Labeling News Topic Threads with Wikipedia Entries

Tomoki Okuoka*, Tomokazu Takahashi†, Daisuke Deguchi*, Ichiro Ide*‡ and Hiroshi Murase*
* *Graduate School of Information Science, Nagoya University*
1 Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
E-mail: okuoka@murase.m.is.nagoya-u.ac.jp, {ddeguchi, ide, murase}@is.nagoya-u.ac.jp
† *Department of Economics and Information, Gifu Shotoku Gakuen University*
1-38 Naka Uzura, Gifu, 500-8288, Japan
E-mail: ttakahashi@gifu.shotoku.ac.jp
‡ *Research Organization of Information and Systems, National Institute of Informatics*
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

***Abstract*—Wikipedia is a famous online encyclopedia. However most Wikipedia entries are mainly explained by text, so it will be very informative to enhance the contents with multimedia information such as videos. Thus we are working on a method to extend information of Wikipedia entries by means of broadcast videos which explain the entries. In this work, we focus especially on news videos and Wikipedia entries about news events. In order to extend information of Wikipedia entries, it is necessary to link news videos and Wikipedia entries. So the main issue will be on a method that labels news videos with Wikipedia entries automatically. In this way, explanations could be more detailed with news videos can be exhibited, and the context of the news events should become easier to understand. Through experiments, news videos were accurately labeled with Wikipedia entries with a precision of 86% and a recall of 79%.**

***Keywords*-Wikipedia; news video; video archive; indexing**

I. Introduction

Advance in data-storage technology has enabled the online archiving of massive amount of broadcast video. Broadcast videos are produced by professionals, and thus usually more reliable than videos produced by amateurs such as those posted on a video sharing website. Especially, news videos are important as a chronological record of events. It will be worth reusing those videos, for example, as video-on-demand services. Meanwhile, Wikipedia is a famous online encyclopedia, where consistent explanations on many items including news events are described by Internet users. However, usually the explanations are summarized, and detailed descriptions are dropped. Moreover, most items are mainly explained by text, so it is necessary to enhance them with multimedia information such as videos or images. Thus we are working on a method to extend information of Wikipedia entries by broadcast videos which explain the entries. In this work, we focus on linking news videos and Wikipedia entries about news events. Hereafter, a "news story" is defined as the minimum semantic unit in a news video covering one event, as defined in the Topic Detection and Tracking (TDT) definition [2].

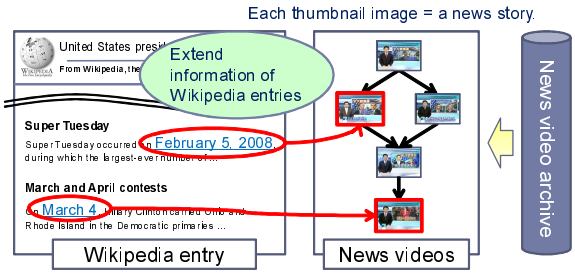

Figure 1. Example of extending information of Wikipedia entries by broadcast videos.

Figure 1 is an example of extending information of Wikipedia entries by using news stories in a video archive. First, related news stories are chained to browse them in time order. Next, date information in Wikipedia texts is associated with the news stories which were broadcast on the date and explain the same event as in the Wikipedia entries; here, the similarity of text information is evaluated. Thus, more detailed explanations can be exhibited by news stories, and the context of the news events should become easier to understand. In order to extend information of Wikipedia entries by news stories, it is necessary to label news stories with each Wikipedia entry. Thus, in this paper, we will propose a method of labeling news stories with Wikipedia entries.

II. Related Works

There are many works on visualizing or complementing the information of Wikipedia. Chan et al. have proposed "Vispedia", a Web-based system that combines visualization and data integration of Wikipedia entries [1].

On the other hand, there are many works on structure analysis and browsing of news videos. Most of them make use of closed caption (hereafter, CC), and analyze relations between videos based on the similarity of the text information. Snoek et al. have proposed "MediaMill" that realizes semantic exploration of news video archives [4]. Meanwhile

978-0-7695-3890-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ISM.2009.67

IEEE computer society

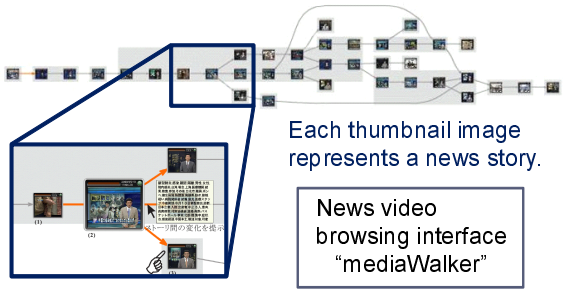

Figure 2. Example of a topic thread structure visualized in the “mediaWalker” interface [3].

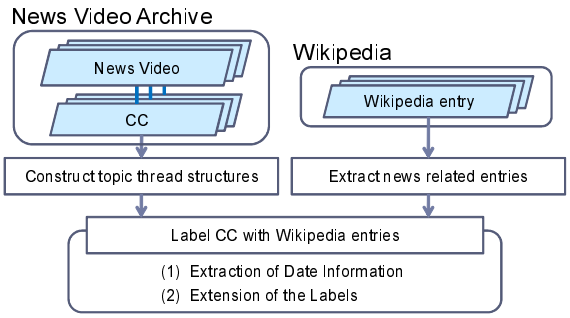

Figure 3. Flow chart of the proposed method.

Ide et al. have proposed a structure analysis method and a browsing interface of a news video archive based on a “topic thread structure”, which is constructed by chaining strongly related news stories [3]. An example of a topic thread structure and the browsing interface are shown in Figure 2. By using a topic thread structure, we can browse related news stories in time order.

III. Labeling News Topic Threads with Wikipedia Entries

A. Overview

The main issue of this paper is on a method that labels news stories with Wikipedia entries. The flow chart of the proposed method is shown in Figure 3. For the news videos, topic thread structures are constructed by analyzing the CC accompanying them. Here, the topic segmentation and threading methods proposed in [3] are used. On the other hand, for Wikipedia entries, entries related to news events (hereafter, news related entries) are extracted. By using both outputs, CC of news stories are labeled with Wikipedia entries by evaluating their similarity, and as a result, news videos are labeled with Wikipedia entries. Here, date information is extracted from the text of news related entries to limit the period of the similarity evaluation. After that, the labels are extended along the topic thread structure. Thus, the labeling coverage should improve.

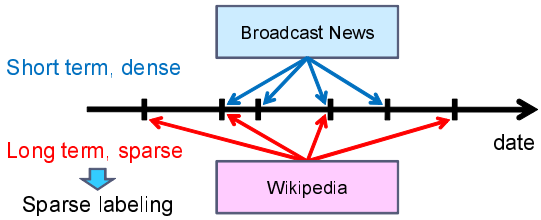

Figure 4. Difference in the description of news between broadcast news and Wikipedia.

B. Extraction of News Related Entries from Wikipedia

In order to avoid incorrect labeling as much as possible, and also to cope with the data size, news related entries are extracted from Wikipedia by taking advantage of the following features. First, a link to “Wikinews[1]” which is a sister project of Wikipedia, often exists in news related entries. Second, they often cite an article in a general news site for reference to prove the credibility of the description.

C. Labeling News Stories with Wikipedia Entries

In order to label news stories with Wikipedia entries, the text-based similarities between the corresponding CC and Wikipedia entries are evaluated. First, texts of both CC and Wikipedia entries are decomposed into morphemes, and the noun frequency vectors are made. Next, the cosine similarity between the vectors is calculated. If it exceeds a threshold, the news story is labeled with the Wikipedia entry.

1) Extraction of Date Information: Evaluating similarities of all CC and all news related entries decreases the labeling accuracy because of incorrect labeling. To avoid this, they are evaluated in the following way: The timing of the occurrence of an event is important in news. In order to restrict the period for similarity evaluation, date information (year, month, day) is extracted in the texts of news related entries.

However, a problem exists in this method. There is a difference in the description of news between broadcast news and Wikipedia. Figure 4 is a conceptual diagram that shows the difference. A news video often picks up a certain topic during a short term. On the other hand, Wikipedia tends to explain a news topic concisely from the beginning to the end. Its description, however, does not always accompany detailed date information. Therefore, if videos are labeled only with date information, the labels will be sparse.

2) Extension of the Labels: To compensate for the sparse labeling problem, the labels are extended along the topic thread structure [3]. Figure 5 shows the conceptual diagram.

For each Wikipedia entry, the labels are extended along the topic thread structure. Since the topic thread structure is constructed by chaining related news stories in time order, it can be considered that adjacent nodes along a topic thread

[1]http://ja.wikinews.org/wiki/

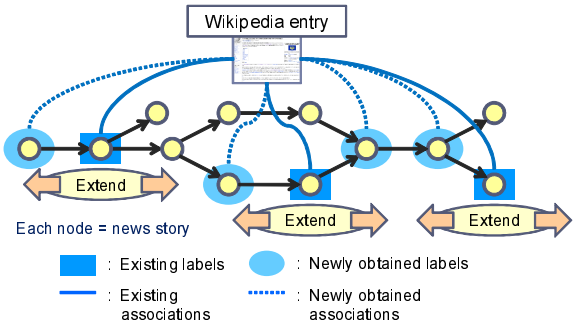

Figure 5. Extending the existing labels along a topic thread structure.

are similar to each other. In fact, a node adjacent to a certain node will be the most similar ones from the point of view of both semantics and time. So, the similarity of the CC of news adjacent to a labeled node on a topic thread structure is evaluated with the focused Wikipedia entry. If it exceeds a threshold, the news story is newly labeled with the Wikipedia entry. This operation is recursively applied along the topic thread sturcture until it encounters a node with low similarity, or a node that has already been labeled with the same entry.

IV. Experiments and Evaluation

For the broadcast news, Japanese news video data accompanied with CC from NHK News7 broadcast during January 1, 2007 and June 30, 2008 were used. They were segmented into stories by the method described in [3]. For Wikipedia, the official archive of the Japanese version[2] was used. The number of all entries at this time was 1,053,561. By the method explained in III-B, 1,645 news related entries were extracted. MeCab[3] was used for Japanese morphological analysis. In this Section, the labeling accuracy is evaluated in IV-A. Next, the number of labeled news stories with a same Wikipedia entry is investigated in IV-B. Finally, an example of extending information of Wikipedia by using video groups is shown in IV-C.

A. Experiment 1 : Labeling Accuracy

We evaluated the labeling accuracy by manually examining CC labeled with each Wikipedia entry. The evaluation was performed against three Wikipedia entries. We defined a correct labeling as when the contents in the CC covers a Wikipedia entry clearly and concretely. For evaluating the precision, all the contents of labeled CC were examined, and for the recall, the ground truth was manually made within a period of three months. For comparison, the following three methods were tested.

Method 1:
: Date information was not used to restrict the period for labeling, and the labels were not extended along the topic thread structure.

Method 2:
: Date information was used to restrict the period for labeling, but the labels were not extended along the topic thread structure.

Method 3 (Proposed method):
: Date information was used to restrict the period for labeling, and the labels were extended along the topic thread structure.

[2]The data was recorded on November 27, 2008.
[3]http://mecab.sourceforge.net/

Table I
Labeling accuracies.

	Method 1	Method 2	Method 3
Date information	–	√	√
Label extension	–	–	√
Precision (%)	43.4	97.4	86.1
Recall (%)	95.1	45.4	79.3

Table I shows the result of the experiment. According to Table 1, the proposed method showed high precision and recall rates, so its effectiveness was confirmed. The precisions of method 2 was also high. In method 1, many mislabelings were observed, since the method did not consider date information, and labeled only by the similarity of the text information. Additionally, the precision of method 2 was higher than that of method 3. This is because CCs were mis-labeled when labels were extended. In order to solve this problem, we should improve the method of evaluating the similarity of text information. On the other hand, the recall of method 1 was also high. The recall of method 2 was low since it greatly depended on whether date information was described in the text of a Wikipedia entry or not. Additionally, the recall of method 1 was higher than that of method 3. This is because the period of the similarity evaluation was limited. If date information was sparsely described in the text of a Wikipedia entry, the recall of method 3 will be lower than that of method 1.

B. Experiment 2 : The Number of Labeled News Stories

We observed the number of news stories labeled with a same Wikipedia entry. The number of Wikipedia entries which were extracted as news related entries was 1,645. A histogram of the number of labeled news stories is shown in Figure 6. The frequency of 0 news story was removed, which had a frequency of 1,305 (79%).

According to Figure 6, in many cases, the number of labeled news stories was under 20. We considered that this number is a little small to explain news events in each Wikipedia entry, so we should improve the labeling coverage further. Moreover, the frequency of 0 news story was high. This is because the number of news stories used in this experiment was small.

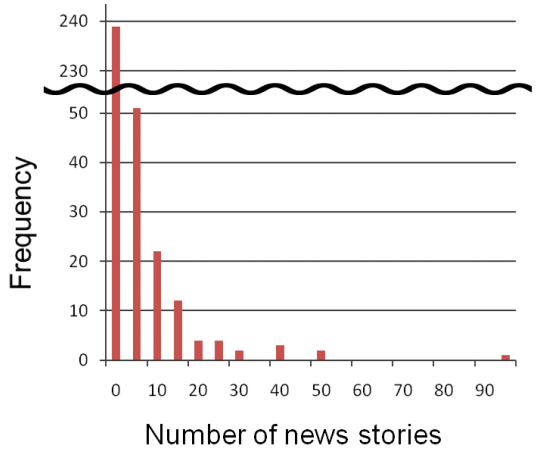

Figure 6. Histogram of the number of news stories labeled with a same Wikipedia entry.

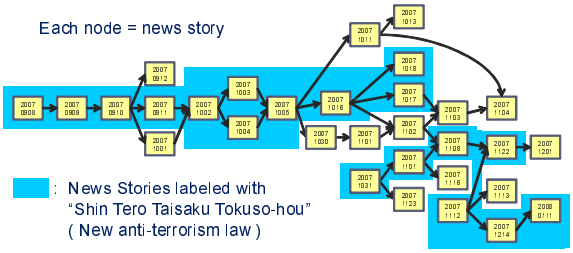

Figure 7. Example of a labeled topic thread structure.

C. Result of extending information of Wikipedia entries

We tried to extend information of Wikipedia entries with news stories. Here, we used the proposed method when labeling news stories. For example, we will examine the result of the Wikipedia entry on "Shin Tero Taisaku Tokuso-hou (New anti-terrorism law)" in this Section. Figure 7 is a part of the topic thread structure of which the starting point is a news story broadcast on September 8, 2007. Here, each node represents a news story. News stories that were labeled as "Shin Tero Taisaku Tokuso-hou" is highlighted in the figure. From this result, video groups are associated with the date information of a Wikipedia entry as shown in Figure 8.

Here, we observed the broadcast dates of news stories labeled with some Wikipedia entries in the topic thread structure. Broadcast dates of the news stories labeled with "Shin Tero Taisaku Tokuso-hou" ranged from September, 2007 to January, 2008. On the other hand, broadcast dates of the news stories labeled with "Shinzo Abe (ex-Japanese Prime Minister)" was on September, 2007 in the same topic thread structure. Then, we could understand the transition of a topic from "Shinzo Abe" to "Shin Tero Taisaku Tokuso-hou" along the topic thread structure. Moreover, we observed

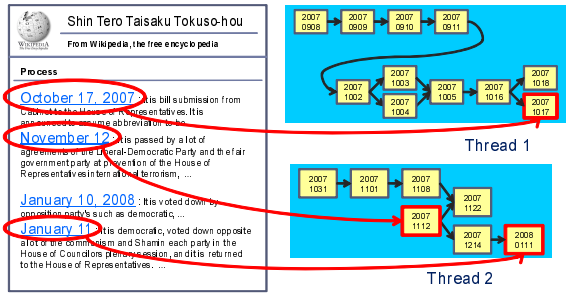

Figure 8. Example of extending information of Wikipedia entry. The Japanese part is translated into English.

that we could also understand the transition of topics well in other cases.

V. Conclusion

In this paper, we proposed a method of labeling news stories in a topic thread structure with Wikipedia entries. The proposed method improved the labeling accuracy. We confirmed by experiments that news stories were accurately labeled with Wikipedia entries with a precision of 86% and a recall of 79%.

As a future work, we will make a browsing interface that intuitively presents news videos related to each Wikipedia entry. Moreover, we will investigate on a method that automatically analyzes the transition of news topics and present it in an interface. Additionally, we will try summarizing news videos by means of labeling them with Wikipedia entries.

Acknowledgment

We thank to National Institue of Informatics (NII) who archived the news videos data used in the experiments. This work was partially supported by the Grants-in-Aid for Scientific Researches from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

References

[1] B. Chan, L. Wu, J. Talbot, M. Cammarano, and P. Hanrahan. Vispedia: Interactive Visual Exploration of Wikipedia Data via Search-Based Integration. *IEEE Trans. on Visualization and Computer Graphics*, 14(6):1213–1220, Oct. 2008.

[2] J. A. Fiscus and G. R. Doddington. Topic Detection and Tracking Evaluation Overview. In J. Allan, editor, *Topic Detection and Tracking: Event-based Information Organization*, pages 17–32. Kluwer Academic Publishers, Feb. 2002.

[3] I. Ide, T. Kinoshita, T. Takahashi, S. Satoh, and H. Murase. mediawalker: A Video Archive Explorer based on Time-Series Semantic Structure. *Proc. 15th ACM Int. Multimedia Conf.*, pages 162–163, Sep. 2007.

[4] C. Snoek, M. Worring, J. van Gemert, J. M. Geusebroek, D. Koelma, G. Nguyen, O. de Rooij, and F. Seinstra. Mediamill: Exploring News Video Archives based on Learned Semantics. *Proc. 13th ACM Int. Multimedia Conf.*, pages 225–226, Nov. 2005.