cross-lingual transfer of relations is performed (which basically represent parts of the predicate-argument structure considered by SRL methods), and Frermann and Bond (2012), where semantic structure matching is used to rank HPSG parses for parallel sentences.

Unsupervised semantic role labeling methods (Lang and Lapata, 2010; Lang and Lapata, 2011; Titov and Klementiev, 2012a; Lorenzo and Cerisara, 2012) present an alternative to the cross-lingual information propagation approaches such as ours, and at least one the methods in this area also makes use of parallel data (Titov and Klementiev, 2012b).

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. We observe consistent improvements over self-training baseline from using joint inference and the experiments suggest that improving the role correspondence model, for example using language-specific prior knowledge or external data sources, may dramatically increase the performance of the resulting system.

Acknowledgments

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

References

Paolo Annesi and Roberto Basili. 2010. 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. Springer-Verlag.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics (ACL-COLING'98)*, pages 86–90, Montreal, Canada.

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): Shared Task*, pages 43–48, Boulder, Colorado, June. Association for Computational Linguistics.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Workshop on Computational Learning Theory (COLT 98)*.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Lin-*

guistics (Coling 2010), pages 89–97, Beijing, China, August. Coling 2010 Organizing Committee.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of LREC 2006*, Genoa, Italy.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2007. Guiding semi-supervision with constraint-driven learning. In *ACL*.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 224–232, Stroudsburg, PA, USA. Association for Computational Linguistics.

Greg Durrett, Adam Pauls, and Dan Klein. 2012. 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. Association for Computational Linguistics.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from United Nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

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.

S. Francis and H. Kucera. 1967. *Computing Analysis of Present-day American English*. Brown University Press, Providence, RI.

Lea Frermann and Francis Bond. 2012. Cross-lingual parse disambiguation based on semantic correspondence. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 125–129, Jeju Island, Korea, July. Association for Computational Linguistics.

Hagen Fürstenau and Mirella Lapata. 2009. 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.

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

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

J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtskỳ, and M. Ševčíková-Razímová. 2006. Prague dependency treebank 2.0. *LDC*.

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. 2009. 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.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English dependency treebank 2.0. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

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

Jena D. Hwang, Rodney D. Nielsen, and Martha Palmer. 2010. Towards a domain independent semantics: Enhancing semantic representation with construction grammar. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 1–8, Los Angeles, California, June. Association for Computational Linguistics.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2010. A cross-lingual annotation projection approach for relation detection. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 564–571, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

Joel Lang and Mirella Lapata. 2010. Unsupervised induction of semantic roles. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 939–947, Los Angeles, California, June. Association for Computational Linguistics.

Joel Lang and Mirella Lapata. 2011. Unsupervised semantic role induction via split-merge clustering. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*.

Adam Lopez, Daniel Zeman, Michael Nossal, Philip Resnik, and Rebecca Hwa. 2008. Cross-language parser adaptation between related languages. In *IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42, Hyderabad, India, January.

Alejandra Lorenzo and Christophe Cerisara. 2012. Unsupervised frame based semantic role induction: application to French and English. In *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, pages 30–35, Jeju, Republic of Korea, July. Association for Computational Linguistics.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. 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. Association for Computational Linguistics.

Paola Merlo and Matthias Leybold. 2001. Automatic distinction of arguments and modifiers: the case of prepositional phrases. In *Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001)*, pages 121–128, Toulouse, France.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 629–637, Jeju Island, Korea, July. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1).

Sebastian Padó and Mirella Lapata. 2005. 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.

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

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

Martha Palmer, Nianwen Xue, Olga Babko-Malaya, Jinying Chen, and Benjamin Snyder. 2005b. A parallel Proposition Bank II for Chinese and English. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, CorpusAnno '05, pages 61–67, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sameer S. Pradhan, Wayne Ward, and James H. Martin. 2008. Towards robust semantic role labeling. *Computational Linguistics*, 34(2):289–310.

Dan Roth and Wen-tau Yih. 2005. Integer linear programming inference for conditional random fields. In *ICML*, pages 736–743.

Anders Søgaard. 2011. 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*, volume 2 of *HLT '11*, pages 682–686, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Ivan Titov and Alexandre Klementiev. 2012a. A Bayesian approach to unsupervised semantic role induction. In *Proc. of European Chapter of the Association for Computational Linguistics (EACL)*.

Ivan Titov and Alexandre Klementiev. 2012b. Crosslingual induction of semantic roles. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, South Korea, July. Association for Computational Linguistics.

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

Tao Zhuang and Chengqing Zong. 2010. 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. Association for Computational Linguistics.