Automatic Comma Insertion for Japanese Text Generation

Masaki Murata
Graduate School of
Information Science,
Nagoya University, Japan
murata@el.itc.nagoya-u.ac.jp

Tomohiro Ohno
Graduate School of
International Development,
Nagoya University, Japan
ohno@nagoya-u.jp

Shigeki Matsubara
Graduate School of
Information Science,
Nagoya University, Japan
matubara@nagoya-u.jp

Abstract

This paper proposes a method for automatically inserting commas into Japanese texts. In Japanese sentences, commas play an important role in explicitly separating the constituents, such as words and phrases, of a sentence. The method can be used as an elemental technology for natural language generation such as speech recognition and machine translation, or in writing-support tools for non-native speakers. We categorized the usages of commas and investigated the appearance tendency of each category. In this method, the positions where commas should be inserted are decided based on a machine learning approach. We conducted a comma insertion experiment using a text corpus and confirmed the effectiveness of our method.

1 Introduction

In Japanese sentences, commas are inserted to mark word boundaries that might be otherwise unclear because Japanese is a non-segmented language. They are also inserted at sharp semantic boundaries to improve the readability of a sentence. While there is a tendency about the positions where commas should be inserted in a Japanese sentence, there is no clear standard for these positions. Therefore, it is hard for non-natives of Japanese such as foreign students to insert commas properly, and the method for automatic comma insertion is required to support sentence generation by such people. In addition, this method is expected to be useful for improving readability of texts generated by automatic speech recognition or machine translation.

This paper proposes a method for automatically inserting commas into Japanese texts. There are several usages of commas, and the positions to insert commas depend on these usages. Therefore, we grouped the usages of commas into nine categories, and investigated the appearance tendency for each category to find the effective features of machine learning by using Japanese newspaper articles. Based on the analysis of comma positions, our method decides whether or not to insert a comma at each *bunsetsu*[1] boundary in an input sentence by machine learning.

We conducted an experiment on comma insertion using the Kyoto Text Corpus (Kurohashi and Nagao, 1998), and obtained higher recall and precision than those of the baseline, leading us to confirm the effectiveness of our method.

This paper is organized as follows: The next section presents related works. Section 3 gives preliminary analyses. Section 4 explains how our comma insertion method works. An experiment and discussions are presented in Sections 5 and 6, respectively.

2 Related Works

There have been many investigations on comma insertion into output texts of speech recognition systems to improve the readability (Christensen et al., 2001; Kim and Woodland, 2001; Liu et al., 2006; Shimizu et al., 2008). Their methods insert commas using pause information of speakers, based on the idea that a point at which a speaker takes a breath partly corresponds to a point where a comma is inserted. However, since pause information cannot be obtained from texts, we cannot use this approach because our targets are written texts.

In addition, there have been some investigations

[1] *Bunsetsu* is a linguistic unit in Japanese that roughly corresponds to a basic phrase in English. A bunsetsu consists of one independent word and zero or more ancillary words.

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 892–901,
MIT, Massachusetts, USA, 9-11 October 2010.

on comma insertion into non-Japanese written texts (White and Rajkumar, 2008; Guo et al., 2010). In Japanese, there are several usages of commas, and some usages are specific to Japanese due to its linguistic nature. Therefore, just adopting the above mentioned methods, which have been developed to process non-Japanese texts, is not sufficient to enable high-quality comma insertion into Japanese sentences. Development of a method based on the detailed analysis of Japanese commas is required.

Furthermore, there have been some investigations on comma insertion into Japanese written texts (Hayashi, 1992; Suzuki et al., 1995). These investigations have adopted rule-based methods. However, the number of their rules is not necessarily sufficient, and no quantitative evaluation has been performed.

3 Analyses on Comma Usages

There have been several discussions on commas, including the draft of "Kutou-hou (punctuation)" made by Archives Division, Minister's Secretariat, Japanese Ministry of Education, Science and Culture in 1906. There are several usages of commas, and depending on the usage, the types of positions where commas are inserted are different. First, we examined some previous publications on commas (Honda, 1982; Inukai, 2002; Shogakukan's editorial department, 2007). Based on the results of the examination, we classified the usages of commas into nine categories shown in Table 1. Here, commas in Japanese sentences and commas in English sentences have some common roles. In Japanese sentences, some commas have the same roles as commas in English sentences, but some commas have roles specific to Japanese due to its linguistic nature such as "Japanese is a non-segmented language" or "Japanese has *kanji* characters and *katakana* characters."

In our study, positions where a comma should be inserted are detected by using machine learning. We investigated the Kyoto Text Corpus version 4.0 (Kurohashi and Nagao, 1998) to find the effective features. The Kyoto Text Corpus is a collection of Japanese articles of Mainichi newspaper. We used the articles on January 1st and from January 3rd to 11th in 1995 as the analysis data. Table 2 shows the size of the data. The data had been manually annotated with information on morphological analysis, bunsetsu segmentation and dependency[2] analysis. Clause boundaries were detected by the clause boundary detection program CBAP (Kashioka and Maruyama, 2004).

Table 1: Categorization of usages of commas

#	usage of comma
1	commas between clauses
2	commas indicating clear dependency relations
3	commas for avoiding reading mistakes and reading difficulty
4	commas indicating the subject
5	commas inserted after a conjunction or adverb at the beginning of a sentence
6	commas inserted between parallel words or phrases
7	commas inserted after an adverbial phrase to indicate time
8	commas emphasizing the adjacent word
9	other

Table 2: Size of the analysis data

sentences	11,821
bunsetsus	117,501
characters	503,970
commas	16,595
characters per sentence	42.63

Out of all the inserted commas, only 1.43% were inserted at positions which were not bunsetsu boundaries. Therefore, we analyzed only commas inserted at bunsetsu boundaries. Of 105,680 bunsetsu boundaries, commas were inserted at 16,357 bunsetsu boundaries, that is, the rate of comma insertion was 15.48%. In the following sections, we focus on morphemes, clause boundaries, dependency relation and the number of characters between commas, and investigate their relations with commas.

3.1 Commas between Clauses

If a sentence consists of several clauses, inserting a comma between clauses makes clear the sentence

[2] A dependency in a Japanese sentence is a modification relation in which a modifier bunsetsu depends on a modified bunsetsu. That is, the modifier bunsetsu and the modified bunsetsu work as a modifier and a modifyee, respectively.

Table 3: Rates of comma insertion according to the clause boundary type

type of clause boundary	ratio of comma insertion (%)	
topicalized element-*wa*	16.94	(1,446/8,536)
adnominal clause	0.72	(43/5,960)
continuous clause	84.57	(2,685/3,175)
compound clause-*te*	23.31	(394/1,690)
quotational clause	4.40	(74/1,680)
supplement clause	17.53	(245/1,398)
discourse marker	60.13	(650/1,081)
compound clause-*ga*	93.85	(946/1,008)
compound clause-*de*	84.52	(606/717)
condition clause-*to*	81.66	(423/518)

structure. Therefore, a clause boundary is considered to be a strong candidate of a position where a comma is inserted. For example, in the following sentence[3]:

- 国連による対イラク制裁解除に向け、関係の深い仏に一層の協力を求めるのが狙いとみられる。
  (Toward lifting the sanctions imposed on Iraq by United Nations, the aim seems to be to request further cooperation from France, which has close ties to Iraq.)

a comma is inserted at the clause boundary right after the continuous clause "国連による対イラク制裁解除に向け (Toward lifting the sanctions imposed on Iraq by United Nations)." Like this example, the same usage of commas is seen in English as well.

In the analysis data, there existed 29,278 clause boundaries excluding sentence breaks. Among them, commas were inserted at 8,805 positions (30.01%). The rate is higher than that of bunsetsu boundaries. This indicates that commas tend to be inserted at clause boundaries.

We investigated the rate of comma insertion about 114 types[4] of clause boundaries. Table 3 shows the top 10 clause boundary types according to the occurrence frequency, and the rates of comma insertion. In cases of "continuous clause" and "compound clause-*de*," the rates were higher than 84%. On the other hand, in cases of "adnominal clause" and "quotational clause," the rates were lower than 5%. This means that the likelihoods of comma insertion are different according to the clause boundary type.

[3]We underlined commas which we mentioned in the example and the corresponding positions in the translation of the example.

[4]In our research, we used the types of clause boundaries defined by the Clause Boundary Annotation Program (Kashioka and Maruyama, 2004).

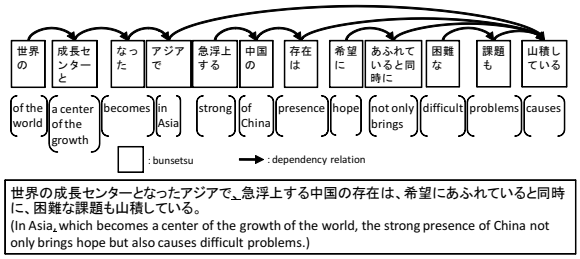

Figure 1: Commas making clear dependency relations

3.2 Commas and Dependency Structure

Commas have a role to make dependency relations clearer. Commas tend to be inserted right after a bunsetsu that depends on a distant bunsetsu. In Figure 1, although the bunsetsu "アジアで (in Asia)" depends on the bunsetsu "山積している (causes)," if the comma right after the bunsetsu "アジアで (in Asia)" is not inserted, the readers might mistakenly understand that the bunsetsu "アジアで (in Asia)" depends on the next bunsetsu "急浮上する (strong)." To avoid the mistake, the comma is inserted.

In the analysis data, there existed 66,984 bunsetsus which depend on the next bunsetsu. Among the bunsetsu boundaries right after them, 2,302 (3.44%) were the positions where a comma was inserted. On the other hand, in the case of a bunsetsu boundary right after a bunsetsu which does not depend on the next bunsetsu, the rate of comma insertion was 36.32% (14,055/38,696).

In addition, when the modifyee of a bunsetsu is located outside the clause containing the bunsetsu, i.e. to the right of the clause end, commas are considered to be more frequently inserted right after the bunsetsu because such bunsetsu causes more complex dependency structure. The rate of comma insertion right after such bunsetsu is 54.24%.

3.3 Commas for Avoiding Reading Mistakes and Reading Difficulty

Although, unlike English, Japanese is a non-segmented language, word boundaries are easy to detect because Japanese has three types of characters; *hiragana* characters, *katakana* characters, and *kanji* characters. However, if the same types of characters appear sequentially, readers may make a reading mistake or feel difficulty in reading them. To avoid such mistakes and difficulty, there is a usage of commas specific to Japanese.

In the following example, a comma is inserted between two sequentially appearing words "焼却 (burned)" and "灰 (ashes)" both of which consist of only *kanji* characters.

- 川崎さんの遺体を群馬県利根郡片品村花咲の知人宅に運んで焼却、灰を同村の山林に捨てたことを認める内容という。(He seemed to acknowledge that he had carried the corpse of Mr. Kawasaki to an acquaintance in Hanasaki, Katashina-mura, Tone-gun, Gunma Prefecture, burned it, and abandoned its ashes in the mountain forest in Katashina-mura.)

The comma was inserted because if there was no comma, the word boundary would become unclear and reading difficulty would be caused. Among 2,409 bunsetsu boundaries over which *kanji* characters appeared sequentially, commas were inserted at 2,188 (90.83%) bunsetsu boundaries. In the case of *katakana* characters, the rate was 97.69% (211/216). Commas tend to be inserted at most bunsetsu boundaries if *kanji* characters or *katakana* characters sequentially appear over a boundary.

3.4 Commas Indicating the Subject

Commas are considered to be inserted right after a bunsetsu that represents the subject of a sentence. For example, in Figure 2, a comma is inserted right after the bunsetsu "戦火は (war)" to indicate that the bunsetsu is the subject of the sentence. Here, we pay attention to the clause boundary of the type "topicalized element-*wa*." The rate that commas were inserted at the clause boundaries "topicalized element-*wa*" was 16.94% (1,446/8,536). This rate is almost the same as that of bunsetsu boundaries. On the other hand, the commas inserted at the clause boundaries "topicalized element-*wa*" accounted for 8.84% (1,446/16,357) of all the inserted commas.

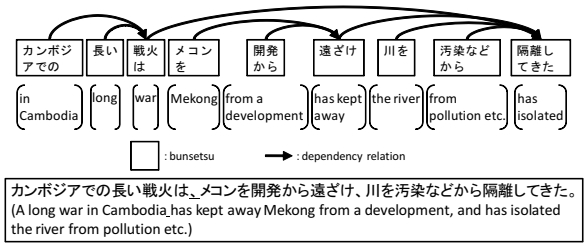

Figure 2: Comma insertion at the clause boundary "topicalized element-*wa*"

In the case of the clause boundary "topicalized element-*wa*" right after a bunsetsu which does not depend on the next bunsetsu (e.g., the bunsetsu "戦火は (war)" in Figure 2), the rate of comma insertion was 20.71% (1,426/6,886). The rate is higher than that of all the clause boundaries "topicalized element-*wa*." This shows that commas tend to be especially inserted at the "topicalized element-*wa*" right after bunsetsus which do not depend on the next bunsetsu.

3.5 Commas after Conjunction or Adverb

Commas tend to be inserted right after a conjunction or an adverb located at the beginning of a sentence. These commas correspond to English commas which are inserted right after a word such as "however" and "furthermore" located at the beginning of a sentence.

- しかし、私はそれに同感する気持ちになれない。(However, I do not feel like agreeing on it.)

In the analysis data, there existed 695 bunsetsus whose rightmost morpheme is a conjunction and which are located at the beginning of a sentence. Among them, commas were inserted right after 498 (71.65%) bunsetsus. In the case of bunsetsus whose rightmost morpheme is an adverb, the rate was 30.97% (140/452).

3.6 Commas Inserted between Parallel Words or Phrases

Commas have a function which makes clear separation between parallel words or phrases. The following example shows commas separating parallel nouns.

- むしろ地球規模の環境、人口、食糧など広範に国連の果たさなければならない役割は大きい。(The United Nations should play a lot of roles in a broad range of fields, such as the global environment, population, and food.)

In this example, commas are inserted to separate parallel nouns "環境 (environment)," "人口 (population)" and "食糧 (food)". In English, there are commas which perform the same role. In fact, commas were inserted between "environment" and "population" and between "population" and "food" in the translation of the above example. When bunsetsus whose rightmost morpheme is a noun appear sequentially, the rate of comma insertion between such bunsetsus is 59.39% (3,330/5,607).

Also, commas are inserted to separate parallel phrases. In the following example,

- メニューは前夜、首相が何を食べたかを調べて同じ献立を避けたり、和食と洋食のバランスを考えたりして決める。(The menu is decided by avoiding the menu the Prime Minister ate on the previous night, and by considering the balance between the Japanese food and the European food.)

a comma is inserted right after the bunsetsu "避けたり (avoiding)" to make clear separation between the parallel phrases "同じ献立を避けたり (by avoiding the menu)" and "和食と洋食のバランスを考えたりして (by considering the balance between the Japanese food and the European food)." The rate of comma insertion between two parallel phrases is 79.89% (751/940). This is much higher than that of bunsetsu boundaries, indicating that commas tend to be inserted when phrases are paralleled.

3.7 Number of Characters between Commas

If there are too many commas at a short distance, the sentence becomes hard to read. Therefore, the number of characters between commas is expected to be not too small. Also, because a long sequence of characters without a comma is generated if the distance between commas is very long, the occurrence frequency of such sequences of characters is considered to be low.

We investigated the number of characters between commas and its occurrence frequency. Figure 3 shows the results of investigation. When the number of characters between commas is either large or small, the occurrence frequency is low.

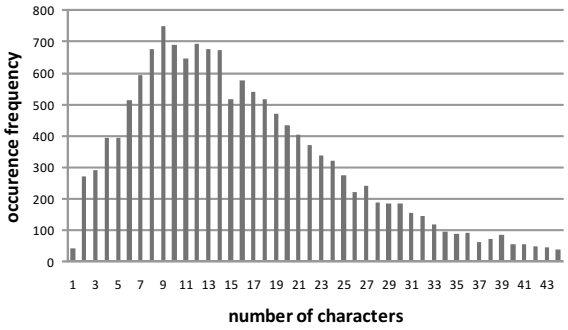

Figure 3: Number of characters between commas and its occurrence frequency

4 Comma Insertion Method

In our method, a sentence, on which morphological analysis, bunsetsu segmentation, clause boundary analysis and dependency analysis have been performed, is considered the input. Our method decides whether or not to insert a comma at each bunsetsu boundary in an input sentence. Based on the analysis results in Section 3, our method adopts the bunsetsu boundaries as candidate positions where a comma is inserted. Our method identifies the most appropriate combination among all combinations of positions where a comma can be inserted, by using the probabilistic model. In this paper, input sentences which consist of n bunsetsus are represented by $B = b_1 \cdots b_n$, and the results of comma insertion by $R = r_1 \cdots r_n$. Here, r_i is 1 if a comma is inserted right after bunsetsu b_i, and 0 otherwise. We indicate the j-th sequence of bunsetsus created by dividing an input sentence into m sequences as $L_j = b_1^j \cdots b_{n_j}^j (1 \leq j \leq m)$, and then, $r_k^j = 0$ if $1 \leq k < n_j$, and $r_k^j = 1$ if $k = n_j$.

4.1 Probabilistic Model for Comma Insertion

When an input sentence B is provided, our method identifies the comma insertion R that maximizes the conditional probability $P(R|B)$. Assuming that whether or not to insert a comma right after a bunsetsu is independent of other commas except the

Table 4: Features used for the maximum entropy method

morphological information	the rightmost independent morpheme, i.e. head word, (part-of-speech and inflected form) and rightmost morpheme (part-of-speech) of a bunsetsu b_k^j
	the rightmost morpheme (a surface form) of b_k^j if the rightmost morpheme is a particle
	the first morpheme (part-of-speech) of b_{k+1}^j
commas inserted between clauses	whether or not a clause boundary exists right after b_k^j
	type of a clause boundary right after b_k^j if there exists a clause boundary
commas indicating clear dependency relations	whether or not b_k^j depends on the next bunsetsu
	whether or not b_k^j depends on a bunsetsu located after the final bunsetsu of the clause including the next bunsetsu of b_k^j
	whether or not b_k^j is depended on by the bunsetsu located right before it
	whether or not the dependency structure of a sequence of bunsetsus between b_k^j and b_1^j is closed
commas avoiding reading mistakes and reading difficulty	whether or not both the rightmost morpheme of b_k^j and first morpheme of b_{k+1}^j are *kanji* characters
	whether or not both the rightmost morpheme of b_k^j and first morpheme of b_{k+1}^j are *katakana* characters
commas indicating the subject	whether or not there exists a clause boundary "topicalized element-*wa*" right after b_k^j and b_k^j depends on the next bunsetsu
	whether or not there exists a clause boundary "topicalized element-*wa*" right after b_k^j and the string of characters right before b_k^j is "では (*dewa*)"
	the number of characters in a phrase indicating the subject[5] if there exists a clause boundary "topicalized element-*wa*" right after b_k^j
	whether or not a clause boundary "topicalized element-*wa*" exists right after b_k^j and a bunsetsu whose rightmost morpheme is a verb depends on the modified bunsetsu of b_k^j
commas inserted after a conjunction or adverb at the beginning of a sentence	whether or not b_k^j appears at the beginning of a sentence and its rightmost morpheme is a conjunction
	whether or not b_k^j appears at the beginning of a sentence and its rightmost morpheme is an adverb
commas inserted between parallel words or phrases	whether or not both the rightmost morphemes of b_k^j and b_{k+1}^j are nouns
	whether or not a predicate at the sentence end is depended on by b_k^j whose rightmost independent morpheme is a verb and by any of the bunsetsus which are located after b_k^j and of which the rightmost independent morpheme is a verb
number of characters from b_1^j to b_k^j	one of the following 4 categories if the number of characters from b_1^j to b_k^j is found there ([num $= 1$], [$2 \leq$ num ≤ 3], [$4 \leq$ num ≤ 21], [$22 \leq$ num])

one appearing immediately before that bunsetsu, $P(R|B)$ can be calculated as follows:

$$
\begin{aligned}
& P(R|B) \qquad (1)\\
=& P(r_1^1 = 0, \cdots, r_{n_1-1}^1 = 0, r_{n_1}^1 = 1, \cdots,\\
& \quad r_1^m = 0, \cdots, r_{n_m-1}^m = 0, r_{n_m}^m = 1|B)\\
\cong& P(r_1^1 = 0|B) \times \cdots\\
& \times P(r_{n_1-1}^1 = 0|r_{n_1-2}^1 = 0, \cdots, r_1^1 = 0, B)\\
& \times P(r_{n_1}^1 = 1|r_{n_1-1}^1 = 0, \cdots, r_1^1 = 0, B) \times \cdots\\
& \times P(r_1^m = 0|r_{n_{m-1}}^{m-1} = 1, B) \times \cdots\\
& \times P(r_{n_m-1}^m = 0|r_{n_m-2}^m = 0, \cdots, r_1^m = 0, r_{n_{m-1}}^{m-1} = 1, B)\\
& \times P(r_{n_m}^m = 1|r_{n_m-1}^m = 0, \cdots, r_1^m = 0, r_{n_{m-1}}^{m-1} = 1, B)
\end{aligned}
$$

where $P(r_k^j = 1|r_{k-1}^j = 0, \cdots, r_1^j = 0, r_{n_{j-1}}^{j-1} = 1, B)$ is the probability that a comma is inserted right after a bunsetsu b_k^j when the sequence of bunsetsus B is provided and the position of j-th comma is identified. Similarly, $P(r_k^j = 0|r_{k-1}^j = 0, \cdots, r_1^j = 0, r_{n_{j-1}}^{j-1} = 1, B)$ is the probability that a comma is not inserted right after a bunsetsu b_k^j. These probabilities are estimated by the maximum entropy method. The result R which maximizes the conditional probability $P(R|B)$ is regarded as the most appropriate result of comma insertion, and calculated by dynamic programming.

4.2 Features on Maximum Entropy Method

To estimate $P(r_k^j = 1|r_{k-1}^j = 0, \cdots, r_1^j = 0, r_{n_{j-1}}^{j-1} = 1, B)$ and $P(r_k^j = 0|r_{k-1}^j = 0, \cdots, r_1^j = 0, r_{n_{j-1}}^{j-1} = 1, B)$ by the maximum entropy method, we used the features in Table 4 based on the analysis described in Section 3.

5 Experiment

To evaluate the effectiveness of our method, we conducted an experiment using a Japanese text corpus.

5.1 Outline of Experiment

As the experimental data, we used the newspaper articles in the Kyoto Text Corpus version 4.0 (Kurohashi and Nagao, 1998). We used the articles from January 14th to 17th as the test data. The training data is same as the analysis data. Table 5 shows the size of the test data. Here, we used the maximum entropy method tool (Le, 2008) with the default options except "-i 2000."

In the evaluation, we obtained the recall, the precision and their harmonic mean, i.e., F-measure. The recall and precision are respectively defined as follows.

$$recall = \frac{\#\ of\ correctly\ inserted\ commas}{\#\ of\ commas\ in\ the\ correct\ data}$$

$$precision = \frac{\#\ of\ correctly\ inserted\ commas}{\#\ of\ inserted\ commas}$$

In our research, to realize automatic comma insertion with high quality, we analyzed each usage of commas and decided the features for the ME method based on the analysis. To confirm the effectiveness of our features, we established the baseline method as a comparative method whereby commas are inserted by the ME method in which only simple morphological information is used. The baseline method uses the morphological information in Table 4 and the information of the rightmost morpheme (a surface form) of a bunsetsu as features.

5.2 Experimental Results

Table 6 shows the experimental results of the baseline and our method. The recall and precision were 69.13% and 84.13% respectively, and we confirmed that our method had higher performance than the baseline method. The percentage of sentences wherein all commas were correctly inserted was 55.81%.

Table 5: Size of test data

sentences	4,659
bunsetsus	46,511
characters	198,899
commas	6,549
characters per sentence	42.69

Table 6: Experimental results

	recall	precision	F-measure
our method	69.13% (4,527/6,549)	84.13% (4,527/5,381)	75.90
baseline	51.38% (3,365/6,549)	70.90% (3,365/4,746)	59.58

Figure 4 shows the comparison between the results of our method and the baseline method. The baseline method was not able to insert commas right after the bunsetsu "浮かんでいるが (are floated)" or "決まらないため (not decided)" but inserted commas at unnatural positions such as between "名乗る (calling himself)" and "副司令官が (the vice commander)." On the other hand, our method was able to insert commas properly at such bunsetsu boundaries.

6 Discussion

6.1 Error Analysis

Among positions where commas existed in the test data, there existed 2,022 positions where our method did not insert commas. Among them, 862 were clause boundaries, and the clause boundary "topicalized element-*wa*" accounted for 53.36% (460/862) of them. There were a lot of clause boundaries of the type "topicalized element-*wa*," and the number of commas inserted at such boundaries was large. But, the rate of comma insertion itself was not very high. We can say that the four features about "topicalized element-*wa*" did not always work well. Ta-

[5] Phrases indicating the subject is a sequence of bunsetsus consisting of b_k^j and all bunsetsus that are connected to b_k^j when we trace their dependency relationship in modifier-to-modifyee direction.

our method:

候補者として石原信雄内閣官房副長官や岩國哲人・島根県出雲市長、鳩山邦夫前労相、作家の堺屋太一氏らの名前が浮かんでいるが、前提となる政党の枠組みが決まらないため、調整は難航。
(Nobuo Ishihara, the deputy chief cabinet secretary, Tetsundo Iwakuni, the mayor of Izumo city in Shimane, Kunio Hatoyama, the former labor minister and the writer Taichi Sakaiya are floated as the candidate, however, since the framework of the predicated political party is not decided, the coordination makes little headway.)

baseline:

候補者として石原信雄内閣官房副長官や岩國哲人・島根県出雲市長鳩山邦夫前労相、作家の堺屋太一氏らの名前が浮かんでいるが前提となる政党の枠組みが決まらないため調整は難航。
(Nobuo Ishihara, the deputy chief cabinet secretary, Tetsundo Iwakuni, the mayor of Izumo city in Shimane_Kunio Hatoyama, the former labor minister and the writer Taichi Sakaiya are floated as the candidate_however, since the framework of the predicated political party is not decided_the coordination makes little headway.)

our method:

マルコスと名乗る副司令官が表に出てくるが、実際の司令官は不明。
(While the vice commander_calling himself Marcos appears in public, an actual commander is uncertain.)

baseline:

マルコスと名乗る、副司令官が表に出てくるが、実際の司令官は不明。
(While the vice commander, calling himself Marcos appears in public, an actual commander is uncertain.)

Figure 4: Comparison of the results of our method and baseline method

ble 7 shows the results of comma insertion at the clause boundaries "topicalized element-*wa*." While there existed 601 commas at such boundaries in the test data, only 141 commas were inserted correctly. We need to consider more effective features about "topicalized element-*wa*."

As for other cases, there existed 130 bunsetsu boundaries between parallel words where commas were not inserted. One example of such case is shown below.

- **correct data:**
  ボウルに豚の背脂、ニンニク、ショウガ、ネギのみじん切りを入れ、彩りの赤ピーマンも加えます。(Put pork backfat, garlic, ginger and shredded green onion in a bowl, and add red bell peppers for color.)

Table 7: Result of comma insertion at the clause boundaries "topicalized element-*wa*."

recall	precision	F-measure
23.46% (141/601)	59.49% (141/237)	33.65

- **our method:**
  ボウルに豚の背脂ニンニク、ショウガ、ネギのみじん切りを入れ、彩りの赤ピーマンも加えます。(Put pork backfat_ garlic, ginger and shredded green onion in a bowl, and add red bell peppers for color.)

In the correct data, a comma was inserted between the bunsetsu "背脂 (backfat)" and "ニンニク (garlic)."

If a comma should be inserted right after the bunsetsu ''背脂 (backfat)," the number of characters between commas would become too small to be judged as appropriate by the proposed method. So, the feature about the number of characters between commas may have had harmful effects there. On the other hand, a comma was inserted properly between the bunsetsu "ニンニク (garlic)" and "ショウガ (ginger)." This is because *katakana* characters appeared sequentially in addition to appearing as parallel nouns.

6.2 Unnatural Comma Insertion

When commas are inserted at obviously unnatural positions, they have a major impact on the understanding of a sentence by readers. Here, we investigated how many commas had been inserted at obviously unnatural positions by our method. For the article on January 14th (217 sentences, 2,349 bunsetsus) in the test data, we examined 47 commas inserted incorrectly. Three persons decided whether or not the inserted commas were obviously unnatural through consultations. Concretely, when all of the three persons felt that an inserted comma would make readers understand wrongly the meaning of the sentence, the comma was judged to be obviously unnatural.

Among 47 commas, 4 commas were judged obviously unnatural. This result shows that our method is capable of inserting commas at natural positions on some level.

Table 8: Comparison with human judgement

	recall	precision	F-measure
by human	78.30% (249/318)	80.58% (249/309)	79.42
our method	71.07% (226/318)	82.78% (226/273)	76.48

6.3 Comparison with Human Judgement

In our experiment, we evaluated the results of comma insertion of our method by comparing them with the correct data. However, the sufficient level to be reached by automatic comma insertion is uncertain. Here, we evaluated our method by comparing them with the results of comma insertion by another person. By using the same data as used in the subsection 6.2, we conducted an experiment on comma insertion by an annotator who was familiar with writing Japanese documents. Table 8 shows the recall, the precision and the F-measure. The second row shows the results of our method for the same data. As the F-measure of the annotator was 79.42, it turned out that comma insertion task was difficult even for humans. For F-measure, our method achieved 96.30% (76.48/79.42) of the annotator's result. Also, the precision of our method was 82.78%. Although the comma insertion task is difficult, our method was able to properly insert commas.

7 Conclusion

This paper proposed a method for inserting commas into Japanese texts. Our method appropriately inserts commas based on the machine learning method using such features as morphemes, dependencies and clause boundaries. An experiment by using the Kyoto Text Corpus (Kurohashi and Nagao, 1998) showed an F-measure of 75.90, and we confirmed the effectiveness of our method.

The analysis of the experimental results showed that our method cannot insert commas of the particular usage. As a future work, it is necessary to find more useful features for commas of this usage and improve the recall of our method. Also, we will examine "commas emphasizing the adjacent word" which were not included in our targets.

Acknowledgments

This research was partially supported by the Grant-in-Aids for Scientific Research (B) (No. 22300051) and Young Scientists (B) (No. 21700157), and by the Continuation Grants for Young Researchers of The Asahi Glass Foundation.

References

Heidi Christensen, Yoshihiko Gotoh, and Steve Renals. 2001. Punctuation annotation using statistical prosody models. In *Proceedings of ISCA Workshop on Prosody in Speech Recognition and Understanding*, pages 35–40.

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.

Yoshihiko Hayashi. 1992. A three-level revision model for improving Japanese bad-styled expressions. In *Proceeding of 14th International Conference on Computational Linguistics*, pages 665–671.

Katsuichi Honda. 1982. *Nihongo no sakubun gijutsu (Japanese writing skill)*. Asahi Shimbun Publications Inc. (In Japanese).

Takashi Inukai. 2002. *Moji hyouki tankyuhou (Method of questioning characters and notations)*. Asakura Pulishing Co., Ltd. (In Japanese).

Hideki Kashioka and Takehiko Maruyama. 2004. Segmentation of semantic unit in Japanese monologue. In *Proceedings of International Conference on Speech Language Technology and Oriental COCOSDA*, pages 87–92.

Ji-hwan Kim and P. C. Woodland. 2001. The use of prosody in a combined system for punctuation generation and speech recognition. In *Proceedings of 7th European Conference on Speech Communication and Technology*, pages 2757–2760.

Sadao Kurohashi and Makoto Nagao. 1998. Building a Japanese parsed corpus while improving the parsing system. In *Proceedings of 1st International Conference on Language Resources and Evaluation*, pages 719–724.

Zhang Le. 2008. Maximum entropy modeling toolkit for python and c++. `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html`. [Online; accessed 1-March-2008].

Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. 2006. Enriching speech recognition with automatic detection of

sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1526–1540.

Tohru Shimizu, Satoshi Nakamura, and Tatsuya Kawahara. 2008. Effect of punctuation marks for speech translation unit boundary detection. *IEICE technical report. Speech*, 108(338):127–131. (In Japanese).

Shogakukan's editiorial department. 2007. *kutoten, kigou, hugou katuyoujiten (dictionary of punctuations and symbols)*. Shogakukan. (In Japanese).

Eiji Suzuki, Shizuo Shimada, Kunio Kondo, and Hisashi Sato. 1995. Automatic recognition of optimal punctuation in Japanese documents. In *Proceedings of 50th National Convention of IPSJ*, 50(3):185–186. (In Japanese).

Michael White and Rajakrishnan Rajkumar. 2008. A more precise analysis of punctuation for broad-coverage surface realization with CCG. In *Proceedings of Workshop on Grammar Engineering Across Frameworks*, pages 17–24.