Unsupervised Induction of Semantic Roles within a Reconstruction-Error Minimization Framework

Ivan Titov **Ehsan Khoddam**
University of Amsterdam
{titov|e.khoddammohammadi}@uva.nl

Abstract

We introduce a new approach to unsupervised estimation of feature-rich semantic role labeling models. Our model consists of two components: (1) an encoding component: a semantic role labeling model which predicts roles given a rich set of syntactic and lexical features; (2) a reconstruction component: a tensor factorization model which relies on roles to predict argument fillers. When the components are estimated jointly to minimize errors in argument reconstruction, the induced roles largely correspond to roles defined in annotated resources. Our method performs on par with most accurate role induction methods on English and German, even though, unlike these previous approaches, we do not incorporate any prior linguistic knowledge about the languages.

1 Introduction

Shallow semantic representations, and semantic role labels in particular, have a long history in linguistics (Fillmore, 1968). More recently, with an 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; Das et al., 2010).

Semantic role 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). Semantic roles 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 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. Moreover, when moved to a new domain (e.g., from news corpora to blogs or biomedical texts), the performance of these models tends to degrade substantially (Pradhan et al., 2008). The scarcity of annotated data has motivated the research into unsupervised learning of semantic representations (Swier and Stevenson, 2004; Grenager and Manning, 2006; Lang and Lapata, 2010; Lang and Lapata, 2011a; Lang and Lapata, 2011b; Titov and Klementiev, 2012a; Fürstenau and Rambow, 2012; Garg and Henderson, 2012). The existing methods have a number of serious shortcomings. First, they make very strong assumptions, for example, assuming that arguments are conditionally independent of each other given the predicate. Second, unlike state-of-the-art supervised parsers, they rely on a very simplistic set of features of a sentence. These factors lead to models being insufficiently expressive to capture the syntax-semantics interface, inadequate

handling of language ambiguity and, overall, introduces a restrictive upper bound on their performance. Moreover, these approaches are especially problematic for languages with freer word order than English, where richer features are necessary to account for interactions between surface realizations, syntax and semantics. For example, the two most accurate previous models (Titov and Klementiev, 2012a; Lang and Lapata, 2011a) both treat the role induction task as clustering of argument signatures: an argument signature encodes key syntactic properties of an argument realization and consists of a syntactic function of an argument along with additional information such as an argument position with respect to the predicate. Though it is possible to design signatures which mostly map to a single role, this set-up limits oracle performance even for English, and can be quite restrictive for languages with freer word order. These shortcomings are inherent limitations of the modeling frameworks used in previous work (primarily generative modeling or agglomerative clustering), and cannot be addressed by simply incorporating more features or relaxing some of the modeling assumptions.

In this work, we propose a method for effective unsupervised estimation of feature-rich models of semantic roles. We demonstrate that reconstruction-error objectives, which have been shown to be effective primarily for training neural networks, are well suited for inducing feature-rich log-linear models of semantics. Our model consists of two components: a log-linear feature-rich semantic role labeler and a tensor-factorization model which captures interaction between semantic roles and argument fillers. When estimated jointly on unlabeled data, roles induced by the model mostly corresponds to roles defined in existing resources by annotators.

Our method rivals the most accurate semantic role induction methods on English and German (Titov and Klementiev, 2012a; Lang and Lapata, 2011a). Importantly, no prior knowledge about the languages was incorporated in our feature-rich model, whereas the clustering counterparts relied on language-specific argument signatures. These languages-specific priors were crucial for their success. For example, using English-specific argument signatures for German with the Bayesian model of Titov and Klementiev (2012a) results in a drop of performance from clustering F1 of 80.9% to considerably lower 78.3% (our model yields 81.4%). This confirms the intuition that using richer features helps to capture the syntax-semantics interface in multilingual settings, reducing the need for language-specific model engineering, as is highly desirable in unsupervised learning.

The rest of the paper is structured as follows. Section 2 begins with a definition of the semantic role labeling task and discusses some specifics of the unsupervised setting. In Section 3, we describe our approach, starting with a general motivation and proceeding to technical details of the model (Section 3.3) and the learning procedure (Section 3.4). Section 4 provides both evaluation and analysis. Finally, additional related work is presented in Section 5.

2 Task Definition

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

(a) [$_{Agent}$ Mary] opened [$_{Patient}$ the door].

(b) [$_{Patient}$ The door] opened.

(c) [$_{Patient}$ The door] was opened [$_{Agent}$ by Mary].

Mary always takes an agent role for the predicate *open*, and *door* is always a patient.

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.

3 Approach

At the core of our approach is a statistical model encoding an interdependence between a semantic role structure and its realization in a sentence. In the unsupervised learning setting, sentences, their syntactic representations and argument positions (denoted by x) are observable whereas the associated semantic roles $\boldsymbol{r}$ are latent and need to be induced by the

model. The idea which underlines much of latent variable modeling is that a good latent representation is the one which helps us to reconstruct x. In practice, we are not interested in predicting x, as x is observable, but rather interested in inducing appropriate latent representations (i.e. $\boldsymbol{r}$). Thus, it is crucial to design the model in such a way that the good $\boldsymbol{r}$ (the one predictive of x) indeed encodes roles, rather than some other form of abstraction.

In what follows, we will refer to roles using their names, though, in the unsupervised setting, our method, as any other latent variable model, will not yield human-interpretable labels for them. We will use the following sentence as a motivating example in our discussion of the model:

$[_{Agent}$ The police] charged $[_{Patient}$ the demonstrators] $[_{Instrument}$ with batons].

The model consists of two components. The first component is responsible for prediction of argument tuples based on roles and the predicate. In our experiments, in this component, we represent arguments as lemmas of their lexical heads (e.g., *baton* instead of *with batons*). We also restrict ourselves to only verbal predicates. Intuitively, we can think of predicting one argument at a time (see Figure 1(b)): an argument (e.g., *demonstrator* in our example) is predicted based on the predicate lemma (*charge*), the role assigned to this argument (i.e. `Patient`) and other role-argument pairs ((`Agent`, *police*) and (`Instrument`, *baton*)). While learning to predict arguments, the inference algorithm will search for role assignments which simplify this prediction task as much as possible. Our hypothesis is that these assignments will correspond to roles accepted in linguistic theories (or, more importantly, useful in practical applications). Why is this hypothesis plausible? Primarily because these semantic representations were introduced as an abstraction capturing the essence of a situation (or a event). And the underlying situation and participant roles in this situation (rather than surface linguistic details like argument order or syntactic functions) are precisely what impose constraints on admissible argument tuples.

The reconstruction component is not the only part of the model. Crucially, what we referred to above as 'searching for role assignments to simplify argument prediction' would actually correspond to learning another component: a semantic role labeler which predicts roles relying on a rich set of sentence features. These two components will be estimated jointly in such a way as to minimize errors in recovering arguments. The role labeler will be the end-product of learning: it will be used to process new sentences, and it will be compared to existing methods in our evaluation.

3.1 Shortcomings of generative modeling

The above paragraph can be regarded as our desiderata; now we discuss how to achieve them. The standard way to approach latent variable modeling is to use the generative framework: that is to define a family of joint models $p(x, y|\theta)$ and estimate the parameters θ by, for example, maximizing the likelihood. Generative models of semantics (Titov and Klementiev, 2012a; Titov and Klementiev, 2011; Modi et al., 2012; O'Connor, 2013; Kawahara et al., 2014) necessarily make very strong independence assumptions (e.g., arguments are conditionally independent of each other given the predicate) and use simplistic features of x and y. Thus, they cannot meet the desiderata stated above. Importantly, they are also much more simplistic in their assumptions than state-of-the-art supervised role labelers (Erk and Pado, 2006; Johansson and Nugues, 2008; Das et al., 2010).

3.2 Reconstruction error minimization

Generative modeling is not the only way to learn latent representations. One alternative, popular in the neural network community, is to instead use autoencoders and optimize the reconstruction error (Hinton, 1989; Vincent et al., 2008). In autoencoders, a latent representation y (their hidden layer) is predicted from x by an encoding model and then this y is used to recover $\tilde{x}$ with a reconstruction model (see Figure 1(a)). Parameters of the encoding and reconstruction components are chosen so as to minimize some form of the reconstruction error, for example, the Euclidean distance $\Delta(x, \tilde{x}) = ||x - \tilde{x}||_2$. Though currently popular only within the deep learning community, latent variable models other than neural networks can also be trained this way, moreover:

- the encoding and reconstruction models can belong to different model families;

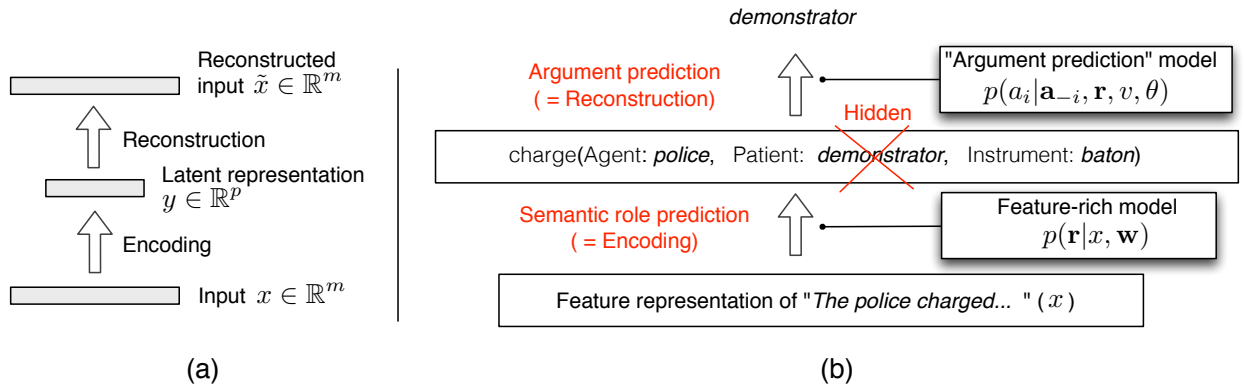

Figure 1: (a) An autoencoder from $\mathbb{R}^m$ to $\mathbb{R}^p$ (typically $p < m$). (b) Modeling roles within the reconstruction-error minimization framework.

- the reconstruction component may be focused on recovering a part of x rather than the entire x, and, in doing so, can rely not only on y but on the remaining part of x.

These observations are crucial as they allow us to implement our desiderata. More specifically, the encoding model will be a feature-rich classifier which predicts semantic roles for a sentence, and the reconstruction model is the model which predicts an argument given its role, and given the rest of the arguments and their roles. The idea of training linear models by minimizing the reconstruction error was previously explored by Daumé (2009) and very recently by Ammar et al. (2014).

3.3 Modeling semantics within the reconstruction-error framework

There are several possible ways to translate the ideas above into a specific method, and we consider one of the simplest instantiations. For simplicity, in the discussion (but not in our experiments), we assume that exactly one predicate is realized in each sentence x. As we mentioned above, we focus only on argument labeling: we assume that arguments $\boldsymbol{a} = (a_1, \ldots, a_N)$, $a_i \in \mathcal{A}$, are known, and only their roles $\boldsymbol{r} = (r_1, \ldots, r_N)$, $r_i \in \mathcal{R}$ need to be induced. For the encoder (i.e. the semantic role labeler), we use a log-linear model:

$$p(\boldsymbol{r}|x, \boldsymbol{w}) \propto \exp(\boldsymbol{w}^T \boldsymbol{g}(x, \boldsymbol{r})),$$

where $\boldsymbol{g}(x, \boldsymbol{r})$ is a feature vector encoding interactions between sentence x and the semantic role representation $\boldsymbol{r}$. Any model can be used here as long as the posterior distributions of roles r_i can be efficiently computed or approximated (we will see why in Section 3.4). In our experiments, we used a model which factorizes over individual arguments (i.e. a set of independent logistic regression classifiers).

The reconstruction component predicts an argument (e.g., the ith argument a_i) given the semantic roles $\boldsymbol{r}$, the predicate v and other arguments $\boldsymbol{a}_{-i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_N)$ with a bilinear softmax model:

$$p(a_i|\boldsymbol{a}_{-i}, \boldsymbol{r}, v, C, \boldsymbol{u}) = \frac{\exp(\boldsymbol{u}_{a_i}^T C_{v,r_i}^T \sum_{j \neq i} C_{v,r_j} \boldsymbol{u}_{a_j})}{Z(\boldsymbol{r}, v, i)}, \quad (1)$$

$\boldsymbol{u}_a \in \mathbb{R}^d$ (for every $a \in \mathcal{A}$) and $C_{v,r} \in \mathbb{R}^{k \times d}$ (for every verb v and every role $r \in \mathcal{R}$) are model parameters, $Z(\boldsymbol{r}, v, i)$ is the partition function ensuring that the probabilities sum to one. Intuitively, embeddings $\boldsymbol{u}_a$, when learned from data, will encode semantic properties of an argument: for example, embeddings for the words *demonstrator* and *protestor* should be somewhere near each other in $\mathbb{R}^d$ space, and further away from that for the word *cat*. The product $C_{v,r}\boldsymbol{u}_a$ is a k-dimensional vector encoding beliefs about other arguments based on the argument-role pair (a, r). For example, seeing the argument *demonstrator* in the `Patient` position for the predicate *charge*, one would predict that the `Agent` is perhaps the word *police*, and the role `Instrument` is filled by the word *baton* or perhaps

(a water) *cannon*. On the contrary, if the `Patient` is *cat* then the `Agent` is more likely to be *dog* than *police*. In turn, the dot product $(C_{v,r_i}\boldsymbol{u}_{a_i})^T C_{v,r_j}\boldsymbol{u}_{a_j}$ is large if these expectations are met for the argument pair (a_i, a_j), and small otherwise. Intuitively, this objective corresponds to scoring argument tuples according to

$$h(\boldsymbol{a}, \boldsymbol{r}, v, C, \boldsymbol{u}) = \sum_{i \neq j} \boldsymbol{u}_{a_i}^T C_{v,r_i}^T C_{v,r_j} \boldsymbol{u}_{a_j}, \quad (2)$$

hinting at connections to (coupled) tensor and matrix factorization methods (Nickel et al., 2011; Yılmaz et al., 2011; Bordes et al., 2011; Riedel et al., 2013) and distributional semantics (Mikolov et al., 2013; Pennington et al., 2014). Note also that the reconstruction model does not have access to any features of the sentence (e.g., argument order or syntax), forcing the roles to convey all the necessary information.

This factorization can be thought of as a generalization of the notion of selection preferences. Selectional preferences characterize the set of arguments licensed for a given role of a given predicate: for example, `Agent` for the predicate *charge* can be *police* or *dog* but not *table* or *idea*. In our generalization, we model soft restrictions imposed not only by the role itself but also by other arguments and their assignment to roles.

In practice, we extend the model slightly: (1) we introduce a word-specific bias (a scalar b_a for every $a \in \mathcal{A}$) in the argument prediction model (equation (1)); (2) we smooth the model by using a sum of predicate-specific and cross-predicate projection matrices $(C_{v,r} + C_r)$ instead of just $C_{v,r}$.

3.4 Learning

Parameters of both model components ($\boldsymbol{w}$, $\boldsymbol{u}$ and C) are learned jointly: the natural objective associated with every sentence would be the following:

$$\sum_{i=1}^{N} \log \sum_{\boldsymbol{r}} p(a_i | \boldsymbol{a}_{-i}, \boldsymbol{r}, v, C, \boldsymbol{u}) p(\boldsymbol{r} | x, \boldsymbol{w}). \quad (3)$$

However optimizing this objective is not practical in its exact form for two reasons: (1) marginalization over $\boldsymbol{r}$ is exponential in the number of arguments; (2) the partition function $Z(\boldsymbol{r}, v, i)$ requires summation over the entire set of potential argument lemmas. We use existing techniques to address both challenges.

In order to deal with the first challenge, we use a basic mean-field approximation. Namely, instead of computing an expectation of $p(a_i|\boldsymbol{a}_{-i}, \boldsymbol{r}, v, C, \boldsymbol{u})$ under $p(\boldsymbol{r}|x, \boldsymbol{w})$, as in (3), we use the posterior distributions $\mu_{is} = p(\boldsymbol{r_i} = s|x, \boldsymbol{w})$ and score the argument predictions as

$$p(a_i|\boldsymbol{a}_{-i}, \boldsymbol{\mu}, v, C, \boldsymbol{u}) = \frac{\exp\left(\phi_i(a_i, \boldsymbol{a}_{-i})\right)}{Z(\boldsymbol{\mu}, v, i)} \quad (4)$$

$$\phi_i(a_i, \boldsymbol{a}_{-i}) = \boldsymbol{u}_{a_i}^T \Big(\sum_s \mu_{is} C_{v,s}\Big)^T \times \sum_{j \neq i} \Big(\sum_s \mu_{js} C_{v,s}\Big) \boldsymbol{u}_{a_j},$$

where $\boldsymbol{\mu}$ are the posteriors for all the arguments, and $\phi_i(a, \boldsymbol{a}_{-i})$ is the score associated with predicting lemma a for the argument i.

In order to address the second problem, the computation of $Z(\boldsymbol{\mu}, v, i)$, we use a negative sampling technique (see, e.g., Mikolov et al. (2013)). More specfically, we get rid of the softmax in equation (4) and optimize the following sentence-level objective:

$$\sum_{i=1}^{N} \Big[\log \sigma(\phi_i(a_i, \boldsymbol{a}_{-i})) - \sum_{a' \in S} \log \sigma(\phi_i(a', \boldsymbol{a}_{-i}))\Big], \quad (5)$$

where S is a random sample of n elements from the unigram distribution of lemmas, and σ is the logistic sigmoid function.

Assuming that the posteriors $\boldsymbol{\mu}$ can be derived in a closed form, the gradients of the objective (5) with respect to parameters of both the encoding component ($\boldsymbol{w}$) and the reconstruction component (C, $\boldsymbol{u}$ and $\boldsymbol{b}$) can be computed using back propagation. In our experiments, we used the AdaGrad algorithm (Duchi et al., 2011) to perform the optimization.

The learning algorithm is quite efficient, as the reconstruction computation is bilinear, whereas the computation of the posteriors $\boldsymbol{\mu}$ (and the computation of their gradients) from the semantic roler labeling component (encoder) is not much more expensive than discriminative supervised learning of

the role labeler. Moreover, the computations can be sped up substantially by observing that the sum $\sum_s \mu_{is} C_{v,s}$ in expression (4) can be precomputed for all i, and reused across predictions of different arguments of the same predicate. At test time, only the linear semantic role labeler is used, so the inference is straightforward.

4 Experiments

4.1 Data and evaluation metrics

We considered English and German in our experiments. For each language, we replicated experimental set-ups used in previous work.

For English, we followed Lang and Lapata (2010) and used the dependency version of PropBank (Palmer et al., 2005) released for the CoNLL 2008 shared task (Surdeanu et al., 2008). The dataset is divided into three segments. As in the previous work on unsupervised role labeling, we used the largest segment (the original CoNLL training set, sections 2-21) both for evaluation and learning. This is permissible as unsupervised models do not use gold labels in training. The two small segments (sections 22 and 23) were used for model development. In our experiments, we relied on gold standard syntax and gold standard argument identification, as this set-up allows us to evaluate against much of the previous work. We refer the reader to Lang and Lapata (2010) for details of the experimental set-up.

There has not been much work on unsupervised induction of roles for languages other than English, perhaps primarily because of the above-mentioned model limitations. For German, we replicate the set-up considered in Titov and Klementiev (2012b). They used the CoNLL 2009 version (Hajič et al., 2009) of the SALSA corpus (Burchardt et al., 2006). Instead of using syntactic parses provided in the CoNLL dataset, they re-parsed it with the MALT dependency parser (Nivre et al., 2004). Similarly, rather than relying on gold standard annotations for argument identification, they used a supervised classifier to predict argument positions. Details of the preprocessing can be found in Titov and Klementiev (2012b).

As in most previous work on unsupervised SRL, we evaluate our model using purity, collocation and their harmonic mean F1. *Purity* (PU) measures the average number of arguments with the same gold role label in each cluster, *collocation* (CO) measures to what extent a specific gold role is represented by a single cluster. More formally:

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

where if 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. Similarly, for collocation:

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

We compute the aggregate PU, CO, and F1 scores over all predicates in the same way as Lang and Lapata (2010): we weight the scores for each predicate by the number of times its arguments occur and compute the weighted average.

4.2 Parameters and features

For the semantic role labeling (encoding) component, we relied on 14 feature patterns used for argument labeling in a popular supervised role labeler (Johansson and Nugues, 2008). These patterns include non-trivial syntactic features, such as a dependency path between the target predicate and the considered argument. The resulting feature space is quite large (49,474 feature instantiations for our English dataset) and arguably sufficient to accurately capture syntax-semantics interface for most languages. We refer the reader to the original publication for details (Johansson and Nugues, 2008: Table 2). Importantly, the dimensionality of the feature space is very different from the one used typically in unsupervised SRL. In principle, any features could be used here but we chose these 14 feature patterns, as they all are fairly simple and generic. They can also be easily extracted from any treebank. We used the same feature patterns both for English and German. However, there is little doubt that some language-specific feature engineering and the use of language-specific priors or constraints (e.g., posterior regularization (Ganchev et al., 2010)) would benefit the performance. Faithful to our goal of constructing the simplest possible feature-rich model,

we use logistic classifiers independently predicting role distribution for every argument.

For the reconstruction component, both for English and German, we set the embedding dimensionality d, the projection dimensionality k and the number of negative samples n to 30, 15 and 20, respectively. The model was not sensitive to the parameter $|\mathcal{R}|$, defining the number of roles as long it was large enough (see Section 4.3 for more discussion). For training, we used uniform random initialization and AdaGrad (Duchi et al., 2011). Any model selections (e.g., choosing the number of epochs) was done on the basis of the respective held-out set.

4.3 Results

4.3.1 English

Table 1 summarizes the results of our method, as well as those of alternative approaches and baselines.

Following (Lang and Lapata, 2010), we use a baseline (*SyntF*) which simply clusters predicate arguments according to the dependency relation to their head. A separate cluster is allocated for each of 20 most frequent relations in the dataset and an additional cluster is used for all other relations. As observed in the previous work (Lang and Lapata, 2011a), this is a hard baseline to beat.

We also compare with previous approaches: the latent logistic classification model (Lang and Lapata, 2010) (labeled *LLogistic*), the agglomerative clustering method (Lang and Lapata, 2011a) (*Agglom*), the graph partitioning approach (Lang and Lapata, 2011b) (*GraphPart*), the global role ordering model (Garg and Henderson, 2012) (*RoleOrdering*). We also report results of an improved version of *Agglom*, recently reported by Lang and Lapata (2014) (*Agglom+*). The strongest previous model is *Bayes*: *Bayes* is the most accurate ('coupled') version of the Bayesian model of Titov and Klementiev (2012a), estimated from the CoNLL dataset without relying on any external data. Titov and Klementiev (2012a) also showed that using Brown clusters induced from a large external corpus resulted in an 0.5% improvement in F1 but that version is not entirely comparable to other systems induced solely from the CoNLL text.

Our model outperforms or performs on par with best previous models in terms of F1. Interestingly, the purity and collocation balance is very different for our model and for the rest of the systems. In fact, our model induces at most 4-6 roles (even if $|\mathcal{R}|$ is much larger). On the contrary, *Bayes* predicts more than 30 roles for the majority of frequent predicates (e.g., 43 roles for the predicate *include* or 35 for *say*). Though this tendency reduces the purity scores for our model, this also means that our roles are more human interpretable. For example, agents and patients are clearly identifiable in the model predictions. Our model has similar purity to the syntactic baseline but outperforms it vastly according to the collocation metric, suggesting that we go substantially beyond recovering syntactic relations.

	PU	CO	F1
Our Model	79.7	86.2	**82.8**
Bayes	89.3	76.6	82.5
Agglom+	87.9	75.6	81.3
RoleOrdering	83.5	78.5	80.9
Agglom	88.7	73.0	80.1
GraphPart	88.6	70.7	78.6
LLogistic	79.5	76.5	78.0
SyntF	81.6	77.5	79.5

Table 1: Results on English (PropBank / CoNLL 2008).

In additional experiments, we observed that our model, in some regimes, starts to induce roles specific to individual verb senses or specific to groups of semantically similar predicates. This suggests that adding a latent variable capturing predicate senses and conditioning the reconstruction component on this variable may not only result in a more informative semantic representation (i.e. include verb senses) but also improve the role induction performance. We leave this exploration for future work.

4.3.2 German

For German, we replicate the experimental set-up previously used by Titov and Klementiev (2012b). As for English, we report results of the syntactic baseline (*SyntF*). The results for all approaches are presented in Table 2. We compare against *Bayes (De)* – the *Bayes* model with argument signatures specialized for German (as reported in Titov and Klementiev (2012b)). We also consider the original

	PU	CO	F1
Our Model	76.4	87.0	**81.4**
Bayes (De)	86.8	75.7	80.9
Bayes (En)	80.6	76.0	78.3
SyntF	83.1	79.3	81.2

Table 2: Results on German (SALSA / CoNLL 2009).

version of the *Bayes* model (denoted as *Bayes (En)*).

Recently, Lang and Lapata (2014) evaluated their *Agglom+* on a version of the same German SALSA dataset. Their best result is F1 of 79.2%, however, this score and our results are not directly comparable. Instead of using the CoNLL dataset, they processed the corpus themselves. They also relied on syntactic features from a constituent parser whereas we used dependency representations.

The overall picture for German closely resembles the one for English. Our method achieves results comparable to the best method evaluated in this setting. Importantly, parameters and features of our model for German and English are identical. On the contrary, one can see that specialization of argument signatures was crucial for the Bayesian model. Also, similarly to English, our method induces less fine-grain sets of semantic roles but achieves much higher collocation scores.

5 Additional Related Work

In recent years, unsupervised approaches to semantic role induction have attracted considerable attention. However, there exist other ways to address lack of coverage provided by existing semantically-annotated resources.

One natural direction is semi-supervised role labeling, where both annotated and unannotated data is used to estimate a model. Previous semi-supervised approaches to SRL can be mostly regarded as extensions to supervised learning by either incorporating word features induced from unnannoted texts (Collobert and Weston, 2008; Deschacht and Moens, 2009) or creating some form of 'surrogate' supervision (He and Gildea, 2006; Fürstenau and Lapata, 2009). Benefits from using unlabeled data were moderate, and more significant for the harder SRL version, frame-semantic parsing (Das and Smith, 2011).

Another important direction includes cross-lingual approaches (Pado and Lapata, 2009; van der Plas et al., 2011; Kozhevnikov and Titov, 2013) which leverage resources from resource-rich languages, as well as parallel data, to produce annotation or models for resource-poor languages. However, both translation shifts and noise in word alignments harm the performance of cross-lingual methods. Nevertheless, even joint unsupervised induction across languages appears to be beneficial (Titov and Klementiev, 2012b).

Unsupervised learning has also been one of the central paradigms for the closely-related area of relation extraction (RE), where several techniques have been proposed to cluster semantically similar verbalizations of relations (Lin and Pantel, 2001; Banko et al., 2007; Yao et al., 2011). Similarly to SRL, unsupervised methods for RE mostly rely on generative modeling and agglomerative clustering.

From the learning perspective, methods which use the reconstruction-error objective to estimate linear models (Ammar et al., 2014; Daumé III, 2009) are certainly related. However, they do not consider learning factorization models, and they also do not deal with semantics. Tensor factorization methods used in the context of modeling knoweldge bases (e.g., (Bordes et al., 2011)) are also close in spirit. However, they do not deal with inducing semantics but rather factorize existing relations (i.e. rely on semantics).

6 Conclusions and Discussion

This work introduces a method for inducing feature-rich semantic role labelers from unannoated text. In our approach, we view a semantic role representation as an encoding of a latent relation between a predicate and a tuple of its arguments. We capture this relation with a probabilistic tensor factorization model. The factorization model (relying on semantic roles) and a feature-rich model (predicting the roles) are jointly estimated by optimizing an objective which favours accurate reconstruction of arguments given the latent semantic representation (and other arguments). Our estimation method yields a semantic role labeler which achieves state-of-the-art results both on English and German.

Unlike previous work on role induction, in our

approach, virtually any computationally tractable structured model can be used as a role labeler, including almost any semantic role labeler introduced in the context of supervised SRL (see, e.g., CoNLL shared tasks (Carreras and Màrquez, 2005; Surdeanu et al., 2008; Hajič et al., 2009)). This opens interesting possibilities to extend our approach to the semi-supervised setting. Previous unsupervised SRL models make too strong assumption and use too limited features to effectively exploit labeled data. For our model, the reconstruction objective can be easily combined with the likelihood objective, yielding a potentially powerful semi-supervised method. We leave this direction for future work.

Acknowledgements

This work is partially supported by a Google focused award on natural language understanding. The authors thank Dipanjan Das, Ashutosh Modi, Alexis Palmer and the anonymous reviewers for their suggestions.

References

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

W. Ammar, C. Dyer, and N. Smith. 2014. Conditional random field autoencoders for unsupervised structured prediction. In *NIPS*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *ACL-COLING*.

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

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

A. Bordes, J. Weston, R. Collobert, and Y. Bengio. 2011. Learning structured embeddings of knowledge bases. In *AAAI*.

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*.

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

R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*.

D. Das and N. A. Smith. 2011. Semi-supervised frame-semantic parsing for unknown predicates. In *ACL*.

D. Das, N. Schneider, D. Chen, and N. A. Smith. 2010. Probabilistic frame-semantic parsing. In *NAACL*.

H. Daumé III. 2009. Unsupervised search-based structured prediction. In *ICML*.

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

K. Deschacht and M.-F. Moens. 2009. Semi-supervised semantic role labeling using the latent words language model. In *Proceedings of EMNLP*.

J. Duchi, E. Hazan, and Y. Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

K. Erk and S. Pado. 2006. Shalmaneser–a toolchain for shallow semantic parsing. In *LREC*.

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

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

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

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

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

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

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

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

J. Hajič, M. Ciaramita, R. Johansson, D. Kawahara, M.A. Martí, L. Màrquez, A. Meyers, J. Nivre, S. Padó, J. Štěpánek, P. Straňák, M. Surdeanu, N. Xue, and Y. Zhang. 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*.

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

G. E. Hinton. 1989. Connectionist learning procedures. *Artificial intelligence*, 40(1):185–234.

R. Johansson and P. Nugues. 2008. Dependency-based syntactic-semantic analysis with PropBank and NomBank. In *CoNLL*.

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

D. Kawahara, D. Peterson, O. Popescu, and M. Palmer. 2014. Inducing example-based semantic frames from a massive amount of verb uses. In *EACL*.

M. Kozhevnikov and I. Titov. 2013. Crosslingual transfer of semantic role models. In *ACL*.

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

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

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

J. Lang and M. Lapata. 2014. Similarity-driven semantic role induction via graph partitioning. *Computational Linguistics*, 40(3):633–669.

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

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

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

A. Modi, I. Titov, and A. Klementiev. 2012. Unsupervised induction of frame-semantic representations. In *NAACL Workshop on Inducing Linguistic Structure*.

M. Nickel, V. Tresp, and H.-P. Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *ICML*.

J. Nivre, J. Hall, and J. Nilsson. 2004. Memory-based dependency parsing. In *Proc. of the Eighth Conference on Computational Natural Language Learning*, pages 49–56, Boston, USA.

B. O'Connor. 2013. Learning frames from text with an unsupervised latent variable model. Technical report, CMU.

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

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

J. Pennington, R. Socher, and C. D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

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

S. Riedel, L. Yao, A. McCallum, and B. M. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *NAACL*.

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)*.

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

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

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

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

I. Titov and A. Klementiev. 2012a. A Bayesian approach to semantic role induction. In *EACL*.

I. Titov and A. Klementiev. 2012b. Crosslingual induction of semantic roles. In *ACL*.

L. van der Plas, J. Henderson, and P. Merlo. 2009. Domain adaptation with artificial data for semantic parsing of speech. In *NAACL*.

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

P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *ICML*.

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

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

L. Yao, A. Haghighi, S. Riedel, and A. McCallum. 2011. Structured relation discovery using generative models. In *EMNLP*.

K. Y. Yılmaz, A. T. Cemgil, and U. Simsekli. 2011. Generalised coupled tensor factorisation. In *NIPS*.