Semi-automatic Korean FrameNet Annotation over KAIST Treebank

Younggyun Hahm, Sunggoo Kwon, Jiseong Kim, Key-Sun Choi
Semantic Web Research Center, School of Computing, KAIST
291 Daehak-ro, Yuseong-gu, Deajeon, South Korea
{hahmyg, fanafa, jiseong, kschoi}@kaist.ac.kr

Abstract

This paper describes a project for constructing FrameNet annotations in Korean over the KAIST treebank corpus to scale up the Korean FrameNet resource. Annotating FrameNet over raw sentences is an expensive and complex task, because of which we have designed this project using a semi-automatic annotation approach. This paper describes the approach and its expected results. As a first step, we built a lexical database of the Korean FrameNet, and used it to learn the model for automatic annotation. Its current scope, status, and limitations are discussed in this paper.

Keywords: FrameNet, Semantic Role Labeling, Corpus Annotation

1. Introduction

FrameNet is a large lexical database that has rich annotations to represent the meanings of text using semantic frames (Baker et al., 1998; Fillmore et al., 2003). FrameNet has been considered a useful resource for various applications such as question answering systems (Shen and Lapata, 2007, Hahm et al., 2016), information extraction (Surdeanu et al., 2003), and dialog systems (Chen et al., 2013). Lately, researchers have shown increasing interest in multilingual FrameNet (Borin et al., 2010; You and Lui, 2005; Meurs et al., 2008; Subirats and Petruck, 2003; Burchardt et al., 2006). A Korean FrameNet project built the Korean FrameNet resource by translating English FrameNet annotations and Japanese FrameNet into its equivalent in Korean (Park et al., 2014; Kim et al., 2016).

One of the purpose of the FrameNet annotation is to build a frame-semantic parser to understand the meaning of a text. Some studies have built frame-semantic parser for English by using full-text annotation and partially annotated exemplar sentences to train their models (Das et al., 2010; Swayamdipta et al., 2017; Yang and Mitchell, 2017). For example, the state-of-the-art frame-semantic parser uses nearly 139k exemplar sentences for training data and it generally introduces a 3–4 F1 gain for parsing (Yang and Mitchell., 2017). In comparison, the Korean FrameNet has full-text annotations for 5,025 sentences and 8,200 lexical units (LUs).

In this paper, we report an ongoing project to construct Korean FrameNet annotations to scale up the amount of annotations. A task to annotate the frame-semantics manually over raw sentences has been formalized by Ruppenhofer et al. (2006). It is an expensive and complex task, because the annotators would need to choose proper frame-semantics for each target word, and its corresponding frame elements for the arguments (1,221 frame-semantics are defined in FrameNet 1.7). Therefore, we have used the existing Korean resources to bootstrap the annotation task. A target corpus for the FrameNet annotation task is the KAIST treebank (Choi et al., 1994). It includes 31,086 sentences with morphological analysis, part-of-speech (POS) tagging, and dependency parsing. It is used in the Korean Universal Dependency treebank (Choi, 2013).

This paper briefly introduces the FrameNet annotation over the KAIST treebank project, and presents its scope in Section 2. The related research tasks are described in Section 3, which are separated into the current state of the research and ongoing tasks. The evaluation of the current progress is discussed in Section 4, and conclusions are presented in Section 5.

2. Problem Definition

2.1 Problem Statement and Workflow

The goal of the project is to annotate FrameNet over the KAIST treebank. Figure 1 shows the workflow of these tasks. In the first step, the semantic frame is automatically annotated over the target corpus, which is the KAIST treebank, by using a model learned from the Korean FrameNet.

First, the target identification module identifies the target words which evoke the frame-semantics, and then the frame identification module and the frame element identification module identify the proper frame-semantics for the given target and its corresponding arguments (subsection 3.2). These modules use the lexical units (LUs) and the full-text annotations in the Korean FrameNet. To support this purpose, we built the lexical database of the Korean FrameNet (subsection 3.1) as a follow-up of the previous study by Park et al. (2014). By this process, the FrameNet annotation over the KAIST treebank was conducted automatically, which is a *silver standard*.

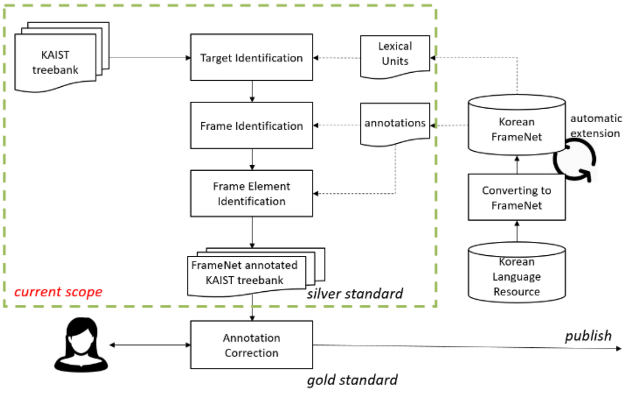

Figure 1: Workflow diagram of FrameNet annotation over the KAIST treebank.

To scale up the LUs in the Korean FrameNet, the task of converting the Sejong Electronic Dictionary (Hong, 2007) to the Korean FrameNet has been included as an ongoing task in this project (subsection 3.3). Then, the annotator would perform corrections of the *silver standard* annotations (subsection 3.4). In this paper, we report the current scope and state of the project, and discuss the ongoing tasks and expected results. Figure 1 shows the workflow of this project.

2.2 Scope

In the FrameNet annotation tasks, the frame-semantics would be assigned to the target words. We considered general nouns, verbs, and adjectives alone in the KAIST treebank as the target words. The converted LUs from the Sejong Dictionary were not considered because it does not meet the gold standard yet. The annotation correction task is outside of the scope of this paper, however, the preparation tasks on ongoing works are briefly described.

3. Task Description

3.1 Building Korean FrameNet Database

Korean FrameNet is a resource that has been manually translated from English and Japanese FrameNet annotations into Korean. To use the Korean FrameNet, which was constructed in our previous work (Kim et al., 2016) as well as the training data, we first collected the LUs from the annotations. In the original annotations in English, the target words would be tokenized by white space; however, the translated Korean target words consist of multiple morphemes. For example, the target word 'visiting' is translated to '방문한', which would be tokenized into morphemes as a noun '방문'(visit), an adjective derivational suffix '하' which transforms the noun to the verb form, and adnominal ending 'ㄴ' which transforms the verb to the adjective form by combining it with other morphemes. To build a dictionary, we collected LUs along with their various form while retaining their grammatical and semantic meaning. We obtained all the target words from the annotations and then pruned specific morphemes, such as endings, *josa* (Korean postpositions), and affixes, as part of a lemmatization task in Korean. Then, 8,200 LUs were collected, which consisted of a lemma, its POS, and its corresponding frame-semantics. In this task, we corrected the errors in the processes, and performed the POS tagging and morphological analysis manually for 450 cases.

FrameNet is a lexical database that includes not only LUs but also syntactic realization patterns (i.e. valence patterns) for the frame elements of each LU. For example, an LU is 'visit' when it is used as a verb, and has a frame-semantics "`Visiting`"; the frame elements are annotated in full-text annotations, such as `agent`, `entity`, `place`, and so on. Moreover, these frame elements have a grammatical role in sentences (e.g., subject of the sentence), and have phrase types (e.g., noun phrase). These patterns are useful resources to the identify frame elements in a text. We parsed full-text annotations in the Korean FrameNet in the dependency syntax, and collected the valence patterns for each frame element of the LUs. For instance, the LU '가르치.v.Educational_teaching' (teach.v) consists of its lemma, POS, and the corresponding frame-semantics. It also has the valence patterns for each frame element; for example, the frame element `teacher` has the role of the subject in its annotations, and it has a *josa* '가,' which gives the nouns a subjective role. All LUs are updated on the Korean FrameNet webpage[1].

3.2 Automatic FrameNet Annotation

The FrameNet annotation task is generally separated into three steps (Das et al., 2010). A system 1) identifies a target word in the text, 2) identifies its proper frame-semantics, and then 3) identifies its frame elements.

As described in Section 2, the target words in the KAIST treebank are specified into three types—general nouns, verbs, and adjectives. Table 1 shows the number of target words of each type. For the target identification task, all words in the three types are considered as the target words.

Contents	Counts (total)	Counts (unique)
# of sentences	31,086	
# of general nouns	99,784	15,180
# of verbs	69,889	6,120
# of adjectives	26,559	1,807

Table 1: Statistics of the KAIST treebank

3.2.1 Frame Identification

Frame identification is a disambiguation task specified in the FrameNet terminology. First, the candidates for the frame-semantics are generated for a given target word, and then the most suitable frame-semantics is selected. In our task, for a given target word, a list of frame-semantics and their annotations were collected from the Korean FrameNet database that was built as described in subsection 3.1.

In the frame identification task, the frame-semantics of the target words would be disambiguated by their surrounding contexts (Baker et al., 1998). It means that if a given target word has a similar context as an LU in the Korean FrameNet, its proper frame-semantics would be a frame-semantics of the LU.

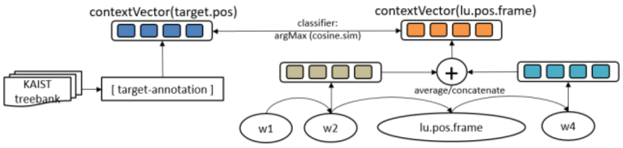

Figure 2: Identify the frame-semantics for a target word by comparing a context vector of the LU, which was learned from the Korean FrameNet annotations.

In this paper, to select a suitable frame-semantics for the target words, we have borrowed ideas from the concepts of synset embedding (Rothe and Schütze, 2015) and doc2vec (Le and Mikolov, 2014). Our model 1) learns the context vectors of each LU (called the *frame embedding*) from the full-text annotation in the Korean FrameNet, 2) generates the context vector for the given target word from the input text in the same way to generate the *frame embeddings*, 3) determines a similarity score between the context vectors of the target word and the *frame embeddings*, and then 4) chooses the most similar context vector and LU to get its

[1] http://framenet.kaist.ac.kr

frame-semantics. Figure 2 shows a snapshot of these operations.

In this paper, the context words are defined as 1-hop connected nodes with a target word in the dependency path, and the context vector is generated by averaging the sum of the word embeddings of each context word. The similarity score is calculated as the cosine similarity between the context vector of the target word and the *frame embeddings*.

Ongoing studies

The method described above uses a limited scope of the surrounding context of a target word and an LU. Hermann et al. (2014) uses a joint model with word embedding and syntactic structure, and Swayamdipta et al. (2017) uses rich syntactic features to identify the frame-semantics. Generating the *frame embeddings* from rich features would be the next step.

3.2.2 Frame Element Identification

As described in subsection 3.1, each LU has a list of valence patterns for its frame elements. In this paper, the valence patterns are applied by a rule-based approach. In other words, if a given target word in a sentence is disambiguated as a specific frame-semantics by the target identification task, its frame elements are identified if only it matched with the valence patterns. For example, if a given target word is 'teach.v' and its frame-semantics is `Educational_teaching`, then the frame element `teacher` is annotated for a phrase which is matched with the valence pattern for the grammatical function and its phrase type.

Ongoing studies

Identifying a boundary of frame elements and its type (i.e., the frame element tag) is still a challenge in the frame-semantic parsing task. Täckström et al. (2015) relied on the dependency features and some heuristic rules, and Yang and Mitchell (2017) used a joint model to jointly assign the frame-semantics and frame elements. In our project, our purpose of using automatic FrameNet annotation is to construct a *silver standard* corpus that would be corrected manually. Next, we are focused on the task that can identify frame elements well. To accomplish this, we are studying methods to generate valence patterns with richer syntactic features than using only grammatical functions and phrase types for high recall performance.

3.3 Automatic Expansion of FrameNet LUs

The Sejong Dictionary is known to have sufficient coverage for Korean corpus such as Wikipedia (Hahm et al., 2014). It includes not only a list of lexemes but also their predicate-arguments structure in PropBank style, their exemplar sentences, and English translations (words or phrases). Our project includes the task of automatic extension of the Korean FrameNet by converting the Sejong Dictionary to FrameNet. As a first step, we have chosen a representative word from the translations of a lexeme, and matched it with LUs in the English FrameNet by string matching approach (Levenshtein similarity > 0.95).

Ongoing studies

In this paper, we do not use the LUs derived from the Sejong Dictionary because it has not yet been manually validated, and the selection of a representative word from a translation is conducted by heuristic rules. Nevertheless, the Sejong Dictionary appears to be a useful resource that can improve the Korean FrameNet in terms of LU and exemplar sentences.

3.4 Manual Annotation Correction

To publish the resource, the FrameNet-annotated KAIST treebank corpus would be validated as a gold standard corpus to prevent error propagation issues in the training process. The result of the annotation correction is beyond the scope of this paper; however, the preparatory work for manual annotation correction is reported in this section.

WebAnno 3.2[2] (Biemann et al., 2017) is considered as the user interface. WebAnno provides a function to suggest proper candidate tags for a given word by using *constraints*. In the FrameNet annotation task, the frame-semantics candidates for a given word are shown at the top of the list of frame-semantics candidates, and it would prevent the annotators from wasting time searching for suitable frame-semantics tags. We generate the *constraints* from the Korean FrameNet database. For example, if a given word exist in the Korean FrameNet, its corresponding frame-semantics tags (i.e., annotated frame-semantics in Korean FrameNet full-text annotations) are shown at the top of the list of candidate tags.

4. Evaluation

For the evaluation of the results, we separate the Korean FrameNet into a training set and a test set. The test set consists of sentences that include one token LU (i.e. excluding the white space in the LU) categorized into three types—general nouns, verbs, and adjectives. The LUs in a test set have more than two frame-semantics candidates. The training set consists of 13,001 sentences, and the test set has 1,816 sentences.

Models	**Accuracy**
Random	67.29
Sentence Embedding	73.57
Frame Embedding	76.21

Table 2: Results of the frame identification task

Table 2 shows the results of the frame identification task that is described in subsubsection 3.2.1. The random model chooses the frame-semantics randomly from a list of candidates, and the sentence embedding model learns sentence embedding from the Korean FrameNet full-text annotations with the frame-semantics annotation on the target words. For example, in the sentence 'I go home,' the word 'go' is assigned to the frame-semantics `Motion`; the sentence embedding model learns the sentence embedding

[2] https://github.com/webanno/webanno

from a list [‘I’, ‘go/`Motion`’, ‘home’] using doc2vec implementation[3]. This model learns all contexts in the sentence without a word order. The *frame embedding* model learns the context vectors for each LU in the Korean FrameNet (subsubsection 3.2.1), therefore, it would keep more relevant contexts as compared to the sentence embedding model. To generate the *frame embeddings*, we used the Korean Wikipedia[4] as a training corpus with the settings of dimensions = 100 and window size = 3.

Types	Total coverage	Word coverage
General Noun	47.82%	8.7%
Verb	68.42%	17.38%
Adjective	40.31%	6.03%

Table 3: Coverage of Korean FrameNet over the KAIST treebank

The frame embedding model also performed the annotation of frame-semantics over the KAIST treebank. For 31,086 sentences, 99,784 frame-semantics were annotated (3.42 per sentence), and 61,579 frame elements were annotated (0.62 per frame-semantics). Table 3 shows the results of the frame-semantics annotations in terms of coverage. Korean FrameNet has an overall coverage of about 57 percent; however, it does not show good coverage for each word shown in the KAIST treebank. It means that the scaling up of LUs is required to annotate FrameNet over the KAIST treebank, as described in subsection 3.3. When compared with the Korean FrameNet, 6.81 frame elements were annotated per frame-semantics in the Korean FrameNet. The task of increasing the coverage of the valence patterns also remains a challenge.

5. Conclusion

This paper describes the FrameNet annotation over the KAIST treebank, and discusses its scope and the current status. To bootstrap the annotation task, we designed the project using a semi-automatic annotation approach. We built the Korean FrameNet database, and used it to learn models to annotate automatically. We discovered that there are some limitations when using only Korean FrameNet, because of which the automatic extension of the Korean FrameNet task and the manual annotation correction task have been included in this project. Several ongoing studies are currently underway to address the challenges identified and described in this paper.

6. Acknowledgements

This work was supported by Institute for Information & Communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (R0124-16-0002, Emotional Intelligence Technology to Infer Human Emotion and Carry on Dialogue Accordingly)

7. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1* (pp. 86-90). Association for Computational Linguistics.

Biemann, C., Bontcheva, K., de Castilho, R. E., Gurevych, I., and Yimam, S. M. (2017). Collaborative Web-based Tools for Multi-layer Text Annotation. *The Handbook of Linguistic Annotation’, Text, Speech, and Technology book series, Springer Netherlands*.

Borin, L., Dannélls, D., Forsberg, M., Gronostaj, M. T., and Kokkinakis, D. (2010). The past meets the present in Swedish FrameNet++. In *14th EURALEX international congress* (pp. 269-281).

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2006). The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)* (pp. 969-974).

Chen, Y. N., Wang, W. Y., and Rudnicky, A. I. (2013). Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing. In *Automatic Speech Recognition and Understanding (ASRU)*, 2013 IEEE Workshop on (pp. 120-125). IEEE.

Choi, J. D. (2013). Preparing Korean data for the shared task on parsing morphologically rich languages. *arXiv preprint arXiv:1309.1649*.

Choi, K. S., Han, Y. S., Han, Y. G., and Kwon, O. W. (1994). KAIST tree bank project for Korean: Present and future development. In *Proceedings of the International Workshop on Sharable Natural Language Resources* (pp. 7-14).

Das, D., Schneider, N., Chen, D., and Smith, N. A. (2010). SEMAFOR 1.0: A probabilistic frame-semantic parser. *Language Technologies Institute, School of Computer Science, Carnegie Mellon University*.

Fillmore, C. J., Johnson, C. R., and Petruck, M. R. (2003). Background to framenet. *International journal of lexicography*, 16(3), 235-250.

Hahm, Y., Nam, S., and Choi, K. S. (2016). QAF: Frame Semantics-based Question Interpretation. In *Proceedings of the Open Knowledge Base and Question Answering Workshop*.

Hahm, Y., Kim, Y., Won, Y., Woo, J., Seo, J., Kim, J., Park, S., Hwang, D., and Choi, K. S. (2014). Toward matching the relation instantiation from DBpedia ontology to Wikipedia text: fusing FrameNet to Korean. In *Proceedings of the 10th International Conference on Semantic Systems* (pp. 13-19). ACM.

Hermann, K. M., Das, D., Weston, J., and Ganchev, K. (2014). Semantic Frame Identification with Distributed Word Representations. In *ACL 2014* (pp. 1448-1458).

Hong, J.S. (2007). Development Research Paper of 21 century Sejong Plan Electronic Dictionary. (11-1370252-000063-10), *the Ministry of Culture and Tourism, The National Institute of the Korean Language, 2007. (in Korean)*

Kim, J., Hahm, Y., and Choi, K. S. (2016). Korean FrameNet Expansion Based on Projection of Japanese FrameNet. In *COLING 2016*.

Kingsbury, P., and Palmer, M. (2002). From TreeBank to PropBank. In *LREC 2002* (pp. 1989-1993).

Le, Q., and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st*

[3] https://radimrehurek.com/gensim/models/doc2vec.html

[4] https://dumps.wikimedia.org/kowiki/20170601/

International Conference on Machine Learning (ICML-14) (pp. 1188-1196).

Meurs, M. J., Duvert, F., Béchet, F., Lefevre, F., and De Mori, R. (2008). Semantic Frame Annotation on the French MEDIA corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Park, J., Nam, S., Kim, Y., Hahm, Y., Hwang, D., and Choi, K. S. (2014). Frame-semantic web: a case study for korean. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272* (pp. 257-260). CEUR-WS. org.

Rothe, S., and Schütze, H. (2015). Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *arXiv preprint arXiv:1507.01127*.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., and Scheffczyk, J. (2006). FrameNet II: Extended theory and practice. *https://framenet.icsi.berkeley.edu/* .

Shen, D., and Lapata, M. (2007). Using Semantic Roles to Improve Question Answering. In *Emnlp-conll* (pp. 12-21).

Subirats, C., and Petruck, M. (2003). Surprise: Spanish FrameNet. In *Proceedings of CIL* (Vol. 17, p. 188).

Surdeanu, M., Harabagiu, S., Williams, J., and Aarseth, P. (2003). Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1* (pp. 8-15). Association for Computational Linguistics.

Swayamdipta, S., Thomson, S., Dyer, C., and Smith, N. A. (2017). Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. *arXiv preprint arXiv:1706.09528*.

Täckström, O., Ganchev, K., and Das, D. (2015). Efficient inference and structured learning for semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3, 29-41.

Yang, B., and Mitchell, T. (2017). A Joint Sequential and Relational Model for Frame-Semantic Parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1258-1267).

You, L., and Liu, K. (2005). Building chinese framenet database. In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on* (pp. 301-306). IEEE.