Extracting a Chinese Learner Corpus from the Web: Grammatical Error Correction for Learning Chinese as a Foreign Language with Statistical Machine Translation

Yinchen ZHAO*, Mamoru KOMACHI* & Hiroshi ISHIKAWA*
*Graduate School of System Design, Tokyo Metropolitan University, Japan
*chou.innchenn@gmail.com

Abstract: In this paper, we describe the TMU system for the shared task of Grammatical Error Diagnosis for Learning Chinese as a Foreign Language (CFL) at NLP-TEA1. One of the main issues in grammatical error correction for CFL is a data bottleneck problem. The Chinese learner corpus at hand (NTNU learner corpus) contains only 1,208 sentences in total, which is obviously insufficient for training supervised techniques. To overcome this problem, we extracted a large-scale Chinese learner corpus from a language exchange site called Lang-8, which results in 95,706 sentences (two million words) after cleaning. We used it as a parallel corpus for a phrase-based statistical machine translation (SMT) system, which translates learner sentences into correct sentences.

Keywords: Chinese learner corpus, web mining, grammatical error correction, statistical machine translation

1. Introduction

Recently, the application of natural language processing techniques to educational purpose is actively studied. For example, grammatical error correction for English as Second Language (ESL) learners has gained large attention in the past few years. Specifically, there were a number of shared tasks of grammatical error correction for ESL learners such as Helping Our Own (HOO) and Conference on Natural Language Learning (CoNLL). However, although there was a shared task of Chinese spelling error correction (Wu, Liu, & Lee, 2013), little attention has been paid to Chinese as a foreign language (CFL). One of the reasons why it is difficult to develop a grammatical error correction system for CFL is the lack of learner corpora. In this paper, we present a method to extract a learner corpus of Chinese from the web, and use it to build a grammatical error correction system for CFL. The main contributions of this paper is as follows:

1. To best of our knowledge, this is the first work that constructs a large-scale learner corpus of Chinese from the web. The corpus contains 100,000 sentences (2,000,000 words) and is annotated with corrections.
2. Although there exist several works which apply phrase-based statistical machine translation (SMT) to Chinese spelling correction task, it is the first work that adopts SMT to grammatical error correction task for CFL. The experimental result shows that our proposed approach is effective to build a precise error correction system.
3. Unlike previous works which use phrase-based SMT for Chinese spelling correction task, we propose to use character-wise tokenization and prove that character-wise tokenization is more robust than word-wise tokenization.
2. Extracting a Chinese Learner Corpus from the Web

To alleviate the problem of shortage of training data, we resort to extract a Chinese learner corpus from the web. We focused on a language exchange social networking service (SNS) called Lang-8\(^1\). Lang-8 offers a wide variety of languages that you can use when you write a blog entry. Fellow users correct your blog entry written in your learning language sentence by sentence, and you in turn may correct other users’ blog entry written in your mother tongue. Lang-8 facilitates the process of mutual “language exchange”. Up to date (August 2014), Lang-8 has about one million users where 50,000 of them are Chinese learners.

3. Grammatical Error Correction with Statistical Machine Translation

We decompose the task of grammatical error correction into two parts. First, we identify the location of errors using statistical machine translation trained on Chinese learner corpus. Second, we classify the type of errors using a simple heuristic rule using dynamic programming.

3.1 Error Identification with Statistical Machine Translation

We followed (Brockett, Dolan, & Gamon, 2006) to make a grammatical error correction system with phrase-based statistical machine translation. One of the advantages of the approach is that we can use off-the-shelf machine translation toolkits to build a grammatical error correction system if we have a learner corpus with sufficient size.

In their paper, grammatical error correction process is modeled using a noisy-channel model as follows:

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

where \(P(e)\) is a language model and \(P(f \mid e)\) is a translation model. In this paper, \(f\) corresponds to a learner sentence and \(e\) corresponds to a corrected sentence, respectively. The phrase-based SMT toolkit we used in this paper actually uses a log linear model which contains the noisy-channel model as follows:

\[
\hat{e} = \arg \max_e w^\top h
\]

where \(w\) is a weight vector and \(h\) is a feature function, respectively.

We propose two types SMT systems: word-based system and character-based system, depending on the pre-processing step of a learner corpus. The intuition behind using a character-wise segmentation is that learners of Chinese tend to write incorrect sentences, which may hurt the accuracy of the word segmentation. Character-based SMT is free from tokenization errors, while it is able to learn word-to-word or phrase-to-phrase correction patterns thanks to the phrase extraction heuristics.

\(^1\) http://lang-8.com/
3.2 Error Classification with Dynamic Programming

Once we identify the location of errors, we classify the type of errors using a simple heuristic rule. We use a dynamic programming algorithm to calculate the number of insertion, deletion and replacement operations for each sentence pair. We then classify the type of errors by the following pseudo-code:

Table 1: Pseudo-code for error type classification.

<table>
<thead>
<tr>
<th>Input: learner sentence $l$, system correction $c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output: error type $t$</td>
</tr>
<tr>
<td>$(i, d, r) \leftarrow \text{get_operations}(l, c)$</td>
</tr>
<tr>
<td>if $d &gt; 0$ and $i &gt; 0$</td>
</tr>
<tr>
<td>$t \leftarrow \text{&quot;Disorder&quot;}$</td>
</tr>
<tr>
<td>else if $r &gt; 0$</td>
</tr>
<tr>
<td>$t \leftarrow \text{&quot;Selection&quot;}$</td>
</tr>
<tr>
<td>else if $d &gt; 0$</td>
</tr>
<tr>
<td>$t \leftarrow \text{&quot;Redundant&quot;}$</td>
</tr>
<tr>
<td>else if $i &gt; 0$</td>
</tr>
<tr>
<td>$t \leftarrow \text{&quot;Missing&quot;}$</td>
</tr>
<tr>
<td>else</td>
</tr>
<tr>
<td>$t \leftarrow \text{&quot;correct&quot;}$</td>
</tr>
<tr>
<td>end if</td>
</tr>
<tr>
<td>return $t$</td>
</tr>
</tbody>
</table>

If a sentence contains only one error, this algorithm correctly returns the "Disorder" error type, while it may fail to classify "Selection" error type and output "Redundant" or "Missing" error types since these error types depend on the alignment of tokens in the original sentence and corrected sentence. In a preliminary experiment, we found that this confusion can be negligible. We did not explore the use of machine learning-based method because the training corpus provided by the organizer contains only 1,000 instances.

4. Experiments

In this section, we describe the experimental settings and results for the NLP-TEA1 shared task.

4.1 Data and Tools

We obtained the Lang-8 Learner Corpora v2.0. The corpora come with “blog id”, “sentence id”, “learning language”, “native language”, “learner sentences” and “corrected sentences”. We extracted blog entries whose “learning language” are set to “Mandarin”. The Chinese portion of the Lang-8 Learner Corpora consists of 29,595 blog entries (441,670 sentences). We discarded following sentences and kept 95,706 sentences at last.

- Too long (more than or equal to 20 words) or too short (less than or equal to 3 words).
- Not written in Chinese.
Any corrected sentence 1.3 times longer or more than the original one.\(^2\)

We used Moses 2.1.1 as a statistical machine translation toolkit with its default parameter. The training and testing was done using the scripts distributed as KFTT Moses Baseline v1.4 (Neubig, 2011). We did not perform minimum error rate training (Och, 2003). We trained an SMT system with two training corpora: the Lang-8 Chinese Learner Corpus with and without segmentation. In other words, we built a grammatical error correction system trained on a character-based phrasal SMT. We used jieba\(^3\)0.32 for Chinese text segmentation.

### 4.2 Results

Table 1 summarizes the false positive rate, accuracy, precision, recall and F1 scores for the formal run. Character-based approach outperformed word-based approach in all evaluation metrics. This confirms the hypothesis that word segmentation errors damage grammatical error correction for CFL.

We ranked the 2\(^{nd}\) at the false positive rate and accuracy out of six groups participated in the shared task. However, these evaluation metrics alone do not show the effectiveness of our approach, since there is a trade-off between these metrics. Note that we only show the accuracy, precision, recall and F1 scores at detection level, since the performance at identification level is almost the same.

Table 1: Experimental results for the formal run at NLP-TEA1. Accuracy, precision, recall and F1 scores are at the detection level.

<table>
<thead>
<tr>
<th></th>
<th>False Positive Rate</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMU-Run1:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Character-based</td>
<td>0.1977</td>
<td>0.5171</td>
<td>0.5399</td>
<td>0.2320</td>
<td>0.3245</td>
</tr>
<tr>
<td>TMU-Run2:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word-based</td>
<td>0.1691</td>
<td>0.5103</td>
<td>0.5287</td>
<td>0.1897</td>
<td>0.2792</td>
</tr>
</tbody>
</table>

### 5. Discussion

Our system achieved the worst (6/6) performance in terms of F1 score. The main reason is that we did not perform any parameter tuning at all even though the error distribution of the test corpus is very skewed (half of the sentences contain errors). In a preliminary experiment, we ran the minimum error rate training using BLEU (Papineni, Roukos, Ward, & Zhu, 2002), but after the optimization the system outputs almost no corrections. This is because the BLEU score will become higher if the system does not change the learner sentence since BLEU counts the n-gram overlaps between the system output and reference, regardless of the number of errors. Although BLEU is used to evaluate grammatical error correction as in (Park & Levy, 2011), it may not adequate to assess the quality of error correction systems. One possible direction is to optimize the SMT system using the F1 score with Z-MERT\(^4\).

Note that the shared task only requires participants to determine whether a given sentence contains an error or not, our system is capable of locating the position of errors. In

---

\(^2\) Some corrected sentences contain comments and annotations, which may harm word alignment for SMT.

\(^3\) https://github.com/fxsjy/jieba

\(^4\) http://cs.jhu.edu/~ozaidan/zmert/
addition, our system can identify multiple errors in a sentence (although it is out of scope of
this shared task).

Also, we would like to emphasize that we did not use any resources provided by the
organizer. It is interesting to use domain adaptation approach such as in … to better reflect
error distribution of the given domain (for example, 50% of the given test corpus contains
errors, which is not often the case in realistic setting).

If a sentence contains more than one error, the proposed error type classification
algorithm will output only one error type. Since the test corpus is controlled to contain only
one error, we opted for a simple rule for the shared task. However, it is possible that these
error types are not identical in real setting, so our future work includes error type
classification for each error.

6. Related Work

Recently, there is a number of works that deal with grammatical error correction for learners
of English as a Second Language.

Lang-8 is considered as one of the invaluable resources for knowledge acquisition for
second language learners. For example, Japanese learner corpus (Mizumoto, Komachi,
Nagata, & Matsumoto, 2011; Kasahara, Komachi, Nagata, & Matsumoto, 2011) and English
learner corpus (Tajiri, Komachi, & Matsumoto, 2012; Mizumoto, Hayashibe, Komachi,
Nagata, & Matsumoto, 2012) can be extracted from Lang-8. It is not surprizing that we can
extract a large corpus of Chinese learners since Chinese (Mandarin) is the third most popular
learning languages in Lang-8\(^5\), followed by English and Japanese.

The use of statistical machine translation techniques to grammatical error correction
was pioneered by (Brockett, Dolan, & Gamon, 2006), and has been adopted to many
researchers in grammatical error correction for ESL (Mizumoto, Hayashibe, Komachi,
Nagata, & Matsumoto, 2012; Buys & van der Merwe, 2013; Yuan & Felice, 2013; Behera &

Recently, similar approach is applied to Chinese spelling error correction as well (Wu,
However, all of these methods use word-based statistical machine translation, even though
some of them use character n-gram language model. One of our proposed models investigates
character-wise segmentation rather than word-wise one, and indicates that character-based
model can learn useful correction patterns if the training corpus is sufficiently large.

Error type classification has gained much attention, for example in English (Swanson &
Yamangil, 2012) and Japanese (Oyama, Komachi & Matsumoto, 2013). Although these
works use linguistically motivated annotation scheme used in previous work and thus
investigate machine learning-based approaches, the error type annotation scheme for the
NTNU learner corpus is based on edit operations and it is more appropriate to use rules rather
than machine learning.

7. Conclusion

In this paper, we described the TMU system for the Grammatical Error Diagnosis for CFL
Shared Task at NLPTEA1. To increase the number of the training corpus, we explored
the web for constructing a learner corpus of Chinese. We extracted 100,000 learner corpus from
the language exchange SNS, Lang-8, and used it to train an SMT-based grammatical error

\(^5\) http://cl.naist.jp/nldata/lang-8/
correction system. Although the system did not use any language resources provided by the shared task organizer, it performed precise error identification.

Acknowledgements

We would like to thank Xi Yangyang for granting use of extracted texts from Lang-8.

References


