Introduction and Motivation

ICALL goal: provide intelligent feedback to learners on language production (cf. Heift and Schulze, 2007)

- 1st step: automatically assign linguistic analysis to sentence
- Requires grammatical description of (in)appropriately-used constructions
  - e.g., subject-verb agreement

Need to carefully consider the appropriate representation for a language to support the analysis of learner constructions

Supporting feedback

Idea: Use corpus annotation to build technology appropriate for distinctions learners know

- Potentially saves time & effort
- Connects to state-of-the-art parsing (e.g., Charniak and Johnson, 2005; Nivre et al., 2007)

But is corpus annotation appropriate for analyzing learner data?

Overarching Goal: provide framework for re-using corpus annotation in a way which supports providing feedback

Modeling learner language

Dickinson and Lee (to appear) outline a framework for converting corpus annotation into an analysis that is desirable

- Promising initial results, but only initial results . . .

Goals for this work-in-progress:

1. Use a real learner corpus for evaluation
2. Adapt other NLP technology—namely, a POS tagger
3. Continue to develop parsing technology

Background: Korean particles

Korean postpositional particles indicate grammatical functions, thematic roles, and locations of people & objects

- Similar to English prepositions, but wider range of functions:

  1) \texttt{Sumi-neun chaek-i pilyohae-yo}
     \texttt{Sumi-TOP book-SBJ need-polite}
     ‘Sumi needs a book.’

- Focus of ICALL systems for Korean & Japanese (Dickinson et al., 2008; Nagata, 1995)

Korean particles: expected errors

Learners of Korean often misuse particles (Ko et al., 2004)

2) \texttt{*Sumi-neun chaek-eul pilyohae-yo}
   \texttt{Sumi-TOP book-OBJ need-polite}
   ‘Sumi needs a book.’

Lee et al. (to appear) & Ko et al. (2004) categorize particle errors by learners of Korean into 6 types; we focus on 2:

- \textit{Omission} & \textit{replacement} errors: 60%+ of particle errors made by beginning learners (Lee et al., to appear)
Usage of Korean particles

We focus on syntactic postpositional particles
- Case markers: indicate relationship between verb & noun
      Sumi-SBJ Jisu-DAT book-OBJ give-PAST-DECL
      ‘Sumi gave Jisu a book.’
- Modifiers (cf. prepositions): indicate specific lexical, syntactic, & semantic information between verb & noun

Chong Min Lee, Soojeong Eom, and Markus Dickinson
Towards Analyzing Korean Learner Particles

Parsing for learner language

What we want: dependencies

We want dependency structures

Constituency-to-dependency conversion is straightforward (cf., e.g., Chung, 2004; Seo, 1993; Kudo and Matsumoto, 2000).
- Dependency relations provide relevant feedback information

Chong Min Lee, Soojeong Eom, and Markus Dickinson
Towards Analyzing Korean Learner Particles

Limitations of current annotation

Particle annotation

KTB has syntactic role particles PCA (case), PAD (adverbial), & PAN (adnominal)
- Each label realizable by several particles
  (5) a. (NP-ADV naenyeon-e/PAD) boneos-reul batneunta next year+at bonus-OBJ receive
  b. (NP-ADV naenyeon-buteo/PAD) boneos-reul batneunta next year+from bonus-OBJ receive
  c. * (NP-ADV naenyeon-eso/PAD) boneos-reul batneunta next year+from bonus-OBJ receive

Chong Min Lee, Soojeong Eom, and Markus Dickinson
Towards Analyzing Korean Learner Particles

Solution: Put particle information into labels
1. Normalization: group particles that function in same manner
   - their selection relies on non-syntactic factors
<table>
<thead>
<tr>
<th>POS</th>
<th>Class</th>
<th>Particles</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>SBJ</td>
<td>-kkeseo, -seo, -ka/-i, -eseo</td>
</tr>
<tr>
<td></td>
<td>OBJ</td>
<td>-eul/-reul</td>
</tr>
<tr>
<td>PAN</td>
<td>UI</td>
<td>-ui</td>
</tr>
<tr>
<td>PAD</td>
<td>EUROSSEO</td>
<td>euroseos</td>
</tr>
<tr>
<td></td>
<td>EUROPUTEO</td>
<td>europuteos</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>-e</td>
</tr>
<tr>
<td></td>
<td>EGE</td>
<td>-ege</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
2. Threshold: focus on particles appearing > 50 times in corpus
Removing information from annotation

But isn’t this highly redundant?
- e.g., EGE will be used whenever e.g., is encountered

However: Labels with particle names predict the presence of particular (type of) particle, even if that particle is not there
- Idea: Remove redundancy for a second model by removing particles from word forms
- Parsing disagreements between models provide platform for error detection (cf. Metcalf and Boyd, 2006)
- Shows success on artificially-created errors in news text

Adapting a learner corpus for evaluation

So far: Evaluated on artificial errors

Next step: Use a Korean learner corpus for evaluation
- annotated for particle errors (Lee et al., to appear)

Learner corpus changes (1)
Data compatibility

To evaluate positives & negatives of error detection before fully moving to unaltered learner data, we make some changes:
1. Correct misspelled/malformed particles (error type 4)
2. Other words are not corrected, to keep the data more real
3. Correct spacing errors in particles (type 6)
   - e.g., particles split from words are merged
4. Fix incorrect sentence boundaries
5. Tokenize punctuation separately

Learner corpus changes (2)
Fine-grained annotation

We do not deal with discourse-based errors: honorifics & topics
- Discourse-based errors can occur within the error types we investigate (substitutions, omissions)

How can we properly evaluate our system on lexical case errors?

Solution: Add error subtype information to the surface-level annotation scheme of Lee et al. (to appear)
- Indicate if error is honorific-based or topic-based

Adapt a POS tagger

So far: Used POS tags from the corpus

Next step: Use POS tagger for Korean (Han and Palmer, 2004)
- Based on same corpus tagset
- Good performance
  - Precision: 95.43%
  - Recall: 95.04%

But tagger is designed for regular language
- How well will the tagger work on learner language?
  - cf. Shih et al. (2000); van Rooy and Schäfer (2002)

Initial tagging vs. hand-cleaned results
New genre

Moving from one genre to another leads to tagging problems:
- Unknown words lead to mis-segmentation & mis-tagging
  - (6) *jungkuk/VV+eo/ECN ⇔ jungkukeo/NNC
    - China+language
  - (7) *hae/NNC+yo/PAU ⇔ ha/VV+yo/EFN
    - to do+verb-ending
- Formal and informal registers
  - Tagger trained on formal newtext: uses da ending
  - Learner data is informal: uses yo ending, e.g., for haeyo:
Initial tagging vs. hand-cleaned results

Underlying forms

Tagger mishypothesizes underlying form (needed for feedback):
- e.g., *deulesesseoyo* in a context to mean 'listen':

\[
\begin{align*}
\text{Tagger mishypothesis} & \quad \Rightarrow \\
\text{Mispaced context} & \quad \Rightarrow \\
\text{Mispaced context} & \quad \Rightarrow \\
\end{align*}
\]

Chong Min Lee, Soojeong Eom, and Markus Dickinson
Towards Analyzing Korean Learner Particles

Summary and Outlook

Summary:
- Examined how to provide parsing model for information about Korean postpositional particles
- Identified challenges & opportunities for using POS tagger
- Began to evaluate on learner data
- Highlighted the need to add more syntactically-annotated data

Outlook:
- Extend the parser to handle a wider range of data
- Integrate tools into a more robust error detection module (cf., e.g., Tetreault and Chodorow, 2008)
- Use dependency labels to perform error diagnosis in a real ICALL setting (Dickinson et al., 2008)

Acknowledgements

Our thanks to:
- Sun-Hee Lee & SeokBae Jang for providing their learner corpus
- Ross Israel for general work & insights
- Rebecca Sachs & Yunkyoung Kang for support on Korean ICALL
- Members of the IU autumn 2009 L700 seminar for feedback on this general line of research


Han, Chung-Hye, Na-Rae Han, Eon-Suk Ko and Martha Palmer (2002). Development and Evaluation of a Korean Treebank and its Application to NLP. In Proceedings of LREC-02.


