Visualizing the Structure of a Large-Scale News Video Corpus Based on Topic Segmentation and Tracking

Ichiro IDE 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
{ide,katayama,satoh}@nii.ac.jp

ABSTRACT

We propose a topic-based inter-video news video corpus structuring method and a visual interface to efficiently browse through a very large-scale corpus. Such inter-video structuring was not deeply sought in previous works. The topic-based structure is analyzed by closed-caption text analysis; topic segmentation and tracking. The visual interface provides the ability to 1)search and select a topic by query terms and 2)track the selected topic interactively, referring to the text analysis results. Although topic retrieval is somewhat similar to conventional video retrieval methods, the combination with topic tracking makes it remarkably easy to narrow down the results that match a user's interest and moreover visualize the latent structure, which is exceptionally important when browsing through a very large-scale video corpus.

1. INTRODUCTION

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

Motivated by such background issues, we have built an automatic news video archiving system to extract high-level knowledge from a large-scale news video corpus. It automatically records video image, audio, and closed-caption text, and archives them in a Oracle database. Up to now, we have archived approximately 250 hours (150GB of MPEG-1 and 900GB of MPEG-2 videos, and 9.5MB of closed-caption text data) from a specific Japanese daily news program. The closed-caption text is tagged with time stamps reflecting the timing of its appearance.

In this paper, we introduce a topic-based news video corpus structuring method and a visual interface to efficiently access the structured corpus. Concretely, daily news videos are segmented into topics, and tracked throughout the entire corpus based on closed-caption text analysis. A visual interface was also built to enable interactive topic tracking.

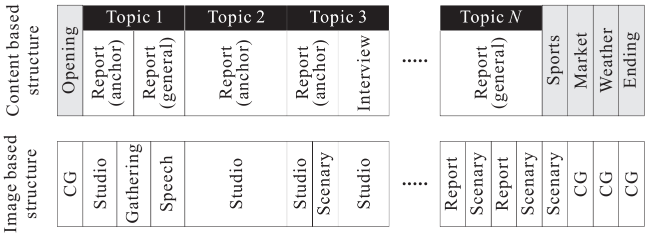

Figure 1: Content and image-based structure of news video.

Topic segmentation and tracking is in general a part of the "Topic Detection and Tracking (TDT) task" defined by NIST. In TDT documents [1, 9], 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 both a *topic* and an *event* in this paper.

First, structure analysis, *i.e.* topic segmentation and tracking methods are described in Section 2, and next the visual interface is introduced in Section 3. Finally, Section 4 summarizes the results and future works.

2. STRUCTURING A NEWS VIDEO CORPUS

2.1 Structure of news video

The general structure of a news video is as shown in Figure 1. In the case of news videos, image and content-based structures are roughly comparable. Thus majority of previous works on structure analysis of news videos has focused on analyzing the image-based structure to subsequently acquire the content-based structure, under the assumption that there are rules that link them. Although this approach works well to some extent, the rules are not always applicable and also depends highly on the editing and designing policy of each program.

Moreover, such structure analyses are limited within a single video, and inter-video structure has not been deeply sought. Since we deal with a very large-scale corpus, content-based inter-video structure analysis (*i.e.* topic tracking) be-

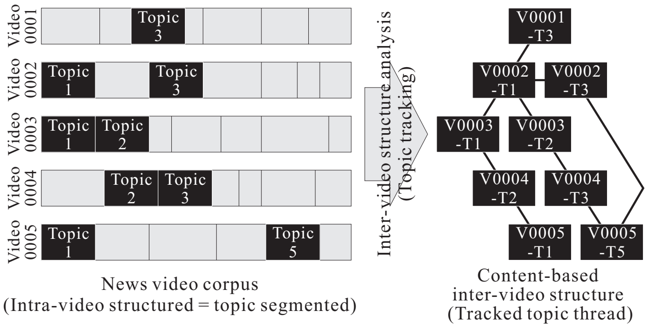

Figure 2: Content-based inter-video structuring.

comes exceptionally important. Such analysis will elucidate the latent content-based structure of the entire corpus which is not simply a huge volume of unrelated data, but data full of rich information in the structure itself. Although there are several works that deal with news video corpora of a similar size to ours such as [5] and the Informedia News-on-Demand project [8], they do not consider inter-video structures.

Figure 2 shows an example of a content-based inter-video structure, *i.e.* a tracked topic thread, in the corpus. Extracting such topic threads throughout the entire corpus elucidates the latent content-based structure that does not emerge from simple intra-video analysis.

As a first step to analyze the structure of the corpus, we analyze closed-caption texts in order to enable topic-based structuring (segmentation and tracking). This is based on the assumption that text contains higher level semantics compared to information easily acquirable from image. Note that although this paper concentrates on text-based structuring, it is a starting point of a multimedia-integrated analysis. Image-based analyses such as topic tracking based on image features [6] will be integrated in future works.

2.2 Topic segmentation and tracking

2.2.1 Related works

Various approaches have been proposed and evaluated in the past TDT workshops, but they are rather tuned to the tasks, and mostly have not been applied to a large-scale real-world data.

As other works, Fukumoto and Suzuki [2] proposed a tracking method that distinguishes *topics* and *events* using a thesaurus, which becomes important to analyze the sub-topic structure.

All the above-mentioned works were applied to English (and partly Chinese) transcript texts, so they are not directly applicable to Japanese news texts, due to difference in linguistic characteristics. Among the few works targeting Japanese news texts, Takao *et al.* [7] proposed a method that segments speech transcription texts.

We do refer to these methods, but since most of them were evaluated with a relatively small data set, we will adopt original methods to deal with a very large-scale corpus.

2.2.2 Topic segmentation

Topic boundaries are detected by applying the following procedure to daily closed-caption texts:

1. Concatenate original time-stamped text lines into single sentences. Sentences are concatenated by detecting a period in the text.

2. Apply morphological analysis to each sentence and extract noun sequences. JUMAN [4], a Japanese morphological analysis software is used.

3. Apply semantic analysis to noun sequences, and generate a keyword vector for each semantic class. Semantic analysis is done by a suffix-based method [3], which classifies noun sequences to 1)general, 2)personal, 3)locational/organizational, or 4)temporal. Thus, four keyword vectors are generated from a sentence: $\vec{k_g}, \vec{k_p}, \vec{k_l}, \vec{k_t}$, respectively. The vectors have keywords (noun sequences) as indices and frequencies as values. Note that this analysis method classifies not only proper noun sequences (*e.g.* Prime Minister Koizumi) but also common noun sequences (*e.g.* fire fighter).

4. Set a window size w, and evaluate the 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 are correctly detected regardless to the window size, there is a large number of over-segmentation. The over-segmentation had the following tendencies:

- Small window size: Tend to over-segment topics longer than its size.
- Large window size: Tend to over-segment topics shorter than its size.

Thus, taking the maximum should compensate for over-segmentation at various window sizes.
Next, weighted sum of relations evaluated in separate semantic attributes defines the overall relation. This approach is taken under the assumption that especially in news texts, certain semantic attributes should be more important than others to evaluate the relations. Multiple linear regression analysis was applied to manually segmented training data (consists of 39 daily closed-caption text, with 384 boundaries), and obtained the following weights:

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

Finally, if $R(i)$ does not exceed a certain threshold (set to 0.17), the boundary between sentences i and $i+1$ is judged as a topic boundary.

[1] 94% of the topics in a manually segmented data ranged from 1 to 10 sentences per topic.

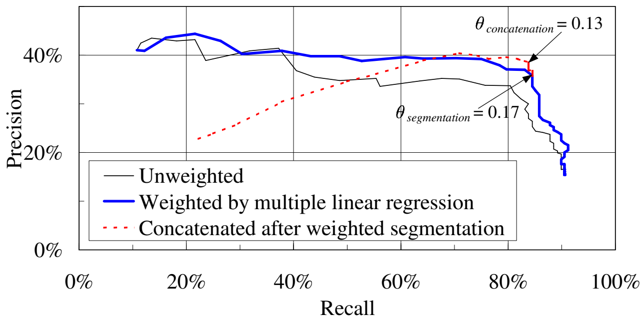

Figure 3: Comparison of segmentation ability.

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:

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

As for a_S, the same weights as defined in (1) were used.
If $R(i,j)$ does not exceed a certain threshold (set to 0.13), the adjoining topics are concatenated. This process is continued until no more concatenation occurs. After the concatenation, topics with only one sentence are deleted since they tend to be either noisy or relatively less important when structuring a large-scale corpus.

Figure 3 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. Superiority of employing the weighted segmentation is shown by comparing it with the unweighted segmentation. The weights derived in (1) indicates that temporal noun sequences are not important in segmentation, and that locational noun sequences are especially important, which matched our assumption. The segmentation threshold 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 the segmentation threshold was 0.17. The concatenation threshold was defined as 0.13 where the sum of recall and precision of the concatenated segmentation curb was maximal.

We applied the procedure to 481 daily closed caption texts ranging from March 16, 2001 to September 24, 2002. This experiment resulted in extracting 5,864 topics with an average of 8.53 sentences per topic.

We examined the effectiveness of the proposed method by applying it to the above-mentioned test data set. Excluding topics with only one sentence, there were 130 manually extracted topics as the ground truth. The result is shown in Table 1. Strictly correct topics are those that both the beginning and the ending point completely match with the ground truth, over-segmented topics are those that begin and end within a ground truth topic. The proposed weighted method shows higher performance than the unwieghted method that does not discriminate noun attributes. Although strictly evaluated recall and precision is low, if over-segmentation could be accepted, the result shows 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.

Table 1: Evaluation of topic segmentation.

	Weighted	Unweighted
Extracted	140	137
(Strictly correct)	47	44
(Over-segmented)	56	53
(Incorrect)	37	40
Overlooked	20	28
Recall (Strict)	36.2%	33.8%
Precision (Strict)	33.6%	32.1%
Recall (Accept over-seg.)	79.2%	74.6%
Precision (Accept over-seg.)	73.6%	70.8%

2.2.3 Topic tracking

Topics are tracked after the segmentation. In order to track topics, relation between two topics needs to be evaluated. Equation (2) was used for this purpose, with the following weights:

$$(a_g, a_p, a_l, a_t) = (0.25, 0.25, 0.25, 0.25)$$

The weights defined in (1) were not used, since relations in segmentation and tracking should have different roles. Although the weights are currently set to an equal value, we are considering to dynamically adjust them depending on the user's initial query terms and tracking history. Such adjustment should provide a user with related topics reflecting his/her intention.

3. VISUALIZING THE TOPIC-BASED STRUCTURE

We implemented a topic browsing interface, namely "Topic Browser" to visualize the topic-based structure analyzed in Section 2. It consists of two different interfaces: "Topic Finder" (Figure 4), and "Topic Tracker" (Figure 5).

The "Topic Finder" interface is somewhat similar to conventional keyword-based news video retrieval, but the combination with the "Topic Tracker" enables to track and narrow down the contents of the retrieval, which is exceptionally important when browsing through a very large-scale corpus. For example, a query "Bin Laden" may return hundreds of topics that contain the name, but with a broad variety of contents, such as "Attacks on September 11", "Bombing in Afghanistan", "Travel safety", and so on. Selecting a specific topic narrows down the query to a small set of the entire corpus, which frees a user from browsing through a tremendous amount of mostly uninteresting topics, and moreover provides the user an interface to view the latent content-based structure of the corpus.

3.1 Topic finding interface: "Topic Finder"

The "Topic Finder" shown in Figure 4 is a portal to the topic browsing interface. First, a user types in a query term (*e.g.* Bin Laden). Then the interface returns topics that contain the query term in chronological order. Each topic segment is represented by a thumbnail (the first video frame of the topic segment) and the first several hundred characters of the closed-caption text of the topic segment. The user can browse through them and select the most relevant one to his/her interest. The right side of the browser displays the video and the closed-caption text corresponding to the selected topic.

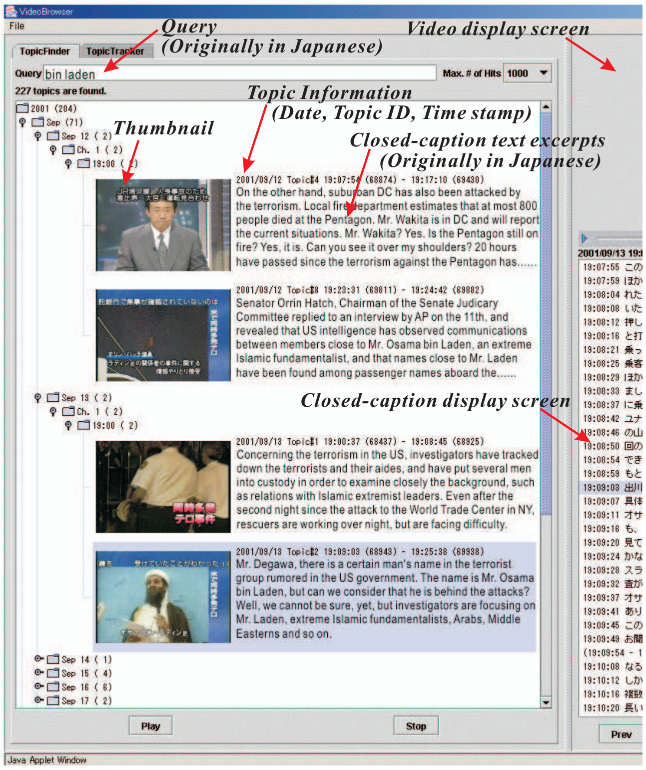

Figure 4: "Topic Finder" interface.

Figure 5: "Topic Tracker" interface.

3.2 Topic tracking interface: "Topic Tracker"

The "Topic Tracker" shown in Figure 5 is an interface to track a topic interactively. Although the initial topic should be selected through the "Topic Finder", the consecutive tracking is done solely within this interface.

The interface displays most relevant topics in relational order, separated in two categories: past news and future news. Here, the terms "past" and "future" represent the time relations with the selected topic. The user could either track anterior or posterior sequence of events concerning the selected topic by selecting a topic among the related ones, and setting it as the next selected topic. Such interactive tracking goes on until the user understands the details of a series of similar topics.

We have found that the tracking interface is very informative in order to follow the transition of a specific topic.

4. CONCLUSION

In this paper, we proposed a topic-based news video structuring method, as a first step to elucidate latent structures within a very large-scale news video corpus. First, methods to segment and track topics by closed-caption text analysis were described and evaluated. Next, visualization of the topic-based structure reflecting the segmentation and tracking was introduced. Although detailed evaluation is yet to be done, the visualized interface showed good browsing ability for users to retrieve and track a topic of interest. We will further investigate on achieving better topic segmentation quality employing better learning methods, and also dynamically adjusting the tracking process in order to adapt to the user's interest. User interface should also be improved in the visual interface to enable smoother tracking.

5. REFERENCES

[1] The 2002 topic detection and tracking (TDT2002) task definition and evaluation plan, May 2002.

[2] F. Fukumoto. Event tracking based on domain dependency. In *Proc. 23rd Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 57–64, July 2000.

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

[4] Kyoto University. Japanese morphological analysis system JUMAN version 3.61, May 1999.

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

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

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

[8] H. D. Wactlar, A. G. Hauptmann, and M. J. Witbrock. Informedia News-on-Demand: using speech recognition to create a digital video library. Technical Report CMU-CS-98-109, Carnegie Mellon University, March 1998.

[9] C. L. Wayne. Multilingual topic detection and tracking: successful research enabled by corpora and evaluation.