E-030

Towards the Integration of Natural Language and Eye Tracking Information for Predicting Comma Placement in Chinese Sentence

Chen Chen[1] Yoshinobu Kano[2] Akiko Aizawa[1, 3]

1. Introduction

This paper investigates a relatively underdeveloped but important subject in natural language processing (NLP) – prediction of punctuation marks. As a logographic language, Chinese provides a case of high contrast for alphabetic systems, because there are no word boundaries (spaces) between Chinese words. (Ren & Yang, 2010). It is sometimes difficult for Chinese learners to identify which characters compose certain words within a Chinese sentence. This characteristic raises an interesting question that how the segmentation cues such as spaces or commas inserted into text influence the course of Chinese reading.

Bai et al. (2008) investigated spaced and unspaced Chinese reading but found no facilitation of spaces in word identification. As a visual mark in written language, a comma is widely used in Chinese writing to provide additional space and serve as a segmentation cue in Chinese text. More importantly, a comma not only has prosodic functions (Kerkhofs, et al. 2008) but can be placed at syntactic boundaries as well (Chafe, 1988). For example, a phrase or a clause in Chinese sentences is usually followed by a comma, although there is a great deal of flexibility in the use of commas. Ren & Yang (2010) showed that prosodic boundaries marked by commas facilitate sentence processing only when they cooperated with syntactic boundaries.

In our present study, a Chinese comma placement predictor is expected to be created by integrating NLP and Eye Tracking technology. Lu & Ng (2010) proposed an approach built on top of dynamic conditional random fields (DCRF) framework, which jointly performs punctuation prediction together with sentence boundary and sentence type prediction on speech utterance without prosodic cues. Zhang, et al. (2009) presented a conditional random fields (CRF) based approach which automates ancient Chinese prose punctuation using the mutual information and the t-test difference as features. Guo & Wang (2010) tried combining sophisticated statistical techniques with linguistic analyses to facilitate generation of punctuation.

However, how to detect a comma placement with prosodic functions remains a problem because it is more related with the human's cognitive system but not only a linguistic problem. Eye tracking technology is considered to be a solution because it can show the intuitive reading process of people and detect hard-to-read point when they are reading Chinese text without comma. After analyzing this difficulty, a comma distribution which more accords with the reader's intuition is expected to be found.

In this paper, we describe our comma predictor of modern Chinese with machine learning (ML) techniques by integrating several linguistic features so that the commas placed at syntactic boundaries can be predicted. We also briefly mention our plan of eye tracking experiments that examine whether eye tracking information can help find commas with prosodic function and then improve the accuracy of prediction.

2. Method

2.1 Model

In this paper, comma annotation is formalized as a binary classification task for each boundary between two consecutive words (including other punctuations). We use the CRF sequence model (Lafferty, et al. 2001), annotating commas in the IOB style that often used in named entity recognition tasks. IOB tags become I-Comma (tag the word after the comma), B-Comma (tag the word before the comma) and O (for other words). The program is based on the CRF toolkit, Mallet, and its application ABNER (Settles, 2004) that predicts IOB tags.

2.2 Features

We used different types of features in CRF including word surface (WS), part-of-speech tags (POS), position of a word in a sentence (PIS) and depth of a word in the parse tree (DIP).

2.3 Corpus

We used Chinese Treebank 7.0 (Xue, et al. 2005) where POS and syntactic brackets annotated for 1,196,329 words. We randomly divided the corpus into training data (90%) and test data with genres of newswire, news magazine, broadcast news and newsgroups/weblogs. And in order to get the upper bound, close set (test data is included in training data) is also tried. We excluded the genre of broadcast conversation due to the great difference between informal spoken Chinese and formal journalistic language. Fig 1 shows an example of the train data. We evaluated the result by the F1 score.

```
对|P|4|O 此|PN|5|B-Comma 浦东|NR|4|I-Comma 不|AD|6|O
是|VC|6|O 简单|VA|9|O 的|DEV|8|O 采取|VV|8|O
*PRO*|-NONE-|12|O “|PU|12|O 干|VV|13|O 一|CD|15|O
段|M|16|O 时间|NN|15|B-Comma 等|P|14|I-Comma
*pro*|-NONE-|17|O 积累|VV|17|O 了|AS|17|O 经验|NN|18|O
```

Fig 1. *Training data with format of WS|POS|DIP|IOB*

3. Experimental Results and Discussion

In this part, the results of three experiments are given in Table 1, Table 2 and Table 3. We used POS-tagged Data, Bracketed Data with and without the null elements respectively to explore the combination of a CRF trained with various subsets of features.

Table 1. *Evaluation of the CRF model with POS-tagged Data*

Result / Feature	**Precision**	**Recall**	***F*-Score**
WS + POS + PIS	77.62%	47.06%	58.60%
WS + POS	75.64%	51.83%	61.51%
WS + POS (closed set)	82.25%	52.53%	64.11%
WS	73.63%	48.94%	58.80%
POS	60.46%	19.44%	29.43%

By investigating Table 1, surprisingly the feature word surface (WS) was found to play an important role contrary to our expectation, while POS does not contribute much for the

[1] The University of Tokyo, Tokyo, Japan , [2] PRESTO, Japan Science and Technology Agency, Tokyo, Japan,
[3] National Institute of Informatics, Tokyo, Japan, {chen, kano, aizawa}@nii.ac.jp

performance improvement. The results indicate that comma placement might be more related to prosodic function or other syntactic factors than POS. The PIS feature is found to decrease the performance as well, which means that it is not necessary to appear a comma in some specific position in a Chinese sentence.

Table 2. *Evaluation of the CRF model with Bracketed Data* (with the null elements such as "*OP*", "*T*" and "*pro", and test data is again selected more randomly from all four kinds of texts)

Result / Feature	**Precision**	**Recall**	***F*-Score**
WS + POS + DIP	80.29%	60.24%	68.84%
WS + POS + DIP (closed set)	85.05%	75.11%	79.78%
WS + DIP	76.22%	60.99%	67.76%
WS + POS	73.33%	42.12%	53.51%
POS + DIP	74.60%	48.31%	58.64%
WS	70.98%	42.54%	53.20%
DIP	66.11%	26.68%	38.01%
POS	64.50%	26.48%	37.55%

Table 2 shows that DIP greatly improves the performance by about 10%. That proves the importance of syntactic factors in Chinese comma prediction. Here null elements (-NONE-), which were created in PTB II and retained in CTB 7.0 as well, were designed to provide traces which may be co-indexed with the relevant source lexical material so as to facilitate predicate-argument interpretation. But the DIP's contribution to the performance is still inferior to that of word surface. This may be because the Chinese word itself owns both prosodic and syntactic information, but the exact reason needs to be further investigated.

Table 3. *Evaluation of the CRF model with Bracketed Data* (without the null elements)

Result / Feature	**Precision**	**Recall**	***F*-Score**
WS + POS + DIP	77.36%	58.94%	66.90%
WS + DIP	73.52%	56.75%	64.05%
WS + POS	64.74%	41.46%	50.55%
POS + DIP	72.59%	43.56%	54.44%
WS	60.05%	28.29%	38.46%
DIP	59.33%	18.14%	27.78%
POS	51.75%	25.00%	33.71%

Almost all the results were worse than that with null elements in Table 2. Because –NONE- covers about 9% of all the data, it may suggest that the amount of data is important to improve the performance. Otherwise, because null can be considered to reveal rich syntactic information, the importance of syntactic features is stressed again. This is compatible with the prior research (Favre et al., 2009).

4. Conclusion and Future Work

Our study showed that performance of modern Chinese prediction can reach around 70% just by using CRF-based predictor with several features, which is at the same level with some prior studies of other methods (Favre et al., 2009; Guo, Wang & Genabith, 2010), although our feature sets are not rich enough and are not chosen elaborately yet. Otherwise, the data sparseness problem was also discovered by comparing the results with and without the null elements. If more annotated data can be available for training from existing corpus or that generated by parsing tools, there would be still much room for improvement. Furthermore, the results are consistent with previous findings concerning the important role of both syntax and prosody to decide the placement of the comma.

After checking the results carefully, a big part of the false positive results (unexpected results) was found to still make sense according to the feedback from native subjects. This suggests that such results are possibly considered reasonable in the next step of this research by using an eye tracker, because they might be acceptable for normal reading, or give the readers even more smooth reading experience.

As a future work, we plan to implement a comma predictor based on the eye tracking data first. We would use the information such as regression, fixation time and saccade length as features to train the comma prediction model. For preliminary plan, CRF sequence model would still be applied until we find some superior methods with better performance. Finally we would combine these two parts in order to find better comma distribution for the readability improvements and to give readers better reading experience.

References

Bai, X., Yan, G., Liversedge, S.P., Zang, C. and Rayner, K. (2008). Reading spaced and unspaced Chinese text: Evidence from eye movements. *JEPHPP*, 34(5), 1277–1287.

Chafe, W. (1988). Punctuation and the prosody of written language. *Written Communication*, 5, 396–426.

Guo, Y.Q., Wang, H.F. and Genabith, J.V. (2010). A linguistically inspired statistical model for Chinese punctuation generation. In *ACM TALIP*, 9 (2), Article 6.

Kerkhofs, R., Vonk, W., Schriefers, H. and Chwilla, D.J. (2008). Sentence processing in visual and auditory modality: Do comma and prosodic break have parallel functions? *Brain Research*, 1224, 102–118.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML '01*, 282-289.

Lu, W. and Ng, H.T. (2010). Better punctuation prediction with dynamic conditional random fields. In *Proceedings of EMNLP '10*, 177-186.

Ren, G.Q. and Yang, Y.F. (2010). Syntactic boundaries and comma placement during silent reading of Chinese text: evidence from eye movements. *JRIR*, 33(2), 168-177.

Settles, B. (2004). Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of JNLPBA '04*, 104-107.

Xue, N., Xia, F., Chiou, F.D. and Palmer, M. (2005). The Penn Chinese TreeBank: Phrase structure annotation of a large corpus, *Natural Language Engineering*, 11(2), 207-238.

Zhang, K.X., Xia, Y.Q. and Yu H. (2009). CRF-based approach to sentence segmentation and punctuation for ancient Chinese prose. *J Tsinghua Univ (Sci & Tech)*, 49, Article 10.