A System for Text Summarization Based on Word Importance Measures

Hiroshi ISHII Rihua LIN Teiji FURUGORI
Department of Computer Science
The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan
{ishii-h,lin}@phaeton.cs.uec.ac.jp, furugori@cs.uec.ac.jp

Abstract

This paper presents a text summarization system based on word importance measures. It produces a summary through a part or all of the following four steps. First, we assign local and global scores to the nouns in each sentence according to the functional roles they play and the significancy they assume in the text for summarization. Second, based on the local and global scores, we calculate the importance value of each sentence. Third, the sentences with the importance values big enough for a summary are selected. Fourth and finally, a coherency test is given and we make the final decision for inclusion or exclusion of the selected sentences into the summary. We compare the summaries thus produced with the ones by humans and other systems.

Keywords: *summarization, sentence extraction, importance values, experiment, evaluation*

1 Introduction

The documents we see everyday are exceeding our capacity of reading them. There is an increasing need to have means to access the right information in a compressed form. Here lies a renewed interest in automatic text summarization [1, 2].

The sentences that constitute a summary can be either extractive or generative. The researchers have tried to extract important sentences through such clues as word frequencies, sentential relations, cue phrases, position of each sentence in a document, and information on word similarities [3, 4, 5, 6].

Summaries by sentence generation have faced problems far too difficult than those of the sentence extraction. Such a summary needs complex and profound analytical and synthetic processes that are cognitive as well as linguistic [e.g. 7, 8, 9].

This paper describes an extractive summarization system that selects sentences using a part or all of the following four steps.

First, we assign local and global scores to the nouns in each sentence according to the functional roles they play and the significancy they assume in the text for summarization. Second, based on the local and global scores, we calculate the importance value of each sentence. Third, we select the sentences with the importance values big enough for a summary. Fourth and finally, we give a coherency test and make the final decision for inclusion or exclusion of the selected sentences into the summary.

2 Word Roles and Significancy

The summaries by sentence extraction are crude ones. Their quality is far from the ones produced by human beings. Nevertheless, such summaries are useful and can be put in practice when considering the need of individuals to get short-handed information from the documents so wide spread in electronic forms, especially on the internet.

If the word x appears n times more than the word y, then x may be n times more important than y in a text. If a word occurs too often, however, it may lose its significance, indicating that such a word is simply functional to form sentences. Many methods that extract sentences have used this kind of information one way or another since the research in automatic summarization started nearly half a century ago [10].

2.1 Observations

The sentences to be extracted must be representing the content of a text in the best way. How can we achieve it? An answer to this question may be in extracting the sentences combining the functional roles the nouns or content words play in the sentence they appear and the significance of each word in text.

Consider the following sentences for the roles the nouns in each sentence play.

(1) The cat caught a mouse.
(猫がねずみを捕まえた。)

(2) A mouse was caught by the cat.
(ねずみが猫に捕まえられた。)

(3) I know the cat that caught a mouse.
(私はネズミを捕まえた猫を知っている。)

We observe that the focus is placed on *cat* in (1) and thus it is more important than *mouse*. This is not true in (2), however: *mouse* is more important than *cat* in this sentence. This is to say that the word in the subject position is more important than the word in the object position. In (3), both *mouse* and *cat* are in the object position. However, an emphasis is put on *cat* rather than on *mouse* in this case. This tells us that the word appearing in the main clause is more important than the word in the subordinate clauses.

It seems reasonable to assume from another observations that the words used repeatedly in text are more important than the words less frequently used. But then how can we count the words and measure the significancy of each word in text?

Words change their forms, but this is not really a problem in counting word frequencies. A big problem here is semantic. We use different words to express the same concept for rhetorical reasons. For instance, it is common for magazines to use a number of verbs to express the meaning in the verb to say: "the party *insists* {*maintains, argues, contends,* ⋯}", instead of "the party *says*."

We use a word with its topically similar words in a text, too. For instance, *doctor* and *nurse* are more likely to appear together than *doctor* and *highway* in medical texts. This means that the significancy of a word depends not only on its simple frequency count but also on its co-occurrence with its topically similar words.

2.2 Identicality and Similarity

Our summarization system relies on the functional roles of nouns in a sentence and the word significancy in text to extract sentences. We see the word identicality and similarity here, as the measurement of the word significancy, depends on them.

Identicality In Japanese one can make compound nouns freely by connecting simple nouns and, as a result, we often use their shortened or abbreviated forms from the beginning or from the second time on in a text. A typical way:

電気**通**信**大**学 (**Den**ki-**tsu**shin-**dai**gaku)
＝**電通大**(Dentsudai)

電通大 (Den-tsu-dai) is the initialism of 電気/通信/大学 (**Den**-ki/**tsu**-shin/**dai**-gaku). Thus, they are identical. One other typical way:

政治**改革**(political reform)
＝改革 (reform)

政治 (politics) and 改革 (reform) makes a compound noun 政治改革 (political reform). And 政治改革 can be identical to 改革 when they co-appear in a text.

Another way of identicality occurs when expressing a person's name. We write a person name in three different ways: full name, surname, or given name, all followed by an honorific suffix. If his or her occupation is important one, then we sometimes use the occupation or position name in the place of the honorific or the occupation, or position name alone to express the person. For instance,

鈴木太郎さん (Suzuki Taro san)
＝鈴木さん (Suzuki san)
＝太郎さん (Taro san)

細川護熙首相 (Hosokawa Morihiro shusho (Prime minister))
＝細川首相 (Hosokawa shusho)
＝首相 (shusho)

Similarity The head noun, the last element in a compound noun, and initialism are to be used for shortened expressions. So, for 政治改革, 政治 (not the head noun) can't be identical to 政治改革. However, one may say that 政治改革 and 政治 are at least closely related when they co-appear in a text. We thus say that a noun is similar to a compound noun if the latter contains the former as its part in certain ways.

3 Calculation of the Word Scores and Importance Value

We call the score a word gets from its functional role the local importance and the score from the word significancy the global importance. We calculate the importance value of each sentence from the local and global importance scores.

3.1 Local Importance Score

A syntactic analysis of each sentence is performed using the KNP analyzer [11] before the calculation of the importance values.

The local importance score of a word $LI(w)$ is defined as :

$$LI(w) = (the\ score\ given\ to\ w) + (the\ score\ given\ to\ the\ clause\ in\ which\ w\ appears)$$

The score given to *w* or the clause in which *w* appears is empirical. We consider that a noun used with

the topical case markers such as は and も is the most important. Then follows a noun used with the subject case marker が, a noun with object case marker を, and a noun with any other case marker. We consider also that the words appearing in the main clause are more important than the words in subordinate clauses.

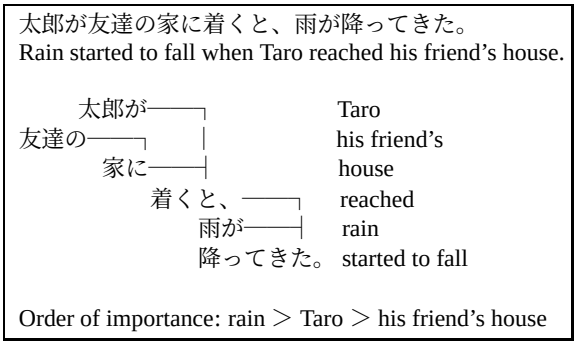

Figure 1. Syntactic analysis and importance scores

Figure 1 shows an output of the syntactic analysis and the order of word importance in a sentence. A noun phrase made by connecting nouns with the particle no(の) is considered to be a noun as a whole as is seen in the case of 友達の家 in Figure 1.

3.2 Global Importance Score

The global importance score of a word $GI(w)$ is calculated by:

$$GI(w) = \sum_{w'} (LI(w') \times Similarity(w, w'))$$

where w' is similar words of w in text.

Here, we decide the similarity between w and w' by:

$$Similarity(w, w') = \frac{CN}{TN}$$

where CN is the number of simple nouns common in the two (compound) nouns and TN the total number of simple words in the two (compound) nouns.

Obviously, the similarity is 1 when w and w' are identical. The similarity between 政治改革 (simple nouns 政治 and 改革) and 政治 is 2/3 and that of 日本映画 (Japanese movies; 日本 and 映画) and アメリカ映画 (American movies; アメリカ and 映画) is 2/4.

3.3 Importance Value of Sentence

The importance value of a sentence is calculated by the local importance and global importance scores. However, we do not calculate it simply by adding the importance values of words in the sentence. If we do this, a long sentence will get higher importance value as it contains more words in it. To avoid this partiality, we divide compound sentence or complex sentence into simple ones, calculate the importance values of simple sentences, and use the highest value among them as the importance value of the compound sentence.

The importance value, IV, of each sentence, s, is defined as:

$$IV(s) = \sum_{w \ in \ s} \sqrt{LI(w) \times GI(w)}$$

4 Summary Generation Processes

We select sentences that get higher importance values to form a summary. Presumably such a summary consists of the sentences containing the words that play central roles in the text. However, a summary by sentence extraction has an avoidable deficiency in coherency. So, in the final step, we try to produce sentences restrictively so that the summary may become coherent and more readable.

4.1 Coherent Relation

There are two main reasons that a summary by sentence extraction becomes incoherent. For a text to be coherent, sentences in it must have certain relations [12]. Consider the following texts.

(1) Taro went to school. He ate Sushi at a cafeteria. The Sushi was good. But the price wasn't right.

(2) Taro went to school. The Sushi was good.

(3) Taro went to school. But the price wasn't right.

It is apparent that (1) is coherent as a text. But (2) and (3) are not. The reason for the incoherency is that the second sentence in (2) appears all over a sudden as nothing related to it is mentioned in the preceding sentence(s). The reason for the incoherency in (3) is the same as for (2), but there is one more irregularity in this case: the conjunction *But* is out of place.

We are able to make up the deficiency in (2) if we choose the second sentence in (1), too, and add it to the summary. When we find no such a sentence, we do nothing more as the sentence selected is the one to introduce new information.

We try to make up the deficiency observed in the use of conjunction in (3) by choosing the sentence just before the sentence in question, too, or rejecting its inclusion into the summary when such a word or phrase in the sentence is one of what we call elaborating connectives (e.g., *consequently, for instance, in other words*, etc.).

1. Select the first sentence from the list of sentences;
2. If the sentence selected contains a connective word or phrase,
   then
     [if it is an elaborating connective,
       then [make the order of importance to be 0; {move the sentence to the end of list} go back to 1]
       else if the sentence just in front of this is already chosen for the summary,
         then include the sentence selected to the summary
         else lower the order of importance by 1; {move the sentence to the 2nd position in the list} go back to 1]
   else include it to the summary;
3. Remove from the list the sentence included into the summary;
4. Select the first sentences in the text that contain the words or concepts in the sentence just taken for the summary and add the sentences to the summary;
5. If the amount of sentences taken does not exceed the summary rate, then go back to 1.

Figure 2. Algorithm to choose sentences for a summary

4.2 Algorithm

Figure 2 is the algorithm to choose the sentences for a summary after making a list of all the sentences in the descending order of importance.

The main process is to repeat selecting and then removing the topmost sentence in the list and selecting its associated sentences until the summary rate is satisfied. Here, the term associated sentences means the first sentences that introduced the words or concepts appearing in the topmost sentence.

A diversion to the main process occurs when the topmost sentence contains a connective word or phrase. We try to exclude such a sentence as unimportant from the summary if the connective used is an elaborating one. Otherwise, to make the summary coherent, we include the sentence into the summary if the preceding sentence in the text is already taken in the summary or lower the order of importance by 1 as the sentence would be structurally dependent on others.

5 Experiment and Results

We tested our ideas by running an experiment. The test data used are provided by National Institute of Informatics(NII). They contain a set of 30 newspaper articles from the Mainichi Shimbun, a daily newspaper.

We set the importance score of each noun with its case marker to be in Table 1. We also set the importance score of a word in the main clause to be 3, in the subordinate clauses to be 2 or 1, depending on the distance from the main clause, and the summary rates(the number of sentences extracted / the number of sentences in the article) to be 10%, 30%, and 50%.

Table 2 shows the results(F-measure) from our experiment. Here, F-measure, Recall, and Precision are defined as:

F-measure = 2 * Recall * Precision / (Recall + Precision)

Table 1. Importance scores of nouns

Case markers	Scores
は or も	4
が	3
を	2
others	1

Table 2. Experimental results (1)

Methods	10%	30%	50%
Local importance	0.167	0.371	0.564
Global importance	0.194	0.433	0.603
Local and Global combined	0.199	0.399	0.589

Recall = (the number of correct sentences by the system) / (the total number of correct sentences by human summarizers)

Precision = (the number of correct sentences by the system) / (the total number of sentences by the system)

The numbers in Table 2 indicate the ratios of sentences extracted by the system and the ”correct sentences” extracted by human summarizers at NII.

Table 3 shows the results from other systems taken under the autopsy of the NII. Here, the result of the system number 8 is ours using the local and global importance combined. The sentences in Lead are taken from the beginning of each article until they satisfy the summary rate. The result of TF is based on term frequency counts alone.

We see from these results the following four points:

(1) No systems work nearly as good as humans do.

(2) Lead performs consistently better than TF.

Table 3. Experimental results (2)

System	10%	30%	50%
1	0.363	0.435	0.589
2	0.337	0.451	0.612
3	0.251	0.447	0.574
4	0.305	0.431	0.568
5	0.282	0.435	0.571
6	0.305	0.473	0.585
7	0.241	0.483	0.578
8	0.199	0.399	0.589
9	0.357	0.420	0.571
10	0.268	0.409	0.570
Lead	0.284	0.432	0.586
TF	0.276	0.367	0.530

(3) The differences in performance are small between the simple systems and others.

(4) Among the systems from 1 to 10, no system performs consistently better or worse than others.

How can we evaluate these facts in the summary production by sentence extraction? From (1), we may conclude that the summary making by sentence extraction has a limitation, however the methods we use.

From (2), it is obvious that the newspaper articles have a certain structure. It tells us that the systems that used cue phrases and/or positional information of the sentences would have performed better. In fact, we say that the bad result of our system, particularly when the summary rate is small, suffered from its methodological generality and the lack of giving consideration on purpose to this point.

The fact in (3) is a serious one. With (1), this indicates a limitation of the summary production by sentence extraction: is there any significance, for instance, between 0.432 by Lead and 0.483 by the system 7, or 0.586 by Lead and 0.612 by the system 2?

It may be from (4) that the real test of the systems is to be seem in applying them to various types of documents. We believe that our system in this respect is sound not only theoretically but also practically as it is independent of textual structures or contents and as it gives a consideration to the coherency of summaries to be produced.

6 Conclusion

We presented a text summarization system and showed some experimental results. Its performances are not very impressive, particularly for the newspaper articles. Nevertheless, we claim that it is usable in certain applications in this era for electric documentations, being the internet search of information a good example.

References

[1] Mani, I. and Maybury, M. eds. (1997): Intelligent scalable text summarization, Universidad Nacional de Educacion a Distancia, Madrid.

[2] Mani, I. and Maybury, M. eds. (1999): Advances in Automatic Text summarization, MIT Press, London.

[3] Zechner, K. (1996) : Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences, In *Proc. of the 16th International Conference on Computational Linguistics*, 986-989.

[4] Miike, S., Itoh, E., Ono, K., Sumita, K. (1994): A Full-Text Retrieval System with a Dynamic Abstract Generation Function, In *Proc. of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 152-161.

[5] Edmundson, H. (1969): New methods in automatic abstracting , *Journal of ACM*, 16 (2), 264-285.

[6] Barzilay, R. and Elhadad, M. (1997): Using lexical chains for text summarization, In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, 10-17.

[7] DeJong, G. F. (1982): An overview of the FRUMP system, In Lehnert W.G. and Ringle M.H. eds., *Strategies for Natural Language Processing*, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

[8] Jacobs, P. S. and Rau, L. F. (1990): SCISOR: extracting information from on-line news, In *Communications of the ACM*, 33 (11), 88-97.

[9] Hovy, E. (1988): Generating natural language under pragmatic constraints. Hillsdale, NJ: Erlbaum.

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

[11] Kurohashi, S. (1998): Nihongo Koubun Kaiseki System KNP Version 2.0b6 Shiyou Setumeisho, Kyoto University (in Japanese).

[12] Mann, W. C. and Thompson, S. A. (1988): Rhetorical structure theory: Toward a functional theory of text organization, *Text*, 8(3), 243-282.