Condition Random Fields-based Grammatical Error Detection for Chinese as Second Language

Jui-Feng Yeh, Chan-Kun Yeh, Kai-Hsiang Yu, Ya-Ting Li, Wan-Ling Tsai
Department of Computer Science and Information Engineering, National Chiayi University
No.300 Syuefu Rd., Chiayi City 60004, Taiwan (R.O.C.)
{ralph, s1030484, s1030495, s1013037, s1013048 }@mail.ncyu.edu.tw

Abstract

The foreign learners are not easy to learn Chinese as a second language. Because there are many special rules different from other languages in Chinese. When the people learn Chinese as a foreign language usually make some grammatical errors, such as missing, redundant, selection and disorder. In this paper, we proposed the conditional random fields (CRFs) to detect the grammatical errors. The features based on statistical word and part-of-speech (POS) pattern were adopted here. The relationships between words by part-of-speech are helpful for Chinese grammatical error detection. Finally, we according to CRF determined which error types in sentences. According to the observation of experimental results, the performance of the proposed model is acceptable in precision and recall rates.

1 Introduction

As the world globalize, travel around the world is quicker than before. With the growth of Chinese market and more and more china town. There are more than 1.3 billion people who speak Chinese. That means there is one speak Chinese out of every five people. Chinese is the most spoken language in the world. Sell products to the Chinese people, study and travel around Asia is much easier than before. To speak with foreigners and trade with foreigners we have to understand their language first. So we believe that learning Chinese is important now.

To learn Chinese as second language we have to know not only pronunciations and glyph of the word, but also grammar and the part of speech of Chinese. For example, there are eight parts of speech (nouns, pronouns, adjectives, verbs, adverbs, prepositions, conjunctions, and interjections) in English. But in Chinese there are ten parts of speech (nouns, adjectives, verbs, adverbs, pronouns, interjections, prepositions, conjunctions, auxiliary words, and quantifiers). Nouns indicate the names of people or things, they can be further divided into four sub sorts, proper nouns, common nouns, abstract nouns, time nouns, place nouns. Adjective show the quality or forms of people or things, or the state of action or behavior. Verbs indicate the behaviors, actions or changes of people or things. They have several subsidiary categories: modal verbs, tendency verbs and deciding verb. Adverb is used in front of verbs or adjectives to show degree, extent, time or negation. Pronoun is replace nouns or numerals. Preposition introduces nouns, pronouns or other linguistic units to verbs or adjectives and show the relationship between time, space, objects or methods. Conjunction connects words, phrases or sentences.

With Chinese become more popular. But it is not an easy language to learn, you will make a fool of yourself even if just one word mistake. More and more people pay their attention to Chinese grammar error. We may not write the right Chinese sentence all the time. Sometimes we make some mistakes such as, overuse preposition, overuse of "a/an", semantic overlap and quantifier error. In the past, the way to detect the grammar mistakes is extremely inefficient. People usually correct grammar mistakes by manual work.

In recent year, there are many researches about Chinese grammar. There are few papers help us as reference. Li et al. (2012) proposed a hierarchical structure of dependency relations based on CDG for Chinese, in which the constraints have been partitioned into three hierarchies: in-the-phrase, between-the-phrase and between-the simple-sentence. And Li, Z., Zhang et al (2014) proposed to integrate the POS (part-of-speech) tagging and parsing can reduce the complexity and improve the accuracy of parsing. Jiang et al. (2012) divided

Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 105–110,
Beijing, China, July 31, 2015.

Table 1. Example of error types.

Error Types	Error Sentence	Correct Sentence
Missing Error	我(Nh) 送(VD) 你(Nh) 那裡(D)	我(Nh) 送(VD) 你(Nh) ***到(VCL)*** 那裡(Ncd)
Redundant Error	他(Nh) 是(SHI) 我(Nh) ***的(DE)*** 以前(Nd) 的(DE) 室友(Na)	他(Nh) 是(SHI) 我(Nh) 以前(Nd) 的(DE) 室友(Na)
Selection Error	吳(Nb) 先生(Na) 是(SHI) 修理(VC) 腳踏車(Na) 的(DE) ***拿手(Nv)***	吳(Nb) 先生(Na) 是(SHI) 修理(VC) 腳踏車(Na) 的(DE) ***好手(Na)***
Disorder Error	所以(Cbb) 我(Nh) 不會(D) 讓(VL) ***失望(VH) 她(Nh)***	所以(Cbb) 我(Nh) 不會(D) 讓(VL) ***她(Nh) 失望(VH)***

Chinese grammar into three groups: morphology of content words, morphology of empty words and syntax. And each group is subdivided. Jiang et al. (2012) proposed the effectiveness of the XML and the goodness of XML structure these two parts compose XML syntax check. They improved local tree grammar of the XML document type definition, and then do XML validity checking in the grammatical structure based on the document type definition. Rozovskaya et al. (2012) presented a linguistically-motivated, holistic framework for correcting grammatical verb mistakes. Describe and evaluate several methods of selecting verb candidates, an algorithm for determining the verb type, and a type-driven verb error correction system. And they gloss a subset of the FCE dataset with gold verb candidates and gold verb type. Lee et al. (2014) develops a sentence judgment system using both rule-based and n-gram statistical methods to detect grammatical errors in sentences written by CFL learners. Users can input Chinese sentences into the proposed system to check for possible grammatical errors. Wu et al. (2012) through examining the collected English-to-Chinese corpus composed of error sentences, in contrast to the errors commonly made by learners of ESL such as the use of articles and prepositions, we found that learners of Chinese whose L1 is English tend to produce sentences with word order, lexical choice, redundancy, and omission errors. And they present an approach using the proposed Relative Position Language Model (RP) and Parse Template Language Model (PT) to deal with the error correction problem, which is especially suitable for the correction of word order errors that comprise about one third of the errors made by learners of Chinese as a Second Language. The four error types considered for correction in their paper are errors of Lexical Choice, Redundancy, Omission, and Word Order. In this paper, we show the four error types in Table 1. Islam et al. (2010) use the Google n-gram data set in a back-off fashion. And it increases the performance of the method. Their method can be applied to other languages for which Google n-grams are available. Sun, X., & Nan, X. (2010) defined the phrase's format to "Modifier + head + complement".

In our method, we tag some labels such as POS and binary variables in the sentences. In the section II, we described the models how to train the corpus by CRF. And show the experiment in the section III.

2 Method

In this section, the architecture of our system is illustrated in Figure 1. Then we will describe the grammatical error detection using CRF in the section 2.2. And the procedures distinguish into two parts: training phase and test phase.

Table 2. Punctuation tagged by CKIP Autotag.

COLONCATEGORY	(：)
COMMACATEGORY	(，)
DASHCATEGORY	(－)
ETCCATEGORY	(…)
EXCLAMATIONCATEGORY	(！)
PARENTHESISCATEGORY	(（ ）)
PAUSECATEGORY	(、)
PERIODCATEGORY	(。)
QUESTIONCATEGORY	(？)
SEMICOLONCATEGORY	(；)
SPCHANGECATEGORY	(‖)

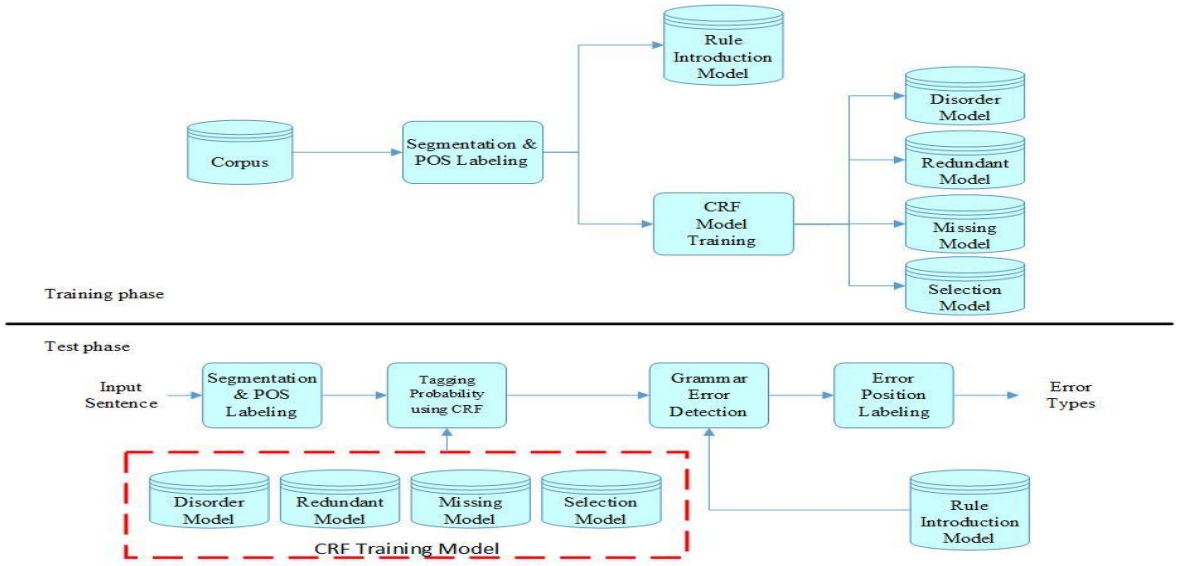

Figure 1. The Architecture of Our System

2.1 Data Preprocessing

CKIP Autotag is a word segment system which made by Taiwan Academia Sinica. CKIP Autotag can segment a long Chinese sentence into Chinese word. And then tag the punctuation (punctuation are listed in Table 2) and the POS to the word. This system classifies Chinese word into 47 different POS.

We use this system to chunking words and tagging the POS on the sentence. We survey the grammar of Chinese sentence by CKIP Autotag. And observe the relation between the words by the POS.

2.2 Condition Random Fields

Conditional random fields (CRFs) is a class of statistical modelling method that is generally applied in machine learning and pattern recognition, where they are used for structured prediction. It was an extension of both Maximum Entropy Model (MEMs) and Hidden Markov Models (HMMs) that was firstly introduced by Lafferty et al., 2001.Whereas an ordinary classifier predicts a label for a single sample without regard to adjacent samples. A CRF can take context into account. It's a discriminative of undirected probabilistic graphical model. It is used to encode known relationships between observations and construct consistent interpretations. Conditional random field defined conditional probability distribution P(Y|X) of given sequence given input sentence. Y is the "class label" sequence and X denotes as the observation word sequence.

A common used special case of CRFs is linear chain, which has a distribution of:

$$P_\Lambda(y|x) = \frac{\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1},y_t,x,t)\right)}{Z_x} \quad (1)$$

Where $f_k(y_{t-1}, y_t, x, t)$ is usually an indicator function. λ_k is the learned weight of the feature and Z_x is the normalization factor that sum the probability of all state sequences. The feature functions can measure any aspect of a state transition, y_{t-1} *to* y_t and the entire observation sequence, x, centered at the current time step, t.

Here we use three conditional random field models to calculate the conditional probability of the missing sentences, redundant sentences, disorder sentences and error selection sentences.

In training phase, we give the matrix {Word, POS, TAG} to denote the sentence of the words in the train set. Such as {去, VCL, T} or {去, D, F}, the word "去(go)" has many part-of-speech in different sentences. The tag "T" means correct word in current sentence and tag "F" means error word in current sentence. Then we use this training data to generate the model by Conditional random fields.

In testing phase, we segment and tag POS labeling by CKIP Autotag. Then we also use the matrix {Word, POS} to denote the words. After preprocessing, we can get the tag's probability of testing words by our training models using CRF++.

For example, input the sentence of "但是(but) 駕駛(driver) 都(neither) 裝作(pretend) 沒(not)

看 到 (see) 或 者 (or) 聽 到 (hear) 我 (me) 了 (interjection)". There are some probabilities from different models, "Missing 0.872773, Redundant 0.465524, Selection 0.832839" and judge it is a redundant error. So we found every word's probability in this sentence. The probability of words are show in Table 3. And we found the error word is "了".

Table 3. Probability of Words in the Sentence.

Word	POS	Probability
但是	Cbb	T/0.963663
駕駛	VC	T/0.986188
都	D	T/0.975163
裝作	VF	T/0.970347
沒	D	T/0.962676
看到	VE	T/0.984734
或者	Caa	T/0.953170
聽到	VE	T/0.988986
我	Nh	T/0.997955
了	T	*F/0.579991*

According the probability of tagging, we can determine what type's error and speculate the position in the sentence.

2.3 Rule Induction

There are many special cases of selection error types in Chinese. Such as quantifier is one case of all. In English, we usually use "a" or "an" to denote quantifier.

But Chinese needs more different quantifiers then the other language. In many cases, Chinese use '個' as a quantifier. There are more times we do not use '個' as a quantifier. About quantifier of human we should use '位' or '個'. About quantifier of animals we should use '隻', '匹', '頭', or '條'. About quantifier of things we should use '件'. About quantifier of buildings we should use '座' or '棟'. About quantifier of transportations we should use '臺', '輛', '架' or '艘' etc.

We also focused on finding the ordering type of the wrong words. There are some rules which we follow to finding ordering error.

- Behind the words "把 (let)" is connected the POS 'Nh' or 'Na' or 'Nep'.
- Behind the POS 'VA' is connected the word "跟(with)", and the POS 'Nh' or 'Na' also is connected behind the words "跟(with)".
- Behind the words "應該(maybe)" or "好像(like)" or "到底(at last)" is connected the POS 'Nh' or 'Na'.
- Behind the word "已經(already)" is connected the POS 'Neqa' or 'Neu', and the POS 'P' or 'Na' or 'VA' is connected behind the POS 'Neqa' or 'Neu'.

Above those rules, we can enhance our method during the detection grammatical errors.

3 Result

To evaluate the performance of our system, we used three parameters: precision, recall and f-score.

Precision is the fraction of retrieved documents that are relevant to the query.

$$\text{Precision} = \frac{tp}{tp + fp}$$

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{Recall} = \frac{tp}{tp + fn}$$

F1-Score is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score.

$$\text{F1} - \text{score} = 2 * \frac{Presision * Recall}{Presision + Recall}$$

The comparative study of all the three cases has done. And a result of these cases are given in the Table 4 and Table 5 with the NLP-TEA 2014 dataset.

In this paper, we collect 2,212 sentences in training dataset. And it contains 622 sentences of missing, 435 sentences of redundant, 849 sentences of selection and 306 sentences of disorder. Then we use two dataset 1,750 sentences from NLP-TEA 2014 and 1,000 sentences from NLP-TEA 2015.

We can find only use CRF it can't find many error but its precision is better. Then add the rule induction can promote the recall means we can find more error from test data. Although its precision is reduced.

Table 4. Detection level

Method	Precision	Recall	F1
CRF	0.6863	0.2000	0.3097
CRF + Rule Induction	0.5257	0.4674	0.4949

Table 5. Identification level

Method	Precision	Recall	F1-Score
CRF	0.5897	0.1314	0.2150
CRF + Rule Induction	0.3549	0.2320	0.2806

Table 6, Table 7, and Table 8 are the performance with the NLP-TEA 2015 dataset and compare the other team

Table 6. Detection level

	Accuracy	Precision	Recall	F1
NCYU	**0.607**	0.6112	0.588	0.5994
CYUT	0.579	0.7453	0.240	0.3631
NTOU	0.531	0.5164	0.976	0.6754

Table 7. Identification level

	Accuracy	Precision	Recall	F1
NCYU	0.463	0.4451	0.300	**0.3584**
CYUT	0.525	0.6168	0.132	0.2175
NTOU	0.225	0.2848	0.364	0.3196

Table 8. Position level

	Accuracy	Precision	Recall	F1
NCYU	0.374	0.2460	0.122	0.1631
CYUT	0.505	0.5287	0.092	0.1567
NTOU	0.123	0.1490	0.160	0.1543

In detection level (see the Table 6.), our recall is better than CYUT's method. It means we can find more error in dataset. And our precision is better than NTOU's method. It means our find correct error rate is better, although we find error quantity less than NTOU.

In identification level (Table 7.), it show who can find most error and error type is correct. In our method, our recall is nearly NTOU's method, it means we find more correct error type than CYUT's method. But our precision is better than NTOU's method. And our F1-Score is the best in this level.

In position level (Table 8.), our method's precision and recall are between the CYUT's method and NTOU's method. It means our method is not illustrious in this level. We consider the reasons are our correction is not enough standard.

4 Conclusion

In this paper, we present a method using conditional random field model for predicting the grammatical error diagnosis for learning Chinese.

After observe the experiment results, our method is acceptable in NLP-TEA 2015. We believe this system is feasible. This system is useful for a foreign who learn Chinese as a second language. Even the people who use Chinese as a first language might use the wrong grammars.

There are some issues should be revise. First, the CRF models can be improved in some ways, such as words tagging or using the parsing tree. Second, increase the ranking mechanism to find the optimal words to correct the sentence.

In the future, we will pay attention to improve the precision and recall rates in this system. And let it can automatic correct the error if the people input the sentences.

Reference

Islam, A., & Inkpen, D. (2010, August). An unsupervised approach to preposition error correction. In Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on (pp. 1-4). IEEE.

Jiang, Y., Wang, T., Lin, T., Wang, F., Cheng, W., Liu, X., ... & Zhang, W. (2012, June). A rule based Chinese spelling and grammar detection system utility. In System Science and Engineering (ICSSE), 2012 International Conference on (pp. 437-440). IEEE.

Jiang, Y., Zhou, Z., Wan, L., Li, M., Zhao, W., Jing, M., & Liu, X. (2012, October). Cross sentence oriented complicated Chinese grammar proofreading method and practice. In Information Management, Innovation Management and Industrial Engineering (ICIII), 2012 International Conference on (Vol. 3, pp. 254-258). IEEE.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional Random Field: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the International Conference on Machine Learning.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In Proceedings of the 3rd Workshop on Very Large

Corpora, pages 82-94. Nocedal, J., and Wright, S. 1999. Numerical optimization. Springer.

Lee, L. H., Yu, L. C., Lee, K. C., Tseng, Y. H., Chang, L. P., & Chen, H. H. (2014). A Sentence Judgment System for Grammatical Error Detection. COLING 2014, 67.

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), Nara, Japan, 30 November, 2014, pp. 42-47.

Li, P., Liao, L., & Li, X. (2012, July). A hierarchy-based constraint dependency grammar parsing for Chinese. In Audio, Language and Image Processing (ICALIP), 2012 International Conference on (pp. 328-332). IEEE.

Li, Z., Zhang, M., Che, W., Liu, T., & Chen, W. (2014). Joint Optimization for Chinese POS Tagging and Dependency Parsing. Audio, Speech, and Language Processing, IEEE/ACM Transactions on, 22(1), 274-286.

Rozovskaya, A., Roth, D., & Srikumar, V. (2014, April). Correcting grammatical verb errors. In Proceedings of EACL.

Chang, R. Y., Wu, C. H., & Prasetyo, P. K. (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.

Sun, X., & Nan, X. (2010, August). Chinese base phrases chunking based on latent semi-CRF model. In Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on (pp. 1-7). IEEE.

Wu, C. H., Liu, C. H., Harris, M., & Yu, L. C. (2010). Sentence correction incorporating relative position and parse template language models. Audio, Speech, and Language Processing, IEEE Transactions on, 18(6), 1170-1181.

Yu, C. H., & Chen, H. H. (2012). Detecting Word Ordering Errors in Chinese Sentences for Learning Chinese as a Foreign Language. In COLING (pp. 3003-3018).

Academia Sinica CKIP. http://ckipsvr.iis.sinica.edu.tw/

CRF++: Yet Another CRF toolkit http://taku910.github.io/crfpp/