A CRF Sequence Labeling Approach to Chinese Punctuation Prediction

Yanqing Zhao, Chaoyue Wang, Guohong Fu
School of Computer Science and Technology, Heilongjiang University
Harbin 150080, China
yanqing_zhao@live.cn, chariey_nlp@yahoo.cn, ghfu@hlju.edu.cn

Abstract

This paper presents a conditional random fields based labeling approach to Chinese punctuation prediction. To this end, we first reformulate Chinese punctuation prediction as a multiple-pass labeling task on a sequence of words, and then explore various features from three linguistic levels, namely words, phrase and functional chunks for punctuation prediction under the framework of conditional random fields. Our experimental results on the Tsinghua Chinese Treebank show that using multiple deeper linguistic features and multiple-pass labeling consistently improves performance.

1 Introduction

Punctuation prediction, also referred to as punctuation restoration, aims at inserting proper punctuation marks at right position of an unpunctuated text (Gravano et al., 2009; Guo et al., 2010). Punctuation is obviously an essential indicator for sentence construction. For Chinese, adding proper punctuation marks can not only enhance the readability of text, but also can provide additional information for further language analysis, such as word segmentation, phrasing and syntactic analysis (Guo et al., 2010; Chen and Huang, 2011; Xue and Yang, 2011). As such, punctuation prediction plays a critical role in many natural language processing applications such as automatic speech recognition (ASR), machine translation, automatic summarization, and information extraction (Matusov et al., 2006; Lu and Ng, 2010).

Over the past years, numerous studies have been performed on the insertion of punctuations in speech transcripts using supervised techniques. However, it is actually very difficult or even impossible to develop a large high-quality corpus to achieve reliable models for predicting punctuations in speech transcripts or ASR outputs (Takeuchi et al., 2007). Furthermore, most previous research on punctuation prediction exploited very shallow linguistic features such as lexical features or prosodic cues (viz. pitch and pause duration) (Lu and Ng, 2010), few studies have been done on the exploration of deeper linguistic features like syntactic structural information for punctuation prediction, particularly in Chinese (Guo et al., 2010).

In this paper we draw our motivation from speech transcripts to written texts. On the one hand, a number of large annotated corpora of written texts are available to date. On the other hand, we intend to examine the role of different linguistic features on Chinese punctuation prediction. To this end, we reformulate Chinese punctuation prediction as a multiple-pass labeling task on word sequences, and then explore multiple features at three linguistic levels, namely words, phrases and functional chunks, for punctuation labeling under the framework of conditional random fields (CRFs). Furthermore, we have also performed evaluation on the Tsinghua Chinese Treebank (Zhou, 2004).

The rest of the paper is organized as follows: In Section 2, we will provide a brief review of the related work on punctuation prediction. In Section

26th Pacific Asia Conference on Language,Information and Computation pages 508–514

3, we will describe in detail a labeling method to Chinese punctuation prediction. Section 4 will summarize the experimental results. Finally in Section 6, we will give our conclusion and some possible directions for future work.

2 Related Work

Punctuation prediction has been well studied in the communities of ASR, and a variety of techniques have been attempted, including n-grams (Takeuchi et al., 2007; Gravano et al., 2009), maximum entropy models (MEMs) (Huang and Zweig, 2002; Guo et al., 2010), and CRFs (Liu et al., 2005; Tomanek et al., 2007; Lu and Ng, 2010).

Current research focuses on seeking informative features for punctuation prediction. Huang and Zweig (2002) attempted to explore POS features and prosodic features for inserting punctuations in automatically recognized speech texts using MEMs. Takeuchi et al. (2007) exploited silence information from ASR systems and head or tail phrases within sentences. They showed that using head and tail phrases could result in improvement of performance in sentence boundary detection. Gravano et al. (2009) examined the effect of different orders of n-grams on performance in punctuation prediction. More recently, Huang and Chen (2011) used CRFs to combine different features for labeling pause and stop in Chinese texts, including the beginning and end features of voice fragments, character features, word features, POS features, syntactic features and topic features.

In addition to speech transcripts or ASR outputs, recently a number of researchers start to study punctuation prediction via written texts. Tomanek et al (2007) employed CRFs to phrase a biological article, and then inserted punctuation to sentences during sentence segmentation. Xue and Yang (2011) used MEMs to explore contextual words, POS features and syntax trees for inserting commas in Chinese texts. Laboreiro and Sarmento (2010) applied support vector machines to exploit multiple cues, including such as characters, character types, symbols and punctuations, for sentence segmentation and punctuation correction in micro-blog texts.

From these studies, it is clear that systems with more and deeper features outperform systems only using simple features. However, most previous studies only used lexical cues for punctuation prediction. This might be that a well-annotated corpus of speech texts is not available to date. As such, in the present study we address the problem of Chinese punctuation prediction from the perspective of written texts. Specially, we attempt to exploit multiple levels of features under the framework of CRF-based sequence labeling and thus examine the role of for Chinese punctuation

3 Approach

This section details the CRFs-based multiple pass labeling method to Chinese punctuation prediction.

3.1 Task Formulation

Chinese punctuation prediction is a process of inserting proper punctuation marks into a raw Chinese text without punctuation marks. In the present study, we reformulate Chinese punctuation prediction as a multiple-pass labeling task on an unpunctuated word string with the help of word pattern tags defined in Table 1. Furthermore, we consider eleven main punctuation marks as shown in Table 2.

Tag	Definition
B	The preceding word of the current punctuation.
A	The following word of the current punctuation.
O	Words not adjacent to the current punctuation.
BOT	The head word of a text.
EOT	The tail word of a text.

Table 1: Patterns of words in punctuation labeling

No.	Name	Punctuation	Tag
1	Comma	，	COM
2	full stop	。	FUL
3	exclamation mark	！	EXC
4	Colon	：	COL
5	Bracket	（）{}[]	BRA
6	question mark	？	QUE
7	Semicolon	；	SEM
8	enumeration comma	、	ENU
9	book title mark	《》 〈〉	BOO
10	quotation mark	“” ‘’	QUO
11	Ellipsis	……	ELL

Table 2: Types of Chinese punctuation marks

In order to reduce the interference between different types of punctuation marks and to simplify the problem of punctuation prediction as well, we take the following order to perform punctuation labelling: sentence-final delimiters (viz. period, question mark and exclamation mark) → sentence-internal delimiters (viz. comma, semicolon, colon, and enumeration comma) → indicators (viz. bracket, book title mark, quotation mark, and ellipsis).

After punctuation labeling, each word within the unpunctuated text will receive a hybrid punctuation tag of the form t_1-t_2, if it is adjacent to a punctuation mark, or is at the beginning or end of a text. Otherwise, it will only receive a tag *O*. Here, t_1 denotes the pattern of the current word in punctuation labeling (as shown in Table 1), and t_2 stands for the type of the punctuation mark (as defined in Table 2) that precedes or follows the current word if applicable.

(a) **Punctuated text:** 执法部门是反腐败斗争 、搞好廉政建设的重点部门之一。

(b) **Unpunctuated word string:** 执法/部门/是/反/腐败/斗争/搞好/廉政/建设/的/重点/部门/之一/

(c) **POS:** 执法/vN 部门/n 是/vC 反/v 腐败/a 斗争/vN 搞好/v 廉政/vN 建设/vN 的/uJDE 重点/n 部门/n 之一/rN

(d) **Phrases:** [np-ZX 执法/vN 部门/n] [vp-SG 是/vC] [np-ZX 反/v 腐败/a 斗争/vN] [vp-SG 搞好/v] [np-ZX 廉政/vN 建设/vN] 的/uJDE [np-ZX 重点/n 部门/n] [np-SG 之一/rN]

(e) **Functional chunks:** [S 执法/vN 部门/n] [P 是/vC] [P 反/v 腐败/a 斗争/vN] [P 搞好/v] [O 廉政/vN 建设/vN] 的/uJDE [H 重点/n 部门/n] [H 之一/rN]

(f) **Punctuation labeling:** 执法/BOT-O 部门/O 是/O 反/O 腐败/O 斗争/B-ENU 搞好/A-ENU 廉政/O 建设/O 的/O 重点/O 部门/O 之一/EOT-FUL

Figure 1: Representation of punctuation labeling

To further illustrate the problem of punctuation labeling, consider the following exemplar text “执法部门是反腐败斗争 、搞好廉政建设的重点部门之一。” (Law enforcement agencies are one of the priority sectors for the fight against corruption and the construction of a clean government.), along with its unpunctuated word string, three levels of linguistic annotations and the corresponding punctuation labeling representation.

It is worth noting the major motivation of this study is to investigate the effects of different levels of linguistic cues on Chinese punctuation prediction. To achieve this, we take the following three steps: First, we remove all punctuation marks within a given original punctuated text like line (a) in Figure 1 and reduce it to an unpunctuated text (viz. line (b)) before punctuation labeling. Then, we explore three levels of linguistic information to restore the removed punctuation marks using the CRF-based multiple-pass labeling strategy. Finally, we evaluate punctuation prediction performance by comparing the automatically restored punctuation marks with the corresponding original ones.

Considering the availability of linguistic information, we perform punctuation prediction on the Tsinghua Chinese Treebank (Zhou, 2004), a corpus of written Chinese with a variety of linguistic annotation information, including word segmentation, POS, phrases and functional chunks. Also, the relevant annotation scheme is used throughout our present study.

3.2 CRFs for Punctuation Labeling

We choose CRFs as the basic framework for punctuation labeling in that CRFs have proven to be one of the most effective techniques for sequence labeling tasks (Lafferty et al., 2001). Compared with other methods, CRFs allow us to exploit numerous observation features as well as state sequence based features or other features to punctuation labeling.

Let $X = (x_1, x_2, \cdots, x_T)$ be an input sequence of Chinese words, $Y = (y_1, y_2, \cdots, y_T)$ be a sequences of punctuation tags as defined in Section 3.1. From a statistical point of view, the goal of punctuation labeling is to find the most likely sequence of punctuation tags $\hat{Y}$ for a given sequence of words X that maximizes the conditional probability $p(Y \mid X)$. CRFs modeling uses Markov random fields to decompose the conditional probability $p(Y \mid X)$ of a tag sequence as a product of probabilities below

$$p(y \mid x) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{T} \sum_{j} \lambda_j f_j(y, x, i)) \quad (1)$$

Where $f_j(y,x,i)$ is the j^{th} feature function at position *i*, associated with a weight λ_j, and $Z(x)$ is a moralization factor that guarantees that the summation of the probability of all sequences of punctuation tags is one, which can be further calculated by

$$Z(x) = \sum_{y} \exp(\sum_{i=1}^{T} \sum_{j} \lambda_j f_j(y,x,i)) \qquad (2)$$

3.3 Features

We explore cues for punctuation prediction from three linguistic levels, namely words, phrases and functional chunks.

At word level, we exploit word forms and their POS tags in a window of three words, including the current word $w_{i,}$ the preceding word w_{i-1} and the following word $w_{i+1,}$ and their respective POS tags t_i, t_{i-1}, and t_{i+1}. Table 3 details the feature template at word level.

No.	Feature	Definition
L1	$w_{i-1}w_i$	The current word and the preceding word.
L2	$w_{i-1}w_{i+1}$	The current word and the following word.
L3	$w_{i-1}t_i$	The preceding word and the current word's POS tag
L4	t_iw_{i+1}	The current word's POS tag and the following word
L5	$t_{i-1}w_i$	The preceding word's POS tag and the current word
L6	w_it_{i+1}	The current word and the following word's POS tag
L7	w_i	The current word

Table 3: Word-level features

At phrase level or functional chunk level, we consider some possible combinations of the current word, the preceding word, the following word and their relevant phrase tags or functional chunk tags as features for punctuation prediction. The templates for phrase-level and functional chunk-level features are given in detail in Table 4 and Table 5, respectively. Where, p_i, p_{i-1} and p_{i+1} denote the category tags of the phrases containing words w_i, w_{i-1} and w_{i+1}, respectively, while p_i, p_{i-1} and p_{i+1} stands for the corresponding functional chunk tags.

No.	Feature	Definition
P1	$w_{i-1}p_{i-1}w_i\ p_i$	The preceding word and its phrase tag, the current word and its phrase tag.
P2	$w_ip_iw_{i+1}\ p_{i+1}$	The current word and its phrase tag, the following word and its phrase tag
P3	$w_{i-1}\ \ p_{i-1}t_i\ p_i$	The preceding word and its phrase tag, the current word's POS and phrase tag
P4	$t_ip_iw_{i+1}\ p_{i+1}$	The current word's POS and phrase tag, the following word and its phrase tag
P5	$t_{i-1}\ \ p_{i-1}w_i\ p_i$	The preceding word's POS and phrase tag, the current word and its phrase tag
P6	$w_ip_it_{i+1}\ p_{i+1}$	The current word and its phrase tag, the following word's POS and phrase tag
P7	$p_{i-1}w_ip_i$	The preceding word's phrase tag, the current word and its phrase tag
P8	$w_i\ p_ip_{i+1}$	The current word and its phrase tag, the following word's phrase tag
P9	$p_{i-1}t_ip_i$	The preceding word's phrase tag, the current word's POS and phrase tag
P10	$t_i\ p_ip_{i+1}$	The current word's POS and phrase tag, the following word's phrase tag

Table 4: Phrase-level features

No.	Feature	Definition
F1	$w_{i-1}f_{i-1}w_i\ f_i$	The preceding word and its functional chunk tag, the current word and its functional chunk tag
F2	$w_if_iw_{i+1}f_{i+1}$	The current word and its functional chunk tag, the following word and its functional chunk tag
F3	$w_{i-1}\ f_{i-1}t_i\ f_i$	The preceding word and its functional chunk tag, the current word's POS and its functional chunk tag
F4	$t_if_iw_{i+1}\ f_{i+1}$	The current word's POS and its functional chunk tag, the following word and its functional chunk tag
F5	$t_{i-1}\ f_{i-1}w_i\ f_i$	The preceding word's POS and functional chunk tag, the current word and its functional chunk tag
F6	$w_if_it_{i+1}\ f_{i+1}$	The current word and its functional chunk tag, the following word's POS and functional chunk tag
F7	$f_{i-1}w_if_i$	The preceding word's functional chunk tag, the current word and its functional chunk tag
F8	$w_i\ f_if_{i+1}$	The current word and its functional chunk tag, the following word's functional chunk tag
F9	$f_{i-1}t_if_i$	The preceding word's functional chunk tag, the current word's POS and its functional chunk tag
F10	$t_i\ f_if_{i+1}$	The current word's POS and its functional chunk tag, the following word's functional chunk tag

Table 5: Functional chunk-level features

4 Experimental Results and Discussions

To assess the effectiveness of our approach, we have conducted several experiments on the Tsinghua University Chinese Treebank (Zhou, 2004). This section will present the relevant results.

4.1 Experiment Setup

In our experiment, we divide the Tsinghua University treebank (Zhou, 2004) into two parts: One for training and the other for testing. Table 6 shows the distribution of different punctuation marks in these datasets.

Punctuation	Training dataset		Test dataset	
	Number	Rate	Number	Rate
comma	25918	44.79	5924	43.83
period	12670	21.90	3350	24.79
enumeration comma	7769	13.43	1896	14.03
quotation mark	5484	9.48	920	6.81
title mark	1656	2.86	360	2.66
bracket	1394	2.41	388	2.87
semicolon	1048	1.81	330	2.44
colon	1009	1.74	223	1.65
dash	260	0.45	44	0.33
question mark	243	0.42	42	0.31
exclamation mark	215	0.37	22	0.16
connective mark	199	0.34	16	0.12
Total	57865	100	13515	100

Table 6: Distribution of different punctuation marks in the experimental datasets

Sentence length	Total	Rate	Average number of punctuation per sentence
< 10	2307	16.19	1.04
10~19	4543	31.89	2.66
20~29	3598	25.25	4.19
30~39	1986	13.94	5.75
40~49	936	6.57	7.51
50~59	430	3.02	9.44
60~69	222	1.56	10.80
≥ 70	226	1.59	15.80
Total	14248	100	4.06

Table 7: Average number of punctuation within sentences of different length in training dataset

Table 7 and Table 8 present the average numbers of punctuations within sentences of different length in the training dataset and the test dataset, respectively. From these two tables, we can see that the number of words in most Chinese sentence is less than 40, and the average number of punctuation marks per sentence in Chinese is about 4.

Sentence length	Total	Rate	Average number of punctuation per sentence
< 10	666	17.76	0.94
10~19	1381	36.82	2.56
20~29	937	24.98	4.05
30~39	447	11.92	5.72
40~49	164	4.37	7.19
50~59	79	2.11	8.75
60~69	27	0.72	11.26
≥ 70	50	1.33	16.94
Total	3751	100	3.60

Table 8: Average number of punctuation marks within sentences of different length in test dataset

In addition, we employ three metrics to score punctuation prediction performance, namely the precision (denoted by P), the recall (denoted by R) and the F-score.

4.2 Effects of Features at Different Levels

Our first experiment intends to examine the effects of different features at different linguistic levels on Chinese punctuation prediction. This experiment is conducted with a single-pass strategy, which performs punctuation labeling in one pass. The results are presented in Tables 9, 10 and 11.

Feature	P	R	F
L1, L2, L7	0.699	0.444	0.543
L4, L5	0.625	0.493	0.551
L4, L5, L7	0.597	0.536	0.565
L1-L5	0.677	0.478	0.560
L1-L6	0.667	0.492	0.566
L1-L7	0.644	0.515	0.572

Table 9: Results for different word-level features under single-pass sequence labeling

As can be seen in these three tables, combining a variety of contextual features can improve the performance of Chinese punctuation prediction. Take the evaluation of word-level features in Table

9 as an example: the F-score is 0.543 when using word unigrams and bigrams only. But when integrating contextual words with their corresponding POS, the F-score can be increased by nearly 3 percents. Furthermore, we can also observe that among the three levels of linguistic cues, using functional chunk cues yields the best performance under the strategy of single-pass sequence labeling.

Feature	P	R	F
P1-P6	0.713	0.464	0.563
P1-P8	0.698	0.489	0.575
P1-P10	0.649	0.640	0.645

Table 10: Results for different phrase-level features under single-pass sequence labeling

Feature	P	R	F
F1-F6	0.788	0.462	0.583
F1-F8	0.782	0.505	0.613
F1-F10	0.738	0.637	0.684

Table 11: Results for different functional chunk-level features under single-pass sequence labeling

4.3 Using Multiple-Pass Sequence Labeling

As we have mentioned above, we employ a multiple-pass sequence labeling strategy to predict different types of punctuation marks in Chinese text. Therefore, our second experiment is designed to examine the effect of using multiple-pass sequence labeling in Chinese punctuation prediction. This experiment is conducted by comparing the outputs of the two labeling strategies, namely multiple-pass sequence labeling and single-pass sequence labeling. The results are given in Table 12.

Feature	Single-pass sequence labeling			Multiple-pass sequence labeling		
	P	R	F	P	R	F
L1-L7	0.644	0.515	0.572	0.785	0.467	0.585
P1-P10	0.649	0.640	0.645	0.773	0.586	0.666
F1-F10	0.738	0.637	0.684	0.817	0.611	0.699

Table 12: Comparing multiple-pass sequence labeling with single-pass sequence labeling

We can observe from Table 12 that compared with single-pass sequence labeling, multiple-pass sequence labeling results in a substantial improvement of precision and F-score, while the recall slightly declines. The reason might be that multiple-pass strategy treats different types of punctuation marks separately and thus can handle their individual characteristics.

4.4 Combining Phrase-Level and Functional Chunk-Level Features

Intuitively, functional chunk features are more informative in short sentence segmentation while phrase-level features are more helpful in tokenization within short sentences. At this point, phrase-level features and functional features might be complementary each other during punctuation prediction. As such, we believe that combining different levels of features would result in further improvement of performance. To prove this, we finally conducted an experiment by comparing the output before and after the combination of phrase-level and functional chunk-level features. The results are presented in Table 13.

Punctuation	P	R	F
comma	0.753	0.743	0.748
period	0.945	0.984	0.964
exclamation mark	0.667	0.09	0.160
colon	0.603	0.184	0.282
bracket	0.829	0.088	0.159
question mark	0.889	0.381	0.533
semicolon	0.529	0.027	0.052
enumeration comma	0.820	0.497	0.619
title mark	0.895	0.047	0.090
quotation mark	0.409	0.03	0.056
Overall	0.820	0.649	0.725

Table 13: Results for combining phrase-level features and functional chunk-level features under multiple-pass sequence labeling

From Table 13 we can see that incorporating functional chunk-level features with phrase-level features can obtain the best overall F-score of 0.725, 2.6 percents higher than that of only using functional chunk-level features (shown in Table 12). This confirms in a sense our intuition.

5 Conclusions

In this paper, we proposed a CRFs-based multiple-pass labeling approach to Chinese punctuation prediction. In particular, we have explored features for punctuation prediction at three levels, namely words, phrases and functional chunks, and thus examined their respective effects on Chinese punctuation prediction through experiments on the Tsinghua Treebank. We show that using multiple deeper features under multiple-pass labeling strategy can result in performance improvement.

Although the proposed method yields good results for periods and commas, the prediction of brackets, quotations and title identifier is still not satisfactory. This might be due to the data sparseness caused by the small number of these punctuations. Another possible reason is that the features in use are not effective or informative for these punctuation marks. Therefore, in the future research we plan to improve our current system by expanding the scale of the training corpus and seeking more informative features for Chinese punctuation prediction.

Acknowledgments

This study was supported by National Natural Science Foundation of China under Grant No.60973081 and No.61170148, the Returned Scholar Foundation of Educational Department of Heilongjiang Province under Grant No.1154hz26, and Harbin Innovative Foundation for Returnees under Grant No.2009RFLXG007, respectively.

References

Agusting Gravano, Martin Jansche, and Michiel Bacchiani. 2009. Restoring punctuation and capitalization in transcribed speech. In Proceedings of ICASSP'09, pp.4741-4744.

Evgeny Matusov, Arne Mauser, and Hermann Ney. 2006. Automatic sentence segmentation and punctuation prediction for spoken language translation. In Proceedings of IWSLT'06, pp.158-165.

Gustavo Laboreiro, and Luís Sarmento. 2010. Tokenizing micro-blogging messages using a text classification approach. In Proceedings of AND'10, pp.81-87.

Hen-Hsen Huang, and Hsin-Hsi Chen. 2011. Pause and stop labeling for Chinese sentence boundary detection. In Proceedings of Recent Advances in Natural Language Processing, pp.146-153.

Hironori Takeuchi, L. Venkata Subramaniam, Shourya Roy, Diwakar Punjani, and Tetsuya Nasukawa. 2007. Sentence boundary detection in conversational speech transcripts using noisily labeled examples. International Journal of Document Analysis and Recognition, 10(3):147-155.

Jing Huang, and Geoffrey Zweig. 2002. Maximum entropy model for punctuation annotation from speech. In Proceedings of ICSLP'02, pp. 917-920.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML'01, pp.282-289.

K. Tomanek, J. Wermter, and U. Hahn. 2007. Sentence and token splitting based on conditional random fields. In Proceedings of PACLING'07, pp.49-57.

Nianwen Xue, and Yaqin Yang. 2011. Chinese sentence segmentation as comma classification. Proceedings of ACL '11, pp. 631-635.

Qiang Zhou. 2004. The annotation scheme for Chinese Treebank. Journal of Chinese information processing, 18(4): 1-8.

Wei Lu, and Hwee Tou Ng. 2010. Better punctuation prediction with dynamic conditional random fields. In Proceedings of EMNLP '10, pp.177-186.

Yang Liu, A. Stolcke, E. Shriberg, and M. Harper. 2005. Using conditional random fields for sentence boundary detection in speech. In Proceedings of ACL '05, pp.451-458.

Yuqing Guo, Haifeng Wang, and J. V. Genabith. 2010. A linguistically inspired statistical model for Chinese punctuation generation. ACM Transactions on Asian Language Information Processing, 9(2): Article 6.