NLPTEA 2017 Shared Task – Chinese Spelling Check

Gabriel Pui Cheong Fung[*], Maxime Debosschere[*], Dingmin Wang[*§], Bo Li[*], Jia Zhu[‡], and Kam-Fai Wong[*]

[*] Department of Sys. Eng. & Eng. Mgt., The Chinese University of Hong Kong, China
{pcfung, dmaxime, boli, kfwong}@se.cuhk.edu.hk

[‡] Department of Computer Science, South China Normal University, China
jzhu@m.scnu.edu.cn

[§] Department of Computer Science, Tsinghua University, China
wangdm15@mails.tsinghua.edu.cn

Abstract

This paper provides an overview along with our findings of the Chinese Spelling Check shared task at NLPTEA 2017. The goal of this task is to develop a computer-assisted system to automatically diagnose typing errors in traditional Chinese sentences written by students. We defined six types of errors which belong to two categories. Given a sentence, the system should detect where the errors are, and for each detected error determine its type and provide correction suggestions. We designed, constructed, and released a benchmark dataset for this task.

1 Introduction

Automatic spelling checking is the task of using machines to automatically detect writing errors (Mays et al., 1991). Most popular word processors have this capability for alphabet-based languages, such as English, French and German, but not for character-based languages, such as Chinese.

In this shared task, we focus on spelling checking Chinese, which is very different than checking alphabetic languages due to several distinct properties of the Chinese language:

1. There is a vast variety of characters. There exist more than 100,000 Chinese characters, around 3,500 of which constitute the common characters of modern Chinese. Many characters have similar shapes and/or similar pronunciations.

2. There are no delimiters between words.

3. Each character has a meaning. Furthermore, the length of words is usually very short, ranging between one and four characters.

4. Depending on their positions in a sentence, identical characters or words can sometimes belong to different kinds of part of speech (verb, noun, etc.).

5. There exist many colloquial words and phrases that do not occur in written Chinese. This property becomes especially obvious in Cantonese, which is a dialect of Chinese. There is a significant number of words, phrases, and sentence structures that are valid in daily conversation, yet are considered invalid when written down.

We observed that publicly available benchmark data for Chinese spelling checking is limited. To make matters worse, no benchmark dataset targets the last of the aforementioned properties. This motivated us to develop a new benchmark dataset which takes colloquialism into account, and which is publicly available in order to promote future research of Chinese spelling checking in this area.

In general, a good spelling checker is able to detect errors and provide correction suggestions for each detected error. Since every character in Chinese has a meaning (i.e., every character can always be regarded as a word, which is very different from alphabetic-based languages), spelling checks must be done within a context, such as a sentence or a long phrase with a certain meaning, rather than within very few words (Mays et al., 1991). Accordingly, we collected a number of students' writings to serve as the benchmark data for this shared task.

For the evaluation, it should be noted that we do not have any widely recognised or standard evaluation schemas specifically designed for evaluating the performance of Chinese spelling checkers. Nonetheless, different evaluation schemas have been proposed for different purposes (Duan et al., 2012; Tseng et al., 2015; Wu et al., 2013; Yu et al., 2008, 2014; Zhao and Liu, 2010). Since we could not find any existing evaluation schema that fulfils all our evaluation criteria, we proposed a new evaluation schema in this task. We understand that this proposed evaluation schema may not be perfect, however it does capture most essential elements for considering whether a spelling checker is effective.

The rest of this paper is organised as follows: Section 2 describes the benchmark dataset, Section 3 presents the tasks, Section 4 outlines the evaluation schema, and Section 5 reports the findings and concludes this paper.

2 Benchmark Dataset

The Hong Kong Applied Science and Technology Research Institute (ASTRI), founded by the Government of the Hong Kong Special Administrative Region in 2000, 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 The Chinese University of Hong Kong to help mark and annotate these writings. Next, we extracted sentences with at least one error, and from this subset we manually filtered all sentences which are semantically meaningful, have a reasonable length, and are easy to understand without additional context. A total of 6,890 sentences met these criteria. Each sentence contains 50 to 150 characters, including punctuation marks. The average number of errors in a sentence is 2.7, and the maximum is 5. Finally, we defined the following six 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 今荀, 今 is a typo and should be replaced by 甘. 今 and 甘 have similar pronunciations in Cantonese.

3. Typo – Mixing simplified and traditional Chinese. E.g., for the word 词語, 词 is simplified Chinese and should be replaced by its traditional counterpart 詞.

4. Colloquialism – Incorrect character. E.g., for the sentence "佢比你高" the character 佢 is colloquial and should be changed into formal writing: either 他 or 她.

5. Colloquialism – Incorrect word or phrase. All characters are proper formal Chinese, but when combined they form a colloquial word. E.g., in the sentence "昨天撞返一個很久沒有見面的小學同學" the word 撞返 is colloquial even though the characters 撞 and 返 are both formal written Chinese. Here, 撞返 should be replaced by 碰見.

6. Colloquialism – Incorrect usage. All characters are properly written without any colloquial characters or words or phrases, but the ordering of some characters or words is incorrect, resulting 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.

We classify the first three types of errors as "typos" and the last three types of errors as "colloquialisms". Since all the writings in our dataset came from Hong Kong students, all colloquialisms in our dataset are Cantonese. Note that it is possible to have any mixture of the above cases, even if just colloquialisms. For example, consider the sentence "大家討論緊這件事". In this context, the character 緊 is a colloquial word and means 正在 (error type 4 in the aforementioned classification). Yet, simply replacing 緊 by 正在 is still wrong since it then triggers error type 5. Instead, the correction should be "大家正在討論這件事".

Since our benchmark dataset also required positive examples, we manually added 3,110 entirely correct sentences from our collection of writings, reaching a round total of 10,000 sentences. Next, we randomly selected 1,000 sentences from our dataset as training data, and another 1,000 sentences as testing data. To the best of our knowledge, there is no publicly available benchmark dataset that takes into account all six types of errors outlined above. We are the first to release such dataset, and it can be obtained from the project website.[1]

[1] https://www.labviso.com/nlptea2017/download/

3 Tasks: Detection and Correction

The objective of this shared task is to develop a computer-assisted system that automatically diagnoses typing errors in traditional Chinese sentences written by native Hong Kong primary students.

3.1 Overview

As mentioned in Section 2, there are two categories of errors: typos and colloquialisms. A sentence may be free of errors, have one error, or have multiple errors. Here are some additional examples:

- No error:
  我很喜歡吃媽媽做的瓜炒蛋飯。
- Typo only:
  我很喜歡吃媽媽做的梁瓜炒蛋飯。
- Colloquialism only:
  我很鍾意吃媽媽做的瓜炒蛋飯。
- Typo and colloquialism:
  我很鍾意吃媽媽做的梁瓜炒蛋飯。
- Multiple typos and multiple colloquialisms:
  我很鍾意食媽媽做的梁瓜炒旦飯。

As this is the first time we have colloquialisms in benchmark data, we provide a Cantonese-Chinese mapping dictionary in this shared task. This dictionary is in JSON format and contains all mappings between Cantonese and formal written Chinese. All Cantonese language that appears in the training and testing datasets is guaranteed to be included in this file. Note that a Cantonese phrase may have more than one possible mapping (depending on the context of the sentence) and different combinations of words in a phrase may yield entirely different results. For example:

```
{"cantonese":"唔", "chinese":["不"]},
{"cantonese":"唔使", "chinese":["不用"]},
{"cantonese":"唔該", "chinese":["請","謝
    謝"]},
{"cantonese":"邊度", "chinese":["哪裏"]},
{"cantonese":"邊處", "chinese":["哪裏"]}
```

We provide the training data, testing data, and their corresponding gold standards. Everything is in JSON format. For example, given the following sentences:

```
{
  "id":"ASTRI01",
  "sentence":"我很喜歡吃媽媽做的凉瓜炒蛋飯。"
},
{
  "id":"ASTRI02",
  "sentence":"我很喜歡吃媽媽做的梁瓜炒蛋飯。"
},
{
  "id":"ASTRI03",
  "sentence":"我很鍾意吃媽媽做的凉瓜炒蛋飯。"
},
{
  "id":"ASTRI04",
  "sentence":"我很鍾意食媽媽做的梁瓜炒旦飯。"
}
```

the corresponding gold standard is:

```
{
  "id":"ASTRI01",
  "typo":null,
  "cantonese":null
},
{
  "id":"ASTRI02",
  "typo":[
    {
      "position":10,
      "correction":["凉"]
    }],
  "cantonese":null,
  "reorder":null
},
{
  "id":"ASTRI03",
  "typo":null,
  "cantonese":[
    {
      "position":3,
      "length":2,
      "correction":["喜歡"]
    }],
  "reorder":null
},
{
  "id":"ASTRI04",
  "typo":[
    {
      "position":10,
      "correction":["凉"]
    },
    {
      "position":13,
      "correction":["蛋"]
    }],
  "cantonese":[
    {
      "position":3,
      "length":2,
      "correction":["喜歡"]
    },
    {
```

```
      "position":5,
      "length":1,
      "correction":["吃"]
    }],
  "reorder":null
}
```

The structure and meaning of the above examples should be self-explanatory. Note that according to Section 2, there are multiple types of colloquialism. This is the reason why the "reorder" field is necessary for colloquialism detection in a sentence. To illustrate this necessity, observe that when given the following sentences:

```
{
  "id":"ASTRI05",
  "sentence":"我走先然後去打球。"
},
{
  "id":"ASTRI06",
  "sentence":"大家討論緊這件事。"
}
```

the corresponding gold standard becomes:

```
{
  "id":"ASTRI05",
  "typo":null,
  "cantonese":null,
  "reorder":[
    {
      "position":1,
      "length":8,
      "correction":["我先走然後去打球"]
    }]
},
{
  "id":"ASTRI06",
  "typo":null,
  "cantonese":[
    {
      "position":5,
      "length":1,
      "correction":["正在"]
    }],
  "reorder":[
    {
      "position":1,
      "length":8,
      "correction":["大家緊討論這件事"]
    }]
}
```

3.2 Task 1 – Detection

Given a sentence, the system should be able to detect where the errors are, and for each detected error determine its type. Note that a sentence may have no errors, one error, or multiple errors (of multiple types).

3.3 Task 2 – Correction

For each detected error, the system should suggest how to correct the error. In contrast to previous similar computerised spelling check tasks (Duan et al., 2012; Tseng et al., 2015; Wu et al., 2013; Yu et al., 2008, 2014; Zhao and Liu, 2010), this shared task allows multiple correction suggestions. This idea originated from the fact that each spelling checker in modern word processing software provides a list of possible corrections for any given error, in order to maximise editing flexibility. Hence, it is reasonable to allow a system to output multiple correction suggestions for an error rather than just one.

4 Evaluation Schema

4.1 Evaluating Detection Performance

For evaluating the detection performance of a system, we compare the system output to the gold standard in terms of types of error and positions. Mathematically,

$$p = \frac{TP}{TP + FP}$$
$$r = \frac{TP}{TP + FN} \quad (1)$$
$$E_{detection} = \frac{2 \times p \times r}{p + r}$$

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. For example, given the following sentences:

```
{
  "id":"ASTRI2000",
  "sentence":"佢想禾你共進免餐。"
},
{
  "id":"ASTRI2001",
  "sentence":"仍記得小學下課的時候，我總愛到
      草推裏捉蠶蟲。"
},
{
  "id":"ASTRI2002",
  "sentence":"我走先然後去打球。"
}
```

and the following result:

```
{
  "id":"ASTRI2000",
  "typo":[
    {
      "position":3,
      "correction":["和"]
    },
    {
      "position":7,
      "correction":["晚", "挽", "行"]
    }],
  "cantonese":[
    {
      "position":1,
      "length":1,
      "correction":["他", "她"]
    }],
  "reorder":[]
},
{
  "id":"ASTRI2001",
  "typo":[
    {
       "position":1,
      "correction":["也"]
    }],
  "cantonese":[],
  "reorder":[]
},
{
  "id":"ASTRI2002",
  "typo":[],
  "cantonese":[],
  "reorder":[]
}
```


then $TP = 3$ (detected the typos 禾 and 免; detected the Cantonese usage 佢), $FP = 1$ (incorrectly suggested 仍 as a typo in ASTRI2001), and $FN = 2$ (did not detect the typo 推 in ASTRI2001 and did not detect the ordering problem of ASTRI2002).

4.2 Evaluating Correction Performance

There may be multiple ways to correct an error in a sentence. Hence, in the gold standard we included as many valid corrections as possible for each error. For example, given the following sentence:


```
{
  "id":"ASTRI2001",
  "sentence":"他想禾你共進免餐。"
}
```


the gold standard is:


```
{
  "id":"ASTRI2001",
  "typo":[
    {
      "position":3,
      "correction":["和"]
    },
    {
      "position":7,
      "correction":["晚", "午"]
    }],
  "cantonese":[],
  "reorder":[]
}
```


In this sentence 免 is a typo. Since 免 and 晚 have similar shapes whereas 免 and 午 have similar pronunciations, we consider both 晚 and 午 to be valid corrections of 免.

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. Consider the above example. If a system suggests a list of corrections [晚, 免] for position 7, then we evaluate that this system successfully detected the corresponding gold standard error, and that it provided one matching and one mismatching correction suggestion.

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,

$$E_{correction} = \frac{1}{|W|} \sum_{\forall i \in W} \frac{|G_i \cap U_i|}{|U_i|} \qquad (2)$$

where W is the set containing all successfully 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.

4.3 Evaluating Overall Performance

In practice, we usually need to obtain a single number to denote the reliability of a system. We suggest to use an evaluation schema similar to F_1 (Sebastiani, 2002):

$$E_{overall} = \frac{2 \times E_{detection} \times E_{correction}}{E_{detection} + E_{correction}} \quad (3)$$

where $E_{detection}$ and $E_{correction}$ are obtained from Sections 4.1 and 4.2, respectively.

5 Discussion and Conclusion

We have seven registered participants from different organisations and universities, including Beijing University of Posts and Telecommunications, National Chia-Yi University, Peking University, and Harvard University. Upon receiving and reviewing the reports, we included the reports "Chinese Spelling Check based on N-gram and String Matching Algorithm" from National Chia-Yi University and "N-gram Model for Chinese Grammatical Error Diagnosis" from Beijing University of Posts and Telecommunications in this proceeding. These two universities used completely different approaches for detection and correction. In terms of results, National Chia-Yi University achieved a detection score of 42.71%, a correction score of 95.47%, and an overall system performance score of 59.01%, which is rather impressive. We encourage our readers to refer to their papers in order to fully appreciate the diversity of their approaches, with their specific advantages and drawbacks.

To conclude, this paper described the Chinese spelling check task at NLPTEA 2017. We illustrated the difficulties of Chinese spelling checking and how it is different from the alphabet-based languages. We released the first ever benchmark dataset which takes the colloquialism property into account, and we proposed a new evaluation schema. The main breakthrough, however, is that we allowed systems to provide multiple correction suggestions, which is a property of most commercial spelling checkers and desirable from the user's perspective, yet missing in existing evaluation schemas and still generally neglected in the research community.

We hope that this shared task will provide inspiration and motivation to advance our knowledge and experience regarding Chinese language processing in general, and to continue the development of state-of-the-art techniques for Chinese spelling checking in particular. We sincerely thank ASTRI and all participants in this shared task.

References

Huiming Duan, Zhifang Sui, Ye Tian, and Wenjie Li. 2012. The CIPS-SIGHAN CLP 2012 Chinese Word Segmentation on MicroBlog Corpora Bakeoff. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2012)*, pages 35–40.

Eric Mays, Fred J Damerau, and Robert L Mercer. 1991. Context Based Spelling Correction. *Information Processing & Management*, 27(5):517–522.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47.

Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to SIGHAN 2015 Bake-off for Chinese Spelling Check. *ACL-IJCNLP 2015*, pages 32–37.

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages 35–42.

Liang-Chih Yu, Lung-Hao Lee, Yuen-Hsien Tseng, and Hsin-Hsi Chen. 2014. Overview of SIGHAN 2014 Bake-off for Chinese Spelling Check. In *Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2014)*, pages 126–132.

Xiaofeng Yu, Wai Lam, Shing-Kit Chan, Yiu Kei Wu, and Bo Chen. 2008. Chinese NER Using CRFs and Logic for the Fourth SIGHAN Bakeoff. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*, pages 102–105.

Hongmei Zhao and Qun Liu. 2010. The CIPS-SIGHAN CLP 2010 Chinese Word Segmentation Bakeoff. In *Proceedings of the First CPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 199–209.