A Word Embedding and a Josa Vector for Korean Unsupervised Semantic Role Induction

Kyeong-Min Nam and **Yu-Seop Kim**
Department of Convergence Software
Hallym University 1 Hallymdaehak-gil, Chuncheon,Gangwon-do, 200-702 Korea
jkre4030@naver.com, yuseop.kim@gmail.com

Abstract

We propose an unsupervised semantic role labeling method for Korean language, one of the agglutinative languages which have complicated suffix structures telling much of syntactic. First, we construct an argument *embedding* and then develop a indicator vector of the suffix such as a *Josa*. And, we construct an argument tuple by concatenating above two vectors. The role induction is performed by clustering the argument tuples. These method which achieves up to a 65.43% of F1-score and 73.35% of accuracy.

Introduction

Semantic Role Labeling (SRL) is the task of identifying the arguments of lexical predicates in a sentence and labeling them with semantic roles. Most of the current statistical approaches to SRL are supervised, requiring large quantities of human annotated data to estimate model parameters. Unsupervised semantic role induction has attracted significant interest recently due to a problem of the limited size of annotated corpora. (Lang and Lapata 2010) used predicate features associated with their syntactic functions. In their later works (Lang and Lapata 2011), they also used the relative position (left/right) of the argument to its predicate. (Grenager and Manning 2006) used an ordering of the linking of semantic roles and syntactic relations.

Agglutinative languages such as Japanese, Korean, and Turkish tend to have a high number of suffixes/morphemes per a word, making them computationally difficult due to word-form sparsity and variable word order (Kim et al. 2014).

1. a.[Chul-su]$_{A0}$ neun [doseogwan]$_{LOC}$ eseo[gongbu]$_{A1}$ rull **handa**
   b.[Doseogwan]$_{LOC}$ eseo [Chul-su]$_{A0}$ neun [gongbu]$_{A1}$ rull **handa**

The semantic roles in examples (1-a) and (1-b) are labeled like PropBank (Palmer, Gildea, and Kingsbury 2005). For both sentences and word order, which plays an important role in English SRL, contributes nothing. The suffixes such as josa($neun, eseo$ and $rull$ in these examples), rather than the position, of arguments determine semantic roles.

In this paper we describe an unsupervised approach to argument classification or role induction that does not make use of role annotated data. We treat role induction as a clustering problem. To resolve the sparsity issue for agglutinative languages, we employ a word representation and a feature vector of the suffixes. Word embedding is, largely due to their ability to capture subtle syntactic and semantic patterns, useful in a variety of NLP tasks, such as semantic relatedness (Baroni, Dinu, and Kruszewski 2014). We construct word embeddings by using Canonical Correlation Analysis (CCA) (Hotelling 1935). We present an unsupervised Korean SRL method which achieves up to a 65.43% of F1-score and 73.35% of accuracy.

Role Labeling Process

Our method will allocate a separate set of clusters for each predicate and assign the arguments of a specific predicate to one of the clusters associated with it. For clustering, we first induce vector representations for all argument types through canonical correlation analysis (CCA) — a powerful and flexible technique for deriving low-dimensional representations. Also, we can induce the Josa representation of the argument as an indicator vector.

Canonical Correlation Analysis (CCA)

CCA is a powerful technique for inducing general representations that operates on a pair of multi-dimensional variables. CCA finds k dimensions in which these random variables are maximally correlated. The new k-dimensional representation of each variable now contains information about the other variable.

Let $x_1 \dots x_m \in \mathbb{R}^n$ and $y_1 \dots y_m \in \mathbb{R}^{n'}$ be m samples of the two variables. Then CCA computes the following for $i = 1 \dots k$:

$$\underset{\substack{u_i \in \mathbb{R}^n,\, v_i \in \mathbb{R}^{n'}: \\ u_i^\top u_{i'} = 0 \;\; \forall i' < i \\ v_i^\top v_{i'} = 0 \;\; \forall i' < i}}{\arg\max} \frac{\sum_{l=1}^{m} (u_i^\top x_l)(v_i^\top y_l)}{\sqrt{\sum_{l=1}^{m} (u_i^\top x_l)^2} \sqrt{\sum_{l=1}^{m} (v_i^\top y_l)^2}} \tag{1}$$

Inducing argument embedding

We now describe how to use CCA to induce vector representations for argument. Using the same notation, let m be the number of instances of argument in the entire data. Let

$x_1 \dots x_m$ be the original representations of the argument samples and $y_1 \dots y_m$ be the original representations of the associated feature set contained in the argument.

Let n be the number of distinct argument types and n' be the number of distinct feature types.

- $x_l \in \mathbb{R}^n$ is a zero vector in which the entry corresponding to the argument type of the l-th instance is set to 1.
- $y_l \in \mathbb{R}^{n'}$ is a zero vector in which the entries corresponding to features activated by the argument are set to 1.

It produces a k-dimensional vector for each argument type corresponding to the CCA projection of the one-hot encoding of that argument.

Features. Features representing argument instances were extracted from ETRI dependency tree corpus. We use and modify features used in (Kim et al. 2014): Part-of-Speech (POS) tags, case information such as SBJ, OBJ or COMP, left or right sibling and the leftmostchild stem, of an argument. In addition, we also use the class number of hierachical noun structure.

Tuple Design. We build an argument embedding with a 50-dimensions vector of real values by using CCA. We concatenate the embedding of an argument and a josa indicator vector and we call this concatenated vector as a *tuple*. Using them as training data, we carry out the K-means clustering.

Experiments and Results

Our experiments are carried out on 10,000 sentences. The data we used is annotated corpus built by (Kim et al. 2014) with 16 semantic roles. One advantage of our semantically annotated corpus is that it is built on top of the ETRI[1], which uses a richer Korean morphological tagging scheme than the Korean Penn Treebank.

The results of the experiment are summarized in Table 1. We only use arguments of verbal predicates, not adjective predicates. We report cluster purity (PU), Collocation (CO), and F1 is the harmonic mean of PU and CO. Finally, we also use all-cluster accuracy. Accuracy is calculated by dividing the number of all instances with the gold role by the number of instances included in each predicate cluster.

Baseline 1: Gold role, which accounts for the biggest ratio in our Korean corpus(majority), is predicted as the role of all arguments. Through the results, we confirm a cluster accuracy of 45.54%.

Baseline 2: We assume that one josa has a single role. We then assign a single role to the arguments having a same josa. It yields 68.35% of accuracy and it shows the josa gives a great impact to the detection of semantic role in korean.

Our Scenario: We verified the performance of our unsupervised approach. We used 5 clusters for each predicate and restricted the set of predicates to those attested with more than 10 instances. Arguments are assigned to clusters based on their tuple which are generated by CCA and a josa indicator vector. We use k-means algorithm in clustering. Each tuple could deal with the sparsity issue for agglutinating languages like Korean. This scenario shows 65.43% of F1-score and improves a 5% points from the *baseline 2* experiment.

[1]http://voice.etri.re.kr/db/db pop.asp?code=88

	Accuracy	PU	CO	F1
Baseline 1	45.54	-	-	-
Baseline 2	68.35	-	-	-
Scenario 1	73.35	67.96	64.21	65.43

Table 1: Clustering results for the various scenarios. Accuracy, purity, collocation, and F1-score are presented in percentages.

Conclusion

We presented an unsupervised semantic role induction for Korean. Our approach is to create a tuple representation with a concatenated argument embedding and josa vector, which gives more morphological and syntactic informative clues of unannotated corpus. Our best model achieved 65.43% of F1-score and 73.35% of accuracy.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (2015R1A2A2A01007333) and by the Ministry of Education, Science and Technology (2010-0010612).

References

Baroni, M.; Dinu, G.; and Kruszewski, G. 2014. Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, 238–247.

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

Hotelling, H. 1935. Canonical correlation analysis (cca). *Journal of Educational Psychology,*.

Kim, Y.-B.; Chae, H.; Snyder, B.; and Kim, Y.-S. 2014. Training a korean srl system with rich morphological features. In *Association for Computational Linguistics (ACL)*.

Lang, J., and Lapata, M. 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*, 939–947. Association for Computational Linguistics.

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

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