A Linguistically-Informed Annotation Strategy for Korean Semantic Role Labeling

Yige Chen[†] KyungTae Lim[‡] Jungyeul Park[¶]
[†]The Chinese University of Hong Kong, Hong Kong [‡]SeoulTech, South Korea
[¶]The University of British Columbia, Canada
yigechen@link.cuhk.edu.hk ktlim@seoultech.ac.kr jungyeul@mail.ubc.ca

Abstract

Semantic role labeling is an essential component of semantic and syntactic processing of natural languages, which reveals the predicate-argument structure of the language. Despite its importance, semantic role labeling for the Korean language has not been studied extensively. One notable issue is the lack of uniformity among data annotation strategies across different datasets, which often lack thorough rationales. In this study, we suggest an annotation strategy for Korean semantic role labeling that is in line with the previously proposed linguistic theories as well as the distinct properties of the Korean language. We further propose a simple yet viable conversion strategy from the Sejong verb dictionary to a CoNLL-style dataset for Korean semantic role labeling. Experiment results using a transformer-based sequence labeling model demonstrate the reliability and trainability of the converted dataset.

Keywords: Semantic role labeling, Korean, Predicate-argument structure

1. Introduction

Semantic role labeling (SRL) is a crucial component of semantic and syntactic analyses in natural language processing (NLP), which concerns the sequence labeling task of identifying the semantic role label for each constituent related to a particular target verb in a parse, revealing the predicate-argument structure of the sentence (Gildea and Jurafsky, 2002; Palmer et al., 2010). SRL is a well-defined task that can be conducted by machines, and there have been shared tasks for SRL, such as CoNLL-2004 shared task for semantic role labeling (Carreras and Màrquez, 2004).

There have been studies attempting to utilize existing datasets, such as the proposition bank (PropBank) (Palmer et al., 2005) or FrameNet (Baker et al., 1998; Ruppenhofer et al., 2010), for SRL tasks. For instance, SRL has been conducted using PropBank for multiple languages, (Akbik et al., 2015) and using FrameNet for Swedish (Johansson and Nugues, 2006). However, research on effective methods to utilize existing language resources in Korean for the purpose of SRL tasks is still lacking.

In the past few decades, there have been debates on the nature of arguments and modifiers as well as the semantic roles concerned in the field of generative grammar. In the context of Categorial Grammar (CG) (Ajdukiewicz, 1935; Bar-Hillel, 1953), researchers distinguish between two types of elements related to the predicate: complements and adjuncts. Complements are obligatory elements that complete the meaning of their head, while adjuncts are optional elements that modify the head's meaning (Dowty, 2003). Some other generativists who adopt frameworks like Principles and Parameters (P&P) (Chomsky, 1986, p.150-151) or Head-driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994) consider a three-way distinction (Carnie, 2002; Sag et al., 2003). In their perspective, the specifier (an immediate daughter of the phrase, usually the subject) and the complement (the sister of the head, usually the object) are unique for a verb phrase. To further clarify the discrepancy, both the 'specifier' and the 'complement' in P&P and HPSG fall in the category of 'complement' in Categorial Grammar. While it is not the study's intention to argue for or against certain linguistic theories, we shall observe the fact that linguistic argumentation serves as the guidance for the construction of language data. It is essential to define the notion of 'argument' and 'modifier' for SRL tasks, as such tasks involve finding arguments, and if applicable, modifiers of the target verb.

In this study, we present well-defined SRL annotation strategies for Korean based on the long-established linguistic argumentation and the well-documented linguistic properties of Korean. We clearly define the notion of 'argument' and 'modifier' for Korean SRL to reduce the ambiguities existing in the NLP community. We also introduce an efficient method to transform example sentences from the Sejong verb dictionary, a resource rich in syntactic and semantic information for verb lexemes, into a CoNLL-style SRL dataset. This method automatically assigns labels to tokens for targets and arguments. Satisfactory experiment results indicate the viability of our conversion ap-

[‡]Corresponding author.

proach.

2. Linguistic Properties of Korean

Korean is considered an agglutinative language with a word order different from English. To be specific, the functional morphemes in Korean are directly attached to the lexical morpheme to form a word, and such functional morphemes appear to succeed the stem of the word. Given their significance with regard to dataset construction and conversion, we discuss two major linguistic properties of Korean in this section.

2.1. Word order in Korean

Korean follows a subject-object-verb (SOV) word order. Typically in a Korean sentence, the object comes before the verb, while the subject is positioned before the phrase consisting of the object and the verb. Adjuncts are also placed at the front of the verb. Below is an example of a Korean sentence in (1), where the two arguments follow the SOV word order of Korean.

(1) 고양이가 쥐를 잡는다
goyangi-ga *jwi-leul* *jabneunda*
cat.nom mouse.acc catch
'A cat catches a mouse.'

where nom stands for *nominative* and acc stands for *accusative*. It can be observed that the subject in the sentence, namely 고양이가 (*goyangi-ga*), appears at the beginning of the sentence, succeeded by the object 쥐를 (*jwi-leul*) and the main verb 잡는다 (*jabneunda*). While syntacticians may argue for different underlying structures, Korean sentences on the surface level generally follow the aforementioned linear order.

2.2. Postpositions

Korean postpositions are suffixes that follow the stem of the word. In terms of the postpositions for nominal words and phrases, they oftentimes indicate the case. Table 1 shows some of such postpositions in Korean and the corresponding cases.

Postposition	Case
은/는 (*n*-)*eun*	topic
이/가 *i/ga*	nominative
을/를 (*l*-)*eul*	accusative
으로/로 (*eu*)*ro*	instrumental
에서 *eseo*	locative/ablative
의 *ui*	genitive

Table 1: Examples of some commonly used postpositions in Korean.

Regarding the previous example sentence (1) in §2.1, the postposition *-ga* specifies the subject being the cat, whereas the postposition *-leul* specifies the object being the mouse. As a result, postpositions serve as good indicators of the semantic roles of the arguments.

2.3. Arguments in Korean

It has been introduced in §1 that linguists hold various ideas regarding the notions and classifications of arguments of the predicate. As our aim is to better address the technical issue of SRL, we adopt a simplified yet consistent definition of arguments for Korean that is mainly based on Categorial Grammar. Arguments of a predicate in Korean are syntactically mandatory and semantically necessary for both the sentence structure and the meaning of the sentence to be completed. Particularly, the arguments need to bear the case, both implicitly and explicitly[1], in Korean. Accordingly, such arguments should be captured and specified in the subcategorization frame of the predicate. All other constituents that are not part of the mandatory elements of the predicate are considered modifiers, and the modifiers do not appear in the subcategorization frame.

3. Korean SRL Data

3.1. Korean PropBank SRL

The SRL dataset converted from the Korean PropBank[2] (Lee et al., 2015) offers rich linguistic annotations of predicate-argument relations for Korean SRL. The Korean PropBank originated from the Virginia Corpus and the Newswire Corpus, and it consists of approximately 186,300 tokens in total. In the SRL dataset converted from the Korean PropBank, although words are further divided into morphemes, the annotations for arguments and target verbs are made at the word level. This indicates that the Korean PropBank SRL dataset treats words as the fundamental units carrying semantic roles.

3.2. NIKL SRL

The NIKL SRL dataset was constructed and organized by the National Institute of Korean Language adopting the annotation strategy from the Electronics and Telecommunications Research Institute of Korea. The dataset contains approximately 2,000,000 tokens stored in JSON format.

3.3. Existing Issues

The two datasets above serve as invaluable resources for Korean SRL. However, it is observed that the annotation strategies of the two datasets

[1] Here, the word 'implicitly' refers to the abstract Case it factually bears, and the word 'explicitly' refers to the morphological case the lexeme is attached to on the surface level (usually realized as postpositions).

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

have potential defects. The SRL dataset converted from Korean PropBank contains word-level semantic role labels, with argument segments limited to single words. However, it is demonstrated in Figure 2 that arguments in Korean can be constituents that consist of more than one word[3]. The annotation strategy of the SRL dataset converted from Korean PropBank overlooks the capability of a semantic role label to span two or more tokens, rendering the dataset incapable of accurately capturing argument boundaries.

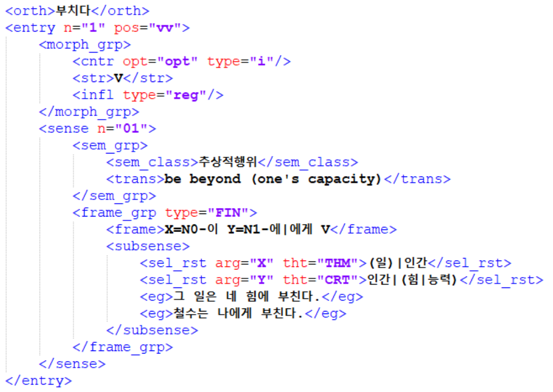
On the other hand, annotations in the NIKL dataset only cover the lexical morphemes of the argument without including the postposition, namely the functional morpheme that carries the case, as part of the argument. We consider such annotation strategies misleading, in that as discussed in §2.3, arguments in Korean should bear cases. It is not uncommon for arguments to contain explicitly marked morphological cases as affixes in natural languages. For instance, Latin nouns and noun phrases bear morphological cases through which abstract Cases are realized (Lacabrese, 1998). Splitting a word into two subsegments and excluding one subsegment from the annotation would both lack rationales and be difficult to implement.

4. Creating a New Dataset

The Sejong dictionary is part of the Sejong corpus organized by the National Institute of Korean Language.[4] We are interested in the verbs in the dictionary, which are sorted in such a way that for every verbal lexeme, a separate entry is created. Such entries consist of the syntactic and semantic information of the verbs for each of the senses included. The possible subcategorization frames and semantic roles of the arguments are provided, along with example sentences. Figure 1 shows an example of the lexeme 부치다 (*buchida*) while the sense included is “be beyond (one’s capacity)”.[5]

4.1. Conversion

To convert the Sejong dictionary data in the XML format into the CoNLL-style format, a set of syntactic and semantic information from the entries is required, including orthography (`orth`), subcategorization frame (`frame`), and semantic roles (`sel_rst`). Based on the information provided in the Sejong dictionary, we annotate semantic role labels onto the CoNLL-formatted example sentences as below.



```
<orth>부치다</orth>
<entry n="1" pos="vv">
    <morph_grp>
        <cntr opt="opt" type="i"/>
        <str>V</str>
        <infl type="reg"/>
    </morph_grp>
    <sense n="01">
        <sem_grp>
            <sem_class>추상적행위</sem_class>
            <trans>be beyond (one's capacity)</trans>
        </sem_grp>
        <frame_grp type="FIN">
            <frame>X=N0-이 Y=N1-에|에게 V</frame>
            <subsense>
                <sel_rst arg="X" tht="THM">(일)|인간</sel_rst>
                <sel_rst arg="Y" tht="CRT">인간|(힘|능력)</sel_rst>
                <eg>그 일은 네 힘에 부친다.</eg>
                <eg>철수는 나에게 부친다.</eg>
            </subsense>
        </frame_grp>
    </sense>
</entry>
```

Figure 1: Example of the lexeme 부치다 (*buchida*) in the Sejong dictionary whose sense is ‘be beyond (one’s capacity)’.

Morphological analysis Example sentences in the Sejong dictionary are tokenized and tagged with their parts of speech using the morpheme-level tagger.[6] The morpheme-based outputs of the tagger are further converted such that each token is a word instead of a morpheme, hence complying with the CoNLL scheme.

Dependency parsing The part-of-speech tagged sentences are further fed to Stanza (Qi et al., 2020), a neural dependency parser, to obtain dependency relations between the words. The dependency relations and heads of the words are saved along with their parts of speech, given that the above information is compatible with the CoNLL format.

Chunking To attach semantic role labels to the tokens, it is crucial to split the sentences into chunks that may represent arguments. Such a chunk can be a single word, a phrase, or a clause. Figure 2 shows an example of an chunked Korean sentence, where the first three chunks are the arguments of the target verb, which appears to be the last chunk. The target verb of a sentence is first extracted, which defines the stopping point of the chunking process. This is because, given the aforementioned word order of Korean, the arguments of a verb precede the verb in most cases. Chunking is, therefore, performed on the segment ahead of the target verb, and it relies on the language-specific parts-of-speech (XPOS) to define the boundaries of the chunks. A subsegment is extracted as a chunk when during the iteration of the tokens, the final token ends with a postposition as suggested by XPOS. While the parsing results are available which may potentially determine the argument chunks, Stanza’s outputs are not satisfactory and are therefore only used when ambiguity occurs.

[3] Also known as *eojeol*, i.e., the natural segmentation of Korean texts that is split by the whitespace.

[4] `https://korean.go.kr`

[5] There are other senses of the lexeme, as well as other lexemes in the same surface form, included in the XML file.

[6] `https://doi.org/10.5281/zenodo.3236528`

산자부 장관은	이 본부장을	본부장직에서	사직시켰다
sanjabu jang-gwan-eun	*i bonbujang-eul*	*bonbujangjig-eseo*	*sajigsikyeossda*
[$_{\text{nom}}$ Minister of Industry]	[$_{\text{acc}}$ Director Lee]	[$_{\text{ajt}}$ position of general manager]	[$_{\text{TARGET}}$ made resign]

‘The Minister of Commerce, Industry and Energy resigned Director Lee from his position as Director.’

Figure 2: Example of a Korean sentence split into chunks where `ajt` stands for *adjunct*.

```
# text = 그 일은 네 힘에 부친다.
# target = 부치다
# frame = X=N0-이 Y=N1-에|에게 V
# arg="X" tht="THM", (일)|인간
# arg="Y" tht="CRT", 인간|(힘|능력)
1  그      그        DET    MM       2  det         B-ARG0
2  일은    일+은     NOUN   NNG+JX   5  dislocated  I-ARG0
3  네      네        DET    MM       4  nummod      B-ARG1
4  힘에    힘+에     NOUN   NNG+JKB  5  obl         I-ARG1
5  부친다  부치+는다 VERB   VV+EF    0  root        TARGET
6  .       .         PUNCT  SF       5  punct       O
```

Figure 3: Converted CoNLL-style instance of an example sentence in Figure 1: *geu il-eun ne him-e buchi-n-da.* (‘The task is beyond your strength.’). (i) X=N0-이 Y=N1-에|에게 V (X=N0-*i* (‘nom’) Y=N1-*e*|*ege* (‘dat’) V) where `dat` stands for *dative*; (ii) arg="X" tht="THM", (일)|인간 (*il*)|*ingan* (‘(work)|human’); (iii) arg="Y" tht="CRT", 인간|(힘|능력) *ingan*|(*him*|*neunglyeog* (‘human|(strength|ability)’)

Chunk-frame alignment The subcategorization frame of a sentence provides the postpositions of all the arguments allowed by a certain sense of the target verb in the sentence. The task of assigning argument labels to the chunks is essentially pairing the suggested arguments in the frame with the extracted chunks. This is achieved by iterating the chunks and annotating each of the frame arguments to the chunk that bears the same postposition. The process is conducted in a linear manner, which means that each frame argument only finds the first unannotated chunk that has the postposition as the frame argument does.

Figure 3 shows the final output of such conversions after the alignment is conducted. We preserve all morphological and syntactic information obtained from the morphological analyzer and the parser, whereas the labels of the target verb and the arguments are appended for the purpose of SRL tasks. The BIO scheme is adopted for the SRL tag set to cope with the fact that an argument may consist of two or more tokens. The indices of the argument tags are inherited from the indices specified in the frame, ranging from 0 to 3.

4.2. Exceptions

While the conversion method detailed above is straightforward and can handle most of the example sentences in the Sejong dictionary, it is noticed that there are some exceptional cases that the conversion method fails to resolve. We hereby list some of such exceptional cases which we hope to address in future work. All exceptions have been removed from the final version of the dataset.

Null postposition The case of a noun in Korean can be sometimes phonologically covert, especially in colloquial speech. This results in null postpositions on the surface form, which means there is no observable postposition in the textual form for the argument. Since the chunking method extensively relies on the postpositions to be the boundaries, an argument bearing the null postposition cannot be properly chunked and is usually concatenated with the succeeding chunk.

The 도 (*do*) postposition Korean possesses an auxiliary postposition, namely 도 (*do*, ‘as well’, `JX`=auxiliary postposition), which occupies the position of any overt case marker. The chunk-frame alignment will therefore fail when the postposition appears, as the prescribed particle in the frame is no longer attached to the argument.

5. Experiments and Results

To validate the quality of the converted data from the Sejong dictionary, we select a subset of the converted CoNLL-style dataset and conduct an SRL experiment using a transformer-based pretrained model. A total number of 20,437 sentences are therefore chosen, and we perform a 2-fold cross-validation with regard to the training set and the test set. The SRL labels include `TARGET` and `ARG`$_n$ under the BIO annotation scheme where n denotes the argument index inherited from the Sejong dictionary.

We use the `KoELECTRA-Base-v3` discriminator model[7] dedicated to the Korean language, and fine-tune the model on the training set of our

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

dataset. The nature of the task is sequence labeling, in that given the target verb (TARGET), the model detects the arguments of the target. We train the model over 6 epochs, using a learning rate of 5e-5. The evaluation strategy is adopted from SemEval'13 (Jurgens and Klapaftis, 2013). We report the exact precision, recall, and F_1 score of the sequence labeling result on the test set using the model that obtained the best F_1 score on the validation set out of 6 training epochs, as in Table 2. The satisfactory experimental results suggest the feasibility of our dataset serving as a trainable and usable source for Korean SRL.

Precision	Recall	F_1
0.946 ± 0.003	0.971 ± 0.002	0.954 ± 0.003

Table 2: Cross-validation results (mean ± standard deviation) of exact matches on test set.

6. Conclusion

In this study, we describe the preferred annotation approach for Korean SRL based on the linguistic features of Korean and previous linguistic research on the nature of the predicate-argument structure. Specifically, we revisit and revise the notion of 'argument' for Korean SRL, hoping to address potential confusion in the NLP community. We further propose an effective method for the conversion from the Sejong verb dictionary to a CoNLL-style SRL dataset. Experiment results suggest that our converted SRL dataset is trainable and reliable.

7. 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.

8. Ethics Statement

We have no ethical concerns.

9. Bibliographical References

Kazimierz Ajdukiewicz. 1935. Die syntaktische Konnexität. *Studia philosophica*, 1:1–27.

Alan Akbik, laura chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating High Quality Proposition Banks for Multilingual Semantic Role Labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China. Association for Computational Linguistics.

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.

Yehoshua Bar-Hillel. 1953. A quasi-arithmetical notation for syntactic description. *Language*, 29(1):47–58.

Andrew Carnie. 2002. *Syntax: A Generative Introduction*. Introducing linguistics. Wiley-Blackwell, New Jersey, United States.

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. In *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 89–97, Boston, Massachusetts, USA. Association for Computational Linguistics.

Noam Chomsky. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Praeger Scientific, New York.

David Dowty. 2003. The dual analysis of adjuncts/complements in Categorial Grammar. In Ewald Lang, Claudia Maienborn, and Cathrine Fabricius-Hansen, editors, *Modifying Adjuncts*, pages 33–66. De Gruyter Mouton, Berlin, Boston.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288.

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.

Andrea Lacabrese. 1998. Some Remarks on the Latin Case System and Its Development in Romance. In *Theoretical analyses on Romance languages: selected papers from the 26th Linguistic Symposium on Romance Languages* (*LSRL XXVI*), *Mexico City, 28-30 March 1996*, pages 71–126. John Benjamins Publishing Company.

Changki Lee, Soo-jong Lim, and Hyun-Ki Kim. 2015. Korean Semantic Role Labeling Using Structured SVM. *Journal of KIISE*, 42(2):220–226.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. *Semantic Role Labeling*. Synthesis Lectures on Human Language Technologies. Springer Cham.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago, Illinois, USA.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. 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.

Ivan A. Sag, Thomas Wasow, and Emily M. Bender. 2003. *Syntactic Theory: A Formal Introduction*, 2nd editio edition. CSLI Lecture Notes. The University of Chicago Press, Chicago, IL, USA.