Unsupervised Acquisition of knowledge about the abbreviation possibility of some of multiple phrases modifying the same verb/noun

Hiroyuki SAKAI Shigeru MASUYAMA
Department of Knowledge-based Information Engineering
Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi 441-8580, Japan
sakai@smlab.tutkie.tut.ac.jp, masuyama@tutkie.tut.ac.jp

Abstract

This paper proposes a statistical method of acquiring knowledge about the abbreviation possibility of some of multiple phrases modifying the same verb/noun. Our method calculates weight values of multiple phrases by mutual information based on the strength of relation between the phrases among the multiple phrases and modified verbs/nouns. Among phrases modifying the same verb/noun, those having relatively low weight value are deleted. The evaluation of our method by experiments shows that the precision attains about 74.0% and the recall attains about 43.0%.

Keywords: *abbreviation of multiple phrases, corpus, summarization.*

1 Introduction

Recent rapid progress of computer and communication technologies enabled us to access enormous amount of machine-readable information easily. This, however, has caused so called the information overload problem. Under these circumstances, the necessity for automatic summarization has been increasing and has been intensively studied recently (see e.g,[6]). No all-purpose automatic summarization method for summarizing any type of documents appropriately exists, thus, in many cases, we must combine several summarization methods as components when we develop an automatic summarization system. Deletion of some unnecessary parts from a sentence is among such important summarization methods[12][3].

In this paper, we propose a statistical learning method which acquires knowledge about the deletion possibility of some of multiple phrases modifying the same verb/noun in order to summarize a document by deleting unimportant segments from a sentence in the document. As a method of summarization which deletes unimportant segments from a sentence, Ohtake et al.[7][8] proposes a method which deletes one of two phrases modifying the same noun by using manually constructed rules. Knight et al.[5] and Jing[2] and Takeuchi et al.[11] propose methods which extract rules for reducing a sentence from aligned corpus of human-made summaries and their original manuscripts. As for such application of supervised learning using aligned corpus, Katoh et al.[4] proposes a method which acquires knowledge about paraphrasing by using an aligned corpus of manuscripts for character broadcasting (i.e., a kind of summaries) and corresponding original TV news manuscripts.

However, these previously proposed methods have the following drawbacks for practical use.

- To make a complete list of rules manually in order to delete unimportant segments from a sentence is a hard task.
- Aligned corpus between original manuscripts and summaries is useful. But it is not necessarily available. Moreover, constructing summaries manually for obtaining such aligned corpus is a time-consuming and costly task.

By these reasons, we propose a statistical learning method which acquires knowledge from a general corpus (e.g., news paper articles), about the deletion possibility of some of multiple phrases modifying the same verb/noun to summarize by deleting unimportant segments from a sentence. Note that multiple phrases denote some phrases modifying the same verb/noun and a phrase denotes one of the multiple phrases throughout this paper unless specified otherwise. An example is shown as follows. In this example, three phrases modify a verb "*wasurereu*(忘れる:forget)";

Example 1: *watasitatiha kessite 1・17wo wasuretehanaranai* (私たちは決して「１・１７」を忘れてはならない: We must never forget "1・17".),
phrase 1: *watasitatiha* (私たちは: We),
phrase 2: *kessite* (決して: never),
phrase 3: *1・17wo* (「１・１７」を: "1・17"),
modified verb: *wasureru* (忘れる: forget),

□

Sakai et al.[9][10] propose statistical methods which acquire knowledge from a general corpus about the deletion possibility of adnominal verb phrases and phrases modifying a verb, respectively. The method proposed in [9] acquires knowledge about the deletion possibility of adnominal verb phrases. The method proposed in [10] acquires knowledge about the deletion possibility of phrases modifying a verb. However, the method [9] is applied only to the adnominal verb phrases and the method [10] is applied only to the phrases modifying a verb, the method to be proposed in this paper applies to all phrases among the multiple phrases.

The method proposed by Ohtake et al.[7][8] deletes one of two phrases modifying the same noun by using manually-constructed rules and the system YELLOW which implemented the method exhibits a good performance at TSC1 task B in NTCIR Workshop 2. However, the method can only be applied to sentences having structures matching the rules and constructing such rules manually is a hard task. In contrast, our method deletes some of multiple phrases having unimportant contents by using statistical information obtained from a single corpus. Consequently, more flexibility in deletion of some of multiple phrases is attained. Moreover our method can be applied to not only multiple phrases modifying the same noun but also multiple phrases modifying the same verb. In general, the multiple phrases modifying the same verb appear more frequently than those modifying the same noun (as will be seen in Table 2).

Our method extracts knowledge about deletion of some of multiple phrases from a single corpus, e.g., news paper articles provided as a machine-readable form, documents obtained from WWW which are easily available.

We participate in the TSC2 task A in NTCIR Workshop 3 to evaluate our method and show results of this task in this paper. Moreover, we also evaluate our method by recall and precision in addition to the evaluation by TSC2.

We introduce our method in Sec. 2 and its implementation and experiments for evaluation are illustrated in Secs. 3 and 4, respectively. We show results of the TSC2 task A in Sec. 5. We analyze the results of the experiments in Sec. 6. Sec. 7 concludes this paper.

2 Proposed method

Our method is based on the intuition that some phrases among the multiple phrases which are easily associated with by modified verb/noun can be deleted. For example, in "Example 1", phrase "*kessite*(決して: never) is easily associated with by modified verb "*wasuresu*(忘れる: forget)".

To reflect the intuition, we calculate the weight value assigned to the strength of the relation between a phrase among multiple phrases and a modified verb/noun by the phrase, and phrases having relatively small weight value comparing with other phrases among the multiple phrases are deleted. And, a phrase having the largest weight value among the multiple phrases is not deleted. The weight value assigned to the strength of the relation is calculated by a formula based on mutual information and it has small weight value if the mutual information has a large value. That is, the combination of the phrase among the multiple phrases and the modified verb/noun, which has a large mutual information value, is frequently contained in a corpus. Consequently, such phrases are easily associated with by the verb/noun.

We introduce a deletion method of some of multiple phrases modifying the same verb in Sec 2.1 and deletion of some of multiple phrases modifying the same noun in Sec 2.2.

2.1 Deletion of some of multiple phrases modifying the same verb

We introduce our deletion method of some of multiple phrases modifying the same verb. An example of multiple phrases modifying the same verb is shown as follows.

Example 2: *watasitatiha kessite 1・17 wo wasurete-hanaranai* (私たちは決して「１・１７」を忘れてはならない: We must never forget "1・17".),
phrase 1: *watasitatiha* (私たちは: We),
phrase 2: *kessite* (決して: never),
phrase 3: *1・17 wo* (「１・１７」を: 1・17),
modified verb: *wasureru* (忘れる: forget),

□

This example shows that verb "*wasureru* (忘れる: forget)" is modified by three phrases, which are *watasitatiha* (私たちは: we), *kessite* (決して: never) and *1・17 wo* (「１・１７」を: 1・17). We define such a structure of a sentence as the multiple phrases and denote one of the multiple phrases as $E(M, c, V)$,
where,

$E(M, c, V)$**:** the phrase containing word M and modified verb V by relation operator c,

M**:** a word contained in the phrase $E(M, c, N)$ and modifying verb V,

c**:** the relation operator showing relation between word M and verb V,

where, M is a word modifying verb V. The part of speech of M is either of a noun, a verb, an adjective, an adverb, or a demonstrative. However, an ancillary words are excluded (for example suffixes and post-positional particles., etc.). For example, when a phrase is *1・17 wo*(「1・17」を: 1・17), M is the word *1・17*. If word M is a noun, M is the semantic code (feature) of the noun, where, we employ the "単語体系 (*Tango taikei*: vocabulary system)" in the thesaurus "日本語語彙大系 (*Nihongo goi taikei*)"[1] as a dictionary of semantic codes (features). However, if word M is a noun which does not have a semantic code (feature), M is the noun. Moreover, if noun N is a compound noun, noun N is replaced with a noun contained at the end of the compound noun. For example, N is "*Tougou*(統合: unification)", when noun N is "*Shijo Tougou*(市場統合: market unification)". (The "*Shijo Tougou*(市場統合: market unification)" is a compound noun composed of a noun:"*Shijo*(市場: market)" and a noun:"*Tougou*(統合: unification)").

c shows the relation between word M and verb V. We define c as a relation operator. For example, when a phrase is *1・17 wo*(「1・17」を: 1・17), c is shown as "ヲ格 (*wo kaku*: *wo*-case)", because a case post-positional particle is "ヲ格 (*wo kaku*: *wo*-case)" in this phrase. Some relation operators are exemplified in Table 4, where, the symbol of relation operator c is based on the results analyzed by KNP[1] version 2.0b6 which we employ as a parser.

2.1.1 Algorithm for deleting some of multiple phrases modifying the same verb

Based on the above observations, our method of deleting some of multiple phrases modifying the same verb is formally described as follows:

Algorithm for deleting some of multiple phrases modifying the same verb

[Step 1:] Weight value $W(E(M,c,V))$ of one of multiple phrase $E(M,c,V)$ is calculated by the following formula :

$$W(E(M,c,V)) = \frac{A(M) \times P(E(M,c,V))}{f(M,c,V)} \cdot \frac{1}{\exp(I(M,c,V))}, \quad (1)$$

$$I(M,c,V) = \log(S \cdot \frac{f(M,c,V)}{f(M,c) \cdot f(c,V)}), \quad (2)$$

where

$f(M,c,V)$**:** frequency of phrases which include word M and modify verb V by relation operator c in the corpus,

$f(M,c)$**:** frequency of phrases which include word M and modify verbs by relation operator c in the corpus,

$f(c,V)$**:** frequency of phrases which modify verb V by relation operator c in a corpus,

S**:** frequency of all multiple phrases in the corpus,

$A(M)$**:** frequency of part of speech of M in the corpus, the part of speech of M is either a noun, a verb, an adjective, a conjunction, an adverb, or a demonstrative pronoun. However, if the part of speech of M is a noun, $A(M)$ is set to be $A(M)$/2, because the frequency of part of speech of noun is too high comparing to those of other parts of speech.

$P(E(M,c,V))$**:** the number of clauses contained in phrase $E(M,c,V)$.

[Step 2:] $Ws(E(M,c_j,V))$ is calculated by the following formula :

$$Ws(E(M_j,c_j,V)) = \frac{W(E(M_j,c_j,V))}{\max_{j=1,2,\ldots,k} W(E(M_j,c_j,V))}, \quad (3)$$

here, verb V is modified by k phrases, that is, verb V is modified by $E(M_1,c_1,V),\ldots,E(M_k,c_k,V)$.

[Step 3:] Delete phrase $E(M_j,c_j,V)$ having weight value $Ws(E(M_j,c_j,V))$ smaller than a threshold value predetermined by trial and error.

□

2.1.2 Explanation of the Algorithm to delete phrases modifying the verb

We calculate the weight value assigned to the strength of the relation between a phrase among multiple phrases and the modified noun in **Step 1**. The relative values to a phrase in the multiple phrases are calculated in **Step 2**. And, the phrases having relatively small weight value comparing to those of other phrases are deleted in **Step 3**. Hence, the phrase having the largest weight value is not deleted among the multiple phrases.

$I(M,c,V)$ is the mutual information between phrase $E(M,c,V)$ containing word M and verb V modified by relation operator c. If the mutual information has a large value, the weight value $Ws(E(M,c,V))$ is small, and as a result, the phrase $E(M,c,V)$ tends to be deleted. The combination of phrase $E(M,c,V)$ and the modified verb V having a large mutual information value is frequently contained in the corpus and such a phrase is easily associated with by verb V.

[1] http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html

Table 1. Some relation operators (multiple phrases modifying the same verb)

part of speech of M	relation operators
noun	ガ格 (*ga kaku*: *ga*-case), ヲ格 (*wo kaku*: *wo*-case)
noun	デ格 (*de kaku*: *de*-case), ニ格 (*ni kaku*: *ni*-case)
verb	複合辞連用 (*hukugouzi renyou*: verb phrases)
adjective	弱連用 (*zyaku renyou*: adjective phrases)

Table 2. Frequency of part of speech of M

part of speech of M	phrases modifying verbs	phrases modifying nouns
noun	1277730	77754
verb	178203	22217
adjective	53918	12709
adverb	70479	2243
demonstrative pronoun	15763	4086
conjunction	19088	108
S	1615181	119117

$A(M)$ is the frequency of the part of speech of M modifying verb V in corpus. Thus, it is difficult to delete phrases containing word M with large $A(M)$ value (for example, the part of speech of M is a noun or a verb). But, it is easy to delete phrases containing word M with small $A(M)$ value (for example, the part of speech of M is an adjective or a conjunction or an adverb). Moreover, the value of $A(M)$ changes with phrases modifying the same verb or phrases modifying the same noun. Thus, degree of deleting some of multiple phrases changes with phrases modifying nouns or phrases modifying verbs. Table 2 shows $A(M)$ when we use 66686 documents from Nikkei newspaper articles from January 1, to June 31, 1993, as a corpus. $P(E(M, c, V))$ is the number of clauses contained in phrase $E(M, c, V)$. For example, when phrase $E(M, c, V)$ is *kessite*(決して: never), $P(E(M, c, V)) = 1$. If the phrase $E(M, c, V)$ has a large $P(E(M, c, V))$ value, the phrase $E(M, c, V)$ is hard to delete because its deletion causes serious information loss. Consequently, phrases with $P(E(M, c, V)) \geq 4$ are not deleted by our method.

In addition, phrases modifying verbs "*suru*(する: do), *naru*(なる: be), *aru*(ある: be)" are not deleted by our method, because, the phrases modifying their verbs are hardly associated with by them. Next, phrases modifying verbs by relation operators shown in Table 3 are not deleted by our method. This is because the phrases containing their relation operators are a subjective case or often have important contents, then deletion causes serious information loss.

Table 3. Relation operators which undeleted phrases modify verbs by

ガ格 (*ga kaku*: *ga*-case)	ヲ格 (*wo kaku*: *wo*-case)
ニ格 (*ni kaku*: *to*-case)	ト格 (*to kaku*: *to*-case)
未格 (*mi kaku*: null-case)	〜と (〜*to*: *to*)

2.2 Deletion of some of multiple phrases modifying the same noun

We introduce a deletion method of some of multiple phrases modifying the same noun. An example of multiple phrases modifying the same noun is shown as follows.

Example 3: *seiken wo obiyakasu yuuryoku na raibaru ga sonnzai sinai* (政権を脅かす有力なライバルが存在しない: There is no strong rival who threatens the administration exists.),
phrase 1: *seikein wo obiyakasu* (政権を脅かす: who threatens the administration),
phrase 2: *yuuryoku na* (有力な: strong),
modified noun: *raibaru* (ライバル: rival),

□

This example shows that noun "*raibaru* (ライバル: rival)" is modified by two phrases, which are *seiken wo obiyakasu* (政権を脅かす: threatening the administration) and *yuuryokuna* (有力な: strong). We define such a structure of a sentence as the multiple phrases, and denote the multiple phrases as $E(M, c, N)$, where,

$E(M, c, N)$**:** the phrase containing word M and modifying noun N by "relation operator" c,

M**:** a word contained in the phrase $E(M, c, N)$ and modifying noun N,

c**:** the "relation operator" showing relation between word M and noun N,

c shows relation between word M and noun N. For example, when the phrase is *seiken wo obiyakasu*(政権を脅かす: threatening the administration), the c is shown as "動詞連体 (*doushi rentai*: adnominal verb phrase)", because the phrase is an adnominal verb phrase. Some relation operators are exemplified in Table 4.

2.2.1 Algorithm for deleting some of multiple phrases modifying the same noun

Based on the above observations, our method of deleting some of multiple phrases modifying the same noun is formally described as follows:

Algorithm for deleting some of multiple phrases modifying the same noun
[Step 1:] Weight value $W(E(M, c, N))$ of one of multiple phrases $E(M, c, N)$ is calculated by the following formula :

$$W(E(M,c,N)) = \frac{A(M)\cdot P(E(M,c,N))}{f(M,c,N)} \cdot \frac{1}{\exp(I(M,c,N))}, \quad (4)$$

$$I(M,c,N) = \log(S \cdot \frac{f(M,c,N)}{f(M,c)\cdot f(c,N)}) \quad (5)$$

where

$f(M, c, N)$**:** frequency of phrases which include word M and modify noun N by relation operator c in the corpus,

$f(M, c)$**:** frequency of phrases which include word M and modify verbs by relation operator c in the corpus,

$f(c, N)$**:** frequency of phrases which include words and modify noun N by relation operator c in the corpus,

S**:** frequency of all multiple phrases in a corpus.

$A(M)$**:** frequency of the part of speech of M in the corpus. However, if the part of speech of M is a noun, $A(M)$ is set to be $A(M)$/2.

$P(E(M, c, N))$**:** the number of clauses constructing phrase $E(M, c, N)$.

Table 5. Relation operators which undeleted phrases modify nouns by

ガ格 (*ga kaku*: *ga*-case)	ヲ格 (*wo kaku*: *wo*-case)
ト格 (*to kaku*: *to*-case)	ノ格 (*no kaku*: *no*-case)
未格 (*mi kaku*: null-case)	隣接 (*rensetu*: adjoin)
～と (～*to*: *to*)	

[Step 2:] $Ws(E(M, c_j, N))$ is calculated by the following formula :

$$Ws(E(M_j,c_j,N)) = \frac{W(E(M_j,c_j,N))}{\max_{j=1,2,\ldots,k} W(E(M_j,c_j,N))}, \quad (6)$$

here, noun N is modified by k phrases, that is, noun N is modified by $E(M_1, c_1, N), \ldots, E(M_k, c_k, N)$.

[Step 3:] Delete phrase $E(M_j, c_j, N)$ having weight value $Ws(E(M_j, c_j, N))$ smaller than a threshold value predetermined by trial and error.

□

In addition, phrases modifying nouns "*koto*(こと: thing), *mono*(もの: thing)" are not deleted by our method. Next, phrases modifying verbs by relation operators shown in Table 5 are not deleted by our method.

3 Implementation

We implemented our method for participating in TSC2 task A. We use 61637 documents from Mainichi newspaper articles from January 1, to June 31, 1998, as a document set. Because, our method is applied to the system developed for participating in TSC2 task A and the Mainichi newspaper articles are employed as a corpus in TSC2 task A. We employ JUMAN[2] version 3.5 as a morphological analyzer, and KNP version 2.0b6 as a parser. Our method acquires knowledge about the deletion possibility of some of multiple phrases, and deletes some of multiple phrases by using the acquired knowledge. The phrases deleted by our method are exemplified as follows. Note that the multiple phrases deleted by our method are underlined.

Example 4: *watasitatiha kessite 1.17 wo wasureteha naranai* (私たちは 決して「１・１７」を忘れてはならない: We must not never forget "1・17".)

Example 5: *tokuni konngo kyouwakokou ga nigiru afurikahyou wo naniganandemo kakuhosuru to tikatta* (とくに コンゴ共和国が握るアフリカ票を なにがなんでも 確保すると誓った: In particular, he promised that the Africa vote which Republic of Congo has was surely secured.)

[2] http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html

Table 4. Some relation operators (multiple phrases modifying the same noun)

part of speech of M	relation operators
noun	ノ格 (*no kaku*: *no*-case)
noun	ト格 (*to kaku*: *to*-case)
verb	動詞連体 (*dousi rentai*: adnominal verb phrase)
adjective	形判連体 (*keihan rentai*: adnominal adjective phrase)

Example 6: *chuugoku deno hatuno gorin kaisai ni syu-unenn wo miseru pekinn nado...* (中国での 初の五輪開催に執念をみせる北京など...: ..., including Beijing which shows a deep attachment to hold the first Olympic Games in China.)

4 Evaluation by TSC2 in NTCIR Workshop 3

We participate in TSC2 in NTCIR Workshop 3 for evaluation of our method and analyze results of this task in this paper. Table 6 shows the results of TSC2 task A by the summarization system, which uses our method as an element constructing it. The system is constructed by an extraction method of important sentences and our deletion method of unnecessary parts from a sentence. We employ the extraction method implemented in a system YELLOW[7][8], which exhibited a good performance at TSC task A in NTCIR Workshop 2. Table 6 also shows the results of TSC2 by the system YELLOW to compare with our method.

5 Experiments for evaluation

We evaluate our method by precision and recall in addition to the evaluation by TSC2 in NTCIR Workshop 3. For this purpose, we implement our method by employing 66686 documents from Nikkei newspaper articles from January 1, to June 31, 1993, as a corpus and we choose documents for deleting some of multiple phrases by our method. We manually make a correct data set which shows multiple phrases appropriate to be deleted among the chosen documents. Note that the precision and the recall are defined as follows.

$$\begin{aligned} Recall\ R &= freq(A)/freq(C), \\ Precision\ P &= freq(A)/freq(M), \end{aligned}$$

where,

freq(A): the frequency of the same phrases shown by the correct data set with phrases deleted by our method,

freq(C): the frequency of phrases shown by the correct data set,

freq(M): the frequency of phrases deleted by our method.

Table 7. Result of comparing our method with YELLOW

Method	Precision(%)	Recall(%)	Deleted
Our method	71.2	49.1	73
Yellow	65.8	48.6	79

5.1 Evaluation of our method of deleting some of multiple phrases modifying the same noun

The method proposed by Ohtake et al.[7][8] deletes one of the two phrases modifying the same noun by using 36 manually-constructed rules and the system by using the method exhibited a good performance at TSC1 task B in NTCIR Workshop 2. We compare our method with the method proposed by Ohtake et al[7][8], and call the method YELLOW. We choose 85 documents from 66686 articles of Nikkei newspaper from January 1, to June 31, 1993. And we manually make a correct data set which shows multiple phrases modifying the same noun appropriate to be deleted among the 85 documents. There are 135 multiple phrases modifying the same noun in the 157 documents. Table 7 shows the results of comparing our method with YELLOW. The threshold values of our method are adjusted so that the recall of our method coincides with that of YELLOW. Next, Table 8 shows a part of the results of precision and recall which are calculated for threshold values changed from 0.12 to 0.32 on our method.

5.2 Evaluation of our method of deleting some of multiple phrases modifying the same verb

We choose 12 documents from 66686 articles of Nikkei newspaper from January 1, to June 31, 1993. And we manually make a correct data set which shows multiple phrases appropriate to be deleted among the 12 documents. There are 199 multiple phrases modifying the same verb in the 12 documents. Table 9 exemplified a part of the results of precision and recall calculated for threshold values changed from 0.12 to 0.32 on our method.

Table 6. The results of TSC2 in NTCIR Workshop 3

System	C 20%	R 20%	C 40%	R 40%
System by using our method	2.53	2.87	2.60	2.77
System Yellow	2.67	2.97	2.50	2.77
System by using tf	3.30	3.30	3.20	3.10
manually summarization	2.33	2.20	2.10	2.03

Table 8. Result of recall and precision at multiple phrases modifying the same noun

Threshold	Precision(%)	Recall(%)	Deleted
0.12	75.4	40.6	57
0.14	76.7	43.4	60
0.16	76.7	43.4	60
0.18	74.6	44.3	63
0.2	75.0	45.3	64
0.22	73.1	46.2	67
0.24	72.5	47.2	69
0.26	70.4	47.2	71
0.28	70.8	48.1	72
0.3	70.8	48.1	72
0.32	71.2	49.1	73
Average	73.4	45.7	66.2

Table 9. Result of recall and precision at multiple phrases modifying the same verb

Threshold	Precision(%)	Recall(%)	Deleted
0.12	75.0	39.0	72
0.14	76.0	41.2	75
0.16	75.0	41.2	76
0.18	75.3	41.9	77
0.2	75.6	42.6	78
0.22	74.7	42.6	79
0.24	75.3	44.0	81
0.26	73.5	44.0	83
0.28	73.5	44.0	83
0.3	74.1	45.5	85
0.32	74.1	45.5	85
Average	74.7	42.9	79.5

6 Discussion

As shown in Table 7, the precision and the recall of the deletion of some of multiple phrases by our method attain about 74.0% and 43.0%, respectively. Moreover, Table 7 shows that the precision of our method is superior to that by the method of YELLOW when the recall of our method is adjusted to coincide with that of YELLOW. Thus our method is promising as a component to delete some of multiple phrases. We consider that the reason why our method outperforms YELLOW is that 36 deletion rules of YELLOW do not cover all types of multiple phrases modifying nouns, as they are manually constructed. In contrast, our method uses the importance for any phrase among multiple phrases in the documents. Note that the importance is increased when the weight value assigned to the strength of the relation between a phrase among multiple phrases and a modified verb/noun has a large value. However, YELLOW does not use such kind of information.

7 Conclusion

We proposed a statistical learning method which acquires knowledge about the deletion possibility of some of multiple phrases from a news paper corpus provided in a machine-readable form in order to summarize by deleting unimportant segments from a sentence. Actually, our method deletes some of multiple phrases which are easily associated with. We evaluate our method, and we conclude that our method is able to delete some of multiple phrases appropriately, because the precision and the recall attain about 74.0% and 43.0%, respectively. Experimental results show that our method is useful for deleting some of multiple phrases.

Acknowledgment

We express our gratitude to Nikkei Shinbun, Inc. and MAINICHI NEWSPAPERS, Inc. who permits us to use the news paper articles in a machine readable form. The authors are also grateful to Nippon Telegraph and Telephone Corporation for permitting use of ALT-JAWS Ver.2.0, a morphological analyzing library for Japanese, to obtain semantic codes from Goitaikei.

References

[1] S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Oyama, and

Y. Hayashi(editors). *Goi-taikei*. Iwanami Publishing, CD-ROM edition, 1999.

[2] H. Jing. Sentence reduction for automatic text summarization. In *Proc. of the 6th Conference on Applied Natural Language Processing*, pages 310–315, 2000.

[3] H. Jing and K. McKeown. Cut and paste based text summarization. In *Proc. of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 178–185, 2000.

[4] N. Katoh and N. Uratani. A new approach to aquiring linguistic knowledge for locally summarizing japanese news sentences. *Natural Language Processing*, 6(7):73–92 (in Japanese), 1999.

[5] K. Knight and D. Marcu. Statistics-based summarization –step one: Sentence compression. In *Proc. of AAAI2000*, pages 703–710, 2000.

[6] I. Mani and M. T.Maybury. *Advances in Automatic Text Summarization*. the MIT Press, 1999.

[7] K. Ohtake, D. Okamoto, M. Kodama, and S. Masuyama. Yet another summarization system with two modules using empirical knowledge. In *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, pages 331–340, 2001.

[8] K. Ohtake, D. Okamoto, M. Kodama, and S. Masuyama. A summarization system yellow for japanese newspaper articles. *IPSJ Transactions on Databases*, 43(SIG2(TOD13)):37–47 (in Japanese), 2002.

[9] H. Sakai and S. Masuyama. Unsupervised knowledge acquisition about the deletion possibility of adnominal verb phrases. In *the Proceedings of Workshop on Multilingual Summarization and Question Answering 2002 (post-conference workshop to be held in conjunction with COLING-2002)*, pages 49–56, 2002.

[10] H. Sakai, N. Shinohara, S. Masuyama, and K. Yamamoto. Knowledge acquisition about the abbreviation possibility of verb phrases. *Natural Language Processing*, 9(3):41–62 (in Japanese), 2002.

[11] K. Takeuchi and Y. Matsumoto. Acquistion of sentence reduction rules for improving quality of text summaries. In *the Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, pages 447–452, 2001.

[12] K. Yamamoto, S. Masuyama, and S. Naito. Green: An experimental system generating summary of japanese editorials by combining multiple discourse characteristics. *Natural Language Processing*, 2(1):39–55 (in Japanese), 1995.