A Chinese Writing Correction System for Learning Chinese as a Foreign Language

Yow-Ting Shiue[1], Hen-Hsen Huang[1], and Hsin-Hsi Chen[1,2]

[1]Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
orina1123@gmail.com,hhhuang@nlg.csie.ntu.edu.tw,hhchen@ntu.edu.tw

Abstract

We present a Chinese writing correction system for learning Chinese as a foreign language. The system takes a wrong input sentence and generates several correction suggestions. It also retrieves example Chinese sentences with English translations, helping users understand the correct usages of certain grammar patterns. This is the first available Chinese writing error correction system based on the neural machine translation framework. We discuss several design choices and show empirical results to support our decisions.

Title and Abstract in Chinese

爲非母語中文學習者設計的中文寫作更正系統

我們建立了一個爲非母語中文學習者設計的中文寫作更正系統，輸入一個錯誤的句子，此系統可以產生數個建議更正，並查詢附有英文翻譯的相關例句，幫助使用者理解某些文法的正確用法。這是第一個基於神經網路機器翻譯框架的中文寫作錯誤更正系統，在此篇論文中我們討論幾個設計上的選擇，呈現幫助我們做決定的實驗數據。

1 Introduction

Grammatical error correction (GEC) helps users check and correct mistakes in their writing. English GEC has been incorporated in commercial software; in contrast, there is far fewer readily usable writing correction tools for Chinese. Chinese has become a popular foreign language to learn worldwide, motivating the development of Chinese writing correction system targeting second language (L2) learners.

Unlike the classification approach, the translation approach to English GEC does not require exact recognition of error types. With many-to-many mappings handled, it is possible to deal with multiple errors of various types with a single translation model. An open-source statistical machine translation (SMT)-based English GEC system is released by Chollampatt and Ng (2017). More recently, neural machine translation (NMT) is applied to English GEC and improvements over the SMT baseline are shown (Yuan and Briscoe, 2016). With the use of distributional word representations, NMT has better ability to generalize to unseen corrections.

The Shared Task for Chinese Grammatical Error Diagnosis (CGED) (Rao et al., 2017) only evaluates detection but not correction performance until 2017. Some studies focus on certain error types of L2 Chinese, such as word ordering errors (Cheng et al., 2014) and word usage errors (Shiue and Chen, 2016; Shiue et al., 2017). Huang et al. (2016) correct preposition errors. Nevertheless, there has not yet been a general model that handles all types of Chinese writing errors.

Given the promising results of translation approaches in English, it is worth investigating their effectiveness in Chinese. Because the machine translation models need to be trained with parallel corpus of wrong-corrected sentences and there is limited amount of Chinese learner data with annotated corrections, we use NMT models and facilitate them with word embeddings pre-trained on large amount of well-formed Chinese text. To our knowledge, we are the first to apply NMT to Chinese error correction.

Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 137–141
Santa Fe, New Mexico, USA, August 20-26, 2018.

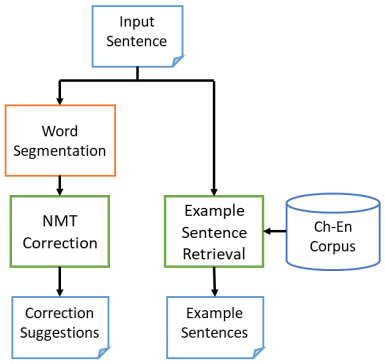

Figure 1: Architecture of our Chinese writing error correction system.

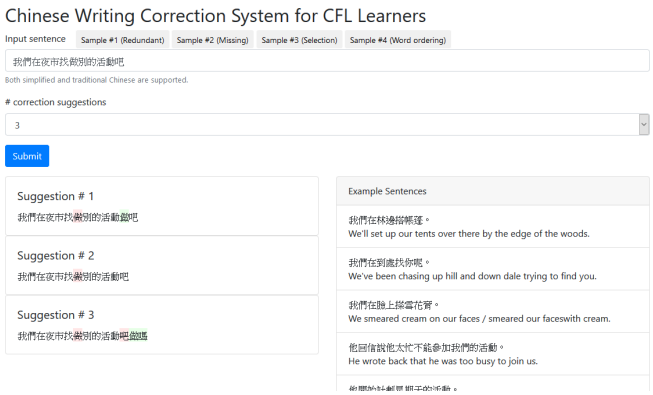

Figure 2: Web-based demonstration of our Chinese error correction system.

To improve writing proficiency, language learners need to know not only how the sentences they wrote are incorrect, but also how to correctly express their intended meanings. Therefore, in addition to correction suggestions, our system provides example sentences related to the input with appropriate level of difficulty. Figure 1 illustrates the overall architecture of our system. The two main components, NMT Correction and Example Sentence Retrieval, will be elaborated in Sections 2 and 3 respectively. All user inputs and system outputs are logged. These records can be utilized to analyze common learner error patterns, and additional training data can be annotated to incrementally improve the system performance. A web-based demonstration of our system is available at `http://nlg6.csie.ntu.edu.tw/CGED-NMT-demo` and a screenshot is shown in Figure 2.

2 Correction with Neural Machine Translation

We treat the error correction task as a translation task from erroneous Chinese to well-formed Chinese. This idea has been widely adopted for English GEC, but we are the first to apply it to the correction of Chinese. A typical NMT model is composed of an encoder and a decoder. The encoder transforms the input sequence into a sequence of hidden states, each of which is calculated with the hidden state of the previous time step and the input of the current time step. The decoder predicts the distribution of words for each time step conditioned on the encoder hidden states and the output of all previous time steps. The encoder-decoder network is trained to maximize the likelihood of the ground-truth translations in the training data. Our system is built on the top of OpenNMT (Klein et al., 2017). We adopt a bidirectional Long-short Term Memory (LSTM) encoder and a two-layer LSTM decoder. Global attention over the sequence of hidden states at the source side is applied. The model generates one to five corrections according to the n-best decoding result. Several design choices will be discussed in Section 2.2.

2.1 Datasets and Evaluation

To train the NMT correction model, we utilize the publicly available datasets of the NLPTEA 14-17 CGED shared tasks[1]. As a whole, there are more simplified Chinese sentences than traditional Chinese ones, so we convert all sentences to simplified Chinese. Each sentence can be completely correct (no correction is needed), or contain one or more errors. The errors are categorized into redundant word, missing word, word selection, and word ordering. However, we build a general correction framework for all types of errors and do not use or predict error type labels. We use the test data of NLPTEA 14 (1,783 sentence) and 15 (1,000 sentences) for validation and testing respectively, and the training data of NLPTEA 14-17 (totally 38,554 sentences) for training. We do not use the test data of NLPTEA 16 and 17 since there are only error type labels but no correction in the datasets.

The correction performance can be evaluated by judging whether a correction is exactly the same as the ground-truth. We report the accuracy as well as hit rates of top candidates. However, hit rates can still be

[1]https://sites.google.com/view/nlptea2018/shared-task

somehow strict since a model will not get any scores even if the top candidate it proposes is only slightly different from the answer. Thus, we also report the General Language Evaluation Understanding (GLEU) metric (Napoles et al., 2015), which is a modification of BLEU that rewards correct modifications while penalizing unnecessary changes. We use the publicly released toolkit[2] to calculate GLEU of n-gram order 4. GLEU is calculated only for the top candidate.

2.2 Design Choices

There are several design choices for building the NMT-based correction system. We discuss the reasons for each decision and show experimental results when necessary. In the experiments, we choose the model with the highest validation GLEU and report the performance on the test set. The GLEU of an output that is completely the same as the source can be regarded as a baseline.

Character-based vs. Word-based Models

Although a word is a more meaningful semantic unit, word-based models might suffer from noise induced by segmentation errors, which might occur more frequently in learners' text than in normal well-formed text. On the other hand, character-based models need to handle longer dependencies. We make the fundamental design decision of treating an input sentence as a sequence of characters or a sequence of words based on empirical results. For word segmentation, we use THU Lexical Analyzer for Chinese (THULAC) (Sun et al., 2016)[3], which results in the best correction performance among several Chinese word segmentation tools.

The performance of the two kinds of models is shown in Table 1. We report character-level GLEUs in order to make the metric values of the two models comparable. As can be seen, the word-based model outperforms the character-based model in all evaluation metrics. A possible reason is that the decoder is trained to output well-formed sentences. Though segmentation errors might affect the understanding of the source sentence, the decoder is still possible to "complete" the output sentence based on partial source information. For example, the erroneous sentence "* 我 覺得 他 是 一個 很好人" (* *I think he is a very good-person*) is corrected to "我 覺得 他 是 一個 很好 的 人" (*I think he is a very good person*). Based on these results, we decide to use the word-based NMT model in our system.

Pre-trained Word Embeddings

Initializing word representations in NMT models with pre-trained word vectors can be useful when the training data is insufficient. In addition to the standard Word2vec continuous bag-of-words (CBOW) and Skip-gram (SG) embeddings (Mikolov et al., 2013), we also experiment with the continuous window (CWIN) and structured skip-gram (Struct-SG) embeddings (Ling et al., 2015), which consider the relative order of context words during training and are shown to be useful for Chinese error detection (Shiue et al., 2017). We segment the Chinese part of ClueWeb[4] with the THULAC toolkit and train the embeddings with it. The embedding size is fixed to 500 and the context window size is 5 for all kinds of embeddings. The results are summarized in Table 1. All pre-trained word embeddings bring improvement over random embeddings. Generally, the NMT correction model with pre-trained Struct-SG embeddings achieves the best performance. Thus, we use Struct-SG embeddings in our final system.

Model	Features	Accuracy	Hit@3	Hit@5	char. GLEU	word GLEU
(Baseline)	-	-	-	-	0.552	0.411
Character-based	Rand. emb.	0.145	0.293	0.341	0.625	-
Word-based	Rand. emb.	0.190	0.327	0.376	0.650	0.558
Word-based	CBOW	0.210	0.368	0.418	0.655	0.564
Word-based	SG	0.194	0.369	0.414	0.657	0.564
Word-based	CWIN	0.214	0.379	**0.433**	0.658	0.566
Word-based	Struct-SG	**0.232**	**0.387**	0.431	**0.668**	**0.580**

Table 1: Performance of NMT-based correction models

[2]https://github.com/cnap/gec-ranking
[3]http://thulac.thunlp.org/
[4]http://lemurproject.org/clueweb09.php

3 Example Sentence Retrieval

Besides giving correction suggestions, our system also shows example sentences to demonstrate how to correctly use the words and grammar patterns in the user input. These example sentences also serve as additional evidence of the correctness of some usage patterns.We adopt UM-Corpus (Tian et al., 2014), a sentence-aligned English-Chinese corpus, as the database of example sentences. We only use sentences in the "Education" domain, which are extracted from online teaching materials. There are 450,000 English-Chinese sentence pairs. We exclude example sentence pairs in which the Chinese sentence is longer than 30 Chinese characters since they usually have complex syntactic structures.

Upon user input, ten example sentences are retrieved. They are ranked by the overlaps of Chinese character bigrams. The more character bigrams an example sentence has in common with the input sentence, the higher score it gets. The score is normalized by the total number of character bigrams. Although more recent retrieval models, such as those based on word embeddings, can handle semantic similarities that are not reflected in the surface form, there is another level of difficulties for foreign language learners to recognize this kind of similarities. Therefore, bigram matching may help to focus on the words and grammar patterns being used in the input sentence.

An example input sentence and the top 3 retrieved example sentences are shown below. As can be seen, the sentences where the phrase "每個月" (*every month*) is used are selected.

Input: ＊在泰國**每個月**天氣都熱 (*In Thailand, the weather is hot every month.*)

Example sentences:

過去十年她**每個月**都在存錢。	*She had been saving money every month for the last ten years.*
你**每個月**的食宿費用是多少?	*How much do you charge a month for room and board?*
每個月25元的月租就是白送錢。	*The monthly rent of 25 yuan per month is white money.*

4 Conclusions

We build a writing correction system for learning Chinese as a foreign language. The system not only provides corrections, but also presents example sentences with English translation, illustrating how to correctly use the words and grammar patterns related to the input sentence. The correction is performed with an NMT model enhanced by pre-trained word representations. On the test set of the NLPTEA 15 CGED shared task, the model achieves GLEU 0.67 and 0.58 at the character and the word levels, respectively. Further research can be conducted on top of our framework, and the web interface can facilitate user evaluation of different back-end models.

Acknowledgements

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-106-2923-E-002-012-MY3, MOST-107-2634-F-002-011- and MOST-107-2634-F-002-019-.

References

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

Shamil Chollampatt and Hwee Tou Ng. 2017. Connecting the Dots: Towards Human-Level Grammatical Error Correction. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 327–333. Association for Computational Linguistics.

Hen-Hsen Huang, Yen-Chi Shao, and Hsin-Hsi Chen. 2016. Chinese Preposition Selection for Grammatical Error Diagnosis. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 888–899. The COLING 2016 Organizing Committee.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Wang Ling, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Two/Too Simple Adaptations of Word2Vec for Syntax Problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground Truth for Grammatical Error Correction Metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593. Association for Computational Linguistics.

Gaoqi Rao, Baolin Zhang, Endong Xun, and Lung-Hao Lee. 2017. IJCNLP-2017 Task 1: Chinese Grammatical Error Diagnosis. In *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 1–8. Asian Federation of Natural Language Processing.

Yow-Ting Shiue and Hsin-Hsi Chen. 2016. Detecting Word Usage Errors in Chinese Sentences for Learning Chinese as a Foreign Language. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Yow-Ting Shiue, Hen-Hsen Huang, and Hsin-Hsi Chen. 2017. Detection of Chinese Word Usage Errors for Non-Native Chinese Learners with Bidirectional LSTM. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 404–410. Association for Computational Linguistics.

Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. THULAC: an efficient lexical analyzer for Chinese. Technical report.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, and Lu Yi. 2014. UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association (ELRA).

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386. Association for Computational Linguistics.