The Effect of Learner Corpus Size in Grammatical Error Correction of ESL Writings

Tomoya Mizumoto†, Yuta Hayashibe†, Mamoru Komachi†, Masaaki Nagata‡, Yuji Matsumoto†
†Nara Institute of Science and Technology, ‡NTT Communication Science Laboratories

Background

Types	%	Types	%
article	19.23	Lexical choice of noun	7.04
noun number	13.88	Lexical choice of verb	6.90
preposition	13.56	pronoun	6.62
tense	8.77	agreement	5.25

Table1. Distribution of errors on KJ corpus[1]

*Spelling errors are excluded from target of annotation in KJ corpus

- A lot of previous works deal with one or a few restricted types of learners' error
  preposition [Rozovskaya+], verb form [Lee+], tense [Tajiri+], spelling/article/preposition/word form [Park+]
- It was not until recently that large scale learner corpora became widely available for error correction
  - little is known about the effect of learner corpus size in ESL error correction

Main Contribution

1. the first attempt to use large scale learner corpus to correct all types error
2. show the effect of learner corpus size on the phrase-based SMT and its advantages and disadvantage

System and Corpus

- Use phrase-based SMT to conduct unrestricted error correction
  - Several studies about grammatical error correction using phase-based SMT [Brockett+ 06], [Mizumoto+ 11], [Ehsan+ 12]
- Use data from a language learning SNS Lang-8[2]
  - Language learners post their writing on the site to be corrected by native speaker
  - Obtain pairs of learner's and corrected sentence in large scale
  - Crawled blog entries found in Lang-8 as of December 2010
    - Used writings written by Japanese ESL learners (509,116 sentence pairs)
    - Filtered noisy sentences -> 391,699 sentence pairs

Summary

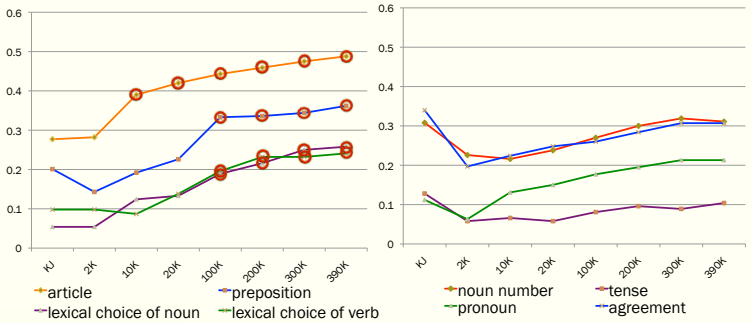
- Show the effect of learner corpus size on the phrase-based SMT and its advantages and disadvantage in grammatical error correction
- Increasing the size of learner corpus;
  - improvement: article, preposition, lexical choice
  - little improvement: noun number, agreement, tense

Experiment and Results

- To see the effect of corpus size, we compare systems
  - Using 1. Lang-8 with different size and 2. KJ (also used test data, 5-fold CV)
- Evaluation metrics (F-measure, precision, recall)
  - P and R for each type of errors are calculated from tp, fp, fn based on error tags

learner：He talked to me _ his life of Kyoto, and he took me _ Kyoto.
correct：He talked to me about his life in Kyoto, and he took me to Kyoto.
system：He talked _ me _ his life on Kyoto, and he took me to Kyoto.

P=1/2, R=1/2



Red circle indicate that the difference of result using Lang-8 and KJ is statistically significant ($p<0.01$)

Discussion

- Classify errors into two types

1. Get better correction by increasing corpus size (article, preposition, lexical choice)

e.g. article

✓ learner：I like a chocolate very much.
correct：I like _ chocolate very much.

e.g. lexical choice of noun

✓ learner：My cycle was injured, but I wasn't.
correct：My bicycle was damaged, but I wasn't.

2. Have little relationship with corpus size. (noun number, tense, pronoun, agreement)

e.g. noun number

✓ learner：I read various type of books.
correct：I read various types of books.

✗ learner：There is a big snoopy dools in my room.
correct：There is a big snoopy doll in my room.

- Word "dools" is displaced from "a big"
- A proper noun "snoopy" is inserted between "a big" and "dools"
  * "dools" is also a spelling error

To solve, needs to conduct generalization using POS or consider dependency relation

e.g. tense

✓ learner：If I'll live in Saitama, I must have ...
correct：If I _ live in Saitama, I must have ...

✗ learner：The weather is very sunny, so we were ...
correct：The weather was very sunny, so we were ...

- System fails to find tense agreement in the complex sentence

To solve, involves global context [Tajiri+ 12]

e.g. agreement

✓ learner：Flowers is very beautiful.
correct：Flowers are very beautiful.

✗ learner：I think, reading comics are not "reading"
correct：I think, reading comics is not "reading"

- Pattern is unseen
- no way to capture the relation between subject "reading" and "are"

To solve, needs to get the subject-verb relation considering a dependency structure

[1] http:www.gsk.or.jp/catalog/GSK2012-A/catalog_e.html [2] http://lang-8.com