Chinese Grammatical Error Diagnosis by Conditional Random Fields

Po-Lin Chen, Shih-Hung Wu*
Chaoyang University of Technology/
Wufeng, Taichung, Taiwan, ROC.
streetcatsky@gmail.com,
*contact author: shwu@cyut.edu.tw

Liang-Pu Chen, Ping-Che Yang, Ren-Dar Yang
IDEAS, Institute for Information Industry/
Taipei, Taiwan, ROC.
{eit, maciaclark, rdyang}@iii.org.tw

Abstract

This paper reports how to build a Chinese Grammatical Error Diagnosis system based on the conditional random fields (CRF). The system can find four types of grammatical errors in learners' essays. The four types or errors are redundant words, missing words, bad word selection, and disorder words. Our system presents the best false positive rate in 2015 NLP-TEA-2 CGED shared task, and also the best precision rate in three diagnosis levels.

1 Introduction

Learning Chinese as foreign language is on the rising trend. Since Chinese has its own unique grammar, it is hard for a foreign learner to write a correct sentence. A computer system that can diagnose the grammatical errors will help the learners to learn Chinese fast (Yu et al., 2014; Wu et al., 2010;Yeh et al., 2014;Chang et al., 2014).

In the NLP-TEA-2 CGED shared task data set, there are four types of errors in the leaners' sentences: Redundant, Selection, Disorder, and Missing. The research goal is to build a system that can detect the errors, identify the type of the error, and point out the position of the error in the sentence.

2 Methodology

Our system is based on the conditional random field (CRF) (Lafferty, 2001). CRF has been used in many natural language processing applications, such as named entity recognition, word segmentation, information extraction, and parsing (Wu and Hsieh, 2012). For different task, it requires different feature set and different labeled training data. The CRF can be regarded as a sequential labeling tagger. Given a sequence data X, the CRF can generate the corresponding label sequence Y, based on the trained model. Each label Y is taken from a specific tag set, which needs to be defined in different task. How to define and interpret the label is a task-depended work for the developers.

Mathematically, the model can be defined as:

$$P(Y|X) = \frac{1}{Z(X)} \exp(\sum_k \lambda_k f_k) \quad (1)$$

where Z(X) is the normalization factor, f_k is a set of features, λ_k is the corresponding weight. In this task, X is the input sentence, and Y is the corresponding error type label. We define the tag set as: {O, R, M, S, D}, corresponding to no error, redundant, missing, selection, and disorder respectively. Figure 1 shows a snapshot of our working file. The first column is the input sentence X, and the third column is the labeled tag sequence Y. Note that the second column is the Part-of-speech (POS) of the word in the first column. The combination of words and the POSs will be the features in our system. The POS set used in our system is listed in Table 1, which is a simplified POS set provided by CKIP[1].

Figure 2 (at the end of the paper) shows the framework of the proposed system. The system is built based on the CRF++, a linear-chain CRF model software, developed by Kudo[2].

```
可是 C O
有 Vt O
一點 DET O
冷 Vi O
了 T R

你 N O
的 T R
過年 Vi O
呢 T O
```

Figure 1: A snapshot of our CRF sequential labeling working file

[1] http://ckipsvr.iis.sinica.edu.tw/

[2] http://crfpp.sourceforge.net/index.html

Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 7–14,
Beijing, China, July 31, 2015.

Simplified CKIP POS	Corresponding CKIP POS
A	非謂形容詞
C	對等連接詞，如：和、跟
	關聯連接詞
POST	連接詞，如：等等
	連接詞，如：的話
	後置數量定詞
	後置詞
ADV	數量副詞
	動詞前程度副詞
	動詞後程度副詞
	句副詞
	副詞
ASP	時態標記
N	普通名詞
	專有名稱
	地方詞
	位置詞
	時間詞
	代名詞
DET	數詞定詞.
	特指定詞
	指代定詞
	數量定詞
M	量詞
Nv	名物化動詞
T	感嘆詞
	語助詞
	的, 之, 得, 地
P	介詞
Vi	動作不及物動詞
	動作類及物動詞
	狀態不及物動詞
	狀態類及物動詞
Vt	動作使動動詞
	動作及物動詞
	動作接地方賓語動詞
	雙賓動詞
	動作句賓動詞
	動作謂賓動詞
	分類動詞
	狀態使動動詞
	狀態及物動詞
	狀態句賓動詞
	狀態謂賓動詞
	有
	是

Table 1: Simplified CKIP POS

2.1 Training phase

In the training phase, a training sentence is first segmented into terms. Each term is labeled with the corresponding POS tag and error type tag. Then our system uses the CRF++ leaning algorithm to train a model. The features used in CRF++ can be expressed by templates. Table 12 (at the end of the paper) shows one sentence in our training set.

Table 13 (at the end of the paper) shows all the templates of the feature set used in our system and the corresponding value for the example. The format of each template is %X[row, col], where row is the number of rows in a sentence and column is the number of column as we shown in Figure 1. The feature templates used in our system are the combination of terms and POS of the input sentences. For example, the first feature template is “Term+POS”, if an input sentence contains the same term with the same POS, the feature value will be 1, otherwise the feature value will be 0. The second feature template is “Term+Previous Term”, if an input sentence contains the same term bi-gram, the feature value will be 1, otherwise the feature value will be 0.

2.2 Test phase

In the Test phase, our system use the trained model to detect and identify the error of an input sentence. Table 2, Table 3, and Table 4 show the labeling results of examples of sentences with error types Redundant, Selection, Disorder, and Missing respectively.

Word	POS	tag	Predict tag
他	N	O	O
是	Vt	O	O
真.	ADV	R	R
很	ADV	O	O
好	Vi	O	O
的	T	O	O
人	N	O	O

Table 2: A tagging result sample of a sentence with error type Redundant

Term	POS	tag	Predict tag
你	N	O	O
千萬	DET	O	O
不要	ADV	O	O
在意	Vt	O	O
這	DET	O	O
個	M	S	S
事情	N	O	O

Table 3: A tagging result sample of a sentence with error type Selection

r

Term	POS	tag	Predict tag
你	N	O	O
什麼	DET	D	D
要.	ADV	D	D
玩	Vt	D	D

Table 4: A tagging result sample of a sentence with error type Disorder

.

Term	POS	Tag	Predict tag
看	Vt	O	O
電影	N	O	O
時候	N	M	M

Table 5: A tagging result sample of a sentence with error type Missing example

If all the system predict tags in the fourth column are the same as the tags in the third column, then the system labels the sentence correctly. In the formal run, accuracy, precision, recall (Clevereon, 1972), and F-score (Rijsbergen,1979) are considered. The measure metrics are defined as follows. The notation is listed in
Table 6.

		System predict tag	
		A	**B**
Known tag	**A**	tpA	eAB
	B	eBA	tpB

Table 6: The confusion matrix.

Precision A =$\frac{tpA}{tpA+eBA}$

Recall A =$\frac{tpA}{tpA+eAB}$

F1-Score A =2 × $\frac{\text{Precision } A \times \text{Recall } A}{\text{Precision } A + \text{Recall } A}$

Accuracy =$\frac{tpA+tpB}{All\ Data}$

3 Experiments

3.1 Data set

Our training data consists of data from NLP-TEA1(Chang et al.,2012)Training Data, Test Data, and the Training Data from NLP-TEA2. Figure 3 (at the end of the paper)shows the format of the data set. Table 7 shows the number of sentences in our training set.

size	NLP-TEA1	NLP-TEA2
Redundant	1830	434
Correct	874	0
Selection	827	849
Disorder	724	306
Missing	225	622

Table 7: Training set size

3.2 Experiments result

In the formal run of NLP-TEA-2 CGED shared task, there are 6 participants and each team submits 3 runs. Table 8 shows the false positive rate. Our system has the lowest false positive rate 0.082, which is much lower than the average.
Table 9, Table 10, and Table 11 show the formal run result of our system compared to the average in Detection level, Identification level, and Position level respectively. Our system achieved the highest precision in all the three levels, but the accuracy of our system is fare. However, the recall of our system is relatively low. The numbers in boldface are the best performance amount 18 runs in the formal run this year.

Submission	False Positive Rate
CYUT-Run1	0.096
CYUT-Run2	**0.082**
CYUT-Run3	0.132
Average of all 18 runs	0.538

Table 8: The false positive rate.

	Detection Level			
	Accuracy	**Precision**	**Recall**	**F1**
CYUT-Run1	0.584	0.7333	0.264	0.3882
CYUT-Run2	0.579	**0.7453**	0.24	0.3631
CYUT-Run3	0.579	0.6872	0.29	0.4079
Average of all 18 runs	0.534	0.560	0.607	0.533

Table 9: Performance evaluation in Detection Level.

	Identification Level			
	Accuracy	**Precision**	**Recall**	**F1**
CYUT-Run1	0.522	0.5932	0.14	0.2265
CYUT-Run2	**0.525**	**0.6168**	0.132	0.2175
CYUT-Run3	0.505	0.5182	0.142	0.2229
Average of all 18 runs	0.335	0.329	0.208	0.233

Table 10: Performance evaluation in Identification Level.

	Position Level			
	Accuracy	**Precision**	**Recall**	**F1**
CYUT-Run1	0.504	0.52	0.104	0.1733
CYUT-Run2	**0.505**	**0.5287**	0.092	0.1567
CYUT-Run3	0.488	0.45	0.108	**0.1742**
Average of all 18 runs	0.263	0.166	0.064	0.085

Table 11: Performance evaluation in Position Level.

4 Error analysis on the official test result

There are 1000 sentences in the official test set of the 2015 CGED shared task. Our system labeled them according to the CRF model that we trained based on the official training set and the available data set from last year.

The number of tag O dominates the number of other tags in the training set for sentences with or without an error. For example, sentence no. B1-0436, a sentence without error:
{上次我坐了 MRT 去了圓山站參觀寺廟了，O(上)，O(次)，O(我)，O(坐)，R(了)，O(MRT)，O(去)，O(了)，O(圓山)，O(站)，O(參觀)，O(寺廟)，O(了)}

And, sentence no. A2-0322, a sentence with an error:
{他們從公車站走路走二十分鐘才到電影院了，O(他們)，O(從)，O(公車站)，O(走路)，O(走)，O(二十)，O(分鐘) ，O(才) ，O(到) ，O(電影院) ，R(了)}

Therefore, our system tends to label words with tag O and it is part of the reason that our system gives the lowest false positive rate this year. Our system also has high accuracy and precision rate, but the Recall rate is lower than other systems. We will analyze the causes and discuss how to improve the fallbacks.

We find that there are 11 major mistake types of our system result.
1. Give two error tags in one sentence.
2. Fail to label the Missing tag
3. Fail to label the Disorder tag
4. Fail to label the Redundant tag
5. Fail to label the Selection tag
6. Label a correct sentence with Missing tag
7. Label a correct sentence with Redundant tag
8. Label a correct sentence with Disorder tag
9. Label a correct sentence with Selection tag
10. Label a Selection type with Redundant tag
11. Label a Disorder type with Missing tag

Analysis of the error cases:

1. Give two error tags in one sentence: In the official training set and test set, a sentence has at most one error type. However, our method might label more than one error tags in one sentence. For example, a system output: {他是很聰明學生，O(他)，R(是)，O(很)，O(聰明)，M(學生)}. Currently, we do not rule out the possibility that a sentence might contain more than one errors. We believe that in the real application, there might be a need for such situation. However, our system might compare the confidence value of each tag and retain only one error tag in one sentence.

2. Fail to label the Missing tag: The missing words might be recovered by rules. For example, a system output: {需要一些東西修理好，O(需要)，**O(一些)**，O(東西)，O(修理好)} should be {需要一些東西修理好，O(需要)，M(一些)，O(東西)，O(修理好)} and the missing word should be ”被” or “把”. A set of rule for ”被” or “把” can be helpful.

3. Fail to label the Disorder tag: The disorder

error is also hard for CRF model, since the named entity (NE) is not recognized first. For example, a system output: {離台北車站淡水不太近，**O(離)，O(台北)，O(車站)，O(淡水)**，O(不)，O(太)，O(近)} should be {離台北車站淡水不太近,D(離),D(台北)，D(車站)，D(淡水)，O(不)，O(太)，O(近)}. The disorder error can only be recognized once the named entities "台北車站" and "淡水" are recognized and then the grammar rule " NE1+離+NE2+近" can be applied.

4. Fail to label the Redundant tag: Some adjacent words are regarded as redundant due to the semantics. Two adjacent words with almost the same meaning can be reduced to one. For example: a system output: {那公園是在台北北部最近新有的，O(那)，O(公園)，O(是)，O(在)，O(台北)，**O(北部)**，O(最近)，O(新)，O(有的)} fail to recognize the redundant word R(台北) or R(北部). In this case, "新有的" is also bad Chinese, it should be "新建的". However, the word segmentation result makes our system hard to detect the error.

5. Fail to label the Selection tag: We believe that it required more knowledge to recognize the selection error than limited training set. For example, a system output: {這是一個很好的新聞，O(這)，O(是)，O(一)，**O(個)**，O(很)，O(好)，O(的)，O(新聞)} fail to recognize the classifiers (also called measure words) for"新聞" should not be "個", the most common Mandarin classifier. It should be "則". A list of the noun to classifier table is necessary to recognize this kind of errors.

6. Label a correct sentence with Missing tag: This case is relative rare in our system. For example, a system output: {一個小時以前我決定休息一下，**M(一)**，O(個)，O(小時)，O(以前)，M(我) ，O(決定)，O(休息) ，O(一下)} accurately contains no error. However our system regard a single "一" should be a missing error according to the trained model.

7. Label a correct sentence with Redundant tag: There are cases that we think our system perform well. For example, our system output: {平常下了課以後他馬上回家,O(平常),O(下),R(了),O(課),O(以後) ,O(他)，O(馬上)，O(回家) }. Where "了" can be regarded as redundant in some similar cases.

8. Label a correct sentence with Disorder tag: This is a rare case in our system. For example, a system output: {以後慢慢知道他這種方式其實是很普通的交朋友的方式，D(以後),D(慢慢),D(知道),D(他),D(這) ，D(種),O(方式),O(其實),O (是) ,O (很)，O (普通),O (的),O (交) ,O (朋友),O (的)，O (方式) }. It is a sentence that cannot be judged alone without enough contexts.

9. Label a correct sentence with Selection tag: In one case, our system output: {今天是個很重要的一天,O(今天),O(是),**S(個)**,O(很)，R(重要) ，O(的)，O(一)，O(天) }, where " 個" is also not a good measure word.

10. Label a Selection type with Redundant tag: Sometimes there are more than one way to improve a sentence. For example, a system output: {下了課王大衛本來馬上回家，O(下),**R(了)**,O(課),O(王大衛),O(本來) ，O(馬上)，O(回家) }, which is no better than {下了課王大衛本來馬上回家，O(下)，O(了)，O(課)，O(王大衛)，S(本來) ，O(馬上)，O(回家) }. Where "本來" should be "就". However, in a different context, it could be "本來想"+"但是…".

11. Label a Disorder type with Missing tag: Since a Disorder error might involve more than two words, comparing to other types, it is hard to train a good model. For example, a system output: {中國新年到了的時候,O(中國)，O(新年)，O(到)，O(了)，M(的) ，O(時候) } should be {中國新年到了的時候，O(中國),D(新年),D(到),D(了),O(的) ，O(時候) }, and the correct sentence should be "到了中國新年的時候". A grammar rule such as "到了"+Event+"的時候" might be help.

5 Conclusion and Future work

This paper reports our approach to the NLP-TEA-2 CGED Shared Task evaluation. Based on the CRF model, we built a system that can achieve the lowest false positive rate and the highest precision at the official run. The

approach uniformly dealt with the four error types: Redundant, Missing, Selection, and Disorder.

According to our error analysis, the difficult cases suggest that to build a better system requires more features and more training data. The system can be improved by integrating rule based system in the future.

Due to the limitation of time and resource, our system is not tested under different experimental settings. In the future, we will test our system with more feature combination on both POS labeling and sentence parsing.

Acknowledgments

This study is conducted under the "Online and Offline integrated Smart Commerce Platform(2/4)" of the Institute for Information Industry which is subsidized by the Ministry of Economic Affairs of the Republic of China .

Reference

Lafferty, A. McCallum, and F. Pereira. (2001) *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. In Intl. Conf. on Machine Learning.

C. W. Clevereon, (1972), *On the inverse relationship of recall and precision*, Workshop on Machine Learning for Information Extraction, pp.195-201.

C. van Rijsbergen, (1979),*Information Retrieval*, Butterworths.

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.

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.

Shih-Hung Wu, Hsien-You Hsieh. (2012). *Sentence Parsing with Double Sequential Labeling in Traditional Chinese Parsing Task*. Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 222–230.

Jui-Feng Yeh, Yun-Yun Lu, Chen-Hsien Lee, Yu-Hsiang Yu, Yong-Ting Chen. (2014). Detecting Grammatical Error in Chinese Sentence for Foreign.

Tao-Hsing Chang, Yao-Ting Sung , Jia-Fei Hong, Jen-I CHANG. (2014). KNGED: a Tool for Grammatical Error Diagnosis of Chinese Sentences.

Yu, L.-C., Lee, L.-H., & Chang, L.-P. (2014). *Overview of grammatical error diagnosis for learning Chinese as a foreign language*. In Proceedings of the 1stWorkshop on Natural Language Processing Techniques for Educational Applications, 42-47.

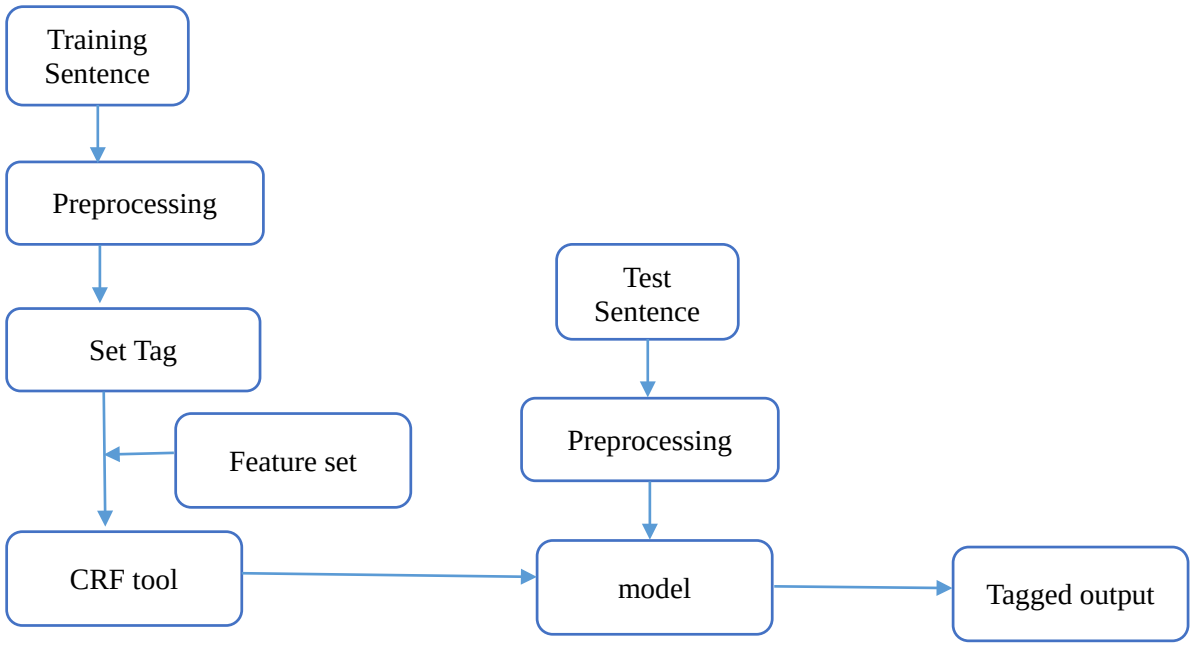

Figure 2: The framework of the proposed system.

```
<root>
<ESSAY title="不能參加朋友找到工作的慶祝會">
<TEXT>
<SENTENCE id="A2-0003-1">
我以前知道妳又很聰明又用功
</SENTENCE>
</TEXT>
<MISTAKE id="A2-0003-1">
<TYPE>Redundant</TYPE>
<CORRECTION>我以前知道妳又聰明又用功</CORRECTION>
</MISTAKE>
</ESSAY>
```

Figure 3: An example of the source data.

	col0	**col1**	**col2**
r-2	他	N	O
r-1	是	Vt	O
r0 (目前 Token)	真	ADV	R
r1	很	ADV	O
r2	好	Vi	O
r3	的	T	O
r4	人	N	O

Table 12: A sample training sentence.

Template Meaning	*Template*	*Feature rule*
Term+POS	%x[0,0]/%x[0,1]	真/ADV
Term+Previous Term	%x[0,0]/%x[-1,0]	真/是
Term+Previous POS	%x[0,0]/%x[-1,1]	真/ Vt
POS+Previous Term	%x[0,1]/%x[-1,0]	ADV/是
POS+Previous POS	%x[0,1]/%x[-1,1]	ADV/ Vt
Term+Previous Term+Previous POS	%x[0,0]/%x[-1,0]/%x[-1,1]	真/是/ Vt
POS+Previous Term+Previous POS	%x[0,1]/%x[-1,0]/%x[-1,1]	ADV/是/ Vt
Term+Second Previous Term	%x[0,0]/%x[-2,0]	真/他
Term+Second Previous POS	%x[0,0]/%x[-2,1]	真/N

POS+Second Previous Term	%x[0,1]/%x[-2,0]	ADV/他
POS+Second Previous POS	%x[0,1]/%x[-2,1]	ADV/N
Term+Second Previous Term+Second Previous POS	%x[0,0]/%x[-2,0]/%x[-2,1]	真/他/N
POS+Second Previous Term+Second Previous POS	%x[0,1]/%x[-2,0]/%x[-2,1]	ADV/他/N
Term+Next Term	%x[0,0]/%x[1,0]	真/很
Term+Next POS	%x[0,0]/%x[1,1]	真/ADV
POS+Next Term	%x[0,1]/%x[1,0]	ADV/很
POS+Next POS	%x[0,1]/%x[1,1]	ADV/ADV
Term+Next Term+Next POS	%x[0,0]/%x[1,0]/%x[1,1]	真/很/ADV
POS+Next Term+Next POS	%x[0,1]/%x[1,0]/%x[1,1]	ADV/很/ADV
Term+Second Next Term	%x[0,0]/%x[2,0]	真/好
Term+Second Next POS	%x[0,0]/%x[2,1]	真/ Vi
POS+Second Next Term	%x[0,1]/%x[2,0]	ADV/好
POS+Second Next POS	%x[0,1]/%x[2,1]	ADV/ Vi
Term+Second Next Term+Second Next POS	%x[0,0]/%x[2,0]/%x[2,1]	真/好/Vi
POS+Second Next Term+Second Next POS	%x[0,1]/%x[2,0]/%x[2,1]	ADV/好/ Vi

Table 13: All the templates and the corresponding value for the sample sentence.