NTOU Chinese Grammar Checker for CGED Shared Task

Chuan-Jie Lin and Shao-Heng Chen
Department of Computer Science and Engineering
National Taiwan Ocean University
No 2, Pei-Ning Road, Keelung 202, Taiwan R.O.C.
{cjlin, shchen.cse}@ntou.edu.tw

Abstract

Grammatical error diagnosis is an essential part in a language-learning tutoring system. Participating in the second Chinese grammar error detection task, we proposed a new system which measures the likelihood of sentences generated by deleting, inserting, or exchanging characters or words. Two sentence likelihood functions were proposed based on frequencies of space-removed version of Google n-grams. The best system achieved a precision of 23.4% and a recall of 36.4% in the identification level.

1 Introduction

Although that Chinese grammars are not defined as clearly as English, Chinese native speakers can easily identify grammatical errors in sentences. This is one of the most difficult parts for foreigners to learn Chinese. They are often uncertain about the proper grammars to make sentences. It is an interesting research topic to develop a Chinese grammar checker to give helps in Chinese learning. There have been several researches focused on Chinese (Wu *et al*., 2010; Chang *et al*., 2012; Yu and Chen, 2012; Tseng *et al*., 2014).

In NLPTEA-1 (Yu *et al*., 2014), the first Chinese grammatical error diagnosis evaluation project, the organizers defined four kinds of grammatical errors: redundant, missing, selection, and disorder. The evaluation was based on detection of error occurrence in a sentence, disregarding its location and correction. We developed an error detection system by machine learning.

However in NLPTEA2-CGED (Lee *et al*., 2015), it is required to report the location of a detected error. To meet this requirement, two new systems were proposed in this paper. The first one was an adaptation of the classifier developed by machine learning where location information was considered. The second one employed hand-crafted rules to predict the locations of errors.

We also designed two scoring functions to predict the likelihood of a sentence. Totally three runs were submitted to NLPTEA2-CGED task. Evaluation results showed that rule-based systems achieved better performance. More details are described in the rest of this paper.

This paper is organized as follows. Section 2 gives the definition of Chinese grammatical error diagnosis task. Section 3 delivers our newly proposed n-gram statistics-based systems. Section 4 gives a brief description about our SVM classifier. Section 5 shows the evaluation results and Section 6 concludes this paper.

2 Task Definition

The task of Chinese grammatical error diagnosis (CGED) in NLPTEA2 is defined as follows. Given a sentence, a CGED system should first decide if there is any of the four types of errors occur in the sentence: redundant, missing, selection, and disorder. If an error is found, report its beginning and ending locations.

Training data provided by the task organizers contain the error types and corrected sentences. Four types of errors are shortly explained here. All examples are selected from the training set where the locations of errors are measured in Chinese characters.

- Redundant: some unnecessary character appears in a sentence

Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 15–19,
Beijing, China, July 31, 2015. ©2015 Association for Computational Linguistics and Asian Federation of Natural Language Processing

[A2-0598, Redundant, 3, 3]
(X) 他是真很好的人
(He is a *really very good man.)
(O) 他是很好的人
(He is a very good man.)

- Missing: some necessary character is missing in a sentence

  [B1-0046, Missing, 4, 4]
  (X) 母親節一個禮拜就要到了
  (Mother's Day is coming in one week.)
  (O) 母親節再一個禮拜就要到了
  (Mother's Day is coming in one more week.)

- Selection: a word is misused and should be replaced by another word

  [B1-1544, Selection, 1, 2]
  (X) 還給原來的地方只花幾秒鐘而已
  (It only takes a few seconds to *return it to its original place.)
  (O) 放回原來的地方只花幾秒鐘而已
  (It only takes a few seconds to put it back to its original place.)

  Note that sometimes a SELECTION error looks like a missing character rather than a misused word. It is because there are many one-character words in Chinese. An example is given as follows.

  [B1-1546, Selection, 5, 5]
  (X) 關於跟你見的事
  (About the seeing with you…)
  (O) 關於跟你見面的事
  (About the meeting with you…)

- Disorder: some words' locations should be exchanged

  [B1-2099, Disorder, 4, 6]
  (X) 當然我會一定開心
  (Of course I will be certainly happy.)
  (O) 當然我一定會開心
  (Of course I will certainly be happy.)

3 N-gram Statistics-Based System

Besides the classifiers developed in the last CGED task (Yu *et al*., 2014), we proposed a new method to build a CGED system based on n-gram statistics from the World Wide Web.

Our assumption is: a corrected sentence has a larger probability than an erroneous sentence. I.e. deleting unnecessary characters, adding necessary characters, and exchanging locations of misplaced words will result in a better sentence. Our system will try to delete, insert, or exchange characters or words in a given sentence to see if the newly generated sentence receives a higher score of likelihood. Steps and details are described in this section.

3.1 Sentence Likelihood Scores

Since our method heavily counts on likelihood of a sentence being seen in Chinese, it is important to choose a good scoring function to measure the likelihood. Although n-gram language model is a common choice, a corpus in a very large scale with word-segmentation information is not easy to obtain. An alternation is to use Google N-gram frequency data.

Chinese Web 5-gram[1] is real data released by Google Inc. who collected from all webpages in the World Wide Web which are unigram to 5-grams. Frequencies of these ngrams are also provided. Some examples from the Chinese Web 5-gram dataset are given here:

Unigram:	稀釋劑	17260
Bigram:	蒸發量 超過	69
Trigram:	能量 遠 低於	113
4-gram:	張貼 色情 圖片 或	73
5-gram:	幸好 我們 發現 得 早	155

We have proposed several sentence likelihood scoring functions when dealing with Chinese spelling errors (Lin and Chu, 2015). But in order to avoid interference of word segmentation errors, we further design some likelihood scoring functions which utilize substring frequencies instead of word n-gram frequencies.

By removing space between n-grams in the Chinese Web 5-gram dataset, we constructed a new dataset containing identical substrings with their web frequencies. For instances, n-grams in the previous example will become:

Length=9:	稀釋劑	17260
Length=15:	蒸發量超過	69
Length=15:	能量遠低於	113
Length=18:	張貼色情圖片或	73
Length=24:	幸好我們發現得早	155

Note that if two different n-gram sets become the same after removing the space, their will merge

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

into one entry with the summation of their frequencies. Simplified Chinese words were translated into Traditional Chinese in advanced.

Given a sentence *S*, let *SubStr*(*S*, *n*) be the set of all substrings in *S* whose lengths are *n* bytes. We define ***Google String Frequency*** *gsf*(*u*) of a string *u* with length *n* to be its frequency data provided in the modified Chinese Web 5-gram dataset. If a string does not appear in that dataset, its *gsf* value is defined to be 0.

Two new sentence likelihood scoring functions are defined as follows. Equation 1 gives the definitions of ***length-weighted string log frequency score*** *SL*(*S*) where each substring in *S* with a length of *n* contributes a score of the logarithm of its Google string frequency multiplied by *n*. We think that short strings are not that meaningful, this function only considers strings no shorter than 6 bytes (i.e. a two-character Chinese words or a bigram of one-character Chinese words.)

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

Equation 2 gives a macro-averaging version of Equation 1 where scores are averaged within each length before summation over different lengths.

$$SLe(S) = \sum_{n=6}^{len(S)} \left(\frac{n \times \sum_{u \in SubStr(S,n)} \log(gsf(u))}{|SubStr(S,n)|} \right) \quad (2)$$

3.2 Character Deletion (Case of Redundant)

To test if a sentence has a redundant character, a set of new sentences are generated by removing characters in the original sentence one by one. If any of the new sentences has a higher likelihood score than the original sentence, it may be the case of redundant-type error.

Because the experimental data are essays written by Chinese-learning foreign students, some redundant errors are commonly seen across different students. Table 1 shows the most frequent redundant errors in the training data.

Char	Freq	Char	Freq	Char	Freq
了	66	去	15	就	6
的	56	在	13	很	6
是	27	會	8	要	6
有	27	得	7	把	5

Table 1. Frequent Redundant Errors

In order not to generate too many new sentences, we only deleted the characters of the frequent redundant errors which occurred at least three times. There were 23 of them which covered 66% of the redundant errors in the training data. Examples of character deletion are as follows where 很 and 到 are frequent redundant errors.

[B1-0764] org: 我很想到跟你見面
new: 我想到跟你見面
new: 我很想跟你見面

3.3 Character Insertion (Case of Missing)

To test if a sentence has a missing character, a set of new sentences are generated by inserting a character into the original sentence at each position (including the beginning and the end). If any of the new sentences has a higher likelihood score than the original sentence, it may be the case of missing-type error.

Similarly, some missing errors are commonly seen across the essays written by Chinese-learning foreign students. Table 2 shows the most frequent missing errors in the training data.

Char	Freq	Char	Freq	Char	Freq
的	74	有	24	要	13
了	65	會	18	在	12
是	44	就	17	過	12
都	34	很	16	讓	11

Table 2. Frequent Missing Errors

In order not to generate too many new sentences, we only inserted the characters of the frequent redundant errors which occurred at least three times. There were 34 of them which covered 73.7% of the missing errors in the training data. Examples of character deletion are as follows.

[B1-1047] org: 我真很怕
new: 的我真很怕
new: 我的真很怕
.....
new: 我真很怕的
new: 了我真很怕
.....
new: 我真很怕買

3.4 Word Exchanging (Case of Disorder)

To test if a sentence has a disorder error, the original sentence is word-segmented, and a set of new sentences are generated by exchanging words in the original sentence, each pair at a time. If any of the new sentences has a higher

likelihood score than the original sentence, it may be the case of disorder-type error. Examples of word exchange are as follows.

[B1-1047] org: 我 真 很 怕
new: 真 我 很 怕
new: 很 真 我 怕
new: 怕 真 很 我
new: 我 很 真 怕
new: 我 怕 很 真
new: 我 真 怕 很

3.5 Error Decision

All the new sentences, whenever generated by removing characters, inserting characters, or exchanging words, are scored by the sentence likelihood functions. The creation type and the modification location of the top-1 new sentence are reported as the error type and error location. If no new sentence's score is higher than the original's, it is reported as a "Correct" case.

3.6 Selection-Error Detection

If a detected error in Section 3.5 is a redundant case, it may also be a Selection-type error. If the deleted character occurs in a multi-character word in the original sentence, report this error as a Selection-type error.

[B1-0764] Redundant => Selection
org: 我 很 想到 跟 你 見面
(I really want to to meet you.)
new: 我 很 想 跟 你 見面
(I really want to meet you.)

Similarly, if a detected error in Section 3.5 is a missing case, it may also be a Selection-type error. To make a decision, the new sentence is also word-segmented. If the inserted character occurs in a multi-character word in the original sentence, report this error as a Selection-type error.

[B1-1047] Missing => Selection
[org] 我 真 很 怕
(I am *real scared.)
[new] 我 真的 很 怕
(I am really scared.)

4 Error Detection by Machine Learning

We also modified our previous CGED system participated in NLPTEA-1 to do error detection. It was a SVM classifier where 3 features were used for error detection:

- $f_{bi\text{-}}$: **number of infrequent word bigrams** appearing in the sentence, where "infrequent bigram" is defined as a bigram whose Google N-gram frequency is less than 100 or not even collected in the Chinese Web 5-gram dataset
- f_{stop}: a Boolean feature denoting the **occurrence of a stop POS bigram** which is often seen in a redundant-type error, such as **VH + T** (a stative intransitive verb followed by a particle) or **Cbb + DE** (a correlative conjunction followed by a function word "的")
- f_{len}: **length of the original sentence,** because a short sentence usually does not have missing- or disorder-type errors

Since the error detection classifier does not provide location information of an error, its location is decided by heuristic rules as follows.

1. If a stop POS bigram appears in the original sentence, the beginning and ending location of the first word matching this bigram are reported.
2. Or, if an infrequent word bigram appears in the original sentence, the beginning and ending location of the first word matching this bigram are reported.
3. Otherwise, simply report "1" as location.

5 Experiments

Three formal runs from our systems were submitted to NLPTEA2-CGED this year. The first run was created by the SVM classifier. The second run as created by the newly proposed CGED system with the original version of the length-weighted string log frequency function. The third run as created by the newly proposed CGED system with the macro-averaging version of the length-weighted string log frequency function.

	NTOU1	NTOU2	NTOU3
Detection Level			
Precision	50.00	**51.64**	50.98
Recall	**100.00**	97.60	98.60
F-1 Score	66.67	**67.54**	67.21
Identification Level			
Precision	18.96	**23.40**	20.95
Recall	28.48	**36.40**	31.96
F-1 Score	26.05	**33.40**	29.27
Position Level			
Precision	0.99	**1.00**	**1.00**
Recall	14.90	**16.00**	15.43
F-1 Score	12.38	**13.40**	12.87

Table 3. Evaluation Results of NTOU Runs

Table 3 shows the evaluation results of our three formal runs. All results suggest that a system using the length-weighted string log frequency function achieves better performance than a SVM classifier.

6 Conclusion

This is the second Chinese grammatical error diagnosis task. We proposed three systems to do the task. One is a SVM classifier where features are length, numbers of infrequent word bigrams, and occurrence of stop POS bigrams. The other two measure the likelihood of newly generated sentences by deleting, inserting, or exchanging characters or words. Two sentence likelihood functions were proposed based on frequencies of space-removed Google n-grams. The second system performed better than the other two which achieved a precision of 23.4% and a recall of 36.4%.

Although the performance seemed not good enough, our system was ranked at the second place in the identification level and the third in the position level, which means that the task is very hard. More rules and features should be studied in the future.

Reference

Ru-Yng Chang, Chung-Hsien Wu, and Philips Kokoh Prasetyo (2012). "Error Diagnosis of Chinese Sentences Using Inductive Learning Algorithm and Decomposition-Based Testing Mechanism," *ACM Transactions on Asian Language Information Processing*, 11(1), article 3, March 2012.

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," *Proceedings of the 25th International Conference on Computational Linguistics* (*COLING '14*), 67-70.

Lung-Hao Lee, Liang-Chih Yu, and Li-Ping Chang (2015). "Overview of Shared Task on Chinese Grammatical Error Diagnosis," *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, to be appeared.

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*, to be appeared.

Chung-Hsien Wu, Chao-Hong Liu, Matthew Harris, and Liang-Chih Yu (2010). "Sentence Correction Incorporating Relative Position and Parse Template Language Models," *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6), 1170-1181.

Chi-Hsin Yu and Hsin-Hsi Chen (2012). "Detecting Word Ordering Errors in Chinese Sentences for Learning Chinese as a Foreign Language," *Proceedings of the 24th International Conference on Computational Linguistics* (*COLING '12*), 3003-3017.

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*), 42-47.