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

LREC-COLING 2024
20-25 May, 2024

Authors

Yige Chen

KyungTae Lim

Jungyeul Park

Background

- Semantic role labeling (SRL): identifying the semantic role label for each constituent related to a particular target verb in a parse, and revealing the predicate-argument structure of the sentence (Gildea and Jurafsky, 2002; Palmer et al., 2010)
  - Oftentimes as a sequence labeling task
  - Existing datasets, such as CoNLL-2004 shared task for semantic role labeling (Carreras and Màrquez, 2004)
- Use of existing non-SRL resources for SRL tasks
  - Proposition bank (PropBank) (Palmer et al., 2005), for multiple languages, (Akbik et al., 2015)
  - FrameNet (Baker et al., 1998; Ruppenhofer et al., 2010) for Swedish (Johansson and Nugues, 2006)

Motivation

- Research on utilizing existing resources in Korean for SRL tasks is still lacking
  - Especially those that take into account the linguistic features of the Korean language
- Despite all kinds of linguistic debates on the nature of arguments and modifiers and the semantic roles concerned, how they should be defined for SRL remains unclear
  - Categorial Grammar (CG) (Ajdukiewicz, 1935; Bar-Hillel, 1953) considers a binary distinction: complements (obligatory elements that complete the meaning of their head) and adjuncts (optional elements that modify the head's meaning) (Dowty, 2003)
  - Principles and Parameters (P&P) (Chomsky, 1986, p.150-151) and Head-driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994) consider a three-way distinction: specifiers, adjuncts, and complements (Carnie, 2002; Sag et al., 2003)

Linguistic Properties of Korean

- Word order: subject-object-verb (SOV)
- Korean 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
- An example, where the postposition *-ga* specifies the subject being the cat, whereas the postposition *-leul* specifies the object being the mouse

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

Arguments in Korean

- 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.
  - The arguments need to bear the case, both implicitly and explicitly, in Korean.
  - 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.

Existing Korean SRL Data

- The SRL dataset converted from the Korean PropBank (Lee et al., 2015)
  - Issues: argument segments are limited to single words, whereas arguments in Korean can be constituents that consist of more than one word (i.e., *eojeols*)
- The NIKL SRL dataset constructed by the National Institute of Korean Language
  - Issues: annotations only cover the lexical morphemes of the argument without postposition (the functional morpheme that carries the case) as part of the argument
  - 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)

Dataset Creation

- We utilize example sentences from the Sejong verb dictionary, which is part of the Sejong corpus organized by the National Institute of Korean Language (`https://korean.go.kr`)
- The data is converted into a CoNLL-style SRL dataset, with a method automatically assigning labels to tokens for targets and arguments
- Format of the Sejong verb dictionary
  - 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

Format of the Sejong Verb Dictionary

```
<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)'.

Conversion

- Preparations
  - A set of syntactic and semantic information from the Sejong dictionary is extracted, including orthography (`orth`), subcategorization frame (`frame`), and semantic roles (`sel_rst`)
- Morphological analysis
  - Example sentences in the Sejong dictionary are tokenized and tagged with their parts of speech using the morpheme-level tagger and converted to word-level
- Dependency parsing
  - The part-of-speech tagged sentences are fed to Stanza (Qi et al., 2020) to obtain dependency relations

Conversion

- Chunking
  - The target verb of a sentence is first extracted, which defines the stopping point of the chunking process
  - Chunking is 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
- Chunk-frame alignment
  - Pairing the suggested arguments in the frame with the extracted chunks by iterating the chunks and annotating each of the frame arguments to the chunk that bears the same postposition

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

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

Converted CoNLL-style Data

```
# 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 2: 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.'), under the BIO annotation scheme.

Exceptions

- Null postposition
  - The case of a noun in Korean can be sometimes phonologically covert, which results in null postpositions on the surface form
  - An argument bearing the null postposition cannot be properly chunked since the chunking method relies on the postpositions to be the boundaries
- 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

Experiments

- Data: we select a subset of the converted CoNLL-style dataset (20,437 sentences)
- Model: `KoELECTRA-Base-v3` discriminator model
- Task: SRL as sequence labeling, in that given the target verb (`TARGET`), the model detects the arguments of the target ($\texttt{ARG}_n$)

Results

- The evaluation strategy is adopted from SemEval'13 (Jurgens and Klapaftis, 2013)

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

Table 1: Cross-validation results (mean $\pm$ standard deviation) of exact matches on test set.

Conclusion

- 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
- 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

References

Ajdukiewicz, K. (1935). Die syntaktische Konnexität. *Studia philosophica*, 1:1–27.

Akbik, A., chiticariu, l., Danilevsky, M., Li, Y., Vaithyanathan, S., and Zhu, H. (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.

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.

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

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

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

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

Dowty, D. (2003). The dual analysis of adjuncts/complements in Categorial Grammar. In Lang, E., Maienborn, C., and Fabricius-Hansen, C., editors, *Modifying Adjuncts*, pages 33–66. De Gruyter Mouton, Berlin, Boston.

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

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.

Lacabrese, A. (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.

Lee, C., Lim, S.-j., and Kim, H.-K. (2015). Korean Semantic Role Labeling Using Structured SVM. *Journal of KIISE*, 42(2):220–226.

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

Palmer, M., Gildea, D., and Xue, N. (2010). *Semantic Role Labeling*. Synthesis Lectures on Human Language Technologies. Springer Cham.

Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago, Illinois, USA.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Celikyilmaz, A. and Wen, T.-H., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. 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.

Sag, I. A., Wasow, T., and Bender, E. M. (2003). *Syntactic Theory: A Formal Introduction*. CSLI Lecture Notes. The University of Chicago Press, Chicago, IL, USA, 2nd editio edition.