Improving Chinese Grammatical Error Correction using Corpus Augmentation and Hierarchical Phrase-based Statistical Machine Translation

Yinchen Zhao **Mamoru Komachi** **Hiroshi Ishikawa**
Graduate School of System Design, Tokyo Metropolitan University, Japan
chou.innchenn@gmail.com
komachi@tmu.ac.jp
ishikawa-hiroshi@tmu.ac.jp

Abstract

In this study, we describe our system submitted to the 2nd Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA-2) shared task on Chinese grammatical error diagnosis (CGED). We use a statistical machine translation method already applied to several similar tasks (Brockett et al., 2006; Chiu et al., 2013; Zhao et al., 2014). In this research, we examine corpus-augmentation and explore alternative translation models including syntax-based and hierarchical phrase-based models. Finally, we show variations using different combinations of these factors.

1 Introduction

The concept of “translating” an error sentence into a correct one was first researched by Brockett et al. (2006). They proposed a statistical machine translation (SMT) system with noisy channel model to correct automatically erroneous sentences for learners of English as a Second Language (ESL).

It seems that a statistical machine translation toolkit has become increasingly popular for grammatical error correction. In the CoNLL-2014 shared task on English grammatical error correction (Ng et al., 2014), four teams of 13 participants each used a phrase-based SMT system. Grammatical error correction using a phrase-based SMT system can be improved by tuning using evaluation metrics such as $F_{0.5}$ (Kunchukuttan et al., 2014; Wang et al., 2014) or even a combination of different tuning algorithms (Junczys-Dowmunt and Grundkiewicz, 2014). In addition, SMT can be merged with other methods. For example, the language model-based and rule-based methods can be integrated into a single sophisticated but effective system (Felice et al., 2014).

For Chinese, SMT has also been used to correct spelling errors (Chiu et al., 2013). Furthermore, as is shown in NLP-TEA-1, an SMT system can be applied to Chinese grammatical error correction if we can employ a large-scale learner corpus (Zhao et al., 2014).

In this study, we extend our previous system (Zhao et al., 2014) to the NLP-TEA-2 shared task on Chinese grammatical error diagnosis, which is based on SMT. The main contribution of this study is as follows:

- We investigate the hierarchical phrase-based model (Chiang et al., 2005) and determine that it yields higher recall and thus F score than does the phrase-based model, but is less accurate.
- We increase our Chinese learner corpus by web scraping (Yu et al., 2012; Cheng et al., 2014) and show that the greater the size of the learner corpus, the better the performance.
- We perform minimum error-rate training (Och, 2003) using several evaluation metrics and demonstrate that tuning improves the final F score.

2 Hierarchical phrase-based model

A hierarchical phase-based model for SMT was first suggested by Chiang et al. (2005). The system first achieves proper word alignment, and instead of extracting phrase alignment, the sys-

Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 111–116,
Beijing, China, July 31, 2015. ©2015 Association for Computational Linguistics and Asian Federation of Natural Language Processing

tem extracts rules in the form of synchronous context-free grammar (SCFG) rules. In a Chinese error correction task, such error-correction rules are extracted as follows:

X → (X_1 一 好消息 X_2,　X_1 一个 好消息 X_2)
(a piece of good news)
X → (我 有,　我 有)
(I have)
X → (告诉 你,　告诉 你)
(to tell you)

The symbols X and X_i here are non-terminal and represent all possible phrases. In addition, glue rules are used to combine a sequence of Xs to form an S.

The glue rules are given as:

S → (X_1, X_1)
S → (S_1X_2, S_1X_2)

A complete derivation of this simple example can then be written:

S → (X_1, X_2)
→ (X_3 一 好消息 X_4,　X_3 一个 好消息 X_4)
→ (我 有 一 好消息 X_4,　我 有 一个 好消息 X_4)
→ (我 有 一 好消息 告诉 你,　我 有 一个 好消息 告诉 你)
(I have a piece of good news to tell you)

To determine a weight of a derivation, this model utilizes features such as generation probability, lexical weights, and phrase penalty. In addition, to avoid too many distinct yet similar translations, rules are constrained by certain filters that, for example, limit the length of the initial phrase the number of non-terminals per rule.

3 Chinese Learner Corpora

3.1 Lang-8 Learner Corpus

The Lang-8 Chinese Learner Corpus was built by extracting error-correct sentence pairs from the Internet (Mizumoto et al., 2011; Zhao et al., 2014). We use it as a training corpus for our SMT-based grammatical error diagnosis system in NLP-TEA-1.

However, after we analyzed edit distance (ED) between error-correct sentence pairs based on word level, we determined it may not be suitable for training our translation model. As Figugre 1 shows, NLP-TEA-2 training data has ED mostly from 1 to 3 whereas Lang-8 Chinese Corpus has many ED longer than 4.

This is reasonable because the NLP-TEA-2 training data are extracted from essays written by high-level Chinese learners and, in most cases, these learners produce only one- or two-word-mistakes. By contrast, Lang-8 is a language exchange social networking website where sentences are written by language learners of any level. If we use this corpus as it currently exists, sentences having too long ED may confuse the SMT system.

Therefore, we cleaned the Lang-8 Chinese Learner Corpus by randomly sampling sentence pairs whose ED is between 4 and 8 and deleting sentences pairs whose ED is longer than 8. This ensures it has a similar ED distribution to that of the NLP-TEA-2 training data. After cleaning, the number of sentences in the corpus decreased from 95,000 to approximately 58,000.

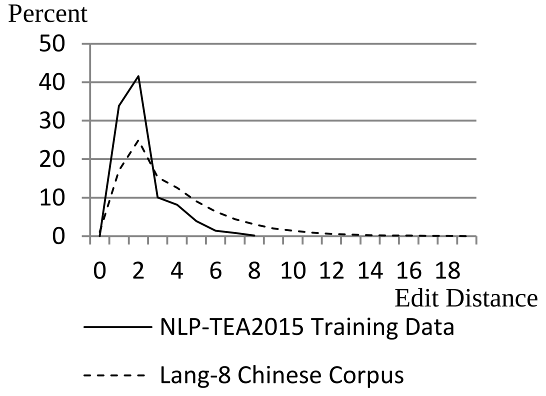

Figure 1: Distribution of ED in different data sets. The distribution of ED in the Lang-8 Chinese Learner Corpus shown here is prior to cleaning.

3.2 HSK Dynamic Essay Corpus

In this shared task, we augment the Chinese learner corpus with another learner corpus extracted from the Internet (Yu et al., 2012; Cheng et al., 2014). The HSK Dynamic Essay Corpus[1] is one such corpus built by Beijing Language and Culture University. In this corpus, approximately 11,000 essays are collected from HSK Chinese tests taken by foreign Chinese language learners, and error sentences are annotated with special marks.

For example:

这就{CQ 要}由有关部门和政策管理制度来控制。

[1] http://nlp.blcu.edu.cn/online-systems/hsk-language-lib-indexing-system.html

where {CQ 要} refers to a redundant word and is revised with the word that follows it.

可是这两个问题同时{CJX}要解决非常不容易。where {CJX} refers to a reordering error.

However, detaching an erroneous sentence and a corresponded correction sentence from an annotated one as above is not easy because we don't know the position information of the reordering error. Moreover, such detachment is also difficult when dealing with some more complex errors, for example, a "ba (把)" error (a special preference of active voice in Chinese) or "bei (被)" error (a special preference of passive voice in Chinese), if we depend only on such marks.

Thus, we extracted sentences having only insertion, deletion, or replacement errors. We also cleaned the HSK corpus by deleting sentences pairs having too long ED as described. As a result, the corpus now contains approximately 59,000 sentences. The distribution of ED in the combined corpus is shown in Figure **2**.

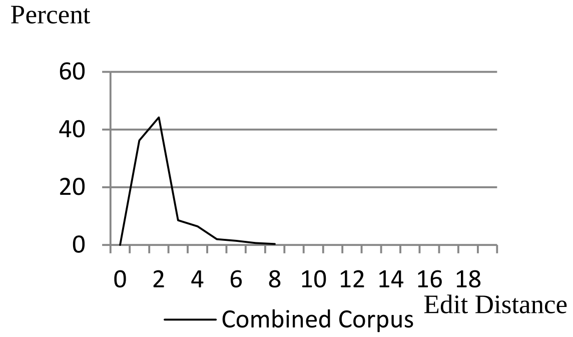

Figure 2: Distribution of ED in combined corpus.

4 Tuning

As previously described, an SMT system with tuning is proved to perform better than one without tuning. Because this shared task uses several evaluation metrics such as accuracy, F1 score, and FP rate, we tune our system using all these metrics with minimum error rate training (MERT) (Och, 2003) at identification level[1]. Our linear evaluation score is computed according to the following:

Score = α *Accuracy+ β * $F_{0.5}$+ γ *(1-FP_rate) where $\alpha + \beta + \gamma = 1.0$.

We conducted a series of preliminary experiments to discover the most effective set of parameters. We followed Kunchukuttan et al. (2014) and Wang et al. (2014) in using $F_{0.5}$ instead of F1. In other words, we expected our system to have high accuracy because, as Ng et al. say in CoNLL-2014, "it is important for a grammar checker that its proposed corrections are highly accurate in order to gain user acceptance." However, we discovered that even when we used a parameter set of α =0.0, β =1.0, and γ =0.0, we still failed to reach a satisfactory correction rate.

Finally, we use α =0.5, β =0.0, and γ =0.5 as a final parameter set for phrase-based and hierarchical phrase-based systems because it produces the greatest number of corrections at identical level among our in-house experiments. In addition, our in-house experiments revealed that an improper parameter set could produce a reasonable but unacceptable result. We discuss this aspect with reference to an experiment regarding a syntax-based system in the next section.

5 Experiment and Results

5.1 Official Runs

We followed the WAT2015[2] baseline system to build phrase-based and hierarchical phrase-based SMT systems. This involves segmenting words using Stanford Word Segmenter version 2014-01-04, running GIZA++ v1.07 on training corpus in both directions, and parsing Chinese sentences with Berkeley parser (for java 1.7). We ran Moses v2.11 for decoding using the same parameters with the WAT2015 baseline. We trained two hierarchical phrase-based systems using different sized corpora according to whether the HSK corpus is included. For error classification, we followed Zhao et al. (2014) to identify error types and locate the positions of errors.

All three runs we submitted are shown in Table 1. In addition, the results of our runs at position level are shown in Table 2. RUN3 produced more corrections and obtained a higher F1 score at position level than did the other runs. However,

[1] Detection level: All error types will be regarded as incorrect. Identification level: All error types should be clearly identified, i.e., Redundant, Missing, Disorder, and Selection. Position level: The system results should be perfectly identical with the quadruples of gold standard.
We tried to tune in position level but we omit these results since this attempt mostly failed.

[2] http://orchid.kuee.kyoto-u.ac.jp/WAT/

it is inferior in terms of accuracy and FP rate compared to RUN2.

At position level, the phrase-based system generated only 15 correct predictions and among them only one Disorder and no Selection types appeared. By contrast, the hierarchical system performed much better, as it successfully predicted seven Disorder and five Selection types. In addition, it produced more correct predictions on Missing and Redundant types.

TMU-RUN1	Lang-8 + hierarchical
TMU-RUN2	Lang-8 + HSK + phrase-based
TMU-RUN3	Lang-8 + HSK + hierarchical

Table 1: Three RUNs submitted by TMU (Tokyo Metropolitan University) team.

	FP rate	Accuracy	Precision	Recall	F1
RUN1	0.478	0.270	0.0363	0.0180	0.0241
RUN2	0.134	0.449	0.1928	0.0320	0.0549
RUN3	0.350	0.362	0.1745	0.07400	0.1039

Table 2: Final test result of TMU RUNs at position level.

5.2 Hierarchical Phrase-based Model

We provide an example of the official test set to explain why hierarchical phrase-based systems appear to be more effective than those that are phrase-based. The following Chinese sentence is used:

B1-1033: 其中有一个人丢护照了。

(One of them lost his passport.)

In a hierarchical-phrase-based system and according to the synchronous CFG rule, the partial derivation of the phrase "丢 护照 了 (lost his passport)" is:

(X, X)→(丢 X_1 ， 丢 X_1)

→(丢 X_2 了, 丢 了 X_2)

→(丢 护照 了, 丢 了 护照)

where X denotes any phrase. Because "X 了" wrongly written as "了 X" is a typical Disorder error in Chinese sentences, the hierarchical phrase-based system extracts the rule X→(X 了, 了 X) and weighs it highly when training on the corpus. This means the model actually examined syntax errors in sentences. By contrast, the phrase-based system lacks the ability to identify syntax errors. Therefore, this translation model is less effective than the hierarchical phrase-based system, as it failed to select a correct translation such as "丢 了 X."

5.3 Corpus Augmentation

According to the results shown in Table 4, expanding the corpus has a beneficial effect. In RUN1, the F1 score of 0.024 means it nearly failed to produce any correction prediction. However, after we increased the corpus size, the F1 score increased to 0.10. The improved F1 score with corpus augmentation is illustrated in Figure 3. Among F1 scores, our RUN3 ranks exactly in the middle of 15 RUNS of all teams.

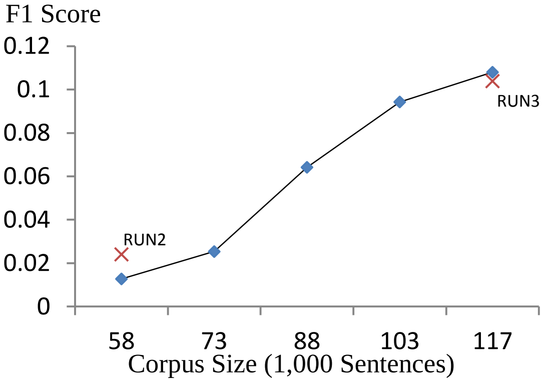

Figure 3: F1 score improved with corpus augmentation. The solid line represents results of our in-house test. The Xs represent results of this open task.

5.4 Tuning

To determine the effect of tuning for improving the two systems, we developed a test on the NLP-TEA-1 training set offered by organizers. Table 3 shows a contrast between tuned and untuned systems. As with the English grammatical error correction task, MERT clearly boosts the F1 score in this task. We tuned the system using the Z-MERT toolkit (Zaidan, 2009).

	F1 Score	
	Phrase-based	Hierarchical-phrase-based
Untuned	0.0513	0.0868
Tuned	0.0701	0.1080

Table 3: F1 score of SMT-based grammatical error correction system on NLP-TEA-1 dataset, with and without tuning.

To compare different syntax-based systems, we also developed a string-to-tree (s2t) SMT system. However, in our attempt to tune it, we failed to obtain a best set of parameters. We first tried a parameter set of (0.5, 0.0, 0.5), which performs most effectively with the phrase-based model. However, it failed to improve the F1 score, as is shown in Table 4.

	FP_Rate	Accuracy	Precision	Recall	F1
Untuned	0.3973	0.4087	0.1042	0.0787	0.0896
Tuned	0.1029	0.4747	0.0480	0.0057	0.0102

Table 4: Tuning result suitable to an evaluation score but unacceptable for its low precision and recall.

The system is clearly optimized to achieve the best performance in terms of FP rate and accuracy. However, this is because, as experiments showed, the system produces nearly all negative predictions, which causes low precision and recall, as increasing true negatives improves both the accuracy and FP rate. We determined that α =0.5, β =0.0, γ =0.5 may not be a “good” parameter set in this situation, even though it seemed acceptable for a preliminary experiment. Unfortunately, we did not identify any parameter sets that can generate more acceptable results than can the s2t system without tuning.

6 Conclusion

We have described a Chinese grammatical error correction system based on SMT for the TMU-NLP team. First, we examined hierarchical phrase-based and string-to-tree translation models of SMT on CGED. Second, we constructed an error-correction parallel corpus based on the HSK Dynamic Essay Corpus, which is nearly equal in size to the Lang-8 Chinese Learner Corpus. We then cleaned and combined the two into a single expanded corpus. Third, we tuned the system with a linear combination of evaluation metrics using MERT. Finally, we showed that the augmented corpus considerably improved performance. In addition, the hierarchical phrase-based translation model generated a higher F1 score than did the phrase-based model.

For future research, we will attempt to expand the corpus further. A possible direction in building a large-scale parallel corpus is to introduce errors artificially to correct sentences. This has already been applied in an English error correction task of Yuan and Felice (2013). In addition, we confirmed that our system produces correct predictions in generated N-best output. However, oracle predictions were not selected during decoding. To solve this, we will employ a much more powerful language model such as the Google n-gram model as well as a re-ranking approach on the N-best output.

Acknowledgments

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

Reference

Chris Brockett, William Dolan, Michael Gamon. (2006). Correcting ESL Errors Using Phrasal SMT Techniques. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 249–256. Sydney, Australia.

Shuk-Man Cheng, Chi-Hsin Yu, Hsin-Hsi Chen. (2014). Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Language Learners. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*: Technical Papers, pages 279–289, Dublin, Ireland.

David Chiang. (2005). A Hierarchical Phrase-Based Model for Statistical Machine Translation. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pages 263–270, Ann Arbor, Michigan.

Hsun-wen Chiu, Jian-cheng Wu, Jason S. Chang. (2013). Chinese Spelling Checker Based on Statistical Machine Translation. *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7)*, pages 49–53, Nagoya, Japan.

Mariano Felice, Zheng Yuan, Øistein E. Andersen, Helen Yannakoudakis, Ekaterina Kochmar. (2014). Grammatical Error Correction using Hybrid Systems and Type Filtering. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 15–24, Baltimore, Maryland.

Marcin Junczys-Dowmunt, Roman Grundkiewicz. (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*, pages 25–33, Baltimore, Maryland.

Anoop Kunchukuttan, Sriram Chaudhury, Pushpak Bhattacharyya. (2014). Tuning a Grammar Correction System for Increased Precision. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 60–64, Baltimore, Maryland.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, Yuji Matsumoto. (2011). Mining Revision Logs of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 148–155, Chiang Mai, Thailand.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond H. Susanto, Christopher Bryant. (2014). The CoNLL-2014 Shared Task on Grammatical Error Correction. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland.

Franz J. Och. (2003). Minimum Error Rate Training in Statistical Machine Translation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167. Sapporo, Japan.

Yiming Wang, Longyue Wang, Derek F. Wong, Lidia S. Chao, Xiaodong Zeng, Yi Lu. (2014). Factored Statistical Machine Translation for Grammatical Error Correction. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland.

Chi-Hsin Yu, Hsin-Hsi Chen. (2012). Detecting Word Ordering Errors in Chinese Sentences for Learning Chinese as a Foreign Language. *Proceedings of COLING 2012: Technical Papers*, pages 3003–3018, Mumbai, India.

Zheng Yuan, Mariano Felice. (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.

Omar F. Zaidan. (2009). Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems. *The Prague Bulletin of Mathematical Linguistics*, pages 79–88

Yinchen Zhao, Mamoru Komachi, Hiroshi Ishikawa. (2014). Extracting a Chinese Learner Corpus from the Web: Grammatical Error Correction for Learning Chinese as a Foreign Language with Statistical Machine Translation. *Proceedings of the 22nd International Conference on Computers in Education*, pages 56–61, Nara, Japan.