Crosslingual Induction of Semantic Roles

Ivan Titov **Alexandre Klementiev**
Saarland University
Saarbrücken, Germany
{titov|aklement}@mmci.uni-saarland.de

Abstract

We argue that multilingual parallel data provides a valuable source of indirect supervision for induction of shallow semantic representations. Specifically, we consider unsupervised induction of semantic roles from sentences annotated with automatically-predicted syntactic dependency representations and use a state-of-the-art generative Bayesian non-parametric model. At inference time, instead of only seeking the model which explains the monolingual data available for each language, we regularize the objective by introducing a soft constraint penalizing for disagreement in argument labeling on aligned sentences. We propose a simple approximate learning algorithm for our set-up which results in efficient inference. When applied to German-English parallel data, our method obtains a substantial improvement over a model trained without using the agreement signal, when both are tested on non-parallel sentences.

1 Introduction

Learning in the context of multiple languages simultaneously has been shown to be beneficial to a number of NLP tasks from morphological analysis to syntactic parsing (Kuhn, 2004; Snyder and Barzilay, 2010; McDonald et al., 2011). The goal of this work is to show that parallel data is useful in unsupervised induction of shallow semantic representations.

Semantic role labeling (SRL) (Gildea and Jurafsky, 2002) involves predicting predicate argument structure, i.e. both the identification of arguments and their assignment to underlying semantic roles. For example, in the following sentences:

(a) [$_{A0}$Peter] blamed [$_{A1}$Mary] [$_{A2}$for planning a theft].

(b) [$_{A0}$Peter] blamed [$_{A2}$planning a theft] [$_{A1}$on Mary].

(c) [$_{A1}$Mary] was blamed [$_{A2}$for planning a theft] [$_{A0}$by Peter]

the arguments *'Peter'*, *'Mary'*, and *'planning a theft'* of the predicate *'blame'* take the agent ($A0$), patient ($A1$) and reason ($A2$) roles, respectively. In this work, we focus on predicting argument roles.

SRL representations have many potential applications in NLP and have recently been shown to benefit question answering (Shen and Lapata, 2007; Kaisser and Webber, 2007), textual entailment (Sammons et al., 2009), machine translation (Wu and Fung, 2009; Liu and Gildea, 2010; Wu et al., 2011; Gao and Vogel, 2011), and dialogue systems (Basili et al., 2009; van der Plas et al., 2011), among others. Though syntactic representations are often predictive of semantic roles (Levin, 1993), the interface between syntactic and semantic representations is far from trivial. Lack of simple deterministic rules for mapping syntax to shallow semantics motivates the use of statistical methods.

Most of the current statistical approaches to SRL are supervised, requiring large quantities of human annotated data to estimate model parameters. However, such resources are expensive to create and only available for a small number of languages and domains. Moreover, when moved to a new domain, performance of these models tends to degrade substantially (Pradhan et al., 2008). Sparsity of annotated data motivates the need to look to alternative

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 647–656,
Jeju, Republic of Korea, 8-14 July 2012.

resources. In this work, we make use of unsupervised data along with parallel texts and learn to induce semantic structures in two languages simultaneously. As does most of the recent work on unsupervised SRL, we assume that our data is annotated with automatically-predicted syntactic dependency parses and aim to induce a model of linking between syntax and semantics in an unsupervised way.

We expect that both linguistic relatedness and variability can serve to improve semantic parses in individual languages: while the former can provide additional evidence, the latter can serve to reduce uncertainty in ambiguous cases. For example, in our sentences (a) and (b) representing so-called *blame* alternation (Levin, 1993), the same information is conveyed in two different ways and a successful model of semantic role labeling needs to learn the corresponding linkings from the data. Inducing them solely based on monolingual data, though possible, may be tricky as selectional preferences of the roles are not particularly restrictive; similar restrictions for patient and agent roles may further complicate the process. However, both sentences (a) and (b) are likely to be translated in German as '$[_{A0}$*Peter*$]$ *beschuldigte* $[_{A1}$*Mary*$]$ $[_{A2}$*einen Diebstahl zu planen*$]$'. Maximizing agreement between the roles predicted for both languages would provide a strong signal for inducing the proper linkings in our examples.

In this work, we begin with a state-of-the-art monolingual unsupervised Bayesian model (Titov and Klementiev, 2012) and focus on improving its performance in the crosslingual setting. It induces a linking between syntax and semantics, encoded as a clustering of syntactic signatures of predicate arguments. The clustering implicitly defines the set of permissible alternations. For predicates present in both sides of a bitext, we guide models in both languages to prefer clusterings which maximize agreement between predicate argument structures predicted for each aligned predicate pair. We experimentally show the effectiveness of the crosslingual learning on the English-German language pair.

Our model admits efficient inference: the estimation time on CoNLL 2009 data (Hajič et al., 2009) and Europarl v.6 bitext (Koehn, 2005) does not exceed 5 hours on a single processor and the inference algorithm is highly parallelizable, reducing inference time down to less than half an hour on multiple processors. This suggests that the models scale to much larger corpora, which is an important property for a successful unsupervised learning method, as unlabeled data is abundant.

In summary, our contributions are as follows.

- This work is the first to consider the crosslingual setting for unsupervised SRL.
- We propose a form of agreement penalty and show its efficacy on English-German language pair when used in conjunction with a state-of-the-art non-parametric Bayesian model.
- We demonstrate that efficient approximate inference is feasible in the multilingual setting.

The rest of the paper is organized as follows. Section 2 begins with a definition of the crosslingual semantic role induction task we address in this paper. In Section 3, we describe the base monolingual model, and in Section 4 we propose an extension for the crosslingual setting. In Section 5, we describe our inference procedure. Section 6 provides both evaluation and analysis. Finally, additional related work is presented in Section 7.

2 Problem Definition

As we mentioned in the introduction, in this work we focus on the labeling stage of semantic role labeling. Identification, though an important problem, can be tackled with heuristics (Lang and Lapata, 2011a; Grenager and Manning, 2006; de Marneffe et al., 2006) or potentially by using a supervised classifier trained on a small amount of data.

Instead of assuming the availability of role annotated data, we rely only on automatically generated syntactic dependency graphs in both languages. While we cannot expect that syntactic structure can trivially map to a semantic representation[1], we can make use of syntactic cues. In the labeling stage, semantic roles are represented by clusters of arguments, and labeling a particular argument corresponds to deciding on its role cluster. However, instead of dealing with argument occurrences directly,

[1] Although it provides a strong baseline which is difficult to beat (Grenager and Manning, 2006; Lang and Lapata, 2010; Lang and Lapata, 2011a).

we represent them as predicate-specific syntactic signatures, and refer to them as *argument keys*. This representation aids our models in inducing high purity clusters (of argument keys) while reducing their granularity. We follow (Lang and Lapata, 2011a) and use the following syntactic features for English to form the argument key representation:

- Active or passive verb voice (`ACT/PASS`).
- Arg. position relative to predicate (`LEFT/RIGHT`).
- Syntactic relation to its governor.
- Preposition used for argument realization.

In the example sentences in Section 1, the argument keys for candidate arguments *Peter* for sentences (a) and (c) would be `ACT:LEFT:SBJ` and `PASS:RIGHT:LGS->by`,[2] respectively. While aiming to increase the purity of argument key clusters, this particular representation will not always produce a good match: e.g. *planning a theft* in sentence (b) will have the same key as *Mary* in sentence (a). Increasing the expressiveness of the argument key representation by using features of the syntactic frame would enable us to distinguish that pair of arguments. However, we keep this particular representation, in part to compare with the previous work. In German, we do not include the relative position features, because they are not very informative due to variability in word order.

In sum, we treat the unsupervised semantic role labeling task as clustering of argument keys. Thus, argument occurrences in the corpus whose keys are clustered together are assigned the same semantic role. The objective of this work is to improve argument key clusterings by inducing them simultaneously in two languages.

3 Monolingual Model

In this section we describe one of the Bayesian models for semantic role induction proposed in (Titov and Klementiev, 2012). Before describing our method, we briefly introduce the central components of the model: the Chinese Restaurant Processes (CRPs) and Dirichlet Processes (DPs) (Ferguson, 1973; Pitman, 2002). For more details we refer the reader to (Teh, 2007).

[2] `LGS` denotes a logical subject in a passive construction (Surdeanu et al., 2008).

3.1 Chinese Restaurant Processes

CRPs define probability distributions over partitions of a set of objects. An intuitive metaphor for describing CRPs is assignment of tables to restaurant customers. Assume a restaurant with a sequence of tables, and customers who walk into the restaurant one at a time and choose a table to join. The first customer to enter is assigned the first table. Suppose that when a client number i enters the restaurant, $i-1$ customers are sitting at each of the $k \in (1,\dots,K)$ tables occupied so far. The new customer is then either seated at one of the K tables with probability $\frac{N_k}{i-1+\alpha}$, where N_k is the number of customers already sitting at table k, or assigned to a new table with the probability $\frac{\alpha}{i-1+\alpha}$, $\alpha > 0$.

If we continue and assume that for each table every customer at a table orders the same meal, with the meal for the table chosen from an arbitrary base distribution H, then all ordered meals will constitute a sample from the Dirichlet Process $DP(\alpha, H)$.

An important property of the non-parametric processes is that a model designer does not need to specify the number of tables (i.e. clusters) a-priori as it is induced automatically on the basis of the data and also depending on the choice of the concentration parameter α. This property is crucial for our task, as the intended number of roles cannot possibly be specified for every predicate.

3.2 The Generative Story

In Section 2 we defined our task as clustering of argument keys, where each cluster corresponds to a semantic role. If an argument key k is assigned to a role r ($k \in r$), all of its occurrences are labeled r.

The Bayesian model encodes two common assumptions about semantic roles. First, it enforces the selectional restriction assumption: namely it stipulates that the distribution over potential argument fillers is sparse for every role, implying that 'peaky' distributions of arguments for each role r are preferred to flat distributions. Second, each role normally appears at most once per predicate occurrence. The inference algorithm will search for a clustering which meets the above requirements to the maximal extent.

The model associates two distributions with each predicate: one governs the selection of argument

fillers for each semantic role, and the other models (and penalizes) duplicate occurrence of roles. Each predicate occurrence is generated independently given these distributions. Let us describe the model by first defining how the set of model parameters and an argument key clustering are drawn, and then explaining the generation of individual predicate and argument instances. The generative story is formally presented in Figure 1.

For each predicate p, we start by generating a partition of argument keys B_p with each subset $r \in B_p$ representing a single semantic role. The partitions are drawn from CRP(α) independently for each predicate. The crucial part of the model is the set of selectional preference parameters $\theta_{p,r}$, the distributions of arguments x for each role r of predicate p. We represent arguments by lemmas of their syntactic heads.[3]

The preference for sparseness of the distributions $\theta_{p,r}$ is encoded by drawing them from the DP prior $DP(\beta, H^{(A)})$ with a small concentration parameter β, the base probability distribution $H^{(A)}$ is just the normalized frequencies of arguments in the corpus. The geometric distribution $\psi_{p,r}$ is used to model the number of times a role r appears with a given predicate occurrence. The decision whether to generate at least one role r is drawn from the uniform Bernoulli distribution. If 0 is drawn then the semantic role is not realized for the given occurrence, otherwise the number of additional roles r is drawn from the geometric distribution $Geom(\psi_{p,r})$. The Beta priors over ψ can indicate the preference towards generating at most one argument for each role.

Now, when parameters and argument key clusterings are chosen, we can summarize the remainder of the generative story as follows. We begin by independently drawing occurrences for each predicate. For each predicate role we independently decide on the number of role occurrences. Then each of the arguments is generated (see **GenArgument**) by choosing an argument key $k_{p,r}$ uniformly from the set of argument keys assigned to the cluster r, and finally choosing its filler $x_{p,r}$, where the filler is the lemma of the syntactic head of the argument.

[3]For prepositional phrases, the head noun of the object noun phrase is taken as it encodes crucial lexical information. However, the preposition is not ignored but rather encoded in the corresponding argument key, as explained in Section 2.

Clustering of argument keys:

for each predicate $p = 1, 2, \ldots$:
  $B_p \sim CRP(\alpha)$ [partition of arg keys]

Parameters:

for each predicate $p = 1, 2, \ldots$:
  for each role $r \in B_p$:
    $\theta_{p,r} \sim DP(\beta, H^{(A)})$ [distrib of arg fillers]
    $\psi_{p,r} \sim Beta(\eta_0, \eta_1)$ [geom distr for dup roles]

Data generation:

for each predicate $p = 1, 2, \ldots$:
  for each occurrence s of p:
    for every role $r \in B_p$:
      if $[n \sim Unif(0, 1)] = 1$: [role appears at least once]
        GenArgument(p, r) [draw one arg]
        while $[n \sim \psi_{p,r}] = 1$: [continue generation]
          GenArgument(p, r) [draw more args]

GenArgument(p, r):
  $k_{p,r} \sim Unif(1, \ldots, |r|)$ [draw arg key]
  $x_{p,r} \sim \theta_{p,r}$ [draw arg filler]

Figure 1: The generative story for predicate-argument structure.

4 Multilingual Extension

As we argued in Section 1, our goal is to penalize for disagreement in semantic structures predicted for each language on parallel data. In doing so, as in much of previous work on unsupervised induction of linguistic structures, we rely on automatically produced word alignments. In Section 6, we describe how we use word alignment to decide if two arguments are aligned; for now, we assume that (noisy) argument alignments are given.

Intuitively, when two arguments are aligned in parallel data, we expect them to be labeled with the same semantic role in both languages. This correspondence is simpler than the one expected in multilingual induction of syntax and morphology where systematic but unknown relation between structures in two language is normally assumed (e.g., (Snyder et al., 2008)). A straightforward implementation of this idea would require us to maintain one-to-one mapping between semantic roles across languages. Instead of assuming this correspondence, we penalize for the lack of isomorphism between the sets of roles in aligned predicates with the penalty dependent on the degree of violation. This softer approach

is more appropriate in our setting, as individual argument keys do not always deterministically map to gold standard roles[4] and strict penalization would result in the propagation of the corresponding over-coarse clusters to the other language. Empirically, we observed this phenomenon on the held-out set with the increase of the penalty weight.

Encoding preference for the isomorphism directly in the generative story is problematic: sparse Dirichlet priors can be used in a fairly trivial way to encode sparsity of the mapping in one direction or another but not in both. Instead, we formalize this preference with a penalty term similar to the expectation criteria in KL-divergence form introduced in McCallum et al. (2007). Specifically, we augment the joint probability with a penalty term computed on parallel data:

$$\sum_{p^{(1)},\ p^{(2)}} \Big(-\gamma^{(1)} \sum_{r^{(1)} \in B_{p^{(1)}}} f_{r^{(1)}} \underset{r^{(2)} \in B_{p^{(2)}}}{\arg\max} \log \hat{P}(r^{(2)}|r^{(1)}) \\ -\gamma^{(2)} \sum_{r^{(2)} \in B_{p^{(2)}}} f_{r^{(2)}} \underset{r^{(1)} \in B_{p^{(1)}}}{\arg\max} \log \hat{P}(r^{(1)}|r^{(2)}) \Big),$$

where $\hat{P}(r^{(l)}|r^{(l')})$ is the proportion of times the role $r^{(l')}$ of predicate $p^{(l')}$ in language l' is aligned to the role $r^{(l)}$ of predicate $p^{(l)}$ in language l, and $f_{r^{(l)}}$ is the total number of times the role is aligned, $\gamma^{(l)}$ is a non-negative constant. The rationale for introducing the individual weighting $f_{r^{(l)}}$ is two-fold. First, the proportions $\hat{P}(r^{(l)}|r^{(l')})$ are more 'reliable' when computed from larger counts. Second, more frequent roles should have higher penalty as they compete with the joint probability term, the likelihood part of which scales linearly with role counts.

Space restrictions prevent us from discussing the close relation between this penalty formulation and the existing work on injecting prior and side information in learning objectives in the form of constraints (McCallum et al., 2007; Ganchev et al., 2010; Chang et al., 2007).

In order to support efficient and parallelizable inference, we simplify the above penalty by considering only disjoint pairs of predicates, instead of summing over all pairs $p^{(1)}$ and $p^{(2)}$. When choosing the pairs, we aim to cover the maximal number of alignment counts so as to preserve as much information from parallel corpora as possible. This objective corresponds to the classic maximum weighted bipartite matching problem with the weight for each edge $p^{(1)}$ and $p^{(2)}$ equal to the number of times the two predicates were aligned in parallel data. We use the standard polynomial algorithm (the Hungarian algorithm, (Kuhn, 1955)) to find an optimal solution.

5 Inference

An inference algorithm for an unsupervised model should be efficient enough to handle vast amounts of unlabeled data, as it can easily be obtained and is likely to improve results. We use a simple approximate inference algorithm based on greedy search. We start by discussing search for the maximum a-posteriori clustering of argument keys in the monolingual set-up and then discuss how it can be extended to accommodate the role alignment penalty.

5.1 Monolingual Setting

In the model, a linking between syntax and semantics is induced independently for each predicate. Nevertheless, searching for a MAP clustering can be expensive: even a move involving a single argument key implies some computations for all its occurrences in the corpus. Instead of more complex MAP search algorithms (see, e.g., (Daume III, 2007)), we use a greedy procedure where we start with each argument key assigned to an individual cluster, and then iteratively try to merge clusters. Each move involves (1) choosing an argument key and (2) deciding on a cluster to reassign it to. This is done by considering all clusters (including creating a new one) and choosing the most probable one.

Instead of choosing argument keys randomly at the first stage, we order them by corpus frequency. This ordering is beneficial as getting clustering right for frequent argument keys is more important and the corresponding decisions should be made earlier.[5] We used a single iteration in our experiments, as we have not noticed any benefit from using multiple iterations.

[4] The average purity for argument keys with automatic argument identification and using predicted syntactic trees, before any clustering, is approximately 90.2% on English and 87.8% on German.

[5] This has been explored before for shallow semantic representations (Lang and Lapata, 2011a; Titov and Klementiev, 2011).

5.2 Incorporating the Alignment Penalty

Inference in the monolingual setting is done independently for each predicate, as the model factorizes over the predicates. The role alignment penalty introduces interdependencies between the objectives for each bilingual predicate pair chosen by the assignment algorithm as discussed in Section 4. For each pair of predicates, we search for clusterings to maximize the sum of the log-probability and the negated penalty term.

At first glance it may seem that the alignment penalty can be easily integrated into the greedy MAP search algorithm: instead of considering individual argument keys, one could use pairs of argument keys and decide on their assignment to clusters jointly. However, given that there is no isomorphic mapping between argument keys across languages, this solution is unlikely to be satisfactory.[6] Instead, we use an approximate inference procedure similar in spirit to annotation projection techniques.

For each predicate, we first induce semantic roles independently for the first language, as described in Section 5.1, and then use the same algorithm for the second language but take the penalty term into account. Then we repeat the process in the reverse direction. Among these two solutions, we choose the one which yields the higher objective value. In this way, we begin with producing a clustering for the side which is easier to cluster and provides more clues for the other side.[7]

6 Empirical Evaluation

We begin by describing the data and evaluation metrics we use before discussing results.

6.1 Data

We run our main experiments on the English-German section of Europarl v6 parallel corpus (Koehn, 2005) and the CoNLL 2009 distributions of the Penn Treebank WSJ corpus (Marcus et al., 1993) for English and the SALSA corpus (Burchardt et al., 2006) for German. As standard for unsupervised SRL, we use the entire CoNLL training sets for evaluation, and use held-out sets for model selection and parameter tuning.

Syntactic annotation. Although the CoNLL 2009 dataset already has predicted dependency structures, we could not reproduce them so that we could use the same parser to annotate Europarl. We chose to reannotate it, since using different parsing models for both datasets would be undesirable. We used MaltParser (Nivre et al., 2007) for English and the syntactic component of the LTH system (Johansson and Nugues, 2008) for German.

Predicate and argument identification. We select all non-auxiliary verbs as predicates. For English, we identify their arguments using a heuristic proposed in (Lang and Lapata, 2011a). It is comprised of a list of 8 rules, which use nonlexicalized properties of syntactic paths between a predicate and a candidate argument to iteratively discard non-arguments from the list of all words in a sentence. For German, we use the LTH argument identification classifier. Accuracy of argument identification on CoNLL 2009 using predicted syntactic analyses was 80.7% and 86.5% for English and German, respectively.

Argument alignment. We use GIZA++ (Och and Ney, 2003) to produce word alignments in Europarl: we ran it in both directions and kept the intersection of the induced word alignments. For every argument identified in the previous stage, we chose a set of words consisting of the argument's syntactic head and, for prepositional phrases, the head noun of the object noun phrase. We mark arguments in two languages as aligned if there is any word alignment between the corresponding sets and if they are arguments of aligned predicates.

6.2 Evaluation Metrics

We use the standard purity (PU) and collocation (CO) metrics as well as their harmonic mean (F1) to measure the quality of the resulting clusters. Purity measures the degree to which each cluster contains arguments sharing the same gold role:

[6] We also considered a variation of this idea where a pair of argument keys is chosen randomly proportional to their alignment frequency and multiple iterations are repeated. Despite being significantly slower than our method, it did not provide any improvement in accuracy.

[7] In preliminary experiments, we studied an even simpler inference method where the projection direction was fixed for all predicates. Though this approach did outperform the monolingual model, the results were substantially worse than achieved with our method.

$$PU = \frac{1}{N} \sum_i \max_j |G_j \cap C_i|$$

where C_i is the set of arguments in the i-th induced cluster, G_j is the set of arguments in the jth gold cluster, and N is the total number of arguments. Collocation evaluates the degree to which arguments with the same gold roles are assigned to a single cluster:

$$CO = \frac{1}{N} \sum_j \max_i |G_j \cap C_i|$$

We compute the aggregate PU, CO, and F1 scores over all predicates in the same way as (Lang and Lapata, 2011a) by weighting the scores of each predicate by the number of its argument occurrences. Since our goal is to evaluate the clustering algorithms, we *do not* include incorrectly identified arguments when computing these metrics.

6.3 Parameters and Set-up

Our models are robust to parameter settings; the parameters were tuned (to an order of magnitude) to optimize the $F1$ score on the held-out development set and were as follows. Parameters governing duplicate role generation, $\eta_0^{(\cdot)}$ and $\eta_1^{(\cdot)}$, and penalty weights $\gamma^{(\cdot)}$ were set to be the same for both languages, and are 100, $1.e$-3 and 10, respectively. The concentration parameters were set as follows: for English, they were set to $\alpha^{(1)} = 1.e$-3, $\beta^{(1)} = 1.e$-3, and, for German, they were $\alpha^{(2)} = 0.1$, $\beta^{(2)} = 1$.

Domains of Europarl (parliamentary proceedings) and German/English CoNLL data (newswire) are substantially different. Since the influence of domain shift is not the focus of work, we try to minimize its effect by computing the likelihood part of the objective on CoNLL data alone. This also makes our setting more comparable to prior work.[8]

6.4 Results

Base monolingual model. We begin by evaluating our base monolingual model *MonoBayes* alone against the current best approaches to unsupervised semantic role induction. Since we do not have access to the systems, we compare on the marginally different English CoNLL 2008 (Surdeanu et al., 2008) shared task dataset used in their experiments. We report the results using gold argument identification and gold syntactic parses in order to focus the evaluation on the argument labeling stage and to minimize the noise due to automatic syntactic annotations. The methods are Latent Logistic classification (Lang and Lapata, 2010), Split-Merge clustering (Lang and Lapata, 2011a), and Graph Partitioning (Lang and Lapata, 2011b) (labeled *LLogistic*, *SplitMerge*, and *GraphPart*, respectively) achieving the current best unsupervised SRL results in this setting. Additionally, we compute the syntactic function baseline (*SyntF*), which simply clusters predicate arguments according to the dependency relation to their head. Following (Lang and Lapata, 2010), we allocate a cluster for each of 20 most frequent relations in the CoNLL dataset and one cluster for all other relations. Our model substantially outperforms other models (see Table 1).

	PU	CO	F1
LLogistic	79.5	76.5	78.0
GraphPart	88.6	70.7	78.6
SplitMerge	88.7	73.0	80.1
MonoBayes	88.1	77.1	**82.2**
SyntF	81.6	77.5	79.5

Table 1: Argument clustering performance with *gold argument identification* and *gold syntactic parses* on CoNLL 2008 shared-task dataset. Bold-face is used to highlight the best F1 scores.

Multilingual extensions. Next, we improve our model performance using agreement as an additional supervision signal during training (see Section 4). We compare the performance of individual English and German models induced separately (*MonoBayes*) with the jointly induced models (*MultiBayes*) as well as the syntactic baseline, see Table 2.[9] While we see little improvement in F1 for English, the German system improves by 1.8%. For German, the crosslingual learning also results in 1.5% improvement over the syntactic baseline, which is considered difficult to outperform (Grenager and Manning, 2006; Lang and Lapata, 2010). Note that recent unsupervised SRL meth-

[8] Preliminary experiments on the entire dataset show a slight degradation in performance.

[9] Note that the scores are computed on correctly identified arguments only, and tend to be higher in these experiments probably because the complex arguments get discarded by the argument identifier.

	English			German		
	PU	CO	F1	PU	CO	F1
MonoBayes	87.5	80.1	83.6	86.8	75.7	80.9
MultiBayes	86.8	80.7	**83.7**	85.0	80.6	**82.7**
SyntF	81.5	79.4	80.4	83.1	79.3	81.2

Table 2: Results on CoNLL 2009 with *automatic argument identification* and *automatic syntactic parses*.

ods do not always improve on it, see Table 1.

The relatively low expressivity and limited purity of our argument keys (see discussion in Section 4) are likely to limit potential improvements when using them in crosslingual learning. The natural next step would be to consider crosslingual learning with a more expressive model of the syntactic frame and syntax-semantics linking.

7 Related Work

Unsupervised learning in crosslingual setting has been an active area of research in recent years. However, most of this research has focused on induction of syntactic structures (Kuhn, 2004; Snyder et al., 2009) or morphologic analysis (Snyder and Barzilay, 2008) and we are not aware of any previous work on induction of semantic representations in the crosslingual setting. Learning of semantic representations in the context of monolingual weakly-parallel data was studied in Titov and Kozhevnikov (2010) but their setting was semi-supervised and they experimented only on a restricted domain.

Most of the SRL research has focused on the supervised setting, however, lack of annotated resources for most languages and insufficient coverage provided by the existing resources motivates the need for using unlabeled data or other forms of weak supervision. This includes methods based on graph alignment between labeled and unlabeled data (Fürstenau and Lapata, 2009), using unlabeled data to improve lexical generalization (Deschacht and Moens, 2009), and projection of annotation across languages (Pado and Lapata, 2009; van der Plas et al., 2011). Semi-supervised and weakly-supervised techniques have also been explored for other types of semantic representations but these studies again have mostly focused on restricted domains (Kate and Mooney, 2007; Liang et al., 2009; Goldwasser et al., 2011; Liang et al., 2011).

Early unsupervised approaches to the SRL task include (Swier and Stevenson, 2004), where the VerbNet verb lexicon was used to guide unsupervised learning, and a generative model of Grenager and Manning (2006) which exploits linguistic priors on syntactic-semantic interface.

More recently, the role induction problem has been studied in Lang and Lapata (2010) where it has been reformulated as a problem of detecting alternations and mapping non-standard linkings to the canonical ones. Later, Lang and Lapata (2011a) proposed an algorithmic approach to clustering argument signatures which achieves higher accuracy and outperforms the syntactic baseline. In Lang and Lapata (2011b), the role induction problem is formulated as a graph partitioning problem: each vertex in the graph corresponds to a predicate occurrence and edges represent lexical and syntactic similarities between the occurrences. Unsupervised induction of semantics has also been studied in Poon and Domingos (2009) and Titov and Klementiev (2011) but the induced representations are not entirely compatible with the PropBank-style annotations and they have been evaluated only on a question answering task for the biomedical domain. Also, a related task of unsupervised argument identification has been considered in Abend et al. (2009).

8 Conclusions

This work adds unsupervised semantic role labeling to the list of NLP tasks benefiting from the crosslingual induction setting. We show that an agreement signal extracted from parallel data provides indirect supervision capable of substantially improving a state-of-the-art model for semantic role induction.

Although in this work we focused primarily on improving performance for each individual language, cross-lingual semantic representation could be extracted by a simple post-processing step. In future work, we would like to model cross-lingual semantics explicitly.

Acknowledgements

The work was supported by the MMCI Cluster of Excellence and a Google research award. The authors thank Mikhail Kozhevnikov, Alexis Palmer, Manfred Pinkal, Caroline Sporleder and the anonymous reviewers for their suggestions.

References

Omri Abend, Roi Reichart, and Ari Rappoport. 2009. Unsupervised argument identification for semantic role labeling. In *ACL-IJCNLP*.

Roberto Basili, Diego De Cao, Danilo Croce, Bonaventura Coppola, and Alessandro Moschitti. 2009. Cross-language frame semantics transfer in bilingual corpora. In *CICLING*.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Pado, and M. Pinkal. 2006. The SALSA corpus: a german corpus resource for lexical semantics. In *LREC*.

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

Hal Daume III. 2007. Fast search for dirichlet process mixture models. In *AISTATS*.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *LREC 2006*.

Koen Deschacht and Marie-Francine Moens. 2009. Semi-supervised semantic role labeling using the Latent Words Language Model. In *EMNLP*.

Thomas S. Ferguson. 1973. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Hagen Fürstenau and Mirella Lapata. 2009. Graph alignment for semi-supervised semantic role labeling. In *EMNLP*.

Kuzman Ganchev, Joao Graca, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research (JMLR)*, 11:2001–2049.

Qin Gao and Stephan Vogel. 2011. Corpus expansion for statistical machine translation with semantic role label substitution rules. In *ACL:HLT*.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labelling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Dan Goldwasser, Roi Reichart, James Clarke, and Dan Roth. 2011. Confidence driven unsupervised semantic parsing. In *ACL*.

Trond Grenager and Christoph Manning. 2006. Unsupervised discovery of a statistical verb lexicon. In *EMNLP*.

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 *CoNLL 2009: Shared Task*.

Richard Johansson and Pierre Nugues. 2008. Dependency-based semantic role labeling of PropBank. In *EMNLP*.

Michael Kaisser and Bonnie Webber. 2007. Question answering based on semantic roles. In *ACL Workshop on Deep Linguistic Processing*.

Rohit J. Kate and Raymond J. Mooney. 2007. Learning language semantics from ambigous supervision. In *AAAI*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit*.

Harold W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

Jonas Kuhn. 2004. Experiments in parallel-text based grammar induction. In *ACL*.

Joel Lang and Mirella Lapata. 2010. Unsupervised induction of semantic roles. In *ACL*.

Joel Lang and Mirella Lapata. 2011a. Unsupervised semantic role induction via split-merge clustering. In *ACL*.

Joel Lang and Mirella Lapata. 2011b. Unsupervised semantic role induction with graph partitioning. In *EMNLP*.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *ACL-IJCNLP*.

Percy Liang, Michael Jordan, and Dan Klein. 2011. Learning dependency-based compositional semantics. In *ACL: HLT*.

Ding Liu and Daniel Gildea. 2010. Semantic role features for machine translation. In *Coling*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Andrew McCallum, Gideon Mann, and Gregory Druck. 2007. Generalized expectation criteria. Technical Report TR 2007-60, University of Massachusetts, Amherst, MA.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *EMNLP*.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *EMNLP-CoNLL*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51.

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

Jim Pitman. 2002. Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, 11:501–514.

Hoifung Poon and Pedro Domingos. 2009. Unsupervised semantic parsing. In *EMNLP*.

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

M. Sammons, V. Vydiswaran, T. Vieira, N. Johri, M. Chang, D. Goldwasser, V. Srikumar, G. Kundu, Y. Tu, K. Small, J. Rule, Q. Do, and D. Roth. 2009. Relation alignment for textual entailment recognition. In *Text Analysis Conference (TAC)*.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *EMNLP*.

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *ACL*.

Benjamin Snyder and Regina Barzilay. 2010. Climbing the tower of Babel: Unsupervised multilingual learning. In *ICML*.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *EMNLP*.

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised multilingual grammar induction. In *ACL*.

Mihai Surdeanu, Adam Meyers Richard Johansson, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Shared Task*.

Richard Swier and Suzanne Stevenson. 2004. Unsupervised semantic role labelling. In *EMNLP*.

Yee Whye Teh. 2007. Dirichlet process. *Encyclopedia of Machine Learning*.

Ivan Titov and Alexandre Klementiev. 2011. A Bayesian model for unsupervised semantic parsing. In *ACL*.

Ivan Titov and Alexandre Klementiev. 2012. A Bayesian approach to unsupervised semantic role induction. In *EACL*.

Ivan Titov and Mikhail Kozhevnikov. 2010. Bootstrapping semantic analyzers from non-contradictory texts. In *ACL*.

Lonneke van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *ACL*.

Dekai Wu and Pascale Fung. 2009. Semantic roles for SMT: A hybrid two-pass model. In *NAACL*.

Dekai Wu, Marianna Apidianaki, Marine Carpuat, and Lucia Specia, editors. 2011. *Proc. of Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*. ACL.