Flexible Japanese Sentence Compression
by Relaxing Unit Constraints

Jun Harashima   Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
`harashima@nlp.ist.i.kyoto-u.ac.jp, kuro@i.kyoto-u.ac.jp`

ABSTRACT

Sentence compression is important in a wide range of applications in natural language processing. Previous approaches of Japanese sentence compression can be divided into two groups. *Word-based methods* extract a subset of words from a sentence to shorten it, while *bunsetsu-based methods* extract a subset of bunsetsu (where a bunsetsu is a text unit that consists of content words and following function words). Basically, bunsetsu-based methods perform better than word-based methods. However, bunsetsu-based methods have the disadvantage that they cannot drop unimportant words from each bunsetsu because they have to follow constraints under which each bunsetsu is treated as a unit. In this paper, we propose a novel compression method to overcome this disadvantage. Our method relaxes the constraints using Lagrangian relaxation and shortens each bunsetsu if it contains unimportant words. Experimental results show that our method effectively compresses a sentence while preserving its important information and grammaticality.

TITLE AND ABSTRACT IN JAPANESE

ユニット制約の緩和による柔軟な日本語文圧縮

文圧縮は，自然言語処理の様々なアプリケーションにおいて重要である．日本語文に対する既存の圧縮手法は二種類に分けられる．単語ベースの手法は文から単語集合を選出し，圧縮文とする．一方，文節ベースの手法は文から文節集合を選出し，圧縮文とする．基本的には後者の方が良く機能する．しかし，文節ベースの手法は，文節をユニットとして扱うという制約があるため，個々の文節を圧縮できない．本稿では，この欠点を克服する新しい圧縮手法を提案する．提案手法はラグランジュ緩和を用いて上の制約を緩和し，各文節を圧縮する．実験の結果，提案手法によって原文の情報を多く保持する文法的な圧縮文を生成できることが分かった．

KEYWORDS: sentence compression, Lagrangian relaxation.

KEYWORDS IN JAPANESE: 文圧縮, ラグランジュ緩和.

1 Introduction

Sentence compression is the task of shortening a sentence while preserving its important information and grammaticality. This task is important in a wide range of applications such as automatic summarization (Jing, 2000; Lin, 2003; Zajic et al., 2007), subtitle generation (Vandeghinste and Pan, 2004), and displaying text on small screens (Corston-Oliver, 2001).

In this paper, we propose a novel compression method for a Japanese sentence. Like other languages, Japanese uses sentences composed of words. However, we can also say that a Japanese sentence is composed of *bunsetsu*. Bunsetsu is a text unit that consists of one or more content words and possibly one or more function words. For example, consider the following sentence.[1]

(1) *nihon to   kanada no   kokusai   kyoudou   kenkyuu guru-pu ga   hakken shita*
    Japan CNJ Canada GEN international collaborative research group   NOM discover did

   (An international collaborative research group between Japan and Canada made a discovery)

This sentence is composed of four bunsetsu: "*nihon to*", "*kanada no*", "*kokusai kyoudou kenkyuu guru-pu ga*", and "*hakken shita*". As seen in this example, a Japanese sentence can be viewed as a bunsetsu sequence as well as a word sequence.

This characteristic of the Japanese language has led researchers to take two compression approaches: *word-based methods* and *bunsetsu-based methods*. Word-based methods view a source sentence as a word sequence and generate a compressed sentence by selecting a subset of words from the source sentence (Hori and Furui, 2004; Hirao et al., 2009). However, the methods do not take account of bunsetsu, and it is thus difficult to generate grammatical compressions. For example, if only content words (or only function words) in a bunsetsu are selected, the grammaticality of the corresponding part in the compressed sentence would be poor.

We can avoid this problem using bunsetsu-based methods. Bunsetsu-based methods view a source sentence as a bunsetsu sequence and generate a compressed sentence by selecting a subset of bunsetsu from the source sentence (Takeuchi and Matsumoto, 2001; Oguro et al., 2002; Yamagata et al., 2006; Nomoto, 2008). Bunsetsu-based methods treat each bunsetsu as a unit. Thus, the methods do not suffer from the above problem and they can generate compressions that are more grammatically correct than those generated by word-based methods.

However, bunsetsu-based methods have a disadvantage in that they cannot shorten each bunsetsu in a source sentence. More precisely, when there is a compound noun in a bunsetsu and the noun contains unimportant words, bunsetsu-based methods cannot drop those words from the noun. Consider the above sentence (1) as an example. The third bunsetsu "*kokusai kyoudou kenkyuu guru-pu ga*" contains a compound noun "*kokusai kyoudou kenkyuu guru-pu*" (international collaborative research group). Suppose that we want to drop the word "*kokusai*" (international) from the noun and to shorten the bunsetsu to "*kyoudou kenkyuu guru-pu ga*". However, bunsetsu-based methods cannot perform such flexible word selection because they have to treat each bunsetsu as it is.

---

[1]This paper uses the abbreviations NOM (nominative), ACC (accusative), DAT (dative), ALL (allative), GEN (genitive), CMI (comitative), CNJ (conjunction), and TOP (topic marker).

In this paper, we propose a novel compression method to overcome the above disadvantage. As described above, the disadvantage results from a constraint for each bunsetsu under which the bunsetsu has to be treated as a unit (hereafter called the *unit constraint*). If we ignore unit constraints, we may be able to avoid the problem. However, if we do so, we again suffer from the problem of word-based methods (i.e., we will not generate grammatical compressions). We therefore do not ignore or adhere to unit constraints, but relax them using Lagrangian relaxation. That is, the proposed method basically follows the constraints and treats each bunsetsu as a unit. However, if a bunsetsu contains unimportant words, our method violates its unit constraint and drops those words from the bunsetsu. In this paper, we formulate this idea using integer linear programming (ILP) and report the effectiveness through experiments.

2   Word-based Method

We first describe word-based methods in detail. Although several word-based methods have been proposed (Hori and Furui, 2004; Hirao et al., 2009), the basic idea behind the methods is the same. We explain the idea in detail and discuss the advantage and disadvantage of word-based methods.

2.1   Idea

In word-based methods, a source sentence is viewed as a word sequence. Let w_i ($i = 1, \ldots, I$) denote a word in a source sentence. The basic idea underlying word-based methods is that the compression is a subset of words with the maximum importance in a source sentence. Through ILP, this idea is formulated as follows.

Sentence Compression (Word-based Formulation)

$$\text{maximize} \quad \sum_{i=1}^{I} x_i \, Score(w_i) \qquad (1)$$

$$\text{subject to} \quad \sum_{i=1}^{I} x_i \, Length(w_i) \leq L \qquad (2)$$

$$x_i = 0 \text{ or } 1 \quad (i = 1, \ldots, I) \qquad (3)$$

where x_i denotes a decision variable of w_i that is 1 if w_i is contained in a compressed sentence, and otherwise 0. $Score(w_i)$ represents the importance of w_i and $Length(w_i)$ represents the length of w_i. L is a predefined maximum length of a compressed sentence. According to Eq. (1), the optimal subset of words in a source sentence is selected as a compressed sentence. In addition, according to Eq. (2), the length of the compressed sentence shall be not more than L.

2.2   Advantage and Disadvantage

The advantage of word-based methods is that the methods can more freely select important words in a source sentence than bunsetsu-based methods. This is because word-based methods do not take account of bunsetsu in a source sentence and are not limited to unit constraints.

However, due to the freeness, word-based methods have the disadvantage that they tend to generate ungrammatical compressions. As described in Section 1, if only content words (or
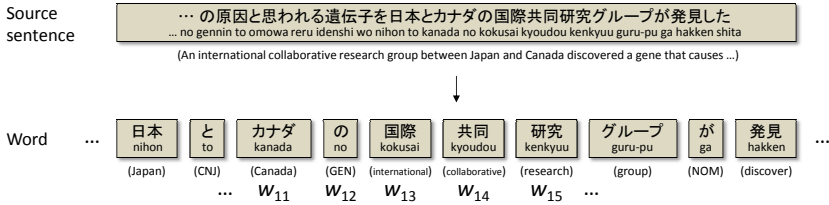
Figure 1: Word-based method. A source sentence is segmented into words and a subset of words in the sentence is selected as the compressed sentence.

only function words) in a bunsetsu are selected, the grammaticality of the corresponding part in the compressed sentence would be poor. Consider the sentence in Figure 1. There is, for example, the bunsetsu "*kanada no*" in the sentence (although word-based methods do not take account of it). If we select only the word "*kanada*" (Canada) from the bunsetsu and do not select the word "*no*" (GEN), the corresponding part of the compressed sentence would not make sense. As just described, it is difficult to generate grammatical compressions using word-based methods.

3   Bunsetsu-based Method

Like word-based methods, most previous bunsetsu-based methods (Oguro et al., 2002; Yamagata et al., 2006; Nomoto, 2008) are based on the same idea. In this section, we explain the idea and discuss the advantages and disadvantage of bunsetsu-based methods.

3.1   Idea

Word-based methods view a source sentence as a word sequence, while bunsetsu-based methods view the sentence as a bunsetsu sequence. Let b_j $(j = 1,\ldots,J)$ denote a bunsetsu in a source sentence. The basic idea underlying bunsetsu-based methods is that the compression is a subset of bunsetsu with the maximum importance in a source sentence. Through ILP, this idea is formulated as follows.

Sentence Compression (Bunsetsu-based Formulation)

$$\text{maximize} \quad \sum_{j=1}^{J} y_j \, Score(b_j) \tag{4}$$

$$\text{subject to} \quad \sum_{j=1}^{J} y_j \, Length(b_j) \leq L \tag{5}$$

$$y_j = 0 \text{ or } 1 \quad (j = 1,\ldots,J) \tag{6}$$

where y_j denotes a decision variable of b_j that is 1 if b_j is contained in a compressed sentence, and otherwise 0. $Score(b_j)$ represents the importance of b_j and $Length(b_j)$ represents the length of b_j. According to Eq. (4), the optimal subset of bunsetsu in a source sentence is selected as a compressed sentence. In addition, according to Eq. (5), the length of the compressed sentence shall be not more than L.

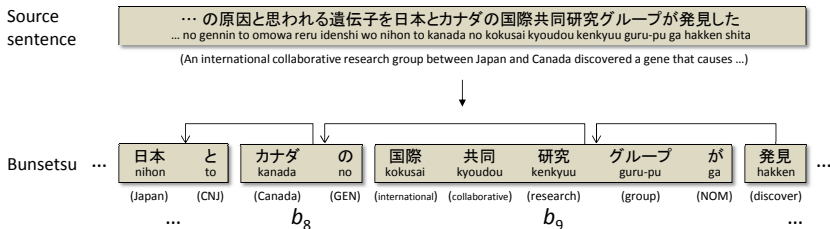| 日本 nihon	と to	カナダ kanada	の no	国際 kokusai	共同 kyoudou	研究 kenkyuu	グループ guru-pu	が ga	発見 hakken
(Japan)	(CNJ)	(Canada)	(GEN)	(international)	(collaborative)	(research)	(group)	(NOM)	(discover)

Figure 2: Bunsetsu-based method. A source sentence is segmented into bunsetsu and a subset of bunsetsu in the sentence is selected as the compressed sentence. Bunsetsu-based methods have the advantage in that they are able to use dependency information.

3.2 Advantages and Disadvantage

One advantage of bunsetsu-based methods is that the methods can generate compressions that are more grammatical than compressions generated by word-based methods. Bunsetsu-based methods select each bunsetsu in a source sentence just as it is. Therefore, the methods do not suffer from the problem of word-based methods (see also Section 2.2).

Bunsetsu-based methods have the another advantage in that they are able to use dependency information in a source sentence. In Japanese, a dependency relation is generally defined between not a pair of words but a pair of bunsetsu. Consider the source sentence in Figure 2. In the sentence, there is the example that bunsetsu b_8 depends on bunsetsu b_9. Bunsetsu-based methods can use this information by adding the following simple constraint to the above formulation.

$$\text{subject to} \quad y_8 \leq y_9 \qquad (7)$$

This constraint ensures that if b_8 is contained in a compressed sentence, b_9 is also contained in the sentence. In this way, bunsetsu-based methods can easily use dependency information in a source sentence. On the other hand, there is a word-based method that defines dependency relations between words in a source sentence and uses the information (Hori and Furui, 2003). However, as described above, a dependency relation is generally defined between a pair of bunsetsu. In the method, bunsetsu dependencies in a source sentence and complex rules are necessary to define the word dependencies, and it is not easy to use the information.

On the other hand, as described in Section 1, the disadvantage of bunsetsu-based methods is that they cannot shorten each bunsetsu in a source sentence. More precisely, when there is a compound noun in a bunsetsu and the noun contains unimportant words, bunsetsu-based methods can not drop those words from the noun. Consider again the sentence in Figure 2. Bunsetsu b_9 contains a compound noun "*kokusai kyoudou kenkyuu guru-pu*" (international collaborative research group). Suppose that we want to drop the word "*kokusai*" (international) from the noun and to shorten the bunsetsu to "*kyoudou kenkyuu guru-pu ga*". However, bunsetsu-based methods cannot perform such flexible word selection because they are limited to unit constraints.

4 Proposed Method

Bunsetsu-based methods basically perform better than word-based methods, especially in terms of the grammaticality of a compressed sentence. However, bunsetsu-based methods have the disadvantage that they cannot shorten each bunsetsu in a source sentence. In this section, we describe a novel compression method that overcomes this disadvantage.

4.1 Idea

The point of our method is to relax unit constraints responsible for the disadvantage. Under the constraints, we have to treat each bunsetsu as a unit. If we ignore the constraints, we may be able to avoid the problem. However, if we do so, we again suffer from the problem of word-based methods (i.e., we will not generate grammatical compressions). Therefore, we select a set of bunsetsu, each containing unimportant words, and relax their unit constraints. Note that each bunsetsu that contains a compound noun (e.g., b_9 in Figure 3) is selected as a bunsetsu that may contain unimportant words. Conversely, we do not shorten each bunsetsu that does not contain a compound noun (e.g., b_8 in Figure 3) because such a bunsetsu has only one content word and does not need to be shortened.

First, let us rewrite the bunsetsu-based formulation in Section 3.1. Using w_i and x_i instead of b_j and y_j, the formulation can be rewritten as follows.

Sentence Compression (Bunsetsu-based Formulation 2)

$$\text{maximize} \quad \sum_{i=1}^{I} x_i \, Score(w_i) \tag{8}$$

$$\text{subject to} \quad \sum_{i=1}^{I} x_i \, Length(w_i) \leq L \tag{9}$$

$$x_i = 0 \text{ or } 1 \quad (i = 1, \ldots, I) \tag{10}$$

$$x_{First(b_j)} = x_{First(b_j)+1} = \cdots = x_{Last(b_j)}$$
$$(w_{First(b_j)}, \ldots, w_{Last(b_j)} \in b_j, \, j = 1, \ldots, J) \tag{11}$$

where $First(b_j)$ represents a function that returns the index of the first word in b_j, while $Last(b_j)$ returns that of the last word in b_j (e.g., in Figure 3, $First(b_9) = 13$ and $Last(b_9) = 17$). In addition, we set $Score(b_j) = \sum_{w_i \in b_j} Score(w_i)$ and $Length(b_j) = \sum_{w_i \in b_j} Length(w_i)$.

The notable aspect of the above formulation is Eq. (11), which is the set of unit constraints. Equation (11) ensures that if we select one word from a bunsetsu in a source sentence, we also select the other words from the bunsetsu. Likewise, if we do not select one word from a bunsetsu, we must also not select the other words from the bunsetsu.

The proposed method does not ignore or adhere to unit constraints but relaxes them. To do this, we use Lagrangian relaxation, which is a classical technique for combinatorial optimization. The technique moves problematic constraints into the objective function and penalizes the function if those constraints are not satisfied. Using the technique, we remove unit constraints for each bunsetsu that contains a compound noun.

Let B_{CN} denotes a subset of bunsetsu, each containing a compound noun. In addition, let us decompose a unit constraint for b_j in B_{CN} (i.e., $x_{First(b_j)} = x_{First(b_j)+1} = \cdots = x_{Last(b_j)}$) into
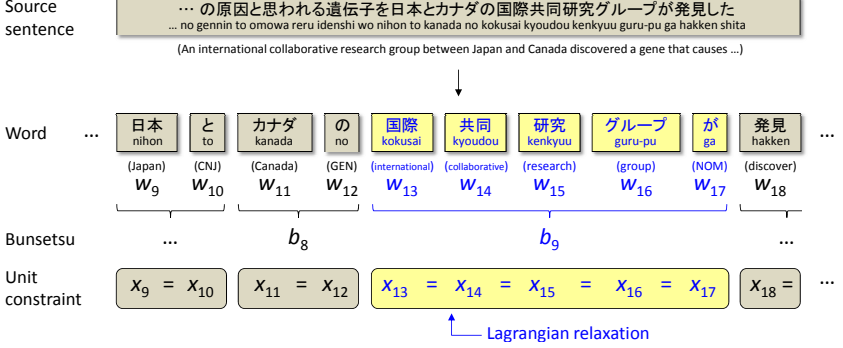
Figure 3: Proposed method. Bunsetsu b_9 contains a compound noun "*kokusai kyoudou kenkyuu guru-pu*" (international collaborative research group). The proposed method relaxes the unit constraint for the bunsetsu using Lagrangian relaxation.

a set of constraints: $x_{First(b_j)} = x_{Last(b_j)}$, $x_{First(b_j)+1} = x_{Last(b_j)}$, ..., and $x_{Last(b_j)-1} = x_{Last(b_j)}$. Our new formulation is given below.

Sentence Compression (Proposed Formulation)

$$\text{maximize} \quad \sum_{i=1}^{I} x_i\, Score(w_i) + \sum_{b_j \in B_{CN}} \sum_{w_i \in b_j \backslash w_{Last(b_j)}} \mu_{i,Last(b_j)}(x_i - x_{Last(b_j)}) \quad (12)$$

$$\text{subject to} \quad \sum_{i=1}^{I} x_i\, Length(w_i) \leq L \quad (13)$$

$$x_i = 0 \text{ or } 1 \quad (i = 1, \ldots, I) \quad (14)$$

$$x_{First(b_j)} \leq x_{First(b_j)+1} \leq \cdots \leq x_{CLast(b_j)} = \cdots = x_{Last(b_j)}$$
$$(w_{First(b_j)}, \ldots, w_{Last(b_j)} \in b_j,\ j = 1, \ldots, J) \quad (15)$$

where $\mu_{i,Last(b_j)}$ is a Lagrangian multiplier provided for a constraint $x_i = x_{Last(b_j)}$. $CLast(b_j)$ in Eq. (15) returns the index of the last content word in b_j (e.g., in Figure 3, $CLast(b_9) = 16$).

A notable aspect is the second term in Eq. (12). For a bunsetsu b_j in B_{CN} (e.g., b_9 in Figure 3), the term penalizes the objective function if a decision variable of a word in b_j (denoted as x_i) is not equal to that of the last word in b_j (denoted as $x_{Last(b_j)}$). For example, if $x_i = 0$ and $x_{Last(b_j)} = 1$, the term penalizes the objective function. Thus, the proposed method basically treats words in b_j as a unit similarly to bunsetsu-based methods. However, now there is no constraint under which $x_{First(b_j)} = x_{First(b_j)+1} = \cdots = x_{Last(b_j)}$. In other words, although we have to consider the penalty, we can set a different value for each decision variable. Suppose that w_{13} in b_9 in Figure 3 have little importance, while the rest words in b_9 have great importance. Unlike bunsetsu-based methods, the proposed method can set $x_{13} = 0$ and the rest decision variables to 1.

Equation (15) represents a constraint that sets the order of preference of the selection of words in a bunsetsu. As described above, for each bunsetsu in B_{CN}, each decision variable

Algorithm 1 Solve the proposed formulation.

1: **for** $b_j \in B_{CN}$ **do**
2:     **for** $w_i \in b_j \backslash w_{Last(b_j)}$ **do**
3:         $\mu_{i,Last(b_j)}^{(0)} \leftarrow 0$
4:     **end for**
5: **end for**
6: **for** $t \in \{1, \ldots, T\}$ **do**
7:     $\boldsymbol{x}^{(t)} \leftarrow \arg \max_{\boldsymbol{x}}$ Eq. (12) (Note that $\mu_{i,Last(b_j)}^{(t-1)}$ is used as $\mu_{i,Last(b_j)}$ in Eq. (12)
8:     **for** $b_j \in B_{CN}$ **do**
9:         **for** $w_i \in b_j \backslash w_{Last(b_j)}$ **do**
10:             $\mu_{i,Last(b_j)}^{(t)} \leftarrow \mu_{i,Last(b_j)}^{(t-1)} - \alpha^{(t)}(x_i^{(t)} - x_{Last(b_j)}^{(t)})$
11:         **end for**
12:     **end for**
13: **end for**
14: return $\boldsymbol{x}^{(T)}$

---

can take a different value from other variables. However, care must be taken in setting each variable. More precisely, latter words in a bunsetsu generally should not be dropped before the earlier words are dropped. One reason for this is that function words are located in the latter part of a bunsetsu. Another is that former words within a Japanese compound noun basically modify the latter words (i.e., the latter words are syntactically more important than the former words). We thus add a constraint to our formulation under which we prioritize latter words in a bunsetsu.

We set $x_{CLast(b_j)} = \cdots = x_{Last(b_j)}$ in the latter part of Eq. (15). The purpose is to treat function words in a bunsetsu in B_{CN} as a unit and select at least one content word from the bunsetsu. Furthermore, using this equation, we can retain a unit constraint for each bunsetsu that does not contain a compound noun (i.e., for such bunsetsu, Eq. (15) is the same as Eq. (11)).

Of course, there are exceptions to Eq. (15), especially proper nouns. For example, if we drop the first word from the compound noun "*murayama tomiichi syusyou*", an unlikely noun "*tomiichi syusyou*" would be generated ("*murayama*", "*tomiichi*", and "*syusyou*" meaning Murayama, Tomiichi, and prime minister, respectively). In this case, we need to recognize the family name ("*murayama*"), the last name ("*tomiichi*"), and the title ("*syusyou*") and drop the words in the following order: last name, family name, and title. In our experiments described in the following section, we handle this exception for person names. However, we do not handle exceptions about other proper nouns such as organization names. We leave this for our future work.

Finally, we present an algorithm to solve our formulation in Algorithm 1. In the algorithm, T denotes the number of iterations and $\alpha^{(t)}$ denotes a parameter that determines a step size to update each Lagrangian multiplier (see (Korte and Vygen, 2008) for detail). Using this algorithm, each multiplier is updated and a subset of words in a source sentence is selected so that a compression produced by our method is as similar to that produced by bunsetsu-based methods as possible. However, as described in the previous paragraphs, if bunsetsu contain unimportant words, our method prioritizes to violate their unit constraints and drop the unimportant words.

4.2 Advantages

Compared with word-based and bunsetsu-based methods, the proposed method has at least three advantages. First, our method can generate compressions that are more grammatical than compressions generated with word-based methods. This is because our method is loosely based on bunsetsu-based methods and basically treats each bunsetsu in a source sentence as it is.

Second, unlike word-based methods, our method can easily use dependency information in a source sentence. This is again because our method is loosely based on bunsetsu-based methods. For example, bunsetsu b_8 in Figure 3 depends on bunsetsu b_9 (see also Figure 2). We can use this information employing the following constraint.

$$\text{subject to} \quad x_{12} \leq x_{17} \tag{16}$$

That is, we introduce a constraint between the last words of the bunsetsu. In this way, when b_8 is contained in a compressed sentence, we can ensure that b_9 is also contained in the sentence regardless of whether the bunsetsu are shortened.

Third, unlike bunsetsu-based methods, our method can shorten a bunsetsu in a source sentence. Since bunsetsu-based methods are limited by unit constraints, they have to treat each bunsetsu as a unit. Thus, even if there are unimportant words in a bunsetsu, the methods do not drop those words from the bunsetsu. On the other hand, our method relaxes unit constraints using Lagrangian relaxation. Thus, our method has the ability to drop unimportant words from a bunsetsu, even though it is loosely based on bunsetsu-based methods.

5 Experiments

In this section, we report two experiments conducted to evaluate the proposed method.

5.1 Test Set

There is no standard test set for Japanese sentence compression. We therefore constructed a test set to evaluate the proposed method. The construction process is as follows.

First, we extracted 240 sentences from Kyoto University Text Corpus (Kurohashi and Nagao, 1998),[2] a parsed corpus of Mainichi Shimbun 1995. More precisely, we extracted sentences that satisfied all of the following three conditions. (1) The sentence was a lead sentence (the first sentence of an article). Lead sentences are often used in experiments for sentence compression because they can be compressed without consideration of their context. (2) The length of the sentence was not too short and not too long. We employed the number of characters in a sentence as the sentence length and extracted sentences not shorter than 51 characters and not longer than 100 characters. (3) The sentence did not contain parentheses. This condition was considered because we found that even human subjects often could not compress content in parentheses (typically speech). From the 409 extracted sentences that satisfied these three conditions, we randomly selected 240 sentences for our experiments.

For each of the 240 sentences, two subjects produced compressed versions. The compression ratio was set to 0.7. For example, if the length of a sentence was 100 characters, each subject was asked to produce a compressed sentence whose length was not longer than 70 characters.

---

[2]http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?Kyoto%20University%20Text%20Corpus

Avg. # of characters per source sentence	75.1
Avg. # of words per source sentence	44.6
Avg. # of bunsetsu per source sentence	14.5
Avg. # of bunsetsu with a compound noun per source sentence	5.3

Table 1: Statistics of our test set.

Finally, from the 240 groups of a source sentence and its two compressed versions, we randomly selected 160 groups as our test set. We used the remaining 80 groups as a development set to tune the proposed method. The statistics of our test set are given in Table 1.

5.2 Methods

In our experiments, we compared the outputs of the following methods.

WORD (RANDOM) Word-based method. This method randomly selected a subset of words from a source sentence as the compressed sentence.

WORD Word-based method described in Section 2.1. The optimal subset of words in a source sentence was selected using ILP.

BNST (RANDOM) Bunsetsu-based method that randomly selected a subset of bunsetsu from a source sentence as the compressed sentence.

BNST Bunsetsu-based method described in Section 3.1. The optimal subset of bunsetsu in a source sentence was selected using ILP.

BNST w/ DPND Bunsetsu-based method. We added dependency constraints to BNST.

PROP Proposed method described in Section 4.1. Using Lagrangian relaxation, unimportant words were dropped from a compound noun in a bunsetsu.

PROP w/ DPND Proposed method. We added dependency constraints to PROP.

HUMAN Human compression. For each source sentence in our test set, one of the two compressed sentences produced by subjects were randomly selected.

For WORD, BNST, BNST w/ DPND, PROP, and PROP w/ DPND, we used lp_solve (a mixed ILP solver[3]). In addition, we set $Score(w_i)$ and $Score(b_j)$ as follows. First, from articles in the newspaper Mainichi Shimbun from 1991 to 2002, we extracted pairs of a lead sentence and a title that could be viewed as a source sentence and its pseudo compression. Note that articles in 1995 were excluded because they overlapped with our test set. Moreover, we viewed a lead sentence and a title as a source sentence and its pseudo compression if the lead sentence contained more than 80% of content words in the title. We then calculated the rate of occurrence of a word in the titles to that in the lead sentences. For example, if a word appeared 50 times in the titles and 100 times in the lead sentences, the rate of occurrence of the word was 0.5. We used this rate as $Score(w_i)$ and set $Score(b_j) = \sum_{w_i \in b_j} Score(w_i)$.

Furthermore, we set $\alpha^{(t)} = \alpha^{(0)}/t$ for PROP and PROP w/ DPND. Note that $\alpha^{(0)}$ was set to 0.02 according to our development set. T was set to 100.

---

[3]lpsolve.sourceforge.net/5.5

	ROUGE 1	ROUGE 2
WORD (RANDOM)	0.690	0.409
WORD	0.736	0.540
BNST (RANDOM)	0.679	0.529
BNST	0.745	0.615
BNST w/ DPND	0.772†	0.653†
PROP	0.751	0.616
PROP w/ DPND	0.796†‡	0.671†‡

Table 2: Information content in a compressed sentence. † and ‡ mark statistically significant improvement over BNST and BNST w/ DPND with $p < 0.01$, respectively.

5.3   Information Content of a Compressed Sentence

In the first experiment, we examined how well our method performed in preserving important information in a source sentence. Using the methods described in the previous section, we compressed each source sentence in our test set. The compression ratio was set to 0.7. We then computed Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores (Lin, 2004) for each of the methods. That is, we measured the n-gram overlap between the outputs of each method and those of human subjects. In the calculation of ROUGE, stopwords were not removed. In addition, each word in a compression was normalized using the Japanese morphological analyzer JUMAN (Kurohashi et al., 1994).

Table 2 gives the results. We can see that PROP significantly outperformed WORD, especially in terms of ROUGE 2. The reason for this is that unlike WORD, PROP basically selected words in each bunsetsu in a source sentence as a unit.

When dependency constraints were not considered, PROP performed better than BNST. However, the performance differences between them were small and only the difference in ROUGE 1 was statistically significant (Wilcoxon signed-rank test, $p < 0.05$). On the other hand, when dependency constraints were considered, PROP significantly outperformed BNST. This time, the differences in ROUGE 1 and ROUGE 2 were both statistically significant ($p < 0.01$).

We found that the differences between PROP w/ DPND and BNST w/ DPND were due to the number of available bunsetsu that each method could select. More precisely, PROP w/ DPND selected 4.6% more (shortened) bunsetsu in a source sentence than BNST w/ DPND. Suppose that bunsetsu b_j is located at a deep node in a dependency tree of a source sentence (i.e., b_j depends on $b_{j'}$ and $b_{j'}$ depends on $b_{j''}$ and ... depends on b_J). To select b_j, both methods first have to select from $b_{j'}$ to b_J owing to the dependency constraints. However, since there is a length constraint, it is usually difficult to select b_j at such a deep node even if b_j contains important words. However, PROP w/ DPND has more chance of selecting b_j than BNST w/ DPND. This is because PROP w/ DPND can make room to select b_j by dropping unimportant words from other bunsetsu. In this way, PROP w/ DPND selected more bunsetsu that contained important words and achieved higher performance than BNST w/ DPND.

In contrast, when dependency constraints were not considered, the number of (shortened) bunsetsu that PROP selected was not so different from the number of that BNST selected (the difference was 2.4%). This is because PROP ignored dependency constraints and almost greedily selected bunsetsu that were composed of many important words. In other words, PROP had less opportunity to drop unimportant words. As a result, the differences between the performances of the two methods were not so large.

	Grammaticality
WORD	1.50
BNST	2.23
BNST w/ DPND	4.15
PROP	2.18
PROP w/ DPND	4.14
HUMAN	4.85

Table 3: Grammaticality of a compressed sentence. The score ranges from 1 to 5: 1 (very poor), 2 (poor), 3 (average), 4 (good), and 5 (very good).

5.4 Grammaticality of a Compressed Sentence

In the second experiment, we examined how well our method performed in producing grammatical compressions. We randomly selected 50 source sentences from our test set and obtained the outputs of six methods for those sentences. Note that the six methods were WORD, BNST, BNST w/ DPND, PROP, PROP w/ DPND, and HUMAN. Then, for each of the 50 source sentences, we presented the six outputs to five subjects and asked them to rate the outputs in terms of grammaticality. The subjects were all native Japanese speakers and did not include the two subjects who constructed our test set. They were told that all outputs were automatically generated. For each of the source sentences, the order of the outputs was randomized.

Table 3 presents the results. From the table, we can confirm that PROP produces compressions that are more grammatically correct than the compressions produced by WORD. This is because PROP was loosely based on bunsetsu-based methods and basically selected each bunsetsu as a unit.

For the same reason, PROP achieved comparable performance with BNST. Note that the performance of PROP was slightly worse than that of BNST. This is because words that should not be dropped from a bunsetsu were dropped by PROP. For example, PROP shortened the bunsetsu "*unyu syou wa*" to "*shou wa*" ("*unyu*", "*syou*", and "*wa*" mean transport, ministry, and TOP, respectively). In Kyoto University Text Corpus, which we used in our experiments, "*shou*" (ministry) was tagged as a noun. However, unlike usual nouns, the word cannot be located at the beginning of a bunsetsu. This is because in Japanese, the word has a strong suffix nature. Thus, we must not drop "*unyu*" (transport) from the bunsetsu and we should treat "*unyu*" (transport) and "*syou*" (ministry) as a unit. We found that most errors arising when employing our method related to such words, which could be viewed as both a noun and suffix (e.g., "*kai*" (meeting), "*jin*" (-ese)). As a future work, we need to properly handle these words.

PROP achieved good performance when we added dependency constraints to the method. The score of PROP w/ DPND was 4.14. This result indicates that the grammaticality of the compressions produced by PROP w/ DPND was generally good. The reason why the performance of PROP w/ DPND was slightly worse than that of BNST w/ DPND is the same as the reason described in the previous paragraph. As seen in the first experiment, dependency constraints were also effective in preserving important information. Thus, we can say that there is no reason for not using the constraints in sentence compression.

To verify that the grading was consistent, we computed correlation coefficients between every pair of our five subjects (i.e., between subjects 1 and 2, 1 and 3, ..., and 4 and 5). Consequently, we found that the average of the coefficients was 0.45 and those coefficients were all statistically significant (t-test, $p < 0.01$). These suggest that the grading was consistent.

```
Source sentence
中米ホンジュラスの/軍司令官は/１５日、/４０年間に/わたり/管轄してきた/警察部隊を/文民の/
指揮下に/置くと/発表した。[52]
(On the 15th, an army commander of Honduras, located in Central America, announced that the
police divisions that the army had controlled for 40 years would be placed under civilian control.)

WORD
中米ホンジュラスの軍司令官は日、４０年間に管轄警察部隊を文民の指揮下にと [36]
Ungrammatical sentence

BNST w/ DPND
中米ホンジュラスの/軍司令官は/警察部隊を/文民の/指揮下に/置くと/発表した。[34]
(An army commander of Honduras, located in Central America, announced that the police divisions
would be placed under civilian control.)

PROP w/ DPND
ホンジュラスの/軍司令官は/１５日、/警察部隊を/文民の/指揮下に/置くと/発表した。[36]
(On the 15th, an army commander of Honduras announced that the police divisions would be placed
under civilian control.)

HUMAN
ホンジュラスの/軍司令官は/１５日、/警察部隊を/文民の/指揮下に/置くと/発表した。[36]
(On the 15th, an army commander of Honduras announced that the police divisions would be placed
under civilian control.)
```

Figure 4: Example of the outputs. For explanation purposes, we inserted slashes between bunsetsu in each of the outputs except for WORD. The values in square bracket denote sentence lengths.

5.5 Example of the Outputs

Figure 4 gives an example of the output of our method (PROP w/ DPND). For comparison, we also show the outputs of WORD, BNST w/ DPND, and HUMAN. In this example, the source sentence comprises 52 characters. Thus, each method was required to produce a compressed sentence whose length was not more than 36 characters ($52 \times 0.7 = 36.4$).

In this example, WORD produced a completely ungrammatical compression. In contrast, BNST w/ DPND produced a good compression, which was grammatical and preserved much information in the source sentence. However, in the output of HUMAN, the first bunsetsu "中米ホンジュラスの" in the source sentence was shortened to "ホンジュラスの" ("中米", "ホンジュラス", and "の" mean Central America, Honduras, and GEN, respectively). BNST w/ DPND could not perform such an operation because the method has to treat each bunsetsu as it is.

On the other hand, PROP w/ DPND could shorten the bunsetsu by dropping word "中米", which had less importance, from the bunsetsu ($Score$("中米") was 0.119). Additionally, using the room that was made by dropping the word, the method could add another bunsetsu "１５日、" to the compression ("１５", "日", and "、" mean 15, day, and comma, respectively). Although there are other bunsetsu that contain compound nouns (e.g., "軍司令官は"), our method did not drop any words from those bunsetsu and selected them as a unit similarly to BNST w/ DPND ("軍", "司令" + "官", and "は" mean army, commander, and TOP, respectively). This is because the words in the bunsetsu had great importance (e.g., $Score$("軍") was 0.378). In this way, PROP w/ DPND could produce the same compression as HUMAN.

6 Related Work

Sentence compression has been widely studied since the early 2000s. For the English language, Jing used multiple knowledge resources to decide which phrases in a source sentence to remove (Jing, 2000). Knight and Marcu modeled a generative process of a source sentence based on a noisy-channel framework and generated a compressed sentence using the model (Knight and Marcu, 2002). Turner and Charniak presented semi-supervised and unsupervised variants of the Knight and Marcu's model (Turner and Charniak, 2005). McDonald employed a discriminative model to learn which words in a source sentence should be dropped (McDonald, 2006). Clarke and Lapata recasted previous methods as ILP and extended those with various constraints (Clarke and Lapata, 2008). Our work differs from these efforts in that we focus on Japanese sentence compression.

For the Japanese language, previous compression methods can be divided into two groups: word-based methods and bunsetsu-based methods. Hori and Furui proposed a word-based method to summarize speech (Hori and Furui, 2004). They extracted a set of important words from an automatically transcribed sentence. Hirao et al. also proposed a word-based method (Hirao et al., 2009). They extended Hori and Furui's method using novel features. Unlike these methods, our method is loosely based on bunsetsu-based methods, and thus easily generates grammatical compressions.

Previous bunsetsu-based methods are given below. Takeuchi and Matsumoto used a support vector machine to acquire rules for dropping unimportant bunsetsu in a source sentence (Takeuchi and Matsumoto, 2001). Oguro et al. and Yamagata et al. defined varying importance of bunsetsu and dependency and extracted the optimal subset of bunsetsu from a source sentence (Oguro et al., 2002; Yamagata et al., 2006). Nomoto generated candidates for a compression by removing bunsetsu from a source sentence and selected the best candidate using a conditional random field (Nomoto, 2008). Our work differs from these efforts in that our method has the ability to drop unimportant words from a bunsetsu.

In our method, we used Lagrangian relaxation to relax unit constraints. Lagrangian relaxation is a well known technique for combinatorial optimization and it has recently been successfully applied to various natural language processing tasks (Koo et al., 2010; M.Rush et al., 2010; Chang and Collins, 2011; M.Rush and Collins, 2011). However, to the best of our knowledge, this is the first work to use the technique for sentence compression.

Conclusions and Future Work

We presented a novel compression method for a Japanese sentence. The proposed method was loosely based on bunsetsu-based methods. Thus, unlike word-based methods, it could easily produce grammatical compressions. Additionally, using Lagrangian relaxation, the proposed method relaxed constraints that troubled bunsetsu-based methods. In this way, unlike bunsetsu-based methods, our method could shorten each bunsetsu if it contained unimportant words. Experimental results showed that the proposed method could preserve more information in a source sentence than word-based and bunsetsu-based methods. Furthermore, we confirmed that our method could produce grammatical compressions similarly to bunsetsu-based methods.

In future work, as described in Section 4.1, we plan to explore a technique to handle proper nouns such as organization names. Additionally, as described in Section 5.4, we need to develop a method to handle words that can be viewed as both a noun and suffix.

References

Chang, Y.-W. and Collins, M. (2011). Exact Decoding of Phrase-Based Translation Models through Lagrangian Relaxation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 26–37.

Clarke, J. and Lapata, M. (2008). Global Inference for Sentence Compression: An Integer Linear Programming Approach. *Journal of Artificial Intelligence Research*, 31:399–429.

Corston-Oliver, S. (2001). Text Compaction for Display on Very Small Screens. In *Proceedings of the Workshop on Automatic Summarization at the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (WAS 2001)*, pages 89–98.

Hirao, T., Suzuki, J., and Isozaki, H. (2009). A Syntax-Free Approach to Japanese Sentence Compression. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009)*, pages 826–833.

Hori, C. and Furui, S. (2003). A New Approach to Automatic Speech Summarization. *IEEE Transactions on Multimedia*, 5(3):368–378.

Hori, C. and Furui, S. (2004). Speech Summarization: An Approach through Word Extraction and a Method for Evaluation. *IEIEC Transactions on Information and Systems*, E87-D(1):15–25.

Jing, H. (2000). Sentence Reduction for Automatic Text Summarization. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP 2000)*, pages 310–315.

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

Koo, T., M.Rush, A., Collins, M., Jaakkola, T., and Sontag, D. (2010). Dual Decomposition for Parsing with Non-Projective Head Automata. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1288–1298.

Korte, B. and Vygen, J. (2008). *Combinatorial Optimization: Theory and Algorithms*. Springer Verlag.

Kurohashi, S. and Nagao, M. (1998). Building a Japanese Parsed Corpus while Improving the Parsing System. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC 1998)*, pages 719–724.

Kurohashi, S., Nakamura, T., Matsumoto, Y., and Nagao, M. (1994). Improvements of Japanese Morphological Analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language*, pages 22–28.

Lin, C.-Y. (2003). Improving Summarization Performance by Sentence Compression — A Pilot Study. In *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages (IRAL 2003)*, pages 1–8.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, pages 74–81.

McDonald, R. (2006). Discriminative Sentence Compression with Soft Syntactic Evidence. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2005)*, pages 297–304.

M.Rush, A. and Collins, M. (2011). Exact Decoding of Syntactic Translation Models through Lagrangian Relaxation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (ACL-HLT 2011)*, pages 72–82.

M.Rush, A., Sontag, D., Collins, M., and Jaakkola, T. (2010). On Dual Decomposition and Linear Programming Relaxations for Natural Language Processing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1–11.

Nomoto, T. (2008). A Generic Sentence Trimmer with CRFs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (ACL-HLT 2008)*, pages 299–307.

Oguro, R., Sekiya, H., Morooka, Y., Takagi, K., and Ozeki, K. (2002). Evaluation of a Japanese Sentence Compression Method Based on Phrase Significance and Inter-Phrase Dependency. In *Proceedings of the 5th International Conference on Text, Speech and Dialogue (TSD 2002)*, pages 27–32.

Takeuchi, K. and Matsumoto, Y. (2001). Acquisition of Sentence Reduction Rules for Improving Quality of Text Summaries. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, pages 447–452.

Turner, J. and Charniak, E. (2005). Supervised and Unsupervised Learning for Sentence Compression. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 290–297.

Vandeghinste, V. and Pan, Y. (2004). Sentence Compression for Automated Subtitling: A Hybrid Approach. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, pages 89–95.

Yamagata, K., Fukutomi, S., Takagi, K., and Ozeki, K. (2006). Sentence Compression Using Statistical Information About Dependency Path Length. In *Proceedings of the 9th International Conference on Text, Speech and Dialogue (TSD 2006)*, pages 127–134.

Zajic, D., J.Dorr, B., Lin, J., and Schwartz, R. (2007). Multi-Candidate Reduction: Sentence Compression as a Tool for Document Summarization Tasks. *Information Processing and Management Special Issue on Summarization*, 43(6):1549–1570.