Proceedings of NTCIR-4, Tokyo, April 2003 - June 2004

Multi-Document Summarization Using Document Set Type Classification

Jun'ichi Fukumoto
Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu-shi, Shiga 525-8577 JAPAN
fukumoto@media.ritsumei.ac.jp

Abstract

In this paper, we propose a summarization system which automatically classifies type of document set and summarizes a document set with its appropriate summarization mechanism. This system will classify a document set into three types: (a) One topic type, (b) multi-topic type, and (c) others. These types will be identified using information of high frequency nouns and Named Entity. In our multi-document summarization system, unnecessary parts are deleted after summarizing each document and then multi-document summary is generated. In type (a), unnecessary parts are similar part between summarized documents by single document summarization. In type (b), unnecessary parts are unsimilar parts in documents. In type (c), unnecessary parts are identified by scores used for single document summarization.

1 Introduction

We have participated Text Summarization Challenge 2 (TSC2) [3] of NTCIR Workshop 3 and had evaluation on single document summarization (Task A) and multi-document summarization (Task B)[1]. Our system applied very simple strategy to generate a summary, that is, TF/IDF based sentence extraction for single document summarization and use of single document summarization for multi-document summarization. In single document summarization, our system performance was a bit higher than the average but in multi-document summarization our system performance was in the lower group.

In our multi-document summarization, we use single document summarization for each document of a document set and removed similar parts between summarized documents for generation of a target summary. This mechanism is based on our assumption that all documents in a document set describe a series of events. The following documents describe additional information of the previous events with a citation of the previous events. That is, in the subsequent documents, some information is repeated one that has already mentioned in the previous documents. Therefore, deletion of similar description will be effective for document summarization of this kind of a document set. However, according to TSC2 evaluation, our assumption could be applied to some document sets but there were different types of multi-document sets which our approach could not. In the TSC-2 test set, there were many types of document sets. Some of them are described about a series of events, that is, the following document describes subsequent event or further information of its previous document. The others are about the same kind of events such as a set of new product information.

In this paper, we propose a summarization system which automatically classifies type of document set and summarizes a document set with its appropriate summarization mechanism. This system will classify a document set into three types: a series of events, a set of the same events and related events. In the first type, the second document shows additional information or subsequent event of the first document, and so on for the following documents. The second type documents describe the same event type such as a set of traffic accidents or new product announcements. The third type of documents is related each other but not classified into the first two types. We have submitted our system for TSC3 and have provisional results of evaluation until now.

2 Document set type

According to the analysis of TSC-2 multi-document test set, we have classified document set into three types as follows:

(a) One topic type

(b) Multi-topic type

(c) Others

In the first type (a), the first document described an event concerning a topic and the following documents describe the event of the first document again

and additional information or related matters of the first event. In this case, important elements of the first documents will be quoted in the following document. These important elements would include proper nouns such as person names, location names and company names because these series of documents describe one event which include these proper nouns.

In the second type (b), a set of documents described the same event type. For example, the first document is news of a new digital camera of some company and the next is also news of a new digital camera of the other company. Multi-document set of this type includes information on the same kinds of elements such new products but different instances. In this case, there are the same product types and organization type but they are different concrete name.

The third type of documents is related each other but not classified into the first two types. We assumed a set of documents collected for the purpose of summarization will be co-related in some aspect. In TSC2 test set of multi-document summarization, there are some classified in this type. We can guess this set of documents is collected with some key words.

3 Single document summarization

In multi-document summarization, each document of a document set will be summarized using TF/IDF based sentence extraction. In the following, we will first describe single document summarization and multi-document summarization method.

3.1 Sentence extraction

In our summarization system, sentence extraction from a document is based on TF/IDF, sentence position in a document and weighing using intention type of a sentence. Weights on sentences of a document are calculated using TF/IDF score of each noun of sentences. Also, information of sentence position such as the number of paragraph and the number of the sentence in a paragraph is used as sentence weight. At first, each sentence is morphologically analyzed using Chasen [4], and then TF/IDF values of nouns of a document are calculated. Weight value of a sentence W_i is obtained from the average number of TF/IDF values of all nouns of the sentence Wt_i. Weight of sentence position Wp_i is calculated using the number of paragraph $pnum$ and the number of sentence $snum$ as follow:

$$Wp_i = \frac{1}{4}(\frac{1}{pnum} + \frac{1}{snum})^2$$

For example, position weight Wp_i of the first sentence of the second paragraph will be 0.562.

In newspaper editorials, intention such as writer's opinion is expressed in a document and handling of such intention is important to generate a summary of this type of documents [2]. In our system, if a sentence has week intention type such as "inferential", its sentence score will have a half of the original score. If a sentence has strong intention type such as "request", "obligation" and "necessity", its sentence score will have 50% additional score of the original sentence score. In Japanese, such intention is expressed as tail expression of a sentence. Examples of tail expressions are presented in Table 1.

Table 1. Examples of Japanese tail expressions

weak	よう、という、そうだ、らしい、みられる、ようだ、といえる、だろう、かもしれない、
strong	たい、ほしい、べきだ、なければならない、大切である、必要だ

Weight W_i of a sentence is calculated in the following formula.

$$W_i = (Wt_i + Wp_i) * Wint_i$$

$$Wint_i = \begin{cases} 1.5 & \text{if sentence } i \text{ has strong intention} \\ 0.5 & \text{if sentence } i \text{ has weak intention} \\ 1.0 & \text{others} \end{cases}$$

3.2 Itemization of extracted sentences

In TF/IDF based sentence extraction, sentences in a summary are not coherently arranged, that is, sentences that have higher score are extracted independently. In order to make this summary more readable one, sentences in a summary are itemized by eliminating a part of tail expression of sentences.

In the current status, we are dealing with the following patterns.

- noun + である。
- noun + する。
- noun + した。

In the above pattern, underlined parts are eliminated, when their neighboring nouns satisfied some conditions, for example, POS code of noun is 17 or 31 ("sahen" verb type).

Moreover, conjunction and conjunctive phrases are also eliminated from extracted sentences. When sentences are itemized in a summary, such conjunctions are not necessary to relate sentences that are independently extracted. Conjunctions are used to express relationships to the previous or former sentences that might not be extracted in a summary.

3.3 Eliminating unnecessary parts of sentences

In order to condense more information into a limited size of summary, our system eliminates unnecessary parts from sentences. Some Japanese letters in parentheses are eliminated from a summary. They are Japanese reading of difficult Kanji letters or some additional information of the previous phrases.

3.4 Single document summarization

Text summarization of a single document is conducted in the following procedures.

1. TF/IDF scores of all sentences in a document are calculated using information of nouns in a document.
2. Sentence position in a paragraph and its paragraph position are used for calculation of sentence position score.
3. Weight on sentence intention is calculated based on tail expression of the sentence, and all sentence scores in a document are calculated.
4. Tail expression of sentences and top conjunctive expressions are eliminated using some patterns.
5. Unnecessary parts of sentences are removed.
6. Sentences are extracted from a document in the order of higher sentence weight until the sum of letters of extracted sentences will reach to the limited number of intended summary.
7. Extracted sentences are sorted in the original order in a document and then these sentences become a summary of the document.

4 Document set type classification

We will classify document set into three types: (a) one topic (b) multi-topic (c) others. Firstly, we will calculate document similarity between documents to identify document set (a). Then, we will identify document set type (b) for the document set not classified to type (a). To identify document set type (b), we will use Named Entity information of documents. Finally, we will identify document set (c) as document not classified into type (a) and (b).

We will extract an important sentence from a document to classify document set type. To extract an important sentence, we use information of nouns which appear in most of documents of a document set and Named Entity information using NExT system. Important sentence is recognized as a sentence which include such nouns and Named Entity elements.

In order to identify document set type, we will use information of Named Entity of important sentences of documents in a document set. Basically, if two sentences include the same Named Entity type and elements of such Named Entity, these two documents are recognized as type (a). If two sentences include the same Named Entity type but elements of such Named Entity are different, these two documents are recognized as type (b). The other case is type(c). Then, highest number of document set types in possible combination of two documents in a document set will be the type of document set.

5 Multi-document summarization

In our multi-document summarization system, the technique of the single document summarization is used for each document. After summarizing all documents, unnecessary parts in the summarized documents are deleted and then multi-document summary is generated.

Recognition of Unnecessary parts are different in each document set type. In type (a), unnecessary parts are similar part between summarized documents by single document summarization. In type (b), unnecessary parts are unsimilar parts in documents. In type (c), unnecessary parts are identified by scores used for single document summarization.

5.1 Summarization of each document

Input documents are co-related and subsequent documents are about subsequent events of the first document. In the subsequent documents, some information is repeated one that has already mentioned in the previous documents. Therefore, our approach to multi-document summarization is to summarize the first document in required summarization ratio and the following documents are summarized in higher ratio of summarization. In the current system, the following documents are summarized in the ratio of 10% more then the first one. For example, if the first document is summarized in 40% and then the followings are in 50%.

5.2 Deletion of unnecessary parts between documents

In order to detect unnecessary parts between documents, sentences in documents are segmented into clauses and similarity values between segmented clauses are calculated. Sentences are segmented by using information of Japanese comma and conjugation form of verb phrases. If the end of a clause (just before Japanese comma) is verb phrase and its conjugation type is “renyou” type, the sentence is segmented into two clauses at the point of the comma.

In order to calculate similarity values of clauses in a document, each clause in subsequent documents is compared with all the clauses in all the previous documents. During this comparison, the highest similarity values will be the similarity value of the clause. Similarity between clauses is calculated using information of the same nouns, adjectives and verbs in the clauses to detect similar parts between summarized documents. Similarity value sv_x of the clause x of document d_j is calculated in the following formula. The clause x have the highest similarity value with a clause in document d_i.

$$sv_x = \frac{\text{the number of shared words between } d_i \text{ and } d_j}{\text{the number of words in } d_j}$$

In type (a), unnecessary parts are clauses which have higher similarity value. Then, clauses are removed in the order of higher similarity value. In type (b), unnecessary parts are clauses which have lower similarity value. Then, clauses are removed in the order of lower similarity value. In type (c),

5.3 Summarization method

Text summarization of a multi-document is conducted in the following procedures.

1. The first document is summarized in the required summarization ratio.
2. The following documents are summarized in the required summarization ratio plus 10%.
3. All the sentences in all the summarized documents are segmented into clauses.
4. Similarity values between clauses are calculated and the highest similarity value will be the score of the clause.
5. Remove the clause which has the highest score until the sum of letters of extracted sentences will reach to the limited number of intended summary. If a clause has subordinate clause in the original sentence, such clause will not be remove.
6. The rest of the clauses are sorted in the original order in a document and then these clauses become a summary of the document.

6 Evaluation results

The results of Extract (short) A is shown in Table 2 and the results of Extract (long) B is shown in Table 3.

Most of document sets used for TSC3 were recognized as document set type (a) in our type classification. However, our method of document set type classification, four sets are recognized as type (a), twenty are type (b), and three are type (c). Our mechanism of document set classification does not work well. This is

Table 2. Results of Extract (short)

topic.length	coverage	precision
0310.6	0.000	0.000
0320.8	0.562	0.571
0340.10	0.350	1.000
0350.10	0.283	0.667
0360.7	0.119	0.750
0370.8	0.125	0.091
0380.13	0.269	0.429
0400.4	0.750	0.286
0410.13	0.064	0.333
0420.8	0.250	0.500
0440.7	0.286	0.667
0450.9	0.056	0.167
0460.5	0.000	0.000
0470.6	0.472	0.400
0480.8	0.125	0.200
0500.5	0.600	0.750
0510.9	0.333	0.400
0520.4	0.500	0.400
0530.12	0.208	0.556
0540.9	0.278	1.000
0550.12	0.153	0.429
0560.10	0.250	0.273
0570.9	0.222	0.750
0580.12	0.292	0.600
0590.7	0.214	0.750
0600.7	0.357	0.800
0610.5	0.000	0.000
0630.16	0.156	0.444
0640.11	0.076	0.667
0650.9	0.167	0.400
average:	0.251	0.476

Table 3. Results of Extract (long)

topic.length	coverage	precision
0310.11	0.273	0.400
0320.14	0.405	0.833
0340.17	0.294	0.818
0350.14	0.214	0.600
0360.16	0.120	0.778
0370.12	0.333	0.250
0380.20	0.300	0.462
0400.10	0.500	0.700
0410.15	0.144	0.400
0420.15	0.233	0.500
0440.11	0.318	0.714
0450.15	0.111	0.500
0460.11	0.136	0.375
0470.16	0.208	0.438
0480.15	0.133	0.400
0500.8	0.625	0.500
0510.16	0.281	0.462
0520.9	0.389	0.444
0530.22	0.273	0.588
0540.20	0.125	0.615
0550.24	0.139	0.583
0560.19	0.289	0.389
0570.16	0.365	0.727
0580.20	0.183	0.353
0590.14	0.143	1.000
0600.15	0.333	0.889
0610.12	0.000	0.000
0630.32	0.156	0.524
0640.21	0.111	0.700
0650.16	0.281	0.455
average:	0.247	0.547

because our current implementation has some system bugs in classification mechanism. So, it is necessary to evaluate our method after improving our system.

Although our classification method does not work well, the results of short and long extractions were not too bad. Also, correctly classified sets and not correctly classified sets have less difference in their results. In our summarization method, all documents in document set are preliminary summarized and unnecessary parts are removed from pre-summarized documents. The margin for removing unnecessary parts is only 10%, so it may be necessary to expand this pre-summarization ratio to higher one.

7 Conclusion

In this paper, we propose a summarization system which automatically classifies type of document set and summarizes a document set with its appropriate summarization mechanism. This system will classify a document set into three types: (a) One topic type, (b) multi-topic type, and (c) others. These types will be identified using information of high frequency nouns and Named Entity. In our multi-document summarization system, unnecessary parts are deleted after summarizing each document and then multi-document summary is generated. In type (a), unnecessary parts are similar part between summarized documents by single document summarization. In type (b), unnecessary parts are unsimilar parts in documents. In type (c), unnecessary parts are identified by scores used for single document summarization.

In the evaluation, our mechanism of document set classification does not work well. This is because our current implementation has some system bugs in classification mechanism. Although our classification method does not work well, the results of short and long extractions were not too bad. It is necessary to evaluate our mechanism after fixing system bugs and improvement of our system.

References

[1] J. Fukumoto. Text summarization based on itemized sentences and similar parts detection between documents. In *Working Notes of the Third NTCIR Workshop Meeting Part V: Text Summarization Challenge 2 (TSC2)*, pages 7–12, 2002.

[2] H. Watanabe. A method for abstracting newspaper articles by using surfece clues. In *Proc. of the 16th International Conference on Computational Linguistics*, pages 974–979, 1996.

[3] TSC http://oku-gw.pi.titech.ac.jp/tsc/.

[4] Chasen URL http://chasen.aist-nara.ac.jp/.