Unsupervised Semantic Role Induction with Global Role Ordering

Nikhil Garg
University of Geneva
Switzerland
nikhil.garg@unige.ch

James Henderson
University of Geneva
Switzerland
james.henderson@unige.ch

Abstract

We propose a probabilistic generative model for unsupervised semantic role induction, which integrates local role assignment decisions and a global role ordering decision in a unified model. The role sequence is divided into *intervals* based on the notion of *primary roles*, and each interval generates a sequence of *secondary roles* and syntactic constituents using local features. The global role ordering consists of the sequence of primary roles only, thus making it a partial ordering.

1 Introduction

Unsupervised semantic role induction has gained significant interest recently (Lang and Lapata, 2011b) due to limited amounts of annotated corpora. A Semantic Role Labeling (SRL) system should provide consistent argument labels across different syntactic realizations of the same verb (Palmer et al., 2005), as in

$(a.)$ [Mark $]_{A0}$ drove [the car $]_{A1}$

$(b.)$ [The car $]_{A1}$ was driven by [Mark $]_{A0}$

This simple example also shows that while certain local syntactic and semantic features could provide clues to the semantic role label of a constituent, non-local features such as predicate voice could provide information about the expected semantic role sequence. Sentence a is in active voice with sequence ($A0$, $PREDICATE$, $A1$) and sentence b is in passive voice with sequence ($A1$, $PREDICATE$, $A0$). Additional global preferences, such as arguments $A0$ and $A1$ rarely repeat in a frame (as seen in the corpus), could also be useful in addition to local features.

Supervised SRL systems have mostly used local classifiers that assign a role to each constituent independently of others, and only modeled limited correlations among roles in a sequence (Toutanova et al., 2008). The correlations have been modeled via *role sets* (Gildea and Jurafsky, 2002), role repetition constraints (Punyakanok et al., 2004), language model over roles (Thompson et al., 2003; Pradhan et al., 2005), and global role sequence (Toutanova et al., 2008). Unsupervised SRL systems have explored even fewer correlations. Lang and Lapata (2011a; 2011b) use the relative position (left/right) of the argument w.r.t. the predicate. Grenager and Manning (2006) use an ordering of the linking of semantic roles and syntactic relations. However, as the space of possible linkings is large, language-specific knowledge is used to constrain this space.

Similar to Toutanova et al. (2008), we propose to use global role ordering preferences but in a generative model in contrast to their discriminative one. Further, unlike Grenager and Manning (2006), we do not explicitly generate the linking of semantic roles and syntactic relations, thus keeping the parameter space tractable. The main contribution of this work is an unsupervised model that uses global role ordering and repetition preferences without assuming any language-specific constraints.

Following Gildea and Jurafsky (2002), previous work has typically broken the SRL task into (i) argument identification, and (ii) argument classification (Màrquez et al., 2008). The latter is our focus in this work. Given the dependency parse tree of a sentence with correctly identified arguments, the aim is to assign a semantic role label to each argument.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 145–149,
Jeju, Republic of Korea, 8-14 July 2012.

Algorithm 1 Generative process

```
——————— PARAMETERS ———————
for all predicate $p$ do
    for all voice $vc \in \{active, passive\}$ do
        draw $\theta^{order}_{p,vc} \sim Dirichlet(\alpha^{order})$
    for all interval $I$ do
        draw $\theta^{SR}_{p,I} \sim Dirichlet(\alpha^{SR})$
        for all adjacency $adj \in \{0, 1\}$ do
            draw $\theta^{STOP}_{p,I,adj} \sim Beta(\alpha^{STOP})$
    for all role $r \in PR \cup SR$ do
        for all feature type $f$ do
            draw $\theta^{F}_{p,r,f} \sim Dirichlet(\alpha^{F})$
——————— DATA ———————
given a predicate $p$ with voice $vc$:
choose an ordering $o \sim Multinomial(\theta^{order}_{p,vc})$
for all interval $I \in o$ do
    draw an indicator $s \sim Binomial(\theta^{STOP}_{p,I,0})$
    while $s \neq STOP$ do
        choose a SR $r \sim Multinomial(\theta^{SR}_{p,I})$
        draw an indicator $s \sim Binomial(\theta^{STOP}_{p,I,1})$
for all generated roles $r$ do
    for all feature type $f$ do
        choose a value $v_f \sim Multinomial(\theta^{F}_{p,r,f})$
```

2 Proposed Model

We assume the roles to be predicate-specific. We begin by introducing a few terms:

Primary Role (PR) For every predicate, we assume the existence of K primary roles (PRs) denoted by $P_1, P_2, ..., P_K$. These roles are not allowed to repeat in a frame and serve as "anchor points" in the global role ordering. Intuitively, the model attempts to choose PRs such that they occur with high frequency, do not repeat, and their ordering influences the positioning of other roles. Note that a PR may correspond to either a core role or a modifier role. For ease of explication, we create 3 additional PRs: $START$ denoting the start of the role sequence, END denoting its end, and $PRED$ denoting the predicate.

Secondary Role (SR) The roles that are not PRs are called secondary roles (SRs). Given N roles in total, there are $(N - K)$ SRs, denoted by $S_1, S_2, ..., S_{N-K}$. Unlike PRs, SRs are not constrained to occur only once in a frame and do not participate in the global role ordering.

Interval An interval is a sequence of SRs bounded by PRs, for instance $(P_2, S_3, S_5, PRED)$.

Ordering An ordering is the sequence of PRs observed in a frame. For example, if the complete role sequence is $(START, P_2, S_1, S_1, PRED, S_3, END)$, the ordering is defined as $(START, P_2, PRED, END)$.

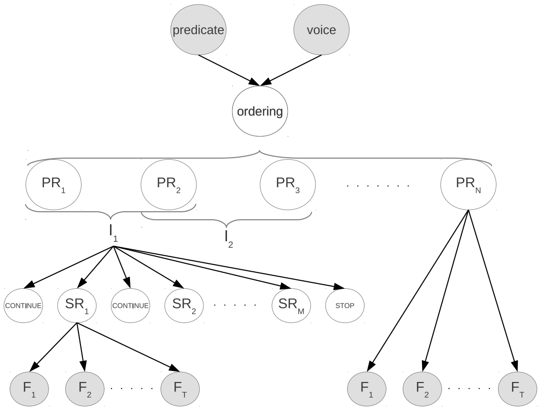

Figure 1: Proposed model. Shaded and unshaded nodes represent visible and hidden variables resp.

Features We have explored 1 frame level (global) feature (i) *voice*: active/passive, and 3 argument level (local) features (i) *deprel*: dependency relation of an argument to its head in the dependency parse tree, (ii) *head*: head word of the argument, and (iii) *pos-head*: Part-of-Speech tag of *head*.

Algorithm 1 describes the generative story of our model and Figure 1 illustrates it graphically. Given a predicate and its voice, an ordering is selected from a multinomial. This ordering gives us the sequence of PRs $(PR_1, PR_2, ..., PR_N)$. Each pair of consecutive PRs, PR_i, PR_{i+1}, in an ordering corresponds to an interval I_i. For each such interval, we generate 0 or more SRs $(SR_{i1}, SR_{i2}, ...SR_{iM})$ as follows. Generate an indicator variable: $CONTINUE/STOP$ from a binomial distribution. If $CONTINUE$, generate a SR from the multinomial corresponding to the interval. Generate another indicator variable and continue the process till a $STOP$ has been generated. In addition to the interval, the indicator variable also depends on whether we are generating the first SR ($adj = 0$) or a subsequent one ($adj = 1$). For each role, primary as well as secondary, we now generate the corresponding constituent by generating each of its features independently $(F_1, F_2, ..., F_T)$.

Given a frame instance with predicate p and voice vc, Figure 2 gives (i) Eq. 1: the joint distribution of the ordering o, role sequence $\mathbf{r}$, and constituent sequence $\mathbf{f}$, and (ii) Eq. 2: the marginal distribution of an instance. The likelihood of the whole corpus is the product of marginals of individual instances.

$$P(o,\mathbf{r},\mathbf{f}|p,vc) = \underbrace{P(o|p,vc)}_{\text{ordering}} \quad * \quad \underbrace{\Pi_{\{r_i\in\mathbf{r}\cap PR\}} P(f_i|r_i,p)}_{\text{Primary Roles}} \quad * \quad \underbrace{\Pi_{\{I\in o\}} P(\mathbf{r}(I),\mathbf{f}(I)|I,p)}_{\text{Intervals}} \tag{1}$$

$$\text{where} \quad P(\mathbf{r}(I),\mathbf{f}(I)|I,p) = \prod_{r_i\in\mathbf{r}(I)} \underbrace{P(continue|I,p,adj)}_{\text{generate indicator}} \underbrace{P(r_i|I,p)}_{\text{generate SR}} \underbrace{P(f_i|r_i,p)}_{\text{generate features}} \quad * \quad \underbrace{P(stop|I,p,adj)}_{\text{end of the interval}}$$

$$\text{and} \qquad P(f_i|r_i,p) = \Pi_t P(f_{i,t}|r_i,p)$$

$$P(\mathbf{f}|p,vc) = \Sigma_o \Sigma_{\{\mathbf{r}\in seq(o)\}} P(o,\mathbf{r},\mathbf{f}|p,vc) \qquad \text{where } seq(o) = \{\text{role sequences allowed under ordering } o\} \tag{2}$$

Figure 2: r_i and f_i denote the role and features at position i respectively, and $\mathbf{r}(I)$ and $\mathbf{f}(I)$ respectively denote the SR sequence and feature sequence in interval I. $f_{i,t}$ denotes the value of feature t at position i.

This particular choice of model is inspired from different sources. Firstly, making the role ordering dependent only on PRs aligns with the observation by Pradhan et al. (2005) and Toutanova et al. (2008) that including the ordering information of only core roles helped improve the SRL performance as opposed to the complete role sequence. Although our assumption here is softer in that we assume the existence of some roles which define the ordering which may or may not correspond to core roles. Secondly, generating the SRs independently of each other given the interval is based on the intuition that knowing the core roles informs us about the expected non-core roles that occur between them. This intuition is supported by the statistics in the annotated data, where we found that if we consider the core roles as PRs, then most of the intervals tend to have only a few types of SRs and a given SR tends to occur only in a few types of intervals. The concept of intervals is also related to the linguistic theory of topological fields (Diderichsen, 1966; Drach, 1937). This simplifying assumption that given the PRs at the interval boundary, the SRs in that interval are independent of the other roles in the sequence, keeps the parameter space limited, which helps unsupervised learning. Thirdly, not allowing some or all roles to repeat has been employed as a useful constraint in previous work (Punyakanok et al., 2004; Lang and Lapata, 2011b), which we use here for PRs. Lastly, conditioning the ($STOP/CONTINUE$) indicator variable on the adjacency value (adj) is inspired from the DMV model (Klein and Manning, 2004) for unsupervised dependency parsing. We found in the annotated corpus that if we map core roles to PRs, then most of the time the intervals do not generate any SRs at all. So, the probability to $STOP$ should be very high when generating the first SR.

We use an EM procedure to train the model. In the E-step, we calculate the expected counts of all the hidden variables in our model using the Inside-Outside algorithm (Baker, 1979). In the M-step, we add the counts corresponding to the Bayesian priors to the expected counts and use the resulting counts to calculate the MAP estimate of the parameters.

3 Experiments

Following the experimental settings of Lang and Lapata (2011b), we use the CoNLL 2008 shared task dataset (Surdeanu et al., 2008), only consider verbal predicates, and run unsupervised training on the standard training set. The evaluation measures are also the same: (i) Purity (PU) that measures how well an induced cluster corresponds to a single gold role, (ii) Collocation (CO) that measures how well a gold role corresponds to a single induced cluster, and (iii) F1 which is the harmonic mean of PU and CO. Final scores are computed by weighting each predicate by the number of its argument instances. We chose a uniform Dirichlet prior with concentration parameter as 0.1 for all the model parameters in Algorithm 1 (set roughly, without optimization[1]). 50 training iterations were used.

3.1 Results

Since the dataset has 21 semantic roles in total, we fix the total number of roles in our model to be 21. Further, we set the number of PRs to 2 (excluding $START$, END and $PRED$), and SRs to 21-2=19.

[1] Removing the Bayesian priors completely, resulted in the EM algorithm getting to a local maxima quite early, giving a substantially lower performance.

	Model	**Features**	**PU**	**CO**	**F1**
0	Baseline[2]	*d*	81.6	78.1	79.8
1a	Proposed	*d*	82.3	78.6	80.4
1b	Proposed	*d,h*	82.7	77.2	79.9
1c	Proposed	*d,p-h*	83.5	78.5	80.9
1d	Proposed	*d,p-h,h*	83.2	77.1	80.0

Table 1: Evaluation. *d* refers to *deprel*, *h* refers to *head* and *p-h* refers to *pos-head*.

# PRs	**PU**	**CO**	**F1**
0	81.67	78.07	79.83
1	82.91	78.99	80.90
2	83.54	78.47	80.93
3	83.68	78.23	80.87
4	83.72	78.08	80.80

Table 2: Performance variation with the number of PRs (excluding $START$, END and $PRED$)

Table 1 gives the results using different feature combinations. Line 0 reports the performance of Lang and Lapata (2011b)'s baseline, which has been shown difficult to outperform. This baseline maps 20 most frequent *deprel* to a role each, and the rest are mapped to the 21st role. By just using *deprel* as a feature, the proposed model outperforms the baseline by 0.6 points in terms of F1 score. In this configuration, the only addition over the baseline is the ordering model. Adding *head* as a feature leads to sparsity, which results in a substantial decrease in collocation (lines 1b and 1d). However, just adding *pos-head* (line 1c) does not cause this problem and gives the best F1 score. To address sparsity, we induced a distributed hidden representation for each word via a neural network, capturing the semantic similarity between words. Preliminary experiments improved the F1 score when using this word representation as a feature instead of the word directly.

Lang and Lapata (2011b) give the results of three methods on this task. In terms of F1 score, the *Latent Logistic* and *Graph Partitioning* methods result in slight reduction in performance over the baseline, while the *Split-Merge* method results in an improvement of 0.6 points. Table 1, line 1c achieves an improvement of 1.1 points over the baseline.

3.2 Further Evaluation

Table 2 shows the variation in performance w.r.t. the number of PRs[3] in the best performing configuration (Table 1, line 1c). On one extreme, when there are 0 PRs, there are only two possible intervals: $(START, PRED)$ and $(PRED, END)$ which means that the only context information a SR has is whether it is to the left or right of the predicate. With only this additional ordering information, the performance is the same as the baseline. Adding just 1 PR leads to a big increase in both purity and collocation. Increasing the number of PRs beyond 1 leads to a gradual increase in purity and decline in collocation, with the best F1 score at 2 PRs. This behavior could be explained by the fact that increasing the number of PRs also increases the number of intervals, which makes the probability distributions more sparse. In the extreme case, where all the roles are PRs and there are no SRs, the model would just learn the complete sequence of roles, which would make the parameter space too large to be tractable.

For calculating purity, each induced cluster (or role) is mapped to a particular gold role that has the maximum instances in the cluster. Analyzing the output of our model (line 1c in Table 1), we found that about 98% of the PRs and 40% of the SRs got mapped to the gold core roles ($A0$,$A1$, etc.). This suggests that the model is indeed following the intuition that (i) the ordering of core roles is important information for SRL systems, and (ii) the intervals bounded by core roles provide good context information for classification of other roles.

4 Conclusions

We propose a unified generative model for unsupervised semantic role induction that incorporates global role correlations as well as local feature information. The results indicate that a small number of ordered primary roles (PRs) is a good representation of global ordering constraints for SRL. This representation keeps the parameter space small enough for unsupervised learning.

Acknowledgments

This work was funded by the Swiss NSF grant 200021_125137 and EC FP7 grant PARLANCE.

[2] The baseline F1 reported by Lang and Lapata (2011b) is 79.5 due to a bug in their system (personal communication).

[3] Note that the system might not use all available PRs to label a given frame instance. #PRs refers to the max #PRs.

References

J.K. Baker. 1979. Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America*, 65:S132.

P. Diderichsen. 1966. *Elementary Danish Grammar*. Gyldendal, Copenhagen.

E. Drach. 1937. *Grundstellung der Deutschen Satzlehre*. Diesterweg, Frankfurt.

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

T. Grenager and C.D. Manning. 2006. Unsupervised discovery of a statistical verb lexicon. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 1–8. Association for Computational Linguistics.

D. Klein and C.D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 478. Association for Computational Linguistics.

J. Lang and M. Lapata. 2011a. Unsupervised semantic role induction via split-merge clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon*.

J. Lang and M. Lapata. 2011b. Unsupervised semantic role induction with graph partitioning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1320–1331, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

L. Màrquez, X. Carreras, K.C. Litkowski, and S. Stevenson. 2008. Semantic role labeling: an introduction to the special issue. *Computational linguistics*, 34(2):145–159.

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

S. Pradhan, K. Hacioglu, V. Krugler, W. Ward, J.H. Martin, and D. Jurafsky. 2005. Support vector learning for semantic argument classification. *Machine Learning*, 60(1):11–39.

V. Punyakanok, D. Roth, W. Yih, and D. Zimak. 2004. Semantic role labeling via integer linear programming inference. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1346. Association for Computational Linguistics.

M. Surdeanu, R. Johansson, A. Meyers, L. Màrquez, and J. Nivre. 2008. The conll-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177. Association for Computational Linguistics.

C. Thompson, R. Levy, and C. Manning. 2003. A generative model for semantic role labeling. *Machine Learning: ECML 2003*, pages 397–408.

K. Toutanova, A. Haghighi, and C.D. Manning. 2008. A global joint model for semantic role labeling. *Computational Linguistics*, 34(2):161–191.