Evaluation of a Japanese Sentence Compression Method Based on Phrase Significance and Inter-Phrase Dependency

Rei Oguro[1], Hiromi Sekiya[1], Yuhei Morooka[1],
Kazuyuki Takagi[1], Kazuhiko Ozeki[1]

The University of Electro-Communications, Chofu, Tokyo 182-8585, Japan
{rei, sekiya, m_yuu, takagi, ozeki}@ice.uec.ac.jp

Abstract. Sentence compression is a method of text summarisation, where each sentence in a text is shortened in such a way as to retain the original information and grammatical correctness as much as possible. In a previous paper, we formulated the problem of sentence compression as an optimisation problem of extracting a subsequence of phrases from the original sentence that maximises the sum of topical importance and grammatical correctness. Based on this formulation an efficient sentence compression algorithm was derived. This paper reports a result of subjective evaluation for the quality of sentences compressed by using the algorithm.

1 Introduction

Text summarisation is an important area in natural language processing, rapidly growing in recent years [1, 2]. To generate an ideal summary, it will be necessary to understand the whole text, and then reconstruct it in a shorter form. Because of technical difficulty to implement this idea, most text summarisation methods reported so far are instead based on the idea of extracting important parts from the original text. Extraction-based text summarisation methods are classified into the following two broad classes depending on the extraction unit chosen. Hybrid methods will also be possible.

1) Extraction of significant sentences from a text to make a shorter text.
2) Extraction of significant words or phrases from a sentence to make a shorter sentence.

This paper is concerned with the latter method, which is referred to as *sentence compression*, or *sentence compaction* [3, 4, 5, 6].

In sentence compression, it is important to retain not only original information but also grammatical correctness as much as possible. In a previous paper [4], we formulated the problem of Japanese sentence compression as a problem of extracting a sequence of phrases from the original sentence that maximises the sum of topical importance and grammatical correctness. This problem can be solved efficiently based on DP-based algorithm. In this formulation, the topical importance is defined on the basis of significance of each phrase, and the grammatical correctness on the basis of dependency strength between phrases. Therefore, in order to get the algorithm to actually work, it is necessary to give phrase significance and dependency strength as *system parameters*. Also, optimum weights for the topical importance and the grammatical correctness need to be given. In the following, methods of determining those system parameters are described. Then, a subjective evaluation result for the performance of the method, together with that for human performance, is presented.

2 Overview of Sentence Compression Method

A Japanese sentence is a sequence of phrases, where a phrase is a syntactic unit called *bunsetsu* in Japanese, consisting of at least one content word followed by (possibly zero) function words such as particles and auxiliary verbs. From a dependency grammatical point of view, the syntactic structure of a Japanese sentence is determined by specifying which phrase modifies which phrase. In other words, the syntactic structure of a phrase sequence $v_0 v_1 \cdots v_{l-1}$ can be represented by a mapping

$$s : \{0, 1, \ldots, l-2\} \to \{1, 2, \ldots, l-1\},$$

which indicates that $v_{s(m)}$ is the phrase modified by v_m. For a normal Japanese sentence, this mapping must satisfy

a) $m < s(m) \quad (\forall m \in \{0, 1, \ldots, l-2\})$,
b) if $m < n$ then $[s(m) \le n$ or $s(n) \le s(m)] \quad (\forall m, n \in \{0, 1, \ldots, l-2\})$.

A mapping satisfying the conditions a) and b) is referred to as a *dependency structure* on a phrase sequence $v_0 v_1 \cdots v_{l-1}$.

Now let $w_0 w_1 \cdots w_{M-1}$ be a sentence to be compressed. The sentence compression problem is formulated as a problem of extracting a *good* subsequence $w_{k_0} w_{k_1} \cdots w_{k_{N-1}}$ of length N ($N < M$) from the sentence. Let $p(w_n, w_m)$ be a function that represents the strength of inter-phrase dependency between w_n and w_m, or the degree of validity for w_n to modify w_m. Then the grammatical correctness of $w_{k_0} w_{k_1} \cdots w_{k_{N-1}}$ can be measured by $\max_s \sum_{n=0}^{N-2} p(w_{k_n}, w_{s(k_n)})$, where s runs over all the dependency structures on the phrase sequence. Let $q(w_n)$ be another function to represent the significance of w_n. Then the topical importance of the phrase sequence can be measured by $\sum_{n=0}^{N-1} q(w_{k_n})$. The total goodness of the phrase sequence $w_{k_0} w_{k_1} \cdots w_{k_{N-1}}$ is then defined as a weighted sum of the grammatical correctness and the topical importance [4]:

$$g(k_0, k_1, \ldots, k_{N-1}) \triangleq \begin{cases} q(w_{k_0}), & \text{if } N = 1; \\ \alpha\{\max_s \sum_{n=0}^{N-2} p(w_{k_n}, w_{s(k_n)})\} + (1-\alpha)\{\sum_{n=0}^{N-1} q(w_{k_n})\}, & \text{otherwise}, \end{cases}$$

where s runs over all the dependency structures on the subsequence of phrases, and α is a parameter to control the weights for the grammatical correctness and the topical importance. An efficient algorithm that maximises $g(k_0, k_1, \ldots, k_{N-1})$ has been reported[4].

3 Corpus and Subjects

Kyoto University Text Corpus [7] was used as language material. This corpus contains 38383 sentences selected from Mainichi Shinbun (Mainichi Newspaper), January ∼ December, 1995. Each sentence is given labels for word and phrase boundary, part-of-speech, as well as dependency structure. From this corpus, various sets of sentences were created for system parameter determination and for final evaluation as shown in Table 1. Table 2 shows subject groups employed for determination of system parameters and for final evaluation.

Table 1. Sentence sets.

Set	#Sentences	Remarks
A	34848	Estimation of dependency strength.
B	200	Estimation of phrase significance. $B \cap A = \phi$.
C	20	Estimation of α. $C \subset B$.
D	200	Final evaluation. $D \cap A = \phi,\ D \cap B = \phi$.

Table 2. Subject groups.

Group	#Subjects	Remarks
X	13	Estimation of phrase significance.
Y	2	Estimation of α. $Y \subset X$.
Z	1	Generation of compressed sentences by human. $Z \subset Y$.
W	5	Final evaluation. $W \subset X$. $Z \cap W = \phi$.

4 Determination of System Parameters

4.1 Inter-Phrase Dependency Strength

Inter-phrase dependency strength was defined on the basis of a morphological *dependency rule* and statistics for dependency distance [8]. Modifying phrases were classified into 219 classes according to the phrase-final word, while modified phrases were classified into 118 classes according to the left-most content word. First, a dependency rule $B(C_k, C_u)$ for modifying phrase class C_k , and modified phrase class C_u was defined using Set A in Table 1 as follows:

$$B(C_k, C_u) \triangleq \begin{cases} T, \text{ if there is a phrase in } C_k \text{ that modifies a phrase in } C_u; \\ F, \text{ otherwise.} \end{cases}$$

Also, the relative frequency $P(x, y)$ of dependency distance between phrases x and y, given the class to which x belongs as well as sentence-final/ non-final distinction for y, was calculated on Set A [8]. Based on the functions $B(C_k, C_u)$ and $P(x, y)$, the inter-phrase dependency strength was defined as

$$p(x, y) \triangleq \begin{cases} \log P(x, y), & \text{if } B(C_k, C_u) = T; \\ -\infty, & \text{if } B(C_k, C_u) = F, \end{cases}$$

where C_k and C_u are classes to which x and y belong, respectively.

4.2 Phrase Significance

In order to estimate the phrase significance, a preliminary experiment was conducted in which the sentences in Set B were compressed by the subjects in Group X. They were asked to compress each sentence at each of 5 compression rates: 80%, 65%, 50%, 35%, and 20%, where the compression rate means the ratio of the number of phrases in the compressed sentence to the number of phrases in the original sentence. The result was analysed statistically. First, phrases were classified into 13 classes according to the part-of-speech of the main content word, and also to the phrase-final function word when the main content word is a noun. Then the remaining rate of each phrase class at each compression rate was computed to define the phrase significance as follows:

1. Count the frequency $C(i)$ of phrases in the class i in the original sentences.
2. Count the frequency $C(i,k)$ of phrases in the class i in the compressed sentences at kth compression rate.
3. Compute the remaining rate of the class i: $R(i,k) = C(i,k)/C(i)$.
4. Normalise the distribution of $R(i,k)$: $F(i,k) = R(i,k)/\sum_i R(i,k)$.
5. Average the distribution over the steps of compression rate: $F(i) = (\sum_k F(i,k))/K$, where $K(=5)$ is the number of steps of compression rate.
6. Define the significance $q(i)$ of the phrase class i as $q(i) = \log F(i)$.

4.3 Parameter α

Another preliminary experiment was carried out to determine the optimum value of the parameter α. By using the phrase significance and the inter-phrase dependency strength obtained as in the previous subsections, automatic sentence compression was carried out for the sentences in Set C. Sentence compression was done at each of 5 compression rates with values of α varying in step of 0.1, and the total impression of each compressed sentence was evaluated with a score 1 (poor) $\sim$ 4 (good) by the subjects in Group Y. Table 3 shows the mean score per compressed sentence as a function of α. Thus, $\alpha = 0.6$ was found to give the best mean subjective score.

Table 3. Mean score as a function of α.

Value of α	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Mean Score	2.25	2.34	2.41	2.43	2.38	2.38	2.24

5 Sentence Compression Experiments

The Set D was used in the final sentence compression experiment. The distribution of sentence length in Set D, measured in the number of phrases, was as follows: 66 sentences of length 7 (short), 68 sentences of length 13 $\sim$ 14 (middle), and 66 sentences of length 18 $\sim$ 50 (long). The phrase significance, the dependency strength, and the value of α were set at values determined in the prelimimary experiments. Sentence compression was done at each of 5 compression rates. For comparison, sentence compression was also conducted by the subject in Group Z (only one subject) for the same sentence set. Furthermore, a random compression experiment was carried out. Thus, 15 compressed sentences (3 ways of compression multiplied by 5 steps of compression rate) were generated for each original sentence.

6 Subjective Evaluation

The total impression of each compressed sentence, taking both retention of the original information and grammatical correctness into account, was evaluated with a score 1 (poor) ~ 6 (good) by the subjects in Group W. They were presented the original sentence, and then the 15 compressed sentences in random order. They were not told how those compressed sentences were generated, and given enough time for evaluation. This process was repeated for the 200 sentences in Set D. Scores were averaged over the compressed sentences at each compression rate for each subject. Then, those scores were further averaged and standard deviation was calculated over the subjects at each compression rate as shown in Table 4.

Table 4. Mean score and standard deviation over the subjects.

	Compression Rate (%)				
	80	65	50	35	20
System	4.76±0.53	3.92±0.49	3.33±0.33	2.85±0.32	2.11±0.36
Human	5.53±0.37	4.88±0.48	4.41±0.38	3.73±0.29	2.61±0.33
Random	4.89±0.55	3.75±0.59	3.04±0.53	2.29±0.46	1.67±0.38

It is seen in Table 4 that compression by a human outperforms both compression by the system and random compression. For compression rates lower than, or equal to 65%, compression by the system is significantly better than random compression. As the compression rate becomes lower, the superiority of compression by the system over random compression becomes more obvious.

Table 5. Evaluation score by Subject 1 and Subject 4.

Subject 1	Compression Rate (%)				
	80	65	50	35	20
System	5.07	4.02	3.38	2.88	2.20
Human	5.82	5.17	4.65	3.80	2.74
Random	5.00	3.60	2.86	2.09	1.56

Subject 4	Compression Rate (%)				
	80	65	50	35	20
System	5.33	4.70	3.90	3.43	2.67
Human	5.86	5.39	4.81	4.15	3.18
Random	5.77	4.87	4.07	3.13	2.35

The standard deviations in Table 4 show that there are considerable variations among the scores by different subjects. It is also noted that in the case of random compression, the standard deviations are larger than those in other two cases. Table 5 shows two examples of scores by different subjects. Subject 1 judged that compression by the system is considerably better than random compression, while Subject 4 did not. Also, Subject 4 tends to give higher scores than Subject 1 on the whole. The large variations among subjects might come from the fact that they were asked to evaluate compressed sentences from a view point of information retention and grammatical correctness altogether. Therefore, it is possible that some subjects payed more attention on grammatical correctness than on information retention, while others did not. To resolve this problem, separate evaluation for information retention and grammatical correctness will be necessary.

7 Conclusion

Based on the algorithm previously proposed, a Japanese sentence compression experiment was conducted. The result of subjective evaluation showed that sentence compression using the algorithm is significantly better than random compression, though not reaching the human performance. Our future work includes

1) Improvement on the definitions of the phrase significance and the dependency strength.
2) Separate evaluation for information retention and grammatical correctness.
3) Employment of more subjects to generate human-compressed sentences.

References

1. Okumura, M., Nanba, H.: Automated text summarization: A survey. Journal of Natural Language Processing **6**(6)(1999) 1–26.
2. Wakao, T., Ehara, T., Shirai, K.: Summarization methods used for captions in TV news programs. Technical Report of Information Processing Society of Japan, 97-NL-122-13 (1997) 83–89.
3. Mikami, M., Masuyama, S., Nakagawa, S.: A Summarization method by reducing redundancy of each sentence for making captions for newscasting. Journal of Natural Language Processing **6**(6) (1999) 65–81.
4. Oguro, R., Ozeki, K., Zhang, Y., Takagi, K.: An efficient algorithm for Japanese sentence compaction based on phrase importance and inter-phrase dependency. Proc. TSD 2000 (LNAI 1902) (2000) 103–108.
5. Knight, K., Marcu, D.: Statistics-based summarization – Step one: Sentence compression. AAAI/IAAI 2000 Proceedings (2000) 703–710.
6. Hori, C., Furui, S.: Advances in automatic speech summarization. Proc. Eurospeech 2001 **3** (2001) 1771–1774.
7. Kyoto University Text Corpus Version 2.0 (1998). http://pine.kuee.kyoto-u.ac.jp/nl-resource/corpus.html
8. Zhang, Y., Ozeki, K.: Dependency analysis of Japanese sentences using the statistical property of dependency distance between phrases. Journal of Natural Language Processing **4**(2) (1997) 3–19.

This article was processed using the LaTeX macro package with LLNCS style