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.

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. 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 a second language, we have to know not only pronunciations and glyph of the word, but also grammar and the part of speech of Chinese.

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).

It is not easy to learn Chinese, people will make a fool of themselves even if one single word mistake. To avoid this problem, more and more people pay their attention to Chinese grammar error.

We detect the grammatical errors by following four common false types: missing, redundant, selection and disorder.

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)***

Method

□ 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. 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:

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

□ **Training phase**

□ 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.

□ **Testing phase**

□ Segment and tagging POS are 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++.

Table. Example of the word’s probability using CRF

Word	POS	Probability
但是(but)	Cbb	T/0.963663
駕駛(driver)	VC	T/0.986188
都(neither)	D	T/0.975163
裝作(pretend)	VF	T/0.970347
沒(not)	D	T/0.962676
看到(see)	VE	T/0.984734
或者(or)	Caa	T/0.953170
聽到(hear)	VE	T/0.988986
我(me)	Nh	T/0.997955
了	T	*F/0.579991*

□ Rule Induction

□ In English, we usually use “a” or “an” to denote quantifier. But Chinese needs more different quantifiers then the other language.

Table. The different quantifiers apply to Chinese

Relationship	Using words
Human	‘位’，‘個’
Animal	‘隻’，‘匹’，‘頭’，‘條’
Event	‘件’
Building	‘座’，‘棟’
Transportation	‘臺’，‘輛’，‘架’，‘艘’

□ There are some rules which follow to finding ordering error.

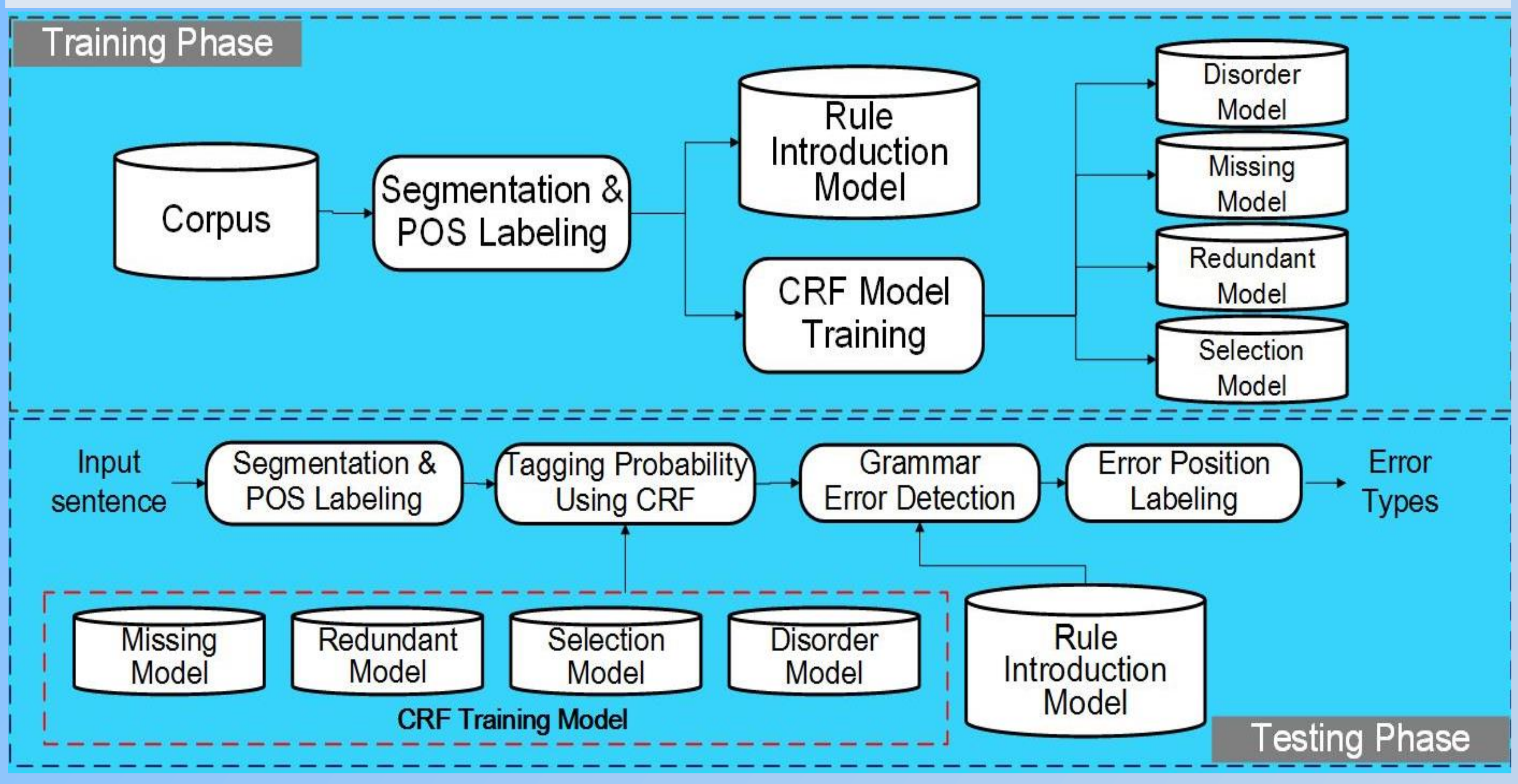
□ Behind the words “把 (let)” is connected the POS ‘Nh’ or ‘Na’ or ‘Nep’.

□ 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’.

System Architecture



Experiment

□**Experiment 1 (CRF system & Hybrid system)**

Only use CRF it can’t find many error but its precision is better. Then add the rule induction can promote the recall means it can find more error from test data.

Detection level

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

Identification level

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

□**Experiment 2 (NLP-TEA 2015 Competition)**

In this experiment, 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.

The performance with the NLP-TEA 2015 testing data and compare the other team show in the table:

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

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

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

Conclusion

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

□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.

Acknowledgements

This work is supported in part by the National Science Council, Taiwan, R.O.C., under the project grant numbers NSC 102-2221-E-415-006-MY3.