Detection of Difference between News Articles on the Same Topic Based on Sequential Comparison

Tomoya NORO and Takehiro TOKUDA
Department of Computer Science, Tokyo Institute of Technology, Japan

Abstract. Currently, a lot of news articles are published on the Web, and it is getting easier for us to read them. However, the number of articles are too large for us to read all of them. Although some Web sites cluster/classify news articles into some topics (categories), it is not enough since a large number of articles are still in each topic. Detecting difference between articles on one topic will be one of the solution to comprehend the whole topic. In this paper, we propose a method for detection of difference between news articles on the same topic. Articles are sequentially compared by three different comparison units: paragraphs, sentences, and simple sentences. Our method is evaluated by applying it to Japanese news articles.

Keywords. difference detection, sequential comparison, multi-document text summarization, news articles, vector space model, dependency structure

Introduction

Currently, a large number of news articles are published on the Web. Since they cover broad range of topics on the earth, we can grasp global situation by reading them. However, the number of articles published on the Web is too large for us to read through. One solution to the problem is clustering or classifying articles into some topics (categories). Many Web sites which collect articles from many news sites and show them, such as Google News [1] and Yahoo! News [2], realize efficient presentation of articles in such a way.

However, we think it is not enough. For example, Google News U.S. collects articles from as many as 4,500 news sites, classifies them into some categories, and groups them by topics. Nevertheless, hundreds/thousands of articles are grouped into one topic and it is still difficult for us to comprehend the whole topic.

Even if two or more articles deal with the same topic, their contents cannot be completely the same except for the case that news source of the articles are the same (e.g. a news site gets an article from another news site and publishes it as it is). Detecting difference between articles on one topic will be a solution other than classification and clustering. Suppose, for example, several news sites published news articles about an accident that an airplane had been hit by an air turbulence. Basic information, such as date, time and place of the accident’s occurrence, the airline company name, and the number of persons injured, is described in all of the articles. However, not all of them include
other information, such as interviews with passengers, comments from the airline company, and information of other past accidents due to turbulence. Efficient presentation of such information will help us comprehend the whole topic.

In this paper, we propose a method for detecting difference between articles on the same topic (event). It could be considered as a kind of multi-document text summarization. However, general multi-document text summarization techniques detect part which appears in many of the given documents, that is, common part. Additionally, they compare documents by a single comparison unit (e.g. sentences, words, etc). Whereas we focus on detecting different part among documents, and our method detects different part efficiently by comparing articles by three different comparison units sequentially: paragraphs, sentences, and simple sentences. The paragraph-by-paragraph comparison and the sentence-by-sentence comparison are based on word occurrence, while the comparison by simple sentences is based on dependency structure as well as word occurrence. Although basic idea of our method is language-independent, we incorporate some Japanese-specific features for improvement. We explain our method taking English examples as far as possible, but we sometimes take Japanese examples.

The organization of the rest of this paper is as follows. Overview of our method is presented in section 1, then details of each step of the method is described in section 2, 3, 4, 5 and 6. Evaluation result is shown in section 7. We discuss related work in section 8, and conclude this paper in section 9.

1. Overview of Our Method

We assume that news articles have already been grouped by topics in some way. Our method takes two news articles, and detects parts common to both of the articles (common part) and parts included in either of the articles (different part). Then, our method calculates significance of each part of the different parts, and outputs the parts with high significance score (Figure 1).

In order to detect common part and different part from two articles, we segment the articles into small units and compare them. We consider the following three different comparison unit.
Paragraph-based comparison: News articles are segmented into paragraphs. For each paragraph of one article is compared with each paragraph of the other article based on word occurrence.

Sentence-based comparison: If an article has only one paragraph, the article cannot be separated into common part and different part by the paragraph-by-paragraph comparison. Additionally, even if two paragraphs are similar to each other as a whole, there may be small difference between them. Since the paragraph-by-paragraph comparison cannot detect such difference, we consider sentence-by-sentence comparison based on word occurrence.

Simple sentence-based comparison: In some cases, what is described in one complex/compound sentence of an article is described in two or more simple sentences of the other article. In the simple sentence-based comparison, we segment each article into simple sentences, and compare two simple sentences based on dependency structure as well as word occurrence.

Since comparison by smaller unit (especially the simple sentence-based comparison) is susceptible to pronouns and ellipsis, anaphora resolution will be needed. However, it takes much calculation cost. Additionally, calculation cost for simple sentence-based comparison is high since a large number of simple sentences have to be compared one another. To avoid this problem, our method applies the three kinds of comparison sequentially as follows (Figure 2).

1. Analyze news articles by morphological/dependency analyzers.
2. Compare the two articles by paragraphs.
3. If similarity score of two paragraphs $\text{Sim}_{\text{para}}$ is higher than a predetermined threshold $t_{\text{para}}$, compare the two paragraphs by sentences.
4. If similarity score of two sentences $\text{Sim}_{\text{sent}}$ is higher than a predetermined threshold $t_{\text{sent}}$, compare the two sentences by simple sentences.
5. If similarity score of two simple sentences $\text{Sim}_{\text{dep}}$ is higher than a predetermined threshold $t_{\text{dep}}$, we judge that the two simple sentences are similar to each other.

6. Filter out some of the pairs so that any unit (simple sentence) may belong to at most one pair.

Details of these steps are described in the following sections.

2. Morphological Analysis and Dependency Analysis

We use two well-known Japanese dependency analyzers, KNP [3] and CaboCha [4]. Although both of them analyze dependency relations between two Japanese phrasal units, called “bunsetsu”, they have different features. KNP, using a Japanese morphological analyzer JUMAN [5] as preprocessing, distinguishes conjunctive structure (parallel relation and appositional relation) from others [6]. It also outputs information about features assigned to each bunsetsu, such as whether it is a predicate or not, surface case information, etc. These information will be useful for our method.

Although CaboCha, including either of two Japanese morphological analyzers ChaSen [7] or MeCab [8], just determines which two bunsetsu are in dependency relation and does not distinguish any types of relation from others [2], it includes named entity recognition module. It recognizes eight classes of named entity defined by IREX [9]. This information will also be useful.

It is said that overall accuracy of CaboCha is generally higher than that of KNP. However, as far as conjunctive structure goes, the accuracy of KNP is higher than that of CaboCha. Fortunately, analyzing process of CaboCha consists of four layers, morphological analysis (tagging), bunsetsu chunking, feature selection, and dependency analysis. We can set an input layer and an output layer. We can also determine some dependency relations after the bunsetsu chunking layer and CaboCha can proceed analysis under the restriction.

In our method, we analyze news articles by KNP and CaboCha as follows.

1. Analyze news articles by KNP, and extract conjunctive relations and some feature information assigned to each bunsetsu.
2. Proceed analysis of the same articles by CaboCha to the bunsetsu chunking layer (We use MeCab for morphological analysis).
3. Give the conjunctive relations extracted in the first step.
4. Continue analysis to the last layer by CaboCha under the restriction.
5. Add the feature information extracted in the first step.

Note that “phrase” indicates bunsetsu in the rest of this paper.

---

1Simply speaking, a bunsetsu consists of a (compound) content word and a (compound) functional word following the content word (In Japanese, a functional word follows a content word). If this definition is applied to English, “The government unveils a plan to support foreign permanent residents” will be divided into 5 bunsetsu: “The government”, “unveils”, “a plan”, “to support”, and “foreign permanent residents”.

2Actually, CaboCha has an option for distinguishing parallel relation and appositional relation from others. However, the accuracy is lower than KNP.

3“ARTIFACT” is added to the seven classes of named entity defined by MUC [10]
3. Comparison by Paragraphs/Sentences

Since the basic idea of the paragraph-based comparison is the same as the sentence-based comparison, we describe only the paragraph-based comparison in this section.

The paragraph-based comparison is based on vector space model: words are extracted from each paragraph and similarity score between the two paragraphs is calculated using the two word lists. A word $w$ (base form) is extracted from a paragraph $P$ if $w$ is one of noun, verb, adjective, number (numeral), and named entity $^4$. In Japanese, a number is usually followed by a numeral classifier. We concatenate a number with the following word if the word is a noun or a suffix. A word list $WL$ extracted from a paragraph $P$ is defined as follows.

$$WL(P) = \{(w, \text{pos}(w), \text{numtype}(w))|\text{tf}(w, P) > 0\}$$

$$\text{numtype}(w) = \begin{cases} \text{pos}(w) & \text{if } \text{pos}(w) \text{ is "DATE" or "TIME"}, \\ w' & \text{else if } w \text{ is a number concatenated with a noun/suffix } w' \\ \text{NULL} & \text{otherwise} \end{cases}$$

$\text{tf}(w, P)$ indicates frequency of a word $w$ appearing in a paragraph $P$. $\text{pos}(w)$ indicates part-of-speech of a word $w$ (i.e. noun, verb, adjective, or number). If $w$ is a named entity, $\text{pos}(w)$ is the named entity class assigned to $w$. Why $\text{numtype}(w)$ is necessary is described later. In the case that $P$ is “Toyota said it expects an operating loss of 150 billion yen for the fiscal year ending in March.”, an extracted word list is as follows.

$$WL(P) = \{(\text{Toyota}, \text{ORGANIZATION}, \text{NULL}), (\text{March}, \text{DATE}, \text{DATE}), (150 \text{ billion yen}, \text{number}, \text{yen}), (\text{loss}, \text{noun}, \text{NULL}), (\text{year}, \text{noun}, \text{NULL}), (\text{say}, \text{verb}, \text{NULL}), (\text{expect}, \text{verb}, \text{NULL}), (\text{operate}, \text{verb}, \text{NULL}), (\text{end}, \text{verb}, \text{NULL}), (\text{fiscal}, \text{adjective}, \text{NULL})\}$$

Assuming that $|WL(P_1)| \leq |WL(P_2)|$, similarity score between two paragraph $P_1$ and $P_2$ is calculated as follows.

$$\text{Sim}_{\text{para}}(P_1, P_2) = \max_{f:WL(P_1)\rightarrow WL(P_2)} \sum_{(w, p, t) \in WL(P_1)} g((w, p, t), f(w, p, t), P_1, P_2) / |WL(P_1)| \cdot |WL(P_2)|$$

where $f$ is an injective mapping. Assuming that $(w_1, p_1, t_1) \in P_1$ and $(w_2, p_2, t_2) \in P_2$,

$^4$A named entity is treated as one word even if it consists of two or more words.
\[ g((w_1, p_1, t_1), (w_2, p_2, t_2), P_1, P_2) = \ \begin{cases} 
\text{weight}((w_1, p_1, t_1), P_1) \cdot \text{weight}((w_2, p_2, t_2), P_2) \\
\text{if } (w_1, p_1, t_1) = (w_2, p_2, t_2) \\
C_{\text{NUM}} \cdot \text{weight}((w_1, p_1, t_1), P_1) \cdot \text{weight}((w_2, p_2, t_2), P_2) \\
\text{else if } t_1 = t_2 \neq \text{NULL} \land (w_1, p_1, t_1) \notin \text{WL}(P_2) \\
\land (w_2, p_2, t_2) \notin \text{WL}(P_1) \\
0 \quad \text{otherwise} 
\end{cases} \]

\[ |\text{WL}(P)| = \sqrt{\sum_{(w, p, t) \in \text{WL}(P)} \text{weight}((w, p, t), P)^2} \]

\[ \text{weight}((w, p, t), P) = \ \begin{cases} 
0 \quad \text{if } (w, p, t) \notin \text{WL}(P) \\
C_{\text{NE}} \quad \text{else if } p \text{ is one of the named entity classes} \\
1 \quad \text{otherwise} 
\end{cases} \]

where \( C_{\text{NUM}} \) and \( C_{\text{NE}} \) are coefficients (\( 0 \leq C_{\text{NUM}} \leq 1 \) and \( C_{\text{NE}} \geq 1 \)). Numeric values, date and time sometimes differ among articles even if they deal with the same topic. For example, one article on an terrorism attack says that a bomb exploded at 10:31 am and 9 people were killed, while another article on the same attack says that it occurred at around 10:30 am and about 10 people were killed. Considering such situation, in calculation of \( g((w_1, p_1, t_1), (w_2, p_2, t_2), P_1, P_2) \), we set difference in numeric values, date and time to aside, and only check their type (numtype) \( t \).

4. Comparison by simple sentences

Although the simple sentence-based comparison is also based on vector space model, it is different from the paragraph/sentence-based comparison: not only word occurrence in each phrase but also dependency structure of each simple sentence is considered in order to compare two simple sentences in more detail.

We represent dependency structure of a simple sentence \( D \) as 2-tuple \((V, M)\), where \( V \) is a phrase including a predicate of the simple sentence and \( M \) is a set of phrases which are in dependency relation with \( V \). Assuming that \(|M_A| \leq |M_B|\), similarity score between two dependency structure \( D_A = (V_A, M_A) \) and \( D_B = (V_B, M_B) \) is calculated as follows.

\[
\text{Sim}_{\text{dep}}(D_A, D_B) = \begin{cases} 
\text{Sim}_V(V_A, V_B) \cdot \text{Sim}_M(M_A, M_B) & \text{if } n \neq 0 \\
0 & \text{otherwise} 
\end{cases}
\]

\[
n = |\{m \in M_A | \text{Sim}_m(m, g(m)) \neq 0\}| 
\]

Similarity score between \( V_A \) and \( V_B \) is calculated as follows.

\[
\text{Sim}_V(V_A, V_B) = C_V + (1 - C_V) \cdot \text{Sim}_{\text{phrase}}(V_A, V_B) 
\]
where $C_V$ is a coefficient and $0 < C_V < 1$. Calculation of $\text{Sim}_\text{phrase}$ is about the same as $\text{Sim}_\text{para}$ and $\text{Sim}_\text{sent}$. The difference is described later. Similarity score between $M_A$ and $M_B$ is calculated as follows.

$\text{Sim}_M(M_A, M_B) = \begin{cases} \sqrt{\frac{\text{Score}(M_A, M_B)}{|M_A|} \frac{\text{Score}(M_A, M_B)}{|M_B|}} & \text{if } |M_A| \neq 0 \text{ and } |M_B| \neq 0 \\ 0 & \text{otherwise} \end{cases}$

$\text{Score}(M_A, M_B) = \max_{h:M_A \rightarrow M_B} \sum_{m \in M_A} \text{Sim}_m(m, g(m))$

where $h$ is an injective mapping.

$\text{Sim}_m(m_1, m_2) = c(m_1, m_2) \cdot \text{Sim}_\text{phrase}(m_1, m_2)$

$c(m_1, m_2) = \begin{cases} 1 & \text{if both/either deptype}(m_1) \text{ and/or deptype}(m_2) \text{ are/is NULL} \\ \text{or deptype}(m_1) = \text{deptype}(m_2) & \text{otherwise} \end{cases}$

where deptype($m$) is dependency type of phrase $m$, and $C_m$ is a coefficient ($0 < C_m < 1$). Dependency type is determined by (a sequence of) postposition(s) which the phrase ends with. If no postposition appears, dependency type of the phrase is “NULL”. In Japanese, ellipsis of postpositions sometimes occurs. If dependency type of a phrase is NULL, $c(m_1, m_2)$ is 1 regardless of dependency type of the other phrase.

An example for simple sentence-based comparison is shown in Figure 3. $\text{Sim}_\text{V}(V_A, V_B)$ is 1. A table at the bottom of Figure 3 shows $\text{Sim}_m$ of two phrases (let $C_m$ be 0.5). Note that $\text{Sim}_m(P_{A2}, P_{B2})$ is 1 since both numtype (“gozen 9 ji 15 fun”) and numtype (“gozen 9 ji 17 fun”) are “TIME”, and $\text{Sim}_m(P_{A3}, P_{B4})$ is 0.5 since dependency type of $P_{A3}$ (“ni”) is different from that of $P_{B4}$ (“he”). From this table, the injective mapping $g$ which maximizes $\sum_{m \in M_A} \text{Sim}_m(m, g(m))$ is $\{(P_{A1}, P_{B1}), (P_{A2}, P_{B2}), (P_{A3}, P_{B4})\}$. Score($M_A, M_B$) is 2.5 and $n$ is 3. Hence, $\text{Sim}_M$ and $\text{Sim}_\text{dep}$ are as follows.

$\text{Sim}_M(M_A, M_B) = \frac{2.5}{3} \times \frac{2.5}{4} = 0.722$

$\text{Sim}_\text{dep}(D_A, D_B) = 1^4 \times 0.722^2 = 0.783$

Although we mentioned above that calculation of $\text{Sim}_\text{phrase}$ is about the same as $\text{Sim}_\text{para}$, there is difference. Basically, for each phrase $P$ in a set of phrases $M$, we create a word list $WL(P)$ in the way described in section 3. However, in some cases, a phrase has some phrases which are in dependency relation. An example is shown in Figure 4 (A dashed arrow indicates conjunctive relation). In this example, phrase $P_1$ and phrase $P_2$ are in conjunctive relation, and phrase $P_3$ modifies phrase $P_1$. In our basic method of

---

5 Simply speaking, surface case of each phrase will be its dependency type.

6 Type of dependency relation (whether conjunctive relation or not) is determined by KNP as described in section 2.
Douki ha 9 ji 17 fun ni Kansai kuukou ni chakuriku shita
(The airplane landed at Kansai International Airport at 9:17)

Douki ha gozen 9 ji 15 fun goro, yotei jikoku doori ni Kansai kuukou he chakuriku shita
(The airplane landed at Kansai International Airport on time at around 9:15 am)

### Figure 3. An example for simple sentence-based comparison

<table>
<thead>
<tr>
<th></th>
<th>$P_{B1}$</th>
<th>$P_{B2}$</th>
<th>$P_{B3}$</th>
<th>$P_{B4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{A1}$</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$P_{A2}$</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$P_{A3}$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.500</td>
</tr>
</tbody>
</table>

word list creation, we extract words only from $P_2$ and $P_4$ since $P_1$ and $P_3$ do not modify $V$ directly. In order to extract words from $P_1$ and $P_3$, we extend the method of word list creation as follows.  

$$W_{L-phrase}(P) = W_{L}(P) \cup \bigcup_{P' \in MP} W_{L-phrase}(P')$$

$$weight_{phrase}((w, p, t), P) = C_{mod}^{N(w, p, t, P)} \cdot weight((w, p, t), P)$$

where $MP$ is a set of phrases which are in dependency relation with $P$, and $N(w, p, t, P)$ is dependency distance (the number of steps of dependency relation except conjunctive relation) between the phrase including the word $w$ and $P$. If the word $w$ appears more than once, we take only the nearest one to $P$. $C_{mod}$ is a coefficient and $0 < C_{mod} < 1$. It means that word weight gets lower with increasing dependency distance between the two phrases, while word weight does not get lower if the two phrases are in conjunctive relation. In the example of Figure 4, $W_{L-phrase}(P_2)$ and $W_{L-phrase}(P_4)$ are as follows.

$$W_{L-phrase}(P_2) = \{(jouin, noun, NULL), (3 nin, number, nin), (joukyaku, noun, NULL), (7 nin, number, nin)\}$$

*In definition of $weight((w, p, t), P)$ in section 3, $W_{L}(P)$ should be replaced with $W_{L-phrase}(P)$ described here.*
Joukyaku 7 nin to jouin 3 nin ga kuukou kuukou no byouin ni hakobareta
(7 passengers and 3 crews were transported to a hospital near the airport)

Figure 4. An example for dependency structure of a simple sentence

\[
W_{\text{phrase}}(P_4) = \{(\text{byouin}, \text{noun}, \text{NULL}), (\text{kuukou}, \text{noun}, \text{NULL}),

(\text{fukin}, \text{noun}, \text{NULL})\}
\]

Weight of each word is as follows.

\[
\begin{align*}
\text{weight}_{\text{phrase}}((\text{jouin}, \text{noun}, \text{NULL}), P_2) &= \text{weight}_{\text{phrase}}((3 \text{ nin}, \text{number}, \text{nin}), P_2) \\
&= \text{weight}_{\text{phrase}}((\text{jyoukyaku}, \text{noun}, \text{NULL}), P_2) \\
&= \text{weight}_{\text{phrase}}((7 \text{ nin}, \text{number}, \text{nin}), P_2) = 1 \\
\text{weight}_{\text{phrase}}((\text{byouin}, \text{noun}, \text{NULL}), P_3) &= 1 \\
\text{weight}_{\text{phrase}}((\text{kuukou}, \text{noun}, \text{NULL}), P_4) &= \text{weight}_{\text{phrase}}((\text{fukin}, \text{noun}, \text{NULL}), P_4) = C_{\text{mod}}
\end{align*}
\]

Since \(P_1 \) and \(P_2\) are in conjunctive relation, dependency distance between the two phrases is 0.

5. Filtering

Common part of articles on the same topic can be extracted by carrying out three comparison methods described above. However, some units each might belong to two or more pairs. An example is shown in Figure 5. In this case, for example, a unit \(U_{A1}\) (simple sentence) in article A is judged to be similar to two units \(U_{B1}\) and \(U_{B7}\) in article B. To solve this problem, we filter out some of the pairs so that any unit may belong to at most one pair as follows.
1. Sort the pairs in descending order of their score $Sim_{dep}$.
2. Let $L$ be an empty set.
3. For each pair $(U_A, U_B)$ in the sorted list, if neither $U_A$ nor $U_B$ is in $L$, keep the pair and add the two units to $L$. Otherwise, discard the pair.

In the case of Figure 5, $(U_{A1}, U_{B7})$ and $(U_{A7}, U_{B1})$ are discarded since $(U_{A1}, U_{B1})$ has already been kept.

**6. Extraction of Significant Part from the Difference**

The units which are not judged to be common part are determined to be different part. However, some units might not be significant as difference between the articles. In this section, we describe our method for picking up significant part from the difference between articles.

The basic idea is that a unit is significant if the part includes some words with high IDF score which do not appear in the common part of the articles. We extract significant part from the difference between articles as follows.

1. Let $CP$ be the common part of the articles (a set of simple sentences in the common part), and Let $DP$ be the different part between the articles (a set of simple sentences in the different part).
2. Create a list of words appearing in the common part $CP$ as follows.

$$WL_{\text{common}} = \bigcup_{D'=(V', M') \in CP} WL_{\text{dep}}(D')$$

$$WL_{\text{dep}}(D') = WL(V') \cup \bigcup_{m' \in M'} WL(m')$$

3. For each $Diff$ in $DP$,
   (a) Create a word list $WL_{\text{dep}}(Diff)$. 

---

**Figure 5.** An example for filtering out some pairs

- Article A
- Article B
- Sorted list by score

- $(U_{A2}, U_{B1}, 1.00)$
- $(U_{A4}, U_{B3}, 1.00)$
- $(U_{A3}, U_{B5}, 1.00)$
- $(U_{A1}, U_{B1}, 0.75)$
- $(U_{A7}, U_{B1}, 0.75)$
- $(U_{A7}, U_{B7}, 0.64)$
- $(U_{A5}, U_{B5}, 0.64)$
- $(U_{A1}, U_{B7}, 0.64)$
- $(U_{A6}, U_{B6}, 0.49)$

- $(U_{A1}, U_{B1}, 0.64)$
- $(U_{A7}, U_{B1}, 0.64)$
- $(U_{A5}, U_{B5}, 0.64)$
- $(U_{A1}, U_{B7}, 0.59)$
- $(U_{A6}, U_{B6}, 0.49)$
(b) Calculate IDF score for each word in \( WL_{dep}(\text{Diff}) - WL_{\text{common}} \), and obtain the top \( N_{\text{top}} \) words (\( N_{\text{top}} > 0 \)). If \(|WL_{dep}(\text{Diff}) - WL_{\text{common}}|\) is less than \( N_{\text{top}} \), obtain all of the words. Let a set of the \( N_{\text{top}} \) words be \( WL_{\text{top}} \).

(c) \( \text{Sig}(\text{Diff}) = \sum_{w \in WL_{\text{top}}} \text{IDF}(w) \)

(d) Let \( P \) be the paragraph including \( \text{Diff} \). If \( \text{Sim}_{\text{para}}(P, P') \) is less than a predetermined threshold \( t_{\text{diff}} \) for any paragraph \( P' \) in the other article, multiply \( \text{Sig}(\text{Diff}) \) by a coefficient \( C_{\text{sig}} \) (i.e. \( \text{Sig}(\text{Diff}) = C_{\text{sig}} \cdot \text{Sig}(\text{Diff}) \)).

(e) If \( \text{Sig}(\text{Diff}) \) is greater than a predetermined threshold \( t_{\text{sig}} \), \( \text{Diff} \) is extracted as the significant part.

If a paragraph in one article is not similar to any paragraphs in the other article, units in the paragraph may be significant. The step 3(d) reflects this intuition.

Although a large number of articles are necessary to calculate IDF, we do not have enough articles at this point. Instead, we use search result of Yahoo! Japan News [11]:

\[
\text{IDF}(w) = \log \frac{N}{\text{Result}(w)}
\]

where \( N \) is the total number of articles stored in the database of Yahoo! Japan News, and \( \text{Result}(w) \) is the number of search results by a word \( w \). Although we do not know the exact value of \( N \), we let \( N \) be 80,000.\(^{8}\)

7. Evaluation

In order to evaluate our method, we collected Japanese news articles on several topics as listed in Table 1. For each of the topics, two articles were collected (let the two articles be article A and B respectively). The number of paragraphs, sentences, and simple sentences of each article are also shown in the table. Thresholds and coefficients described previously are set as shown in Table 2.

The number of units judged as common part and the number of units judged as significant different part of each article are shown in Table 3 (the second, third, and fourth columns). We manually determined common part and compared with the result. The fifth column indicates the number of units of the common part determined manually, and the sixth column indicates the number of units judged correctly. Precision and recall are shown in the seventh and eighth columns.

From the result, precision of common part recognition is high, while recall is low. Some causes of the error are as follows.

- Synonyms are judged as different words. For example, a verb “hakobu” is a synonym for “hansou suru” (transport). In our method, if a verb of one unit (i.e. a simple sentence) is different from a verb of the other unit, similarity score between the two units tends to be low. Some language resources such as thesauri are necessary to solve this problem.

---

\(^{8}\)Yahoo Japan! News keeps news articles for at most 90 days, and the total number of articles in the database will change day-by-day. In our experience, we guess the total number of articles is between 75,000 and 80,000.
**Table 1.** Topics of collected news articles

<table>
<thead>
<tr>
<th>Topic</th>
<th>Paragraph</th>
<th>Sentence</th>
<th>Simple</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>Topic 1</td>
<td>An airplane accident due to a turbulence</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Topic 2</td>
<td>Shoplifting by more than 20 high school students on a school trip</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Topic 3</td>
<td>Arrest of a leader of a group of loan sharks on extortion</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Topic 4</td>
<td>Files on a police officer were sent to prosecutor as hit-and-run accident in drink-driving</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Topic 5</td>
<td>An arson fire and burn death of two children</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Topic 6</td>
<td>Arrest of a university student on indecent assault</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Topic 7</td>
<td>An announcement about a plan to prohibit students from bringing mobile phones to school by the education board of Osaka</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Topic 8</td>
<td>A doctor was acquitted of negligent homicide in the death of a four year child due to medical accident</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

**Table 2.** Value of the thresholds and the coefficients

<table>
<thead>
<tr>
<th>Thresholds</th>
<th>t_{para}</th>
<th>t_{sent}</th>
<th>t_{dep}</th>
<th>t_{diff}</th>
<th>t_{tag}</th>
<th>Coefficients</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.2</td>
<td>0.35</td>
<td>0.2</td>
<td>20</td>
<td>C_{NUM}</td>
<td>C_{NE}</td>
</tr>
</tbody>
</table>

**Table 3.** The number of units judged as common part and the number of units judged as significant different part

<table>
<thead>
<tr>
<th>Common</th>
<th>Different</th>
<th>A</th>
<th>B</th>
<th>Manual</th>
<th>Correct</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>7</td>
<td>26</td>
<td>11</td>
<td>8</td>
<td>7</td>
<td>1.00</td>
<td>0.86</td>
</tr>
<tr>
<td>Topic 2</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>9</td>
<td>4</td>
<td>1.00</td>
<td>0.44</td>
</tr>
<tr>
<td>Topic 3</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>8</td>
<td>7</td>
<td>1.00</td>
<td>0.88</td>
</tr>
<tr>
<td>Topic 4</td>
<td>3</td>
<td>3</td>
<td>14</td>
<td>3</td>
<td>3</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Topic 5</td>
<td>8</td>
<td>11</td>
<td>22</td>
<td>7</td>
<td>6</td>
<td>0.75</td>
<td>0.86</td>
</tr>
<tr>
<td>Topic 6</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td>7</td>
<td>1.00</td>
<td>0.88</td>
</tr>
<tr>
<td>Topic 7</td>
<td>6</td>
<td>19</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>0.83</td>
<td>1.00</td>
</tr>
<tr>
<td>Topic 8</td>
<td>11</td>
<td>9</td>
<td>6</td>
<td>14</td>
<td>9</td>
<td>0.82</td>
<td>0.64</td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>63</td>
<td>49</td>
<td></td>
<td></td>
<td>0.92</td>
<td>0.78</td>
</tr>
</tbody>
</table>

In some cases, syntactically similar but semantically different units are judged as common part. For example, “Tokyo chisai ha muzai wo iiwatashita” (The Tokyo district court declared not guilty) and “Tokyo kousai ha muzai wo iiwatashita” (The Tokyo high court declared not guilty) seem similar, but they are completely different since subjects of the two sentence are different. In the case of the example of Figure 3 in section 4, a phrase corresponding to the phrase $P_{B3}$ in the simple sentence B is missing from the simple sentence A. However, the difference is not as significant as the difference between “the Tokyo district court” and “the Tokyo high court”. We need to distinguish between these cases to solve this problem.
There is some error in morphological analysis and dependency analysis. Due to the error, some units could not be extracted correctly.

In some cases, a unit in one article corresponds to two or more units in the other article. In the case of the topic 3, name of the arrested person and the reason are separately described in different units of the article A, while they are described together in one unit of the article B.

- A leader of a loan shark group was arrested on extortion. The arrested person is Kouji Kamei.
- Kouji Kamei, leader of a loan shark group, was arrested on extortion.

In our evaluation, if one of the two units of the article A is judged as a similar unit to the unit of the article B, we judge “correct”.

Next, we look at the units judged as different parts. Summary of the common part and the different part of two articles of each topic judged by our method is shown in Table 4 and 5. Sentences in parentheses indicate that they were judged as different part although they should have been judged as common part \(^9\). Basic information of each topic, such as date, time and place of the event’s occurrence, relevant people and organization, etc, is described in both of the articles. On the other hand, some additional information is described in either of the articles. In the case of the topic 1, information about time and place of the accident occurrence and the airline company’s name is commonly described, while the airline company’s comment, interviews with passengers, behaviour of rescue staff and passengers at the airport, and other accidents due to turbulence occurred in the past are described only either of the articles. We can see such difference by using our method.

8. Related Work

Researches on multi-document text summarization have been done in the past. Multi-document text summarization is the process of creating a compressed version of a given set of documents. Generally, they detect parts (e.g. words, sentences) which appears in many of the given documents \([12,13]\) – they focus on detecting common part rather than different part. Whereas we focus on detecting different part among documents since we think that different part among documents includes a lot of useful information.

Some researches on multi-document text summarization use information extraction technique \([14,15]\). In their cases, domain (topic) of a given set of documents and information which should be extracted are defined in advance. For example, they collect documents about terrorism (the documents do not have to deal with exactly the same terrorism), and extract predefined information, such as date and place of the event, the number of victims, name of the terrorist organization, etc. Whereas we do not define either domain of a given document set or information which should be extracted. Assumption of our method is that given documents deal with exactly the same event.

The UNIX operating system has a file comparison utility, “diff”. It is based on solving the longest common subsequence problem. It compares two sequences by only one

---

\(^9\)In the case of the topic 4, information about the accident appears in both of the articles. However, their contents are different from each other. The same holds for the topic 5.
comparison unit ("diff" makes line-by-line comparison), and the common subsequence should basically appear in the same relative order. Our method sequentially compares articles by three different comparison units, and considers dependency structure as well as word occurrence. Additionally, relative order of the common parts do not have to be the same.

9. Conclusion

In this paper, we presented a method for detecting difference between articles on the same topic. In our method, articles are sequentially compared by three different comparison units: paragraphs, sentences, and simple sentences. We evaluated our method on Japanese news articles, and show that we can see the difference between articles by using our method.

At this point, we do not use any language resources such as thesauri and so on. Fortunately, the first version of the Japanese WordNet [16] has been released recently. Some errors described in section 7 could be solved by using it.

Our method can detect difference between only two articles. We are planning to construct a system that shows difference among three or more articles depending on user behaviour. The process will be as follows (Figure 6).

1. System shows a list of articles.
2. User selects one of the articles (let the article be $A$), and read it (Information that the user selected $A$ is automatically sent to the system).
3. System detects difference between $A$ and each of the rest of the articles, and shows a list of the articles with the detected difference.
4. User selects another one of the articles (let the article be “article $B$”), and read it (Information that the user selected $B$ is automatically sent to the system).
5. System detects difference between articles that the user has already read (i.e. $A$ and $B$) and each of the rest of the articles, and shows a list of the articles with the detected difference.
6. Repeat the process until the user reads all of the articles or stops reading articles.

In order to realize the system, we will consider how to show difference among articles and how to order the articles.

References

<table>
<thead>
<tr>
<th>Topic</th>
<th>Article</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic 1</strong> Common</td>
<td></td>
<td>KLM Flight 867 from Amsterdam to Osaka was hit by a turbulence at 0:43 am. The airplane continued to fly, and arrived at Osaka at 9:15 am. The airplane left Amsterdam at 10:40 pm the previous day. The airplane was hit by the turbulence about 2 hours after the departure.</td>
</tr>
<tr>
<td>A</td>
<td>The airline company’s message about the reason why the airplane did not make an emergency landing or return to the nearest airport. Interviews with some passengers. Behaviour of rescue staff and passengers at the airport. (7 Japanese passengers and 3 Dutch crews were transported to a hospital.)</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Some passengers were participants and conductors of a tour offered by a travel agency, and one of them got injured. Other past accidents due to turbulence. (7 Japanese passengers and 3 Dutch crews were transported to a hospital.)</td>
<td></td>
</tr>
<tr>
<td><strong>Topic 2</strong> Common</td>
<td></td>
<td>High school students who were on a school trip shoplifted in Los Angels. 21 students were suspended from school for 5 days. A baseball team of the high school has participated in the high school baseball tournament five times. Some of the 21 students were members of the baseball team.</td>
</tr>
<tr>
<td>A</td>
<td>NONE</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>108 high school students were traveling the US from November 7th to 12th. 8 students shoplifted 33 items from duty-free shops in the LA International Airport. A message from the vice principal of the high school.</td>
<td></td>
</tr>
<tr>
<td><strong>Topic 3</strong> Common</td>
<td></td>
<td>A leader of a loan shark group and a woman were arrested on extortion. He is a wanted criminal for violation of the investment law related to a family suicide in the wake of repayment pressure in Osaka. They extorted about 400 thousand yen from a company managing national chain pet shops. Someone shot a gun to a pet shop owned by the company in April, and link between the case and the extortion is under investigation.</td>
</tr>
<tr>
<td>A</td>
<td>He demanded up to 270 times as high rate as the legal interest rate on loans. Detail information about the family suicide. A new law to tighten controls over loan sharks was established after the family suicide.</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>NONE</td>
<td></td>
</tr>
<tr>
<td><strong>Topic 4</strong> Common</td>
<td></td>
<td>Files on a police officer were sent to prosecutor as violation of the traffic control law. He drove drunk, caused a minor collision, and fled.</td>
</tr>
<tr>
<td>A</td>
<td>Information about the accident.</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Information about the accident. The police officer was dismissed in disgrace.</td>
<td></td>
</tr>
<tr>
<td>Topic</td>
<td>Article</td>
<td>Contents</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>----------</td>
</tr>
</tbody>
</table>
| Topic 5 | Common | A fire broke out in a taxi driver’s two-story wooden house, and burnt it down.  
A family of four lived there.  
The taxi driver said that he spread heating oil and set the fire.  
Interviews about the children who were killed in the fire.  
B The police suspects that cause of the fire is a domestic issue.  
Interviews about the children who were killed in the fire. |
| Topic 6 | Common | A university student belonging to the track team was arrested on molestation of a female high school student in a train.  
She caught him at the scene, and handed him over to a station employee.  
He admitted to the charges.  
He was expelled from a track team of the university.  
A He was on the way from a training camp to the campus for taking classes.  
(The track team qualified for the university marathon relay race the following year.)  
B (The track team qualified for the university marathon relay race the following year.) |
| Topic 7 | Common | The education board of Osaka announced a plan to prohibit public elementary/junior high school students from bringing mobile phones to school.  
Public high school students will be prohibited from using mobile phones in school.  
In Osaka, 88% of public elementary schools and 94% of public junior high schools now prohibit students from bringing mobile phones.  
A Cyberbullying is rife over, and using mobile phones causes students to lose concentration.  
Objections against the plan by a vice-principal of a high school and a member of the educational board in a certain city.  
Results of survey on students’ use of mobile phones conducted by the educational board in Osaka.  
Fewer than half of the public high schools now prohibit students from bringing mobile phones in Hyogo, next to Osaka.  
B NONE |
| Topic 8 | Common | Tokyo High Public Prosecutors’ Office decided not to appeal to the Supreme Court against an acquittal of a doctor who was accused of negligent homicide  
A prosecutor’s comment on the decision.  
A (A nursery school toddler died due to sticking a chopstick into his throat.)  
A comment by his mother.  
B (A nursery school toddler died due to sticking a chopstick into his throat.)  
The prosecutor’s claim in the court case. |