Training a Korean SRL System with Rich Morphological Features

Young-Bum Kim, Heemoon Chae, Benjamin Snyder and Yu-Seop Kim*
University of Wisconsin-Madison, Hallym University*
{ybkim, hmchae21, bsnyder}@cs.wisc.edu, yskim01@hallym.ac.kr⋆

Abstract

In this paper we introduce a semantic role labeler for Korean, an agglutinative language with rich morphology. First, we create a novel training source by semantically annotating a Korean corpus containing fine-grained morphological and syntactic information. We then develop a supervised SRL model by leveraging morphological features of Korean that tend to correspond with semantic roles. Our model also employs a variety of latent morpheme representations induced from a larger body of unannotated Korean text. These elements lead to state-of-the-art performance of 81.07% labeled F1, representing the best SRL performance reported to date for an agglutinative language.

1 Introduction

Semantic Role Labeling (SRL) is the task of automatically annotating the predicate-argument structure in a sentence with semantic roles. Ever since Gildea and Jurafsky (2002), SRL has become an important technology used in applications requiring semantic interpretation, ranging from information extraction (Frank et al., 2007) and question answering (Narayanan and Harabagiu, 2004), to practical problems including textual entailment (Burchardt et al., 2007) and pictorial communication systems (Goldberg et al., 2008).

SRL systems in many languages have been developed as the necessary linguistic resources become available (Taulé et al., 2008; Xue and Palmer, 2009; Böhmová et al., 2003; Kawahara et al., 2002). Seven languages were the subject of the CoNLL-2009 shared task in syntactic and semantic parsing (Hajič et al., 2009). These languages can be categorized into three broad morphological types: fusional (4), analytic (2), and one agglutinative language.

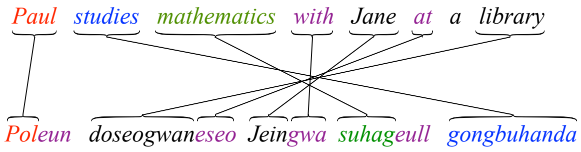

Figure 1: English (SVO) and Korean (SOV) words alignment. The subject, verb, and object are highlighted as red, blue, and green, respectively. Also, prepositions and suffixes are highlighted as purple.

Björkelund et al. (2009) report an average labeled semantic F1-score of 80.80% across these languages. The highest performance was achieved for the analytic language group (82.12%), while the agglutinative language, Japanese, yielded the lowest performance (76.30%). Agglutinative languages such as Japanese, Korean, and Turkish are computationally difficult due to word-form sparsity, variable word order, and the challenge of using rich morphological features.

In this paper, we describe a Korean SRL system which achieves 81% labeled semantic F1-score. As far as we know, this is the highest accuracy obtained for Korean, as well as any agglutinative language. Figure 1 displays a English/Korean sentence pair, highlighting the SOV word order of Korean as well as its rich morphological structure. Two factors proved crucial in the performance of our SRL system: *(i)* The analysis of fine-grained morphological tags specific to Korean, and *(ii)* the use of latent stem and morpheme representations to deal with sparsity. We incorporated both of these elements in a CRF (Lafferty et al., 2001) role labeling model.

Besides the contribution of this model and SRL system, we also report on the creation and availability of a new semantically annotated Korean corpus, covering over 8,000 sentences. We used this corpus to develop, train, and test our Korean SRL model. In the next section, we describe the process of corpus creation in more detail.

2 A Semantically Annotated Korean Corpus

We annotated predicate-argument structure of verbs in a corpus from the Electronics and Telecommunications Research Institute of Korea (ETRI).[1] Our corpus was developed over two years using a specialized annotation tool (Song et al., 2012), resulting in more than 8,000 semantically annotated sentences. As much as possible, annotations followed the PropBank guidelines for English (Bonial et al., 2010).

We view our work as building on the efforts of the Penn Korean PropBank (PKPB).[2] Our corpus is roughly similar in size to the PKPB, and taken together, the two Korean corpora now total about half the size of the Penn English PropBank. One advantage of our corpus is that it is built on top of the ETRI Korean corpus, which uses a richer Korean morphological tagging scheme than the Penn Korean Treebank. Our experiments will show that these finer-grained tags are crucial for achieving high SRL accuracy.

All annotations were performed by two people working in a team. At first, each annotator assigns semantic roles independently and then they discuss to reduce disagreement of their annotation results. Initially, the disagreement rate between two annotators was about 14%. After 4 months of this process, the disagreement rate fell to 4% through the process of building annotation rules for Korean. The underlying ETRI syntactically-annotated corpus contains the dependency tree structure of sentences with morpho-syntactic tags. It includes 101,602 multiple-clause sentences with 21.66 words on average.

We encountered two major difficulties during annotation. First, the existing Korean frame files from the Penn Korean PropBank include 2,749 verbs, covering only 13.87% of all the verbs in the ETRI corpus. Secondly, no Korean PropBanking guidelines have previously been published, leading to uncertainty in the initial stages of annotation. These uncertainties were gradually resolved through the iterative process of resolving inter-annotator disagreements.

Table 1 shows the semantic roles considered in our annotated corpus. Although these are based on the general English PropBank guidelines (Bonial et al., 2010), they also differ in that we used only 4 numbered arguments from ARG0 to ARG3 instead of 5 numbered arguments. We thus consider 17 semantic roles in total. Four of them are numbered roles, describing the essential arguments of a predicate. The other roles are called modifier roles that play more of an adjunct role.

Roles	Definition	Rate
ARG0	Agent	10.02%
ARG1	Patient	26.73%
ARG2	Start point / Benefactive	5.18%
ARG3	Ending point	1.10%
ARGM-ADV	Adverbial	1.26%
ARGM-CAU	Cause	1.17%
ARGM-CND	Condition	0.36%
ARGM-DIR	Direction	0.35%
ARGM-DIS	Discourse	28.71%
ARGM-EXT	Extent	4.50%
ARGM-INS	Instrument	1.04%
ARGM-LOC	Locative	4.51%
ARGM-MNR	Manner	8.72%
ARGM-NEG	Negation	0.26%
ARGM-PRD	Predication	0.27%
ARGM-PRP	Purpose	0.77%
ARGM-TMP	Temporal	5.05%

Table 1: Semantic roles in our annotated corpus.

We have annotated semantic roles by following the PropBank annotation guideline (Bonial et al., 2010) and by using frame files of the Penn Korean PropBank built by Palmer et al. (2006). The PropBank and our corpus are not exactly compatible, because the former is built on constituency-based parse trees, whereas our corpus uses dependency parses.

More importantly, the tagsets of these corpora are not fully compatible. The PKPB uses much coarser morpho-syntactic tags than the ETRI corpus. For example, the PCA tag in PKPB used for a case suffix covers four different functioning tags used in our corpus. Using coarser suffix tags can seriously degrade SRL performance, as we show in Section 6, where we compare the performance of our model on both the new corpus and the older PKPB.

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

[2] http://catalog.ldc.upenn.edu/LDC2006T03

3 Previous Work

Korean SRL research has been limited to domestically published Korean research on small corpora. Therefore, the most direct precedent to the present work is a section in Björkelund et al. (2009) on Japanese SRL. They build a classifier consisting of 3 stages: predicate disambiguation, argument identification, and argument classification.

They use an L_2-regularized linear logistic regression model cascaded through these three stages, achieving F1-score of 80.80% on average for 7 languages (Catalan, Chinese, Czech, English, German, Japanese and Spanish). The lowest reported performance is for Japanese, the only agglutinative language in their data set, achieving F1-score of 76.30%. This result showcases the computational difficulty of dealing with morphologically rich agglutinative languages. As we discuss in Section 5, we utilize these same features, but also add a set of Korean-specific features to capture aspects of Korean morphology.

Besides these morphological features, we also employ latent continuous and discrete morpheme representations induced from a larger body of unannotated Korean text. As our experiments below show, these features improve performance by dealing with sparsity issues. Such features have been useful in a variety of English NLP models, including chunking, named entity recognition (Turian et al., 2010), and spoken language understanding (Anastasakos et al., 2014). Unlike the English models, we use individual morphemes as our unit of analysis.

4 Model

For the semantic role task, the input is a sentence consisting of a sequence of words $x = x_1, \dots, x_n$ and the output is a sequence of corresponding semantic tags $y = y_1, \dots, y_n$. Each word consists of a stem and some number of suffix morphemes, and the semantic tags are drawn from the set $\{\text{NONE}, \text{ARG0}, \dots, \text{ARGM-TMP}\}$. We model the conditional probability $p(y|x)$ using a CRF model:

$$Z(x)^{-1} \prod_{i=1}^{x} \exp \sum_{m} \lambda_m f_m(y_{i-1}, y_i, x, i),$$

where $f_m(y_{i-1}, y_i, x, i)$ are the feature functions. These feature functions include transition features that identify the tag bigram (y_{i-1}, y_i), and emission features that combine the current semantic tag (y_i) with instantiated feature templates extracted from the sentence x and its underlying morphological and dependency analysis. The function Z is the normalizing function, which ensures that $p(y|x)$ is a valid probability distribution. We used 100 iteration of averaged perceptron algorithm to train the CRF.

5 Features

We detail the feature templates used for our experiments in Table 2. These features are categorized as either general features, Korean-specific features, or latent morpheme representation features. Korean-specific features are built upon the morphological analysis of the suffix agglutination of the current word x_i.

Korean suffixes are traditionally classified into two groups called *Josa* and *Eomi*. Josa is used to define nominal cases and modify other phrases, while Eomi is an ending of a verb or an adjective to define a tense, show an attitude, and connect or terminate a sentence. Thus, the Eomi and Josa categorization plays an important role in signaling semantic roles. Considering the functions of Josa and Eomi, we expect that numbered roles are relevant to Josa while modifier roles are more closely related to Eomi. The one exception is adverbial Josa, making the attached phrase an adverb that modifies a verb predicate.

For all feature templates, "A-" or "P-" are used respectively to signify that the feature corresponds to the argument in question (x_i), or rather is derived from the verbal predicate that the argument depends on.

General features: We use and modify 18 features used for Japanese from the prior work of Björkelund et al. (2009), excluding SENSE, POSITION, and re-ranker features.

- Stem: a stem without any attachment. For instance, the first word *Poleun* at the Figure 1 consists of a stem *Pol* plus Josa *eun*.

- POS_Lv1: the first level (coarse classification) of a POS tag such as noun, verb, adjective, or adverb.

Feature	Description
A-Stem, P-Stem	Stem of an argument and a predicate
A-POS_Lv1, P-POS_Lv1	Coarse-grained POS of A-Stem and P-Stem
A-POS_Lv2, P-POS_Lv2	Fine-grained POS of A-Stem and P-Stem
A-Case, P-Case	Case of A-Stem and P-Stem
A-LeftmostChildStem	Stem of the leftmost child of an argument
A-LeftSiblingStem	Stem of the left sibling of an argument
A-LeftSiblingPOS_Lv1	Coarse-grained POS of A-LeftSiblingStem
A-LeftSiblingPOS_Lv2	Fine-grained POS of A-LeftSiblingStem
A-RightSiblingPOS_Lv1	Coarse-grained POS of a stem of the right sibling of an argument
A-RightSiblingPOS_Lv2	Fine-grained POS of a stem of the right sibling of an argument
P-ParentStem	Stem of a parent of a predicate
P-ChildStemSet	Set of stems of children of a predicate
P-ChildPOSSet_Lv1	Set of coarse POS of P-ChildStemSet
P-ChildCaseSet	Set of cases of P-childStemSet
A-JosaExist	If 1, Josa exists in an argument, otherwise 0.
A-JosaClass	Linguistic classification of Josa
A-JosaLength	Number of morphemes consisting of Josa
A-JosaMorphemes	Each morpheme consisting of Josa
A-JosaIdenity	Josa of an argument
A-EomiExist	If 1, Eomi exists in an argument, otherwise 0.
A-EomiClass_Lv1	Linguistic classification of Eomi
A-EomiClass_Lv2	Another linguistic classification of Eomi
A-EomiLength	Number of morphemes consisting of Eomi
A-EomiMorphemes	Each morpheme consisting of Eomi
A-EomiIdentity	Eomi of an argument
A-StemRepr	Stem representation of an argument
A-JosaRepr	Josa representation of an argument
A-EomiRepr	Eomi representation of an argument

Table 2: Features used in our SRL experiments. Features are grouped as General, Korean-specific, or Latent Morpheme Representations. For the last group, we employ three different methods to build them: (i) CCA, (ii) deep learning, and (iii) Brown clustering.

- POS_Lv2: the second level (fine classification) of a POS tag. If POS_Lv1 is *noun*, either a proper noun, common noun, or other kinds of nouns is the POS_Lv2.

- Case: the case type such as SBJ, OBJ, or COMP.

The above features are also applied to some dependency children, parents, and siblings of arguments as shown in Table 2.

Korean-specific features: We have 11 different kinds of features for the Josa (5) and Eomi (6). We highlight several below:

- A-JosaExist: an indicator feature checking any Josa whether or not exists in an argument. It is set to 1 if any Josa exists, otherwise 0.

- A-JosaClass: the linguistic classification of Josa with a total of 8 classes. These classes are adverbial, auxiliary, complemental, connective, determinative, objective, subjective, and vocative.

- A-JosaLength: the number of morphemes consisting of Josa. At most five morphemes are combined to consist of one Josa in our data set.

- A-JosaMorphemes: Each morpheme composing the Josa.

- A-JosaIdentity: The Josa itself.

- A-EomiClass_Lv1: the linguistic classification of Eomi with a total of 14 classes. These 14 classes are adverbial, determinative, coordinate, exclamatory, future tense, honorific, imperative, interrogative, modesty, nominal, normal, past tense, petitionary, and subordinate.

- A-EomiClass_Lv2: Another linguistic classification of Eomi with a total of 4 classes. The four classes are closing, connection, prefinal, and transmutation. The EomiClass_Lv1 and Lv2 are combined to display the characteristic of Eomi such as 'Nominal Transmutation Eomi', but not all combinations are possible.

Corpus	Gen	Gen+Kor	Gen+Kor+LMR			
			CCA	Deep	Brown	All
PKPB	64.83%	75.17%	75.51%	75.43%	75.55%	75.54%
Our annotated corpus	66.88%	80.33%	80.88%	80.84%	80.77%	**81.07%**
PKPB + our annotated corpus	64.86%	78.61%	79.32%	79.44%	78.91%	79.20%

Table 3: Experimental F1-score results on every experiment. Abbreviation on features are Gen: general features, Kor: Korean specific features, LMR: latent morpheme representation features.

Latent morpheme representation features: To alleviate the sparsity, a lingering problem in NLP, we employ three kinds of latent morpheme representations induced from a larger body of unsupervised text data. These are (i) linear continuous representation through Canonical Correlation Analysis (Dhillon et al., 2012), (ii) non-linear continuous representation through Deep learning (Collobert and Weston, 2008), and (iii) discrete representation through Brown Clustering (Tatu and Moldovan, 2005).

The first two representations are 50 dimensional continuous vectors for each morpheme, and the latter is a set of 256 clusters over morphemes.

6 Experiments and Results

We categorized our experiments by the scenarios below, and all results are summarized in Table 3. The F1-score results were investigated for each scenario. We randomly divided our data into 90% training and 10% test sets for all scenarios.

For latent morpheme representations, we used the Donga news article corpus.[3] The Donga corpus contains 366,636 sentences with 25.09 words on average. The Domain of this corpus covers typical news articles such as health, entertainment, technology, politics, world and others. We ran Kokoma Korean morpheme analyzer[4] on each sentence of the Donga corpus to divide words into morphemes to build latent morpheme representations.

1st Scenario: We first tested on general features in previous work (2nd column in Table 3). We achieved 64.83% and 66.88% on the PKPB and our corpus. When the both corpora were combined, we had 64.86%.

2nd Scenario: We then added the Korean-specific morphological features to signify its appropriateness in this scenario. These features increased greatly performance improvements (3rd column in Table 3). Although both the PKPB and our corpus had improvements, the improvements were the most notable on our corpus. This is because PKPB POS tags might be too coarse. We achieved 75.17%, 80.33%, and 78.61% on the PKPB, our corpus, and the combined one, respectively.

3rd Scenario: This scenario is to reveal the effects of the different latent morpheme representations (4-6th columns in Table 3). These three representations are from CCA, deep learning, and Brown clustering. The results gave evidences that all representations increased the performance.

4th Scenario: We augmented our model with all kinds of features (the last column in Table 3). We achieved our best F1-score of 81.07% over all scenarios on our corpus.

7 Conclusion

For Korean SRL, we semantically annotated a corpus containing detailed morphological annotation. We then developed a supervised model which leverages Korean-specific features and a variety of latent morpheme representations to help deal with a sparsity problem. Our best model achieved 81.07% in F1-score. In the future, we will continue to build our corpus and look for the way to use unsupervised learning for SRL to apply to another language which does not have available corpus.

Acknowledgments

We thank Na-Rae Han and Asli Celikyilmaz for helpful discussion and feedback. This research was supported by the Basic Science Research Program of the Korean National Research Foundation (NRF), and funded by the Korean Ministry of Education, Science and Technology (2010-0010612).

[3] http://www.donga.com

[4] http://kkma.snu.ac.kr/

References

Tasos Anastasakos, Young-Bum Kim, and Anoop Deoras. 2014. Task specific continuous word representations for mono and multi-lingual spoken language understanding. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 43–48. Association for Computational Linguistics.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The prague dependency treebank. In *Treebanks*, pages 103–127. Springer.

Claire Bonial, Olga Babko-Malaya, Jinho D Choi, Jena Hwang, and Martha Palmer. 2010. Propbank annotation guidelines. *Center for Computational Language and Education Research, CU-Boulder*.

Aljoscha Burchardt, Nils Reiter, Stefan Thater, and Anette Frank. 2007. A semantic approach to textual entailment: system evaluation and task analysis. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pages 10–15, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Paramveer Dhillon, Jordan Rodu, Dean Foster, and Lyle Ungar. 2012. Two step cca: A new spectral method for estimating vector models of words. *arXiv preprint arXiv:1206.6403*.

Anette Frank, Hans-Ulrich Krieger, Feiyu Xu, Hans Uszkoreit, Berthold Crysmann, Brigitte Jörg, and Ulrich Schäfer. 2007. Question answering from structured knowledge sources. *Journal of Applied Logic*, 5(1):20 – 48. Questions and Answers: Theoretical and Applied Perspectives.

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

Andrew B Goldberg, Xiaojin Zhu, Charles R Dyer, Mohamed Eldawy, and Lijie Heng. 2008. Easy as abc?: facilitating pictorial communication via semantically enhanced layout. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 119–126. Association for Computational Linguistics.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18. Association for Computational Linguistics.

Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. 2002. Construction of a japanese relevance-tagged corpus. In *LREC*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Srini Narayanan and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martha Palmer, Shijong Ryu, Jinyoung Choi, Sinwon Yoon, and Yeongmi Jeon. 2006. Korean propbank. *Linguistic data consortium.*

Hye-Jeong Song, Chan-Young Park, Jung-Kuk Lee, Min-Ji Lee, Yoon-Jeong Lee, Jong-Dae Kim, and Yu-Seop Kim. 2012. Construction of korean semantic annotated corpus. In *Computer Applications for Database, Education, and Ubiquitous Computing*, pages 265–271. Springer.

Marta Tatu and Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 371–378. Association for Computational Linguistics.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *LREC*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

Nianwen Xue and Martha Palmer. 2009. Adding semantic roles to the chinese treebank. *Natural Language Engineering*, 15(01):143–172.