International Journal of Software Engineering and Its Applications
Vol.8, No.6 (2014), pp.229-240
http://dx.doi.org/10.14257/ijseia.2014.8.6.18

Effectively Detecting Topic Boundaries in a News Video by Using Wikipedia

Jong Wook Kim[1] and Sae-Hong Cho[2]

[1]*Dept. of Media Software at Sangmyung University*
20-Gil, Hongji-dong, Jongno-gu, Seoul, Korea
[2]*Dept. of Multimedia Engineering at Hansung University*
389 Samsun-Dong 3-Ga, Sungbuk-Gu, Seoul, Korea
[1]*jkim@smu.ac.kr*, [2]*chosh@hansung.ac.kr*

Abstract

With the development of internet technology, traditional TV news providers start sharing theirs news videos on the Web. As the number of TV news videos on the Web is constantly increasing, there is an impending need for effective mechanisms that are able to reduce the navigational overhead significantly with a given collection of TV news videos. Naturally, a TV news video contains a series of stories that are not related to each other, and thus building indexing structures based on the entire contents of it might be ineffective. An alternative and more promising strategy is to first find topic boundaries in a given news video based on topical coherence, and then build index structures for each coherent unit. Thus, the main goal of this paper is to develop an effective technique to detect topic boundaries of a given news video. The topic boundaries identified by our algorithm are then used to build indexing structures in order to support effective navigation guides and searches. The proposed method in this paper leverages Wikipedia to map the original contents of a news video from the keyword-space into the concept-space, and finds topic boundaries by using the contents represented in the concept-space. The experimental results show that the proposed technique provides significant precision gains in finding topic boundaries of a news video.

Keywords: *Wikipedia, Semantic interpretation, Topic boundary detection*

1. Introduction

As internet technology develops, traditional TV news providers start sharing theirs news videos on the Web. With the constantly increasing number of TV news videos shared on such websites, it is becoming more and more difficult for users to find the information that they need in a collection of TV news videos. Thus, there is an impending need for the effective mechanism that enables to significantly reduce the navigational overhead.

Naturally, a TV news video contains a series of stories that are not related to each other, and thus building indexing structures based on the entire contents of it might be ineffective. An alternative and more promising strategy is to find topic boundaries in a given news video based on topical coherence, to split it into multiple segments such that each of which contains one coherent story, and to build index structures for the content of each segment. Bearing this in mind, the very first step to provide effective guides and searches to users is to identify topic boundaries in a given news video. Thus, the main goal of this paper is to develop an effective technique to find topic boundaries of a given news video. The topic boundaries identified by our algorithm are then used to build indexing structures in order to support effective navigation guides and searches.

ISSN:1738-9984 IJSEIA

1.1. Contributions of this Paper

Detecting shot boundaries in a video has been extensively studied in the literatures [12-15]. Shot boundary detection in a video is usually performed by comparing various features of neighborhoods of frames, such as color, texture and shape. However, the same cannot be said about a TV news video shared on the Web. As a TV news video contains a series of stories that are not related to each other, it is more interesting to find topic boundaries in a news video rather than to detect shot boundaries. Thus, the goal of this paper is to develop the topic segmentation algorithm which effectively is able to find topic boundaries of a given TV news video. The topic boundaries obtained by our algorithm are then used to build index structures for a collection of TV news videos in order to support effective navigation and search for users. Major contributions of this paper can be summarized as follows:

- In Subsection 4.2, we map the original closed-caption texts of a news video from the keyword-space into the concept-space by leveraging Wikipedia based semantic interpreter.
- In Subsection 4.3, we develop the topic boundary detection algorithm which exploits semantically enriched contents with Wikipedia.

The rest of this paper is structured as follows. In the next section, we formally define the problem. In Section 3, we describe Wikipedia-based semantic interpreter which will be used when developing the proposed topic boundary detection algorithm. In Section 4, we introduce our algorithm that can effectively identify topic boundaries in a news video. In Section 5, we experimentally evaluate our approach using real data sets. In Section 6, we conclude the paper.

2. Problem Statement

Closed-caption data, which are texts displayed on a screen, are included in most of news videos in order to provide additional information. The propose technique in this paper uses closed-caption data in order to identify topic boundaries of a TV news video. In this paper, closed-caption data is treated as a stream of sentences consisting of s_1, s_2,……, s_n. Given a stream of n sentences, the topic boundary detection problem in a news video can be rewritten as finding the e special sentences, each of which corresponds to the first sentence that introduces a new story in a news video (where $e \leq n$).

3. Background

In this section, we describe Wikipedia-based semantic interpreter [2, 3], which will be used in the next section to develop the algorithm proposed in this paper.

3.1. Wikipedia-based Keyword-Concept Matrix

Wikipedia [1] contains more than 4 million articles in the English version and is the largest encyclopedias freely available on the Web. Each article in Wikipedia represents a concept which belongs to at least one concept-category. We model the concepts in Wikipedia by using a $l \times m$ keyword-concept matrix, C. Here, l equals to the number of distinct keywords in the dictionary, and m is the total number of concepts in Wikipedia. Given the keyword-concept matrix, C, let assume that $C_{i,r}$ denote the weight of the i-th keyword, t_i, in the r-th concept, c_r. Let further assume that the r-th concept vector $\vec{c_r}$ be represented as $\vec{c_r} = C_{-,r} = [w_{1,r}, w_{2,r}, w_{3,r}, w_{4,r}, \ldots\ldots, w_{l,r}]^T$. Without loss of

generality, we assume that each concept-vector, $\vec{c}_r$, is normalized into a unit length (*i.e.*, $|\vec{c}_r| = 1$).

3.2. Wikipedia-based Semantic Interpreter

Wikipedia has been successfully used as a semantic interpreter in diverse application domains, including link analysis [5], clustering [6], classification [7], word disambiguation [2], user profile creation [8], video labeling [9], topic detection [10] and topic segmentation [4,11]. Wikipedia-based semantic interpreter enriches original documents by leveraging the concepts of Wikipedia. As shown in Figure 1, such semantic reinterpretation is to map original documents from the keyword-space into the concept-space [2]. The mapping between the original keyword-space and the Wikipedia concept- space is performed as follows: We

1. first identify Wikipedia concepts that are related to the keywords appearing in the given document and,
2. next replace the keywords in the document with these related concepts found in the previous step.

Given Wikipedia-based keyword-concept matrix, *C*, and a document vector, $\vec{d} = [w_1, w_2, w_3, \ldots\ldots, w_l]^T$ in the keyword-space, a semantically re-interpreted document vector with Wikipedia concepts, $\overrightarrow{dc} = [w'_1, w'_2, w'_3, \ldots\ldots, w'_m]^T$ is formally defined as [2]

$$\overrightarrow{dc} = \vec{d}C.$$

By definition of matrix multiplication, we compute the contribution of the concept c_r in the vector $\overrightarrow{dc}$ as follows:

$$w'_r = \sum_{1 \le i \le l} w_i \times C_{i,r}.$$

Wikipedia-based semantic enrichment has the potential to ensure that documents mapped into the Wikipedia concept-space are semantically informed. See [2] for more detailed description of Wikipedia-based semantic interpreter.

4. Algorithm

In this section, we describe the proposed technique to effectively identify topic boundaries of a given news video being composed of a series of different stories. Generally, closed-caption texts of a news video are too short to be meaningful by themselves. Therefore, in this paper, we leverage Wikipedia-based semantic interpreter to enrich the original closed-caption texts of a given news video. Figure 2 provides an overview of the proposed approach in this paper:

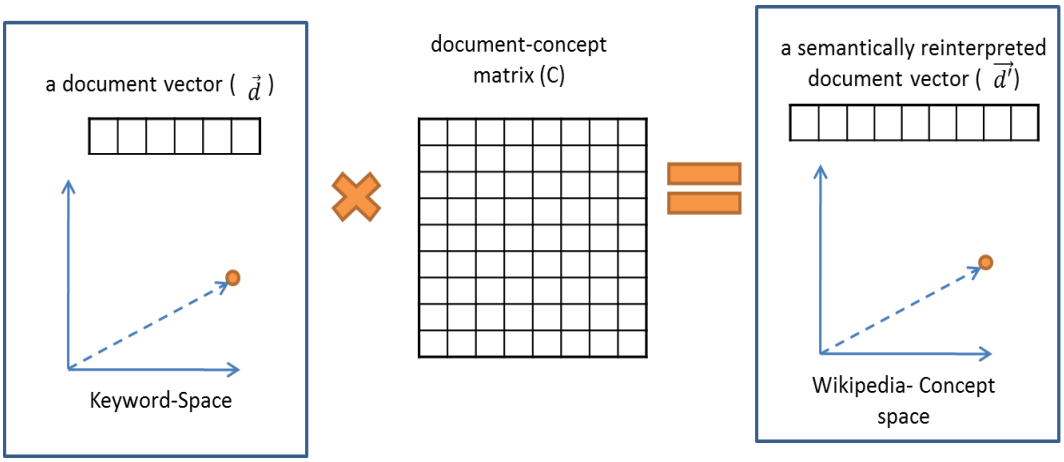

Figure 1. An Overview of Wikipedia-based Semantic Interpreter

- Given a stream of closed caption data consisting of n-sentences, s_1, s_2, …, s_n, extracted from a new video, first we generate (n - p) sentence sequence vectors, $\vec{s}_{\langle 1,p\rangle}$, $\vec{s}_{\langle 2,p+1\rangle}, \vec{s}_{\langle 3,p+2\rangle}, \ldots\ldots, \vec{s}_{\langle n-p+1,n\rangle}$, by passing over the stream with a sliding window of size p (Subsection 4.1).
- We first compute the semantically enriched vectors, $\overrightarrow{sc}_{\langle 1,p\rangle}$, $\overrightarrow{sc}_{\langle 2,p+1\rangle}, \overrightarrow{sc}_{\langle 3,p+2\rangle}$, $\overrightarrow{sc}_{\langle 4,p+3\rangle}, \ldots\ldots, \overrightarrow{sc}_{\langle n-p+1,n\rangle}$, by mapping the original sentence sequence vectors from the keyword-space into the Wikipedia concept-space by using Wikipedia-based Semantic Interpreter in [3] (Subsection 4.2).
- Then, we identify topic boundaries of a given news video by relying on these semantically enriched vectors space (Subsection 4.3).
  - We first perform a low-granularity analysis to identify the rough area where the topic boundaries will be likely found (Subsection 4.3.1).
  - Finally, a higher-granularity analysis is carried out in order to detect the sentences to be marked as topic boundaries within these areas (Subsection 4.3.2).

We now explain and describe each of these steps in detail.

4.1. Preliminary

In this paper, we treat the closed caption data of a news video as a stream of n sentences, s_1, s_2, …, s_n. Given a dictionary L, a stream of n sentences is represented as $n \times l$ sentence-keyword matrix, S, where n is the number of sentences and l is the number of distinct keywords in the dictionary. Let $S_{r,i}$ denote the weight of the i-th keyword, t_i, in the r-th sentence, s_r. Given a closed-caption stream, let $\vec{s}_{\langle r,r+p-1\rangle}$ be the r-th sentence sequence vector which is formed by concatenating p consecutive sentences, s_r, s_{r+1}, …, s_{r+p-1}. Without loss of generality, we assume that each sentence sequence vector is normalized into a unit length.

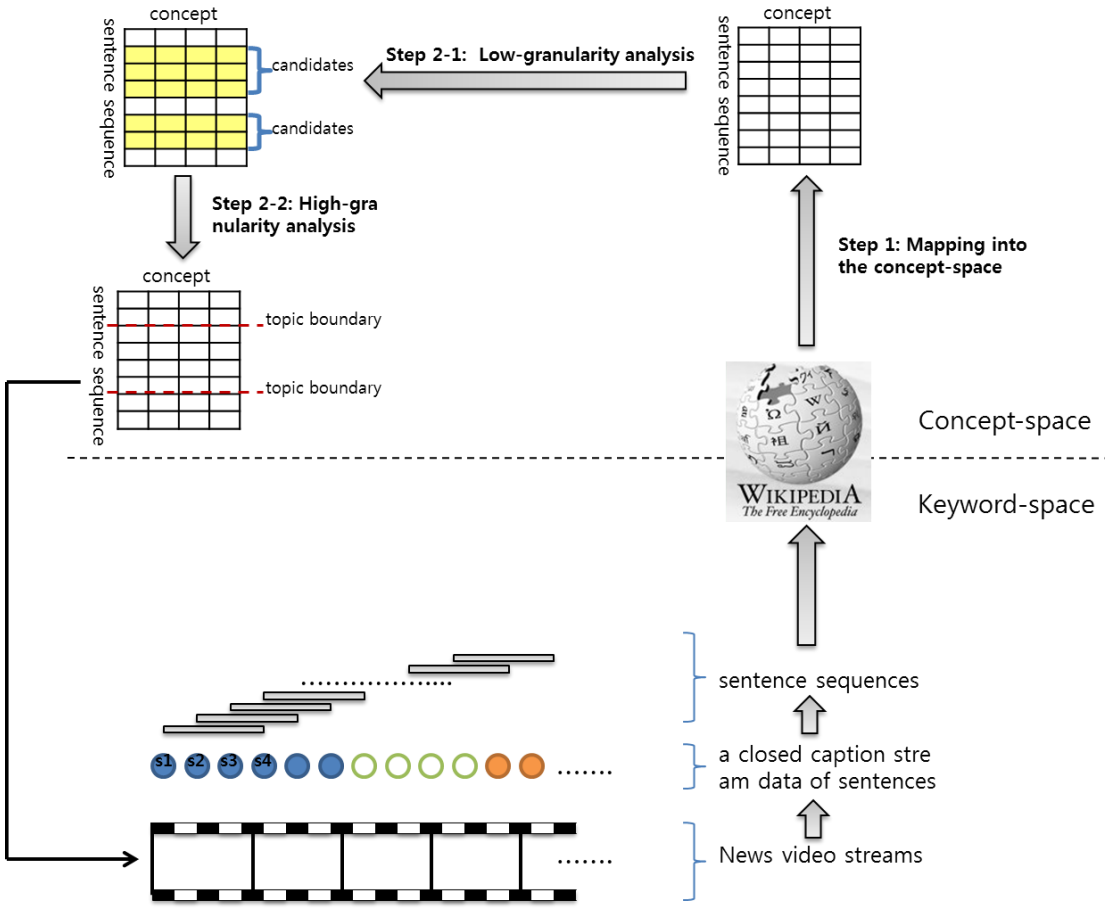

Figure 2. An Overview of Identifying Topic Boundaries in a News Video by using Wikipedia-based Semantic Interpreter

4.2. Step 1: Mapping from the Keyword-Space into the Concept-Space

Although the keyword-based model is commonly used in many applications, it also has limitations, such as a lack of semantic relationship between keywords. As discussed previously, one way to tackle these limitations is to enrich the individual data with the background knowledge obtained from Wikipedia [3, 4]. However, considering the huge number of concepts (articles) in Wikipedia, it is not feasible to enrich original data with all possible concepts of Wikipedia. More importantly, concepts in Wikipedia are not equally important to a given data. Therefore, in this step, we *enrich each sentence sequence with the best k concepts in Wikipedia that are relevant to it*.

Given a *r*-th sentence sequence vector, $\vec{s}_{\langle r,r+p-1\rangle}$, let Bs_r be a set of *k* concept vectors in Wikipeda, such that the following holds:

$$\forall_{\vec{c}_a \in Bs_r, \vec{c}_b \notin Bs_r} \; sim(\vec{s}_{\langle r,r+p-1\rangle}, \vec{c}_a) \geq sim(\vec{s}_{\langle r,r+p-1\rangle}, \vec{c}_b),$$

where sim($\vec{s}_{\langle r,r+p-1\rangle}$, $\vec{c}_a$) is the similarity score between two vectors, $\vec{s}_{\langle r,r+p-1\rangle}$ and $\vec{c}_a$. In other words, Bs_r contains *k* concept vectors in Wikipeda whose similarity scores to

$\vec{s}_{\langle r,r+p-1\rangle}$ belong to the first k highest scores. Then, given a r-th sentence sequence vector, $\vec{s}_{\langle r,r+p-1\rangle} = [w_1, w_2, \ldots\ldots, w_l]$, in the keyword-space, a corresponding semantically enriched vector, $\overrightarrow{sc}_{\langle r,r+p-1\rangle} = [w'_1, w'_2, \ldots\ldots, w'_m]$ with the k most important Wikipedia concept vectors, Bs_r, is defined as

- if $\vec{c}_a \in B_{sr}$, $then\ w'_a = \sum_{1\le i\le l}(w_i \times C_{i,a})$
- Otherwise, $w'_a = 0$.

See [3,4] for more detailed description of the algorithm which computes the semantically enriched vector with the concepts having the k highest similarity scores with the original vector.

4.3. Step 2: Detecting Topic Boundaries in the Concept-space

In this subsection, we describe the proposed approach which identifies topic boundaries by using semantically enriched vectors. We also note that our framework is general in that semantically enriched vectors with Wikipedia concepts can be used with existing topic boundary detection algorithms, such as [17,18].

4.3.1. Step 2-1: Finding Candidate Sentence-Clusters: In this step, we perform a low-granularity analysis to detect a set of sentence-clusters in which entry sentences that diverge significantly from previous news story are likely to be found. Like topic detection and tracking systems, which mainly focus on detecting and tracking events in a streaming data, our goal in this paper is to identify events significantly different from those events seen before. Naturally, it is common that sentence sequences consisting of p consecutive sentences may contain more than one topic, when they are located on near topic boundaries. Then, as one computes the similarity score between any two sentence sequences whose distance equals to the size of the sliding window (*i.e.*, p), it is likely that topic boundaries are found at the local minima of the similarities-curve (Figure 3).

Based on this observation, we first identify the candidate sentence that may diverge significantly from an original theme and thus, introduce an entirely new story as following. Let us given two sentence sequence vectors, $\vec{s}_{\langle r,r+p-1\rangle}$, and $\vec{s}_{\langle r+p,r+2p-1\rangle}$, whose distance equals to the size of sliding windows, p. Let further assume that $\overrightarrow{sc}_{\langle r,r+p-1\rangle}$, and $\overrightarrow{sc}_{\langle r+p,r+2p-1\rangle}$ be the corresponding semantically enriched concept vectors respectively, as explained in Subsection 3.2. Then, we mark s_{r+p} as a candidate sentence when the following holds:

$$\mathrm{sim}\left(\overrightarrow{sc}_{\langle r,r+p-1\rangle}, \overrightarrow{sc}_{\langle r+p,r+2p-1\rangle}\right) < \theta_c$$

Here, θ_c is the threshold value that can be learned from training data. Naturally, as shown in Figure 3, a cluster of candidate sentences is formed near topic boundary, implying that a large cluster provides a fairly good indicator of topic divergence. Therefore, in order to be able to benefit from this observation when detecting topic boundaries, the candidate sentences further need to be clustered based on the distance among them, as shown in Figure 4. In this paper, we employ agglomerative hierarchical based clustering algorithm in order to choose the candidate sentences to be clustered iteratively [16]. However, we note that other clustering algorithms can also be used.

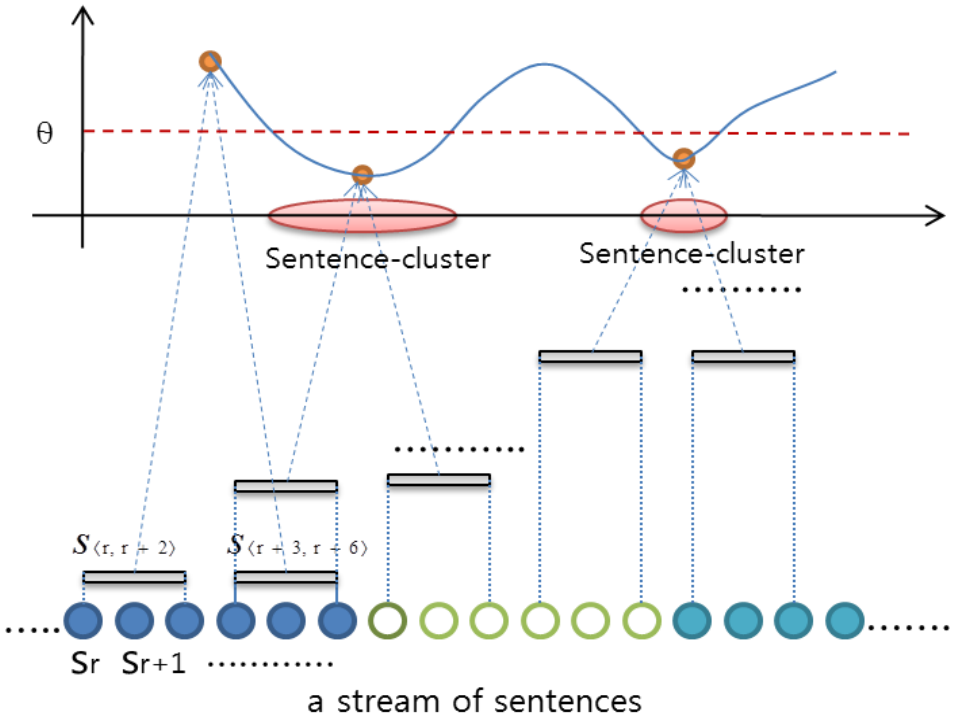

Figure 3. Similarity Score between Sentence Sequence Vectors whose Distance Equals to the Size of the Sliding Window (p)

The algorithm in Figure 4 starts with a set of candidate sentences, Cnd_{sen}, identified as described earlier and returns a set of sentence-clusters, Cnd_{cls}. In the initialization step, each sentence cluster in Cnd_{cls} is initialized (line 1-4). The algorithm iteratively merges sentence-clusters cls_a and cls_b into cls_c by leveraging agglomerative hierarchical clustering method (line 5-9). Here, the distance between two sentence-clusters, cls_a and cls_b, is computed as following:

$$\text{dis}(cls_a, cls_b) = \max_{si \in clsa, sj \in clsb} \left(dis(s_i, s_j) \right),$$

where $dis(s_i, s_j) = |i - j|$. This process iteratively is repeated until $\text{dis}(cls_a, cls_b)$ is less than the threshold, θ_{dis}. In line 11-13, sentence-cluster whose size is less than the threshold, θ_{size}, is removed from Cnd_{cls}. The reason behind this is to reduce false positive errors: the algorithm indicates a topic boundary, but in fact there is no topic divergence.

Input: a set of candidate sentences, Cnd_{sen}
Output: a set of sentence-clusters, Cnd_{cls}

```
    /* initialize each cluster in Cnd_cls */
 1: Cnd_cls ← ∅
 2: for each s_r ∈ Cnd_sen do
 3:     cls_r ← {s_r}
 4:     Cnd_cls ← Cnd_cls ∪ {cls_r}
    /* merge step */
 5: repeat
 6:   Apply agglomerative hierarchical clustering method to
      select candidates, cls_a and cls_b, in Cnd_cls that are
      merged into cls_c
 7:   if dis(cls_a, cls_b) < θ_dis then
 8:     Cnd_cls ← Cnd_cls ∪ {cls_c}
 9:     Cnd_cls ← Cnd_cls − {cls_a, cls_b}
10: until dis(cls_a, cls_b) < θ_dis
    /* filtering step */
11: for each cls_r ∈ Cnd_cls do
12:   if |cls_r| < θ_size then
13:     Cnd_cls ← Cnd_cls − {cls_r}
14: return Cnd_cls
```

Figure 4. Pseudo-code for Clustering Candidate Sentences

4.3.2. Step 2-2: Identifying Topic Boundarie: Given a sentence-cluster, $cls_a \in Cnd_{cls}$, we first need to identify two sentences lying on the boundaries of the cluster. Let s_{la} and s_{ra} be the sentences such that they correspond to the left and right boundaries of cls_a. Here, la and ra are obtained as follows:

$$\text{la} = argmin_{sr \in clsa}(\text{r})$$
$$\text{ra} = argmax_{sr \in clsa}(\text{r})$$

Then, the sentence-cluster, cls_a, can be thought as a stream of the candidate sentences, s_{la}, s_{la+1}, …… , s_{ra}. Intuitively, the topic boundary will best divide this stream into two segments, s_{la}, s_{la+1},…… , s_{t-1} and s_t,…… , s_{ra-1}, s_{ra}. Bearing this intuition in mind, the sentence, s_t, which diverges significantly from the original news story and introduces an entirely new one, is identified as following:

$$\text{t} = argmin_{la+1 \le i \le ra}\ (sim\ (\overrightarrow{sc}_{\langle la,i-1\rangle}, \overrightarrow{sc}_{\langle i,ra\rangle}))$$

where $\overrightarrow{sc}_{\langle la,i-1\rangle}$ and $\overrightarrow{sc}_{\langle i,ra\rangle}$ are computed as described in Subsection 4.3.

Once we find a set of sentences which are the entry points that introduce new topics, the physical topic boundaries are determined by the shots that are associated with these sentences (Figure 2).

5. Experiment

In this section, we describe the experiments we carried out to evaluate the proposed method in Section 4.

5.1. Experimental Setup

Naturally, the most reasonable way to evaluate effectiveness of the proposed approach in this paper is to compare the topic boundaries detected by the algorithm with ground truth generated by the human common sense. Therefore, for evaluation purposes, we used data that has similar structures to the ones we would like to examine. First, we collected 500 TV news transcripts from [19], each of which has a different news topic. We, then, concatenate those news transcripts together so as to simulate a stream of closed-caption texts of a TV news video.

In order to implement Wikipedia-based semantic interpreter described in [3,4], we used the Wikipedia dump which was released at September 2009 that contains about 4M articles. As in [3, 4], we eliminated redirect pages and articles containing insufficient non-stopword from the Wikipedia data set.

5.2. Results and Discussion

The purpose of the first set of experiments is to examine the effectiveness of the low-granularity analysis presented in Subsection 4.3.1. Since the low-granularity analysis aims to find a set of sentence-clusters where the topic boundaries will be likely found, we measure the precision, recall and F1 of the proposed method as followings:

Precision
$= \#\ sentence\ clusters\ containing\ actual\ \ topic\ boundaries\ /\#\ sentence\ clusters\ ,$

Recall
$= \#\ sentence\ clusters\ containing\ actual\ \ topic\ boundaries\ /\#\ actual\ topic\ boundaries$

$$\mathrm{F1} = (2\ \times \mathrm{precision}\ \times\ \mathrm{recall})/\ (precision + recall)$$

In the experiments, we set θ_{size} to 4 and θ_{dis} to 3. In order to observe the precision, recall and F1 values on varying θ_c, we varied θ_c from 0.05 to 0.2. Table 1 shows the precision, recall, and F1 values of the low-granularity analysis presented in Subsection 4.3.1 on varying θ_c. As expected, the precision values increase, as θ_c increases. This is because as θ_c increases, out algorithm presented in Subsection 4.3.1 will generate a larger number of sentence clusters, which leads to a higher false rate, and thus a lower precision. On the other hand, the recall values decrease, as θ_c increases. As can be seen in the table, we have the highest F1 value when θ_c set to 0.15. Thus, in the next set of experiments, we set θ_c to 0.15.

Table 1. Precision, Recall and F1 Values on Varying θ_c

	$\theta_c = 0.05$	$\theta_c = 0.1$	$\theta_c = 0.15$	$\theta_c = 0.2$
Precision	0.620	0.841	0.923	0.903
Recall	0.980	0.976	0.960	0.922
F1	0.759	0.903	0.941	0.912

Table 2. The Performance Comparison of *Seg_Wiki* and *Seg_Non_Wiki*

	Precision	Recall	F1
Seg_Wiki	0.769	0.800	0.784
Seg_Non_Wiki	0.640	0.666	0.652

In the next set of experiments, we evaluate the performance of the topic boundary detection algorithm proposed in Section 4. For evaluation purposes, we report results obtained by using following two alternatives:

- case *Seg_Wiki*: This is the proposed scheme. In this case, topic boundaries of a given news video are computed based on semantically enriched sentence sequence vectors with Wikipedia, as described in Section 4.
- case *Seg_Non_Wiki*: This is the same method as *Seg_Wiki*, excepting that it relies on the original sentence sequence vectors represented in the *keyword-space*, instead of the enriched sentence sequence vector represented in the *Wikipedia concept-space*.

Table 2 shows the precision, recall, and F1 results obtained by using two different approaches, *Seg_Wiki* and *Seg_Non_Wiki*. As shown in the table, the proposed *Seg_Wiki* significantly outperforms *Seg_Non_Wiki* in precision and recall. This verifies the usefulness of using Wikipedia-based semantic enrichment strategy in topic boundary detection problem.

6. Conclusion

In this paper, we proposed a novel algorithm for effectively identifying topic boundaries in a news video. Our proposed approach detects topic boundaries based on the concept-space obtained from Wikipedia, instead of the original keyword-space. Experiment results show that the proposed technique significantly improves the effectiveness of topic boundary detection task. Future work will include investigation of the proposed method for other types of data.

Acknowledgements

This research was supported by a 2014 Research Grant from Sangmyung University

References

[1] Wikipedia, the free encyclopedia.http://www.wikipedia.org.
[2] E. Gabrilovich and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing", Journal of Artificial Intelligence Research, vol. 34, **(2009)**.
[3] J. W. Kim, A. Kashyap, D. Li and S. Bhamidipati, "Efficient Wikipedia-based semantic interpreter by exploiting top-k processing", Proceedings of the ACM conference on Information and knowledge management, Toronto, Canada, **(2010)** October 26-30.
[4] J. W. Kim, A. Kashyap and S. Bhamidipati, "Wikipedia-Based Semantic Interpreter Using Approximate Top-k Processing and Its Application", Journal of Universal Computer Science, vol. 18, no. 5, **(2012)**.
[5] D. Milne and I. H. Witten, "Learning to link with Wikipedia", Proceedings of the 17th ACM conference on Information and knowledge management, Napa Valley, California, USA, **(2008)** October 26-30.
[6] D. Carmel, H. Roitman and N. Zwerdling, "Enhancing cluster labeling using Wikipedia", Proceedings of the 32nd international ACM SIGIR conference, Boston, MA, USA, **(2009)** July 19-23.
[7] P. Wang and C. Domeniconi, "Building semantic kernels for text classification using Wikipedia", Proceedings of the 14th ACM SIGKDD international conference, Las Vegas, NV, USA, **(2008)** August 24-27.
[8] K. Ramanathan and K. Kapoor, "Creating user profiles using Wikipedia", Proceedings of the 28th International Conference on Conceptual Modeling, Gramado, Brazil, **(2008)** November 9-12.
[9] T. Okuoka, T. Takahashiy, D. Deguchi, I. Ide and H. Murase, "Labeling news topic threads with Wikipedia entries", Proceedings of the IEEE International Symposium on Multimedia, San Diego, California, USA, **(2009)** December 14-16.
[10] P. Schonhofen, "Identifying document topics using the wikipedia category network", Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, **(2006)** December 18-22.
[11] J. W. Kim and S. H, Cho, "Content-based Topic Segmentation in a News Video by Leveraging Wikipedia", Proceedings of the 3rd Multimedia Workshop, Jeju, South Korea, **(2014)** April 15-19.
[12] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram and D. Zhong, "VideoQ: An automated content based video search system using visual cues", Proceedings of the ACM Multimedia 1997, Seattle, WA, USA, **(1997)** November 9-13.
[13] A. Hampapur, R. Jain and T. E. Weymouth, "Indexing video databases", Proceedings of the Storage and Retrieval for Image and Video Databases 1995, San Jose, CA, USA, **(1995)**.
[14] N. V. Patel and I. K. Sethi, "Video shot detection and characterization for video databases", Pattern Recognition, vol. 30, **(1997)**, pp. 583.
[15] H. Zhang, A. Kankanhalli and S. Somaliar, "Automatic partitioning of full-motion video", Multimedia Systems, vol. 1, no. 10, **(1993)**.
[16] W. H. E. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering method", Journal of Classification, vol. 1, no. 1, **(1984)**.
[17] J. Allan, C. Wade and A. Bolivar, "Retrieval and novelty detection at the sentence level", Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval, New York, NY, USA, **(2003)** July 28-August 01.
[18] F. Y. Y. Choi, "Advances in domain independent linear text segmentation", Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, **(2000)**.
[19] CNN Transcripts, http://transcripts.cnn.com/TRANSCRIPTS/.

Authors

Jong-Wook Kim
Assistant Professor, Sangmyung University
Dept. of Media Software
Interest Area: Web Data Mining, Information Retrieval, DB Systems
e-mail: jkim@smu.ac.kr

Sae-Hong Cho
Professor, Hansung University
Dept. of Multimedia Engineering
Interest Area: Multimedia, Virtual Reality, Big Data, Digital Contents
e-mail: chosh@hansung.ac.kr