Semi-Supervised Semantic Role Labeling: Approaching from an Unsupervised Perspective

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

ABSTRACT
Reducing the reliance of semantic role labeling (SRL) methods on human-annotated data has become an active area of research. However, the prior work has largely focused on either (1) looking into ways to improve supervised SRL systems by producing surrogate annotated data and reducing sparsity of lexical features or (2) considering completely unsupervised semantic role induction settings. In this work, we aim to link these two veins of research by studying how unsupervised techniques can be improved by exploiting small amounts of labeled data. We extend a state-of-the-art Bayesian model for unsupervised semantic role induction to better accommodate for annotated sentences. Our semi-supervised method outperforms a strong supervised baseline when only a small amount of labeled data is available.

KEYWORDS: semantic role labeling, semi-supervised learning, shallow semantics, Bayesian model.

Proceedings of COLING 2012: Technical Papers, pages 2635–2652,
COLING 2012, Mumbai, December 2012.

1 Introduction

Shallow representations of meaning, and semantic role labels in particular, have a long history in linguistics (Fillmore, 1968). More recently, with the emergence of large annotated resources such as PropBank (Palmer et al., 2005) and FrameNet (Baker et al., 1998), automatic semantic role labeling (SRL) has attracted a lot of attention (Gildea and Jurafsky, 2002; Carreras and Màrquez, 2005; Surdeanu et al., 2008; Hajič et al., 2009).

SRL representations encode the underlying predicate-argument structure of sentences, or, more specifically, for every predicate in a sentence they identify a set of arguments and associate each argument with an underlying *semantic role*, such as an agent (an initiator or doer of the action) or a patient (an affected entity). SRL representations have many potential applications in NLP and have 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., 2009), among others.

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). Scarcity of annotated data has motivated the research into techniques capable of exploiting unlabeled data, that is, semi-supervised and unsupervised learning.

The existing semi-supervised approaches to SRL can largely be regarded as extensions to supervised techniques, as they use supervised learning as sub-routines in the estimation process. These include self-training and co-training methods (He and Gildea, 2006b; Lee et al., 2007; Kaljahi and Samad, 2010), mono-lingual and cross-lingual annotation projection (Fürstenau and Lapata, 2009; Pado and Lapata, 2009; van der Plas et al., 2011), and methods which exploit or induce word representations to reduce the sparsity of lexicalized features (He and Gildea, 2006a; Deschacht and Moens, 2009; Collobert et al., 2011). Most of these approaches, especially the bootstrapping-style methods (He and Gildea, 2006b; Lee et al., 2007; Kaljahi and Samad, 2010; Fürstenau and Lapata, 2009), have achieved minimal or even no improvement from using unlabeled data. Consequently, the development of effective semi-supervised techniques remains an important and largely unresolved problem.

Another vein of research exploiting unlabeled data for shallow semantic parsing has focused on purely unsupervised set-ups (Swier and Stevenson, 2004; Grenager and Manning, 2006; Lang and Lapata, 2010, 2011a,b; Titov and Klementiev, 2012; Garg and Henderson, 2012; Fürstenau and Rambow, 2012). The unsupervised setting is important in itself, and the development of these methods arguably provides interesting insights into modeling implicit supervision signals present in unlabeled data. However, given that small amounts of labeled data are often easy to obtain, it is surprising that no previous work that we are aware of looked into integration of labeled data into unsupervised SRL systems.[1] Moreover, due to the inherent difference in the clustering metrics used for unsupervised SRL and the labeled accuracy scores used to evaluate supervised SRL methods, they have so far never been properly compared. These are the gaps addressed by this paper.

In this work, we show how a state-of-the-art unsupervised Bayesian model (BayesSRL) (Titov

[1] This semi-supervised learning setting is sometimes referred to as semi-*un*supervised (Daumé III, 2009).

and Klementiev, 2012) can be used in a semi-supervised set-up. BayesSRL is especially appropriate for our study as it automatically induces a common representation encoding properties of the syntax-semantics interface that are valid across predicates, contrasting much of other research on unsupervised SRL where separate models were induced for each predicate (Grenager and Manning, 2006; Lang and Lapata, 2010, 2011a,b; Garg and Henderson, 2012; Fürstenau and Rambow, 2012). These models would not be able exploit sparse labeled data effectively, as they would essentially split this scarce data into even smaller (and often empty) training sets.

A straightforward way of integrating labeled data into learning of a generative model would amount to maximizing joint probability of labeled and unlabeled data. However, due to hard constraints in the BayesSRL model and the great disbalance between the amount of labeled and unlabeled data, we argue that a different approach is preferred. Namely, we use labeled data to construct an informed prior over the potential semantic representations and also modify the model to integrate the labels as soft constraints on admissible semantic structures.

We compare the semi-supervised approach we propose to a state-of-the-art supervised method (Johansson and Nugues, 2008a). Though the BayesSRL model exploits a cross-predicate representation, it does not align roles across predicates which prevents us from using supervised evaluation metrics. Consequently, we evaluate the methods using clustering measures: the harmonic mean of purity and collocation, a common metric for unsupervised SRL evaluation (Lang and Lapata, 2010), and the information-theoretic V-Measure (Rosenberg and Hirschberg, 2007).

The semi-supervised method outperforms its supervised counterpart when the amount of labeled data is small. Unsurprisingly, it does not fare as well when the amount of data increases. We believe that this is primarily due to the overly coarse modeling of the syntax-semantics interface, as it is optimized for the unsupervised setting. Nevertheless, these results strongly suggest that approaching the semi-supervised learning setting for SRL from an unsupervised perspective is a promising research direction and that the existing unsupervised SRL methods are already mature enough to be useful for low resource languages with little or no labeled data available.

2 Background

In this section, we begin by formally defining the semantic role labeling task, and then discuss the distance-dependent Chinese Restaurant process (Blei and Frazier, 2011), used as a component in the BayesSRL model and crucial for effective learning in the semi-supervised setting. We conclude the section with a short description of the BayesSRL model.

2.1 Task Definition

The SRL task involves prediction of predicate argument structure, i.e. both identification of arguments as well as assignment of labels according to their underlying semantic role. For example, in the following sentences:

(a) [$_{A0}$ Mary] opened [$_{A1}$ the door].
(b) [$_{A1}$ The door] opened.
(c) [$_{A1}$ The door] was opened [$_{A0}$ by Mary].

Mary always takes an agent role (*A*0 in the PropBank notation (Palmer et al., 2005)) for the predicate *open*, and *door* is always a patient (*A*1).

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), with unsupervised techniques (Abend et al., 2009) or potentially by using a supervised classifier trained on a small amount of data. In our experiments we use the heuristic identifier of Lang and Lapata (2011a). Also, as in much of the previous work on supervised and unsupervised SRL, we rely on automatically generated syntactic dependency trees.

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, in BayesSRL they are represented as predicate-specific syntactic signatures, called *argument keys*. The following syntactic features are used to form the argument key representation:

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

In the above example, the argument keys for candidate arguments *Mary* 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. *door* in sentence (b) will have the same key as *Mary* in sentence (a). Consequently, this introduces an upper bound on the model performance: in our experimental set-up the upper bound on the purity of clustering was equal to 91.7%.

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 representation, in part to compare with previous work and in part because we are primarily interested in set-ups with little annotated data where this upper bound would not be as limiting.

The clustering implicitly defines the set of permissible *alternations*, or changes in the syntactic realization of the argument structure of the verb. For example, passivization can be roughly represented with the clustering of the key `ACT:LEFT:SBJ` with `PASS:RIGHT:LGS->by` and `ACT:RIGHT:OBJ` with `PASS:LEFT:SBJ`.

In sum, BayesSRL treats the unsupervised semantic role labeling task as clustering of argument keys. Thus, argument occurrences in the corpus whose keys are clustered together are assigned to the same semantic role. The objective of this work is to study how argument key clusterings can be improved by using small amounts of annotated data.

2.2 Distance-dependent CRPs

The Chinese Restaurant Process (CRP), a standard component in non-parametric Bayesian modeling, defines a probability distributions over partitions of a set of objects. It encodes general rich-get-richer dynamics and, as such, is often useful in modeling long tail distributions. CRPs do not distinguish between individual objects and, consequently, prior probability that two objects would end up in the same subset is constant for any choice of objects. Distant-dependent

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

CRPs (dd-CRPs) (Blei and Frazier, 2011) use a similarity function d_{ij} in generating partitions: they prefer to place pairs (i, j) with larger similarity d_{ij} in a single subset. More formally, each object i chooses itself a partner c_i with the probability

$$p(c_i = j|D, \alpha) \propto \begin{cases} d_{ij}, & i \neq j \\ \alpha, & i = j \end{cases} \tag{1}$$

where α is a non-negative concentration parameter. The resulting partition is defined by connected components in the directed graph encoded by the partnership relation c. Unlike normal CRP, dd-CRP lacks the exchangeability property and the probability of a given partition cannot be efficiently computed. Nevertheless, efficient inference is possible with MCMC techniques or approximate MAP search methods.

The prior is invariant under joint rescaling of the concentration parameter and the similarity scores, and the proportion of the concentration parameter to the distance parameters can be regarded as a parameter controlling granularity of clustering. We use a slight extension to the original dd-CRP by allowing the concentration parameter to be different for every example and, therefore, write it as $\alpha_i = d_{ii}$.

The similarities D can be fixed and used to encode prior knowledge about the problem (Blei and Frazier, 2011; Socher et al., 2011; Duan et al., 2007; Jensen and Shore, 2011) or can be induced automatically by sharing them across several instances of the clustering problem in a multi-task setting (Titov and Klementiev, 2012). In this work, as discussed in Section 3.2, we use the dd-CRP priors to fill both of these roles.

2.3 BayesSRL Model

In this section we describe the Bayesian model which we use as a basis for our semi-supervised learning approach. For more detailed and formal description of the model we refer the reader to Titov and Klementiev (2012). In this work we use the *coupled* version of the BayesSRL model, that is the model which induces cross-predicate representations.

In Section 2.1 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.

As we argued in Section 2.1, clusterings of argument keys implicitly encode the pattern of alternations for a predicate. The set of permissible alternations is predicate-specific,[3] but still most of the alternation are shared across several or many predicates (e.g., passivization or dativization). Consequently, BayesSRL regards semantic role induction as a multi-task clustering problem and encodes the relative 'popularity' of alternations by quantifying how likely a pair of keys is to be clustered. These scores (d_{ij} for every pair of argument keys i and j) are induced automatically within the model, and treated as latent variables shared across predicates.

[3] Or, at least specific to a class of predicates (Levin, 1993).

```
Parameters:

$D \sim NonInform$                                                    [similarity graph]
for each predicate $p = 1, 2, \ldots$:
  $B_p \sim dd\text{-}CRP(\alpha, D)$                                 [partition of arg keys]
  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 BayesSRL model.

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.

The generation starts by choosing a graph D with non-negative weights $d_{i,j}$ on edges from a non-informative prior, in other words, uniformly over the space of such graphs. Then for each predicate p, a partition of argument keys B_p is drawn from a distance-dependent Chinese Restaurant Process dd-CRP(α, D), with each subset $r \in B_p$ representing a single semantic role.

Next, the parameters are generated from the corresponding prior distributions. For details, we refer the reader to Titov and Klementiev (2012).

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.

In sum, the properties of the BayesSRL model most relevant to the discussion of the semi-supervised extension are (1) induction of predicate-specific hard clustering of argument keys and (2) learning of a cross-predicate similarity measure D over pairs of argument keys.

```
**GenArgument**$(p, r)$:
  $b \sim Bernoulli(\epsilon)$
  if $b = 1$:
    $k_{p,r} \sim H^{(K)}$                                  [noisy arg key]
  else
    $k_{p,r} \sim Unif(1, \ldots, |r|)$                     [true arg key]
    $x_{p,r} \sim \theta_{p,r}$                             [draw arg filler]
```

Figure 2: A modified model of argument generation.

3 Semi-Supervised Extension

In this section, we discuss two ways that the labeled data can be exploited in estimating the BayesSRL model. In practice, we found that their combination yields the best result.

3.1 Adding Labels

The integration of labeled data in a generative model is usually trivial and amounts to maximizing the joint likelihood of the observable data. In practice, it implies that the observable labels will be clamped in the estimation process. The straightforward application of this idea to our set-up is problematic. The BayesSRL method makes hard decisions about the clustering of argument keys, and, given the imperfect purity of argument keys and potential annotation errors, no single clustering would be entirely compatible with the labeled data, resulting in zero probability for any model state. Intuitively, one would want to relax this compatibility assumption by allowing for some inconsistency between induced clusterings and labeled data, while still favoring more compatible configurations.

A standard trick to achieve this behavior within the generative framework is to assume that with some small probability ϵ the true outcome is substituted with a random pick. The parameter ϵ would serve as a penalty for inconsistency, the smaller the probability ϵ, the more severe is the penalty. In our case, it translates into modifying the **GenArgument**(p, r) by introducing the possibility of drawing the random argument key from some base distribution $H^{(K)}$, instead of choosing it from the set of keys associated with r (See Figure 2). We use the normalized counts of argument keys in the corpus as the base distribution $H^{(K)}$.

Labeled data integrated in this relaxed BayesSRL model would affect the induced shared prior D and, consequently, the information present in the labeled data would be propagated across different predicates. Unfortunately, there are two problems with using this approach which negatively affect practical results.

The first deficiency is connected with the fact that in practice the amount of unlabeled data vastly exceeds the amount of labeled data nullifying the effect of the latter during estimation. A standard heuristic approach to mitigate this deficiency is to reweigh the data to put an extra emphasis on the labeled part. This technique is unlikely to be very effective here as the argument key clusterings are drawn from dd-CRP(α, D) once for each predicate, not once per predicate occurrence, and the proportion of predicates in labeled and unlabeled data would remain unaffected by instance reweighting.

Another problem is more subtle. As discussed in Titov and Klementiev (2012), their method induces the pairwise clustering preferences D but does not attempt to learn the concentration parameters α_k and also enforces a form of normalization on the pairwise similarity, effectively 'freezing' the granularity of clusterings. This is fairly natural for an unsupervised setting where the model designer should have some form of control over the granularity but not as desirable in the semi-supervised setting where the granularity should be learned from the annotated data. In fact, as we will see in Section 4, labeled data mostly provides evidence for combining clusters (thus increasing collocation), and, consequently, the ability to learn granularity is crucial.

Thus, a compromise is necessary between (a) learning the granularity from labeled data and (b) limiting the influence of unlabeled data on cluster granularity. We implement this idea by using annotated examples to construct an informed prior.

3.2 Constructing Informed Priors

An alternative approach to directly incorporating the labeled data in the objective function would be to use the data to define an informed 'prior' over argument key clusterings. To this end, we estimate from the labeled data how likely argument keys k and k' are to belong to the same role and how likely a specific key k is to be left unclustered. We use the former to set the similarity $\hat{d}_{kk'}$, and the latter to set the concentration parameter α_k for the dd-CRP model. More precisely, we estimate both the predicate-specific similarities $\hat{d}^{(p)}$ and the cross-predicate similarities $\hat{d}$. When generating partitions B_p (see Figure 1), we multiply $\hat{d}^{(p)}$, $\hat{d}$, and the automatically induced prior d and use the resulting combined similarity in the dd-CRP process. The concentration parameters are combined in the same way as similarities. This techniques corresponds to the standard product-of-expert combination approach (Hinton, 2002). The remaining part of the section describes this idea more formally.

Initially we will consider individual predicates and then we will generalize the approach to cross-predicate similarities. Consider a predicate p, and assume that we have K different argument keys and R different roles,[4] and that each argument key k appears N_k times in the labeled data, and is annotated $N_{k,r}$ times with role r. In order to estimate the required probabilities we need to make assumptions about the joint generation of labels and argument keys.

We assume that there exists a fixed latent mapping g from argument keys to semantic roles and any such mapping is a-priori equiprobable, $P(g) = const$. However, when generating a label $g(k)$ for a key k, we assume that it can be replaced by any of the remaining $R-1$ roles with small probability γ. The probability of the set of labeled examples X_k associated with the key k given a mapping g can be written as

$$P(X_k|g(k)=r) = (1-\gamma)^{N_{k,r}} \left(\frac{\gamma}{R-1}\right)^{N_k - N_{k,r}}.$$

The joint probability of the sets of labeled examples X_k and $X_{k'}$ under the assumptions that either (1) the two keys belong to the same (any) role or (2) belong to two different roles can

[4] In our experiments we set R to 21, the number of distinct roles in PropBank, and K to the number of argument keys appearing both in labeled and unlabeled data for the considered predicate.

be computed by summing over the roles:[5]

$$\begin{aligned}&P(X_k, X_{k'}|g(k) = g(k'))\\&\quad = \sum_r P(X_k|g(k) = r)P(X_{k'}|g(k') = r)\\&P(X_k, X_{k'}|g(k) \neq g(k'))\\&\quad = \sum_r P(X_k|g(k) = r)\sum_{r' \neq r} P(X_{k'}|g(k') = r')\end{aligned}$$

The posterior probability that two keys belong to the same role $P(g(k) = g(k')|X)$, where X is the entire labeled dataset, is given by renormalizing the two likelihoods above. As the distance $d^{(p)}_{kk'}$ in dd-CRP essentially encodes how much more likely the two keys are clustered together than by random chance, we compute the similarity as

$$\hat{d}^{(p)}_{kk'} = \frac{P(g(k) = g(k')|X)}{P(g(k) = g(k'))}, \tag{2}$$

where $P(g(k) = g(k'))$ is the prior probability that two keys are labeled with the same role, equal to $1/R$.

A very similar algebra is used to derive the probability that an argument key k is the only key assigned to some role $P(\nexists k', k \neq k' : g(k) = g(k')|X)$. The concentration parameter $\hat{\alpha}_k$ is set to

$$\hat{\alpha}^{(p)}_k = \frac{P(\nexists k', k \neq k' : g(k) = g(k')|X)}{P(\nexists k', k \neq k' : g(k) = g(k'))}, \tag{3}$$

where the denominator is the prior probability of not sharing the role with any other argument key, $(R-1)^{k-1}/R^{k-1}$. Note that if no labeled data is available for the considered predicate p, equations (2) and (3) would yield 1 and, as desired, the prior would not affect prediction of other experts in the product-of-expert combination.

The above approach induces predicate specific priors but this is insufficient for all but very frequent predicates. Consequently, we use a similar approach to define cross-predicate similarities $\hat{d}$ but with a larger γ', thus penalizing less severely for violations. For the cross-predicate similarities, the assumption is that (independently over pairs of keys) each pair of keys either shares a role in all the predicates or the two keys are labeled with a different role in all the predicates. This implies that the similarities can be computed by multiplying the results of computations (2) over all the predicates, while using the parameter γ' instead of the original γ. The same multiplication is done for the concentration parameter.

Note that in this approach we never attempted to encode cross-predicate correspondence between labeled semantic roles, the prior (and the model as whole) is invariant under any renaming of roles for individual predicates.

Admittedly, this method is not a proper estimation method for the BayesSRL model but rather the use of an extrinsic probabilistic model to set the similarity scores in the dd-CRP prior. This is in line with much of the work on using dd-CRPs where the similarities were used to encode prior or external knowledge (Blei and Frazier, 2011; Socher et al., 2011; Duan et al., 2007;

[5] Note that we use here the fact that the mappings are equiprobable.

Jensen and Shore, 2011) rather than estimated as in the multi-tasking set-up of Titov and Klementiev (2012).

To induce the model in the semi-supervised set-up, we use the same approximate MAP search algorithm, as originally proposed in Titov and Klementiev (2012) for the unsupervised setting.

4 Experiments

4.1 Data

Datasets. We evaluate our semi-supervised approach on the CoNLL 2009 distribution (Hajič et al., 2009) of the Penn Treebank WSJ corpus (Marcus et al., 1993). We split the CoNLL training set roughly in half: we draw annotated sentences from the first part (20,000 sentences), and evaluate on the remaining 19,279 sentences. All, but the drawn annotated sentences are used as unsupervised training data as standard for unsupervised SRL.

Syntactic annotation. We annotate the data with dependency structures predicted by the syntactic component of the LTH system (Johansson and Nugues, 2008b), a more realistic setup than making use of the gold syntactic annotation.

Predicate and argument identification. We select all non-auxiliary verbs as predicates.[6] We identify their arguments using a heuristic proposed in (Lang and Lapata, 2011a). Since our goal is to evaluate the argument labeling stage of semantic role labeling, we use this argument identification procedure for all of the systems in our experiments. The quality of argument identification on CoNLL 2009 using predicted syntactic analyses was F1 82.7% (P 83.3% / R 82.0%).

4.2 Evaluation Metrics

We cannot use supervised metrics to evaluate our models, since we do not have an alignment between gold labels and clusters induced in the unsupervised and semi-supervised set-up.[7] Instead, we use the following two standard sets of clustering metrics for our evaluation:

Purity, Collocation, and F1. 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 and collocation evaluates the degree to which arguments with the same gold roles are assigned to a single cluster, see (Lang and Lapata, 2010).

Homogeneity, Completeness, and V-Measure. Additionally, we also evaluate with the information-theoretic V-Measure (V) (Rosenberg and Hirschberg, 2007). It is defined as the harmonic mean of homogeneity (H) and completeness (C) scores, which attempt to measure similar characteristics of the induced clustering as purity and collocation, respectively.

We compute the aggregate scores for all metrics 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] In this work we do not disambiguate predicate senses.

[7] Our BayesSRL extension does not propagate role labels between predicates which we would need to compute supervised metrics.

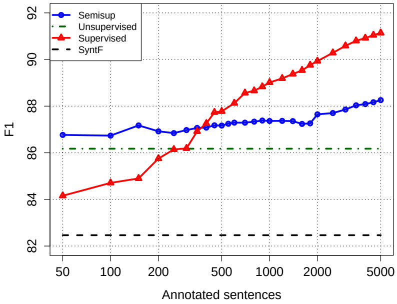

Figure 3: Performance (F1) of supervised (*Supervised*) and semi-supervised (*SemiSup*) systems vs. the number of annotated sentences, along with the original unsupervised model (*Unsupervised*) and the syntactic baseline (*SyntF*).

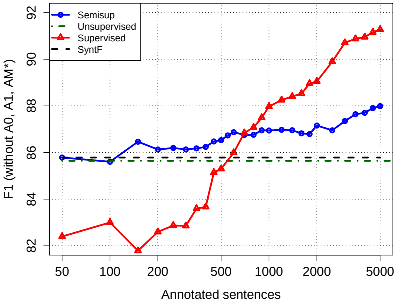

Figure 4: Performance (F1) evaluated on all roles except *A*0, *A*1, and *AM*∗ (modifier arguments) vs. the number of annotated sentences.

4.3 Model Parameters

The unsupervised model and the semi-supervised extension are robust to parameter settings. While they could be tuned by visual inspection of the induced argument roles, as in much of the previous work, we instead tuned them on the standard CoNLL held-out set primarily for replicability reasons.

4.4 Systems

In our experiments, we compare the performance of three systems: our semi-supervised extension (*SemiSup*) to the original state-of-the-art unsupervised model (*Unsupervised*) of Titov and Klementiev (2012), as well as the best CoNLL-08 shared task supervised SRL system (*Supervised*) of Johansson and Nugues (2008b). We also compare against the syntactic function baseline (*SyntF*), which is considered difficult to outperform in the unsupervised setting (Grenager and Manning, 2006; Lang and Lapata, 2010). It simply clusters predicate arguments according to the dependency relation to their head. As in previous work, we allocate a cluster for each of 20 most frequent relations in the CoNLL dataset and one cluster for all other relations.

4.5 Discussion

Figure 3 summarizes the results for the three systems and the syntactic baseline. The semi-supervised model outperforms the supervised counterpart when up to about 350 annotated sentences are available for training. It also continues to improve over the original unsupervised model as more annotated sentences are used for training. Table 1 details the single point of 300 labeled sentences on Figure 3 and breaks up the evaluation of the three systems and the syntactic baseline. It also shows the effect of the two ways of exploiting labeled data we

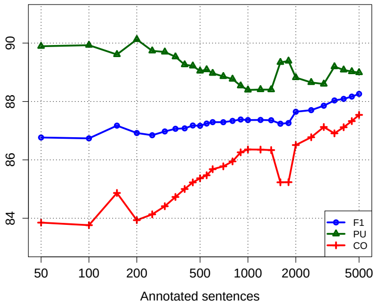

Figure 5: Purity, Collocation, and F1 of our semi-supervised extension (*SemiSup*) vs. the number of annotated sentences.

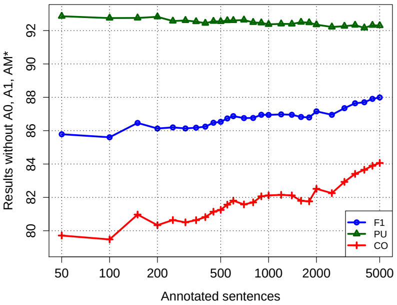

Figure 6: Purity, Collocation, and F1 of our semi-supervised extension (*SemiSup*) evaluated on all roles except *A*0, *A*1, and *AM*∗ vs. the number of annotated sentences.

	PU	CO	**F1**	H	C	**V**
Supervised	88.0	84.5	**86.2**	79.6	74.7	**77.0**
Unsupervised	89.6	83.0	**86.2**	83.8	73.3	**78.2**
SemiSup	89.7	84.4	**87.0**	83.6	74.9	**79.0**
SemiSup-l	89.5	84.2	**86.8**	83.3	74.6	**78.7**
SemiSup-p	90.0	82.5	**86.1**	84.4	72.8	**78.2**
SyntF	83.3	81.6	**82.5**	73.4	70.4	**71.9**

Table 1: Purity (PU), Collocation (CO), and F1, as well as Homogeneity (H), Completeness (C), and V-Measure (V) for for a single point (300 labeled sentences) on Figure 3. Results are for the syntactic baseline (*SyntF*), the supervised system (*Supervised*), the unsupervised model (*Unsupervised*), our semi-supervised extension (*SemiSup*), as well as our extension without adding labeled data to the generative story (*SemiSup-l*), and without the informed prior (*SemiSup-p*).

proposed in Section 3. *SemiSup-l* and *SemiSup-p* is our semi-supervised (*SemiSup*) method without adding labeled data to the generative story, and without informed priors, respectively. Note, that while adding labeled data alone does not improve over the performance of the unsupervised model for this number of labeled examples, the combination of the two methods yields a substantial improvement both in terms of F1 and V-Measure.

*A*0 and *A*1 arguments are annotated in PropBank based on the proto-role theory presented in (Dowty, 1991) and correspond to proto-agents and proto-patients, respectively, while arguments receiving an *AM*∗ label are supposed to be adjuncts, and the roles they express are consistent across all verbs. In order to evaluate the model performance on arguments which do not necessarily express consistent semantic roles across verbs, we next exclude *A*0, *A*1, and *AM*∗ from evaluation (Figure 4). The semi-supervised extension again substantially outperforms the supervised model when fewer than about 700 annotated examples are available.

Finally, Purity / Collocation breakdown for our semi-supervised extension (*SemiSup*) evaluated an all roles and all roles except $A0$, $A1$, and $AM*$ is shown on Figure 5 and Figure 6, respectively. Labeled data mostly provides evidence for combining clusters, so more labeled data implies collocation improvements albeit with some drop in purity.

Our semi-supervised method outperforms the state-of-the-art supervised model when the number of labeled sentences is relatively small, but falls behind when the amount of annotated data grows. This is likely due to the simplistic and overly coarse representation and modeling of the linking between syntax and semantics which places an upper bound on how well the original unsupervised model and the semi-supervised extension can do. However, our results strongly suggest that approaching semi-supervised SRL by exploiting labeled data in unsupervised methods is a promising research direction. Existing state-of-the-art methods can already be used for languages and domains for which little or no annotated data is available.

5 Additional Related Work

Additionally to the semi-supervised approaches to SRL discussed in the introduction, semi-supervised and weakly-supervised techniques have also been explored for other types of semantic representations but these studies have mostly focused on restricted domains (Kate and Mooney, 2007; Liang et al., 2009; Titov and Kozhevnikov, 2010; Goldwasser et al., 2011; Liang et al., 2011). Similarly, unsupervised induction for other shallow semantic formalisms include Poon and Domingos (2009, 2010) and Titov and Klementiev (2011).

A related problem of inducing script knowledge, or narrative event chains, has recently received a considerable attention (Chambers and Jurafsky, 2008; Manshadi et al., 2008; Chambers and Jurafsky, 2009; Regneri et al., 2010, 2011) with approaches mostly considering unsupervised or weakly-supervised setting due to scarcity of labeled data. Though in this paper we focus on the labeling of arguments the complementary task of unsupervised argument identification was considered in Abend et al. (2009).

Unsupervised learning has been one of the central paradigms for the closely-related area of relation extraction, where several techniques have been proposed to cluster semantically similar verbalizations of relations (Lin and Pantel, 2001; Banko et al., 2007). Similarly to SRL, semi-supervised approaches in this area are also typically based on bootstrapping techniques (e.g., (Agichtein and Gravano, 2000; Rosenfeld and Feldman, 2007)) and often achieve impressive results. However, their set-up is arguably different from ours as relation extractors are generally more precision-oriented, focus primarily on binary relations and can partially sidestep the complexity of language.

6 Conclusions

In this work, we demonstrated that unsupervised techniques can be improved by exploiting small amounts of labeled data yielding SRL parsers competitive with supervised approaches in a low resource setting. We also uncovered some of the deficiencies of the existing unsupervised approaches; namely, overly coarse modeling of syntax-semantics interface resulting in a lower asymptotic performance in semi-supervised settings. These results motivate further research into design of generative models appropriate for semi-supervised learning of shallow semantics.

Acknowledgements

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

References

Abend, O., Reichart, R., and Rappoport, A. (2009). Unsupervised argument identification for semantic role labeling. In *ACL-IJCNLP*.

Agichtein, E. and Gravano, L. (2000). Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (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.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *IJCAI*.

Basili, R., Cao, D. D., Croce, D., Coppola, B., and Moschitti, A. (2009). Cross-language frame semantics transfer in bilingual corpora. In *CICLING*.

Blei, D. M. and Frazier, P. (2011). Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12:2461–2488.

Carreras, X. and Màrquez, L. (2005). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *CoNLL*.

Chambers, N. and Jurafsky, D. (2008). Unsupervised learning of narrative event chains. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 789–797.

Chambers, N. and Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. In *Proc. of the Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Daumé III, H. (2009). Semi-supervised or semi-unsupervised? In *NAACL HLT Workshop on Semisupervised Learning for Natural Language Processing*.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *LREC 2006*.

Deschacht, K. and Moens, M.-F. (2009). Semi-supervised semantic role labeling using the Latent Words Language Model. In *EMNLP*.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Duan, J., Guindani, M., and Gelfand, A. (2007). Generalized spatial dirichlet process models. *Biometrika*, 94:809–825.

Fillmore, C. J. (1968). The case for case. In E., B. and R.T., H., editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart, and Winston, New York.

Fürstenau, H. and Lapata, M. (2009). Graph alignment for semi-supervised semantic role labeling. In *EMNLP*.

Fürstenau, H. and Rambow, O. (2012). Unsupervised induction of a syntax-semantics lexicon using iterative refinement. In *In Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*.

Gao, Q. and Vogel, S. (2011). Corpus expansion for statistical machine translation with semantic role label substitution rules. In *ACL:HLT*.

Garg, N. and Henderson, J. (2012). Unsupervised semantic role induction with global role ordering. In *ACL*.

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

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

Grenager, T. and Manning, C. (2006). Unsupervised discovery of a statistical verb lexicon. In *EMNLP*.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., and Zhang, Y. (2009). The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009), June 4-5*.

He, S. and Gildea, D. (2006a). Integrating cluster information for cross-frame semantic role labeling. Technical report, Technical Report 892, University of Rochester.

He, S. and Gildea, D. (2006b). Self-training and co-training for semantic role labeling: Primary report. Technical report, Technical Report 891, University of Rochester.

Hinton, G. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Jensen, S. and Shore, S. (2011). Semiparametric bayesian modeling of income volatility heterogeneity. *Journal of the American Statistical Association*, 106(496):1280–1290.

Johansson, R. and Nugues, P. (2008a). Dependency-based semantic role labeling of PropBank. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 69–78.

Johansson, R. and Nugues, P. (2008b). Dependency-based semantic role labeling of PropBank. In *EMNLP*.

Kaisser, M. and Webber, B. (2007). Question answering based on semantic roles. In *ACL Workshop on Deep Linguistic Processing*.

Kaljahi, Z. and Samad, R. (2010). Adapting self-training for semantic role labeling. In *Proceedings of the ACL 2010 Student Research Workshop*, pages 91–96. Association for Computational Linguistics.

Kate, R. J. and Mooney, R. J. (2007). Learning language semantics from ambigous supervision. In *AAAI*.

Lang, J. and Lapata, M. (2010). Unsupervised induction of semantic roles. In *ACL*.

Lang, J. and Lapata, M. (2011a). Unsupervised semantic role induction via split-merge clustering. In *ACL*.

Lang, J. and Lapata, M. (2011b). Unsupervised semantic role induction with graph partitioning. In *EMNLP*.

Lee, J., Song, Y., and Rim, H. (2007). Investigation of weakly supervised learning for semantic role labeling. In *Advanced Language Processing and Web Information Technology, 2007. ALPIT 2007. Sixth International Conference on*, pages 165–170. IEEE.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Liang, P., Jordan, M., and Klein, D. (2011). Learning dependency-based compositional semantics. In *ACL: HLT*.

Liang, P., Jordan, M. I., and Klein, D. (2009). Learning semantic correspondences with less supervision. In *ACL-IJCNLP*.

Lin, D. and Pantel, P. (2001). DIRT – discovery of inference rules from text. In *KDD*.

Liu, D. and Gildea, D. (2010). Semantic role features for machine translation. In *Coling*.

Manshadi, M., Swanson, R., and Gordon, A. (2008). Learning a probabilistic model of event sequences from internet weblog stories. In *Proceedings of the 21st FLAIRS Conference*.

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

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

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Poon, H. and Domingos, P. (2009). Unsupervised semantic parsing. In *EMNLP*.

Poon, H. and Domingos, P. (2010). Unsupervised ontology induction from text. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 296–305. Association for Computational Linguistics.

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

Regneri, M., Koller, A., and Pinkal, M. (2010). Learning script knowledge with web experiments. In *Proceedings of ACL 2010*, Uppsala, Sweden. Association for Computational Linguistics.

Regneri, M., Koller, A., Ruppenhofer, J., and Pinkal, M. (2011). Learning script participants from unlabeled data. In *Proceedings of RANLP 2011*, Hissar, Bulgaria.

Rosenberg, A. and Hirschberg, J. (2007). V-Measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 EMNLP-CoNll Joint Conference*, pages 410–420.

Rosenfeld, B. and Feldman, R. (2007). Using corpus statistics on entities to improve semi-supervised relation extraction from the web. In *ACL*.

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

Shen, D. and Lapata, M. (2007). Using semantic roles to improve question answering. In *EMNLP*.

Socher, R., Maas, A., and Manning, C. (2011). Spectral chinese restaurant processes: Non-parametric clustering based on similarities. In *AISTATS*.

Surdeanu, M., Johansson, A. M. R., Màrquez, L., and Nivre, J. (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Shared Task*.

Swier, R. and Stevenson, S. (2004). Unsupervised semantic role labelling. In *EMNLP*.

Titov, I. and Klementiev, A. (2011). A Bayesian model for unsupervised semantic parsing. In *ACL*.

Titov, I. and Klementiev, A. (2012). A Bayesian approach to semantic role induction. In *Proc. EACL*, Avignon, France.

Titov, I. and Kozhevnikov, M. (2010). Bootstrapping semantic analyzers from non-contradictory texts. In *ACL*.

van der Plas, L., Henderson, J., and Merlo, P. (2009). Domain adaptation with artificial data for semantic parsing of speech. In *Proc. 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 125–128, Boulder, Colorado.

van der Plas, L., Merlo, P., and Henderson, J. (2011). Scaling up automatic cross-lingual semantic role annotation. In *ACL*.

Wu, D., Apidianaki, M., Carpuat, M., and Specia, L., editors (2011). *Proc. of Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*. ACL.

Wu, D. and Fung, P. (2009). Semantic roles for SMT: A hybrid two-pass model. In *NAACL*.