Korean Particle Error Detection via Probabilistic Parsing

Markus Dickinson
Indiana University
md7@indiana.edu

Chong Min Lee
Georgetown University
cml54@georgetown.edu

1. ICALL for Korean

- Parsing learner input only developed for a small number of languages (cf. Van de Weert & Altman, 2003)
- Korean presents challenges:
  - Scrambling language, allowing for freer word order
  - Morphological units combine into a word/phrase level (`어절`)

2. Korean particles as target form

- Postpositional particles relate a verb & its arguments
  - No one-to-one mapping with English
  - `어` = `to`, `에` = `with`, `을` = `to`
  - `어` = `to`, `in`
- Particles are difficult to learn (Ko et al., 2004)

3. Context: Korean ICALL system

- Online chat between 2 language learners (Dickinson et al., 2008)
  - Beginning students of Korean learning particles.
- Picture-based information-gap task & word bank used to limit the range of input vocabulary.

4. Error detection & Annotated corpora

4.1 Constraint relaxation approaches

- Treat grammar as a set of constraints
- Identify types of constraints for which learners may vary—e.g.,
  - subject-verb agreement
- Allow these constraints to be relaxed

4.2 Error grammar approaches

- Identify rules for which learners may vary
- Add error rules (mal-rules) to the grammar
  - e.g., S → NP, VP

Goal: Explore the use of annotated corpora for assisting in detecting Korean particle errors

Why use annotated corpora to train probabilistic technology?

- Annotation represents significant, reusable linguistic analysis
- Saves time in constructing a grammar
- Technology trained on corpus annotation well-understood & state-of-the-art

What are the potential pitfalls in using annotated corpora?

- Annotation may not be the most appropriate for task
  - e.g., missing properties such as agreement features
- Probabilistic parser does not distinguish grammaticality
  - Provides analysis for any sentence, even ill-formed ones

How can we get the grammar we want from the annotation we have?

5. Making the grammar fit

5.1 Add more information

   - Distinguish subject from object NPs based on parent: NP'S vs. NP'VP
2. Intuition: Use hand-crafted linguistic generalizations
   - Fill in agreement properties for NPs based on pronoun type
   - (1) He/PRP laughs/VBZ → He/PRP-3s laughs/VBZ
3. External source:
   - Use additional technology, corpora, or knowledge bases
     - POS tags trained on PTB and SUSANNE corpora
     - Add argument relations from semantic annotation

5.2 Use less information

- Non-predictive (sparse) information: Less information makes better predictions
  - Subtagging: verb tense in Russian does not predict noun case (Hana et al., 2004)
- Complementary information: A less informative model could show different patterns
  - Lexicalized & unlexicalized PCFGs have different patterns (Metcalf and Boyd, 2006)

6. A case study: The Korean Treebank (V2.0)

How can we tell whether a particle is being used correctly?

1. What is the main verb, and what are the surrounding NPs? → Available
2. Which NPs are dependent upon the verb? → Partially available
   - (2) (NP (S (NP-ADV 지난 1866 년 병원방역 당시)) (VP (어설령 오지 않은 년간의 전염병)) (VP (NP-ADV 병원방역 당시)) (NP (NP-ADV 이순신의 시절에)) (NP (NP-ADV 지난 1866 년 병원방역 당시)) (NP (NP-ADV 이순신의 시절에)) (NP (NP-ADV 지난 1866 년 병원방역 당시)))
   - Old documents that the French army stole at an event in 1866:
     - Which verb (VV) takes subject NP requires additional knowledge
3. What is the relation between the verb and its NPs? → Partially available
   - (3) (S (VP (S (NP-ADV 이순신의 시절에)) (NP (NP-ADV 지난 1866 년 병원방역 당시)) (NP (NP-ADV 이순신의 시절에)) (NP (NP-ADV 지난 1866 년 병원방역 당시))) (NP (NP-ADV 이순신의 시절에)) (NP (NP-ADV 지난 1866 년 병원방역 당시)) (NP (NP-ADV 이순신의 시절에)) (NP (NP-ADV 지난 1866 년 병원방역 당시)))
   - Both companies decided to sue each other before the WTO:
     - COMP is a general grammatical term, realizable by several particles

7. Recovering information not in annotation

7.1 NPs dependent upon the verb

Solution: Recover via head rules

- Convert to dependency structures for parser training
  - Directly identify grammatical relations between words (regardless of word order)
  - Indicate direct argument requirements of verbs
- Korean dependency structure defined in terms of relationship between 어절.

7.2 Grammatical relation between the verb and surrounding NPs

Solution: Extract from Korean PropBank

- Korean PropBank labeled with extended relations:
  - EXT (extent), DIR (direction), LOC (location), TEMP (temporal), ...
- 3,800 predicate tokens annotated out of 23,700 in Korean TreeBank 2.0.

References


