Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education

Description of NTOU Chinese Grammar Checker in CFL 2014

Chuan-Jie Lin and Shao-Heng Chan
Department of Computer Science and Engineering,
National Taiwan Ocean University, Taiwan, R.O.C
cjlin@ntou.edu.tw

Abstract: This paper describes our first Chinese grammar checker participating in CFL 2014. Several features related to grammatical errors were proposed, including numbers of infrequent word bigrams and POS bigrams. Two SVM classifiers were trained and two formal runs were submitted, where the best F-scores were 66.67% in detection level and 34.36% in identification level.

Keywords: Chinese grammar checking, foreign language learning, machine learning

1. Introduction

Grammar checking for learning Chinese as a foreign language is a new challenge. Mistakes made by foreign students may greatly differ from the ones made by native speakers. It is necessary to study how to build a grammar checker for text written by students who learn Chinese as a foreign language.

As a shared task of ICCE, CFL 2014 (Grammatical Error Diagnosis for Learning **C**hinese as a **F**oreign **L**anguage) attempts to provide a benchmark to develop techniques on Chinese grammar checking. Four types of errors were defined in this task: redundant word, missing word, word disorder, and word selection problems. In this year, the task only focused on error detection and classification.

- **Redundant word**: a word should be deleted from this sentence
- **Missing word**: a missing word should be added into this sentence
- **Word disorder**: at least one word should change its location in this sentence
- **Word selection**: a word should be replaced into another word

This paper is organized as follows. Section 2 expresses some ideas we have got after observing examples in the training set. Section 3 gives definitions of features. Section 4 delivers experimental results and Section 5 concludes the paper.

2. Observation in Training Data

After observing example sentences in the training set, we found that the occurrence of low-frequency bigrams in the sentence is helpful. We used Google Web 1T 5-grams[1] (Google N-grams for short hereafter) as the resource of bigram frequencies. Some examples selected from the training set are provided here to illustrate our hypothesis.

(1) Example of redundant word problem

A redundant word is often unlikely to appear in that context. Moreover, removing this redundant word will create a higher-frequency bigram. For example,

[1] https://catalog.ldc.upenn.edu/LDC2006T13

[Sentence A2-0019-1]
可是 現在* 最近 我 工作 很 忙
(But now* recently my work is-very busy)

The word “現在” (now) has similar meaning with “最近” (recently), thus it is redundant. As an evidence, the bigram “現在+最近” is not collected in Google N-grams but the frequency of the bigram “可是+最近” is 218250.

(2) Example of missing word problem

Some examples provided in the training set are more like “a missing characters”, not “a missing word”. For example,

[Sentence A2-0026-1]
聽說 你 準備 開 一個 祝 會
(It-is-said-that you prepare to-have a wish* meeting*)
[Correct sentence]
聽說 你 準備 開 一個 慶祝會
(It-is-said-that you prepare to-have a celebration)

The character “慶” is missing, so the word “慶祝會” (celebration) cannot be correctly identified and is broken into two words “祝” and “會”. As an evidence, the bigram “祝+會” is not collected in Google N-grams.

(3) Example of word disorder problem

Word disorder means that order of the words should be re-arranged into a correct sentence. For example,

[Sentence A2-0027-1]
你* 很 久 以前 找 工作 很 幸* 苦
(You very long ago found jobs very lucky* difficult)
[Correct sentence]
很 久 以前 你 找 工作 很 辛苦
(Very long ago you found jobs very not-easily)

In Chinese, a long temporal phrase (“很久以前”, “very long ago” in this example) often appears in front of a complete sentence or, in another word, in front of a subject (“你”, “you” in this example). As an evidence, “以前+找” is not collected in Google N-grams but the frequency of the bigram “你+找” is 305477.

(4) Example of word selection problem

Word selection problem is that at least one word should be replaced with another, more appropriate word. For example,

[Sentence A2-0047-1]
我 真的 高興 你 找到 一* 新 的 工作 了
(I really am-happy you found one* new DE job LE)
[Correct sentence]
我 真的 高興 你 找到 一個 新 的 工作 了
(I really am-happy you found a new DE job LE)

(Note that DE and LE are function words without carrying much meaning)

When mentioning a countable noun in Chinese, quantifiers (量詞) should be used. For example, to say "a job", you use "一 個 工作" (one+GE+job), not "一 工作" (one+job). The character "個" (GE) in this example serves as a quantifier.

However, according to CNS14366, the Segmentation Standard for Chinese Natural Language Processing (中央標準局中文分詞標準, Huang *et al.*, 1997) in Taiwan, a number and a succeeding quantifier are segmented into two words, not grouped as one word. Such example is more like a missing problem rather than a word selection problem to us.

3. Error Detection Features

According to the observations described in Section 2, we defined several features to detect grammar errors as follows.

f_{bi-}:**number of infrequent bigrams** appearing in the sentence, where "infrequent bigram" is defined as a bigram NOT collected in Google N-grams. We expect that an erroneous sentence containing more infrequent bigrams. We are also interested to see if the number of infrequent bigrams is related to error types.

f_{POS-}: **number of infrequent POS bigrams** appearing in the sentence, where "infrequent POS bigrams" were trained from ASBC, a large POS-tagged corpus. Considering a POS bigram p_1p_2, if the probability P(p_2 | p_1) is less than 0.01, this bigram is an infrequent POS bigram.

f_{Nf-}: a Boolean feature denoting the **occurrence of a number without a succeeding quantifier**, where quantifiers are POS-tagged as Nf.

f_{stop}: a Boolean feature denoting the **occurrence of a stop POS bigram**. We defined a stop list of POS bigrams. POS bigrams in the stop list are:

- **VH + T**, a stative intransitive verb (mostly adjective in English) followed by a particle
- **Cbb + DE**, a correlative conjunction followed by a function word "的"
- **VC + Nd_DATE**, an active transitive verb followed by a date expression

f_D: a Boolean feature denoting the **occurrence of a key POS,** where key POS includes adverbs (D) and temporal nouns (Nd). Examples of disordered words often fall into these two POS classes.

$f_{bi=}$: **normalized number of infrequent bigrams,** i.e. f_{bi-} divided by the length of this sentence.

4. Run Submission and Results

Two runs were submitted to the CFL shared task this year. They were classification results from two different classifiers. System01 uses 5 features, f_{bi-}, f_{POS-}, f_{Nf-}, f_{stop}, and f_D. System02 also uses 5 features, $f_{bi=}$, f_{POS-}, f_{Nf-}, f_{stop}, and f_D. The only difference is the normalization of the first feature. Classifiers were trained by using LIBSVM (Chang and Lin, 2011).

Table 1 shows the performance of these two classifiers on training sets. The performances of the two systems are quite similar. Unfortunately, none of the classifiers can identify any word disorder case. System02 can correctly identify 4 word selection cases thus outperforms System01 a little.

Table 1: Performance in Training.

	System01			System02		
#	Prec.	Recl.	F-1	Prec.	Recl.	F-1
Redundant word	43.53	41.73	42.61	41.59	41.73	41.66
Missing word	43.71	75.51	55.43	44.71	75.51	56.22
Word disorder	0.00	0.00	0.00	0.00	0.00	0.00

Word selection	0.00	0.48	0.00	100.00	0.48	0.96
All	21.81	29.48	24.51	46.58	29.48	24.71

Table 2 shows the performance of two formal runs predicted by these two classifiers. The two systems have the same ability to detect errors. In fact, all sentences were predicted as "YES" but only half of them were correct. However, System02 achieved better performance in error-type classification thus outperforms System01 again.

Table 2: Performance of formal runs.

Submission	FP Rate	Detection Level				Identification Level			
		Acc.	Prec.	Recl.	F-1	Acc.	Prec.	Recl.	F-1
NTOU-Run1	100	50	50	100	66.67	16.00	24.24	32.00	27.59
NTOU-Run2	100	50	50	100	66.67	20.74	29.32	41.49	34.36

5. Conclusion

This paper describes our first Chinese grammar checker participating in CFL 2014. Six features related to grammatical errors were proposed, including numbers of infrequent word bigrams and POS bigrams. F-scores of formal runs were 66.67% in detection level and 34.36% in identification level. Normalized features seem outperform original numbers.

Because it was our first attempt to build a Chinese grammar checker, the performance was not satisfied. More studies and more features are needed for building a better system in the future.

References

Chang, C.-C. and C.-J. Lin (2011) LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, 27:1-27.

Huang, C.-R., K.-J. Chen, and Lili Chang (1997). Segmentation Standard for Chinese Natural Language Processing. *Computational Linguistics and Chinese Language Processing*, 2(2), 47-62.