Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education

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

Yinchen ZHAO[a], Mamoru KOMACHI[a*] & Hiroshi ISHIKAWA[a]
[a]*Graduate School of System Design, Tokyo Metropolitan University, Japan*
*komachi@tmu.ac.jp

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 obstacles 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 supervised learning-based techniques. To overcome this problem, we extract a large-scale Chinese learner corpus from a language exchange site called Lang-8, which results in 95,706 sentences (two million words). We use 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, educational applications of natural language processing techniques are actively studied. For example, grammatical error correction for English as a 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, 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. It contains 100,000 sentences (2M words) annotated with corrections.
2. It is the first work that adopts statistical machine translation (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 using 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 focus on a language exchange social networking service (SNS) called Lang-8[1]. Lang-8 offers a wide variety of languages that you can use to write a blog entry. Other users correct your blog entry written in your learning language, and you in turn 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 a 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 follow (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 an off-the-shelf machine translation toolkit to build a grammatical error correction system if we have a learner corpus with sufficient size.

In their paper, the 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 use in this paper actually uses a log linear model which contains the noisy-channel model as follows:

$$\hat{e} = \arg\max_{e} \mathbf{w}^{\mathrm{T}} \mathbf{h}$$

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.

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:

[1] http://lang-8.com/

Table 1: Pseudo-code for error type classification.

```
Input: learner sentence l, system correction c
Output: error type t
(i, d, r) ← get_operations(l, c)
if  d > 0 and i > 0
  t ← “Disorder”
else if r > 0
  t ← “Selection”
else if d > 0
  t ← “Redundant”
else if i > 0
  t ← “Missing”
else
  t ← “correct”
end if
return t
```

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. In a preliminary experiment, we found that this confusion can be negligible. We did not explore the use of machine learning-based classification 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.

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” is 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 word segmentation. In other words, we built a grammatical error correction

[2] Some corrected sentences contain comments and annotations, which may harm word alignment for SMT.

system trained on a character-based phrasal SMT in addition to a word-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 2nd at the false positive rate and accuracy out of six groups participated in the shared task. However, these evaluation metrics alone cannot verify the effectiveness of our approach, since there is a trade-off between these metrics. Note that we only reports the 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.

	False Positive Rate	Accuracy	Precision	Recall	F1
TMU-Run1: Character-based	0.1977	0.5171	0.5399	0.2320	0.3245
TMU-Run2: Word-based	0.1691	0.5103	0.5287	0.1897	0.2792

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. 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 addition, our system can identify multiple errors in a sentence (although it is out of scope of this shared task).

One of the side effects of using the Lang-8 corpus is that the error correction system misclassifies correct sentences as "Missing" errors since it tends to use commas where applicable. However, commas often make more natural Chinese expressions than original. For instance, consider the following example. The system output is more fluent than the original, but it is different from the gold standard annotation, which deteriorates performance.

Gold: 今天的天氣很好不怎麼熱　（Today's weather is good, not very hot.）
System: 今天的天氣很好，不怎麼熱

[3] https://github.com/fxsjy/jieba
[4] http://cs.jhu.edu/~ozaidan/zmert/

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 (Imamura, Saito, Sadamitsu & Nishikawa, 2012) 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

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 & Bhattacharyya, 2013; Junczys-Dowmunt & Grundkiewicz, 2014).

Recently, similar approach is applied to Chinese spelling error correction as well (Wu, Liu & Lee, 2013; Wu, Chiu & Chang, 2013; Liu, Cheng, Luo, Duh & Matsumoto, 2013). 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 proposed in previous work, 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 NLP-TEA1. To increase the number of training corpus, we explored the web for constructing a learner corpus of Chinese. We extracted 100,000 learner sentences paired with their correction from the language exchange SNS, Lang-8, and used it to train an SMT-based grammatical error correction system. We compared two types of segmentation for phrasal SMT and found that character-based SMT outperforms word-based SMT for CFL grammatical error correction. The system achieved moderate performance even though it did not use any language resources from the target domain.

[5] http://cl.naist.jp/nldata/lang-8/

Acknowledgements

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

References

Behera, B. & Bhattacharyya, P. (2013). Automated Grammar Correction Using Hierarchical Phrase-Based Statistical Machine Translation. *Proceedings of the International Joint Conference on Natural Language Processing*, (pages 937-941). Nagoya, Japan.

Brockett, C., Dolan, B. W., & Gamon, M. (2006). Correcting ESL Errors Using Phrasal SMT Techniques. *Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the ACL*, (pages 249-256). Sydney, Australia.

Buys, J. & van der Merwe, B. (2013). A Tree Transducer Model for Grammatical Error Correction. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, (pages 43–51). Sofia, Bulgaria.

Imamura, K., Saito, K., Sadamitsu, K. & Nishikawa, H. (2012). Grammatical Error Correction Using Pseudo-Error Sentences and Domain Adaptation. *Proceedings of the Annual Meeting of the ACL*, (pages 388-392). Jeju Island, Korea.

Junczys-Dowmunt, M. & Grundkiewicz, R. (2014). The AMU System in the CoNLL-2014 Shared Task: Grammatical Error Correction by Data-Intensive and Feature-Rich Statistical Machine Translation. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task (CoNLL-2014 Shared Task)*, (pages 25-33). Baltimore, USA.

Kasahara, S., Komachi, M., Nagata, M., & Matsumoto, Y. (2011). Correcting Romaji-Kana Conversion for Japanese Language Education. *Proceedings of the Workshop on Text Input Methods*, (pages 38-42). Chiang Mai, Thailand.

Liu, X., Cheng, F., Luo, Y., Duh, K., & Matsumoto, Y. (2013). A Hybrid Chinese Spelling Correction Using Language Model and Statistical Machine Translation with Reranking. *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7)*, (pages 54–58), Nagoya, Japan.

Mizumoto, T., Hayashibe, Y., Komachi, M., Nagata, M., & Matsumoto, Y. (2012). The Effect of Learner Corpus Size in Grammatical Error Correction of ESL Writings. *Proceedings of the International Conference on Computational Linguistics*, (pages 863-872). Mumbai, India.

Mizumoto, T., Komachi, M., Nagata, M., & Matsumoto, Y. (2011). Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. *Proceedings of the International Joint Conference on Natural Language Processing*, (pages 147-155). Chiang Mai, Thailand.

Neubig, G. (2011). The Kyoto Free Translation Task. http://www.phontron.com/kftt.

Och, J. F. (2003). Minimum Error Rate Training in Statistical Machine Translation. *Proceedings of the Annual Meeting of the ACL*, (pages 160-167). Sapporo, Japan.

Oyama, H., Komachi, M., & Matsumoto, Y. (2013). Towards Automatic Error Type Classification of Japanese Language Learners' Writing. *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation*, (pages 163-172). Taipei, Taiwan.

Papineni, K., Roukos S., Ward, T., & Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the Annual Meeting of the ACL*, (pages 311-318). Philadelphia, USA.

Park, Y. A., & Levy, R. (2011). Automated Whole Sentence Grammar Correction Using a Noisy Channel Model. *Proceedings of the Annual Meeting of the ACL*, (pages 934-944). Ohio, USA.

Swanson, B., & Yamangil, E. (2012). Correction Detection and Error Type Selection as an ESL Educational Aid. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pages 357-361). Montreal, Canada.

Tajiri, T., Komachi, M., & Matsumoto, Y. (2012). Tense and Aspect Error Correction for ESL Learners Using Global Context. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, (pages 198-202). Jeju Island, Korea.

Wu, J.-C., Chiu, H.-W., & Chang, J. S. (2013). Integrating Dictionary and Web N-grams for Chinese Spell Checking. *Computational Linguistics and Chinese Language Processing*, 18(4), (pages 17-30).

Wu, S.-H., Liu, C.-L., & Lee, L.-H. (2013). Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013. *Proceedings of the SIGHAN Workshop on Chinese Language Processing*, (pages 35-42). Nagoya, Japan.

Yuan, Z. & Felice, M. (2013). Constrained Grammatical Error Correction Using Statistical Machine Translation. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, (pages 52–61). Sofia, Bulgaria.