A Structured Model for Joint Learning of Argument Roles and Predicate Senses

Yotaro Watanabe
Graduate School of Information Sciences
Tohoku University
6-6-05, Aramaki Aza Aoba, Aoba-ku,
Sendai 980-8579, Japan
yotaro-w@ecei.tohoku.ac.jp

Masayuki Asahara **Yuji Matsumoto**
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma,
Nara, 630-0192, Japan
{masayu-a, matsu}@is.naist.jp

Abstract

In predicate-argument structure analysis, it is important to capture non-local dependencies among arguments and inter-dependencies between the sense of a predicate and the semantic roles of its arguments. However, no existing approach explicitly handles both non-local dependencies and semantic dependencies between predicates and arguments. In this paper we propose a structured model that overcomes the limitation of existing approaches; the model captures both types of dependencies simultaneously by introducing four types of factors including a global factor type capturing non-local dependencies among arguments and a pairwise factor type capturing local dependencies between a predicate and an argument. In experiments the proposed model achieved competitive results compared to the state-of-the-art systems without applying any feature selection procedure.

1 Introduction

Predicate-argument structure analysis is a process of assigning *who* does *what* to *whom*, *where*, *when*, etc. for each predicate. Arguments of a predicate are assigned particular *semantic roles*, such as *Agent*, *Theme*, *Patient*, etc. Lately, predicate-argument structure analysis has been regarded as a task of assigning semantic roles of arguments as well as word senses of a predicate (Surdeanu et al., 2008; Hajič et al., 2009).

Several researchers have paid much attention to predicate-argument structure analysis, and the following two important factors have been shown. Toutanova et al. (2008), Johansson and Nugues (2008), and Björkelund et al. (2009) presented importance of capturing non-local dependencies of core arguments in predicate-argument structure analysis. They used argument sequences tied with a predicate sense (e.g. AGENT-buy.01/Active-PATIENT) as a feature for the re-ranker of the system where predicate sense and argument role candidates are generated by their pipelined architecture. They reported that incorporating this type of features provides substantial gain of the system performance.

The other factor is inter-dependencies between a predicate sense and argument roles, which relate to selectional preference, and motivated us to jointly identify a predicate sense and its argument roles. This type of dependencies has been explored by Riedel and Meza-Ruiz (2008; 2009b; 2009a), all of which use Markov Logic Networks (MLN). The work uses the global formulae that have atoms in terms of both a predicate sense and each of its argument roles, and the system identifies predicate senses and argument roles simultaneously.

Ideally, we want to capture both types of dependencies simultaneously. The former approaches can not explicitly include features that capture inter-dependencies between a predicate sense and its argument roles. Though these are implicitly incorporated by re-ranking where the most plausible assignment is selected from a small subset of predicate and argument candidates, which are generated independently. On the other hand, it is difficult to deal with core argument features in MLN. Because the number of core arguments varies with the role assignments, this type of features cannot be expressed by a single formula.

Thompson et al. (2010) proposed a generative model that captures both predicate senses and its argument roles. However, the first-order markov assumption of the model eliminates ability to capture non-local dependencies among arguments. Also, generative models are in general inferior to discriminatively trained linear or log-

Proceedings of the ACL 2010 Conference Short Papers, pages 98–102,
Uppsala, Sweden, 11-16 July 2010.

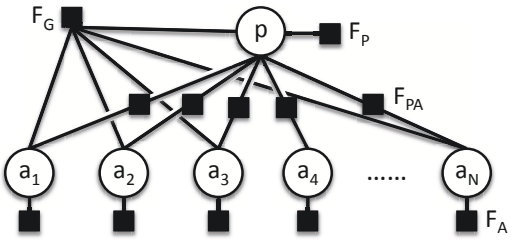

Figure 1: Undirected graphical model representation of the structured model

linear models.

In this paper we propose a structured model that overcomes limitations of the previous approaches. For the model, we introduce several types of features including those that capture both non-local dependencies of core arguments, and inter-dependencies between a predicate sense and its argument roles. By doing this, both tasks are mutually influenced, and the model determines the most plausible set of assignments of a predicate sense and its argument roles simultaneously. We present an exact inference algorithm for the model, and a large-margin learning algorithm that can handle both local and global features.

2 Model

Figure 1 shows the graphical representation of our proposed model. The node p corresponds to a predicate, and the nodes $a_1, ..., a_N$ to arguments of the predicate. Each node is assigned a particular predicate sense or an argument role label. The black squares are *factors* which provide scores of label assignments. In the model, the nodes for arguments depend on the predicate sense, and by influencing labels of a predicate sense and its argument roles, the most plausible label assignment of the nodes is determined considering all factors.

In this work, we use linear models. Let $\mathbf{x}$ be words in a sentence, p be a sense of a predicate in $\mathbf{x}$, and $\mathcal{A} = \{a_n\}_1^N$ be a set of possible role label assignments for $\mathbf{x}$. A predicate-argument structure is represented by a pair of p and $\mathcal{A}$. We define the score function for predicate-argument structures as $s(p, \mathcal{A}) = \sum_{F_k \in \mathcal{F}} F_k(\mathbf{x}, p, \mathcal{A})$. $\mathcal{F}$ is a set of all the factors, $F_k(\mathbf{x}, p, \mathcal{A})$ corresponds to a particular factor in Figure 1, and gives a score to a predicate or argument label assignments. Since we use linear models, $F_k(\mathbf{x}, p, \mathcal{A}) = \mathbf{w} \cdot \Phi_k(\mathbf{x}, p, \mathcal{A})$.

2.1 Factors of the Model

We define four types of factors for the model.

Predicate Factor F_P scores a sense of p, and does not depend on any arguments. The score function is defined by $F_P(\mathbf{x}, p, \mathcal{A}) = \mathbf{w} \cdot \Phi_P(\mathbf{x}, p)$.

Argument Factor F_A scores a label assignment of a particular argument $a \in \mathcal{A}$. The score is determined independently from a predicate sense, and is given by $F_A(\mathbf{x}, p, a) = \mathbf{w} \cdot \Phi_A(\mathbf{x}, a)$.

Predicate-Argument Pairwise Factor F_{PA} captures inter-dependencies between a predicate sense and one of its argument roles. The score function is defined as $F_{PA}(\mathbf{x}, p, a) = \mathbf{w} \cdot \Phi_{PA}(\mathbf{x}, p, a)$. The difference from F_A is that F_{PA} influences both the predicate sense and the argument role. By introducing this factor, the role label can be influenced by the predicate sense, and vise versa.

Global Factor F_G is introduced to capture plausibility of the whole predicate-argument structure. Like the other factors, the score function is defined as $F_G(\mathbf{x}, p, \mathcal{A}) = \mathbf{w} \cdot \Phi_G(\mathbf{x}, p, \mathcal{A})$. A possible feature that can be considered by this factor is the mutual dependencies among core arguments. For instance, if a predicate-argument structure has an agent (A0) followed by the predicate and a patient (A1), we encode the structure as a string *A0-PRED-A1* and use it as a feature. This type of features provide *plausibility* of predicate-argument structures. Even if the highest scoring predicate-argument structure with the other factors misses some core arguments, the global feature demands the model to fill the missing arguments.

The numbers of factors for each factor type are: F_P and F_G are 1, F_A and F_{PA} are $|\mathcal{A}|$. By integrating the all factors, the score function becomes $s(p, \mathcal{A}) = \mathbf{w} \cdot \Phi_P(\mathbf{x}, p) + \mathbf{w} \cdot \Phi_G(\mathbf{x}, p, \mathcal{A}) + \mathbf{w} \cdot \sum_{a \in \mathcal{A}} \{\Phi_A(\mathbf{x}, a) + \Phi_{PA}(\mathbf{x}, p, a)\}$.

2.2 Inference

The crucial point of the model is how to deal with the global factor F_G, because enumerating possible assignments is too costly. A number of methods have been proposed for the use of global features for linear models such as (Daumé III and Marcu, 2005; Kazama and Torisawa, 2007). In this work, we use the approach proposed in (Kazama and Torisawa, 2007). Although the approach is proposed for sequence labeling tasks, it

can be easily extended to our structured model. That is, for each possible predicate sense p of the predicate, we provide N-best argument role assignments using three local factors F_P, F_A and F_{PA}, and then add scores of the global factor F_G, finally select the argmax from them. In this case, the argmax is selected from $|\mathcal{P}_l|N$ candidates.

2.3 Learning the Model

For learning of the model, we borrow a fundamental idea of Kazama and Torisawa's perceptron learning algorithm. However, we use a more sophisticated online-learning algorithm based on the Passive-Aggressive Algorithm (PA) (Crammer et al., 2006).

For the sake of simplicity, we introduce some notations. We denote a predicate-argument structure $\mathbf{y} = \langle p, \mathcal{A} \rangle$, a local feature vector as $\Phi_L(\mathbf{x}, \mathbf{y}) = \Phi_P(\mathbf{x}, p) + \sum_{a \in \mathcal{A}} \{\Phi_A(\mathbf{x}, a) + \Phi_{PA}(\mathbf{x}, p, a)\}$, a feature vector coupling both local and global features as $\Phi_{L+G}(\mathbf{x}, \mathbf{y}) = \Phi_L(\mathbf{x}, \mathbf{y}) + \Phi_G(\mathbf{x}, p, \mathcal{A})$, the argmax using Φ_{L+G} as $\hat{\mathbf{y}}^{L+G}$, the argmax using Φ_L as $\hat{\mathbf{y}}^{L}$. Also, we use a loss function $\rho(\mathbf{y}, \mathbf{y}')$, which is a cost function associated with $\mathbf{y}$ and $\mathbf{y}'$.

The margin perceptron learning proposed by Kazama and Torisawa can be seen as an optimization with the following two constrains.

(A) $\mathbf{w} \cdot \Phi_{L+G}(\mathbf{x}, \mathbf{y}) - \mathbf{w} \cdot \Phi_{L+G}(\mathbf{x}, \hat{\mathbf{y}}^{L+G}) \geq \rho(\mathbf{y}, \hat{\mathbf{y}}^{L+G})$

(B) $\mathbf{w} \cdot \Phi_L(\mathbf{x}, \mathbf{y}) - \mathbf{w} \cdot \Phi_L(\mathbf{x}, \hat{\mathbf{y}}^{L}) \geq \rho(\mathbf{y}, \hat{\mathbf{y}}^{L})$

(A) is the constraint that ensures a sufficient margin $\rho(\mathbf{y}, \hat{\mathbf{y}}^{L+G})$ between $\mathbf{y}$ and $\hat{\mathbf{y}}^{L+G}$. (B) is the constraint that ensures a sufficient margin $\rho(\mathbf{y}, \hat{\mathbf{y}}^{L})$ between $\mathbf{y}$ and $\hat{\mathbf{y}}^{L}$. The necessity of this constraint is that if we apply only (A), the algorithm does not guarantee a sufficient margin in terms of local features, and it leads to poor quality in the N-best assignments. The Kazama and Torisawa's perceptron algorithm uses constant values for the cost function $\rho(\mathbf{y}, \hat{\mathbf{y}}^{L+G})$ and $\rho(\mathbf{y}, \hat{\mathbf{y}}^{L})$.

The proposed model is trained using the following optimization problem.

$$\mathbf{w}_{new} = \arg \min_{\mathbf{w}' \in \Re^n} \frac{1}{2} ||\mathbf{w}' - \mathbf{w}||^2 + C\xi$$

$$\begin{cases} \text{s.t. } l_{L+G} \leq \xi,\ \xi \geq 0 & \text{if } \hat{\mathbf{y}}^{L+G} \neq \mathbf{y} \\ \text{s.t. } l_{L} \leq \xi,\ \xi \geq 0 & \text{if } \hat{\mathbf{y}}^{L+G} = \mathbf{y} \neq \hat{\mathbf{y}}^{L} \end{cases} \quad (1)$$

$$l_{L+G} = \mathbf{w} \cdot \Phi_{L+G}(\mathbf{x}, \hat{\mathbf{y}}^{L+G}) - \mathbf{w} \cdot \Phi_{L+G}(\mathbf{x}, \mathbf{y}) + \rho(\mathbf{y}, \hat{\mathbf{y}}^{L+G}) \quad (2)$$

$$l_{L} = \mathbf{w} \cdot \Phi_{L}(\mathbf{x}, \hat{\mathbf{y}}^{L}) - \mathbf{w} \cdot \Phi_{L}(\mathbf{x}, \mathbf{y}) + \rho(\mathbf{y}, \hat{\mathbf{y}}^{L}) \quad (3)$$

l_{L+G} is the loss function for the case of using both local and global features, corresponding to the constraint (A), and l_L is the loss function for the case of using only local features, corresponding to the constraints (B) provided that (A) is satisfied.

2.4 The Role-less Argument Bias Problem

The fact that an argument candidate is not assigned any role (namely it is assigned the label "NONE") is unlikely to contribute predicate sense disambiguation. However, it remains possible that "NONE" arguments is biased toward a particular predicate sense by F_{PA} (i.e. $\mathbf{w} \cdot \Phi_{PA}(\mathbf{x}, sense_i, a_k = \text{"NONE"}) > \mathbf{w} \cdot \Phi_{PA}(\mathbf{x}, sense_j, a_k = \text{"NONE"})$.

In order to avoid this bias, we define a special sense label, $sense_{any}$, that is used to calculate the score for a predicate and a roll-less argument, regardless of the predicate's sense. We use the feature vector $\Phi_{PA}(\mathbf{x}, sense_{any}, a_k)$ if $a_k =$ "NONE" and $\Phi_{PA}(\mathbf{x}, sense_i, a_k)$ otherwise.

3 Experiment

3.1 Experimental Settings

We use the CoNLL-2009 Shared Task dataset (Hajič et al., 2009) for experiments. It is a dataset for multi-lingual syntactic and semantic dependency parsing [1]. In the SRL-only challenge of the task, participants are required to identify predicate-argument structures of only the specified predicates. Therefore the problems to be solved are predicate sense disambiguation and argument role labeling. We use Semantic Labeled F1 for evaluation.

For generating N-bests, we used the beam-search algorithm, and the number of N-bests was set to $N = 64$. For learning of the joint model, the loss function $\rho(\mathbf{y}_t, \mathbf{y}')$ of the Passive-Aggressive Algorithm was set to the number of incorrect assignments of a predicate sense and its argument roles. Also, the number of iterations of the model used for testing was selected based on the performance on the development data.

Table 1 shows the features used for the structured model. The global features used for F_G are based on those used in (Toutanova et al., 2008; Johansson and Nugues, 2008), and the features

[1] The dataset consists of seven languages: Catalan, Chinese, Czech, English, German, Japanese and Spanish.

F_P	Plemma of the predicate and predicate's head, and ppos of the predicate
	Dependency label between the predicate and predicate's head
	The concatenation of the dependency labels of the predicate's dependents
F_A	Plemma and ppos of the predicate, the predicate's head, the argument candidate, and the argument's head
	Plemma and ppos of the leftmost/rightmost dependent and leftmost/rightmost sibling
	The dependency label of predicate, argument candidate and argument candidate's dependent
	The position of the argument candidate with respect to the predicate position in the dep. tree (e.g. CHILD)
	The position of the head of the dependency relation with respect to the predicate position in the sentence
	The left-to-right chain of the deplabels of the predicate's dependents
	Plemma, ppos and dependency label paths between the predicate and the argument candidates
	The number of dependency edges between the predicate and the argument candidate
F_{PA}	Plemma and plemma&ppos of the argument candidate
	Dependency label path between the predicate and the argument candidates
F_G	The sequence of the predicate and the argument labels in the predicate-argument structure (e.g. A0-PRED-A1)
	Whether the semantic roles defined in frames exist in the structure, (e.g. *CONTAINS:A1*)
	The conjunction of the predicate sense and the frame information (e.g. *wear.01&CONTAINS:A1*)

Table 1: Features for the Structured Model

	Avg.	Ca	Ch	Cz	En	Ge	Jp	Sp
F_P+F_A	79.17	78.00	76.02	85.24	83.09	76.76	77.27	77.83
F_P+F_A+F_{PA}	79.58	78.38	76.23	85.14	83.36	78.31	77.72	77.92
F_P+F_A+F_G	80.42	79.50	76.96	85.88	84.49	78.64	78.32	79.21
ALL	80.75	79.55	77.20	**85.94**	84.97	79.62	**78.69**	79.29
Björkelund	**80.80**	80.01	**78.60**	85.41	**85.63**	**79.71**	76.30	79.91
Zhao	80.47	**80.32**	77.72	85.19	85.44	75.99	78.15	**80.46**
Meza-Ruiz	77.46	78.00	77.73	75.75	83.34	73.52	76.00	77.91

Table 2: Results on the CoNLL-2009 Shared Task dataset (Semantic Labeled F1).

	SENSE	ARG
F_P+F_A	89.65	72.20
F_P+F_A+F_{PA}	89.78	72.74
F_P+F_A+F_G	89.83	74.11
ALL	90.15	74.46

Table 3: Predicate sense disambiguation and argument role labeling results (average).

used for F_{PA} are inspired by formulae used in the MLN-based SRL systems, such as (Meza-Ruiz and Riedel, 2009b). We used the same feature templates for all languages.

3.2 Results

Table 2 shows the results of the experiments, and also shows the results of the top 3 systems in the CoNLL-2009 Shared Task participants of the *SRL-only* system.

By incorporating F_{PA}, we achieved performance improvement for all languages. This results suggest that it is effective to capture local interdependencies between a predicate sense and one of its argument roles. Comparing the results with F_P+F_A and F_P+F_A+F_G, incorporating F_G also contributed performance improvements for all languages, especially the substantial F1 improvement of $+1.88$ is obtained in German.

Next, we compare our system with top 3 systems in the CoNLL-2009 Shared Task. By incorporating both F_{PA} and F_G, our joint model achieved competitive results compared to the top 2 systems (Björkelund and Zhao), and achieved the better results than the Meza-Ruiz's system [2]. The systems by Björkelund and Zhao applied feature selection algorithms in order to select the best set of feature templates for each language, requiring about 1 to 2 months to obtain the best feature set. On the other hand, our system achieved the competitive results with the top two systems, despite the fact that we used the same feature templates for all languages without applying any feature engineering procedure.

Table 3 shows the performances of predicate sense disambiguation and argument role labeling separately. In terms of sense disambiguation results, incorporating F_{PA} and F_G worked well. Although incorporating either of F_{PA} and F_G provided improvements of $+0.13$ and $+0.18$ on average, adding both factors provided improvements of $+0.50$. We compared the predicate sense dis-

[2]The result of Meza-Ruiz for Czech is substantially worse than the other systems because of inappropriate preprocessing for predicate sense disambiguation. Excepting Czech, the average F1 value of the Meza-Ruiz is 77.75, where as our system is 79.89.

ambiguation results of $F_P + F_A$ and ALL with the McNemar test, and the difference was statistically significant ($p < 0.01$). This result suggests that combination of these factors is effective for sense disambiguation.

As for argument role labeling results, incorporating F_{PA} and F_G contributed positively for all languages. Especially, we obtained a substantial gain ($+4.18$) in German. By incorporating F_{PA}, the system achieved the F1 improvements of $+0.54$ on average. This result shows that capturing inter-dependencies between a predicate and its arguments contributes to argument role labeling. By incorporating F_G, the system achieved the substantial improvement of F1 ($+1.91$).

Since both tasks improved by using all factors, we can say that the proposed joint model succeeded in *joint learning* of predicate senses and its argument roles.

4 Conclusion

In this paper, we proposed a structured model that captures both non-local dependencies between arguments, and inter-dependencies between a predicate sense and its argument roles. We designed a linear model-based structured model, and defined four types of factors: predicate factor, argument factor, predicate-argument pairwise factor and global factor for the model. In the experiments, the proposed model achieved competitive results compared to the state-of-the-art systems without any feature engineering.

A further research direction we are investigating is exploitation of unlabeled texts. Semi-supervised semantic role labeling methods have been explored by (Collobert and Weston, 2008; Deschacht and Moens, 2009; Fürstenau and Lapata, 2009), and they have achieved successful outcomes. However, we believe that there is still room for further improvement.

References

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *CoNLL-2009*.

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

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *JMLR*, 7:551–585.

Hal Daumé III and Daniel Marcu. 2005. Learning as search optimization: Approximate large margin methods for structured prediction. In *ICML-2005*.

Koen Deschacht and Marie-Francine Moens. 2009. Semi-supervised semantic role labeling using the latent words language model. In *EMNLP-2009*.

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

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *CoNLL-2009*, Boulder, Colorado, USA.

Richard Johansson and Pierre Nugues. 2008. Dependency-based syntactic-semantic analysis with propbank and nombank. In *CoNLL-2008*.

Jun'Ichi Kazama and Kentaro Torisawa. 2007. A new perceptron algorithm for sequence labeling with non-local features. In *EMNLP-CoNLL 2007*.

Ivan Meza-Ruiz and Sebastian Riedel. 2009a. Jointly identifying predicates, arguments and senses using markov logic. In *HLT/NAACL-2009*.

Ivan Meza-Ruiz and Sebastian Riedel. 2009b. Multilingual semantic role labelling with markov logic. In *CoNLL-2009*.

Sebastian Riedel and Ivan Meza-Ruiz. 2008. Collective semantic role labelling with markov logic. In *CoNLL-2008*.

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 *CoNLL-2008*.

Synthia A. Thompson, Roger Levy, and Christopher D. Manning. 2010. A generative model for semantic role labeling. In *Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics (to appear)*.

Kristina Toutanova, Aria Haghighi, and Christopher D. Manning. 2008. A global joint model for semantic role labeling. *Computational Linguistics*, 34(2).