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

LREC-COLING 2024
20-25 May, 2024

Background

- FrameNet: a rich linguistic resource that reveals the frame semantics of natural languages (Baker et al., 1998; Lönneker-Rodman and Baker, 2009; Ruppenhofer et al., 2010)
  - FrameNet datasets in languages other than English, e.g., Japanese (Ohara et al., 2003), Chinese (You and Liu, 2005), Italian (Lenci et al., 2010), Swedish (Johansson and Nugues, 2006), multilingual (Hartmann and Gurevych, 2013)
  - FrameNet parsing has been made possible for English FrameNet (Bauer et al., 2012)
- FrameNet for Korean
  - Park et al. (2014): a Korean FrameNet dataset converted from English FrameNet sentences originated from English Propbank
  - Kim et al. (2016): similar to Park et al. (2014), by projecting the Japanese FrameNet to translated Korean texts
  - Hahm et al. (2018): a Korean FrameNet dataset based on the KAIST treebank (Choi et al., 1994)

Motivation

- A problem of existing Korean FrameNet datasets
  - Korean is an agglutinative language, and its functional morphemes are attached to the lexical morphemes to form the natural segments of the language, i.e., *eojeols*
  - Functional morphemes hardly contribute to the semantics of the sentence, and a great number of tokens will be introduced to the vocabulary if *eojeols* is considered the basic unit during tokenization
- Previous attempts at morpheme-based schemes for Korean NLP
  - Part-of-speech tagging (Park and Tyers, 2019)
  - Dependency parsing (Chen et al., 2022)
  - Named entity recognition (Chen et al., 2023)
- Nevertheless, how the morpheme-based approach can be employed in annotating Korean FrameNet datasets has not been extensively studied

Data

- Originally developed and published by KAIST (Park et al., 2014; Kim et al., 2016; Hahm et al., 2018), including multiple sources from which the data are collected
- We choose parts of the whole dataset originating from three sources
  - Korean FrameNet data from Korean PropBank (PKFN)
  - Korean FrameNet data from the Japanese FrameNet (JKFN)
  - Korean FrameNet data from the Sejong Dictionary (SKFN)

# 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 1: The number of frames per lexical unit (LU) for each of the Korean FrameNet datasets. An LU is a word with its part-of-speech.

The PKFN Dataset

- Sourced from the Korean PropBank (Palmer et al., 2006)
- Contains mainly verbal targets, along with a few adjectival targets

	북한이	국무부의 테러지원국 명단에	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 JKFN Dataset

- Projected from the Japanese FrameNet (Ohara et al., 2003) by Kim et al. (2016)
- Given the syntactic similarities between Korean and Japanese, the JKFN data are direct and literal translations from the original word chunks in the Japanese FrameNet, in which way the projected JKFN data preserves the boundaries of the frames

	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` TARGET
	↓		↓	↓
Korean FrameNet	`Theme`		`Path`	`Path_shape` TARGET
	초등학생이	파란 불에	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 SKFN Dataset

- Based on the example sentences in the Sejong dictionary
- The example sentences in the dictionary are usually short, and a sentence in the SKFN data carries a single frame only
- All frame targets in SKFN are verbs
- Automatic detection and mapping between frame elements and arguments for the frame of the given predicate are conducted

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

	저 사람은	사사건건	우리 일에	개입한다.
	jeo salam-eun	*sasageongeon*	*uli il-e*	*gaeibhanda*
	'that person-TOP'	'everything'	'our affairs-OBL'	'interfere'
FrameNet	`Participant_1`	`Manner`	`Event`	`Participation` 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.

Morphologically Enhanced FrameNet

- We propose a morpheme-based scheme for Korean FrameNet data that leverages the linguistic properties of the Korean language
- Issues concerning the previous *eojeol*-based scheme
    - Korean's natural segmentation (*eojeol*) can consist of both the lexical morpheme and its postposition, such as a particle that marks tense or case
    - In other words, a single token is a mixture of the lexical part and the functional part
    - ...whereas only the lexical morphemes contribute to the semantic meaning and determine the lexical units the targets instantiate and the semantic frame they evoke
    - 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

Morphologically Enhanced FrameNet

- The sentence is decomposed into morphemes as the basic unit of tokens following the CoNLL-U format
- The frames are annotated on morphemes instead of *eojeols*

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...').

Morphologically Enhanced FrameNet

- We neither exclude the functional morphemes from the annotated targets or arguments, nor do we introduce additional labels to annotate them
  - 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)
  - Introducing additional labels would potentially confuse the parser, worsening the model performance
  - Separation between lexical morphemes and functional morphemes can be performed in postprocessing steps if necessary
- 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
- 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

Experiments

- We perform semantic frame parsing on the proposed datasets and the original datasets respectively
- 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
- 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 instances
    - 5 classes: `O`, `B-FrameTarget`, `I-FrameTarget`, `B-Argument`, and `I-Argument`, following the BIO tagging scheme

Experiments

- Parsers: the `KoELECTRA-Base-v3` discriminator model[1] and the `KR-BERT-char16424` model (Lee et al., 2020)[2]
- Fine-tuned for the argument detection task using our proposed datasets
- While the models have their own tokenizers, they process the already segmented *eojeols* and morphemes from our proposed datasets

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

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

Results

- We use measurements as suggested in SemEval'13 (Jurgens and Klapaftis, 2013)
- The morpheme-based outputs are converted back into the *eojeol*-based format for fair comparisons of the results

		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

Figure 5: 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.

- 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 which potentially causes some discrepancies

Conclusion

- We propose a morphologically enhanced scheme to annotate Korean FrameNet datasets motivated by the linguistic features of the Korean language
- We convert existing Korean FrameNet data into our proposed format through an alignment algorithm
- Parsers are trained 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

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*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Bauer, D., Fürstenau, H., and Rambow, O. (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).

Chen, Y., Jo, E. L., Yao, Y., Lim, K., Silfverberg, M., Tyers, F. M., and Park, J. (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.

Chen, Y., Lim, K., and Park, J. (2023). Korean Named Entity Recognition Based on Language-Specific Features. *Natural Language Engineering*, FirstView:1–25.

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*, pages 7–14, Nara Institute of Science and Technology. Nara Institute of Science and Technology.

Hahm, Y., Kim, J., Kwon, S., and Choi, K.-S. (2018). Semi-automatic Korean FrameNet Annotation over KAIST Treebank. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hartmann, S. and Gurevych, I. (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.

Johansson, R. and Nugues, P. (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.

Jurgens, D. and Klapaftis, I. (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.

Kim, J.-u., Hahm, Y., and Choi, K.-S. (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.

Lee, S., Jang, H., Baik, Y., Park, S., and Shin, H. (2020). KR-BERT: A Small-Scale Korean-Specific Language Model. *ArXiv*, abs/2008.03979.

Lenci, A., Johnson, M., and Lapesa, G. (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).

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

Ohara, K. H., Fujii, S., Saito, H., Ishizaki, S., Ohori, T., and Suzuki, R. (2003). The Japanese FrameNet project: A preliminary report. In *Proceedings of pacific association for computational linguistics*, pages 249–254.

Palmer, M., Ryu, S., Choi, J., Yoon, S., and Jeon, Y. (2006). Korean PropBank.

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

Park, J., Nam, S., Kim, Y., Hahm, Y., Hwang, D., and Choi, K.-S. (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.

Park, J. and Tyers, F. (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.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., and Scheffczyk, J. (2010). FrameNet II: Extended Theory and Practice. Technical report, International Computer Science Institute, Berkeley, CA.

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