Modeling Comma Placement in Chinese Text for Better Readability using Linguistic Features and Gaze Information

Tadayoshi Hara[1] **Chen Chen**[2*] **Yoshinobu Kano**[3,1] **Akiko Aizawa**[1]
[1]National Institute of Informatics, Japan [2]The University of Tokyo, Japan
[3]PRESTO, Japan Science and Technology Agency
{harasan, kano, aizawa}@nii.ac.jp

Abstract

Comma placements in Chinese text are relatively arbitrary although there are some syntactic guidelines for them. In this research, we attempt to improve the readability of text by optimizing comma placements through integration of linguistic features of text and gaze features of readers.

We design a comma predictor for general Chinese text based on conditional random field models with linguistic features. After that, we build a rule-based filter for categorizing commas in text according to their contribution to readability based on the analysis of gazes of people reading text with and without commas.

The experimental results show that our predictor reproduces the comma distribution in the Penn Chinese Treebank with 78.41 in F_1-score and commas chosen by our filter smoothen certain gaze behaviors.

1 Introduction

Chinese is an ideographic language, with no natural apparent word boundaries, little morphology, and no case markers. Moreover, most Chinese sentences are quite long. These features make it especially difficult for Chinese learners to identify composition of a word or a clause in a sentence.

Punctuation marks, especially commas, are allowed to be placed relatively arbitrarily to serve as important segmentation cues (Yue, 2006) for providing syntactic and prosodic boundaries in text; commas indicate not only phrase or clause boundaries but also sentence segmentations, and they capture some of the major aspects of a writer's prosodic intent (Chafe, 1988). The combination of both aspects promotes cognition when reading text (Ren and Yang, 2010; Walker et al., 2001).

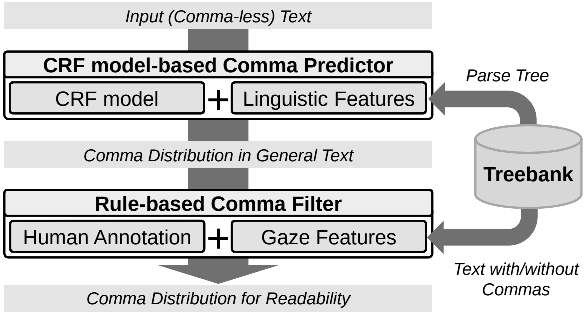


Figure 1: Our approach

However, although there are guidelines and research on the syntactic aspects of comma placement, prosodic aspects have not been explored, since they are more related with cognition. It is as yet unclear how comma placement should be optimized for reading, and it has thus far been up to the writer (Huang and Chen, 2011).

In this research, we attempt to optimize comma placements by integrating the linguistic features of text and the gaze features of readers. Figure 1 illustrates our approach. First, we design a comma predictor for general Chinese text based on conditional random field (CRF) models with various linguistic features. Second, we build a rule-based filter for classifying commas in text into ones facilitating or obstructing readability, by comparing the gaze features of persons reading text with and without commas. These two steps are connected by applying our rule-based filter to commas predicted by our comma predictor. The experimental results for each step validate our approach.

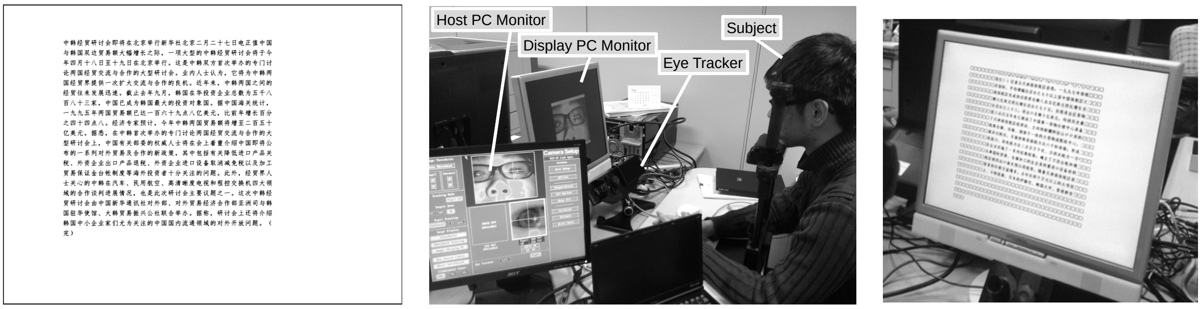
Related work is described in Section 2. The functions of Chinese commas are described in Section 3. Our CRF model-based comma predictor is examined in Section 4, and our rule-based comma filter is constructed and examined in Section 5 and 6. Section 7 contains a summary and outlines future directions of this research.

[*]The Japan Research Institute, Ltd. (from April, 2013)

Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, pages 49–58,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

[Case 1] When a pause between a subject and a predicate is needed. (* (,) means the original or comparative position of the comma in Chinese text.) e.g. **我们看得见的星星，绝大多数是离地球非常远的恒星。**(The stars we can see (,)* are mostly fixed stars that are far away from the earth.)
[Case 2] When a pause between an inner predicate and an object of a sentence is needed. e.g. **应该看到，科学需要一个人贡献出毕生的精力。**(We should see that (,) science needs a person to devote all his/her life to it.)
[Case 3] When a pause after an inner (adverbial, prepositional, etc.) modifier of a sentence is needed. e.g. **对于这个城市，他并不陌生。**(He is no stranger (,) to this city.) (The order of the modifier and the main clause is opposite in the English translation.)
[Case 4] When a pause between clauses in a complex sentence is needed, besides the use of semicolon (；). e.g. **据说苏州园林有一百多处，我到过的不过十多处。**(It is said that there are more than 100 Suzhou traditional gardens, (,) no more than 10 of which I have been to.)
[Case 5] When a pause between phrases of the same syntactic type is needed. e.g. **学生比较喜欢年轻，有活力的教师** (The students prefer young (,) and energetic teachers.)

Table 1: Five main usages of commas in Chinese text



中韩经贸研讨会即将在北京举行新华社北京二月二十七日电正值中国与韩国双边贸易额大幅增长之际，一项大型的中韩经贸研讨会将于今年四月十八日至十九日在北京举行。这是中韩双方首次举办的专门讨论两国经贸交流与合作的大型研讨会。业内人士认为，它将为中韩两国经贸界提供一次扩大交流与合作的良机。近年来，中韩两国之间的经贸往来发展迅速。截止去年九月，韩国在华投资企业总数为五千八百八十三家，中国已成为韩国最大的投资对象国。据中国海关统计，一九九五年两国贸易额已达一百六十九点八亿美元，比前年增长百分之四十四点八。经济专家预计，今年中韩两国贸易额将增至二百五十亿美元。据悉，在中韩首次举办的专门讨论两国经贸交流与合作的大型研讨会上，中国有关部委的权威人士将在会上着重介绍中国即将公布的一系列对外贸易及合作的新政策，其中包括有关降低进口产品关税、外资企业出口产品退税、外资企业进口设备取消减免税以及加工贸易保证金台帐制度等海外投资者十分关注的问题。此外，经贸界人士关心的中韩在汽车、民用航空、高清晰度电视和程控交换机四大领域的合作谈判进展情况，也是此次研讨会主要议题之一。这次中韩经贸研讨会由中国新华通讯社对外部、对外贸易经济合作部亚洲司与韩国驻华使馆、大韩贸易振兴公社联合举办。据称，研讨会上还将介绍韩国中小企业家们尤为关注的中国国内流通领域的对外开放问题。（完）

(a) Screenshot of a material (b) Scene of the experiment (c) Window around a gaze point

Figure 3: Settings for eye-tracking experiments

WS	Word surface
POS	POS tag
DIP	Depth of a word in the parse tree
STAG	Syntactic tag
OIC	Order of the clause in a sentence that a word belongs to
WL	Word length
LOD	Length of fragment with specific depth in a parsing tree

Table 2: Features used in our CRF model

2 Related Work

Previous work on Chinese punctuation prediction mostly focuses on sentence segmentation in automatic speech recognition (Shriberg et al., 2000; Huang and Zweig, 2002; Peitz et al., 2011).

Jin et al. (2002) classified commas for sentence segmentation and succeeded in improving parsing performance. Lu and Ng (2010) proposed an approach built on a dynamic CRF for predicting punctuations, sentence boundaries, and sentence types of speech utterances without prosodic cues. Zhang et al. (2006) suggested that a cascade CRF-based approach can deal with ancient Chinese prose punctuation better than a single CRF. Guo et al. (2010) implemented a three-tier maximum entropy model incorporating linguistically motivated features for generating commonly used Chinese punctuation marks in unpunctuated sentences output by a surface realizer.

```
((IP (PP (P 据)
     (NP (NN 了解)))
     (PU ，)
     (NP-TPC (CP (WHNP-1 (-NONE- *OP*))
          (CP (IP (NP-SBJ (-NONE- *T*-1))
               (VP (VV 新建)))
               (DEC 的)))
          (QP (CD 五)
          (CLP (M 家)))
          (NP (NN 保险)
          (NN 公司)) )
     (PU ，)
     (IP (IP (QP-SBJ (CD 三)
               (CLP (M 家)))
          (VP (VC 是)
          (DNP-PRD (ADJP (JJ 全国性))
                (DEG 的))))
     (PU ，)
     (IP (QP-SBJ (CD 两)
               (CLP (M 家)))
          (VP (VC 是)
          (DNP-PRD (ADJP (JJ 地区性))
                (DEG 的)))))
     (PU 。)))
```

(a)

```
WS|POS|STAG|DIP|OIC|WL|LOD|IOB-tag

据|P|PP|4|I0|1|1|O
了解|NN|NP|5|I0|2|3|B-Comma
*OP*|-NONE-|NP|6|C0|4|0|I-Comma
*T*-1|-NONE-|NP|8|I1|5|0|O
新建|VV|VP|8|I1|2|2|O
的|DEC|CP|6|C1|1|3|O
五|CD|QP|5|C0|1|4|O
家|M|CLP|6|C0|1|2|O
保险|NN|NP|5|I0|2|3|O
公司|NN|NP|5|I0|2|4|B-Comma
三|CD|QP|6|I2|1|5|I-Comma
家|M|CLP|7|I2|1|2|O
是|VC|VP|6|I2|1|2|O
全国性|JJ|ADJP|8|I2|3|3|O
的|DEG|DNP|7|I2|1|7|B-Comma
两|CD|QP|6|I2|1|5|I-Comma
家|M|CLP|7|I2|1|2|O
是|VC|VP|6|I2|1|2|O
地区性|JJ|ADJP|8|I2|3|3|O
的|DEG|DNP|7|I2|1|14|O
。|PU|IP|3|I1|1|27|O
```

(b)

Figure 2: Example of a parse tree (a) and its corresponding training data (b) with the features

3 Functions of Chinese Commas

There are five main uses of commas in Chinese text, as shown in Table 1. Cases 1 to 4 are from ZDIC.NET (2005), and Case 5 obviously exists in Chinese text. The first three serve the function of emphasis, while the latter two indicate coordinating or subordinating clauses or phrases.

In Cases 1 and 2, a comma is inserted as a kind of pause between a short subject and a long predicate, or between a short remainder predicate, such as **看到** (see/know), **説明/表明** (indicate), **発**

Feature	F_1 (P/R)	A
WS	59.32 (72.67/50.12)	95.45
POS	32.51 (69.06/21.26)	94.08
DIP	34.14 (68.65/22.72)	94.13
STAG	22.44 (64.00/13.60)	93.67
OIC	9.27 (66.56/ 4.98)	93.42
WL	10.70 (75.24/ 5.76)	93.52
LOD	35.32 (59.20/25.17)	93.81
WS+POS	63.75 (79.93/53.01)	96.03
WS +DIP	70.06 (83.27/60.47)	96.61
WS +STAG	57.42 (81.94/44.19)	95.67
WS +OIC	60.35 (77.98/49.22)	95.73
WS +WL	60.90 (76.39/50.63)	95.71
WS +LOD	70.85 (78.87/64.31)	96.53
WS+POS+DIP	73.41 (84.62/64.82)	96.93
WS+POS+DIP+STAG	74.58 (83.66/67.27)	97.01
WS+POS+DIP +OIC	76.87 (84.29/70.65)	97.23
WS+POS+DIP +WL	70.18 (83.33/60.62)	96.63
WS+POS+DIP +LOD	76.61 (82.61/71.43)	97.16
WS+POS+DIP+STAG+OIC	76.62 (84.48/70.09)	97.21
WS+POS+DIP+STAG +WL	74.12 (84.00/66.33)	96.98
WS+POS+DIP+STAG +LOD	77.64 (85.11/71.38)	97.33
WS+POS+DIP +OIC+WL	75.43 (84.76/67.95)	97.11
WS+POS+DIP +OIC +LOD	78.23 (84.23/73.03)	97.36
WS+POS+DIP +WL+LOD	74.01 (85.80/65.06)	97.02
WS+POS+DIP+STAG+OIC+WL	77.25 (83.97/71.53)	97.26
WS+POS+DIP+STAG+OIC +LOD	77.31 (86.36/69.97)	97.33
WS+POS+DIP+STAG +WL+LOD	76.55 (85.24/69.46)	97.23
WS+POS+DIP +OIC+WL+LOD	77.60 (84.30/71.89)	97.30
WS+POS+DIP+STAG+OIC+WL+LOD	78.41 (83.97/73.54)	97.36

F_1: F_1-Score, **P**: precision (%), **R**: recall (%), **A**: accuracy (%)

Table 3: Performance of the comma predictor

Article ID	(A) #Characters, (B) #Punctuations, (C) #Commas			(C) / (A)	(C) / (B)	Subjects
6	692	49	28	4.04%	57.14%	L, T, C
7	335	30	15	4.48%	50.00%	L, T, C
10	346	18	7	2.02%	38.89%	L, T, C, Z
12	221	18	7	3.17%	38.89%	L, T, C
14	572	33	14	2.45%	42.42%	L, T, C
18	471	36	13	2.76%	36.11%	C, Z
79	655	53	28	4.27%	52.83%	Z
82	471	30	13	2.76%	43.33%	Z
121	629	41	19	3.02%	46.34%	Z
294	608	50	24	3.95%	48.00%	Z
401	567	43	21	3.70%	48.84%	L, T, C
406	558	39	18	3.23%	46.15%	Z
413	552	52	22	3.99%	42.31%	T, C, Z
423	580	49	26	4.48%	53.06%	L, C, Z
438	674	46	28	4.15%	60.87%	Z
Average	528.73	39.13	18.87	3.57%	48.22%	-

Table 4: Materials assigned to each subject

見 (find) etc., and following long clause-style objects. English commas, on the other hand, seldom have such usages (Zeng, 2006). In Cases 3 and 4, commas instead of conjunctions sometimes connect two clauses in a relation of either coordination or subordination. English commas, on the other hand, are only required between independent clauses connected by conjunctions (Zeng, 2006).

Liu et al. (2010) proved that Chinese commas can change the syntactic structures of sentences by playing lexical or syntactic roles. Ren and Yang (2010) claimed that inserting commas as clause boundaries shortens the fixation time in post-comma regions. Meanwhile, in computational linguistics, Xue and Yang (2011) showed that Chinese sentence segmentation can be viewed as detecting loosely coordinated clauses separated by commas.

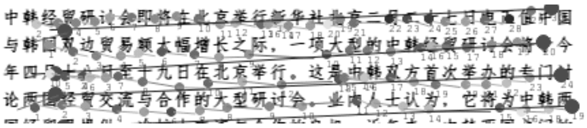

中韩经贸研讨会即将在北京举行新华社北京二月二十七日电 在中国
与韩国双边贸易额大幅增长之际，一项大型的中韩经贸研讨会将于今
年四月十七日至十九日在北京举行。这是中韩双方首次举办的专门讨
论两国经贸交流与合作的大型研讨会。业内人士认为，它将为中韩两

Figure 4: Obtained eye-movement trace map

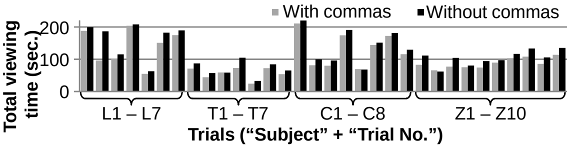


Figure 5: Total viewing time

4 CRF Model-based Comma Predictor

We first predict comma placements in existing text. The prediction is formalized as a task to annotate each word in a word sequence with an IOB-style tag such as I-Comma (following a comma), B-Comma (preceding a comma) or O (neither I-Comma nor B-Comma). We utilize a CRF model for this sequential labeling (Lafferty et al., 2001).

4.1 CRF Model for Comma Prediction

A conditional probability assigned to a label sequence Y for a particular sequence of words X in a first-order linear-chain CRF is given by:

$$P_\lambda(Y|X) = \frac{\exp(\sum_w^n \sum_i^k \lambda_i f_i(Y_{w-1}, Y_w, X, w))}{Z_0(X)}$$

where w is a word position in X, f_i is a binary function describing a feature for Y_{w-1}, Y_w, X, and w, λ_i is a weight for that feature, and Z_0 is a normalization factor over all possible label sequences.

The weight λ_i for each f_i is learned on training data. For f_i, the linguistic features shown in Table 2 are derived from a syntactic parse of a sentence[1]. The first three were used initially; the rest were added after we got feedback from construction of our rule-based filters (see Section 5). Figure 2 shows an example of a parsing tree and its corresponding training data.

[1] Some other features or tag formats which worked well in the previous research, such as bi-/tri-gram, a preceding word (L-1) or its POS (POS-1), and IO-style tag (Leaman and Gonzalez, 2008) were also examined, but they did not work that well, probably because of the difference in task settings.

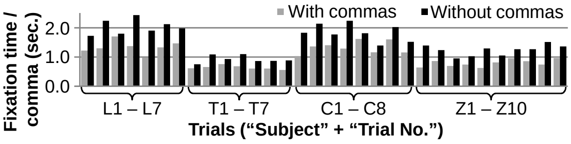


Figure 6: Fixation time per comma

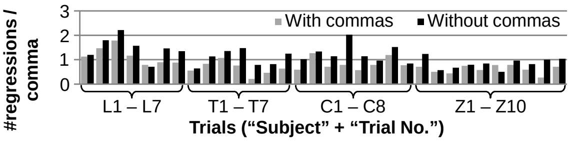


Figure 7: Number of regressions per comma

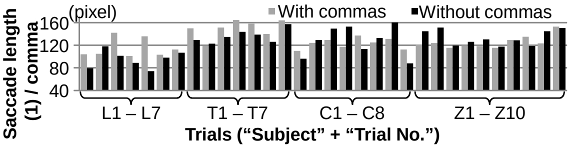


Figure 8: Saccade length (1) per comma

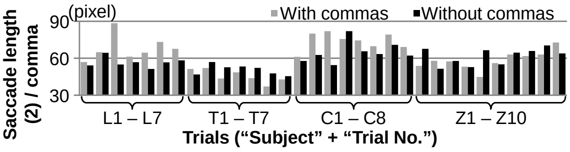


Figure 9: Saccade length (2) per comma

4.2 Experimental Settings

The Penn Chinese Treebank (CTB) 7.0 (Naiwen Xue and Palmer, 2005) consists of 2,448 articles in five genres. It contains 1,196,329 words, and all sentences are annotated with parse trees. We selected four genres for written Chinese (newswire, news magazine, broadcast news and newsgroups/weblogs) from this corpus as our dataset. These were randomly divided into training (90%) and test data (10%). We also corrected errors in tagging and inconsistencies in the dataset, mainly by solving problems around strange characters tagged as PU (punctuation). The commas and characters after this preprocessing numbered 63,571 and 1,533,928 in the training data and 4,116 and 111,172 in the test data.

MALLET (McCallum, 2002) and its application ABNER (Settles, 2005) were used to train the CRF model. We evaluated the results in terms of precision (P $= tp/(tp + fp)$), recall (R $= tp/(tp+fn)$), F$_1$-score (F$_1 = 2\mathrm{PR}/(\mathrm{P}+\mathrm{R})$), and accuracy (A $= (tp + tn)/(tp + tn + fp + fn)$), where tp, tn, fp and fn are respectively the number of true positives, true negatives, false positives and false negatives, based on whether the model and the corpus provided commas at each location.

4.3 Performance of the CRF Model

Table 3 shows the performance of our CRF model[2]. We can see that WS contributed much more to the performance than other features, probably because a word surface itself has a lot of information on both prosodic and syntactic functions. Combining WS with other features greatly improved performance, and as a result, with all features (WS + POS + STAG + DIP + OIC + LOD + WL), precision, recall, F$_1$-score and accuracy were 83.97%, 73.54%, 78.41 and 97.36%.

We also found that a large number of false positives seemed helpful according to native speakers (see the description of the subjects in Section 5 and 6). Although these commas do not appear in the CTB text, they might smoothen the reading experience. We constructed a rule-based filter in order to pick out such commas.

5 Rule-based Comma Filter

We constructed a rule-based comma filter for classifying commas in text into ones facilitating (positive) or obstructing (negative) the reading process as follows:
[Step 1]: Collect gaze data from persons reading text with or without commas (Section 5.1).
[Step 2]: Compare gaze features around commas to find those features that reflect the effect of comma placement. (Section 5.2).
[Step 3]: Annotate commas with categories based on the obtained features (Section 5.3), and devise rules to explain the annotation (Section 5.4).

5.1 Collecting Human Eye-movement Data

Eye-movements during reading contain rich information on how the document is being read, what the reader is interested in, where difficulties happen, etc. The movements are characterized by fixations (short periods of steadiness), saccades (fast movements), and regressions (backward saccades) (Rayner, 1998). In order to analyze the effect of commas on reading through the features, we collected gaze data from subjects reading text in the following settings.

[**Subjects and Materials**] Four native Man-

[2] Precision, recall, F$_1$-score, and accuracy with WS + POS + DIP + L-1 + POS-1 were 82.96%, 65.04%, 72.91 and 96.84%, respectively (lower than those with WS+POS+DIP).

Categories	**Effect on readability**	**Outward manifestation**
Positive (◯)	Can improve readability.	Presence would cause GF+.
Semi-positive (△)	Might be necessary for readability, but the importance is not as obvious as a positive comma.	Absence might cause GF-.
Semi-negative (□)	Might be negative, but its severity is not as obvious as a negative comma.	Absence might cause GF+.
Negative (×)	Thought to reduce a document's readability.	Presence would cause GF-.

GF+/GF-: values of eye-tracking features that represent good/poor readability

Table 5: Comma categories

Subject	**Positive (◯)**	**Semi-positive (△)**	**Semi-negative (□)**	**Negative (×)**	**Adjustment formula**
L	$\Delta\mathrm{FT}'>800$	$500<\Delta\mathrm{FT}'\leq 800$	$-100<\Delta\mathrm{FT}'\leq 500$	$\Delta\mathrm{FT}'<-100$	$\Delta\mathrm{FT}' = \Delta\mathrm{FT} + \Delta\mathrm{RT} \times 200$
C	$\Delta\mathrm{FT}'>900$	$600<\Delta\mathrm{FT}'\leq 900$	$-200<\Delta\mathrm{FT}'\leq 600$	$\Delta\mathrm{FT}'<-200$	$\Delta\mathrm{FT}' = \Delta\mathrm{FT} + \Delta\mathrm{RT} \times 275$
T	$\Delta\mathrm{FT}'>600$	$300<\Delta\mathrm{FT}'\leq 600$	$-300<\Delta\mathrm{FT}'\leq 300$	$\Delta\mathrm{FT}'<-300$	$\Delta\mathrm{FT}' = \Delta\mathrm{FT} + \Delta\mathrm{RT} \times 250$
Z	$\Delta\mathrm{FT}'>650$	$350<\Delta\mathrm{FT}'\leq 650$	$-250<\Delta\mathrm{FT}'\leq 350$	$\Delta\mathrm{FT}'<-250$	$\Delta\mathrm{FT}' = \Delta\mathrm{FT} + \Delta\mathrm{RT} \times 250$

$\Delta\mathrm{FT} = [\text{ fixation time (without commas) [ms]}] - [\text{ fixation time (with commas) [ms]}]$
$\Delta\mathrm{RT} = [\text{ \#regressions (without commas) }] - [\text{ \#regressions (with commas) }]$

Table 6: Estimation formula for judging the contribution of commas to readability

ID	◯	△	□	×
6	13	6	4	5
7	8	6	1	0
10	5	0	1	1
12	1	4	2	0
14	4	4	5	1
18	5	1	4	3
79	11	4	9	4
82	5	6	2	0
121	11	2	6	0
294	9	9	4	1
401	10	7	2	2
406	5	6	5	2
413	8	5	6	3
423	11	4	7	4
438	6	16	6	0
Total	112	80	64	26

Table 7: Categories of annotated commas

darin Chinese speakers (graduate students and researchers) read 15 newswire articles selected from CTB 7.0 (included in the test data in Section 4.2). Table 4 and Figure 3(a) show the materials assigned to each subject and a screenshot of one material. Each article was presented in 12-15 points of bold-faced Fang-Song font occupying 13×13, 14×15, 15×16 or 16×16 pixels along with a line spacing of 5-10 pixels[3].

[**Apparatus**] Figure 3(b) shows a scene of the experiment. An EyeLink 1000 eye tracker (SR Research Ltd., Toronto, Canada) with a desktop mount monitored the movements of a right eye at 1,000 Hz. The subject's head was supported at the chin and forehead. The distance between the eyes and the monitor was around 55 cm, and each Chinese character subtended a visual angle 1°. Text was presented on a 19" monitor at a resolution of 800×600 pixels, with the brightness adjusted to a comfortable level. The displayed article was masked except for the area around a gaze point (see Figure 3(c)) in order to confirm that the gaze point was correctly detected and make the subject concentrate on the area (adjusted for him/her).

[**Procedure**] Each article was presented twice (once with/once without commas) to each subject. The one without commas was presented first[4] (not necessarily in a row). We did not give any comprehension test after reading; we just asked the subjects to read carefully and silently at their normal or lower speed, in order to minimize the effect of the first reading on the second. The subjects were informed of the presence or absence of commas beforehand. The apparatus was calibrated before the experiment and between trials. The experiment lasted around two hours for each subject.

[**Alignment of eye-tracking data to text**] Figure 4 shows an example of the obtained eye-movement trace map, where circles and lines respectively mean fixation points and saccades, and color depth shows their duration. The alignment of the data to the text is a critical task, and although automatic approaches have been proposed (Martínez-Gómez et al., 2012a; Martínez-Gómez et al., 2012b), they do not seem robust enough for our purpose. Accordingly, we here just compared the entire layout of the gaze point distribution and that of the actual text, and adjusted them to have relatively coherent positions on the x-axis; i.e., the beginning and end of the gaze point sequence in a line were made as close as possible to those of the line in the text.

5.2 Analysis of Eye-movement Data

The gaze data were analyzed by focusing on regions around each comma or where each one should be (three characters left and right to the comma[5]).

[3] These values, as well as the screen position of the article, were adjusted for each subject.

[4] If we had used the reversed order, the subject would have knowledge about original comma distribution, and this would cause abnormally quick reading of the text without commas. With the order we set, conflicts between false segmentations (made in first reading) and correct ones might bother the subject, which is trade-off (though minor) in the second reading.

[5] When a comma appeared at the beginning of a line, two characters to the left and right of the comma and one charac-

1. If L_Seg and R_Seg are both very long, a comma must be put between them.
2. If two △ appear serially, one is necessary whereas the other might be optional or judged negative, but it still depends on the lengths of the siblings.
3. If two neighboring commas appear very close to each other, one of them is judged as negative whereas judgment on the other one is reserved.
4. If several (more than 2) ×s appear continually, one or more ×s might be reserved in consideration of the global condition.
5. A comma is always needed after a long sentence or clause without any syntactically significant punctuation with the function of segmentation.
6. If a △ appears near a ◯, it might be judged as negative with a high probability. However, the judgment process is always from the bottom up, which means × → □ → △ → ◯. For example, if a □ appears near a △, we judge □ first (to be positive or negative), then judge the △ in the condition with or without the comma of □.

Table 8: General rules for reference

Figure 5, 6 and 7 respectively show the total viewing time, fixation time (duration for all fixations and saccades in a target region) per comma, and number of regressions per comma[6] for each trial. We can see a general trend wherein the former two were shorter and the latter was smaller for the articles with commas than without. The diversity of the subjects was also observed in Figure 6.

Figure 8 and 9 show the saccade length per comma for different measures. The former (latter) figure considers a saccade in which at least one edge (both edges) was in the region. We cannot see any global trend, probably because of the difference in global layout of materials brought by the presence or absence of commas.

5.3 Categorization of Commas

Using the features shown to be effective to represent the effect of comma placement, we analyzed the statistics for each comma in order to manually construct an estimation formula for judging the contribution of each comma to readability. The contribution was classified into four categories (Table 5), and the formula is described in Table 6[7]. The adjustment formula was based on our observation that the number of regressions could only be regarded as an aid. For example, for subject C, if ΔFT=200ms and ΔRT $=-2$, ΔFT$'=-350$, and therefore, the comma is annotated as negative. All parameters were decided empirically and manually checked twice (self-judgment and feedback from the subjects).

On the basis of this estimation formula, all articles in Table 4 were manually annotated. Table 7 shows the distribution of the assigned categories[8].

5.4 Implementation of Rule-based Filter

The annotated commas were classified into Cases 1 to 5 in Table 1, based on the types of left and right segment conjuncts (L_Seg and R_Seg, which were obtained from the parse trees in CTB). For each of the five cases, the reason for the assignment of a category (◯, △, □ or ×) to each comma was explained by a manually constructed rule which utilized information about L_Seg and R_Seg. The rules were constructed so that they would cover as many instances as possible. Table 8 shows the general rules utilized as a reference, and Table 9 shows the finally obtained rules. The rightmost column in this table shows the number of commas matching each rule. These rules were then implemented as a filter for classifying commas in a given text.

For several rules (◯10, □8, □10, □11 and □12), there were only single instances. In addition, although our rules were built carefully, a few exceptions to the detailed threshold were found. Collecting and investigating more gaze data would help to make our rules more sophisticated.

6 Performance of the Rule-based Filter

We assumed that our comma predictor provides a CTB text with the same distribution as the original one in CTB (see Figure 1). Accordingly, we examined the quality of the comma categorization by our rule-based filter through gaze experiments.

6.1 Experimental Settings

Another five native Mandarin Chinese speakers were invited as test subjects. The CTB articles assigned to the subjects are listed in Table 10. These articles were selected from the test data in Section 4.2 in such a way that 520<*#characters*<700, *#commas*>17, *#commas*/*#punctuations*>38%, and *#commas*/*#characters*>3.1%, since we needed articles of appropriate length with a fair number of commas. After that, we manually chose articles that seemed to attract the subjects' interest from those that satisfied the conditions.

ter to the left and right of the final character of the last line were analyzed.

[6] Calculated by counting the instances where the x-position of [a fixation / end point of a saccade] was ahead of [the former fixation / its start point]. Although the counts of these two types were almost the same, by counting both of them, we expected to cover any possible regression.

[7] One or two features are used to judge the category of a comma. We will explore more features in the future.

[8] In the case of severe contradictions, the annotators discussed them and resolved them by voting.

Case 1: L_Subject + R_Predicate		**#commas**
◯6	L_IP-SBJ + R_VP (length both<14 (In_Seg_Len))	2
△7	L_IP-SBJ/NP-SBJ (Org_Len>13, Ttl_Len>15)	7
×6	L_NP-SBJ/IP-SBJ (<14) + R_VP (≥25)	2
Case 2: L_Predicate + R_Object		**#commas**
◯9	Long frontings (Modifier/Subject, >7) + short L_predicate (VV/VRD/VSB· · · , ≤3) + Longer R_object (IP-OBJ, >28)	6
△8	Short frontings (<5) + short L_predicate (<3) + moderate-length R_object (IP-SBJ, <20)	4
□6	Short frontings (<6) + short L_predicate (≤3) + long R_object (IP-SBJ, >23)	9
Case 3: L_Modifier		**#commas**
◯3	Short frequently used L_modifier (2-3, 经···, 据···, etc.) + moderate-length/long R_SPO (≥w18p10)	13
◯7	Short L_(PP/LCP)-TMP (5, 6) + long R_NP (≥10)	4
◯10	Long L_CP-CND (e.g., 若···, >18) + moderate-length R_Seg (SPO, IP, etc. <18)	1
△1	Long L_modifier (PP(-XXX, P+Long NP/IP), IP-ADV, ≥17)	6
△4	Moderate-length/short L_modifier (PP(-XXX, P+IP, There is IP inside, >6<15, cf. □6 (NP))	9
△9	Long L_(PP/LCP)-TMP (Ttl_Len≥10), short R_Seg (NP/ADVP, <3)	4
△10	Short L_(LCP/PP)-LOC (<8)	2
□2	Long L_LOC (or there is LCP inside PP, >10)	5
□3	Very short frequently used L_ADVP/ADV (2)	8
□5	Short L_(PP/LCP/NP)-TMP (4;5-6, when R_Seg is short (<10))	12
□4	Moderate-length PP(-XXX, P+NP, >8 ≤13) + R_Seg (SPO, IP, VO, MSPO, etc.)	6
□8	Short L_IP-CND (<8)	1
□11	Long L_PP-DIR (>20) + short R_VO (≤10)	1
×2	Very short L_(QP/NP/LCP)-TMP (≤3)	8
×5	Short frequently used L_modifier (as in ◯3, ≤3) + short/moderate-length R_Seg (SPO etc., <c20w9)	1
Case 4: L_c + R_c		**#commas**
◯2	L_c & R_c are both long (In_Seg_Len≥15; or one>13, the other near 20)	39
◯8	L_c is the summary of R_c	2
△2	Moderate-length L_c + R_c (both ≥10≤15; or one≥17, the other≤12)	25
△3	Moderate-length clause (>10), but connected with familiar CC or ADVP	6
△5	Three or more consecutive moderate-length clauses (all<15, and at least one ≤10)	12
×7	Very short L_c + R_c (both <5), something like slogan)	1
Case 5: L_p + R_p		**#commas**
◯1	Short coordinate modifiers (Both side <5)	4
◯4	Short L_p+R_p (both<c15w5, and at least one <10), but pre-L_p (e.g., SBJ) is too long (>18)	2
◯5	Between two moderate-length/long phrases (both ≥15; or L_p≥17, R_p=10-14; Or L_p=10-14, R_p>20)	39
◯11	Long pre-L_p (SBJ /ADV, etc. >16) + short L_p (≤5) + long R_p (≥18)	2
(△3	Moderate-length phrase (>10), but connected with familiar CC or ADVP)	(6)
△6	Three or more consecutive short/moderate-length phrases (both<15, at least one<8)	5
□1	Between short phrases (both ≤c13w5), and pre-L_p (SBJ/ADV, etc.) is short/moderate-length (<11)	13
□7	Coordinate VPs, and L_VP is a moderate-length VP (PP-MNR VP)	4
□9	Phrasal coordination between a long (≥18) and a short (<10) phrase	3
□10	Moderate-length coordinate VPs (>10<15), and R_VP has the structure like VP (MSP VP)	1
□12	Between two short/moderate-length NP phrases (both ≤15, e.g., L_NP-TPC+R_NP-SBJ)	1
×1	Moderate-length/short phrase ((i) c:one>10<18, The other >5≤10, w:one≤5, the other>5≤10; (ii) c:both≥10<15, w:both>5≤7), and pre-L_p (SBJ/ADV, etc.) is short (≤5)	13

· L_x/R_x: the left/right segment of a target comma which is x.
(x can be "p" (phrase) / "c" (clause), syntactic tags (with function tags) such as "VP" and "IP-SBJ", or general functions such as "Subject" and "Predicate".)
· Org_Len: the number of characters in a segment (including other commas or punctuation inside).
· In_Seg_Len/Ttl_Len: the number of characters between the comma and nearest punctuation (inside a long/outside a short target segment).
· SPO: subject + predicate + object, belonging to the outermost sentence. The length is defined in the similar way as In_Seg_Len.
· MSPO: modifier + subject + predicate + object. The length is defined in the similar way as In_Seg_Len.
· -XX or -XXX: arbitrary type of possible functional tag (or without any functional tag) connected with the former syntactic tag.
· $\leq ciwj$: #characters$\leq i$ and #words$\leq j$.
· In some cases (in Case 3, 4 and 5), the length is calculated after negative (or judged negative) commas are eliminated.
· The rules related with TMP are applied faster than ones related with LCP (in Case 3).
· △3 appears in both Case 4 (clause) and Case 5 (phrase). The number of commas is given by the sum of those in both cases.

Table 9: Entire classification of rules based on traditional comma categories

Article ID	(A) #Characters,	(B) #Punctuations,	(C) #Commas	(C) / (A)	(C) / (B)	Subjects
6	692	49	28	4.04%	57.14%	L, S, H
11	672	48	21	3.13%	43.75%	L, S, F
15	674	67	26	3.86%	38.81%	L, S, H
16	547	43	22	4.02%	51.16%	L, S, F
56	524	43	18	3.44%	41.86%	L, H, M
73	595	46	28	4.71%	60.87%	S, H, F, M
79	655	53	28	4.27%	52.83%	H, F, M
99	671	55	24	3.58%	43.64%	F, M
Average	628.75	50.50	24.38	3.88%	48.27%	-

Table 10: Materials assigned to each subject

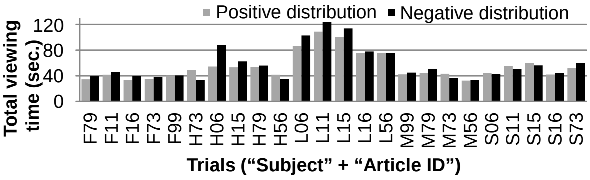


Figure 10: Total viewing time for two distributions

Our rule-based filter was applied to the commas of each article[9], and the commas were classified into two distributions: a positive one (positive + semi-positive commas) and a negative one (negative + semi-negative commas). Two types of materials were thus generated by leaving the commas in one distribution and removing the others.

[9] Instances of incoherence among the applied rules were manually checked and corrected.

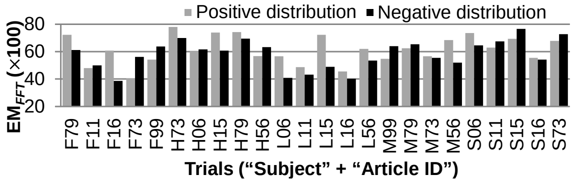


Figure 11: EM_{FFT} for two distributions

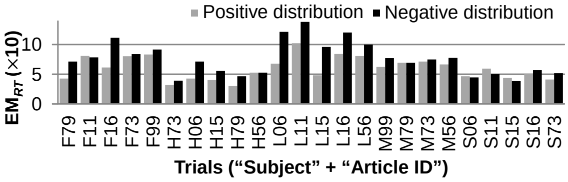


Figure 13: EM_{RT} for two distributions

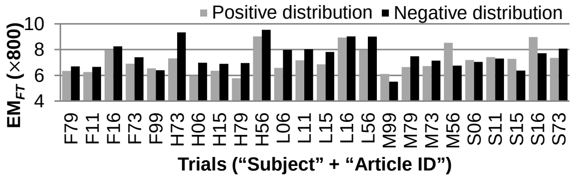


Figure 12: EM_{FT} for two distributions

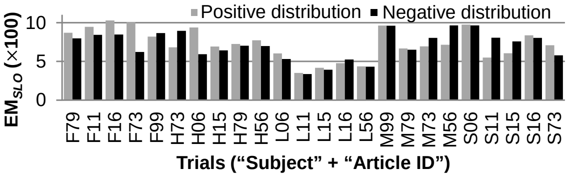


Figure 14: EM_{SLO} for two distributions

The apparatus and procedure were almost the same as those in Section 5.1, whereas, on the basis of the feedback from the previous experiments, the font size, number of characters in a line, and line spacing were fixed to single optimized values, respectively, 14-point Fang-Song font occupying 15×16 pixels, 33 characters and 7 pixels.

6.2 Evaluation Metrics

We examined whether our positive/negative distributions really facilitated/obstructed the subjects' reading process by using the following metrics:

TT, $\mathrm{EM}_{FFT} = \frac{\mathrm{FFT}}{\mathrm{FT}}$[10], $\mathrm{EM}_{FT} = \frac{\mathrm{FT}}{\mathrm{CN}\cdot\mathrm{TT}}$[11], $\mathrm{EM}_{RT} = \frac{\mathrm{RT}}{2\cdot\mathrm{CN}}$[12], $\mathrm{EM}_{\mathrm{SLO}} = \frac{\mathrm{SLO}}{2\cdot\mathrm{TT}}$,

where TT, FT, RT and CN are total viewing time, fixation time, number of regressions, and number of commas respectively, as described in Section 5.2. FFT and SLO are additionally introduced metrics respectively for the “total duration for all first-pass fixations in a target region that exclude any regressions” and for the “length of saccades from inside a target region to the outside”[13]. All of the areas around commas appearing in the original article were considered target areas for the metrics. The other settings were the same as in Section 5.

6.3 Contribution of Categorized Commas

Figure 10, 11, 12, 13 and 14 respectively show TT, EM_{FFT}, EM_{FT}, EM_{RT} and EM_{SLO} for two types of comma distributions in each trial.

For TT, we cannot see any general trend, mainly because this time, the reading order of the text was random, which spread out the second reading effect evenly between the two distributions. For EM_{FFT}, we cannot reach a conclusion either. In contrast, in more than half of the trials, EM_{FFT} was larger for positive distributions, which would imply that the positive commas helped to prevent the reader's gaze from revisiting the target regions. For most trials, except for subject S whose calibration was poor and reading process was poor in M56, EM_{FT} and EM_{RT} decreased and EM_{SLO} increased for positive distributions, which implies that the positive commas smoothed the reading process around the target regions.

7 Conclusion

We proposed an approach for modeling comma placement in Chinese text for smoothing reading. In our approach, commas are added to the text on the basis of a CRF model-based comma predictor trained on the treebank, and a rule-based filter then classifies the commas into ones facilitating or obstructing reading. The experimental results on each part of this approach were encouraging.

In our future work, we would like see how commas affect reading by using much more material, and thereby refine our framework in order to bring a better reading experience to readers.

Acknowledgments

This research was partially supported by Kakenhi, MEXT Japan [23650076] and JST PRESTO.

[10] Ratio to the total fixation time in the target areas (FT).

[11] Normalized by the total viewing time (TT).

[12] Two types of RT count (see Section 5.2) were averaged.

[13] Respectively to reflect “the early-stage processing of the region” and “the information processed for a fixation and a decision of the next fixation point” (Hirotani et al., 2006).

References

Wallace Chafe. 1988. Punctuation and the prosody of written language. *Written Communication*, 5:396–426.

Yuqing Guo, Haifeng Wang, and Josef van Genabith. 2010. A linguistically inspired statistical model for Chinese punctuation generation. *ACM Transactions on Asian Language Information Processing*, 9(2):6:1–6:27, June.

Masako Hirotani, Lyn Frazier, and Keith Rayner. 2006. Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language*, 54(3):425–443.

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*, pages 146–153.

Jing Huang and Geoffrey Zweig. 2002. Maximum entropy model for punctuation annotation from speech. In *Proceedings of the International Conference on Spoken Language Processing*, pages 917–920.

Mei xun Jin, Mi-Young Kim, Dongil Kim, and Jong-Hyeok Lee. 2002. Segmentation of Chinese long sentences using commas. In *Proceedings of the Third SIGHAN Workshop on Chinese Language Processing*, pages 1–8.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Robert Leaman and Graciela Gonzalez. 2008. BANNER: An executable survery of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing (PSB'08)*, pages 652–663.

Baolin Liu, Zhongning Wang, and Zhixing Jin. 2010. The effects of punctuations in Chinese sentence comprehension: An erp study. *Journal of Neurolinguistics*, 23(1):66–68.

Wei Lu and Hwee Tou Ng. 2010. Better punctuation prediction with dynamic conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*, pages 177–186.

Pascual Martínez-Gómez, Chen Chen, Tadayoshi Hara, Yoshinobu Kano, and Akiko Aizawa. 2012a. Image registration for text-gaze alignment. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (IUI '12)*, pages 257–260.

Pascual Martínez-Gómez, Tadayoshi Hara, Chen Chen, Kyohei Tomita, Yoshinobu Kano, and Akiko Aizawa. 2012b. Synthesizing image representations of linguistic and topological features for predicting areas of attention. In Patricia Anthony, Mitsuru Ishizuka, and Dickson Lukose, editors, *PRICAI 2012: Trends in Artificial Intelligence*, pages 312–323. Springer.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit.

Fu-dong Chiou Naiwen Xue, Fei Xia and Marta Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Stephan Peitz, Markus Freitag, Arne Mauser, and Hermann Ney. 2011. Modeling punctuation prediction as machine translation. In *Proceedings of International Workshop on Spoken Language Translation*, pages 238–245.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Gui-Qin Ren and Yufang Yang. 2010. Syntactic boundaries and comma placement during silent reading of Chinese text: evidence from eye movements. *Journal of Research in Reading*, 33(2):168–177.

Burr Settles. 2005. ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154.

Judy Perkins Walker, Kirk Fongemie, and Tracy Daigle. 2001. Prosodic facilitation in the resolution of syntactic ambiguities in subjects with left and right hemisphere damage. *Brain and Language*, 78(2):169–196.

Nianwen Xue and Yaqin Yang. 2011. Chinese sentence segmentation as comma classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers*, pages 631–635.

Ming Yue. 2006. Discursive usage of six Chinese punctuation marks. In *Proceedings of the COLING/ACL 2006 Student Research Workshop*, pages 43–48.

ZDIC.NET. 2005. Commonly used Chinese punctuation usage short list. *Long Wiki*, Retrieved Dec 10, 2012, from http://www.zdic.net/appendix/f3.htm. (in Chinese).

X. Y. Zeng. 2006. The comparison and the use of English and Chinese comma. *College English*, 3(2):62–65. (in Chinese).

Kaixu Zhang, Yunqing Xia, and Hang Yu. 2006. CRF-based approach to sentence segmentation and punctuation for ancient Chinese prose. *Journal of Tsinghua Univ (Science and Technology)*, 49(10):1733–1736. (in Chinese).