# Text summarization based on itemized sentences and similar parts detection between documents

Junichi FUKUMOTO
Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu-shi, Shiga 525-8577, Japan
fukumoto@cs.ritsumei.ac.jp

## Abstract

*In this paper, we propose a text summarization system for a single document and multiple documents. The system for a single one extracts sentences from a document and itemizes them to generate a summary. We applied this mechanism for Task A (single document summarization). We also utilized this mechanism for multi-document summarization (Task B) except for itemization mechanism. The system for multi-document detects similar parts between all summarized documents and eliminates them. In the Formal Run evaluation of TSC2, our system got better evaluation for a single document summarization but multi-document summarization was not so good.*
**Keywords:** *TF/IDF, intention, itemize, similarity detection.*

## 1 Introduction

It is our first participation to Text Summarization Challenge (TSC) [2] and we do not have enough research background in this area. Our system applied very simple strategy to generate a summary, that is, TF/IDF based sentence extraction for a single document summarization and use of single document summarization for multiple document summarization.

Sentences extraction technique is widely used for summary generation, however, sentences are extracted independently using some sort of term or sentence weighting method [3]. Our approach to generate a summary is to make independently extracted sentences readable one by itemizing the sentences.

In multiple text summarization, we applied single document summarization without itemizing mechanism and remove similar part from generated multiple summaries. There are many viewpoints to gather several documents such as the same kind of events, a series of events unrelated events and so on. In our assumption, input documents are related and following 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.

## 2 Single document summarization

### 2.1 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 [1], 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 [4]. 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.

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

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

## 2.2 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.

## 2.3 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.

## 2.4 Summarization method (single)

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 remove.
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.

Results of a short summary and long summary are shown in Figure 1 and Figure 2, respectively. Both are chosen from TSC2 Formal Run test set. In the both examples, itemized sentences are arranged in one line.

```
<SUM-RESULT>
<DOCNO>990624052</DOCNO>
<SUMLENGTH-C>250</SUMLENGTH-C>
<SUMTEXT>
成立に熱心な自由党とは対照的に、自民党には延長国会で全力をあげて成立させようという熱意が感じられない。
自民党は法案の成立はおろか、審議入りにさえ慎重だ。
理由は、自民党が連立のパートナーの自由党と、これから連立を組む公明党向けの二つの顔を見せたことにある。
公明党はこれに真っ向から反対。
定数削減問題は自自公連立へ進む大きなトゲであった。
民主党はこの点に着目して、定数削減案に賛成しようという動きを見せている。
自民党は、自由党と公明党の板挟みになったと言えるが、このようなことは連立政治の常であろう。
</SUMTEXT>
</SUM-RESULT>
```

**Figure 1. Example of a short summary (single document)**

```
<SUM-RESULT>
<DOCNO>990624052</DOCNO>
<SUMLENGTH-C>495</SUMLENGTH-C>
<SUMTEXT>
自民、自由両党が２３日、衆院比例代表の定数を５０減らす公選法改正案を国会に提出。
成立に熱心な自由党とは対照的に、自民党には延長国会で全力をあげて成立させようという熱意が感じられない。
自民党は法案の成立はおろか、審議入りにさえ慎重だ。
理由は、自民党が連立のパートナーの自由党と、これから連立を組む公明党向けの二つの顔を見せたことにある。
定数５０削減のための法案提出は、１月の自自連立政権発足にあたり、両党間で合意したことだ。
公明党はこれに真っ向から反対。
公明党との連携強化、さらに連立参加を意図してのことだった。
収まらないのは自由党。
自民党に約束通り法案の今国会提出を迫った。
定数削減問題は自自公連立へ進む大きなトゲであった。
民主党はこの点に着目して、定数削減案に賛成しようという動きを見せている。
連立政権は、歴史も理念・政策も異なる政党が共に政権を運営すること。
自民党は、自由党と公明党の板挟みになったと言えるが、このようなことは連立政治の常であろう。
政党にとって死活的な意味も持つ選挙制度の調整が難しいことは理解できる。
</SUMTEXT>
</SUM-RESULT>
```

**Figure 2. Example of a long summary (single document)**

## 3 Multi-document summarization

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

### 3.1 Summarization of each document

Single document summarization utilized for multi-document summarization is almost the same one as the one used for single document summarization mentioned in the previous section except for the mechanism of itemization of sentences.

There are many types of document sets for summarization, for example, a series of related documents, a set of same kinds of event information, non-related documents and so on. In our assumption, 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%.

### 3.2 Deletion of similar parts between documents

In order to detect similar 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 $d_i$.

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

### 3.3 Summarization method (multi)

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.

Results of a multi-document summary which includes short and long versions is shown in Figure 3.

## 4 Evaluation results

The results of Task A is shown in Table 2 and the results of Task B is shown in Table 3. “C” and “R” means *content-based evaluation* and *readability evaluation*, respectively.

As for Task A, the results of contents-based evaluation are not good but the results of readability evaluation are rather good, compared with the other systems. Our system needs more improvement on sentence extraction mechanism. However, we might conclude that itemization of extracted sentences works a little bit better because a summary consists of independently extracted sentences could be consider readable one. As for Task B, this mechanism does not work well. We need more improvement on the mechanism of sentence extraction from multiple documents.

## 5 Conclusion

In this paper, we propose a text summarization system that extracts sentences from a document and itemize them. We applied this system to single document

```
<TOPIC>
<TOPIC-ID>0030</TOPIC-ID>
<SUM-RESULT>
<SUMLENGTH-C>1000</SUMLENGTH-C>
<SUMTEXT>
２７日午後９時半ごろ、大阪市東淀川区豊里７の路上で、軽トラックの荷台の新聞紙とほろが炎上。２７日午後９時半ごろ、大阪市東淀川区豊里７の路上で、軽トラックの荷台の新聞紙とほろが炎上。２７日午後９時半ごろ、大阪市東淀川区豊里７の路上で、軽トラックの荷台の新聞紙とほろが炎上。４日午前２時ごろ、大阪府摂津市鶴野３の倉庫会社「日本トランスシティ」の倉庫から出火、庫内の携帯用ガスボンベが次々に破裂、炎上し、鉄骨スレートぶき平屋建て約７０００平方メートルを全焼した＝写真は午前４時半、三村政司写す。　【因幡健悦】　　大阪府の淀川以北の大阪市東淀川区、吹田市、摂津市を含む半径６キロ圏内で、今年に入って計１７０件の不審火が集中発生していることが１９日、分かった。そのほとんどがごみ置き場のごみや廃材、自転車などの「建造物等以外放火」だが、本格的な火事に発展したケースも。■写真説明　淀川以北で集中発生している不審火。大阪府摂津市鶴野で今月４日、延べ約７０００平方メートルの会社倉庫などが全焼した不審火で、大阪府警摂津署は２１日、出火場所を倉庫南東部分の資材置き場付近と特定。現場は倉庫などの事業所が建ち並び、夜間は街灯もなく人通りもほとんどないところ。今年に入り不審火が多発していた大阪府吹田市で今月に入りほぼ連日、二十数件にのぼる不審火が発生する異常事態となっている。捜査１課は、吹田市を含めた２区３市で起きた約１００件の連続放火を分析した結果、焼け跡から油成分がほとんど検出されておらず、簡易ライターなどで点火したと判断。</SUMTEXT>
</SUM-RESULT>
<SUM-RESULT>
<SUMLENGTH-C>500</SUMLENGTH-C>
<SUMTEXT>
２７日午後９時半ごろ、大阪市東淀川区豊里７の路上で、軽トラックの荷台の新聞紙とほろが炎上。今年に入って計１７０件の不審火が集中発生していることが１９日、分かった。そのほとんどがごみ置き場のごみや廃材、自転車などの「建造物等以外放火」だが、本格的な火事に発展したケースも。■写真説明　淀川以北で集中発生している不審火。大阪府摂津市鶴野で今月４日、延べ約７０００平方メートルの会社倉庫などが全焼した不審火で、大阪府警摂津署は２１日、出火場所を倉庫南東部分の資材置き場付近と特定。現場は倉庫などの事業所が建ち並び、夜間は街灯もなく人通りもほとんどないところ。今年に入り不審火が多発していた大阪府吹田市で今月に入りほぼ連日、二十数件にのぼる不審火が発生する異常事態となっている。捜査１課は、吹田市を含めた２区３市で起きた約１００件の連続放火を分析した結果、焼け跡から油成分がほとんど検出されておらず、簡易ライターなどで点火したと判断。</SUMTEXT>
</SUM-RESULT>
</TOPIC>
```

**Figure 3. Example of a multi-document summary**

**Table 2. Results of Task A (F0106)**

| Text ID | C 20% | R 20% | C 40% | R 40% |
|---|---|---|---|---|
| 990109032 | 3 | 2 | 3 | 3 |
| 990117039 | 1 | 1 | 1 | 1 |
| 990120043 | 3 | 3 | 3 | 3 |
| 990129047 | 3 | 3 | 3 | 3 |
| 990130032 | 3 | 3 | 3 | 3 |
| 990201036 | 3 | 3 | 3 | 3 |
| 990202041 | 3 | 3 | 3 | 3 |
| 990105044 | 3 | 3 | 3 | 3 |
| 990211049 | 3 | 3 | 2 | 3 |
| 990305053 | 2 | 2 | 2 | 2 |
| 990311036 | 4 | 3 | 2 | 2 |
| 990313042 | 1 | 1 | 1 | 1 |
| 990313046 | 3 | 3 | 3 | 3 |
| 990402040 | 4 | 3 | 3 | 3 |
| 990403032 | 3 | 3 | 3 | 3 |
| 990410033 | 3 | 3 | 3 | 3 |
| 990428029 | 3 | 3 | 2 | 2 |
| 990430039 | 3 | 2 | 3 | 2 |
| 990501040 | 3 | 3 | 3 | 3 |
| 990502043 | 3 | 3 | 3 | 3 |
| 990531030 | 3 | 3 | 3 | 3 |
| 990603040 | 3 | 3 | 1 | 1 |
| 990604040 | 3 | 3 | 3 | 3 |
| 990605036 | 3 | 3 | 3 | 3 |
| 990616038 | 1 | 1 | 2 | 2 |
| 990618040 | 3 | 3 | 3 | 3 |
| 990624050 | 2 | 2 | 3 | 3 |
| 990624052 | 3 | 1 | 3 | 2 |
| 990629039 | 2 | 3 | 3 | 4 |
| 990630039 | 2 | 2 | 3 | 4 |
| avg | 2.73 | 2.57 | 2.63 | 2.67 |

**Table 3. Results of Task B (F0206)**

| TopicID | C short | R short | C long | R long |
|---|---|---|---|---|
| 0010 | 4 | 2 | 3 | 3 |
| 0020 | 2 | 2 | 1 | 1 |
| 0030 | 2 | 1 | 1 | 1 |
| 0040 | 4 | 4 | 4 | 4 |
| 0050 | 4 | 4 | 4 | 4 |
| 0060 | 4 | 4 | 4 | 4 |
| 0070 | 4 | 4 | 4 | 4 |
| 0080 | 2 | 1 | 3 | 1 |
| 0090 | 3 | 2 | 3 | 2 |
| 0100 | 2 | 3 | 4 | 4 |
| 0110 | 4 | 3 | 3 | 3 |
| 0120 | 4 | 4 | 4 | 4 |
| 0130 | 1 | 2 | 2 | 4 |
| 0140 | 4 | 4 | 4 | 4 |
| 0150 | 3 | 1 | 4 | 4 |
| 0160 | 3 | 4 | 4 | 4 |
| 0170 | 2 | 1 | 3 | 2 |
| 0180 | 4 | 3 | 4 | 2 |
| 0190 | 4 | 4 | 4 | 4 |
| 0200 | 4 | 4 | 4 | 4 |
| 0210 | 2 | 2 | 4 | 4 |
| 0220 | 4 | 4 | 4 | 4 |
| 0230 | 4 | 4 | 4 | 4 |
| 0240 | 2 | 1 | 2 | 1 |
| 0250 | 3 | 3 | 3 | 3 |
| 0260 | 4 | 4 | 4 | 4 |
| 0270 | 4 | 4 | 4 | 4 |
| 0280 | 3 | 3 | 4 | 4 |
| 0290 | 2 | 4 | 4 | 4 |
| 0300 | 4 | 4 | 4 | 4 |
| avg | 3.20 | 3.00 | 3.47 | 3.30 |

summarization (Task A). We also applied this mechanism for multi-document summarization (Task B) except for itemization mechanism. This system detects similar part between summarized documents and eliminates them. In the Formal Run evaluation of TSC2, our system got better evaluation for a single document summarization on readability and not on content-based evaluation.

According to this evaluation, itemization of sentences in a summary works better but sentence extraction does not work better. It is necessary to improve sentence extraction for better selection of important sentences from a document and to develop more itemization patterns. As for multi-document summarization, the results were bad one. So, we have to analyze the results and might be better to improve the whole system architecture.

## References

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

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

[3] H. N. Manabu Okumura. Automated text summarization: A servey. *Journal of Natural Language Processing*, 6(6):1–26, 1999.

[4] 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.