Unsupervised frame based semantic role induction: application to French and English

Alejandra Lorenzo
Lorraine University/LORIA Nancy
alejandra.lorenzo@loria.fr

Christophe Cerisara
CNRS/LORIA Nancy
christophe.cerisara@loria.fr

Abstract

This paper introduces a novel unsupervised approach to semantic role induction that uses a generative Bayesian model. To the best of our knowledge, it is the first model that jointly clusters syntactic verbs arguments into semantic roles, and also creates verbs classes according to the syntactic frames accepted by the verbs. The model is evaluated on French and English, outperforming, in both cases, a strong baseline. On English, it achieves results comparable to state-of-the-art unsupervised approaches to semantic role induction.

1 Introduction and background

Semantic Role Labeling (SRL) is a major task in Natural Language Processing which provides a shallow semantic parsing of a text. Its primary goal is to identify and label the semantic relations that hold between predicates (typically verbs), and their associated arguments (Màrquez et al., 2008).

The extensive research carried out in this area resulted in a variety of annotated resources, which, in time, opened up new possibilities for supervised SRL systems. Although such systems show very good performance, they require large amounts of annotated data in order to be successful. This annotated data is not always available, very expensive to create and often domain specific (Pradhan et al., 2008). There is in particular no such data available for French. To bypass this shortcoming, "annotation-by-projection" approaches have been proposed (Pado and Lapata, 2006) which in essence, (i) project the semantic annotations available in one language (usually English), to text in another language (in this case French); and (ii) use the resulting annotations to train a semantic role labeller. Thus Pado and Pitel (2007) show that the projection-based annotation framework permits bootstrapping a semantic role labeller for FrameNet which reaches an F-measure of 63%; and van der Plas et al. (2011) show that training a joint syntactic-semantic parser based on the projection approach permits reaching an F-measure for the labeled attachment score on PropBank annotation of 65%.

Although they minimize the manual effort involved, these approaches still require both an annotated source corpus and an aligned target corpus. Moreover, they assume a specific role labeling (e.g., PropBank, FrameNet or VerbNet roles) and are not generally portable from one framework to another.

These drawbacks with supervised approaches motivated the need for unsupervised methods capable of exploiting large amounts of unannotated data. In this context several approaches have been proposed. Swier and Stevenson (2004) were the first to introduce unsupervised SRL in an approach that used the VerbNet lexicon to guide unsupervised learning. Grenager and Manning (2006) proposed a directed graphical model for role induction that exploits linguistic priors for syntactic and semantic inference. Following this work, Lang and Lapata (2010) formulated role induction as the problem of detecting alternations and mapping non-standard linkings to cannonical ones, and later as a graph partitioning problem in (Lang and Lapata, 2011b). They also proposed an algorithm that uses successive splits and merges of semantic roles clusters in order to improve

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 30–35,
Jeju, Republic of Korea, 12 July 2012.

their quality in (Lang and Lapata, 2011a). Finally, Titov and Klementiev (2012), introduce two new Bayesian models that treat unsupervised role induction as the clustering of syntactic argument signatures, with clusters corresponding to semantic roles, and achieve the best state-of-the-art results.

In this paper, we propose a novel unsupervised approach to semantic role labeling that differs from previous work in that it integrates the notion of verb classes into the model (by analogy with VerbNet, we call these verb classes, frames). We show that this approach gives good results both on the English PropBank and on a French corpus annotated with VerbNet style semantic roles. For the English PropBank, although the model is more suitable for a framework that uses a shared set of role labels such as VerbNet, we obtain results comparable to the state-of-the-art. For French, the model is shown to outperform a strong baseline by a wide margin.

2 Probabilistic Model

As mentioned in the introduction, semantic role labeling comprises two sub-tasks: argument identification and role induction. Following common practice (Lang and Lapata, 2011a; Titov and Klementiev, 2012), we assume oracle argument identification and focus on argument labeling. The approach we propose is an unsupervised generative Bayesian model that clusters arguments into classes each of which can be associated with a semantic role. The model starts by generating a frame assignment to each verb instance where a frame is a clustering of verbs and associated roles. Then, for each observed verb argument, a semantic role is drawn conditioned on the frame. Finally, the word and dependency label of this argument are generated. The model admits a simple Gibbs algorithm where the number of latent variables is proportional to the number of roles and frames to be clustered.

There are two key benefits of this model architecture. First, it directly encodes linguistic intuitions about semantic frames: the model structure reflects the subcategorisation property of the frame variable, which also groups verbs that share the same set of semantic roles, something very close to the VerbNet notion of frames. Second, by ignoring the "verb-specific" nature of PropBank labels, we reduce the need for a large amount of data and we better share evidence across roles.

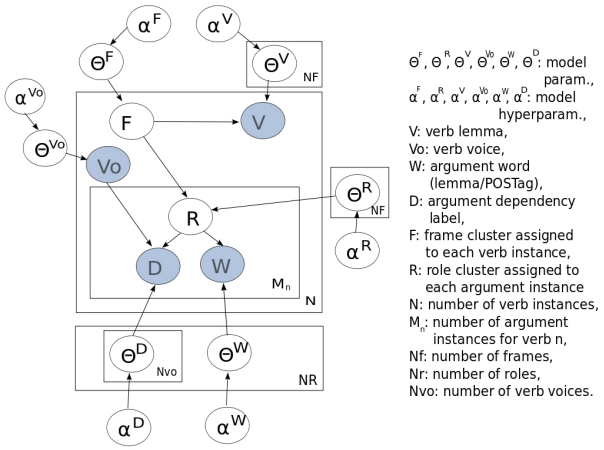

Figure 1: Plate diagram of the proposed directed Bayesian model.

In addition, because it is unsupervised, the model is independent both of the language and of the specific semantic framework (since no inventory of semantic role is a priori chosen).

2.1 Model description

The goal of the task is to assign argument instances to clusters, such that each argument cluster represents a specific semantic role, and each role corresponds to one cluster. The model is represented in the form of a plate diagram in Figure 1. The observed random variables are the verb V (lemma), its voice Vo (active or passive), the words W (lemma) that are arguments of this verb, and the syntactic dependency labels D that link the argument to its head. There are two latent variables: the frame F that represents the class of the verb, and the role R assigned to each of its arguments. The parameters θ of all multinomial distributions are Dirichlet distributed, with fixed symmetric concentration hyper-parameter α. The frame plays a fundamental role in this setting, since it intends to capture classes of verbs that share similar distributions of role arguments.

The model's generative story is described next, followed by a description of the inference algorithm used to apply the model to an unannotated corpus.

2.2 Generative story

For each verb instance, the proposed model first generates a frame cluster, a voice (active or passive), and

then a verb lemma from the distribution of verbs in this frame. The number of arguments is assumed fixed. For each argument, a role is sampled conditioned on the frame. Then, a word is sampled from the distribution of words associated to this role, and finally a dependency label is generated, conditioned both on the role and the voice. All multinomial parameters are collapsed, and thus not sampled. All Dirichlet hyper-parameters are assumed constant.

To identify words, we use either word lemmas or part-of-speech tags. In order to avoid data sparseness issues, we consider the word lemma only in cases where there are more than 9 instances of the word lemma in the corpus. Otherwise, if the number of word lemma instances is less than 10, we use the part-of-speech tags.

2.3 Learning and Inference

A collapsed Gibbs sampler is used to perform posterior inference on the model. Initially, all frames F_i are sampled randomly from a uniform distribution, while the roles $R_{i,j}$ are assigned either randomly or following the deterministic syntactic function baseline, which simply clusters predicate arguments according to their syntactic function. This function is described in detail in Section 3.

The Gibbs sampling algorithm samples each latent variable (F_i and $R_{i,j}$) in turn according to its posterior distribution conditioned on all other instances of this variable (noted $F_{\neg i}$ and $R_{\neg(i,j)}$ respectively) and all other variables. These posteriors are detailed next.

In the following, $R_{i,j}$ represents the random variable for the jth role of the ith verb in the corpus: its value is $R_{i,j} = r_{i,j}$ at a given iteration of the sampling algorithm. $nr_{f,r}$ is the count of occurrences of $(F_i = f, R_{i,j} = r)$ in the whole corpus, excluding the ith instance when the superscript $^{-i}$ is used. A star $*$ matches any possible value. The joint probability over the whole corpus with collapsed multinomial parameters is:

$$
\begin{aligned}
& p(F, R, V, W, D, Vo|\alpha) \\
&= \frac{\prod_{i=1}^{N_f} \Gamma(nf_i + \alpha^F)}{\Gamma(\sum_{i=1}^{N_f} nf_i + \alpha^F)} \frac{\Gamma(\sum_{i=1}^{N_f} \alpha^F)}{\prod_{i=1}^{N_f} \Gamma(\alpha^F)} \times \\
&\quad \prod_{i=1}^{N_f} \frac{\prod_{j=1}^{N_v} \Gamma(nv_{i,j} + \alpha^V)}{\Gamma(\sum_{j=1}^{N_v} nv_{i,j} + \alpha^V)} \frac{\Gamma(\sum_{j=1}^{N_v} \alpha^V)}{\prod_{j=1}^{N_v} \Gamma(\alpha^V)} \times \\
&\quad \prod_{i=1}^{N_f} \frac{\prod_{j=1}^{N_r} \Gamma(nr_{i,j} + \alpha^R)}{\Gamma(\sum_{j=1}^{N_r} nr_{i,j} + \alpha^R)} \frac{\Gamma(\sum_{j=1}^{N_r} \alpha^R)}{\prod_{j=1}^{N_r} \Gamma(\alpha^R)} \times \\
&\quad \prod_{i=1}^{N_{vo}} \prod_{j=1}^{N_r} \frac{\prod_{k=1}^{Nd} \Gamma(nd_{i,j,k} + \alpha^D)}{\Gamma(\sum_{k=1}^{N_d} nd_{i,j,k} + \alpha^D)} \frac{\Gamma(\sum_{k=1}^{N_d} \alpha^D)}{\prod_{k=1}^{N_d} \Gamma(\alpha^D)} \times \\
&\quad \prod_{i=1}^{Nr} \frac{\prod_{j=1}^{Nw} \Gamma(nw_{i,j} + \alpha^W)}{\Gamma(\sum_{j=1}^{N_w} nw_{i,j} + \alpha^W)} \frac{\Gamma(\sum_{j=1}^{N_w} \alpha^W)}{\prod_{j=1}^{Nw} \Gamma(\alpha^W)} \times \\
&\quad \frac{\prod_{i=1}^{N_{vo}} \Gamma(nvo_i + \alpha^{Vo})}{\Gamma(\sum_{i=1}^{N_{vo}} nvo_i + \alpha^{Vo})} \frac{\Gamma(\sum_{i=1}^{N_{vo}} \alpha^{Vo})}{\prod_{i=1}^{N_{vo}} \Gamma(\alpha^{Vo})}
\end{aligned}
$$

The posterior from which the frame is sampled is derived from the joint distribution as follows:

$$
\begin{aligned}
& p(F_i = y|F_{\neg i}, R, V, W, Vo) \qquad (1) \\
&\propto \frac{p(F, R, V, W, D, Vo)}{p(F_{\neg i}, R_{\neg i}, V_{\neg i}, W_{\neg i}, D_{\neg i}, Vo_{\neg i})} \\
&= \frac{(nf_y^{-i} + \alpha^F)}{(\sum_{z=1}^{N_f} nf_z^{-i} + \alpha^F)} \times \frac{(nv_{y,v_i}^{-i} + \alpha^V)}{(\sum_{j=1}^{N_v} nv_{y,j}^{-i} + \alpha^V)} \times \\
&\quad \frac{\prod_{r \in r_{i,*}} \prod_{x=0}^{nr_r^{+i}-1} (nr_{y,r}^{-i} + \alpha^R + x)}{\prod_{x=0}^{M_i} (\sum_{r=1}^{N_r} nr_{y,r}^{-i} + \alpha^R + x)}
\end{aligned}
$$

where nr_r^{+i} is the count of occurrences of role r in the arguments of verb instance i ($M_i = \sum_r nr_r^{+i}$).

The update equation for sampling the role becomes:

$$
\begin{aligned}
& p(R_{i,j} = y|R_{\neg(i,j)}, F, V, W, D, Vo) \qquad (2) \\
&\propto \frac{p(F, R, V, W, D, Vo)}{p(F_{\neg i}, V_{\neg i}, R_{\neg(i,j)}, W_{\neg(i,j)}, D_{\neg(i,j)}, Vo_{\neg(i,j)})} \\
&= \frac{(nr_{f_i,y}^{-(i,j)} + \alpha^R)}{(\sum_{k=1}^{N_r} nr_{f_i,k}^{-(i,j)} + \alpha^R)} \times \frac{(nd_{vo_i,y,d_{i,j}}^{-(i,j)} + \alpha^D)}{(\sum_{k=1}^{N_d} nd_{vo_i,y,k}^{-(i,j)} + \alpha^D)} \times \\
&\quad \frac{(nw_{y,w_{i,j}}^{-(i,j)} + \alpha^W)}{(\sum_{k=1}^{N_w} nw_{y,k}^{-(i,j)} + \alpha^W)}
\end{aligned}
$$

After T iterations, the process is stopped and the expected value of the sampled frames and roles after the burn-in period (20 iterations) is computed. With deterministic (syntactic) initialization, T is set to 200, while it is set to 2000 with random initialization because of slower convergence.

3 Evaluations and results

We evaluate our model both on English to situate our approach with respect to the state of the art; and on French to demonstrate its portability to other languages.

3.1 Common experimental setup

The model's parameters have been tuned with a few rounds of trial-and-error on the English development corpus: For the hyper-parameters, we set

$\alpha^F = 0.5$, $\alpha^R = 1.e^{-3}$, $\alpha^V = 1.e^{-7}$, $\alpha^{Vo} = 1.e^{-3}$, $\alpha^D = 1.e^{-8}$ and $\alpha^W = 0.5$. For the evaluation on French, we only changed the α^F and α^W parameters. In order to reflect the rather uniform distribution of verb instances across verb classes we set α^F to 1. Moreover, we set α^W to 0.001 because of the smaller number of words and roles in the French corpus. The number of roles and frames were chosen based on the properties of each corpus. We set number of roles to 40 and 10, and the number of frames to 300 and 60 for English and French respectively. As done in (Lang and Lapata, 2011a) and (Titov and Klementiev, 2012), we use purity and collocation measures to assess the quality of our role induction process. For each verb, the purity of roles' clusters is computed as follows:

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

where C_i is the set of arguments in the ith cluster found, G_j is the set of arguments in the jth gold class, and N is the number of argument instances. In a similar way, the collocation of roles' clusters is computed as follows:

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

Then, each score is averaged over all verbs. In the same way as (Lang and Lapata, 2011a), we use the micro-average obtained by weighting the scores for individual verbs proportionally to the number of argument instances for that verb. Finally the F1 measure is the harmonic mean of the aggregated values of purity and collocation:

$$F1 = \frac{2 * CO * PU}{CO + PU}$$

3.2 Evaluations on French

To evaluate our model on French, we used a manually annotated corpora consisting on sentences from the Paris 7 Treebank (Abeillé et al., 2000), containing verbs extracted from the gold standard V-GOLD (Sun et al., 2010)[1]. For each verb, at most 25 sentences from the Paris 7 Treebank were randomly selected and annotated with VerbNet-style thematic roles. In some cases, the annotated roles were obtained by merging some of the VerbNet roles (e.g., Actor, Actor1 and Actor2 are merged); or by grouping together classes sharing the same thematic grids. The resulting roles assignment groups 116 verbs into 12 VerbNet classes, each associated with a unique thematic grid. Table 1 shows the set of roles used and their relation to VerbNet roles. This constitutes our gold evaluation corpus.

Role	**VerbNet roles**
Agent	Agent, Actor, Actor1, Actor2
Experiencer	Experiencer
Theme	Stimulus, Theme, Theme1, Theme2
Topic	Proposition, Topic
PredAtt	Predicate, Attribute
Patient	Patient, Patient1, Patient2
Start	Material, Source
End	Product, Destination, Recipient
Location	Location
Instrument	Instrument
Cause	Cause
Beneficiary	Beneficiary
Extent	Asset, Extent, Time, Value

Table 1: VerbNet role groups (French).

The baseline model is the "syntactic function" used for instance in (Lang and Lapata, 2011a), which simply clusters predicate arguments according to the dependency relation to their head. This is a standard baseline for unsupervised SRL, which, although simple, has been shown difficult to outperform. As done in previous work, it is implemented by allocating a different cluster to each of the 10 most frequent syntactic relations, and one extra cluster for all the other relations. Evaluation results are shown in Table 2. The proposed model significantly outperforms the deterministic baseline, which validates the unsupervised learning process.

	PU	CO	F1
Synt.Func. (baseline)	78.9	73.4	76.1
Proposed model - rand. init	74.6	82.9	78.5

Table 2: Comparison of the Syntactic Function baseline with the proposed system initialized randomly, evaluated with gold parses and argument identification (French).

3.3 Evaluations on English

We made our best to follow the setup used in previous work (Lang and Lapata, 2011a; Titov and Kle-

[1] V-GOLD consists of 16 fine grained Levin classes with 12 verbs each (translated to French) whose predominant sense in English belong to that class.

mentiev, 2012), in order to compare with the current state of the art.

The data used is the standard CoNLL 2008 shared task (Surdeanu et al., 2008) version of Penn Treebank WSJ and PropBank. Our model is evaluated on gold generated parses, using the gold PropBank annotations. In PropBank, predicates are associated with a set of roles, where roles A2-A5 or AA are verb specific, while adjuncts roles (AM) are consistent across verbs. Besides, roles A0 and A1 attempt to capture Proto-Agent and Proto-Patient roles (Dowty, 1991), and thus are more valid across verbs and verb instances than A2-A5 roles.

Table 3 reports the evaluation results of the proposed model along with those of the baseline system and of some of the latest state-of-the-art results.

	PU	CO	F1
Synt.Func.(LL)	81.6	77.5	79.5
Split Merge	88.7	73.0	80.1
Graph Part.	88.6	70.7	78.6
TK-Bay.1	88.7	78.1	83.0
TK-Bay.2	89.2	74.0	80.9
Synt.Func.	79.6	84.6	82.0
Proposed model - rand. init	82.2	83.4	82.8
Proposed model - synt. init	83.4	84.1	83.7

Table 3: Comparison of the proposed system (last 2 rows) with other unsupervised semantic role inducers evaluated on gold parses and argument identification.

We can first note that, despite our efforts to reproduce the same baseline, there is still a difference between our baseline (Synt.Func.) and the baseline reported in (Lang and Lapata, 2011a) (Synt.Func.(LL)) [2].

The other results respectively correspond to the Split Merge approach presented in (Lang and Lapata, 2011a) (Split Merge), the Graph Partitioning algorithm (Graph Part.) presented in (Lang and Lapata, 2011b), and two Bayesian approaches presented in (Titov and Klementiev, 2012), which achieve the best current unsupervised SRL results. The first such model (TK-Bay.1) clusters argument fillers and directly maps some syntactic labels to semantic roles for some adjunct like modifiers that are explicitly represented in the syntax, while the second model (TK-Bay.2) does not include these two features.

[2]We identified afterwards a few minor differences in both experimental setups that partly explain this, e.g., evaluation on the test vs. train sets, finer-grained gold classes in our case...

Two versions of the proposed model are reported in the last rows of Table 3: one with random (uniform) initialization of all variables, and the other with deterministic initialization of all R_i from the syntactic function. Indeed, although many unsupervised system are very sensitive to initialization, we observe that in the proposed model, unsupervised inference reaches reasonably good performances even with a knowledge-free initialization. Furthermore, when initialized with the strong deterministic baseline, the model still learns new evidences and improves over the baseline to give comparable results to the best unsupervised state-of-the-art systems.

4 Conclusions and future work

We have presented a method for unsupervised SRL that is based on an intuitive generative Bayesian model that not only clusters arguments into semantic roles, but also explicitly integrates the concept of frames in SRL. Previous approaches to semantic role induction proposed some clustering of roles without explicitly focusing on the verb classes generated. Although there has been work on verb clustering, this is, to the best of our knowledge, the first approach that jointly considers both tasks.

In this work in progress, we focused on the role induction task and we only evaluated this part, leaving the evaluation of verb classes as future work. We successfully evaluated the proposed model on two languages, French and English, showing, in both cases, consistent performances improvement over the deterministic baseline. Furthermore, its accuracy reaches a level comparable to that of the best state-of-the-art unsupervised systems.

The model could be improved in many ways, and in particular by including some penalization term for sampling the same role for several arguments of a verb instance (at least for core roles). Moreover, we believe that our model better fits within a framework that allows roles sharing between verbs (or frames), such as VerbNet, and we would like to carry out a deeper evaluation on this concept.

Acknowledgments

The authors wish to thank Claire Gardent for her valuable suggestions and Ingrid Falk for providing the data for the evaluation on French.

References

A. Abeillé, L. Clément, and A. Kinyon. 2000. Building a treebank for French. In *Proceedings of the LREC 2000*.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67:547–619.

Trond Grenager and Christopher D. Manning. 2006. Unsupervised discovery of a statistical verb lexicon. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joel Lang and Mirella Lapata. 2010. Unsupervised induction of semantic roles. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 939–947, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joel Lang and Mirella Lapata. 2011a. Unsupervised semantic role induction via split-merge clustering. In *ACL*, pages 1117–1126. Association for Computer Linguistics.

Joel Lang and Mirella Lapata. 2011b. Unsupervised semantic role induction with graph partitioning. In *EMNLP*, pages 1320–1331. Association for Computer Linguistics.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: an introduction to the special issue. *Comput. Linguist.*, 34(2):145–159, June.

Sebastian Pado and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of ACL-COLING 2006*, pages 1161–1168, Sydney, Australia.

Sebastian Pado and Guillaume Pitel. 2007. Annotation précise du franais en sémantique de rôles par projection cross-linguistique. In *Proceedings of TALN-07*, Toulouse, France.

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

L. Sun, A. Korhonen, T. Poibeau, and C. Messiant. 2010. Investigating the cross-linguistic potential of VerbNet-style classification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1056–1064, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim 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*, CoNLL '08, pages 159–177, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robert S. Swier and Suzanne Stevenson. 2004. Unsupervised Semantic Role Labelling. In *EMNLP*, pages 95–102. Association for Computational Linguistics.

Ivan Titov and Alexandre Klementiev. 2012. A bayesian approach to unsupervised semantic role induction. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, April.

Lonneke van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up cross-lingual semantic annotation transfer. In *Proceedings of ACL/HLT*, pages 299–304.