Towards Standardized Annotation and Parsing for Korean FrameNet

Yige Chen[1] **Jae Ihn**[2**] **KyungTae Lim**[3*] **Jungyeul Park**[2]
[1]The Chinese University of Hong Kong, Hong Kong
[2]The University of British Columbia, Canada
[3]SeoulTech, South Korea
yigechen@link.cuhk.edu.hk jae@therocketbrew.com ktlim@seoultech.ac.kr
jungyeul@mail.ubc.ca

Abstract

Previous research on Korean FrameNet has produced several datasets that serve as resources for FrameNet parsing in Korean. However, these datasets suffer from the problem that annotations are assigned on the word level, which is not optimally designed based on the agglutinative feature of Korean. To address this issue, we introduce a morphologically enhanced annotation strategy for Korean FrameNet datasets and parsing by leveraging the CoNLL-U format. We present the results of the FrameNet parsers trained on the Korean FrameNet data in the original format and our proposed format, respectively, and further elaborate on the linguistic rationales of our proposed scheme. We suggest the morpheme-based scheme to be the standard of Korean FrameNet data annotation.

Keywords: FrameNet, Korean, data standardization

1. Introduction

As a rich linguistic resource that reveals the frame semantics of natural languages, FrameNet (Baker et al., 1998; Lönneker-Rodman and Baker, 2009; Ruppenhofer et al., 2010) has been widely adopted in natural language processing, especially for semantic parsing. While the earliest FrameNet project focuses on the English language only, various FrameNet datasets in languages other than English, such as Japanese (Ohara et al., 2003), Chinese (You and Liu, 2005), Italian (Lenci et al., 2010), Swedish (Johansson and Nugues, 2006), as well as multilingual FrameNet datasets (Hartmann and Gurevych, 2013), have been constructed. Learning frame semantics through parsing has also been made possible for English FrameNet, where Bauer et al. (2012) develop a dependency-parsed FrameNet dataset based on which parsers can be trained to predict the frame arguments.

There has been research on Korean FrameNet as well. Park et al. (2014) create a Korean FrameNet dataset by converting existing English FrameNet sentences originated from English Propbank into Korean. Kim et al. (2016) follow the same approach and develop additional Korean FrameNet data by projecting the Japanese FrameNet to translated Korean texts. Hahm et al. (2018) further construct a Korean FrameNet dataset based on the KAIST Treebank (Choi et al., 1994). However, all existing Korean FrameNet datasets suffer from a shared problem, which is rooted in the linguistic property of the Korean language. Since Korean is an agglutinative language, its functional morphemes are attached to the lexical morphemes to form segments of the language. These functional morphemes hardly contribute to the semantics of the sentence, and a great number of tokens will be introduced to the vocabulary if the natural segmentation, which can be complex combinations of various morphemes, is considered to be the basic unit during tokenization. While morpheme-based schemes have been proven effective in other Korean processing tasks such as part-of-speech tagging (Park and Tyers, 2019), dependency parsing (Chen et al., 2022) and named entity recognition (Chen et al., 2023), how the morpheme-based approach can be employed in annotating Korean FrameNet datasets has not been extensively studied.

To fill up the gap, we provide morphologically enhanced FrameNet datasets for Korean based on existing Korean FrameNet datasets. We also train parsers on the original data and the morphologically enhanced data to compare their performance to show the benefit of the morphologically enhanced annotation, and further demonstrate the rationales of our proposed scheme in reference to the linguistic features of Korean. We suggest that the morpheme-based scheme be the standardized way of representing Korean FrameNet data.

2. Korean FrameNet Dataset

The dataset we use was originally developed and published by KAIST (Park et al., 2014; Kim et al., 2016; Hahm et al., 2018), and it includes multiple sources from which the data are collected. We

[*]Corresponding author. [**]Currently at Rocketbrew Inc., Canada.

choose parts of the whole dataset originating from three sources for the purpose of this study, which are the Korean FrameNet data from Korean PropBank (pkfn), the Japanese FrameNet (jkfn), and the Sejong Dictionary (skfn). While the Korean FrameNet data from the English PropBank (ekfn) is also available, we noticed that the tokenization scheme does not agree with other datasets, and decided not to adopt it to the current study. Table 1 introduces statistics describing the distribution of the lexical units (LUs). Table 2 presents the number of frames per LU, which measures the degree of ambiguity in the lexical units within the three subsets. Table 3 shows the total number of sentences and instances in each subset, in which identical sentences with different frames count as a single sentence but as separate instances.

# of LUs	pkfn	jkfn	skfn
Noun	0	755	0
Verb	644	500	2,252
Adjective	6	155	0
Others	0	14	0
Total	650	1,424	2,252

Table 1: Distributions of the lexical units (LUs) of the targets in 3 Korean FrameNet datasets. An LU is a word with its part-of-speech.

# of frames per LU	pkfn	jkfn	skfn
Noun	0	1.109	0
Verb	1.183	1.276	1.274
Adjective	1.167	1.290	0
Others	0	1.286	0
Overall	1.183	1.189	1.274

Table 2: The number of frames per lexical unit for each of the Korean FrameNet datasets.

	pkfn	jkfn	skfn
# of sentences	1,767	1,357	5,703
# of instances	2,350	2,919	5,703
# of frames per sentence	1.330	2.151	1.000

Table 3: Numbers of sentences and instances in the 3 Korean FrameNet datasets.

pkfn The pkfn data in the Korean FrameNet dataset was sourced from the Korean PropBank (Palmer et al., 2006). The dataset contains mainly verbal targets, along with a few adjectival targets. Figure 1 illustrates how a single sentence is labeled in the Korean PropBank and the Korean FrameNet dataset respectively, where the FrameNet annotation inherits the predicate-argument relation from PropBank and re-analyzes the sentence using frame semantics.

jkfn The jkfn data, as presented in Kim et al. (2016), was projected from the Japanese FrameNet (Ohara et al., 2003). Given the syntactic similarities between Korean and Japanese, the jkfn data are direct and literal translations from the original word chunks separated by frame data in the Japanese FrameNet, in which way the projected jkfn data preserves the boundaries of the frames (Kim et al., 2016) as shown in Figure 2. The dataset contains a large number of nominal targets and a considerable number of verbal targets, whereas adjectival targets are also present in the dataset.

skfn The skfn data is based on the example sentences in the Sejong dictionary. The major characteristic that differentiates skfn from the above two subsets is that the example sentences in the dictionary are usually short, and as a result, a sentence in the skfn data carries a single frame only. All frame targets in skfn are verbs with no exception. Figure 3 presents an example of the frame-based information in the Sejong dictionary and how its example sentence is annotated in the FrameNet data. Note that 이 (-*i*) denotes any nominative particle in Sejong Dictionary. As a result, X corresponds to the nominative noun phrase *jeo salameun* (that person), and Y corresponds to the event *uli il-e* (out affairs), in the example. The boundaries of frame arguments cannot be inherited from the original source because the Sejong dictionary did not explicitly specify such boundaries. Instead, automatic detection and mapping between frame elements and arguments for the frame of the given predicate are conducted.

3. Morphologically Enhanced FrameNet Dataset

We propose a morpheme-based scheme for Korean FrameNet data that leverages the linguistic properties of the Korean language. As an agglutinative language, Korean possesses the feature that the natural segmentation, namely an *eojeol*, can consist of both the lexical morpheme and its postposition, such as a particle that marks tense or case. This poses challenges in Korean FrameNet parsing, as the parser is not able to distinguish the arguments from their functional morphemes given the *eojeol*-based segmentation. In other words, the smallest unit (i.e., *eojeol*) as a single token is a mixture of the lexical part and the functional part, and a sequence labeling model is not able to learn from the *eojeol*-based data and tell what the lexical morphemes are in an *eojeol*. Since the lexical morphemes contribute to the semantic meaning of the *eojeol* on a large scale and determine the lexical units the targets instantiate and the semantic frame they evoke, it is essential to separate them from their postpositions during processing.

As illustrated in Figure 4, the sentence is decomposed into morphemes as the basic unit of tokens.

	북한이	국무부의테러지원국명단에	빠지게	되면...
	bughan-i	*gugmubuui teleojiwongug myeongdan-e*	*ppajige*	*doemyeon ...*
	'North Korea-nom'	'Department of State's list of state sponsors of terrorism-obl'	'exclude'	'if/when' ...
PropBank	(thing-excluded)$_{\text{ARG1}}$	(excluded-from)$_{\text{ARG2}}$	(exclude)$_{\text{PRED}}$	
	↓	↓	↓	
FrameNet	Theme	Source	Removing$_{\text{TARGET}}$	

Figure 1: Comparisons between annotations on the same instance in Korean PropBank and the Korean FrameNet dataset. The meaning of the above instance is 'if North Korea were excluded from the Department of State's list of state sponsors of terrorism...', which is part of a sentence in the Korean PropBank.

	小学生が	青信号で	横断歩道を	渡る。
	shōgakusei-ga	*aoshingō-de*	*ōdan hodō-o*	*wataru*
Japanese FrameNet	Theme		Path	Path_shape$_{\text{TARGET}}$
	↓		↓	↓
Korean FrameNet	Theme		Path	Path_shape$_{\text{TARGET}}$
	초등학생이	파란불에	횡단보도를	건넌다.
	chodeunghagsaeng-i	*palan bul-e*	*hoengdanbodoleul*	*geonneonda*
	'elementary school students-nom'	'green light-obl'	'crosswalk-acc'	'cross'

Figure 2: Comparisons between annotations on the same instance in the Japanese FrameNet dataset and the Korean FrameNet dataset. The meaning of the above instance is 'elementary school students cross a crosswalk on the green light'.

Sejong 개입하다 (*gaeibhada*, to intervene)
Frame: X=N0-이 Y=N1-에 V
X: AGT (individual|group); Y: LOC (abstract object|event|action)

	저사람은	사사건건	우리일에	개입한다.
	jeo salam-eun	*sasageongeon*	*uli il-e*	*gaeibhanda*
	'that person-top'	'everything'	'our affairs-obl'	'interfere'
FrameNet	Participant_1	Manner	Event	Participation$_{\text{TARGET}}$

Figure 3: Comparisons between the corresponding information in Sejong Dictionary and the annotation in the Korean FrameNet dataset with regard to a single instance. The meaning of the above instance is "that person interferes in our affairs constantly and meddles in everything".

On the other hand, the information on its natural segmentation is preserved by keeping the *eojeols* at the top of the morphemes that are split from the corresponding *eojeol* following the CoNLL-U format. The frames are therefore annotated on morphemes instead of *eojeols*, and lexical morphemes and functional morphemes are split into separate tokens. Although whether a token is lexical or functional is not explicitly annotated, the morphologically enhanced annotation scheme allows the parser to subconsciously distinguish functional components from the lexical morphemes that trigger semantic frames. This is in line with the aforementioned agglutinative feature of the Korean language.

We neither exclude the functional morphemes from the annotated targets or arguments, nor do we introduce additional labels to annotate them. This is because (1) functional morphemes are parts of the targets/arguments (Park and Kim, 2023) that a parser should identify (therefore must not be labeled as O's), (2) introducing additional labels would potentially confuse the parser, worsening the model performance, and (3) separation between lexical morphemes and functional morphemes can be performed in postprocessing steps if necessary. Based on the above, we implement a script that automatically converts existing Korean FrameNet datasets into the morpheme-based format, and back-converts our morpheme-based format into the original format. Conversions in both directions rely on alignments between *eojeols* and morphemes and assignments of tags on the aligned tokens. The morphologically enhanced FrameNet datasets are therefore prepared using the aforementioned script for further experiments.

4. Experiments and Results

We perform semantic frame parsing on the proposed datasets and the original datasets respectively. Specifically, we focus only on the argument extraction task with the assumption that the frame target and the frame itself have already been given to the parsers as inputs. This allows us to approach the problem as a sequence labeling task, where the tokens are the lexical units and the classes are frame elements. We remap the frame-specific elements into general arguments given that the Korean FrameNet datasets contain more than 2,000 unique frame elements which are hard to be classified with the limited

index	word	lexeme	target	frame	annotation
...					
16	30	30	_	_	B-Time
17-19	여년간	_	_	_	_
17	여	여	_	_	I-Time
18	년	년	_	_	I-Time
19	간	간	_	_	I-Time
20-21	오스트리아를	_	_	_	_
20	오스트리아	오스트리아	_	_	B-Dependent_entity
21	를	을	_	_	I-Dependent_entity
22-24	통치한	_	_	_	_
22	통치	통치	통치하다.v	Being_in_control	B-FrameTarget
23	하	하	_	_	I-FrameTarget
24	ㄴ	은	_	_	I-FrameTarget
25-26	좌익이	_	_	_	_
25	좌익	좌익	_	_	B-Controlling_entity
26	이	이	_	_	I-Controlling_entity
...					

Figure 4: Example of the morphologically enhanced FrameNet data: *30yeonyeongan oseuteulialeul jibaehan jwaigi...* ('The left wing that ruled Austria for over 30 years...')

		KoELECTRA-Base			KR-BERT-char16424		
		pkfn	jkfn	skfn	pkfn	jkfn	skfn
exact	*eojeol*	0.2523 ± 0.0215	0.3968 ± 0.0445	0.8091 ± 0.0003	0.2964 ± 0.0229	0.3493 ± 0.0281	0.8041 ± 0.0009
	morph	0.3319 ± 0.0807	0.6528 ± 0.0135	0.6054 ± 0.0056	0.3070 ± 0.0868	0.6256 ± 0.0127	0.5343 ± 0.0042
partial	*eojeol*	0.3051 ± 0.0224	0.4438 ± 0.0444	0.8279 ± 0.0003	0.3475 ± 0.0226	0.4010 ± 0.0267	0.8241 ± 0.0008
	morph	0.4091 ± 0.0694	0.7152 ± 0.0096	0.7373 ± 0.0047	0.4094 ± 0.0677	0.6929 ± 0.0083	0.6627 ± 0.0036

Table 4: The cross validation mean ± standard deviation of exact and partial F_1 scores on *eojeol*- and morpheme-based variants of pkfn, jkfn and skfn datasets.

instances. Hence, our classification is over five classes: `O`, `B-FrameTarget`, `I-FrameTarget`, `B-Argument`, and `I-Argument`, following the BIO tagging scheme.

Our parsers are based on the pre-trained `KoELECTRA-Base-v3` discriminator model[1] and the `KR-BERT-char16424` model (Lee et al., 2020)[2], and are fine-tuned for the argument detection task using our proposed datasets. The models have their own tokenizers whereas they process the already segmented *eojeols* and morphemes from our proposed datasets. The hyperparameter settings are as follows:

Epochs	3
Learning Rate	5e-5
Batch Size (train)	128
Batch Size (eval)	256
Evaluation Strategy	epoch

For evaluation of the parsers' performance, we use measurements as suggested in SemEval'13 (Jurgens and Klapaftis, 2013). Specifically, we use the exact F_1 score to choose our best epoch out of three training epochs. The morpheme-based outputs are converted back into the *eojeol*-based format for fair comparisons of the results.

The exact and partial F_1 scores of parsers trained on *eojeol*- and morpheme-based data using 2-fold cross-validation is summarized in Table 4. It is observed that the parsers trained on the morpheme-based datasets substantially outperform those trained on the *eojeol*-based alternatives with regard to the pkfn and jkfn data. The disagreement from skfn may be owning to the fact that the argument boundaries are not direct inheritances from its source data, as discussed in Section 2. This potentially causes some discrepancies within the skfn dataset, and the discrepancies further hinder the morpheme-based parsers from obtaining satisfactory performance since morphemes as smaller units than *eojeols* are more sensitive to the boundaries. Overall, we find our proposed scheme an effective approach to representing Korean FrameNet data as previous work suggested in other Korean language processing tasks. As future work, resolving the discrepancies within skfn will necessitate a comprehensive strategy. Primarily, it is essential to conduct a more thorough investigation into the underlying causes of these inconsistencies, as detailed in Section 2, with the goal of fortifying the dataset's reliability. This may involve the refinement of argument boundary derivation processes or the exploration of alternative methods to ensure greater precision and consistency in annotations.

5. Conclusion

We propose a morphologically enhanced scheme to annotate Korean FrameNet datasets, which is motivated by the linguistic features of the Korean

[1] https://github.com/monologg/KoELECTRA

[2] https://github.com/snunlp/KR-BERT

language. We convert existing Korean FrameNet data into our proposed format through an alignment algorithm, and further train parsers on the standardized morpheme-based data as well as the original word-based data for the comparison purpose. The results show that the Korean FrameNet data, once enhanced morphologically, improves the parsing outcomes when using datasets in which annotations are securely inherited from their sources. We consider the proposed morpheme-based scheme a standardized way to annotate Korean FrameNet datasets for parsing.

6. Acknowledgements

This work was supported by National Research Foundation of Korea grants funded by the Korean government (Ministry of Science and ICT) (2021R1F1A1063474) for KyungTae Lim.

7. Ethics Statement

We have no ethical concerns.

8. Bibliographical References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 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*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Daniel Bauer, Hagen Fürstenau, and Owen Rambow. 2012. The Dependency-Parsed FrameNet Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3861–3867, Istanbul, Turkey. European Language Resources Association (ELRA).

Yige Chen, Eunkyul Leah Jo, Yundong Yao, KyungTae Lim, Miikka Silfverberg, Francis M Tyers, and Jungyeul Park. 2022. Yet Another Format of Universal Dependencies for Korean. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5432–5437, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yige Chen, KyungTae Lim, and Jungyeul Park. 2023. Korean Named Entity Recognition Based on Language-Specific Features. *Natural Language Engineering*, FirstView:1–25.

Key-Sun Choi, Young S. Han, Young G. Han, and Oh W. Kwon. 1994. KAIST Tree Bank Project for Korean: Present and Future Development. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, pages 7–14, Nara Institute of Science and Technology. Nara Institute of Science and Technology.

Younggyun Hahm, Jiseong Kim, Sunggoo Kwon, and Key-Sun Choi. 2018. Semi-automatic Korean FrameNet Annotation over KAIST Treebank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Silvana Hartmann and Iryna Gurevych. 2013. FrameNet on the Way to Babel: Creating a Bilingual FrameNet Using Wiktionary as Interlingual Connection. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1363–1373, Sofia, Bulgaria. Association for Computational Linguistics.

Richard Johansson and Pierre Nugues. 2006. A FrameNet-Based Semantic Role Labeler for Swedish. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 436–443, Sydney, Australia. Association for Computational Linguistics.

David Jurgens and Ioannis Klapaftis. 2013. SemEval-2013 task 13: Word sense induction for graded and non-graded senses. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299, Atlanta, Georgia, USA. Association for Computational Linguistics.

Jeong-uk Kim, Younggyun Hahm, and Key-Sun Choi. 2016. Korean FrameNet Expansion Based on Projection of Japanese FrameNet. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 175–179, Osaka, Japan. The COLING 2016 Organizing Committee.

Sangah Lee, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. 2020. KR-BERT: A Small-Scale Korean-Specific Language Model. *ArXiv*, abs/2008.03979.

Alessandro Lenci, Martina Johnson, and Gabriella Lapesa. 2010. Building an Italian FrameNet through Semi-automatic Corpus Analysis. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Birte Lönneker-Rodman and Collin F. Baker. 2009. The FrameNet model and its applications. *Natural Language Engineering*, 15(3):415–453.

Kyoko Hirose Ohara, Seiko Fujii, Hiroaki Saito, Shun Ishizaki, Toshio Ohori, and Ryoko Suzuki. 2003. The Japanese FrameNet project: A preliminary report. In *Proceedings of pacific association for computational linguistics*, pages 249–254.

Martha Palmer, Shijong Ryu, Jinyoung Choi, Sinwon Yoon, and Yeongmi Jeon. 2006. Korean PropBank.

Jungyeul Park and Mija Kim. 2023. A role of functional morphemes in Korean categorial grammars. *Korean Linguistics*, 19(1):1–30.

Jungyeul Park, Sejin Nam, Youngsik Kim, Younggyun Hahm, Dosam Hwang, and Key-Sun Choi. 2014. Frame-Semantic Web : a Case Study for Korean. In *ISWC-PD'14: Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272*, pages 257–260, Riva del Garda, Italy. International Semantic Web Conference.

Jungyeul Park and Francis Tyers. 2019. A New Annotation Scheme for the Sejong Part-of-speech Tagged Corpus. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 195–202, Florence, Italy. Association for Computational Linguistics.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. FrameNet II: Extended Theory and Practice. Technical report, International Computer Science Institute, Berkeley, CA.

Liping You and Kaiying Liu. 2005. Building Chinese FrameNet database. In *2005 International Conference on Natural Language Processing and Knowledge Engineering*, pages 301–306.