Chinese Spelling Check based on N-gram and String Matching Algorithm

Jui-Feng Yeh*, Li-Ting Chang, Chan-Yi Liu , Tsung-Wei Hsu

Department of Computer Science & Information Engineering,
National Chia-Yi University,
Chiayi city, Taiwan (R.O.C.)
{ralph, 1060416, 1060417, 1050480}@mail.ncyu.edu.tw

Abstract

This paper presents a Chinese spelling check approach based on language models combined with string match algorithm to treat the problems resulted from the influence caused by Cantonese mother tone. N-grams first used to detecting the probability of sentence constructed by the writers, a string matching algorithm called Knuth-Morris-Pratt (KMP) Algorithm is used to detect and correct the error. According to the experimental results, the proposed approach can detect the error and provide the corresponding correction.

1 Introduction

In recent years, Chinese became more and more popular. Not only the Asian learning it as a mother language, but also people around the world learning as the second language. But Chinese is not easy to learn for almost every kinds of people. Due to the diversity of Chinese characters and flexibility of grammars, Chinese spelling correction plays an important role in both foreign learners and Natural Language Processing researchers.

Traditional Chinese is a sophisticated art. With Cantonese, it could be even harder. The high-educated may sometimes spell it wrong, not to mention those who are learning it. Because of the following reasons, Traditional Chinese is more difficult to learn than other languages. First, the Chinese's grammar is more complicated and more flexible than English's. For Cantonese usage, the orderings of some characters are sometimes incorrect. Second, Chinese characters are evolved from by Hieroglyphic. Every single character has its own meanings, but two or more characters combined to a word can express a whole different meaning. Therefore, we can simply classify the errors to Typo, which is spelling error, Cantonese usage, and ordering error. In this paper, we proposed the Chinese spelling check with Cantonese correction system using N-gram language model and Knuth-Morris-Pratt (KMP) algorithm. It can detect the spelling error and Cantonese usage form a sentence, then give the suggested correction.

In NLP-TEA 2017, a share task for Chinese spelling check (CSC) aims to detect and correct the spell errors in text. There are three error types defined in Table 1. 'Typo' for spelling error, 'Cantonese' for only Cantonese usage, 'reorder' for incorrect order in Cantonese usage. Some typical examples of the errors are shown in Table 1. For this shared task, each input could have multiple errors, which means that it might need several phrases to process each input sentence. As the result of that, the proposed system is divided as two phrases, Preprocessing phrase and Chinese spelling check with Cantonese correction phrase.

Error types	Examples of erroneous sentences	Examples of correct sentences
拼字錯誤 (Typo)	我喜歡吃**梁**瓜炒蛋飯。	我喜歡吃**涼**瓜炒蛋飯。
粵語字詞 (Cantonese)	**佢**比你高	**他**比你高
語序錯誤 (Ordering)	我**走先**然後去打球	我**先走**然後去打球

Table 1: Some typical examples of the errors

Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications, pages 35–38,
Taipei, Taiwan, December 1, 2017

For each input sentence, the system should output the location, length, and the corresponding correction. In order to address the problem, we need to establish Cantonese to Traditional Chinese dictionary. NLP-TEA 2017 released a dictionary about 1000 words, then we according to an open traditional Chinese and Cantonese dictionary (http://kaifangcidian.com/han/yue) added up to 8000 words. Figure 1 shows the example of Cantonese usage errors detected by the proposed system and Figure 2 shows the example of typo error errors detected by the proposed system.

- Example 1
  Input：這個週末，我們全家去渡假村泡溫泉。
  Output：0

- Example 2
  Input：媽媽不停地催速我：「快點穿衣服！不然就趕不上校車了！」
  Output：1 19 1 便

- Example 3
  Input：她細聲地對我說：「我借你一枝鉛筆。」
  Output：1 2 2 小聲

- Example 4
  Input：我在雪糕店工作那時，我們曾招待外藉人士的家庭，他們成日光顧，可是我們不喜歡他們。
  Output：1 3 2 冰淇淋

- Example 5
  Input：為什麼所有人也蝕錢而主角可以賺錢？
  Output：1 8 2 廚本,廚錢

Figure 1: Examples of Cantonese usage errors output by proposed system

- Example 1
  Input：我們栽歌栽舞，歡慶新一年到來。
  Output：載 3 載 5

- Example 2
  Input：媽媽不停地催速我：「快點穿衣服！不然就趕不上校車了！」
  Output：促 7

- Example 3
  Input：每年的情人節，爸爸都會帶著他的朋友逛花店，碰巧就是媽媽工作的花店，爸爸會讓媽媽挑選最美麗的一束玫瑰花送給朋友。
  Output：玫 48

- Example 4
  Input：爺爺的農莊裏有一些小胡蘆，我常用它們來裝水。
  Output：葫 11

- Example 5
  Input：現代科技的發展日新月義，不斷進步。
  Output：異 11

Figure 2: Examples of Typo errors output by proposed system

This paper is organized as follows: Section 2 describe the proposed method of Chinese spelling check with Cantonese correction system. Section 3 we analyze the performance in experimental results of the proposed system. Finally, Section 4 the conclusion of this paper.

2 The Proposed Method

In this section, the processing flow is shown as follow. The proposed system is divided as two main phrases: Preprocessing phrase and Chinese spelling check with Cantonese correction phrase. In Pre-processing phrase, we first extracting the information from the data that NLP-TEA 2017 released. First CKIP (Chinese Knowledge and Information Processing) Auto tag was applied to do word segmentation and obtain part-of-speech(POS). Then we proceed to Chinese spelling check phrase. Based on the POS, we determine whether the words in the sentence is correct, detail will describe in section 2.1. In section 2.2, we will describe the process of Cantonese correction and the algorithm we applied.

2.1 Language Models for Error Checking

In this section, we will explain how the works, and the method we used for checking words in sentences.

Equation 1 is the possibility of a string of characters from N-gram language model. For example, a string “我(I) 吃了(ate) 漢堡(hamburger)”, can obtain the possibility P(“我(I) 吃了(ate) 漢堡(hamburger)”)is equal to the production of P(“我(I)”)P(“吃了(ate)” | ”我(I)”) and P(“漢堡(hamburger)” | ”我(I) 吃了(ate)”). In which P(“漢堡(hamburger)”|”我(I) 吃了(ate)”), “我(I)” “吃了(ate)” is the left context dependency of ”漢堡(hamburger)”. With this approach, we can count the word and calculate the words’ possibility. Furthermore, use this possibility to measure whether the word is correct or not.

$$
\begin{aligned}
&P(S) = P(w_1 w_2 \dots w_n) \\
&= P(w_n | w_1 w_2 \dots w_{n-1}) P(w_1 w_2 \dots w_{n-1}) \\
&= P(w_n | w_1 w_2 \dots w_{n-1}) P(w_{n-1} | w_1 w_2 \dots w_{n-2}) \\
&\quad P(w_1 w_2 \dots w_{n-2}) \\
&= P(w_1) \prod_{i=2}^{n} P(w_i | w_1 w_2 \dots w_{i-1})
\end{aligned}
$$

Equation 1: Equation of the N-gram language model

After, we introduced the E-HowNet and a dictionary of Similar Pronunciation & Shape in Chinese character. We input the pre-processed data into proposed system, comparing with E-HowNet dictionary, processing flow shown as Figure 3.

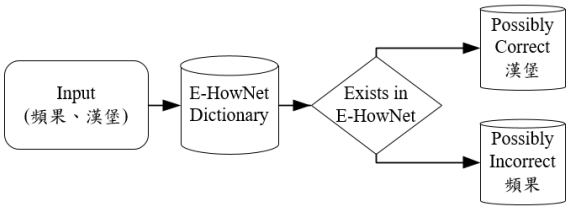

Figure 3: Example of comparing flow

With the collection of possibly incorrect words, we find the similar character from the dictionary of Similar Pronunciation & Shape in Chinese character to exchange the character in word separately. Then compare with E-HowNet dictionary again, if the exchanged word exists, means the exchanged word is correct and save it into a file.

The following Figure 4 shows an example of word exchanging, we exchange a character separately, then compare with E-HowNet dictionary, if the exchange result exists in E-HowNet, the process will stop and move on to the next word.

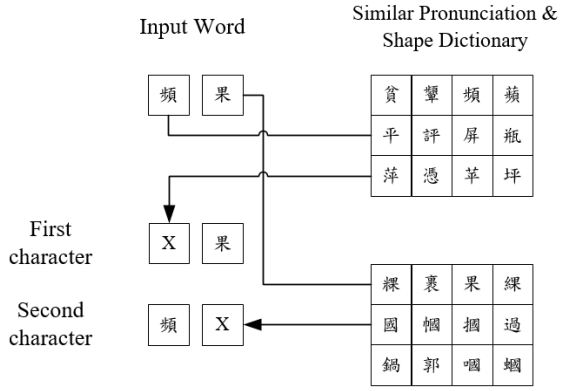

Figure 4: The exchange process of the word

2.2 Knuth-Morris-Pratt Algorithm for Correction

The correction system's main algorithm is based on KMP string matching algorithm. The KMP algorithm is used to search the specific string in a sentence. KMP is known for its highly effectiveness because it can search the specific string without starting over. In the middle of searching, we can also note down the position of the target string, which helps us to find out where the Cantonese usage is. Figure 5 shows an example of KMP algorithm, as the figure shows, we can skip lots of searching iteration.

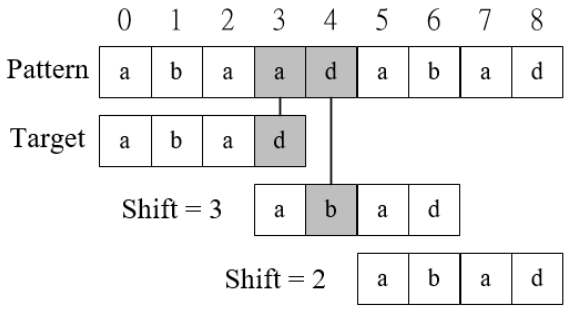

Figure 5: Example of KMP algorithm

When it comes to Cantonese correction, it is a lot easier than English string, because most of the Cantonese word are the combination of two or three character, word's length is often shorter than the English word.

By the using of KMP algorithm, any words that built in the dictionary appears in sentences, we will note the positon of the word and the translation of Cantonese usage. After search all the data, we will output a file that contain every result of the sentences, and according to this file output in the form of demanded format.

3 Experiments and Results

This experiment is based on the training data and testing data from NLP-TEA 2017. Each of the data contains 1000 sentences. After this experiment those sentences would be labeled with error tags and revise them or correct tags. And the results would determine by the performance standard given by NLP-TEA 2017 workshop. The evaluation matrices are shown as follow:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$Performance_{Detection} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

The confusion matrix is shown below in Table 2.

	System Results	
	Positive	Negative
Golden Standard Positive	TP (True Positive)	FP (False Negative)
Golden Standard Negative	FN (False Positive)	TN (True Negative)

Table 2: Confusion Matrix

Table 3 shows the detection performance evaluated by NLP-TEA 2017 workshop. TP, FP and FN

are 243, 143, and 509 respectively. By the formulas above, we can obtain precision, recall and performance are 62.9534%, 32.3138% and 42.7065% respectively.

Metric	Value
TP	243
FP	143
FN	509
Precision	62.9534%
Recall	32.3138%
Performance	42.7065%

Table 3: Detection performance by proposed system

There are other two standards for evaluations given by NLP-TEA 2017 workshop, Correction Performance and Overall System Performance. Correction performance is for each detected error, contestants' system would deliver one or more (five at most) correction suggestions, if correction suggestions are correct means that the union between the golden standard suggestions and the contestants' suggestions is not null. Following is the formula for Correction performance,

$$Performance_{Correction} = \frac{1}{|W|}\sum_{\forall_{i\in w}} \frac{|G_i \cap U_i|}{|U_i|} \quad (5)$$

Overall system performance means NLP-TEA 2017 ranked performance of the system as follows,

$$Performance_{Overall} = \frac{2*Performance_{Detection}*(Performance_{Correction})}{(Performance_{Detection})+Performance_{Correction}} \quad (6)$$

The result of proposed system evaluated by NLP-TEA 2017 is shown as follow:

Type	Value
Correction Performance	95.4737%
Overall System Performance	59.0149%

Table 4. Correction performance and overall system performance by our system

4 Conclusions

This paper main purpose is solving the NLP-TEA 2017 shared task for Chinese spelling check with Cantonese usage. Our proposed methods include KMP algorithm string matching and N-gram language model. According to the result of proposed system, it shows our system have some weakness should be revised. Recall rate is not ideal, it means that the proposed system should revised the algorithm of detecting Chinese spelling. But there is an advantage in the proposed system, the performance evaluation given by NLP-TEA 2017 shows that the correction performance of our system archives 95%. Therefore, we believe the proposed system is feasible. In the future, we will make an effort to improve the overall performance, especially on error detection to increase the recall rate.

References

CKIP AutoTag http://ckipsvr.iis.sinica.edu.tw/

E-HowNet http://ehownet.iis.sinica.edu.tw/

Rasool et al. 2012. Parallelization of KMP string matching algorithm on different SIMD architectures: Multi-core and GPGPU's. *International Journal of Computer Applications*, *49*(11).

Ku, D. T., and Chang, C. C. 2012. Development of context awareness learning system for elementary Chinese language learning. In *Genetic and Evolutionary Computing (ICGEC), 2012 Sixth International Conference*:538-541.

Lin, C. C., and Tsai, R. T. H. 2012. A Generative Data Augmentation Model for Enhancing Chinese Dialect Pronunciation Prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(4):1109-1117.

Yeh et al. 2015. Condition Random Fields-based Grammatical Error Detection for Chinese as Second Language. In *Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications*:105-110.

Lee et al. 2016. Overview of NLP-TEA 2016 Shared Task for Chinese Grammatical Error Diagnosis. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*:40-48.

Matthews, S., and Yip, V. 2013. *Cantonese: A comprehensive grammar*. Routledge.

Kingsbury et al 2013. A high-performance Cantonese keyword search system. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference*:8277-8281.

Cooper, A., and Wang, Y. 2012. The influence of linguistic and musical experience on Cantonese word learning. *The Journal of the Acoustical Society of America*, *131*(6):4756-4769.

Gao et al. 2000. Acoustic modeling for Chinese speech recognition: A comparative study of Mandarin and Cantonese. In *ICASSP'00. Proceedings. 2000 IEEE International Conference on* (Vol. 3): 1261-1264.