The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

A New Benchmark and Evaluation Schema for Chinese Typo Detection and Correction*

Dingmin Wang,[‡⋆] Gabriel Pui Cheong Fung,[‡†]
Maxime Debosschere,[‡†] Shichao Dong,[‡] Jia Zhu,[∓] Kam-Fai Wong[‡†]
[‡]The Chinese University of Hong Kong, Hong Kong, China
[†]MoE Key Laboratory of High Confidence Software Technologies (Sub-Laboratory, CUHK), Hong Kong, China
[⋆]Tsinghua University, Beijing, China [∓]South China Normal University, Guangzhou, China
{dmwang, pcfung, dmaxime, scdong, kfwong}@se.cuhk.edu.hk, jzhu@m.scnu.edu.cn

Abstract

Despite the vast amount of research related to Chinese typo detection, we still lack a publicly available benchmark dataset for evaluation. Furthermore, no precise evaluation schema for Chinese typo detection has been defined. In response to these problems: (1) we release a benchmark dataset to assist research on Chinese typo correction; (2) we present an evaluation schema which was adopted in our NLPTEA 2017 Shared Task on Chinese Spelling Check; and (3) we report new improvements to our Chinese typo detection system ACT.

1 Introduction

Automatic typo detection is an important prerequisite step for many Natural Language Processing (NLP) applications to better understand the underlying semantics of sentences. Error detection applications for Chinese are not yet well developed, unlike applications for alphabet-based languages such as English and French. This is partly due to a lack of benchmark data, and due to the absence of a precise evaluation schema. Our contributions described in this paper are:

- Constructed and released a benchmark dataset for Chinese typo detection;
- Proposed a new schema for evaluating the performance of a typo detection system;
- Improved a system for automatically detecting typos in Chinese, which is an extension of our previous work (Dong et al. 2016).

2 Benchmark Dataset

The Hong Kong Applied Science and Technology Research Institute first collected more than 5,000 writings by Hong Kong primary students. We then invited researchers from the Department of Chinese Language and Literature at CUHK to help us mark and annotate these writings. Next, we selected a total of 6,890 sentences of a reasonable length (50–150 characters, including punctuation) which contain at least one error. The average number of errors in a sentence is 2.7, and the maximum is 5. Since our benchmark dataset also requires positive examples, we manually added 3,110 entirely correct sentences for a round total of 10,000.

Our benchmark dataset contains the following types of errors: (1) Typo – Similar shape (e.g., in the word 辨論, 辨 is a typo and should be written as 辯. 辨 and 辯 have similar shapes); (2) Typo – Similar pronunciation (e.g., in the word 合諧, 合 is a typo and should be replaced by 和. 合 and 和 have similar pronunciations in Cantonese); (3) Colloquialism – Incorrect character (e.g., the character 佢 is colloquial and should be changed into 他); (4) Colloquialism – Incorrect phrase (e.g., the word 撞返 is colloquial even though the characters 撞 and 返 both are not. Here, 撞返 should be replaced by 碰見); (5) Incorrect word ordering (no characters or phrases are colloquial, but the ordering of some characters or words results in colloquial language. E.g., in the sentence "我走先了" the word 走先 is colloquial and should be written as 先走); (6) Mixing Simplified Chinese and Traditional Chinese (e.g., for the word 词語, 词 is simplified Chinese and should be replaced by its traditional counterpart 詞); (7) Errors in poems and idioms (e.g., in the idiom 天生我才必有用, 才 should be replaced by 材).

口语化的

Note that it is possible to have any mixture of the above cases. For example, consider the sentence 大家討論緊這件事. In this context the character 緊 is a colloquial word which means 正在 (error type 3). Yet simply replacing 緊 with 正在 is still wrong since it then triggers error type 4. Instead, the correction should be 大家正在討論這件事.

To the best of our knowledge, there is no publicly available benchmark dataset that takes into account error types 3 to 7. We are the first to release such dataset, and it can be obtained from the CUHK MOE lab website.[1]

For the NLPTEA 2017 Shared Task on Chinese Spelling Check (Fung et al. 2017), we assembled two sets of 1,000 randomly selected sentences from the benchmark dataset. Each of these two sets then had a corresponding *gold standard*: the best solution that any spell checking system can possibly give. The gold standard includes as many valid corrections as possible for each error. To the best of our knowledge, these are the first datasets with this property.

*This work was partially supported by the MoE Key Laboratory of High Confidence Software Technologies (Sub-Laboratory, CUHK), NSFC (61772211, 61750110516), S&T Projects of Guangdong Province (2016A030303055, 2016B030305004, 2016B010109008) and Guangzhou S&T Project (201604046017).

[1]www1.se.cuhk.edu.hk/~moelab/act/act-dataset.zip

3 Evaluation Schema

The evaluation schema described in this section has been implemented in our NLPTEA 2017 Shared Task on Chinese Spelling Check (Fung et al. 2017).

3.1 Evaluating Detection Performance

We apply the standard approach of combining *precision* and *recall* to measure typo detection performance. Mathematically,

$$precision = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN}$$

$$P_{detection} = \frac{2 \times precision \times recall}{precision + recall}$$

where TP is the number of characters that are correctly identified as errors; FP is the number of characters that are incorrectly identified as errors; and FN is the number of errors that remained undetected.

3.2 Evaluating Correction Performance

Spelling checkers in modern word processing software usually provide multiple possible corrections for any given error, which maximises editing flexibility. In order to allow for this, we included as many valid corrections as possible in our gold standard.

A correction in the gold standard is considered *successfully detected* when a system provided a correction suggestion for the same position. For every successfully detected error, a system is expected to deliver one or more appropriate correction suggestions. However, in order to avoid the case where a system provides long lists of correction suggestions in order to cover all gold standard corrections, a penalty proportional to the number of invalid provided suggestions is imposed. Mathematically,

$$P_{correction} = \frac{1}{|W|} \sum_{\forall i \in W} \frac{|G_i \cap U_i|}{|U_i|}$$

where W is the set containing all correctly detected errors; G_i is the set containing the gold standard suggestions for error $i \in W$; and U_i is the set containing the system correction suggestions for error $i \in W$. For G_i and U_i, major cases are:

- $G_i \cap U_i = G_i = U_i$: all system suggestions are in the gold standard corrections, and vice versa;
- $G_i \cap U_i = \emptyset$: no system suggestions are in the gold standard corrections;
- $G_i \cap U_i = U_i$ and $|G_i| \gneq |U_i|$: all system suggestions are in the gold standard corrections, but not all gold standard corrections are in the system suggestions;
- $G_i \cap U_i \neq \emptyset$ and $|G_i \cap U_i| \lneq |U_i|$: at least one system suggestion is in the gold standard corrections, and at least one system suggestion is not in the gold standard corrections.

3.3 Evaluating Overall System Performance

In order to obtain a single number to denote the reliability of a system, we suggest to use an evaluation schema similar to F_1:

$$P_{overall} = \frac{2 \times P_{detection} \times P_{correction}}{P_{detection} + P_{correction}}.$$

4 Automatic Chinese Typo Detection System

We developed a system called ACT, which is an extension of our previous work (Dong et al. 2016). Since our first deployment we have made several enhancements, the most important of which is briefly explained below.

ACT conducts segmentation and part-of-speech tagging on sentences to find words that do not fit in the context. Based on big data mining in conjunction with a unique intelligent algorithm and a new scoring mechanism, ACT efficiently identifies errors in sentences and offers replacement suggestions. For example, if 元素 is incorrectly written as 原素, then both segmentation and part-of-speech tagging return nonsensical results, which reveal the presence of an error.

While this framework is efficient, it handles reduplication inadequately. Occasionally two or more identical characters need to be replaced simultaneously in order for an error to get corrected. For example, given 惇惇教誨, 惇惇 should be replaced by 諄諄. However, sometimes only one of the identical characters is incorrect (e.g., for 實事求事, only the second 事 should be replaced by 是). We are not aware of any algorithm that is able to solve this problem effectively. On encountering two or more identical characters, ACT calculates the most viable correction by considering different possible replacements by means of language models (Heafield 2011). In terms of the above evaluation schema, we achieved a detection score of 69.06%, a correction score of 91.47%, and an overall system performance score of 78.7%, which we regard as satisfactory. We opened our ACT website to the public for testing and utilisation.[2]

5 Conclusions and Future Work

We released a benchmark dataset and an evaluation schema for Chinese typo detection, both of which will be useful for future research. Furthermore, we improved our Chinese typo detection system ACT. Future work includes developing new ways to automatically detect redundancy (e.g., detecting that one 論 in the sentence 他正在寫論論文 is redundant), and finding ways to fill in missing characters (e.g., filling in the missing verb in the sentence 他正在論文). Both problems require an understanding of the underlying semantics in order to obtain appropriate solutions.

References

Dong, S.; Fung, G. P. C.; Li, B.; Peng, B.; Liao, M.; Zhu, J.; and Wong, K.-F. 2016. ACE: Automatic Colloquialism, Typographical and Orthographic Errors Detection for Chinese Language. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016): System Demonstrations*, 194–197.

Fung, G. P. C.; Debosschere, M.; Wang, D.; Li, B.; Zhu, J.; and Wong, K.-F. 2017. NLPTEA 2017 Shared Task – Chinese Spelling Check. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*. In Press.

Heafield, K. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, 187–197. Association for Computational Linguistics.

[2] www1.se.cuhk.edu.hk/~moelab/act/