Automatic Grammatical Error Detection for Chinese based on Conditional Random Field

Yajun Liu, Yingjie Han, Liyan Zhuo, Hongying Zan
Natural Language Processing Laboratory
College of Information and Engineering, Zhengzhou University, China
liuyanjun_gz@163.com, ieyjhan@zzu.edu.cn
1967331775@qq.com, iehyzan@zzu.edu.cn

Abstract

In the process of learning and using Chinese, foreigners may have grammatical errors due to negative migration of their native languages. Currently, the computer-oriented automatic detection method of grammatical errors is not mature enough. Based on the evaluating task ---- CGED2016, we select and analyze the classification model and design feature extraction method to obtain grammatical errors including Mission(M), Disorder(W), Selection (S) and Redundant (R) automatically. The experiment results based on the dynamic corpus of HSK show that the Chinese grammatical error automatic detection method, which uses CRF as classification model and n-gram as feature extraction method. It is simple and efficient which play a positive effect on the research of Chinese grammatical error automatic detection and also a supporting and guiding role in the teaching of Chinese as a foreign language.

1 Introduction

As China's status is improved and its influence in the world is increasing, more and more foreigners begin to learn Chinese. The HSK is an international standardized test for Chinese language proficiency of non-native speakers. From the analysis of the examination papers for many years, we can see that foreigners who study Chinese often make grammatical errors such as Mission(M), Disorder(W), Selection (S) and Redundant (R), owing to their language negative migration, over-generalization, teaching methods, learning strategies and other reasons.

Automatic detection of Chinese grammatical errors is really a challenge for many researchers. There is no space between word and word in Chinese corpus. If words in Chinese corpus are separated from each other, we can use combination of multiple features such as words, part of speech tagging (POS) and word frequency to detect grammatical errors, automatically. But errors of word segmentation and part of speech tagging will be accumulated in, and then have a negative effect on automatic detection of grammatical errors.

Examples are as follows:

The original sentence:

a)	他现在的工作是研究生物	His present job is studying biology
b)	他站起身来	He stands up
c)	他明天起身去北京	He leaves Beijing tomorrow

After word segmentation:

a)	他/现在/的/工作/是/研究/生物	His / present / job / is / studying / biological
b)	他站/起 /身/来	He stands / up
c)	他/明天/起身/去 /北京	He / leaves for / Beijing / tomorrow

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications, pages 57–62, Osaka, Japan, December 12 2016.

In example a, the "研究生物 (study biology)" will arise segmentation ambiguity, in example b and c, "起身 (get up)" has two different way of divisions which has a bad effect on automatic detection of grammatical errors. In this respect, current automatic detection methods have poor performance, we need actively explore effective automatic detection methods which can help reduce workload of artificial detection and play a positive guiding role in teaching Chinese as a foreign language. With some grammatical errors and error cause found by these methods, teaching Chinese as a foreign language will be well guided.

For many researchers, CGED evaluating task provides a platform to study automatic detection of Chinese grammatical errors. CGED 2016 evaluating task divides the Chinese grammatical errors into four categories: Mission(M), Disorder(W), Selection (S) and Redundant (R), and includes three tasks such as Detection Level, Identification Level and Position Level.

In order to achieve Chinese grammatical error automatic detection, we first consider the problem of Chinese grammatical errors as a classification problem, and then use rule-based method, statistical learning method or the fusion of multiple classification methods. Through analysis and comparison, we use CRF to complete three tasks including Detection Level, Identification Level and Position Level.

The rest of this paper is organized as follows: Section 2 briefly introduces related work in this field. Section 3 introduces the statistical learning method CRF and its tools. Section 4 discusses the realization of Chinese grammatical error automatic detection which includes data preprocessing, data feature extraction, model selection and result analysis. Finally, conclusion and prospects are arranged.

2 Related work

In the aspect of automatic detection of grammatical errors, the study of English is more deep. Anubhav Gupta (2014) proposed a rule-based approach that relies on the difference in the output of two POS taggers, to detect verb forms, lexical and spelling errors, but fuzzy or erroneous input of the POS tagger could result in an erroneous output. In order to solve context-sensitive spelling correction, an algorithm combining Winnow variable and weighted majority voting was proposed by Andrew R. Golding (1999), but in this way we need to improve the adaptability of the algorithm to unfamiliar test sets. Anoop Kunchukuttan (2014) proposed two enhancements based on statistical machine translation for all types of errors. Although it is possible to use a simple set of methods to increase recall rate, it also leads to a decrease in precision.

Relevant works related to Chinese grammatical error detection are much less compared with that of English. Chi Hsin Yu and Hsin-Hsi (2012) proposed a classifier based on CRF model to detect Chinese text disorder. Shuk-Man Cheng (2014) proposed a support vector machine model to further explore the problem of word order reordering. Yang Xiang and Xiaolong Wang (2015) used an ensemble learning method which learns and trains the corpus to identity the grammatical errors and error types, but the detection of the error location is not ideal. Xiupeng Wu and Peijie Huang (2015) used a hybrid model that integrates rule-based and N-gram statistical method to detect the Chinese grammatical errors, which can identify the error types well and point out error position, but rules are needed summarizing manually. Lung-Hao Lee and Liang-Chih Yu (2014) introduced a sentence-level detection system that integrates multiple rules and N-gram statistical features. Generally speaking, relevant rule are needed in most of the Chinese grammatical error automatic detection summarizing manually, and these existing methods on the error position are not ideal at present.

3 CRF

3.1 CRF model

CRF (Random Field Conditional) is a distinguished indirect graph model. In an indirect graph $G = (V, E)$, where V be the set of end point, E be the set of indirect edges, $Y = \{Y_v | v \in V\}$, that is, each node in V corresponds to a random variable which is in the range of possible tag set {y}. If we observe the sequence X as a condition and each random variable Y_v satisfies the following Markov characteristic:

$$p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v) \quad (1)$$

where denotes that two nodes are adjacent in graph G, then (X, Y) is a conditional random field. Model diagram is shown in Figure 1.

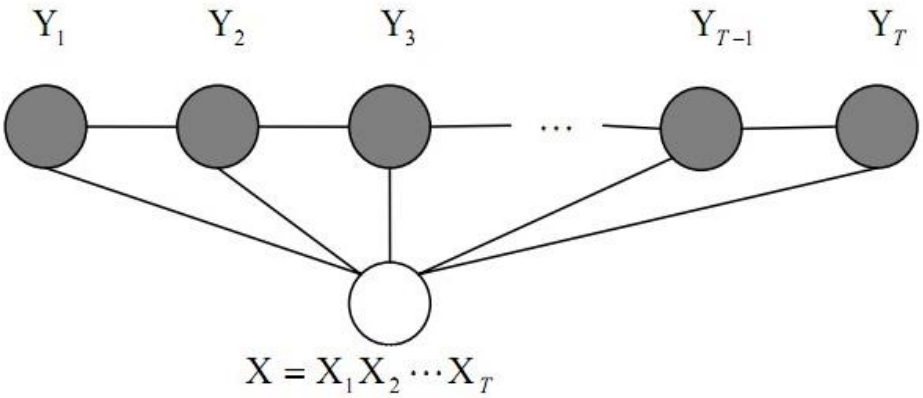

Fig.1 Schematic diagram of conditional random field model

For the first time, Lafferty introduced CRF into natural language processing, and the choice of CRF potential function is greatly influenced by the maximum entropy model, and first-order chain structure is applied to construct the CRF model. In graph $G = (V, E)$, the largest group which is the edge in graph G contains only two adjacent nodes.

We define the form of each potential function as follows:

$$\emptyset_{y_c}(y_c) = \exp(\sum_k \lambda_k f_k(c, y|c, x)) \quad (2)$$

Where $y|c$ denotes the random variable corresponding to the node in the C group, $f_k(c, y|c, x)$ is a Boolean feature function, then p (y | x) is:

$$p(y|x) = \frac{1}{Z(x)} \exp(\sum_{c \epsilon C} \sum_k \lambda_k f_k(c, y_c, x)) \quad (3)$$

where $Z_{(x)}$ is the normalization factor.

$$Z_{(x)} = \sum_y \exp(\sum_{c \epsilon C} \sum_k \lambda_k f_k(c, y_c, x)) \quad (4)$$

3.2 CRF ++ tool selection

CRF-based tools are currently available such as crf++, flexcrf, pocket crf.

First of all, crf++ is the first order crf, flexcrf is the second order crf, because n-order crf training time required (p is the number of markers, T is the first order crf training time, N is the order), so compared with crf++, flexcrf needs more training time.

Second, pocket crf does not provide a command line, and there is only one example that shows how to complete the training and testing, and pocket crf does not identify the space, so pocket crf string input file must be strictly separated by 0x09. In contrast, crf++ has a command line, and can ignore all spaces and 0x09 between the columns. So in this experiment, we use the CRF++ tool 0.58 version[1].

4 CRF-based automatic detection of Chinese grammatical errors

According to the above discussion, we choose CRF as the statistical learning method and CRF++ as the tool of automatic detection. Through data preprocessing, feature selection, training and cross validation, the automatic detection result of test data, result analysis are given.

4.1 Data preprocessing

The main work of data preprocessing is to preprocessing the training data set of CGED 2016, and then the training corpus format is adjusted to the input format required by CRF++. The correct sentences and the wrong sentences are extracted from the corpus according to artificial annotation, and then the corpus are automatically marked according to the wrong position and the wrong type.

Example of original training corpus:

<DOC>

<TEXT id="20040510952310036O_2_6x2">

[1] https://code.google.com/archive/p/crfpp/downloads.

妈妈对爸爸劝戒烟的原因就是我的健康。非吸烟者比吸烟者得病率更高，这个所有的人知道。
</TEXT>
<CORRECTION>
妈妈劝爸爸戒烟的理由就是我的健康。非吸烟者比吸烟者得病率更高，这个所有的人都知道。
</CORRECTION>
<ERROR start_off="3" end_off="3" type="R"></ERROR>
<ERROR start_off="4" end_off="6" type="W"></ERROR>
<ERROR start_off="10" end_off="11" type="S"></ERROR>
<ERROR start_off="39" end_off="39" type="M"></ERROR>
</DOC>

Preprocessed training corpus:

妈/C 妈/C 对/R 爸/W 爸/W 劝/W 戒/C 烟/C 的/C 原/S 因/S 就/C 是/C 我/C 的/C 健/C 康/C 。/C 非/C 吸/C 烟/C 者/C 比/C 吸/C 烟/C 者/C 得/C 病/C 率/C 更/C 高/C ，/C 这/C 个/C 所/C 有/C 的/C 人/C 知/M 道/C 。/C

妈/C 妈/C 劝/C 爸/C 爸/C 戒/C 烟/C 的/C 理/C 由/C 就/C 是/C 我/C 的/C 健/C 康/C 。/C 非/C 吸/C 烟/C 者/C 比/C 吸/C 烟/C 者/C 得/C 病/C 率/C 更/C 高/C ，/C 这/C 个/C 所/C 有/C 的/C 人/C 都/C 知/C 道/C 。/C

Note: 1. /C, /M, /W, /S, /R represent Correct, Mission, Disorder, Selection and Redundant.
2. punctuation, letters, etc. are also followed by the corresponding label.

4.2 Feature Selection

In practice, the feature selection directly affects the performance of the model. The more features are selected, the more time is required when the feature is analyzed and the model is trained, may be the more complex the model is. Therefore, selecting better features not only can simplify the model, but also can reduce the running time. In the statistical machine learning method CRF, this experiment adopts feature length of 5 and 7, then uses bi-gram and tri-gram model to extract features. We conduct cross validation for two kinds of sequence length features, and results are shown in the table1.

Sequence length		5	7
False Positive Rate		0.0518	0.0811
Detection Level	Precision	0.7192	0.6623
	Recall	0.1284	0.1489
	F1-Score	0.2179	0.2431
Identification Level	Precision	0.6142	0.5588
	Recall	0.0798	0.0962
	F1-Score	0.1413	0.1641
Position Level	Precision	0.3981	0.4286
	Recall	0.0332	0.0569
	F1-Score	0.0612	0.1005

Table 1 cross validation results

Through the comparison of Precision, Recall and F1-Score, False Positive Rate has an increase of 2.93% when the sequence length is 7, but there are different levels of promotion in the recall rate of Detection Level, Identification Level and Position Level, F1 is also better than the sequence length 5.

4.3 Results and analysis

The results of the closed test with the training data of CGED2016 are shown in Table 2. Considering the influence of the size of the training data on the model, we add 2015 TOCFL training data to 2016 HSK training data for closed test, the results are shown in Table 2.

Training corpus		2016 HSK data	2015 TOCFL and 2016 HSK data
False Positive Rate		0.0759	0.0596
Detection Level	Precision	0.7055	0.6515
	Recall	0.1323	0.1361
	F1-Score	0.2227	0.2252
Identification Level	Precision	0.6258	0.5516
	Recall	0.0923	0.0896
	F1-Score	0.1609	0.1541
Position Level	Precision	0.4381	0.3414
	Recall	0.0430	0.0377
	F1-Score	0.0784	0.0680

Table 2 closed test results

We compare and analyze the closed results, and then select the HSK data of 2016 and TOCFL data of 2015 as training data, as shown in Table 3.

Training corpus	Correct	Error				Sum
		R	S	M	W	
2016 HSK	10072	5532	10942	6619	1691	20144
2015 TOCFL	2205	430	849	620	306	4410

Table 3 Training data distribution table

Note: 1. Each of these error statements may contain multiple types of errors or include multiple identical types of errors.
2. 2015 TOCFL corpus is converted to HSK for use

In the three results we submitted, SKY_Run2.txt and SKY_Run3.txt are generated by model which is strained by feature template with the sequence length of 5 and 7. These two submitted results have best performance on all three tasks, especially False Positive Rate, Accuracy and Precision indicators, but work badly in recall rate. Our team achieved the lowest false positive rate of 0.0481 in 2016 CGED evaluating task.

The evaluation results are as follows:

Submission results		SKY_Run1.txt	SKY_Run2.txt	SKY_Run3.txt
False Positive Rate		0.0695	**0.0481**	0.0559
Detection Level	Accuracy	0.6523	0.6579	**0.6659**
	Precision	0.8326	**0.8746**	0.8652
	Recall	0.3614	0.3505	0.3750
	F1-Score	0.5040	0.5005	0.5232
Identification Level	Accuracy	0.6605	0.6765	**0.6849**
	Precision	0.8235	**0.8821**	0.8744
	Recall	0.2732	0.2972	0.3815
	F1-Score	0.4132	0.4446	0.4669
Position Level	Accuracy	0.6073	0.6376	**0.6477**
	Precision	0.6153	0.7054	**0.7144**
	Recall	0.1783	0.2217	0.2430
	F1-Score	0.2765	0.3373	0.3627

Table 4 Evaluation results

From the analysis of the results, we can see that feature templates with two kinds of sequence length use bigram and trigram models to extract features and select more features, thus greatly improve Precision, but have a serious impact on recall rate.

As for Position Level task, SKY_Run3.txt plays better than SKY_Run2.txt, and has good performance on Accuracy, Precision, Recall and F1-Score indicators, so feature template with sequence length

7 plays better. When the length of the sequence becomes longer, the effect of position level task is better. But if the length is too long, the learning process will become difficult and the model will become more complex. Compared with the first two tasks, the results in Position Level are not ideal. Since the open source tool based on the statistical machine learning method CRF only supports chained sequences, when the sequence length is 5 and 7, the long sentences can't be analyzed on the whole, which affects the automatic detection of Chinese grammatical errors.

5 Conclusion and prospect

In this paper, we use statistical learning method CRF and n-gram feature extraction method to achieve Chinese grammatical error automatic detection. It can be seen from the evaluation results that the CRF model has a good performance in the automatic detection of Chinese grammatical errors, especially False Positive Rate and Precision. But in the overall quantity of Chinese grammatical errors, the errors that are detected are too few, which affects the overall performance.

In general, CRF has great potential in automatic detection of Chinese grammatical errors. Compared with HMM (Hidden Markov Model), it has no strict independence assumption, and its feature design is flexible. Compared with the regular method, it can predict more flexible grammatical errors. It is also simpler than multiple classifier fusion methods. The only thing we need to do is to manually mark the corpus for CRF learning. In the following work, we will collect more corpus of Chinese grammatical errors to improve the performance of the model, and we will also consider mutual information and other methods for feature extraction.

Reference

Chang, Ru-Yng, 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 (TALIP), 11(1), 3.

Cheng, Shuk-Man, Chi-Hsin Yu, and Hsin-Hsi Chen. 2014. Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Language Learners. In COLING (pp. 279-289).

Della Pietra, Stephen, Vincent Della Pietra, and John Lafferty. 1997. Inducing features of random fields. IEEE transactions on pattern analysis and machine intelligence, 19(4), 380-393.

Golding, Andrew R., and Dan Roth. 1999. A winnow-based approach to context-sensitive spelling correction. Machine learning, 34(1-3), 107-130.

Gupta, Anubhav. 2014. Grammatical Error Detection and Correction Using Tagger Disagreement. CoNLL-2014, 21860(26282), 49.

Kunchukuttan, Anoop, Sriram Chaudhury, and Pushpak Bhattacharyya. 2014, May. Tuning a Grammar Correction System for Increased Precision. In CoNLL Shared Task (pp. 60-64).

Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001, June. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the eighteenth international conference on machine learning, ICML (Vol. 1, pp. 282-289).

Lee, Lung-Hao, Liang-Chih Yu, and Li-Ping Chang. 2015. Overview of the NLP-TEA 2015 Shared Task for Chinese Grammatical Error Diagnosis. ACL-IJCNLP 2015, 1.

Lee, Lung-Hao, et al. 2014, July. A Sentence Judgment System for Grammatical Error Detection. In COLING (Demos) (pp. 67-70).

Ng, Hwee Tou, et al. 2014, May. The CoNLL-2014 Shared Task on Grammatical Error Correction. In CoNLL Shared Task (pp. 1-14).

Sha, Fei, and Fernando Pereira. 2003, May. Shallow parsing with conditional random fields. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 134-141). Association for Computational Linguistics.

Xiang, Yang, et al. 2015. Chinese grammatical error diagnosis using ensemble learning. ACL-IJCNLP 2015, 99.

Yu, Liang-Chih, Lung-Hao Lee, and Li-Ping Chang. 2014, November. Overview of grammatical error diagnosis for learning Chinese as a foreign language. In Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA'14), Nara, Japan (pp. 42-47).