

A Linguistically-Informed Annotation Strategy for Korean SRL

Data available through: <https://www.korean.go.kr>

k-SRL

 Yige Chen[†]  KyungTae Lim[‡]  Jungyeul Park[¶]

[†]The Chinese University of Hong Kong, Hong Kong, [‡]SeoulTech, South Korea, and [¶]The University of British Columbia, Canada.

```
<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>
```

It shows an example of the lexeme 부치다 (*buchida*) while the sense included is “be beyond (one’s capacity)”. There are other senses of the lexeme, as well as other lexemes in the same surface form, included in the XML file.

산자부 장관은 이 본부장을 본부장직에서 사직시켰다
sanjabu jang-gwan-eun i bonbujang-eul bonbujangjig-eseo sajigsikyeosda
 [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.’
 Example of a Korean sentence split into chunks where ajt stands for *adjunct*.

```
# text = 그 일은 네 힘에 부친다. geu il-eun ne him-e buchi-n-da. ('The task is beyond your strength.')
# 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

Converted CoNLL-style instance

Acknowledgement: 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.

Experiments and Results

- 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 (ARG_n)
- The evaluation strategy is adopted from SemEval’13

Precision	Recall	F ₁
0.946 ± 0.003	0.971 ± 0.002	0.954 ± 0.003

Cross-validation results (mean ± 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



Download the paper →

