Extracting Key Paragraph based on Topic and Event Detection — Towards Multi-Document Summarization

Fumiyo Fukumoto and Yoshimi Suzuki†
Department of Computer Science and Media Engineering,
Yamanashi University
4-3-11 Takeda, Kofu 400-8511 Japan
{*fukumoto@skye.esb, ysuzuki@alps1.esi*†}*.yamanashi.ac.jp*

Abstract

This paper proposes a method for extracting key paragraph for *multi*-document summarization based on distinction between a topic and an event. A topic and an event are identified using a simple criterion called domain dependency of words. The method was tested on the TDT1 corpus which has been developed by the TDT Pilot Study and the result can be regarded as promising the idea of domain dependency of words effectively employed.

1 Introduction

As the volume of online documents has drastically increased, summarization techniques have become very important in IR and NLP studies. Most of the summarization work has focused on a single document. This paper focuses on *multi*-document summarization: broadcast news documents about the same *topic*. One of the major problems in the multi-document summarization task is how to identify differences and similarities across documents. This can be interpreted as a question of how to make a clear distinction between an *event* and a *topic* in documents. Here, an event is the subject of a document itself, i.e. a writer wants to express, in other words, notions of who, what, where, when, why and how in a document. On the other hand, a topic in this paper is some unique thing that happens at some specific time and place, and the unavoidable consequences. It becomes *background* among documents. For example, in the documents of 'Kobe Japan quake', the event includes early reports of damage, location and nature of quake, rescue efforts, consequences of the quake, and on-site reports, while the topic is Kobe Japan quake. The well-known past experience from IR that notions of who, what, where, when, why and how may not make a great contribution to the topic detection and tracking task (Allan and Papka, 1998) causes this fact, i.e. a topic and an event are different from each other[1] .

[1] Some topic words can also be an event. For instance, in the document shown in Figure 1, 'Japan' and 'quake' are topic words and also event words in the document. However, we regarded these words as a topic, i.e. not be an event.

In this paper, we propose a method for extracting key paragraph for multi-document summarization based on distinction between a topic and an event. We use a simple criterion called *domain dependency of words* as a solution and present how the idea of domain dependency of words can be utilized effectively to identify a topic and an event, and thus allow multi-document summarization.

The basic idea of our approach is that whether a word appeared in a document is a topic (an event) or not, depends on the domain to which the document belongs. Let us take a look at the following document from the TDT1 corpus.

(1-2) Two Americans known dead in Japan quake

1. The number of [Americans] known to have been killed in Tuesday's earthquake in Japan has risen to two, the [State] [Department] said Thursday.
2. The first was named Wednesday as Voni Lynn Wong, a teacher from California. [State] [Department] spokswoman Christine Shelly declined to name the second, saying formalities of notifying the family had not been completed.
3. With the death toll still mounting, at least 4,000 people were killed in the earthquake which devastated the Japanese city of Kobe.
4. [U.S.] diplomats were trying to locate the several thousand-strong [U.S.] community in the area, and some [Americans] who had been made homeless were found shelter in the [U.S.] consulate there, which was only lightly damaged in the quake.
5. Shelly said an emergency [State] [Department] telephone number in Washington to provide information about private [American] citizens in Japan had received over 6,000 calls, more than half of them seeking direct assistance.
6. The Pentagon has agreed to send 57,000 blankets to Japan and [U.S.] ambassador to Tokyo Walter Mondale has donated a $25,000 discretionary fund for emergencies to the Japanese Red Cross, Shelly said.
7. Japan has also agreed to a visit by a team of [U.S.] experts headed by Richard Witt, national director of the Federal Emergency Management Agency.

Figure 1: The document titled 'Two Americans known dead in Japan quake'

Figure 1 is the document whose topic is 'Kobe Japan quake', and the subject of the document (event

words) is 'Two Americans known dead in Japan quake'. Underlined words denote a topic, and the words marked with '[ ]' are events. '1~7' of Figure 1 is paragraph *id*. Like Luhn's technique of keyword extraction, our method assumes that an event associated with a document appears throughout paragraphs (Luhn, 1958), but a topic does not. This is because an event is the subject of a document itself, while a topic is an event, along with all directly related events. In Figure 1, event words 'Americans' and 'U.S.', for instance, appears across paragraphs, while a topic word, for example, 'Kobe' appears only the *third* paragraph. Let us consider further a broad coverage domain which consists of a small number of sample news documents about the same topic, 'Kobe Japan quake'. Figure 2 and 3 are documents with 'Kobe Japan quake'.

(1-1) Quake collapses buildings in central Japan

1. At least two people died and dozens were injured when a powerful earthquake rolled through central Japan Tuesday morning, collapsing buildings and setting off fires in the cities of Kobe and Osaka.
2. The Japan Meteorological Agency said the earthquake, which measured 7.2 on the open-ended Richter scale, rumbled across Honshu Island from the Pacific Ocean to the Japan Sea.

.....

Figure 2: The document titled 'Quake collapses buildings in central Japan'

(1-3) Kobe quake leaves questions about medical system

1. The earthquake that devastated Kobe in January raised serious questions about the efficiency of Japan's emergency medical system, a government report released on Tuesday said.
2. 'The earthquake exposed many issues in terms of quantity, quality, promptness and efficiency of Japan's medical care in time of disaster,' the report on health and welfare said.

.....

Figure 3: The document titled 'Kobe quake leaves questions about medical system'

Underlined words in Figure 2 and 3 show the topic of these documents. In these two documents, 'Kobe' which is a topic appears in every document, while 'Americans' and 'U.S.' which are events of the document shown in Figure 1, does not appear. Our technique for making the distinction between a topic and an event explicitly exploits this feature of the domain dependency of words: how strongly a word features a given set of data.

The rest of the paper is organized as follows. The next section provides domain dependency of words which is used to identify a topic and an event for broadcast news documents. We then present a method for extracting topic and event words, and describe a paragraph-based summarization algorithm using the result of topic and event extraction. Finally, we report some experiments using the TDT1 corpus which has been developed by the TDT (Topic Detection and Tracking) Pilot Study (Allan and Carbonell, 1998) with a discussion of evaluation.

2 Domain Dependency of Words

The domain dependency of words that how strongly a word features a given set of data (documents) contributes to event extraction, as we previously reported (Fukumoto et al., 1997). In the study, we hypothesised that the articles from the *Wall Street Journal* corpus can be structured by three levels, i.e. *Domain*, *Article* and *Paragraph*. If a word is an event in a given article, it satisfies the two conditions: (1) The dispersion value of the word in the *Paragraph* level is smaller than that of the *Article*, since the word appears throughout paragraphs in the *Paragraph* level rather than articles in the *Article* level. (2) The dispersion value of the word in the *Article* is smaller than that of the *Domain*, as the word appears across articles rather than domains.

However, there are two problems to adapt it to multi-document summarization task. The first is that the method extracts only events in the document. Because the goal of the study is to summarize a single document, and thus there is no answer to the question of how to identify differences and similarities across documents. The second is that the performance of the method greatly depends on the structure of a given data itself. Like the *Wall Street Journal* corpus, (i) if a given data can be structured by three levels, *Paragraph*, *Article* and *Domain*, each of which consists of several paragraphs, articles and domains, respectively, and (ii) if *Domain* consists of different subject domains, such as 'aerospace', 'environment' and 'stock market', the method can be done with satisfactory accuracy. However, there is no guarantee to make such an appropriate structure from a given set of documents in the multi-document summarization task.

The purpose of this paper is to define domain dependency of words for a number of sample documents about the same topic, and thus for multi-document summarization task. Figure 4 illustrates the structure of *broadcast news* documents which have been developed by the TDT (Topic Detection and Tracking) Pilot Study (Allan and Carbonell, 1998). It consists of two levels, Paragraph and Document. In Document level, there is a small number of sample news documents about the same topic. These documents are arranged in chronological order such as, '(1-1) Quake collapses buildings in central Japan (Figure 2)', '(1-2) Two Americans known dead in Japan quake (Figure 1)' and '(1-3) Kobe quake leaves questions about medical system (Figure 3)'. A particular document consists of several

paragraphs. We call it Paragraph level. Let words within a document be an event, a topic, or among others (We call it a *general* word).

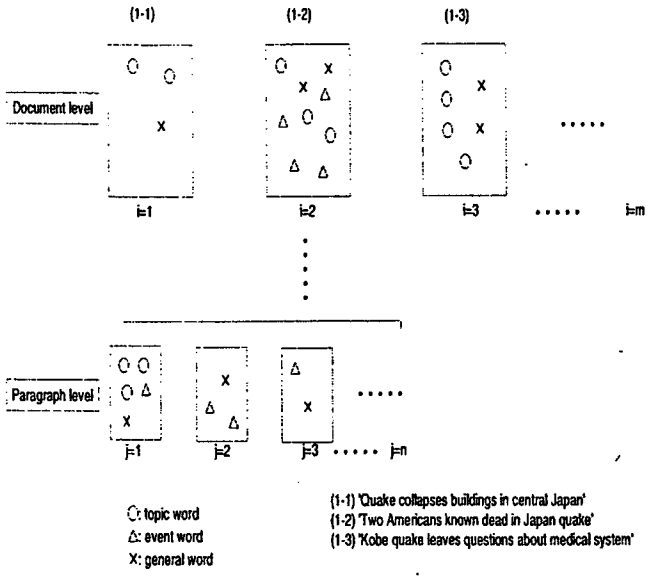

Figure 4: The structure of *broadcast news* documents (event extraction)

Given the structure shown in Figure 4, how can we identify every word in document (1-2) with an event, a topic or a general word? Our method assumes that an event associated with a document appears across paragraphs, but a topic word does not. Then, we use domain dependency of words to extract event and topic words in document (1-2). Domain dependency of words is a measure showing how greatly each word features a given set of data.

In Figure 4, let '○', '△' and '×' denote a topic, an event and a general word in document (1-2), respectively. We recall the example shown in Figure 1. '△', for instance, 'U.S.' appears *across paragraphs*. However, in the Document level, '△' *frequently* appears *in document*, (1-2) *itself*. On the basis of this example, we hypothesize that if word i is an event, it satisfies the following condition:

[1] Word i greatly depends on a particular document in the Document level rather than a particular paragraph in the Paragraph.

Next, we turn to identify the remains (words) with a topic, or a general word. In Figure 5, a topic of documents (1-1) $\sim$ (1-3), for instance, 'Kobe' appears in a *particular* paragraph in each level of Paragraph1, Paragraph2 and Paragraph3. Here, (1-1), (1-2) and (1-3) corresponds to Paragraph1, Paragraph2 and Paragraph3, respectively. On the other hand, in Document level, a topic *frequently* appears *across documents*. Then, we hypothesize that if word i is a

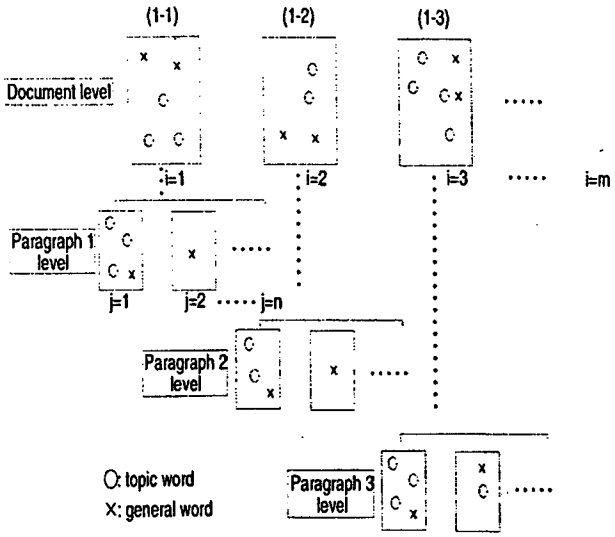

Figure 5: The structure of *broadcast news* documents (topic extraction)

topic, it satisfies the following condition:

[2] Word i greatly depends on a particular paragraph in each Paragraph level rather than a particular document in Document.

3 Topic and Event Extraction

We hypothesized that the domain dependency of words is a key clue to make a distinction between a topic and an event. This can be broken down into two observations: (i) whether a word appears *across paragraphs* (*documents*), (ii) whether or not a word appears *frequently*. We represented the former by using *dispersion value*, and the latter by *deviation value*. Topic and event words are extracted by using these values.

The first step to extract topic and event words is to assign weight to the individual word in a document. We applied TF*IDF to each level of the Document and Paragraph, i.e. Paragraph1, Paragraph2 and Paragraph3.

$$W\mathrm{d}_{it} = TF\mathrm{d}_{it} * \log \frac{N}{N\mathrm{d}_t} \quad (1)$$

$W\mathrm{d}_{it}$ in formula (1) is TF*IDF of term t in the i-th document. In a similar way, $W\mathrm{p}_{it}$ denotes TF*IDF of the term t in the i-th paragraph. $TF\mathrm{d}_{it}$ in (1) denotes term frequency of t in the i-th document. N is the number of documents and $N\mathrm{d}_t$ is the number of documents where t occurs. The second step is to calculate domain dependency of words. We defined it by using formula (2) and (3).

$$Disp\mathsf{D}_t = \sqrt{\frac{\sum_{i=1}^{m}(W\mathsf{d}_{it} - mean_t)^2}{m}} \quad (2)$$

$$Devd_{it} = \frac{(W\mathsf{d}_{it} - mean_t)}{Disp\mathsf{D}_t} * 10 + 50 \quad (3)$$

Formula (2) is dispersion value of term t in the level of Document which consists of m documents, and denotes how frequently t appears across documents. In a similar way, $Disp\mathsf{P}_t$ denotes dispersion of term t in the level of Paragraph. Formula (3) is the deviation value of t in the i-th document and denotes how frequently it appears in a particular document, the i-th document. $Devp_{it}$ is deviation of term t in the i-th paragraph. In (2) and (3), $mean_t$ is the mean of the total TF*IDF values of term t in the level of Document.

The last step is to extract a topic and an event using formula (2) and (3). We recall that if t is an event, it satisfies [1] described in section 2. This is shown by using formula (4) and (5).

$$Disp\mathsf{P}_t < Disp\mathsf{D}_t \quad (4)$$

$$for\ all\ \mathsf{p}_j \in \mathsf{d}_i \quad Devp_{jt} < Devd_{it} \quad (5)$$

Formula (4) shows that t frequently appears across paragraphs rather than documents. In formula (5), d_i is the i-th document and consists of the number of n paragraphs (see Figure 4). p_j is an element of d_i. (5) shows that t frequently appears in the i-th document d_i rather than paragraphs p_j ($1 \leq j \leq n$). On the other hand, if t satisfies formula (6) and (7), then propose t as a topic.

$$Disp\mathsf{P}_t \geq Disp\mathsf{D}_t \quad (6)$$

$$for\ all\ \mathsf{d}_i \in D,$$

$$\mathsf{p}_{jt}\ exists\ such\ that \quad Devp_{jt} \geq Devd_{it} \quad (7)$$

In formula (7), D consists of the number of m documents (see Figure 5). (7) denotes that t frequently appears in the particular paragraph p_j rather than the document d_i which includes p_j.

4 Key Paragraph Extraction

The summarization task in this paper is paragraph-based extraction (Stein et al., 1999). Basically, paragraphs which include not only event words but also topic words are considered to be significant paragraphs. The basic algorithm works as follows:

1. For each document, extract topic and event words.
2. Determine the paragraph weights for all paragraphs in the documents:
   (a) Compute the sum of topic weights over the total number of topic words for each paragraph.
   (b) Compute the sum of event weights over the total number of event words for each paragraph.
   A topic and an event weights are calculated by using $Devd_{it}$ in formula (3). Here, t is a topic or an event and i is the i-th document in the documents.
   (c) Compute the sum of (a) and (b) for each paragraph.
3. Sort the paragraphs according to their weights and extract the N highest weighted paragraphs in documents in order to yield summarization of the documents.
4. When their weights are the same, Compute the sum of all the topic and event word weights. Select a paragraph whose weight is higher than the others.

5 Experiments

Evaluation of extracting key paragraph based on multi-document is difficult. First, we have not found an existing collection of summaries of multiple documents. Second, the manual effort needed to judge system output is far more extensive than for single document summarization. Consequently, we focused on the TDT1 corpus. This is because (i) events have been defined to support the TDT study effort, (ii) it was completely annotated with respect to these events (Allan and Carbonell, 1997). Therefore, we do not need the manual effort to collect documents which discuss about the target event.

We report the results of three experiments. The first experiment, Event Extraction, is concerned with event extraction technique. In the second experiment, Tracking Task, we applied the extracted topics to tracking task (Allan and Carbonell, 1998). The third experiment, Key Paragraph Extraction is conducted to evaluate how the extracted topic and event words can be used effectively to extract key paragraph.

5.1 Data

The TDT1 corpus comprises a set of documents (15,863) that includes both newswire (Reuters) 7,965 and a manual transcription of the broadcast news speech (CNN) 7,898 documents. A set of 25 target events were defined [2] .

All documents were tagged by the tagger (Brill, 1992). We used nouns in the documents.

[2] http://morph.ldc.upenn.edu/TDT

5.2 Event Extraction

We collected 300 documents from the TDT1 corpus, each of which is annotated with respect to one of 25 events. The result is shown in Table 1.

In Table 1, 'Event type' illustrates the target events defined by the TDT Pilot Study. 'Doc' denotes the number of documents. 'Rec' (Recall) is the number of correct events divided by the total number of events which are selected by a human, and 'Prec' (Precision) stands for the number of correct events divided by the number of events which are selected by our method. The denominator 'Rec' is made by a human judge. 'Accuracy' in Table 1 is the total average ratio.

In Table 1, recall and precision values range from 55.0/47.0 to 83.3/84.2, the average being 71.0/72.2. The worst result of recall and precision was when event type was 'Serbs violate Bihac' (55.0/59.3). We currently hypothesize that this drop of accuracy is due to the fact that some documents are against our assumption of an event. Examining the documents whose event type is 'Serbs violate Bihac', 3 (one from CNN and two from Reuters) out of 16 documents has discussed the same event, i.e. 'Bosnian Muslim enclave hit by heavy shelling'. As a result, the event appears across these three documents. Future research will shed more light on that.

5.3 Tracking Task

Tracking task in the TDT project is starting from a few sample documents and finding all subsequent documents that discuss the same event (Allan and Carbonell, 1998), (Carbonell et al., 1999). The corpus is divided into two parts: training set and test set. Each of the documents is flagged as to whether it discusses the target event, and these flags ('YES', 'NO') are the only information used for training the system to correctly classify the target event. We applied the extracted topic to the tracking task under these conditions. The basic algorithm used in the experiment is as follows:

1. Create a single document S_{tp} and represent it as a term vector

   For the results of topic extraction, all the documents that belong to the same topic are bundled into a single document S_{tp} and represent it by a term vector as follows:

$$S_{tp} = \begin{bmatrix} t_{tp1} \\ t_{tp2} \\ \vdots \\ t_{tpn} \end{bmatrix} \text{ s.t. } t_{tpj} = \begin{cases} f(t_{tpj}) & \text{if } t_{tpj} \text{ is a topic of } S_{tp} \\ 0 & \text{otherwise} \end{cases}$$

   $f(w)$ denotes term frequency of word w.

2. Represent other training and test documents as term vectors

   Let $S_1, \cdots, S_m$ be all the other training documents (where m is the number of training documents which does not belong to the target event) and S_x be a test document which should be classified as to whether or not it discusses the target event. $S_1, \cdots, S_m$ and S_x are represented by term vectors as follows:

$$S_i = \begin{bmatrix} t_{i1} \\ t_{i2} \\ \vdots \\ t_{in} \end{bmatrix} \text{ s.t. } t_{ij} = \begin{cases} f(t_{ij}) & \text{if } t_{ij} \ (1 \le i \le m) \text{ appears in } S_i \text{ and not be a topic of } S_{tp} \\ 0 & \text{otherwise} \end{cases}$$

$$S_x = \begin{bmatrix} t_{x1} \\ t_{x2} \\ \vdots \\ t_{xn} \end{bmatrix} \text{ s.t. } t_{xj} = \begin{cases} f(t_{xj}) & \text{if } t_{xj} \text{ appears in } S_x \\ 0 & \text{otherwise} \end{cases}$$

3. Compute the similarity between a training document and a test document

   Given a vector representation of documents S_1, $\cdots$, S_m, S_{tp} and S_x, a similarity between two documents S_i $(1 \le i \le m, tp)$ and the test document S_x would be obtained by using formula (8), i.e. the inner product of their normalized vectors.

$$Sim(S_i, S_x) = \frac{S_i \cdot S_x}{| S_i || S_x |} \qquad (8)$$

   The greater the value of $Sim(S_i, S_x)$ is, the more similar S_i and S_x are. If the similarity value between the test document S_x and the document S_{tp} is largest among all the other pairs of documents, i.e. $(S_1, S_x), \cdots, (S_m, S_x)$, S_x is judged to be a document that discusses the target event.

We used the standard TDT evaluation measure [3]. Table 2 illustrates the result.

Table 2: The results of tracking task

N_t	%Miss	%F/A	F1	%Rec	%Prec
1	32.5	0.16	0.68	67.5	70.0
2	23.7	0.06	0.80	76.3	87.8
4	23.1	0.05	0.81	76.9	90.1
8	12.0	0.08	0.87	88.0	91.4
16	13.7	0.06	0.89	86.3	93.6
Avg	21.0	0.08	0.76	79.0	86.6

In Table 2, 'N_t' denotes the number of positive training documents where N_t takes on values 1, 2, 4, 8

[3] http://www.nist.gov/speech/tdt98.htm

Table 1: The results of event words extraction

Event type	Doc	Avg Rec/Avg Prec	Event type	Doc	Avg Rec/Avg Prec
Aldrich Ames	8	61.7/70.5	Karrigan/Harding	2	64.7/55.5
Carlos the Jackal	8	60.7/73.3	Kobe Japan quake	16	74.5/75.0
Carter in Bosnia	16	76.3/79.1	Lost in Iraq	16	75.7/68.8
Cessna on White House	8	65.7/80.0	NYC Subway bombing	16	68.0/84.2
Clinic Murders	16	75.9/80.0	OK-City bombing	16	78.8/47.0
Comet into Jupiter	16	65.2/61.9	Pentium chip flaw	4	81.1/72.9
Cuban riot in Panama	2	65.2/73.9	Quayle lung clot	8	63.6/74.4
Death of Kim Jong	16	83.3/71.4	Serbians down F-16	16	78.6/75.0
DNA in OJ trial	16	78.7/72.9	Serbs violate Bihac	16	55.0/59.3
Haiti ousts observers	8	62.0/74.0	Shannon Faulker	4	71.4/82.4
Hall's copter	16	78.5/75.0	USAir 427 crash	16	72.6/86.3
Humble, TX, flooding	16	80.4/70.2	WTC Bombing trial	16	62.6/70.1
Justice-to-be Breyer	8	75.9/72.2			
Accuracy			71.0/72.2		

and 16. 'Miss' means Miss rate, which is the ratio of the documents that were judged as YES but were not evaluated as YES for the run in question. 'F/A' shows false alarm rate and 'F1' is a measure that balances recall and precision. 'Rec' denotes the ratio of the documents judged YES that were also evaluated as YES, and 'Prec' is the percent of the documents that were evaluated as YES which correspond to documents actually judged as YES.

Table 2 shows that more training data helps the performance, as the best result was when we used $N_t = 16$.

Table 3 illustrates the extracted topic and event words in a sample document. The topic is 'Kobe Japan quake' and the number of positive training documents is 4. '$Devp_{1t}$', '$Devd_{1t}$', '$DispP_t$' and '$DispD_t$' denote values calculated by using formula (2) and (3).

Table 3: Topic and event words in 'Kobe Japan quake'

Topic word	$Devp_{1t}$	$Devd_{1t}$	$DispP_t$	$DispD_t$
earthquake	53.5	50.0	12.3	10.3
Japan	69.8	50.0	13.3	9.8
Kobe	56.6	50.0	8.6	6.4
fire	57.0	46.4	2.3	1.5

Event word	$Devp_{1t}$	$Devd_{1t}$	$DispP_t$	$DispD_t$
emergency	50.0	74.7	0.9	1.5
area	40.6	50.0	0.6	1.0
worker	50.0	66.1	0.4	1.0
rescue	43.3	50.0	2.3	3.4

In Table 3, 'Event' denotes event words in the first document in chronological order from $N_t = 4$, and the title of the document is 'Emergency Work Continues After Earthquake in Japan'. Table 3 clearly demonstrates that the criterion, domain dependency of words effectively employed.

Figure 6 illustrates the DET (Detection Evaluation Tradeoff) curves for a sample event (event type is 'Comet into Jupiter') runs at several values of N_t.

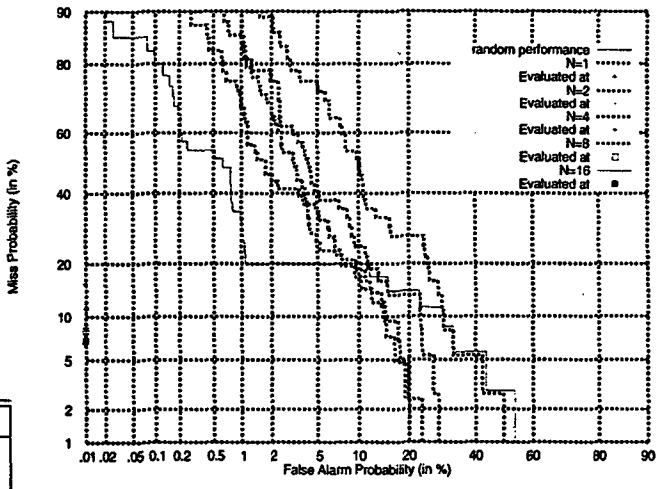

Figure 6: DET curve for a sample tracking runs

Overall, the curves also show that more training helps the performance, while there is no significant difference among $N_t = 2$, 4 and 8.

5.4 Key Paragraph Extraction

We used 4 different sets as a test data. Each set consists of 2, 4, 8 and 16 documents. For each set, we

5.2 Event Extraction

We collected 300 documents from the TDT1 corpus, each of which is annotated with respect to one of 25 events. The result is shown in Table 1.

In Table 1, 'Event type' illustrates the target events defined by the TDT Pilot Study. 'Doc' denotes the number of documents. 'Rec' (Recall) is the number of correct events divided by the total number of events which are selected by a human, and 'Prec' (Precision) stands for the number of correct events divided by the number of events which are selected by our method. The denominator 'Rec' is made by a human judge. 'Accuracy' in Table 1 is the total average ratio.

In Table 1, recall and precision values range from 55.0/47.0 to 83.3/84.2, the average being 71.0/72.2. The worst result of recall and precision was when event type was 'Serbs violate Bihac' (55.0/59.3). We currently hypothesize that this drop of accuracy is due to the fact that some documents are against our assumption of an event. Examining the documents whose event type is 'Serbs violate Bihac', 3 (one from CNN and two from Reuters) out of 16 documents has discussed the same event, i.e. 'Bosnian Muslim enclave hit by heavy shelling'. As a result, the event appears across these three documents. Future research will shed more light on that.

5.3 Tracking Task

Tracking task in the TDT project is starting from a few sample documents and finding all subsequent documents that discuss the same event (Allan and Carbonell, 1998), (Carbonell et al., 1999). The corpus is divided into two parts: training set and test set. Each of the documents is flagged as to whether it discusses the target event, and these flags ('YES', 'NO') are the only information used for training the system to correctly classify the target event. We applied the extracted topic to the tracking task under these conditions. The basic algorithm used in the experiment is as follows:

1. Create a single document S_{tp} and represent it as a term vector

   For the results of topic extraction, all the documents that belong to the same topic are bundled into a single document S_{tp} and represent it by a term vector as follows:

$$S_{tp} = \begin{bmatrix} t_{tp1} \\ t_{tp2} \\ \vdots \\ t_{tpn} \end{bmatrix} \text{ s.t. } t_{tpj} = \begin{cases} f(t_{tpj}) & \text{if } t_{tpj} \text{ is a topic of } S_{tp} \\ 0 & \text{otherwise} \end{cases}$$

   $f(w)$ denotes term frequency of word w.

2. Represent other training and test documents as term vectors

   Let $S_1, \cdots, S_m$ be all the other training documents (where m is the number of training documents which does not belong to the target event) and S_x be a test document which should be classified as to whether or not it discusses the target event. $S_1, \cdots, S_m$ and S_x are represented by term vectors as follows:

$$S_i = \begin{bmatrix} t_{i1} \\ t_{i2} \\ \vdots \\ t_{in} \end{bmatrix} \text{ s.t. } t_{ij} = \begin{cases} f(t_{ij}) & \text{if } t_{ij} \ (1 \le i \le m) \text{ appears in } S_i \text{ and not be a topic of } S_{tp} \\ 0 & \text{otherwise} \end{cases}$$

$$S_x = \begin{bmatrix} t_{x1} \\ t_{x2} \\ \vdots \\ t_{xn} \end{bmatrix} \text{ s.t. } t_{xj} = \begin{cases} f(t_{xj}) & \text{if } t_{xj} \text{ appears in } S_x \\ 0 & \text{otherwise} \end{cases}$$

3. Compute the similarity between a training document and a test document

   Given a vector representation of documents S_1, $\cdots$, S_m, S_{tp} and S_x, a similarity between two documents S_i $(1 \le i \le m, tp)$ and the test document S_x would be obtained by using formula (8), i.e. the inner product of their normalized vectors.

$$Sim(S_i, S_x) = \frac{S_i \cdot S_x}{\mid S_i \mid\mid S_x \mid} \qquad (8)$$

   The greater the value of $Sim(S_i, S_x)$ is, the more similar S_i and S_x are. If the similarity value between the test document S_x and the document S_{tp} is largest among all the other pairs of documents, i.e. $(S_1, S_x), \cdots, (S_m, S_x)$, S_x is judged to be a document that discusses the target event.

We used the standard TDT evaluation measure [3]. Table 2 illustrates the result.

Table 2: The results of tracking task

N_t	%Miss	%F/A	F1	%Rec	%Prec
1	32.5	0.16	0.68	67.5	70.0
2	23.7	0.06	0.80	76.3	87.8
4	23.1	0.05	0.81	76.9	90.1
8	12.0	0.08	0.87	88.0	91.4
16	13.7	0.06	0.89	86.3	93.6
Avg	21.0	0.08	0.76	79.0	86.6

In Table 2, 'N_t' denotes the number of positive training documents where N_t takes on values 1, 2, 4, 8

[3] http://www.nist.gov/speech/tdt98.htm

Table 1: The results of event words extraction

Event type	Doc	Avg Rec/Avg Prec	Event type	Doc	Avg Rec/Avg Prec
Aldrich Ames	8	61.7/70.5	Karrigan/Harding	2	64.7/55.5
Carlos the Jackal	8	60.7/73.3	Kobe Japan quake	16	74.5/75.0
Carter in Bosnia	16	76.3/79.1	Lost in Iraq	16	75.7/68.8
Cessna on White House	8	65.7/80.0	NYC Subway bombing	16	68.0/84.2
Clinic Murders	16	75.9/80.0	OK-City bombing	16	78.8/47.0
Comet into Jupiter	16	65.2/61.9	Pentium chip flaw	4	81.1/72.9
Cuban riot in Panama	2	65.2/73.9	Quayle lung clot	8	63.6/74.4
Death of Kim Jong	16	83.3/71.4	Serbians down F-16	16	78.6/75.0
DNA in OJ trial	16	78.7/72.9	Serbs violate Bihac	16	55.0/59.3
Haiti ousts observers	8	62.0/74.0	Shannon Faulker	4	71.4/82.4
Hall's copter	16	78.5/75.0	USAir 427 crash	16	72.6/86.3
Humble, TX, flooding	16	80.4/70.2	WTC Bombing trial	16	62.6/70.1
Justice-to-be Breyer	8	75.9/72.2			
Accuracy			71.0/72.2		

and 16. 'Miss' means Miss rate, which is the ratio of the documents that were judged as YES but were not evaluated as YES for the run in question. 'F/A' shows false alarm rate and 'F1' is a measure that balances recall and precision. 'Rec' denotes the ratio of the documents judged YES that were also evaluated as YES, and 'Prec' is the percent of the documents that were evaluated as YES which correspond to documents actually judged as YES.

Table 2 shows that more training data helps the performance, as the best result was when we used $N_t = 16$.

Table 3 illustrates the extracted topic and event words in a sample document. The topic is 'Kobe Japan quake' and the number of positive training documents is 4. '$Devp_{1t}$', '$Devd_{1t}$', '$DispP_t$' and '$DispD_t$' denote values calculated by using formula (2) and (3).

Table 3: Topic and event words in 'Kobe Japan quake'

Topic word	$Devp_{1t}$	$Devd_{1t}$	$DispP_t$	$DispD_t$
earthquake	53.5	50.0	12.3	10.3
Japan	69.8	50.0	13.3	9.8
Kobe	56.6	50.0	8.6	6.4
fire	57.0	46.4	2.3	1.5

Event word	$Devp_{1t}$	$Devd_{1t}$	$DispP_t$	$DispD_t$
emergency	50.0	74.7	0.9	1.5
area	40.6	50.0	0.6	1.0
worker	50.0	66.1	0.4	1.0
rescue	43.3	50.0	2.3	3.4

In Table 3, 'Event' denotes event words in the first document in chronological order from $N_t = 4$, and the title of the document is 'Emergency Work Continues After Earthquake in Japan'. Table 3 clearly demonstrates that the criterion, domain dependency of words effectively employed.

Figure 6 illustrates the DET (Detection Evaluation Tradeoff) curves for a sample event (event type is 'Comet into Jupiter') runs at several values of N_t.

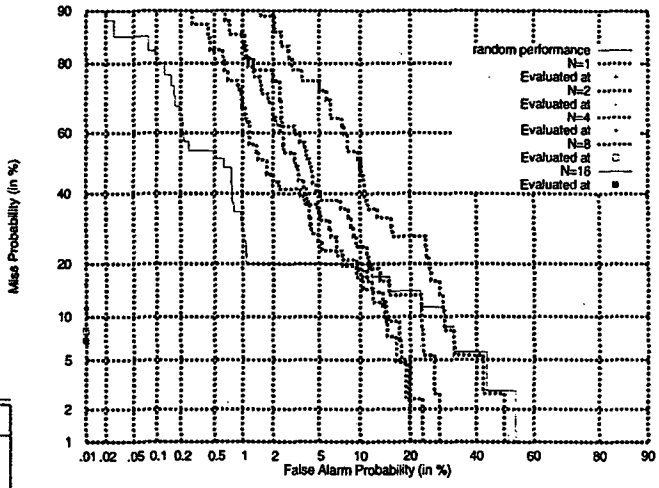

Figure 6: DET curve for a sample tracking runs

Overall, the curves also show that more training helps the performance, while there is no significant difference among $N_t = 2$, 4 and 8.

5.4 Key Paragraph Extraction

We used 4 different sets as a test data. Each set consists of 2, 4, 8 and 16 documents. For each set, we

extracted 10% and 20% of the full-documents paragraph length (Jing et al., 1998). Table 4 illustrates the result.

In Table 4, 'Num' denotes the number of documents in a set. 10 and 20% indicate the extraction ratio. 'Para' denotes the number of paragraphs extracted by a human judge, and 'Correct' shows the accuracy of the method.

The best result was 77.7% (the extraction ratio is 20% and the number of documents is 2).

We now turn our attention to the main question: how was the contribution of making the distinction between a topic and an event for summarization task? Figure 7 illustrates the results of the methods which used (i) the extracted topic and event words, i.e. our method, and (ii) only the extracted event words.

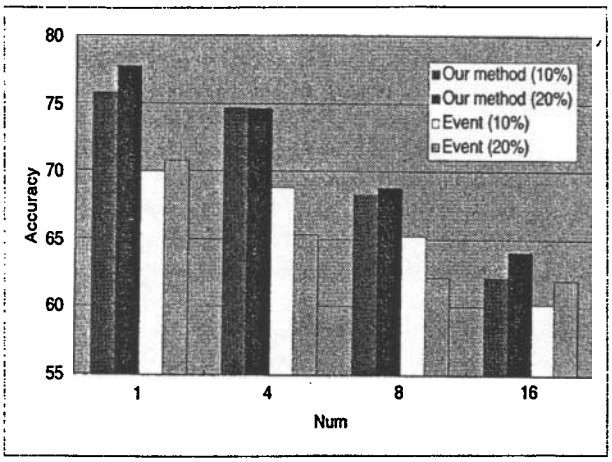

Figure 7: Accuracy with each method

In Figure 7, '(10%)' and '(20%)' denote the extracted paragraph ratio. 'Event' is the result when we used only the extracted event words. Figure 7 shows that our method consistently outperforms the method which used only the extracted events. To summarize the evaluation:

1. Event extraction effectively employed when each document discusses different subject about the same topic. This shows that the method will be applicable to other genres of corpora which consist of different subjects.
2. The result of tracking task (79.0% average recall and 86.6% average precision) is comparable to the existing tracking techniques which tested on the TDT1 corpus (Allan and Carbonell, 1998).
3. Distinction between a topic and an event improved the results of key paragraph extraction, as our method consistently outperforms the method which used only the extracted event words (see Figure 7).

6 Related Work

The majority of techniques for summarization fall within two broad categories: Those that rely on template instantiation and those that rely on passage extraction.

Work in the former approach is the DARPA-sponsored TIPSTER program and, in particular, the message understanding conferences has provided fertile ground for such work, by placing the emphasis of document analysis to the identification and extraction of certain core entities and facts in a document, while work on template-driven, knowledge-based summarization to date is hardly domain or genre-independent (Boguraev and Kennedy, 1997).

The alternative approach largely escapes this constraint, by viewing the task as one of identifying certain passages(typically sentences) which, by some metric, are deemed to be the most representative of the document's content. A variety of approaches exist for determining the salient sentences in the text: statistical techniques based on word distribution (Kupiec et al., 1995), (Zechner, 1996), (Salton et al., 1991), (Teufell and Moens, 1997), symbolic techniques based on discourse structure (Marcu, 1997) and semantic relations between words (Barzilay and Elhadad, 1997). All of their results demonstrate that passage extraction techniques are a useful first step in document summarization, although most of them have focused on a single document.

Some researchers have started to apply a single-document summarization technique to multi-document. Stein et. al. proposed a method for summarizing multi-document using single-document summarizer (Stralkowsik et al., 1998), (Stralkowski et al., 1999). Their method first summarizes each document of multi-document, then groups the summaries in clusters and finally, orders these summaries in a logical way (Stein et al., 1999). Their technique seems sensible. However, as she admits, (i) the order the information should not only depend on topic covered, (ii) background information that helps clarify related information should be placed first. More seriously, as Barzilay and Mani claim, summarization of multiple documents requires information about similarities and differences *across* documents. Therefore it is difficult to identify these information using a single-document summarizer technique (Mani and Bloedorn, 1997), (Barzilay et al., 1999).

A method proposed by Mani et. al. deal with the problem, i.e. they tried to detect the similarities and differences in information *content* among documents (Mani and Bloedorn, 1997). They used a spreading activation algorithm and graph matching in order to identify similarities and differences across documents. The output is presented as a set of paragraphs with similar and unique words highlighted. However, if the same information is men-

Table 4: The results of Key Paragraph Extraction

Num	Accuracy					
	%10		%20		Total	
	Para	Correct(%)	Para	Correct(%)	Para	Correct(%)
2	58	44(75.8)	117	91(77.7)	175	135(77.1)
4	107	80(74.7)	214	160(74.7)	321	240(74.7)
8	202	138(68.3)	404	278(68.8)	606	416(68.6)
16	281	175(62.2)	563	361(64.1)	844	536(63.5)
Total	648	437(67.4)	1,298	890(68.5)	1,946	1,327(68.1)

tioned several times in different documents, much of the summary will be redundant.

Allan et. al. also address the problem and proposed a method for event tracking using *common words* and *surprising features* by supplementing the corpus statistics (Allan and Papka, 1998) (Papka et al., 1999). One of the purpose of this study is to make a distinction between an event and an *event class* using surprising features. Here event class features are broad news areas such as politics, death, destruction and warfare. The idea is considered to be necessary to obtain high accuracy, while Allan claims that the surprising words do not provide a broad enough coverage to capture all documents on the event.

A more recent approach dealing with this problem is Barzilay et. al's approach (Barzilay et al., 1999). They used paraphrasing rules which are manually derived from the result of syntactic analysis to identify theme intersection and used language generation to reformulate them as a coherent summary. While promising to obtain high accuracy, the result of summarization task has not been reported.

Like Mani and Barzilay's techniques, our approach focuses on the problem that how to identify differences and similarities across documents, rather than the problem that how to form the actual summary (Sparck, 1993), (McKeown and Radev, 1995), (Radev and McKeown, 1998). However, while Barzilay's approach used paraphrasing rules to eliminate redundancy in a summary, we proposed *domain dependency of words* to address robustness of the technique.

7 Conclusion

In this paper, we proposed a method for extracting key paragraph for summarization based on distinction between a topic and an event. The results showed that the average accuracy was 68.1% when we used the TDT1 corpus. TIPSTER Text Summarization Evaluation (SUMMAC) proposed various methods for evaluating document summarization and tasks (Mani et al., 1999). Of these, participants submitted two summaries: a fixed-length summary limited to 10% of the length of the source, and a summary which was not limited in length. Future work includes quantitative and qualitative evaluation. In addition, our method used single words rather than phrases. These phrases, however, would be helpful to resolve ambiguity and reduce a lot of noise, i.e. yield much better accuracy. We plan to apply our method to phrase-based topic and event extraction, then turn to focus on the problem that how to form the actual summary.

Acknowledgments

The authors would like to thank the reviewers for their valuable comments. This work was supported by the Grant-in-aid for the Japan Society for the Promotion of Science(JSPS, No.11780258) and Tateisi Science and Technology Foundation.

References

J. Allan and J. Carbonell. 1997. The tdt pilot study corpus documentation. In *TDT.Study.Corpus, V1.3.doc.*

J. Allan and J. Carbonell. 1998. Topic detection and tracking pilot study: Final report. In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop.*

J. Allan and R. Papka. 1998. On-line new event detection and tracking. In *Proc. of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–45.

R. Barzilay and M. Elhadad. 1997. Using lexical chains for text summarization. In *Proc. of ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17.

R. Barzilay, K. R. McKeown, and M. Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proc. of 37th Annual Meeting of Association for Computational Linguistics*, pages 550–557.

B. Boguraev and C. Kennedy. 1997. Salience-based content characterization of text documents. In *Proc. of ACL Workshop on Intelligent Scalable Text Summarization*, pages 2–9.

E. Brill. 1992. A simple rule-based part of speech tagger. In *Proc. of the 3rd Conference on Applied Natural Language Processing*, pages 152–155.

J. Carbonell, Y. Yang, and J. Lafferty. 1999. CMU report on TDT-2: Segmentation, detection and tracking. In *Proc. of the DARPA Broadcast News Workshop.*

F. Fukumoto, Y. Suzuki, and J. Fukumoto. 1997. An automatic extraction of key paragraphs based on context dependency. In *Proc. of the 5th Conference on Applied Natural Language Processing*, pages 291–298.

H. Jing, R. Barzilay, K. R. McKeown, and M. Elhadad. 1998. Summarization evaluation methods: Experiments and analysis, intelligent text summarization. In *Proc. of 1998 American Association for Artificial Intelligence Spring Symposium*, pages 51–59.

J. Kupiec, J. Pedersen, and F. Chen. 1995. A trainable document summarizer. In *Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM journal*, 2(1):159–165.

I. Mani and E. Bloedorn. 1997. Multi-document summarization by graph search and matching. In *Proc. of the 15th National Conference on Artificial Intelligence*, pages 622–628.

I. Mani, T. Firmin, and B. Sundheim. 1999. The TIPSTER SUMMAC text summarization evaluation. In *Proc. of Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–85.

D. Marcu. 1997. From discourse structures to text summaries. In *Proc. of ACL Workshop on Intelligent Scalable Text Summarization*, pages 82–88.

K. R. McKeown and D. R. Radev. 1995. Generating summaries of multiple news articles. In *Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82.

R. Papka, J. Allan, and V. Lavrenko. 1999. UMASS approaches to detection and tracking at TDT2. In *Proc. of the DARPA Broadcast News Workshop.*

D. R. Radev and K. R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

G. Salton, J. Allan, C. Buckley, and A. Singhal. 1991. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, 164:1421–1426.

K. J. Sparck. 1993. What might be in a summary? In *Proc. of Information Retrieval93*, pages 9–26.

G. C. Stein, T. Strzalkowski, and G. B. Wise. 1999. Summarizing multiple documents using text extraction and interactive clustering. In *Proc. of the Pacific Association for Computational Linguistics1999*, pages 200–208.

T. Stralkowsik, G. C. Stein, and G. B. Wise. 1998. A text-extraction based summarizer. In *Proc. of Tipster Workshop.*

T. Stralkowski, G. C. Stein, and G. B. Wise. 1999. Getracker: A robust, lightweight topic tracking system. In *Proc. of the DARPA Broadcast News Workshop.*

S. Teufell and M. Moens. 1997. Sentence extraction as a classification task. In *Proc. of ACL Workshop on Intelligent Scalable Text Summarization*, pages 58–65.

K. Zechner. 1996. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In *Proc. of the 16th International Conference on Computational Linguistics*, pages 986–989.