Topic-Based Structuring of a Very Large-Scale News Video Corpus

Ichiro IDE   Hiroshi MO   Norio KATAYAMA   Shin'ichi SATOH
National Institute of Informatics
2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430, Japan
TEL: +81-3-4212-2585   FAX: +81-3-3556-1916
E-Mail: {ide,mo,katayama,satoh}@nii.ac.jp

Abstract

We introduce a topic-based inter-video structuring method that considers application to a very large-scale news video corpus as well as user interfaces that provide the users with the ability to efficiently browse through the corpus based on the topic structure. Although the proposed method is a multimedia-integrated method that refers to both text and image based information, this paper focuses on text-based topic segmentation and tracking / threading. First, topic segmentation is performed referring to inter-sentence keyword vector relations within a single video. Next, topic tracking and threading is performed referring to inter-topic keyword vector relations throughout the entire video corpus. Such analysis should reveal the underlying structure of the entire corpus which is not simply a large volume of unrelated data, but data full of rich information in the content-based relational structure itself. The segmentation method evaluated by applying the proposed method showed realistic ability. The proposed method was then applied to 555 daily news video (approximately 270 hours) obtained from a specific Japanese news program. Although detailed evaluation is yet to be done, the user interfaces showed good browsing ability for users to retrieve and track a topic thread of interest.

Introduction

Due to the recent advance in telecommunication technology, large amounts of videos have become available through various media. Such video data contain a broad range of human activities, which could be considered as valuable cultural and social heritage of the human race. From this viewpoint, news videos contain such information most densely. Nonetheless, building and analyzing a large-scale news video corpus has not been thoroughly examined until recently, due to limitation of computation power and storage size.

Motivated by such background issues, we have built an automatic news video archiving system where important topics can be searched and tracked easily. It automatically records video image, audio, and closed-caption text, and archives them in an Oracle database as shown in Figure 1. Up to now, approximately 270 hours (150GB of MPEG-1 and 925GB of MPEG-2 videos, and 11MB of closed-caption text data) have been obtained and archived from a specific Japanese daily news program.

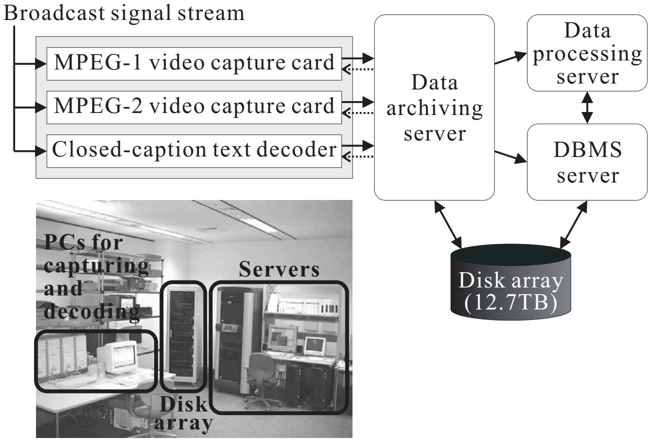
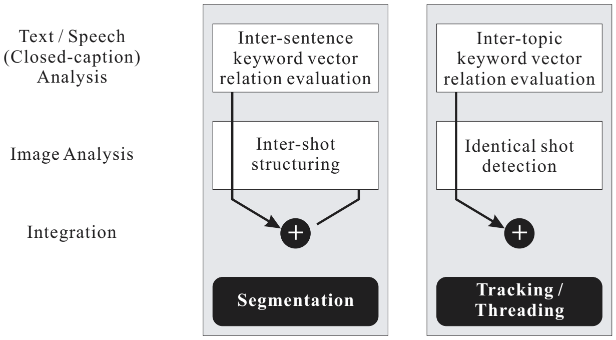


Figure 1: The automatic broadcast video archiving system.

In this paper, we introduce a topic-based news video corpus structuring method as well as user interfaces that provide the users with the ability to efficiently browse through the corpus based on the topic structure.

Topic segmentation and tracking is in general a part of the "Topic Detection and Tracking (TDT) task" defined by NIST (National Institute of Standards and Technology 2002). In TDT documents (Wayne 2000), a *topic* is defined as "A seminal *event* or activity, along with all directly related *events* and activities". Nonetheless, as the term "topic" generally stands for the TDT defined *event* in news video analysis, we will use the term "topic" to indicate an *event* and the term "(topic) thread" a *topic* in this paper.

The overall scheme of the proposed structuring method is shown in Figure 2. First, topic segmentation is performed referring to inter-sentence keyword vector relations and image-based inter-shot structuring within a single video. Next, topic tracking and threading is performed referring to inter-topic keyword vector relations and image-based identical shot detection (Satoh 2002) throughout the entire video corpus. Such integrated media analysis should compensate for the limits of single medium analyses. Note that although the scheme integrates both text and image based informa-


Text / Speech
(Closed-caption)
Analysis
Image Analysis
Integration
Inter-sentence keyword vector relation evaluation
Inter-shot structuring
Segmentation
Inter-topic keyword vector relation evaluation
Identical shot detection
Tracking / Threading

4. Set a window size w, and evaluate relations between w preceding and succeeding vectors at each sentence boundary. The relation at the boundary between sentences i and $i+1$ is defined as follows:

$$R_{S,w}(i) = \frac{\sum_{m=i-w+1}^{i} \vec{k_S}(m) \cdot \sum_{n=i+1}^{i+w} \vec{k_S}(n)}{\left|\sum_{m=i-w+1}^{i} \vec{k_S}(m)\right| \left|\sum_{n=i+1}^{i+w} \vec{k_S}(n)\right|}$$
$$(i = w, w+1, ..., i_{max} - w)$$

where $S = \{g, p, l, t\}$ and i_{max} stands for the number of sentences in a daily closed-caption text. We set $w = 1, 2, ..., 10$ in the following experiment[1].

5. Evaluate the following function to detect topic boundaries:

$$R(i) = \sum_{S=\{g,p,l,t\}} a_S \max_{w} R_{S,w}(i)$$

First, the maximum of $R_{S,w}(i)$ along the w axis is taken. According to a preliminary observation, although most boundaries were correctly detected regardless to the window size, there was a large number of over-segmentation. The over-segmentation had the following tendencies:

- Small w: Tends to over-segment long topics
- Large w: Tends to over-segment short topics

Thus, taking the maximum should compensate for over-segmentation at various window sizes. An example comparing these tendencies and the effect of taking the maximum is shown in Figure 4.

Next, weighted sum of relations evaluated in separate semantic attributes is defined to evaluate the overall relation. This approach is taken under the assumption that especially in news texts, certain semantic attributes should be more important than others when considering topic segmentation. Multiple linear regression analysis was applied to manually segmented training data (consists of 39 daily closed-caption texts, with 384 boundaries), which resulted in obtaining the following weights:

$$(a_g, a_p, a_l, a_t) = (0.23, 0.21, 0.48, 0.08) \qquad (1)$$

The obtained weights indicate that temporal noun sequences are not important in segmentation, and that locational / organizational noun sequences are especially important, which matches with our intuition.

Finally, if $R(i)$ does not exceed a certain threshold θ_{seg}, the boundary between sentences i and $i+1$ is judged as a topic boundary.

Figure 5 shows the process of procedures 3. to 5.

6. Create a keyword vector $\vec{K_S}$ for each detected topic, and re-evaluate the relations between adjoining topics i and $j(= i+1)$ by the following function to concatenate over-segmented topics:

$$R(i,j) = \sum_{S=\{g,p,l,t\}} a_S \frac{\vec{K_S}(i) \cdot \vec{K_S}(j)}{\left|\vec{K_S}(i)\right| \left|\vec{K_S}(j)\right|} \qquad (2)$$

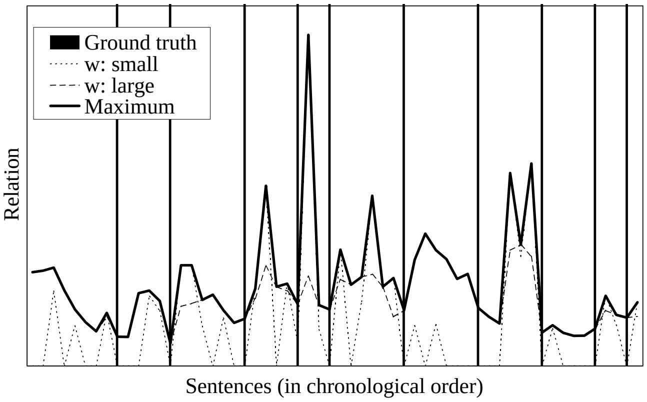


Figure 4: Over-segmentation tendencies according to window sizes.

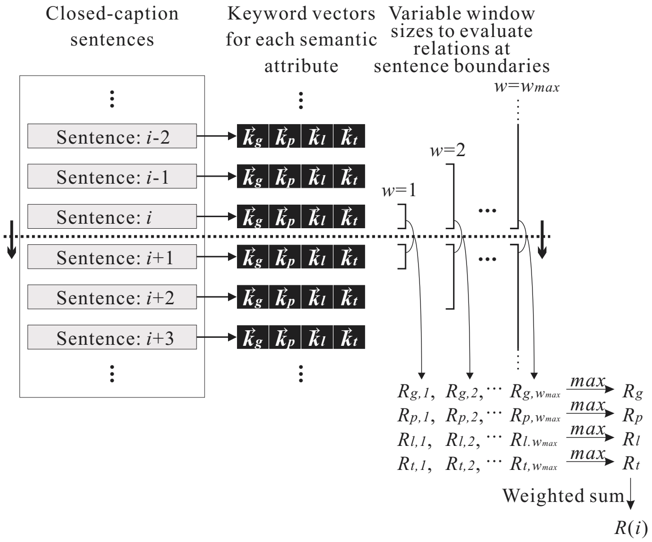

Closed-caption sentences
Keyword vectors for each semantic attribute
Variable window sizes to evaluate relations at sentence boundaries
$w=w_{max}$
Sentence: i-2
Sentence: i-1
Sentence: i
Sentence: i+1
Sentence: i+2
Sentence: i+3
$\vec{k_g}$ $\vec{k_p}$ $\vec{k_l}$ $\vec{k_t}$
w=2
w=1
$R_{g,1}$, $R_{g,2}$, ⋯ $R_{g,w_{max}}$ $\xrightarrow{max}$ R_g
$R_{p,1}$, $R_{p,2}$, ⋯ $R_{p,w_{max}}$ $\xrightarrow{max}$ R_p
$R_{l,1}$, $R_{l,2}$, ⋯ $R_{l,w_{max}}$ $\xrightarrow{max}$ R_l
$R_{t,1}$, $R_{t,2}$, ⋯ $R_{t,w_{max}}$ $\xrightarrow{max}$ R_t
Weighted sum
$R(i)$

Figure 5: Evaluation of relations at sentence boundaries.

As for a_S, the same weights as defined in (1) were used. If $R(i,j)$ does not exceed a certain threshold θ_{cat}, the adjoining topics are concatenated. This process is continued until no more concatenation occurs. After the concatenation, topics with only one sentence are excluded since they tend to be either noisy or relatively less important in a large-scale corpus.

Figure 6 shows the recall-precision curb of topic boundary detection derived from applying the proposed method to a test data set (14 days) independent from the training data set. The superiority of employing the weighted segmentation is shown by comparing it with the unweighted segmentation. θ_{seg} was defined as 0.17 where the sum of recall and precision of the weighted segmentation curb was maximal. The dotted curb is the recall-precision curb of topic boundary detection after over-segmented topic concatenation when θ_{seg} was set to 0.17. θ_{cat} was defined as 0.13 where the sum of recall and precision of the concatenated segmentation curb was maximal.

[1] The range was set reflecting the fact that 94% of the topics in a manually segmented data ranged from 1 to 10 sentences per topic.

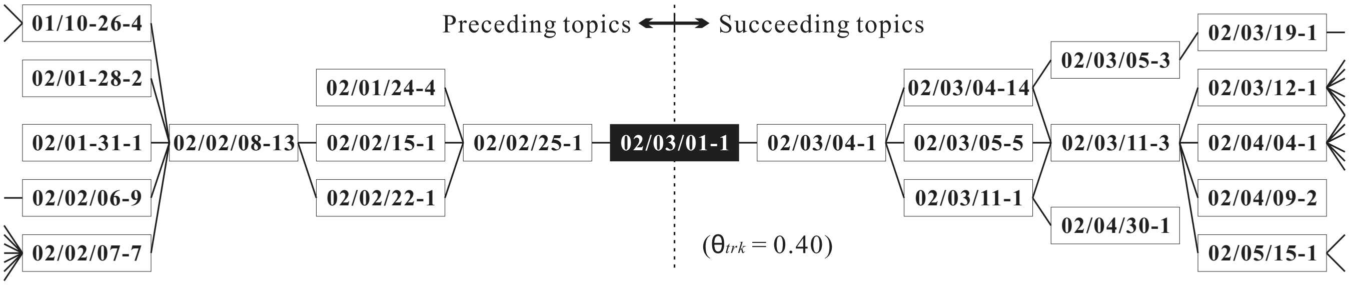


Figure 7: Example of topic threads extracted from the corpus. Topics are named in the following format: "Year/Month/Day-Topic#". Summaries of actual contents are shown in detail in Figure 11.

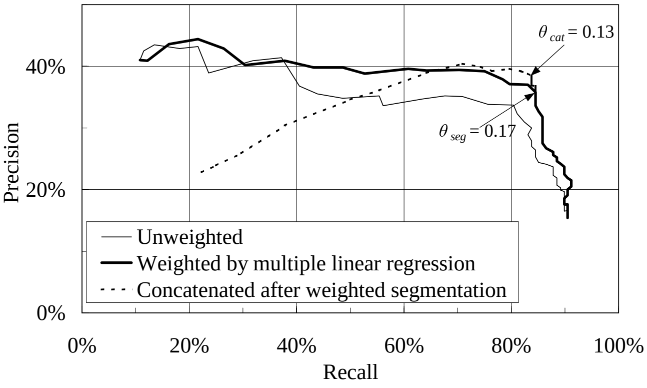


Figure 6: Comparison of segmentation ability, and definition of thresholds from the recall-precision curb.

Table 1: Evaluation of topic segmentation.

	Weighted	Unweighted
Extracted	140	137
(Strictly correct)	47	44
(Over-segmented)	56	53
(Incorrect)	37	40
Overlooked	27	33
Recall (Strict)	36.2%	33.8%
Recall (Accept over-seg.)	79.2%	74.6%
Precision (Strict)	33.6%	32.1%
Precision (Accept over-seg.)	73.6%	70.8%

After defining the thresholds as above, the procedure was applied to 555 daily closed-caption texts ranging from March 16, 2001 to December 13, 2002. This experiment resulted in extracting 6,696 topics with an average of 8.59 sentences per topic, after excluding topics with only one sentence.

We examined the segmentation ability by applying it to the above-mentioned test data set. Excluding topics with only one sentence, there were 130 manually extracted topics as a ground truth. The result is shown in Table 1. Strictly correct topics are those that both the beginning and the ending sentence boundaries completely match with the ground truth, over-segmented topics are those that begin and end within a ground truth topic segment. The proposed weighted method shows higher performance than the unweighted method that does not discriminate semantic attributes. Although strictly evaluated recall and precision is low, if over-segmentation could be accepted, the result could be considered as a realistic ability. Since even over-segmented topics consist of at least two sentences, they should be sufficient to represent the contents of topics to some extent.

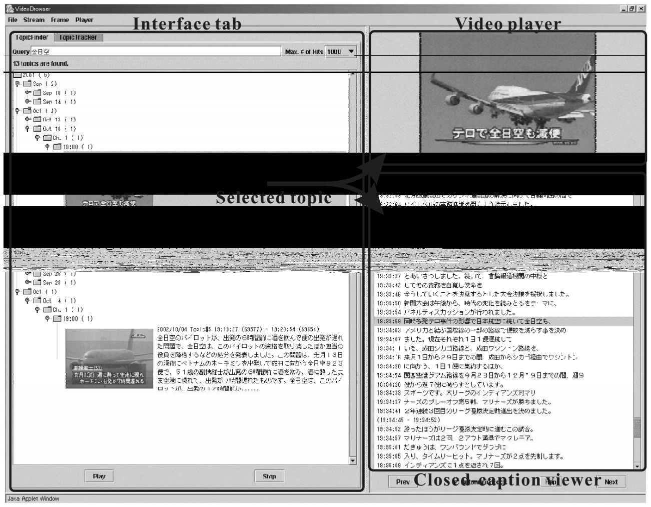
Topic tracking and threading Topics are tracked and thread structures are analyzed after the segmentation. In order to track topics, relations between any two topic pairs need to be evaluated beforehand. Equation 2 was used for this purpose, with the same weights as defined in Equation 1.

Topic threads are created by tracking up and down (in chronological order) topic after topic. Only links between topics with weighted relations ($R(i,j)$) larger than a certain threshold (θ_{trk}) are tracked. Figure 7 shows an example of actual topic threads involving topic #1 on March 1, 2002, extracted from the corpus. Only linked topics are considered as related, and topics placed in parallel are considered not related among themselves. The threads are structured so that always the topics linked to the left should be preceding, and those to the right should be succeeding ones. For reference, Figure 11 shows the summaries of actual contents of the topics shown in Figure 7.

An interactive topic-browsing interface

We implemented a topic browsing interface, namely the "Topic Browser" to provide the users with access to the topics based on the analyzed structure. It consists of two interfaces: the "Topic Finder" and the "Topic Tracker". Figure 8 shows the "Topic Bowser" interface. The left side of the browser displays each interface (inter-switchable by selecting the corresponding tab). The right side displays the video and the closed-caption text corresponding to the topic selected in the tab.

The "Topic Finder" (Figure 9) is a portal to the topic browsing interface. First, a user inputs a query term. Then the interface returns topics that contain the query term in chronological order. Each topic is represented by a thumbnail image (the first frame of the video segment corresponding to the segment) and an excerpt of a closed-caption text.



[redacted]

[redacted]

Figure 8: The “Topic Browser” interface.

A user browses through them and selects the most relevant one to his/her interest to set it as an initial topic for the tracking process.

Next, the “Topic Tracker” (Figure 10) is an interface to track up and down a topic thread interactively. Although the initial topic should be selected through the “Topic Finder”, the consecutive tracking is done solely within this interface. The interface displays relevant topic threads in chronological order, separated in two categories: preceding topics and succeeding topics, reflecting the topic thread structure as exemplified in Figure 7. Here, the terms “preceding” and “succeeding” represent the chronological relations with the selected topic. A user could either track anterior or posterior sequence of topics by selecting one of the presented threads, and setting it as the next selected topic. Such interactive tracking goes on topic after topic. Figure 10 shows the list of topic threads starting with the topic selected in Figure 9. Each thread represents a different subject related to the series of terrorist attacks on September 11 and its aftermath.

The “Topic Finder” may seem somewhat similar to conventional keyword-based news video retrieval, but the combination with the “Topic Tracker” narrows down the result according to the users’ interests and intentions which is important when dealing with a large-scale corpus. On the other hand, while narrowing down the results, the tracking could also be considered as a query expansion process. Having the two seemingly opposite characteristics, the tracking process provides a user with a topic thread that matches their interests and intentions to the maximal extent. Moreover, it reveals chronological transition, divergence, and merger of topic threads, which will be effective for thorough understanding of the story related to the initial query. Figure 12 shows an example of the tracking by one of the authors trying to recall the path to the attack to Afghanistan after the series of terrorist attacks on September 11. We found the above-mentioned features very effective and informative after trying to track down several topics of interest.

Similar approaches are taken in (Christel *et al.* 2002), although they fix the browsing interfaces within individual semantic attributes. On the contrary, our interface is designed to cope flexibly with various and vague interests and intentions of the users.

Conclusions

In this paper, we proposed a topic-based news video structuring method as a first step to reveal the underlying content-based structure in a very large-scale news video corpus. First, methods to segment and track / thread topics by closed-caption text analysis were described and evaluated. Next, interactive user interfaces that provide the users with the ability to browse the corpus based on topic structure were introduced. Although detailed evaluation is yet to be done, the interface showed good browsing ability for users to retrieve and track a topic thread of interest.

We will further investigate on achieving better topic segmentation quality by referring to image-based video structures to compensate for the limitation of text-based analysis (*e.g.* lack of distinctive keywords). Topic tracking will be improved by referring to graphically identical shots (Satoh 2002). This is based on the assumption that the same video footage is played over and over in news videos discussing the same topic, due to the limitation of data source. Here, we consider to employ identical shots, since current image-based similar shot retrieval technology is not robust enough to retrieve similar contents sufficiently. The tracking interface will also be improved by dynamically adjusting the weights used in Equation 2, depending on the user’s initial query terms and tracking history in a relevance feedback manner. Such adjustment should provide the user with related topics reflecting his/her intention.

References

Christel, M. G.; Hauptmann, A. G.; Wactler, H. D.; and Ng, T. D. 2002. Collages as dynamic summaries for news video. In *Proc. 10th ACM Intl. Conf. on Multimedia*, 561–569.

Ide, I.; Hamada, R.; Sakai, S.; and Tanaka, H. 1999. Semantic analysis of television news captions referring to suffixes. In *Proc. 4th Intl. Workshop on Information Retrieval with Asian Languages*, 37–42.

Kyoto Univ. 1999. Japanese morphological analysis system JUMAN version 3.61.

Merlino, A.; Morey, D.; and Maybury, M. 1997. Broadcast news navigation using story segmentation. In *Proc. 5th ACM Intl. Conf. on Multimedia*, 381–391.

National Institute of Standards and Technology. 2002. The 2002 topic detection and tracking (TDT2002) task definition and evaluation plan.

Satoh, S. 2002. News video analysis based on identical shot detection. In *Proc. 2002 IEEE Intl. Conf. on Multimedia and Expo*, volume 1, 69–72.

Takao, S.; Ogata, J.; and Ariki, Y. 2000. Topic segmentation of news speech using word similarity. In *Proc. ACM Multimedia 2000 Workshops*, 195–200.

Wactlar, H. D.; Christel, M. G.; Gong, Y.; and Hauptmann, A. G. 1999. Lessons learned from building a Terabyte digital video library. *IEEE Computer* 32(2):66–73.

Wayne, C. L. 2000. Multilingual topic detection and tracking: successful research enabled by corpora and evaluation. In *Proc. 2nd Intl. Conf. on Language Resources and Evaluation*, volume 3, 1487–1493.

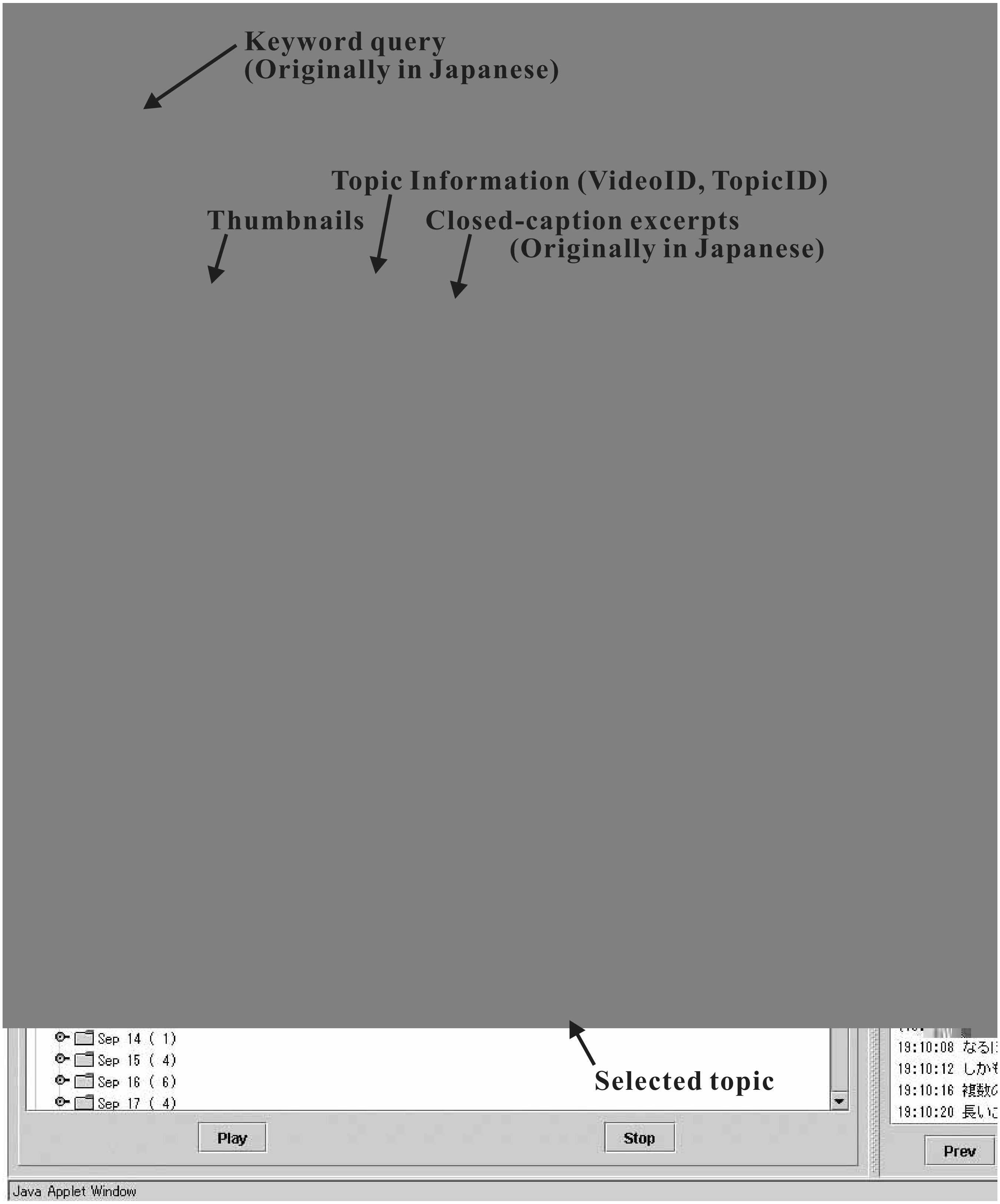


Figure 9: The “Topic Finder” interface. Result of a query “Bin Laden”.

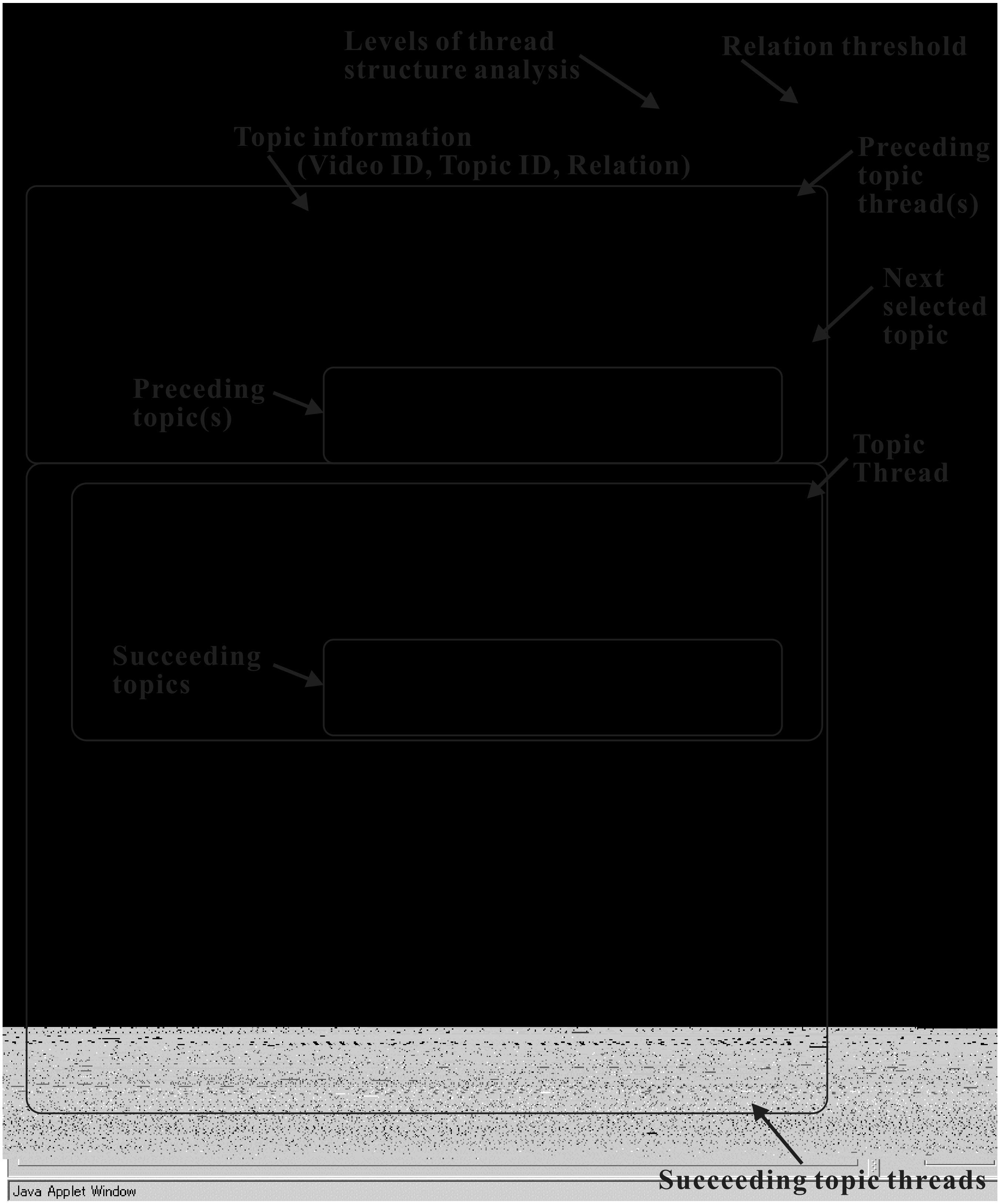


Figure 10: The “Topic Tracker” interface. [Preceding topic(s)] Thread 1: Report on the series of terrorist attacks on September 11. [Succeeding topics] Thread 1: Investigation on the attack; Thread 2: Travel warning by the Ministry of Foreign Affairs; Thread 3: Beginning of Ramadan in Egypt.

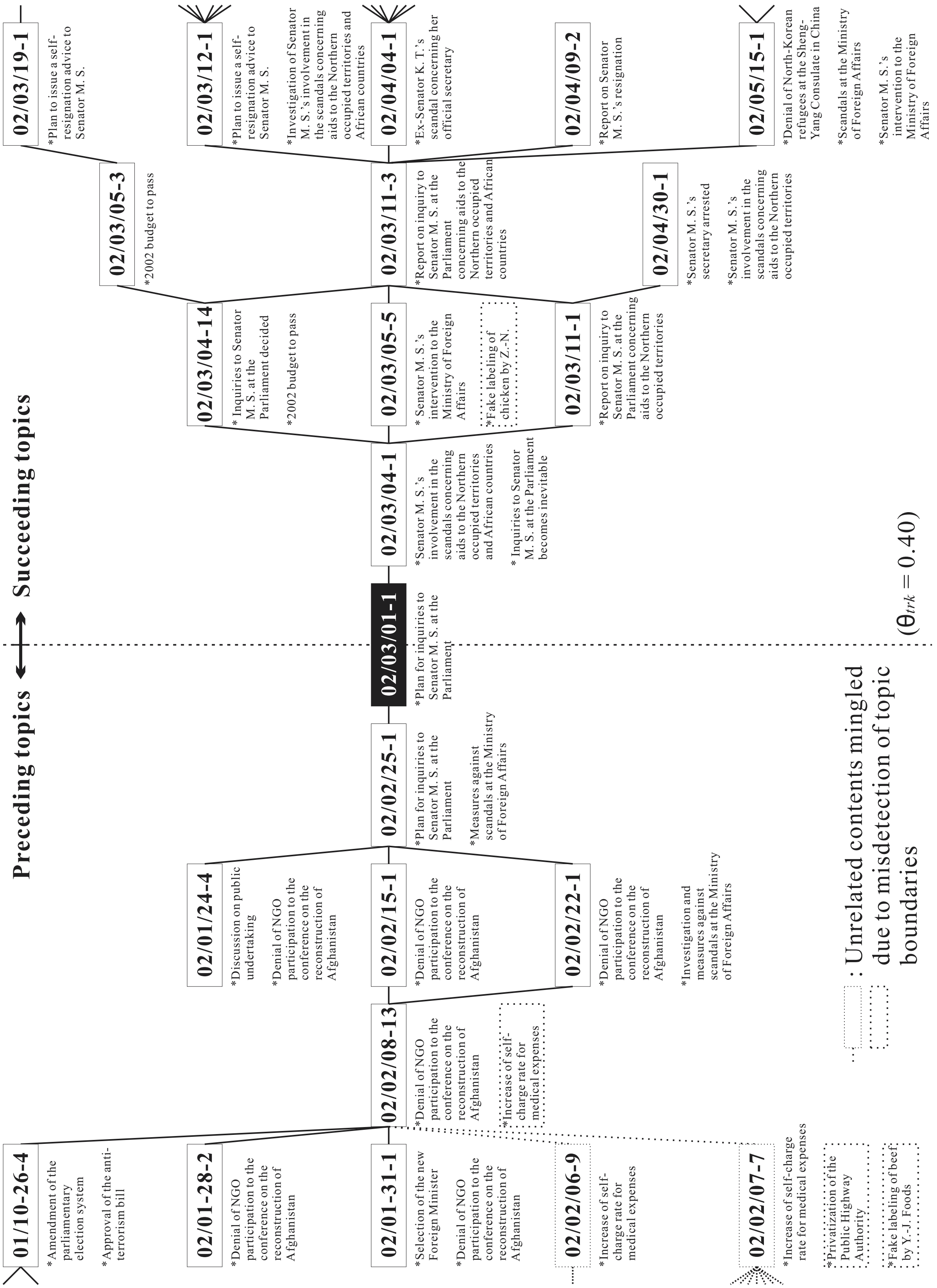
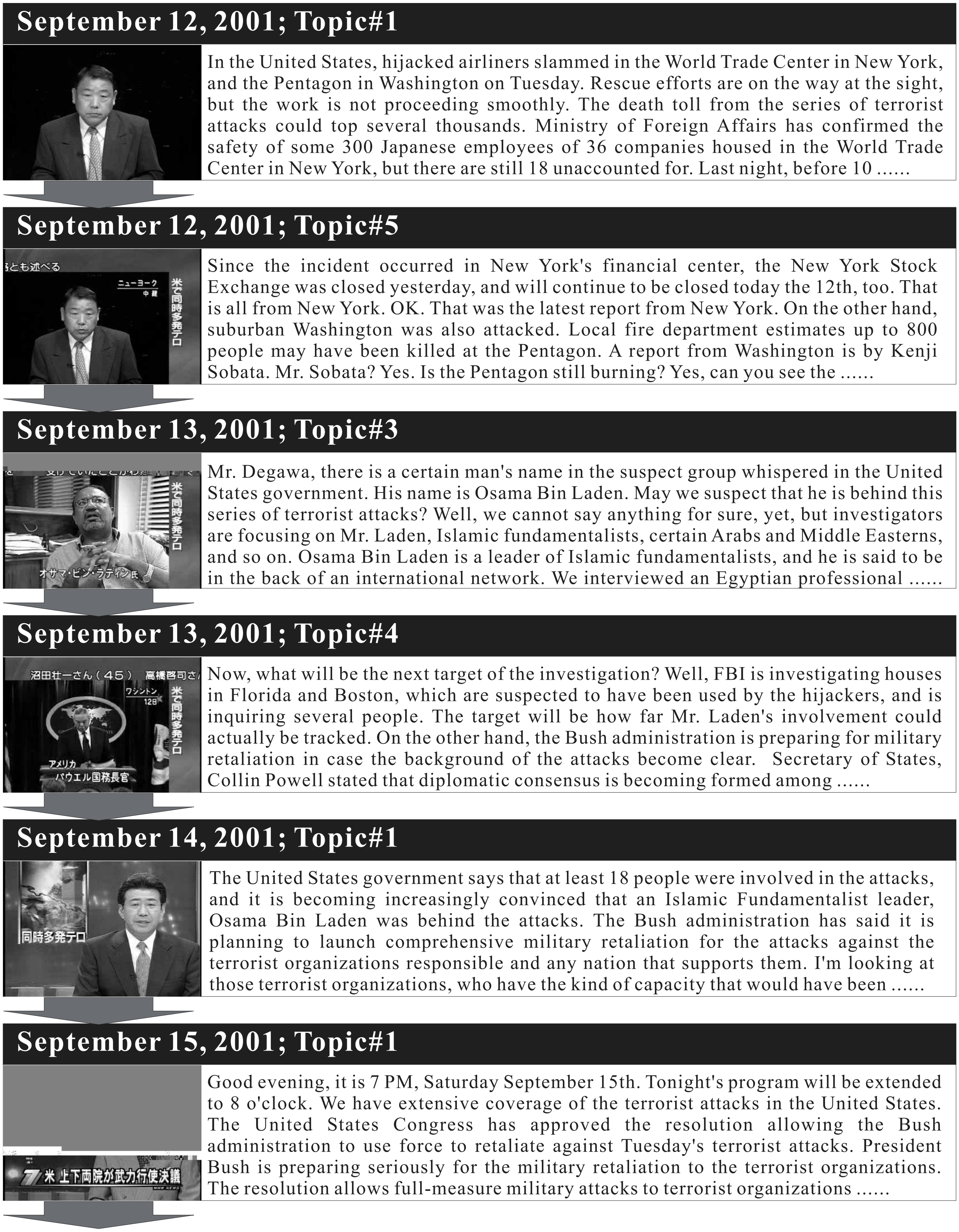

Preceding topics
Succeeding topics
01/10-26-4
*Amendment of the parliamentary election system
*Approval of the anti-terrorism bill
02/01-28-2
*Denial of NGO participation to the conference on the reconstruction of Afghanistan
02/01-31-1
*Selection of the new Foreign Minister
*Denial of NGO participation to the conference on the reconstruction of Afghanistan
02/02/06-9
*Increase of self-charge rate for medical expenses
02/02/07-7
*Increase of self-charge rate for medical expenses
*Privatization of the Public Highway Authority
*Fake labeling of beef by Y.-J. Foods
02/02/08-13
*Denial of NGO participation to the conference on the reconstruction of Afghanistan
*Increase of self-charge rate for medical expenses
02/01/24-4
*Discussion on public undertaking
*Denial of NGO participation to the conference on the reconstruction of Afghanistan
02/02/15-1
*Denial of NGO participation to the conference on the reconstruction of Afghanistan
02/02/22-1
*Denial of NGO participation to the conference on the reconstruction of Afghanistan
*Investigation and measures against scandals at the Ministry of Foreign Affairs
02/02/25-1
*Plan for inquiries to Senator M. S. at the Parliament
*Measures against scandals at the Ministry of Foreign Affairs
02/03/01-1
*Plan for inquiries to Senator M. S. at the Parliament
02/03/04-1
*Senator M. S.'s involvement in the scandals concerning aids to the Northern occupied territories and African countries
* Inquiries to Senator M. S. at the Parliament becomes inevitable
02/03/04-14
* Inquiries to Senator M. S. at the Parliament decided
*2002 budget to pass
02/03/05-5
* Senator M. S.'s intervention to the Ministry of Foreign Affairs
*Fake labeling of chicken by Z.-N.
02/03/11-1
*Report on inquiry to Senator M. S. at the Parliament concerning aids to the Northern occupied territories
02/03/05-3
*2002 budget to pass
02/03/11-3
*Report on inquiry to Senator M. S. at the Parliament concerning aids to the Northern occupied territories and African countries
02/04/30-1
*Senator M. S.'s secretary arrested
*Senator M. S.'s involvement in the scandals concerning aids to the Northern occupied territories
02/03/19-1
*Plan to issue a self-resignation advice to Senator M. S.
02/03/12-1
*Plan to issue a self-resignation advice to Senator M. S.
*Investigation of Senator M. S.'s involvement in the scandals concerning aids to the Northern occupied territories and African countries
02/04/04-1
*Ex-Senator K. T.'s scandal concerning her official secretary
02/04/09-2
*Report on Senator M. S.'s resignation
02/05/15-1
*Denial of North-Korean refugees at the Sheng-Yang Consulate in China
*Scandals at the Ministry of Foreign Affairs
*Senator M. S.'s intervention to the Ministry of Foreign Affairs
: Unrelated contents mingled due to misdetection of topic boundaries
$(\theta_{trk} = 0.40)$

Figure 11: Detailed example of topic threads extracted from the corpus. Topics are named in the following format: "Year/Month/Day-Topic#".



September 12, 2001; Topic#1

In the United States, hijacked airliners slammed in the World Trade Center in New York, and the Pentagon in Washington on Tuesday. Rescue efforts are on the way at the sight, but the work is not proceeding smoothly. The death toll from the series of terrorist attacks could top several thousands. Ministry of Foreign Affairs has confirmed the safety of some 300 Japanese employees of 36 companies housed in the World Trade Center in New York, but there are still 18 unaccounted for. Last night, before 10 ......

September 12, 2001; Topic#5

Since the incident occurred in New York's financial center, the New York Stock Exchange was closed yesterday, and will continue to be closed today the 12th, too. That is all from New York. OK. That was the latest report from New York. On the other hand, suburban Washington was also attacked. Local fire department estimates up to 800 people may have been killed at the Pentagon. A report from Washington is by Kenji Sobata. Mr. Sobata? Yes. Is the Pentagon still burning? Yes, can you see the ......

September 13, 2001; Topic#3

Mr. Degawa, there is a certain man's name in the suspect group whispered in the United States government. His name is Osama Bin Laden. May we suspect that he is behind this series of terrorist attacks? Well, we cannot say anything for sure, yet, but investigators are focusing on Mr. Laden, Islamic fundamentalists, certain Arabs and Middle Easterns, and so on. Osama Bin Laden is a leader of Islamic fundamentalists, and he is said to be in the back of an international network. We interviewed an Egyptian professional ......

September 13, 2001; Topic#4

Now, what will be the next target of the investigation? Well, FBI is investigating houses in Florida and Boston, which are suspected to have been used by the hijackers, and is inquiring several people. The target will be how far Mr. Laden's involvement could actually be tracked. On the other hand, the Bush administration is preparing for military retaliation in case the background of the attacks become clear. Secretary of States, Collin Powell stated that diplomatic consensus is becoming formed among ......

September 14, 2001; Topic#1

The United States government says that at least 18 people were involved in the attacks, and it is becoming increasingly convinced that an Islamic Fundamentalist leader, Osama Bin Laden was behind the attacks. The Bush administration has said it is planning to launch comprehensive military retaliation for the attacks against the terrorist organizations responsible and any nation that supports them. I'm looking at those terrorist organizations, who have the kind of capacity that would have been ......

September 15, 2001; Topic#1

Good evening, it is 7 PM, Saturday September 15th. Tonight's program will be extended to 8 o'clock. We have extensive coverage of the terrorist attacks in the United States. The United States Congress has approved the resolution allowing the Bush administration to use force to retaliate against Tuesday's terrorist attacks. President Bush is preparing seriously for the military retaliation to the terrorist organizations. The resolution allows full-measure military attacks to terrorist organizations ......

Figure 12: Detailed example of topic tracking: The thread was selected by one of the authors. Starting from the report on the attack, the process of the investigation and the path to the attack to Afghanistan could be understood.