Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners

Tomoya Mizumoto†, Mamoru Komachi†, Masaaki Nagata‡, Yuji Matsumoto†

† Nara Institute of Science and Technology
‡ NTT Communication Science Laboratories
Background

- Number of Japanese language learners has increased
  - 3.65 million people in 133 countries and regions

- Only 50,000 Japanese language teachers overseas
  - High demand to find good instructors for writers of Japanese as a Second Language (JSL)
Recent error correction for language learners

- NLP research has begun to pay attention to second language learning
  - Most previous research deals with restricted type of learners’ errors
  - E.g. research for JSL learners’ error
    - Mainly focus on Japanese case particles
  - Real JSL learners’ writing contains various errors
    - Spelling errors
    - Collocation errors
Error correction using SMT [Brockett et al., 2006]

- Proposed to correct ESL learners’ errors using statistical machine translation
- Advantage is that it doesn’t require expert knowledge
- Learns a correction model from learners’ and corrected corpora
  - Not easy to acquire large scale learners’ corpora
- Japanese sentences is not segmented into words
  - JSL learners’ sentences are hard to tokenize
Purpose of our study

1. Solve the knowledge acquisition bottleneck
   - Create a large-scale learners’ corpus from error revision logs of language learning SNS

2. Solve the problem of word segmentation errors caused by erroneous input using SMT techniques with extracted learners’ corpus
SNS sites that helps language learners

- **smart.fm**
  - Helps learners’ practice language learning

- **Livemocha**
  - Offers course of grammar instructions, reading comprehension exercises and practice

- **Lang-8**
  - Multi-lingual language learning and language exchange SNS
  - Soon after the learners write a passage in a learning language, native speakers of the language correct errors in it
**Lang-8 data**

<table>
<thead>
<tr>
<th>Language</th>
<th>English</th>
<th>Japanese</th>
<th>Mandarin</th>
<th>Korean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sentences</td>
<td>1,069,549</td>
<td>925,588</td>
<td>136,203</td>
<td>93,955</td>
</tr>
</tbody>
</table>

- Sentence of JSL learners: 925,588
- Corrected sentences: 1,288,934

Example of corrected sentence from Lang-8

Learner: ビデオゲームをやまシた。
Correct: ビデオゲームをやりまシた。

Pairs of learners sentence and corrected sentence
Types of correction

- Correction by insertion, deletion and substitution
  Learner: ビデオゲームをやました。  
  Correct: ビデオゲームをやりましした。

- Correction with a comment
  Learner: 銭湯行った。  
  Correct: 銭湯行った。 いつ行ったかがあるほうがいい  

- Exist “corrected” sentences to which only the word “GOOD” is appended at the end
  Learner: 銭湯に行った。  
  Correct: 銭湯に行った。 GOOD

- Removing comments
  Number of sentence pair: 1,288,934 → 849,894
### Comparison of Japanese learners’ corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Data size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Lang-8 corpus</td>
<td>849,894 sentences 448 MB</td>
</tr>
<tr>
<td>Teramura error Data (1990)</td>
<td>4,601 sentences 420 KB</td>
</tr>
<tr>
<td>Ohso Database (1998)</td>
<td>756 files 15 MB</td>
</tr>
<tr>
<td>JSL learners parallel database (2000)</td>
<td>1,500 JSL learners’ writings</td>
</tr>
</tbody>
</table>
Error correction using SMT

\[ \hat{e} = \arg \max_e P(e|f) = \arg \max_e P(e)P(f|e) \]

**SMT**

- **e**: target sentences
- **f**: source sentences
- **P(e)**: probability of the language model
- **P(f|e)**: probability of the translation model
Error correction using SMT

\[ \hat{e} = \arg \max_{e} P(e|f) = \arg \max_{e} P(e)P(f|e) \]

**SMT**
- e: target sentences
- f: source sentences
- \( P(e) \): probability of the language model
- \( P(f|e) \): probability of the translation model

**Error correction**
- e: corrected sentences
- f: Japanese learners’ sentences
- \( P(e) \): probability of the language model
- \( P(f|e) \): probability of the translation model

- Can be learned from a monolingual corpus of the language to be learned
- Can be learned from the sentence-aligned learners’ corpus
Difficulty of handling the JSL learners’ sentences

- Word segmentation is usually performed as a preprocessing

- JSL learners’ sentences contain many errors and hiragana (phonetic characters)
  - hard to tokenize by traditional morphological analyzer
Difficulty of handling the JSL learners’ sentences

E.g.

Learner: でもじょうずじゃりません

Correct: でもじょうずじゃありません

tokenize

Learner: でも じょうずじゃりません

Correct: でも じょうず じゃ ありません
Character-wise model

- Character-wise segmented
  - e.g. でもじょずじゃりません
    → でもじょずじゃりません

- Not affected by word segmentation errors
  - Expected to be more robust

Learner: でもじょずじゃりません
Correct: でもじょうずじゃりません
Experiment

- Carried out an experiment to see
  1. the effect of corpus size
  2. the effect of granularity of tokenization
Experimental setting

Methods

- Baseline: Word-wise model
- Proposed method: Character-wise model
  - Language model: 3-gram
  - Language model: 5-gram

Data

- Extracted from revision logs of Lang-8
  - 849,849 sentences
  - Test: 500 sentences
  - Re-annotated 500 sentences to make gold-standard
Evaluation metrics

- **BLEU**
  - Adopted to BLEU for automatic assessment of ESL errors. [Park and Levy, 2011]
  - Followed their use of BLEU in the error correction task of JSL learners
    - JSL learners’ sentences are hard to tokenize by morphological analyzer
    - Character-based BLEU
Larger the corpus, the higher the BLEU

- Character-wise model: Character 5-gram

The difference is not statistically significant
Character-wise models are better than word-wise model

- **TM Training corpus:** 0.3M sentences
  - Achieves the best result

<table>
<thead>
<tr>
<th></th>
<th>Word 3-gram</th>
<th>Character 3-gram</th>
<th>Character 5-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80.72</td>
<td>81.63</td>
<td>81.81</td>
</tr>
</tbody>
</table>
Both 0.1M and 0.3M model corrected

Learner: またど もう ありがとう
(Thanks, Mantadomou (OOV))
Correct: まだど うも ありがとう
(Thank you again)

Learner: TRUTH わ 美しいです
(TRUTH wa beautiful)
Correct: TRUTH は 美しいです
(TRUTH is beautiful)
Learner: 学生なら たる たら 学校に行ける
(The learner made an error in conjunction form)
Correct: 学生なら たら 学校に行ける
(Becoming a student, I can go to school)
0.1M : 学生なら ため 学校に行ける
(I can go to school to be student)
0.3M : 学生なら たら 学校に行ける
(Becoming a student, I can go to school)
Conclusions

- Make use of a large-scale corpus from the revision logs of a language learning SNS
- Adopted SMT approaches to alleviate the problem of erroneous input from learners
  - Character-wise model outperforms the word-wise model
- Apply method using SMT techniques with extracted learners’ corpus to error correction of English as a second language learners
Handling the comment

- Conduct the following three pre-processing steps
  1. If the corrected sentence contain only “GOOD” or “OK”, we don’t include it in the corpus
  2. If edit distance between the learner’s sentences and corrected sentences is larger than 5, we simply drop the sentence for the corpus
  3. If the corrected sentence ends with “OK” or “GOOD”, we remove it and retain the sentence pair.
Statistical machine translation

Parallel Corpus
- English
- Japanese

Japanese Corpus

I like English 一 私は英語が好き ...

Translation Model

English sentence

Japanese sentence

Language Model

Japanese sentence

Japanese sentence

TM is learned from sentence-aligned parallel corpus

LM is learned from Japanese monolingual corpus
Japanese error correction

Learners’ Corpus

Learner
Correct

Japanese Corpus

Correct sentence

Learner’s sentence
Translation Model

Correct sentence

Correct sentence

Correct sentence

Language Model

TM is learned from sentence-aligned learners’ corpus

LM is learned from Japanese monolingual corpus

私が英語が好き 一 私は英語が好き ...

Image