Detecting Grammatical Errors in the NTOU CGED System by Identifying Frequent Subsentences

Chuan-Jie Lin and Shao-Heng Chen

Department of Computer Science and Engineering
National Taiwan Ocean University
{cjlin, shchen.cse}@mail.ntou.edu.tw

Abstract

The main goal of Chinese grammatical error diagnosis task is to detect word errors in the sentences written by Chinese-learning students. Our previous system would generate error-corrected sentences as candidates and their sentence likelihood were measured based on a large scale Chinese n-gram dataset. This year we further tried to identify long frequently-seen subsentences and label them as correct in order to avoid propose too many error candidates. Two new methods for suggesting missing and selection errors were also tested.

1 Introduction

The CGED (Chinese grammatical error diagnosis) tasks have been organized for 5 years (Yu *et al.*, 2014; Lee *et al.*, 2015; Lee *et al.*, 2016; Rao *et al.*, 2017). This task focuses on four kinds of errors in writing Chinese: using redundant words, missing words, arranging words in a wrong order, or using similar but incorrect words.

In our previous attempts in this task, our systems generated corrected-sentence candidates by different methods according to different error types. These candidates were scored by substring scoring functions (Lin and Chen, 2015). Although these systems were ranked in the middle place in the subtask of identification level, they tended to propose too many errors thus achieved rather low precisions.

This year we tried another approach to detect correct parts in a sentence before guessing positions of errors. The proposed methods in early and this year's tasks are explained in the following sections.

2 Identifying Frequent Subsentences

The main stage of the CGED tasks is to correct sentences written by Chinese-learning foreign students. The corrections were provided by Chinese teachers.

In our experience, corrections can be given in two levels. The first level is to make a sentence "correct" both in syntax and semantics. The second level is to make a sentence "better", which means the original sentence is also correct but there is a better paraphrase commonly used in Chinese. Unfortunately, our previous systems cannot distinguish the two different types of corrections. They will still propose suggestions when the original sentence is already a correct one.

In order to decrease the number of suggestions, we decided to trust the original sentences more. A simple approach is to detect frequently-seen long subsentences in Chinese. Only the positions not covered by the frequent subsentences can be candidates of grammatical errors. Our referencing database of frequent subsentences is the Chinese Web 5-gram dataset[1], which collects substrings occurring more than 20 times on the Internet.

The steps to identify frequent subsentences are described as follows. All substrings (with at least three Chinese characters) in the original sentence are looked up in the Chinese Web 5-gram dataset. All matched substrings in the original sentence are considered "correct". If two substrings overlap with at least one Chinese character (or two characters for substrings no longer than 4 Chinese characters), they are merged into one longer substring. After the matching process, only those words not covered by any substring can be deleted (as redundant errors), replaced (as selection errors), or

[1] https://catalog.ldc.upenn.edu/LDC2010T06

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 203–206
Melbourne, Australia, July 19, 2018.

switched (as disorder errors). And only the positions not inside any matched substring can have additional words being inserted (as missing errors).

An example of identifying frequent subsentences is given here. The second sentence in the Query 200405109523201166_2_1x2 in the training data is “我認為吸煙的壞處比長處更多”. We can match three subsentence in the Google 5-gram dataset:

我認為吸煙的	128
吸煙的壞處	2111
處更多	25635

The first two are further merged into one larger subsentence. So the identified frequent subsentences in the original sentence can be shown in brackets as [我認為吸煙的壞處]比長[處更多].

After substituting “長處” (advantage) with its synonym “好處” (advantage) by the methods described in Section 3.4, a longer subsentence “好處更多” can fully cover the previous identified frequent subsentence “處更多”. Therefore, an error will be reported as a Selection Error here.

好處更多	12938

3 Correction Candidate Generation

3.1 Character or Word Deletion (Case of Redundant)

Generating correction candidates in the case of Redundant type is quite straightforward: simply removing any substring in an arbitrary length. However, in order not to generate too many unnecessary candidates, we only do the removal under three special cases: removing one character, removing two-adjacent characters, and removing one word whose length is no longer than two Chinese characters. This method is the same as in the previous CGED tasks.

3.2 Character Insertion (Case of Missing)

The idea of generating correction candidates in the case of Missing type is to insert a character or a word into the given sentence. But it is impractical to enumerate candidates by inserting every known Chinese characters or words. We observed the CGED 2015 training set (Lin and Chen, 2015) and collected 34 characters which were frequently missing in the essays written by Chinese-learning foreign students, as they occurred at least three times and covered 73.7% of the missing errors in the CGED 2015 training set. Insertion happens between characters or words as usual.

A new idea to find insertion candidates was tested this year. Instead of inserting frequently missing characters, we directly discovered the n-gram string with the highest frequency in the Google 5-gram dataset. Take the sentence of the Query 200405205525200106_2_2x1 “這個團體的目的是減少邊走邊抽的人” as an example. When considering inserting characters in the position between “抽” and “的”, we found the longest most-frequent n-gram string is “邊抽煙的人” (a person smoking at the same time) which is the correct Missing Error.

3.3 Substring Moving (Case of Disorder)

Generating correction candidates in the case of Disorder type is also straightforward: simply moving any substring in any length to another position to its right (not to its left so that no duplication will be produced). This method is the same as in the previous CGED tasks.

3.4 String Substitution (Cases of Selection)

The first case of selection errors is the misuse of prepositions. To generate the correction candidates for preposition substitutions, we first extracted all prepositions in the Academia Sinica Balanced Corpus (ASBC for short hereafter, cf. Chen *et al.*, 1996). An input sentence is word-segmented and POS-tagged automatically beforehand. Correction candidates are generated by replacing each preposition (whose POS is “P”) in the given sentence by other prepositions.

The second case of selection errors is the misuse of synonyms. As we known, even synonyms cannot freely replace each other without considering context.

To generate the correction candidates for synonym substitutions, we consulted a Chinese thesaurus, Tongyici Cilin[2] (the extended version; Cilin for short hereafter). A given sentence is word-segmented beforehand. Correction candidates are generated by replacing each word in the given sentence by its synonyms in Cilin if any.

[2] http://ir.hit.edu.cn/
http://www.ltp-cloud.com/

The third case of selection errors is the misuse of words which were lexically similar to the correct ones. It is possible that the writer tried to use a word but misused another word with similar looking, such as “仔細” (carefully) and “細節” (details).

To generate but not over-generate the correction candidates for similar string substitutions, we first collected all 2-character words in the Google 5-gram dataset. Correction candidates are generated by replacing each 2-character word in the given sentence by another 2-character word having at least one character in common, such as “仔細” and “細節” where “細” appears in both words, or “合適” (suitable, *adjective*) and “適合” (suiting, *verb*) where both characters appear in both words. Examples of similar string substitution are given in the next page.

A new idea to find selection candidates was tested this year. We searched the Google 5-gram dataset and extracted the n-gram string with the highest frequency which differed with the original sentence with only one or two characters.

Take the second sentence of the Query 200405109523200578_2_1x2 “吸煙也是各人的人權” as an example. When considering replacing the character “各”, we found the longest most-frequent n-gram string is “是個人的人權” (is a personal human right) which is the correct Selection Error.

4 Substring Scoring Functions

In our previous work (Lin and Chen, 2015), we have defined a sentence likelihood scoring function to measure the likelihood of a sentence to be common and correct. This function uses frequencies provided in the Chinese Web 5-gram dataset in a way described as follows.

Chinese Web 5-gram consists of real data released by Google Inc. which were collected from a large amount of webpages in the World Wide Web. Entries in the dataset are unigrams to 5-grams. Frequencies of these n-grams are also provided. Some examples from the Chinese Web 5-gram dataset are given in the left part of Table 1.

In order to avoid interference of word segmentation errors, we decided to use substrings instead of word n-grams as the scoring units of likelihood. When scoring a sentence, frequencies of all substrings in all lengths are used to measure the likelihood.

Frequencies of substrings are derived by removing space between n-grams in the Chinese Web 5-gram dataset. For instances, n-grams in the left part of Table 1 will become the strings in the right part, where length of a substring is measured in bytes and a Chinese character often occupies 3 bytes in UTF-8 encoding. Note that if two or more different n-grams are transformed into the same substring after removing the space, they become one entry and its new frequency is the summation of their original frequencies. Simplified Chinese words were translated into Traditional Chinese in advanced.

Some notations are explained as follows. Given a sentence *S*, let *SubStr*(*S*, *n*) be the set of all substrings in *S* whose lengths are *n* bytes, and ***Google String Frequency*** *gsf*(*u*) be the frequency of a string *u* in the modified Chinese Web 5-gram dataset. If a string does not appear in that dataset, its *gsf* value is defined to be 1 (so that its logarithm becomes 0).

Equation 1 gives the equation of ***length-weighted string log-frequency score*** *SL*(*S*). Each substring *u* in *S* contributes a score of the logarithm of its Google string frequency weighted by *u*’s length *n*. The value of *n* starts from 6, because most content words are not shorter than 6 bytes (i.e. two Chinese characters).

$$SL(S) = \sum_{n=6}^{len(S)} \left(n \times \sum_{u \in SubStr(S,n)} \log\left(gsf(u)\right) \right) \qquad \text{Eq 1.}$$

This function was also explained in the work of Lin and Chu (2015). Please refer to that paper for examples of how to compute the sentence generation likelihood scores.

5 Run Submission

We planned to submit two runs this year. One run was produced with the previous system, i.e. generating error-correction candidates and choosing the ones with the highest length-weighted substring scores. The other run was produced by identifying frequent subsentences and then proposing errors containing in longer, more frequent n-gram strings found by new candidate generating methods.

Unfortunately, due to some errors in our procedures, only the one run was produced which reported as many errors as our previous system. We will finish the correct experiment as soon as possible to see the real performance of the newly proposed methods.

Two different strategies to identify frequent subsentences have been observed on the training data, where two thresholds are defined as follows. The ***length threshold*** (*lenTh*) defines the confident level of a subsentence in length (in Chinese characters). All subsentences no shorter than the length threshold are marked as "correct". The ***frequency threshold*** (*frqTh*) defines the confident level of a subsentence in frequency. All subsentences with a high frequency are also marked as "correct", even though their lengths might be short. The correction-candidates inside these "correct" subsentences are discarded.

Table 1 shows the evaluation results at the position level in the training data with different combination of length thresholds and frequency thresholds. The results suggest that trusting subsentences with at least 8 Chinese characters or appearing at least 900000 times in the Internet can reduced the erroneous proposal of corrections in a best way.

lenTh	frqTh	F-score (%)
20	1000000	8.9160
10	1000000	8.9244
8	1000000	8.9823
5	1000000	7.9004
20	900000	9.1603
10	900000	9.1647
8	**900000**	**9.2144**
5	900000	7.9651
20	500000	8.8235
10	500000	8.8280
8	500000	8.9185
5	500000	7.9507

Table 1: Performance of Error Proposal at the Position Level in the Training Data.

6 Conclusion

This paper describes the design of our Chinese grammatical error diagnosis system. This is our fourth attempt in the CGED tasks. Long frequent subsentences in the original sentences were identified in the first step. An error could be proposed only if it was not covered by a longer "correct" subsentences. Two runs were planned to be submitted. One run was produced with the previous system, i.e. generating error-correction candidates and choosing the ones with the highest length-weighted substring scores. The other run was produced by identifying frequent subsentences and then proposing errors containing in longer, more frequent n-gram strings found by new candidate generating methods.

References

Liang-Chih Yu, Lung-Hao Lee, and Li-Ping Chang (2014) "Overview of Grammatical Error Diagnosis for Learning Chinese as a Foreign Language," *Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications* (*NLPTEA '14*), pp. 42-47.

Lung-Hao Lee, Liang-Chih Yu, and Li-Ping Chang (2015) "Overview of the NLP-TEA 2015 Shared Task for Chinese Grammatical Error Diagnosis," *Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications* (*NLP-TEA 2*), *the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing* (*ACL-IJCNLP 2015*), pp. 1-6.

Lung-Hao Lee, Gaoqi Rao, Liang-Chih Yu, Endong Xun, Baolin Zhang, and Li-Ping Chang (2016). "Overview of the NLP-TEA 2016 Shared Task for Chinese Grammatical Error Diagnosis," *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications* (*NLPTEA '16*), pp. 40-48.

Gaoqi Rao, Baolin Zhang, Endong Xun (2017) "IJCNLP-2017 Task 1: Chinese Grammatical Error Diagnosis," *Proceedings of the 8th International Joint Conference on Natural Language Processing* (*IJCNLP*), *Shared Tasks*, pp. 1-8.

Chuan-Jie Lin and Shao-Heng Chen (2015) "NTOU Chinese Grammar Checker for CGED Shared Task," *Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications* (*NLP-TEA 2*), *the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing* (*ACL-IJCNLP 2015*), pp. 15-19.

Chuan-Jie Lin and Wei-Cheng Chu (2015) "A Study on Chinese Spelling Check Using Confusion Sets and N-gram Statistics," *International Journal of Computational Linguistics and Chinese Language Processing* (*IJCLCLP*), Vol. 20, No. 1, pp. 23-48.