

Video Scene Title Generation based on Explicit and Implicit Relations among Caption Words

Jeong-Woo Son, Wonjoo Park, Sang-Yun Lee, Sun-Joong Kim

Smart Media Research Group,

Broadcasting and Media Research Laboratory

ETRI (Electronics and Telecommunications Research Institute), Korea

[jwson, wjpark, syllee, kimsj]@etri.re.kr

Abstract— Titles of videos are the most important aspect to provide various services. Recently, scene based or video fragments based services have been launched. Since videos in such services are often automatically generated by segmenting a video, these contents cannot have their own titles. As a result, titles of the video fragments are annotated by human hands. To reduce the cost for manual annotation of video titles, this paper proposes a novel method to generate titles of videos by selecting informative sentences from closed captions. The proposed method utilizes explicit and implicit relations among words occurred in closed captions by constructing a stochastic matrix. And then, the proposed method picks important words up based on their weights estimated with TextRank. A title is generated by selecting a sentence with important words from closed captions. Experimental results shows several cases with well-known Korean TV programs.

Keywords— Scene title, video clip, text rank, stochastic matrix, closed caption, keyword extraction

I. INTRODUCTION

Short videos and clips have been played a central role in content consumption on the Internet recently. Even though there exist a huge amount of video contents produced by content providers such as production companies, broadcasting stations and so on, they do not meet the condition to make a service on the Internet. Ordinary video contents have a long play time and do not contain any informative tag or meta-data. Thus, such contents are provided to audiences by segmenting and analyzing them automatically.

When a set of videos or clips is suggested, an audience selects one of them to play a video. This is a common scenario for a video clip service. At that time, a title of a video is the most important factor to make a decision of an audience. Most scene titles in contemporary video clip services are manually annotated. Thus, such manual process make a hurdle to provide for audiences various contents existed in the repositories of content providers.

The advance of machine learning, more specifically deep learning suggest various methods to generate descriptions for images and videos [1,2]. However, these image and video captioning methods cannot fully support the needs of video

based services. Figure 1 shows an example of a video clip service. As shown in this figure, a manually generated title contain specific information like the name of persons, situation, location, and so on. Thus, it is difficult to adopt ordinary methods which aim to make general descriptions.



Figure 1. An example of a video service¹

This paper proposes a novel method to generate titles of videos, clips and scenes. In the proposed method, closed captions in a video are used as its title. Since there exist several closed captions, it demands a method to select the best one among closed captions. The proposed method basically uses words and their relations observed in closed captions. First, by using morphological analysis, nouns and verbs are extracted from closed captions. Second, explicit relations

¹ NAVERTV, <http://tv.naver.com>

among words are extracted based on their distances in a closed caption. Then, a stochastic matrix is constructed and implicit relations are derived from the stochastic matrix. TextRank [3] algorithm is used to determine weights of words with the stochastic matrix. Finally, a closed caption is selected with the average weight of its words. In experiments, we proved the effectiveness of the proposed method with well-known Korean TV programs.

II. TITLE GENERATION

When a video clip and its closed caption are given, nouns and verbs are first extracted by using a morphological analyzer [4]. Then, a kernel matrix is constructed with words and their co-occurrences. The stochastic matrix is then constructed by computing the long-term average of transitions in the kernel matrix. Weights of all words are determined from the stochastic matrix by using TextRank. Finally, the title is generated by selecting a sentence from closed captions with respect to average word weights.

A. Extracting words

A closed caption C is composed of $\{S, T\}$ where S is a set of sentences and T denotes time stamps of sentences in S . That is, $s \in \{S\}$ and $t_s \in \{T\}$. A set of Words W is constructed by extracting nouns and verbs from all s in S . Then, a sentence s can be represented as $\{w_1, w_2, \dots, w_n\}$ and $\{f_1, f_2, \dots, f_n\}$ where f_i denote the normalized frequency of i -th word and n is the size of W . If there exist several sentences in the same time stamp, the word and frequency vectors are constructed by merging all vectors from the sentences.

B. Constructing stochastic matrix

When the word and frequency vectors are constructed, the kernel matrix can be generated by using the co-occurrences of words. Co-occurrences between words are obtained from the word vector. After normalizing the kernel matrix, the normalized version of the constructed kernel matrix K can be regarded as a stochastic matrix with direct transitions. These transition probabilities are explicit relations among extracted words. The implicit relations among words are revealed by determining long term average of the stochastic matrix. That is,

$$K' = \sum_{i=1}^m K^i,$$

where m denotes the maximum length of indirect transition between words. With K' , we can compute expectations to reach an arbitrary word through a number of transitions by $E = F \cdot K' \cdot I$, where F is the frequency vector and I represents a column matrix of all ones that acts as a sum over transitions. The expectations of transitions E can be regarded as word weights. However, it is not enough to represent an importance of a word. The importance of a word should be determined words appeared the path to the word. If all words in the transitions to the word are much important, the word

should be important. On the other hand, even though there exist various paths to a word, the word is not important when all words in the paths are meaningless. To reflect it, we applied TextRank with K' .

C. TextRank and selecting the title

TextRank is a well-known method to extract keywords among words with their relations. Basically, TextRank has been adopted in text summarization [5]. The role of TextRank in text summarization is to extract keywords. TextRank determines centralities of words with an iterative process. Then, a set of words are extracted with respect to their centralities. TextRank estimates the weights of words as,

$$TR(w_i) = (1 - d) + d \cdot \sum_{w_j \in In(w_i)} \frac{K'_{ji}}{\sum_{w_k \in Out(w_j)} K'_{jk}} TR(w_j),$$

where $TR(w)$ denotes the weight of w while $In(w)$ and $Out(w)$ are in-bound and out-bound connections in K' respectively. d is a damping factor to decide where a transition is occurred or not. Generally, d is set as 0.85. In here, the initial weight of all words are determined with E in the previous section. Through iterations, the weight of a word is updated by using weights of its neighbours.

When all words in W has their own weights TR , the importance of a sentence in closed captions can be determined with the average of word weights. When a video has several sentences, we can select one of them with respect to their importance. The selected sentence is used as the title of the video.

III. EXPERIMENTS

There is no benchmark dataset for video title generation. Thus, in this paper, we suggest several cases that generate video titles with video clips. All video clips are automatically generated from well-known Korean TV programs by using [6].

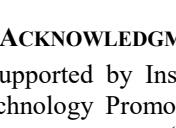
Table 1 shows 10 results of title generation. Since all captions in the experimental video are written in Korean, we manually translate them in English. As shown in this table, the proposed method select informative sentences from closed captions. Especially, for the second program “알쓸신잡”, even though all the keyframes are similar each other (since same persons are appeared in a room), the proposed method can generate different sentences for video clips. This kind of results cannot be generated from scene description [7, 8] or scene captioning methods [9]. This is the advantage of the proposed method comparing with other sentence generation methods.

IV. CONCLUSIONS

This paper proposes a novel title generation method for video clips. Since titles are efficient ways to connect audiences for videos, it is a much importance factor for various services. In the proposed method, the title of a video is automatically generated with sentences in closed captions. The importance of a sentence is determined with word weights.

The word weights are estimated with the stochastic matrix and TextRank. By adopting both algorithm, the proposed method can reflect direct and indirect relations among words in their weights. Experimental results shows the real cases of title generation with the proposed method. Often, keyframes or thumbnails cannot represent the context of the videos. At that case, the proposed method can generate proper titles for videos.

Table 1. Title generation results

PROGRAM	KEYFRAME	TITLE
“태양의 후예”		HEY, YOU BASTARD! WHAT CAN I DO NOW THAT I CANNOT HELP YOU?
		NO MORE THAN MAIN GATE
		IT'S NOT THAT I WAS GOING TO GO, BUT THE INTERNET WILL BE FASTER HERE.
		IS THE OPERATION AREA A MEDICUBE FIELD HOSPITAL?
		LOW BLOOD PRESSURE ON ABDOMINAL DISTENSION? HIMOFENIE?
“알쓸신잡”		KILLER, KILLER
		I'VE NEVER HEARD ANYONE SAY THIS BEFORE.
		BECAUSE OF THIS <TAEBAEK MOUNTAIN RANGE>
		THAT'S ABOUT 200 PAGES OF MANUSCRIPTS.
		ALPHA VESUS ALPHA, WHO WILL WIN IF THEY FIGHT?

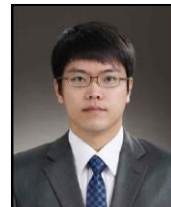
ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) [2015-0-00219,

Development of smart broadcast service platform based on semantic cluster to build an open media ecosystem]

REFERENCES

- [1] H. Yu, J. Wang, Z. Huang, Y. Yang, W. Xu, “Video paragraph captioning using hierarchical recurrent neural networks,” in *Proc. CVPR*, 2016, pp. 4584-4593.
- [2] P. Pan, Z. Xu, Y. Yang, F. Wu, Y. Zhuang, “Hierarchical recurrent neural encoder for video representation with application to captioning,” in *Proc. CVPR*, 2016, pp. 1029-1038.
- [3] R. Mihalcea, P. Tarau, “TextRank: Bringing Order into Texts,” in *Proc. EMNLP*, 2014.
- [4] J. Shin, “Komoran,” <https://github.com/shin285/KOMORAN>
- [5] G. Erkan and D. R. Radev, “LexRank: Graph-based Lexical Centrality as Saliency in Text Summarization”, Volume 22, pages 457-479, 2004.
- [6] J.-W. Son, W. Park, S.-Y. Lee, J. Kim, S.-J. Kim, “Smart Media Generation System for Broadcasting Contents,” in *Proc. SIGIR 2017*, pp. 1297-1300, 2002.
- [7] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, “Long-Term Recurrent Convolutional Networks for Visual Recognition and Description,” in *Proc. CVPR*, 2015, pp. 2625-2634.
- [8] Rohrbach A., Rohrbach M., Qiu W., Friedrich A., Pinkal M., Schiele B. “Coherent Multi-sentence Video Description with Variable Level of Detail,” In *Proc. GCPR 2014*.
- [9] R. Shetty and J. Laaksonen, “Video captioning with recurrent networks based on frame- and video-level features and visual content classification,” *CoRR*, 2015.



Jeong-Woo Son received his MS and Ph.D. degrees in computer engineering from Kyungpook National University, Daegu, Rep. of Korea in 2007 and 2012 respectively. Since 2013, he has been with Broadcasting & Telecommunications Media Research Lab., Electronics and Telecommunications Research Institute (ETRI), Daejeon, Rep. of Korea. He focuses on machine learning, NLP, and information retrieval.



Wonjoo Park received her MS degrees in information and communication engineering from Chungnam National University, Daejeon, Rep. of Korea in 2000. She joined ETRI, Rep. of Korea in 2000, where she is currently senior researcher. Her research interests includes data mining, topic model, and ontology.



Sang-Yun Lee received his MS and Ph.D. degrees in electronics and telecommunications engineering from Hanyang University, Seoul, Rep. of Korea in 1996 and 2007 respectively. Since 2011, he has been with Broadcasting & Telecommunications Media Research Lab., Electronics and Telecommunications Research Institute (ETRI), Daejeon, Rep. of Korea. He focuses on embedded software, smart TV software platform, and Web technology.



Sun-Joong Kim received her BS degree in computational statistics and her MS degree in computer science from Chungnam National University, Daejeon, Rep. of Korea, in 1989 and 2000 respectively. In February 1989, she joined ETRI, Daejeon, Rep. of Korea, where she is currently principal researcher and director. Her research interests includes convergence service control, smart TV, content knowledge mining.