ACE: Automatic Colloquialism, Typographical and Orthographic Errors Detection for Chinese Language

Shichao Dong[1], **Gabriel Pui Cheong Fung**[2]
Binyang Li[4], **Baolin Peng**[1], **Ming Liao**[1], **Jia Zhu**[3] and **Kam-fai Wong**[1]

[1]Department of SEEM, The Chinese University of Hong Kong {scdong,blpeng,mliao,kfwong}@se.cuhk.edu.hk
[2]Director, Lab Viso Limited gabriel@labviso.com
[3]School of Computer Science, South China Normal University jzhu@m.scnu.edu.cn
[4]University of International RelationsUniversity of International Relations byli@uir.edu.cn

Abstract

We present a system called ACE for **A**utomatic **C**olloquialism and **E**rrors detection for written Chinese. ACE is based on the combination of N-gram model and rule-base model. Although it focuses on detecting colloquial Cantonese (a dialect of Chinese) at the current stage, it can be extended to detect other dialects. We chose Cantonese becauase it has many interesting properties, such as unique grammar system and huge colloquial terms, that turn the detection task extremely challenging. We conducted experiments using real data and synthetic data. The results indicated that ACE is highly reliable and effective.

1 Introduction

In general, there are two kinds of writing errors, typographical error (a.k.a. spelling errors) and orthographic error (a.k.a. cognitive error) (Damerau, 1964; Min et al., 2000). Typographical error means incorrectly substituting a right character with a wrong one, whereas orthographic error happens during the process of cognition. For colloquialism, there are two kinds as well: colloquial word and colloquial usage. For example, the word "返工" (means "back to work") is a colloquial Cantonese word. Its formal counterpart is "上班" (note: the characters of both words are completely different). On the other hand, the phrase "吃飯先" (go to dinner first), is a colloquial Cantonese usage and its formal counterpart is "先吃飯" (note: all characters in both words are the same but the position of the character "先" is different).

In this paper, we proposed a system called ACE (**A**utomatic **C**olloquialism and Spelling **E**rror Detector) to deal with all the errors stated previously. In ACE, there are three functions: (1) Identify the colloquial Cantonese words and usage; (2) Identify the potential spelling errors; (3) Provide correction suggestions.

To the best of our knowledge, there is no work related to automatically identify colloquial Cantonese. We do not aware any work on colloquialism in other language as well. For the work related to Chinese spelling error, (Lee et al., 2014) applied N-gram model and rule-based system to judge a sentence based on large number of data and experts knowledge. (Xie et al., 2015) builds a system using both N-gram model and Language model, and implements a dynamic programming to increase the efficiency. (Chang et al., 2015) implements a rule-base model and a linear regression model to tackle the task with the help of Chinese Orthographic Database. We observed that large training corpus is one of the key element for a reliable model (Tseng et al., 2015). Unfortunately, such setup is difficult to apply in our scenario because of the lack of Cantonese corpus.

2 System Description

ACE has two main modules: Cantonese detector and spelling error detector. Here is an outline of ACE: (Step 1) Identify over-segment parts in a sentence; (Step 2) Apply the Cantonese detector to check if there is any colloquial Cantonese (both usages and words); (Step 3) Apply the

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 194–197, Osaka, Japan, December 11-17 2016.

spelling error detector to check if there is any spelling error; (Step 4) Give correction suggestions for the errors detected in Step 2 and Step 3. In the followings we briefly describe the major elements within ACE.

2.1 Over-segment Parts

It is well proven that after sentence segmentation, the over-segment parts is an effective indicator to indicate potential spelling errors (Wu et al., 2010). Consider: "現在簡介有關香港電台數碼地面電視廣播法展概況", and its segmentation result: "現在/簡介/有關/香港電台 /數碼/地面/電視廣播/法/展/概況". The spelling error is "法" (the 4th last character). The correct character is"發". Note that the last four characters are segmented into three parts: "法/展/概況". If this sentence is written correctly as "... 發展概況", then the segmentation result will become ".../發展/概況". Hence two parts are resulted. By identifying the over-segment parts, we may have some cues if there is any potential spelling error. There are many different kinds of segmentation algorithms, such as HMM and Maximum Probability. In ACE, we use Maximum Probability as it performs that best empirically. Note that not all single-character word are regarded as over-segment part. Details will be discussed in Section 2.3.

2.2 Cantonese Detector

The Cantonese detector has two elements: (1) Build a large dictionary, and (2) Build a rule-base system. To build the large dictionary, apart from collecting the official Cantonese characters from the Hong Kong Information Office(http://www.gov.hk/tc/about/helpdesk) public education resources – "Hong Kong Extra Adding font collection", we further collect some "hot" and "trendy" Cantonese characters from online, such as OpenRice(http://www.openrice.com/zh/hongkong). There are totally more than 11000 words in our Cantonese dictionary. To build the rule-base system, we apply some Cantonese linguistic rules and use pos-tagging to describe these rules. Accordingly, we build eight rules as a start for eight basic Cantonese sentence structure. A rule usually has two parts: a flag word and a part-of-speech-tagging pattern. For example: the phrase "吃飯先" has no Cantonese characters, and it can be tagged as "吃飯/v 先/d"(ACE follows the ICTCLAS(http://ictclas.nlpir.org) part-of-speech-tagging standards). The rule can be organized as "1-先 v/d", the "1" indicates the position of the flag character, in this case "先", and "v/d" is the part-of-speech-tagging pattern of a certain phrase.

2.3 Spelling Error Detector

To detect the spelling errors in a sentence and offer replacement suggestions, a typical way is to employ an recursion algorithm as follows: (1) Check if there is any single-character word. A single-character word will be regarded as an over-segment part; (2) Replace the characters in the over-segmented parts by their corresponding confusion sets one by one. The confusion set of a character is the set of characters that are similar to the character typographically or orthographically; (3) Reassemble a new sentence and justify if the character replacement is appropriate. Unfortunately, we encountered several problems with such approach. First, there are many single correct characters in a Chinese sentence. For example, "是" (mean"is"), "地" (similar to append " ing" in a word, mean something is continuing) and "的" (similar to append "'s" in a noun) are all single-character word and usually appear in a sentence. They will always be segmented as a single-character word. If we perform the recursion algorithm as stated above, the whole system will be slow down dramatically and become useless in practice because many correct single-character words are regarded as over-segment parts. Second, unnecessary replacements may happen, because some single-character words have high-frequent replacement candidates according to the training data. For example, word "白" has a replace candidate "的" from the confusion set, but "的" has much higher frequency comparing with "白" in the training data, then an unnecessary replacement from "白" to "的" may happen despite of the context.

	Precision	Recall	F1
High	0.4843	0.7764	0.5839
Medium	0.4239	0.7872	0.5368
Low	0.2647	0.7505	0.3770

Table 1: The performance for large corps

To deal with these problems, we assign a score to every sentence based on its segmented words after the sentence segmentation. The score is computed based on a language model: the more frequent a word appears in the training data (e.g., the word "是"), the higher score it is and the higher co-occurrence of words combinations get higher score. Setting thresholds has been proven a useful method (Ferraro et al., 2011). We regard a single-character word as an over-segment part if and only if its score is higher than a predefined threshold. The threshold is computed based on the minmax principle: the smallest score of the most frequent word in the training data. In addition, we set bias on the sentence scores, if the length of words list becomes shorter, which means the number of over-segmented parts in a sentence decrease, ACE will add a positive bias on the score to make it higher. In contrast, the score of the sentence will become lower with adding a negative bias if the list of words of the sentence become longer.

In addition, in ACE, unlike the existing approaches which usually try to do the character replacement immediately once they identified a potential spelling error, we regard the consecutive over-segment parts as one candidate set and perform the replacement for all characters within such set. This can effectively help us to identify some spelling errors where two characters in a word are both spelling wrong. For example, if "政策" is incorrectly written as "正束" (both characters are written incorrectly), then ACE is possible to detect the error, whereas the existing approaches may not necessary able to do so.

To justify whether a replacement is appropriate, we follow the existing approaches by: (1) Reassemble the sentence after character replacement, (2) Score the sentence, and (3) If the new score is higher than the previous score, we say that the replacement is justifiable.

3 Experiments

We conducted experiments on synthetic data and real data. For synthetic data, we collected 500 error-free compositions from school students. For each composition, we randomly pick N Chinese terms from a predefined dictionary and replace them with the corresponding colloquial Cantonese. Next, we randomly pick M characters from the composition and replace them with one of the characters from their corresponding confusion sets. We vary M and N to test the sensitiveness of ACE. We set $M+N$ equals to 4, 8, 10 to denote low, medium and high level of errors. For real error data, we collect 411 sentences from Hong Kong school students. Each of them may have more than one spelling errors or colloquial Cantonese usage.

3.1 Evaluation Results

We compute precision, recall and F-1 using true positive (the no. of spelling errors that are correctly detected), false positive (the no. of non-existent errors are identified) and false negative (the no. of spelling errors cannot be detected) . Table 1 shows the results using synthetic data. The result is satisfactory and comparable to the latest existing works.

Table 2 shows some sample results using the real data. For the sentence "從令天開始，我就成為一名小學生啦！", "今天" is written as "令天". The sentence "快到聖誕節了，我和媽媽一同去構買聖誕禮物。" shares similar error. The sentence "表姐專門來送結婚請貼給爸爸媽媽。", "請帖" is written as "請貼", the wrong character is **not** the first character of the word. This indicate the ACE could be able to select the best candidate using its recursion replacement algorithm. There are some sentences have colloquial Cantonese usages and spelling errors in the same sentence. For example, "今天返工，突然下起的大雨淋得他混身都濕透了。" , "返工" is a

Sentence	Correction
從令天開始，我就成為一名小學生啦！	令 -> 今
我們不能隨便丟棄電弛，否則會污染環境。	弛 -> 池
快到聖誕節了，我和媽媽一同去構買聖誕禮物。	構 -> 購
今天返工，突然下起的大雨淋得他混身都濕透了。	返工 -> 下班, 混 -> 浑
我最愛吃媽媽包的交子啦！讓我吃飯先	交子 -> 餃子, 吃飯先 -> 先吃飯
表姐專門來送結婚請貼給爸爸媽媽。	請貼 -> 請帖
她是一位名付其實的好老師，學生們都很喜歡。	付 -> 符
常年累月的辛勞，使外公的腰越來越彎了。	常年累月 -> 長年累月
他誇耀自己的時候，總是眉飛色武，說個不停。	武 -> 舞

Table 2: Result Examples

colloquial word and "混身" is an error. ACE detects both errors successfully. ACE also detect the colloquial Cantonese *usage* (*not* Cantonese *word*). For example, "我最愛吃媽媽包的交子啦！讓我吃飯先", "交子" should be written as"餃子" and a colloquial usage "吃飯先". Finally, for a complex context such as "常年累月的辛勞，使外公的腰越來越彎了。", ACE could also detect the errors.

4 Conclusions

In this paper, we introduced ACE (**A**utomatic **C**olloquialism, Typographical and Orthographic **E**rror Detection) to detect the spelling errors and colloquial Cantonese from written Chinese, and to provide correction suggestions. The results indicated that ACE is effective and efficient.

Acknowledgment

This project is funded by the Hong Kong Applied Science and Technology Research Institute (ASTRI) (project code: 7050854), the fundamental research fund for the central universities (3262014T75, 3262015T70) and the National Natural Science Foundation of China (61502115, U1536207, 61370165, 61572043).

References

Tao-Hsing Chang, Cheng-Han Yang, and Hsueh-Chih Chen. 2015. Introduction to a proofreading tool for chinese spelling check task of sighan-8. In *ACL-IJCNLP 2015*, page 50.

Fred J Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Gabriela Ferraro, Rogelio Nazar, and Leo Wanner. 2011. Collocations: a challenge in computer assisted language learning. In *MTT*, pages 69–79.

Lung-Hao Lee, Liang-Chih Yu, Kuei-Ching Lee, Yuen-Hsien Tseng, Li-Ping Chang, and Hsin-Hsi Chen. 2014. A sentence judgment system for grammatical error detection. In *COLING*.

Kyongho Min, William H. Wilson, and Yoo-Jin Moon. 2000. Typographical and orthographical spelling error correction. In *LREC*.

Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to sighan 2015 bake-off for chinese spelling check. In *ACL-IJCNLP 2015*, page 32.

Shih-Hung Wu, Yong-Zhi Chen, Ping che Yang, Tsun Ku, and Chao-Lin Liu. 2010. Reducing the false alarm rate of chinese character error detection and correction. In *CLP 2010*, pages 54–61.

Weijian Xie, Peijie Huang, Xinrui Zhang, Kaiduo Hong, Qiang Huang, Bingzhou Chen, and Lei Huang. 2015. Chinese spelling check system based on n-gram model. In *ACL-IJCNLP 2015*, page 128.