Trainable Automatic Text Summarization Using Segmentation of Sentence

Kai ISHIKAWA Shin-ichi ANDO Shin-ichi DOI Akitoshi OKUMURA
Multimedia Research Laboratories, NEC Corporation
4-1-1 Miyazaki Miyamae-ku Kawasaki-shi Kanagawa 216-8555, Japan
{k-ishikawa@dq, s-ando@cw, s-doi@ah, a-okumura@bx}.jp.nec.com

Abstract

In this paper, we propose an automatic summarization method combining conventional sentence extraction and trainable classifier based on Support Vector Machine. To make extraction unit smaller than the original sentence extraction, we also introduce sentence segmentation process in our method. The evaluation results show that our system achieves the best result among all the other systems with regard to contents, and closer to the human constructed summaries (upper bound) at 20% summary rate. On the other hand, the system needs to improve readability of its summary output.

Keywords: *automatic text summarization, manually constructed summary, unit of extraction, machine learning, dividing sentences into clauses, attribute values, Support Vector Machine.*

1 Introduction

One of the challenging issues of our summarization study is the improvement of accuracy in evaluating importance of parts in articles to compose summaries. The basic concepts of the extraction method are well known [10][2]. When we participated in the previous NTCIR Workshop, i.e. NTCIR-2 TSC task-A (sentence extraction task), we tried to improve the precision of sentence extraction by optimizing the combination of heuristic rules. We utilized headline information, term frequencies, and sentence position as useful information to calculate a sentence importance, and studied the way of combining then into a score and obtained improvements.

In our previous work, we have the following two problems left. The first one is the degradation of sentence score accuracy by using a linear combined score of multiple score functions based on heuristic rules. In this formulation, it is difficult to make one heuristic rule most credible among the heuristics rules to judge the importance of an extraction unit. We should introduce some automatic method to determine the coupling coefficients which we optimized manually according to some experimental results in the previous work. The second one is that the minimum extracting unit is a sentence, which needs to be smaller when we have longer sentences in an article containing both important and unimportant parts.

In this paper, we try to improve these problems in our previous summarization work by introducing the following two procedures: (1) Machine learning using human constructed summaries. (2) Using clauses/sub-sentence as extraction unit. Before going through the procedures (1) and (2) in details, we want to review some of related works in corpus base extraction methods. In these types of methods, trainable classifiers are used to classify sentences into important class and un-important class. Kupiec, et.al. [9] proposed a method using statistical classifier based on Bayes' rule. Nomoto, et.al. [13] and Okumura [14] applied decision tree learning (C4.5 [15]) to obtain a sentence classifier.

Rhetorical Structure Theory (RST) proposed by Mann and Thompson [11] is applied to automatic summarization by Marcu [12]. An elemental unit of RST corresponds to clauses/sub-sentence. In Marcu's method, importance of each unit is calculated according to the hierarchical depth of a rhetorical structure. Kobori et.al. [8]also applied RST in their summarization method. His method uses phrase as an extraction unit, which is much smaller than the unit of RST. They use morphological information beside rhetorical relations, and applied C4.5 to obtain a classifier for phrases from training data composed of 1000 paragraphs from newspapers and academic papers randomly selected. They report the problem of producing ungrammatical sentences in summaries.

Knight [7] proposed a statistical summarization method based on noisy channel model, its mathematical formulation is almost the same with that of statistical machine translation. This method composes a summary sentence composed of selected words from the word sequence of an input sentence, and generates

much precise summaries than by phrase extraction. However, large amount of parallel corpus (word alignment of source sentences and summary sentences is obtained) is necessary to construct a statistical model, and the information of discourse level is not yet considered in the present formalism.

As we can see from these related works, reducing the size of extraction unit entails increase of cost to construct training data and risk to generate ungrammatical sentences as well as increases the preciseness of extraction to generate summary sentences. Our method processes summarization based on our conventional extraction method using heuristic rules, combined with automatic trainable classifiers. A sentence or a clause is used as an extracted unit. The advantage of this method is the use of manually summarized data as its training data. The method also resolves the problem of summarization particles being too coarse with long sentences contained in the text.

2 System Description

2.1 Approaches

Our summarization system uses a conventional key sentence extraction method, where we use a clause as a minimum unit of extraction as opposed to a sentence commonly used as a minimum unit. The summarizer operates segmentation of compound sentences based on cue words like "connecting expressions", and obtains clauses / sub-sentences as units of extraction. Each segmented unit is "repaired" to improve its readablity by converting the form of declinable word at the end of the units to a complete form. When a sentence is not segmented, it becomes the unit of extraction itself.

The summarizer is combined with classifiers obtained by machine learning using Support Vector Machine (SVM). Summarization by sentence extraction is considered as a classification task, where each sentence is classified as relevant or non-relevant for the summary extracted patterns. The system automatically learns the manually annotated summaries. Here, the vector attribute factors used in the SVM learning are; position of the sentences in the text, tf scores, similarity with the headline sentence, occurrence of cue words such as conjunctions, and document genres. The SVM classifier assigns real value $\sigma(-1 \leq \sigma \leq 1)$ to each unit as classification result. We consider this σ value as a score of relevance for each unit, and combine it into the new score by calculating a linear combination of σ and the relevance score of conventional method. Those with highest scores are extracted and compose a summary text.

2.2 Process flow

The following process flow depicts our method, grouped into (A) training process and (B) summary generation process.

(A) Training process

(A-1) Apply morphological analysis to the sentences in the articles of training data.

(A-2) Extract attribute values from the sentences (refer to table 1).

(A-3) Generate a classifier by training the sentences in the training data (decision of extracting each sentence while summarization are mannually annotated) and the attributes values using SVM.

(B) Summary generation process

(B-1) Apply morphological analysis to the sentences in the input article.

(B-2) Obtain the units of extraction from the sentences using segmentation of comound sentences at clause boundaries.

(B-3) Repair the segmented units by converting them to a complete end-form.

(B-4) Extract attribute values from the each unit of extraction (refer to table 1).

(B-5) Calculate the scores of each units based on a conventional extraction method.

(B-6) Calculate the relevancy σ using SVM classifier.

(B-7) Combine the scores of conventional method and the SVM classifier.

(B-8) Generate summary text by extracting the higher ranked units of extraction in their original order in the input article.

2.3 Segmentation of sentences into clauses / sub-sentences

It is known that the connecting expressions in Japanese, which connect clauses, can be classified into some conjunction levels based on their "clause connection strength" to the main sentence [6] [1]. In our approach, we use the expressions of the highest connection level to detect the sentence boundaries. Segmented sub-sentences become independent sentences (or clauses) as described in the followings.

Method:

1. Find connecting expressions in a sentence that indicate the break points for clauses/sub-sentences.
2. Evaluate the independency of each component in their contexts.
   - Inhibit segmentations within the expression "～ば… (if～then…)".
   - Inhibit segmentations within the parenthesis.
3. Repair each segmented unit by converting it to a complete end-form to improve their readability.

The next two examples show cases where (a) the segmentation is applied and (b) the segmentation is not applied.

(a) An example of sentence where segmentation process IS APPLIED:

Before segmentation process:
[1 大手会社は直営の映画館も持っており、(Since major companies own movie theaters under their direct management,)[2 客の入ら/ない映画館を維持するためには、高額の製作費がかかる自社作品を作るより、独立プロが製作した作品を有利な契約で上映する方が安上がりだし、(it is cheaper to show the movies produced by independent productions under profitable contract than to make in-house products with high production cost, to maintain the low-profit theaters, and plus,)[3 大きなリスクを背負うこともない。(they don't face any big risk.)]

After segmentation and repairment:

1st-part: 大手会社は直営の映画館も持っている。(Major companies own movie theaters under their direct management.)

2nd-part: 客の入らない映画館を維持するためには、高額の製作費がかかる自社作品を作るより、独立プロが製作した作品を有利な契約で上映する方が安上がりだ。(It is cheaper to show the movies produced by independent productions under profitable contract than to make in-house products with high production cost, to maintain the low-profit theaters.)

3rd-part: 大きなリスクを背負うこともない。(They don't face any big risk.)

(b) An example of sentence which sentence segmentation process IS NOT APPLIED:

[1 文化振興基金からの映画製作助成金の増額、製作資金の長期貸出制度の設立、スタッフ養成機関の設置、保証なしで作品を発表できる公設劇場の建設など、国の予算措置があれば、(If there exists budgetary support by the federal government, such as Cultural Society's fund increase in movie-making, formation of long-term loan for producers, establishment of training institutes, and constructions of public theaters to provide opportunities for film presentations,)[2 実現可能な振興策はいくつか考えられよう。(then several promoting measures are conceivable.)]

2.4 SVM classifier

Our system uses SVM_{light} (Ver.4.00) [5], which is an implementation of Vapnik's Support Vector Machine. There are some other choices in the machine learning frameworks that are known to be effective in the related summarization works, i.e.; probabilistic classifier based on Bayes' rules [9], decision tree C4.5 [4] [13] [14] [8], and perceptron [4].

Aside from SVM method, we were inspired by these works and performed preliminary testing using C4.5. Decision tree classifier proved effective in extracting most important and unimportant parts (10 to 20 % of the document) from a document. However, summarization rate in this method seemed uncontrolable, because we could not observe the difference among the summaries from three decision trees each trained with 10%, 30%, and 50% summaries respectivelly. Observing the trees beeing trained we found that the order of heuristic rules applied to the decision trees seemed indifferent and causing a confusion while classifying sentences with medium relevancy. Therefore, we used the SVM classifier here, which gives values of -1 to 1 as σ that is convenient to controll the summarization ratios of the outputs. The following attributes are used.

Attributes (16) $rhead$, (17) tf, and (18) $tfisf$, are assigned based on the following equations, where we represent a document as D, a unit of extraction (clause or sub-sentence) in the document as S, and a term as t respectively.

$$rhead(S) = \sum_{t_1 \in headline} \sum_{t_2 \in S} \delta_{t1,t2} / \sum_{t \in S} 1, \qquad (1)$$

here, δ is the Kronecher's delta. This $rhead(S)$ represents the rate of occurrence of headline words in a unit of extraction S.
$tf(S)$ is given as follows, where $f_D(t)$ is the number of times t appeared in the document D:

$$tf(S) = \sum_{t \in S} f_D(t) \qquad (2)$$

$tfisf(S)$ is given as follows:

Table 1. The set of attributes used for the SVM.

No.	Feature name	value	Definitions
1	top page	0 / 1	Genre of the article. The value is 1 for 1st, 2nd, and 3rd page of the articles.
2	general	0 / 1	Genre of the article. The value is 1 for general articles.
3	editorial	0 / 1	Genre of the article. The value is 1 for editorial articles.
4	nsent	Positive integer	The number of sentences in the article.
5	dloc	0 to 1 Real	Linear scale of sentence position in the article. 0 for the first, 1 for the last.
6	ploc	0 to 1 Real	Linear scale of sentence position in the paragraph, 0 for the first, 1 for the last.
7	Conj-SB	0 / 1	Conjunction at the beginning of the sentence. Value = 1 if present.
8	Demo-SB	0 / 1	Demonstrative pronoun at the beginning of the sentence. Value = 1 if present.
9	Mark-SB	0 / 1	Mark at the beginning of the sentence. Value = 1 if present.
10	Unpunct-SE	0 / 1	Value = 1 if no punctuation mark at the end.
11	Conj-NSB	0 / 1	Conjunction at the beginning of the following next sentence. Value = 1 if present.
12	Demo-NSB	0 / 1	Demonstrative pronoun at the beginning of the following next sentence. Value = 1 if present.
13	Mark-NSB	0 / 1	Mark at the beginning of the following next sentence. Value = 1 if present.
14	Last-NS	0 / 1	Value = 1 when the following next sentence is the last sentence in the article.
15	nchar	Positive integer	The number of characters in the sentence.
16	rhead	Positive real	Similarity with the headline. The value is given by the equation 1.
17	tf	Positive real	Normalized term frequency. The value is given by the equation 2.
18	tfisf	Positive real	Product of term frequency and inverted sentence frequency. The value is given by the equation 3.

$$tfisf(S) = \sum_{t \in S} f_D(t) \cdot isf(t), \qquad (3)$$

where,

$$isf(t) = \frac{log \sum_{S \in D} 1 - log \sum_{S \in D} \sum_{t_1 \in S} \delta_{t,t_1}}{log \sum_{S \in D} 1}. \qquad (4)$$

2.5 Summary generation

In summary generation process, the score for each unit of extraction, (a sentence or a clause obtained by segmentation process) is calculated and outputs higher ranked units as a summary output under constraint of the assigned summary rate. We use a combined score with the conventional score based on heuristics $Score_{conventional}(s)$and the output value σ of the SVM classifier. The following is used to calculate the conventional score;

$$Score_{conventional}(s) = tfisf(s) + rhead(s). \qquad (5)$$

This score is combined with the output value σ of SVM classifier into the score $Score(s)$;

$$Score(s) = Score_{conventional}(s) + \alpha \cdot \sigma(s) \qquad (6)$$

Here, α is a coupling constant. If we have multiple sets of training data in different summarization rates, r1, r2, and r3, we should obtain multiple classifiers by training from them independently. In this case, multiple outputs of classifiers, σ_{r1}, σ_{r2}, and σ_{r3} are combined into the score $Score(s)$:

$$\begin{aligned} Score(s) &= Score_{conventional}(s) \\ &+ \alpha \cdot (\sigma_{r1}(s) + \sigma_{r2}(s) + \sigma_{r3}(s)) \end{aligned} \qquad (7)$$

3 Evaluation

3.1 Submitted results

We implemented a summarization system as described in the previous section and submitted the results (at 20% and 40% summarization rate for 30 articles) of our system to TSC-2 task-A [3]. To carry out

Table 2. Ranking of summaries for each system.

	Content			Reliability		
	20%	40%	Ave.	20%	40%	Ave.
System 1	2.53	2.60	2.57	2.87	2.77	2.82
System 2	2.67	2.50	2.59	2.97	2.77	2.87
System 3	2.80	2.90	2.85	2.93	2.90	2.92
System 4	2.77	2.80	2.79	2.73	2.90	2.82
System 5	2.70	2.60	2.65	2.73	2.77	2.75
System 6	2.73	2.63	2.68	2.57	2.67	2.62
System 7	2.70	2.50	2.60	2.60	2.53	2.57
Our System	**2.40**	**2.60**	**2.50**	**2.83**	**2.77**	**2.80**
TF	3.30	3.20	3.25	3.30	3.10	3.20
Human	2.33	2.10	2.22	2.20	2.03	2.12

the training of SVM classifier, we utilized NTCIR-2 TSC (previous TSC) task-A test set collection of dry-run and formal-run. This is composed of 60 newspaper articles with answer summaries of each article at three summarization rates, 10%, 30%, and 50%. We trained SVM classifier on this data set and obtained three classifiers corresponding to each summarization rate. These classifiers are used in the sentence score calculation described in the equation (7).

3.2 TSC-2 evaluation results

The feasibility and the efficiency of our method are shown in the evaluation results given to participants groups by TSC-2 task organizer. Table 2 shows the results of ranking evaluation. In this evaluation, the following four types of summaries are compared and manually ranked.
(1) Extraction base summary constructed by human
(2) Free style summary by human (upper bound)
(3) Summary result by a summarization system (target)
(4) Summary result by lead method (baseline)

The ranking of summarization systems performed on each article is shown in the table 2. The ranking was based on:
Content-Important contents of the original article are covered in summary.
Readability-Summary remains to be readabile and meaningful for comprehension.

The results show that, our system achieved the best result among all the other systems with regard to contents, and performed closer to the human process especially at 20% summarization rate. The system didn't excel with regard to readability when compared to other top systems. Our summarization system as more advantage in extracting key contents while more work is necessary to improve the readability. Table 3 shows the correlation of rankings on content and readability for each article at 20% and 40% summarization rates.

Table 3. Correlation between the rankings according to content and readability.

		Content							
		20% summaries				40% summaries			
		1	2	3	4	1	2	3	4
Readability	1	2	0	0	0	4	0	0	0
	2	1	2	2	0	0	1	2	0
	3	3	0	14	0	1	1	15	0
	4	2	0	2	0	0	0	4	0

The results are pretty consistent for the content and the readability. Some summaries at 20% rate have value of 1 for the content while having different values from 1 to 4 for the readability. To further investigate the difference, we pick the two extreme summary samples: the article 990905036 with readability 4 shown in table 4, and the article 990618040 with readability 1 shown in table 5. Both have the content ranking of 1.

入学した旧制湘南中学は私にはいやな学校でね。(That old education system's Shonan Junior High, I hated it.) そんな信用できない俗物的な教師が多く、勉強ばかりの学校で大嫌いだった。(I hated it because it was all work, and because of these untrustworthy snobby teachers that I mentioned.) 終戦直後、湘南中の第 1 期生の大蔵官僚が学校に招かれ「これが我が校の出世頭だ」と紹介された。(After the war, this high-ranking bureaucrat was invited as one of the first graduate of the Shonan Junior High and introduced as "the most successful alumni".) 奥野先生は「人間は感覚が一番大事なのだ」と教えてくれた。(Professor Okuno taught us that the senses are most important to us humans.) 奥野先生は「極端な話、四角のものでも丸く見えたら丸を描け」と感覚の大切さを教えてくれた。(Professor Okuno also taught us the importance of the senses that we can even draw a rectangle a circle if it appeared to be round to our senses.) 先生に勝る師はいない。(No teacher is better than that professor.) 大学では「学校はいいもんだ」と思った。(In college, I felt the school isn't all that bad.)

Table 4. A summary result with 1 for contents and 4 for readability, generated from the article DOCID: 990605036.

By comparing the two summaries, we found that the summary generated with better readability (article

会期制が設けられたのは、多数党が数にたのんで国会日程を延ばしたりして、法案の成立をごり押しすることを避けるためだ。(The fixed session was introduced to avoid the majority party bulldozing the law establishment by extending the congressional sessions.) もっとも同じ国会法で、通常国会の場合は１回、特別国会と臨時国会の場合は２回まで、それぞれ延長できることが規定されている。(The same Diet law also permits the one time extension of regular session and two times for the special and the extraordinary sessions.) しかし、政府の言う通り、会期延長の主目的が、雇用創出や景気対策を早期に実施するためというのであれば、大方の国民も支持するのではないか。(But as the government official says, if the session extension's main purpose is job creation and to quickly implement the economic measures, most Japanese are likely to support the decision.) ただし、会期延長の狙いが、それだけでないこともまた、はっきりしている。(But it is clear that the purpose of the extension is not limited to these.) 延長国会で与野党が取り組むべき課題はまだまだ多い。(Politicians of the ruling and opposition parties have a long way to go in the coming extended session.)

Table 5. A summary result with 1 for contents and 1 for readability, generated from the article DOCID: 990618040.

990618040) is more readable, meaningful and naturally combined. The other summary with lower readability ranking (article 990605036) is not comprehensive enough and often disruptive, with demonstrative pronoun presented without its precedent. One drawback of our system is lack of mechanism to deal with the discourse structure of an article.

4 Conclusion

In this paper, we propose an automatic summarization method combining conventional sentence extraction and trainable classifier based on Support Vector Machine. To make extraction unit smaller than the original sentence extraction, we also introduce sentence segmentation process in our method. The feasibility and the efficiency of our method are shown in the evaluation results given to participants groups by TSC-2 task organizer. The results show that, our system achieves the best results among all others with regard to contents, even close to the human process (upper bound) at the summary rate of 20%. On the other hand, our system didn't excel in readability improvement compared to other top systems. More detailed assessment will be necessary to evaluate the efficiency of each procedure.

References

[1] S. Doi, K. Iwata, K. Muraki, and Y. Mitome. Pause control in japanese text-to-speech with lexical discourse grammar. In *Proceedings of 3rd International Conference on Spoken Language Processing (ICSLP 94)*, pages 743–746, 1994.

[2] H. P. Edmundson. New methods in automatic extracting. *ACM*, 16(2):264–285, 1969.

[3] T. Fukusima and M. Okumura. Text summarization challenge: Text summarization evaluation at ntcir workshop2. In *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization.*, pages 45–50, 2001.

[4] E. Hovy and C.-Y. Lin. Automated Text Summarization in SUMMARIST. Chapter 8 in: *Advances in Automatic Text Summarization*, I. Mani and M. T. Maybury (ed.), The MIT Press, 1999.

[5] T. Joachims. Making Large-Scale SVM Learning Practical. Chapter 11 in: *Advances in Automatic Text Summarization*, B. Schoelkopf and C. J. C. Burges and A. J. Smola (ed.), The MIT Press, 1999.

[6] S. Kamei and K. Muraki. Proposal of lexical discourse grammer. *IEICE Technical Report*, NLC86-7:1–5, 1986.

[7] K. Knight and D. Marcu. Statics-based summarization - step one: Sentence compression. In *Proceedings of the 17th National Conference on Artificial Intelligence*, pages 703–710, 2000.

[8] M. Kobori and N. Tamura. Automatic summarization based on the rhetorical structure of each paragraph and the order of importance between paragraphs. *IPSJ SIG Notes*, NL-136(11):79–86, 2000.

[9] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR Conference*, pages 68–73, 1995.

[10] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[11] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[12] D. Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, 2000.

[13] T. Nomoto and Y. Matsumoto. The reliability of human coding and effects on automatic abstructing. *IPSJ SIG Notes*, NL-120(11):71–76, 1997.

[14] M. Okumura, Y. Haraguchi, and H. Mochizuki. Some observations on automatic text summarization based on decision tree learning. *IPSJ 59th Annual Convention*, 2:393–394, 1999.

[15] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman Publishers, 1993.