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. Vandeventer Faltin, 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
    - ∗ 'to' ≈ *-에*, *-에게*, *-께*, *-한테* *-에* ≈ '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_{pl} VP_{sg}

Commonalities

Require grammar capturing relevant language properties
- Akin to annotation found in a corpus

Anticipation-based in some ways
- Need to know how (learner) language can vary

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

1. **Implicit annotation:** Recover latent annotation (cf. Klein and Manning, 2003)
   - distinguish subject from object NPs based on parent: NPˆS vs. NPˆVP
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 taggers 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 병인양요 당시) (S (NP-SBJ 프랑스+군+이) (VP
a historic event in 1866 France soldier-SBJ
(NP-OBJ *T*-1) (CV (VV 약탈+하+어) 가/VV+은)))) (NP-1 외규장각 고문서))
steal go old document

'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-SBJ 양측+이) (VP (NP-COMP WTO+에) (VV 제소+하+기+으로)))) 하+었+다)
both side-SBJ WTO-DAT sue do

'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)
  - directly indicate argument requirements of verbs
- Korean dependency structure defined in terms of relationship between *어절*s.

7.2 Grammatical relation between the verb and surrounding NPs

Solution: **Extract from Korean PropBank**

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

Particle Error Detection

Use multiple models (and less information)

First model: original corpus
- Probabilities of words with specific particles
- Influenced by surface forms of words

Second model: corpus without particles
- General argument relations between verb and particular words
- Less influenced by surface forms

Both models can add annotation to data, informing error diagnosis
→ MISMATCHES MIGHT INDICATE ERRORS

References

Dickinson, Markus, Rebecca Sachs, Yunkyoung Kang and Soojeong Eom (2008). Integrating ICALL into synchronous CMC. 25th Annual CALICO Conference. San Francisco.

Hana, Jirka, Anna Feldman and Chris Brew (2004). A Resource-light Approach to Russian Morphology: Tagging Russian using Czech resources. In *Proceedings of EMNLP-04*. Barcelona.

Klein, Dan and Christopher D. Manning (2003). Accurate Unlexicalized Parsing. In *Proceedings of ACL-03*. Sapporo, Japan.

Ko, S., M. Kim, J. Kim, S. Seo, H. Chung and S. Han (2004). *An analysis of Korean learner corpora and errors*. Hanguk Publishing Co.

Metcalf, Vanessa and Adriane Boyd (2006). Head-lexicalized PCFGs for Verb Subcategorization Error Diagnosis in ICALL. In *Workshop on Interfaces of Intelligent Computer-Assisted Language Learning*. Columbus, OH.

Vandeventer Faltin, Anne (2003). Syntactic error diagnosis in the context of computer assisted language learning. Thèse de doctorat, Université de Genève, Genève.