Topic-based Generation of Keywords and Caption for Video Content

Masanao Okamoto[1], Kiichi Hasegawa[1], Sho Sobue[1], Akira Nakamura[2], Satoshi Tamura[1], Satoru Hayamizu[1]
[1]Gifu University
[2]Advanced Technology Research Center, SANYO Electric Co.,Ltd.
1-1 Yanagido,Gifu,Gifu,501-1193,Japan
E-mail: okamoto@asr.info.gifu-u.ac.jp Tel: +81-58-293-2763

***Abstract*— This paper studies usage of both keywords and captions in one scene for video content. Captions show the spoken content and are renewed in a sentence unit. A method is proposed to extract keywords automatically from transcribed texts. The method estimates topic boundary, extracts keywords by Latent Dirichlet Allocation (LDA) and presents them in speech balloon captioning system. The proposed method is evaluated by experiments from the viewpoint of easy to view and helpfulness to understand the video content. Adding keywords and captions obtained favorable scores by subjective assessments.**

I. INTRODUCTION

For decades, the closed captions have been used in TV programs. Some cognitive experiments of news captioning for hearing-impaired people were conducted. They show that all of the spoken texts (350-400 characters per minute) are too many to understand [1]. The guideline of TV closed caption is constituted based on these experiments. On the other hand, more video contents are nowadays around us. Captions are mostly used as an assistance to understand video contents. Most captions are obtained from manually transcribed or summarized texts, which require a lot of costs.

Recently, speech recognition technology is used to produce texts for captions from audio part of video [3,4]; for example, a speech balloon captioning system is proposed for information support on meetings with multiple speakers [2]. In the context of summarization research, automatic text summarization is studied for TV news programs [5]. Other related works are about statistic-based text summarization [6], speech summarization [7] and integration of linguistic and visual information [8]. Since keywords, or important words, express the summarized content and convey information more efficiently and effectively than whole texts, keywords are usually attached by authors or automatically extracted from whole texts [9].

This paper studies automatic keyword extraction as well as topic boundary detection. A speech balloon captioning system is also proposed. In the system, keywords are presented with different colors for each topic. Topic changes are estimated by features from LDA (Latent Dirichlet Allocation) [10] models. The whole framework of keyword usage and captioning is evaluated by experiments from the viewpoint of easy to view and helpfulness to understand the video content.

The rest of this paper is organized as follows: Section II describes the speech balloon captioning system and related technologies used in the system. Keyword extraction and topic change detection are introduced in Section III. Evaluation experiments are shown in Section IV. Finally Section V concludes this paper.

II. A SPEECH BALLOON CAPTIONING SYSTEM

A. Overview

We develop a captioning system using speech balloons. Balloon captions are presented in a movie, indicating a transcription of a speaker's utterance. Keywords are extracted from the text, and topic boundaries are also estimated. The keywords are emphasized using a same color for one topic, in order to help user's understanding.

Fig.1 shows an outline of the system proposed in this paper. At first, audio data are extracted from a video file. Secondary, beginning points of each utterance in a time domain are determined by a Voice Activity Detection (VAD) technique. A caption text is obtained from the audio data, then keyword and topic extraction are conducted. Finally, our system integrates the movie file, the beginning points and the caption text to display a captioned presentation using Adobe Flash. Note that a transcribed text is used in this paper since we'd like to evaluate a balloon captioning system as well as a keyword detection method described in Section III.

B. Voice Activity Detection

Detecting speech segments in audio data is necessary not only to synchronize the audio with corresponding texts but also to conduct speech recognition. A simple VAD method is implemented in our system; a power coefficient is calculated in each audio frame to detect speech segments by a particular threshold. Afterwards, we also executed hangover processing to reduce detection errors.

C. Sentence Boundary Estimation

In the proposed system, sentence boundaries are needed in order to fill a speech balloon with a suitable amount of texts. Texts are segmented and tagged with part-of-speech classes by morphological analysis. For each word, a probability of its

Proceedings of the Second APSIPA Annual Summit and Conference, pages 605–608,
Biopolis, Singapore, 14-17 December 2010.

label can be expressed by the conditional probability with model parameters. Here, the label is either a final word or a non-final word in one sentence. A sequence of five words is a unit to be considered. It consists of two preceding words, the focused word, and two following words. The feature set is the surface expression and the part-of-speech tag for each word in the sequence. These are decided by preliminary experiments.

The conditional probability of label for each input word can be estimated by this feature set and model parameters of Conditional Random Fields (CRF) [11]. The model parameters were trained with news paper articles and transcribed speech corpus. Words before end punctuation are labeled as final words. In the following experiments, the transcribed texts from speech are used instead of speech recognition results.

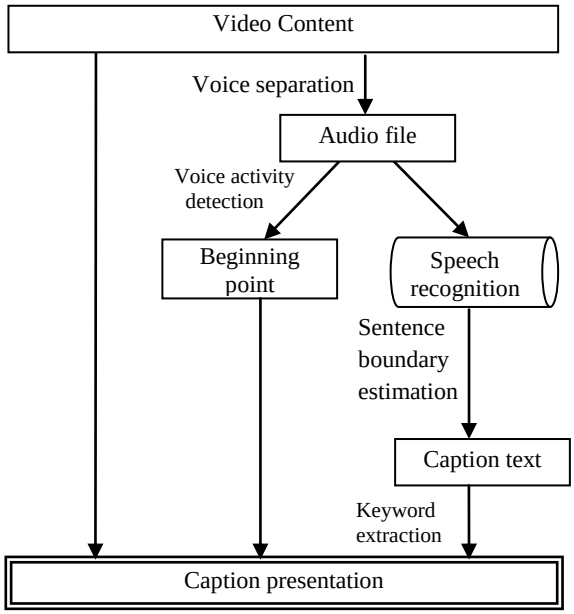

Fig.1 System outline.

D. *Display Method of Captions*

Caption of television (TV) and the movie are often displayed in the lower part of the screen. In contrast, speech balloon captioning system displays the caption near speaker's face. The latter captioning method makes it easier to know who and when spoke, as well as what was spoken in the case of multiple speakers, Fig.2 shows examples of TV captioning system (left) and speech balloon captioning system (right).

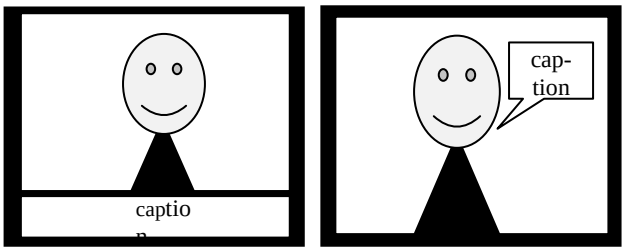

Fig.2 Caption systems (left: TV system, right: balloon system).

III. AUTOMATIC KEYWORD EXTRACTION

The system extracts keywords from a caption text while estimating topic boundaries. The keywords of each topic can be extracted by estimating the topic boundary. The keywords are highlighted so that the user can quickly understand the outline of a video content. LDA is employed for topic boundary estimation and keyword extraction. MeCab [12] is used as a part-of-speech and morphological analyzer.

A. *Latent Dirichlet Allocation*

LDA is a probabilistic document model that assumes the distribution θ over C latent topics $(z_1, z_2 \ldots z_C)$ is given by the Dirichlet distribution Dir($\theta|\alpha$) for each document. A probability of a document $d=(w_1, w_2 \ldots w_{|d|})$ is expressed by:

$$P(d \mid \alpha, \beta) = \int Dir(\theta \mid \alpha) \left(\prod_{n=1}^{|d|} \sum_{k=1}^{C} P(w_n \mid z_k, \beta) P(z_k \mid \theta) \right) d\theta \qquad (1)$$

where α and β are model parameters of LDA and β_{kj} denotes $P(w_j|z_k)$; the unigram probability of w_j in topic z_k ($1 \leqq j \leqq V$; V:vocaburary size).

B. *Topic Boundary Estimation*

When an unknown text d is observed, $P(z_k|d)$ can be computed by LDA. A topic mixture ratio vector, or a C-dimensional vector each of whose element is $P(z_k|d)$, represents the mixture proportion of the latent topics contained in the text. Similarity between two texts can be measured by:

$$cos(t_1, t_2) = \frac{t_1 \cdot t_2}{\|t_1\| \|t_2\|} \qquad (2)$$

where t_1 and t_2 denotes topic mixture ratio vectors of each text respectively. The system detects a topic boundary when the similarity between two adjacent text blocks is below a threshold. Fig.3 shows an example of topic mixture ratio vectors.

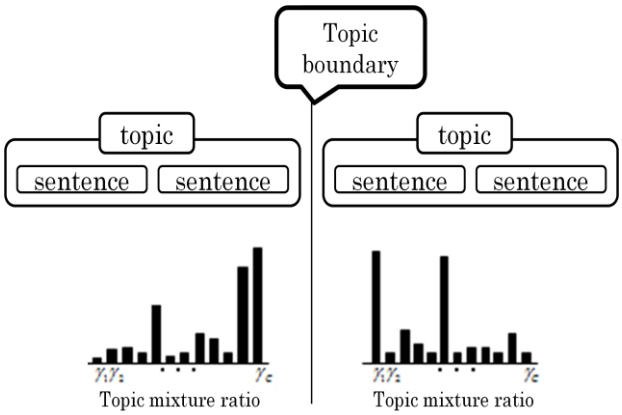

Fig.3 An example of topic mixture ratio vectors.

C. *Compound Word*

A compound word is a word composed of two or more words. A number of common compound words are analyzed as separate words when the standard IPA lexicon for MeCab is employed. Hence we added approximately 90,000 compound nouns used as the titles of Wikipedia articles to the lexicon. In addition, 56,000 sequences of two nouns appeared in the training corpus are added as compound words.

D. *Keyword Extraction*

A word related to a specific topic is extracted from a caption as a keyword. In this study, a word w is considered to belong to a topic z_k when $P(w|z_k)$ is largest among C latent topics. After a document is segmented by the detected topic boundaries, the system extracts keywords that belong to any of the several dominant topics, of which mixture proportion in the current segment is above a threshold. In this paper, the number of latent topics is set at 100. The keywords are highlighted in different colors according to the topics they belong to. Fig.4 shows some examples of extracted keywords.

In the 15th COP of UNFCCC, the government …
---
occupies one fourth of all materials in the universe, …
---
emphasized we don't give up a dream of world peace which never rely on any force…

----- Topic boundary

Fig.4 Examples of extracted keywords.

IV. EVALUATION EXPERIMENTS

A. *Evaluation of Balloon Captioning System*

To test the effectiveness of our captioning system, we conducted the subjective experiments on 5-point scales about helpfulness to understand the content. Evaluation items for the first experiment are shown as follows:

[Evaluation item]
(1) Timing of switching caption
(2) Sentence boundary
(3) Easy to view
(4) Number of printing character
(5) Total evaluation

Table 1 shows details of a video content used in the experiments. Fig.5 shows results for the above five evaluation items. Higher score indicates better evaluation results.

Table 1 Conditions of evaluation experiments.

Speaker	1
Character's number of caption	Not limit
Length of content	3 minutes
Number of subjects evaluated	14
Caption system	Balloon

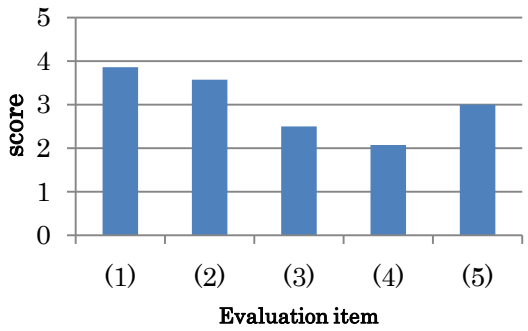

Fig.5 Evaluation for a balloon captioning system.

The evaluation about sentence boundary (item 2) indicates a good estimation result and effectiveness of VAD and CRF. The number of characters varies for each caption. It becomes sometimes several lines long and makes the balloon too large. In such cases, the balloon caption system obtained unfavorable evaluation results (item 3 and item 4). They show that improvement is needed to insert appropriate boundaries in one sentence, or to limit the number of characters in one caption, or to summarize the whole sentence. This experiment was about speech balloon captioning system for one speaker. Effectiveness for multiple speakers should be explored further. In order to extend this captioning system, it is considered to utilize keywords in the captions.

B. *Evaluation of Keyword Extraction*

A subjective experiment is conducted to evaluate the keyword extraction system. Fig.6 shows sample content with a caption and emphasized keywords proposed in this paper. A movie was presented to test subjects (14 students), who were then asked to assess them in the form of a five-grade scale. Other experimental conditions were the same as the preliminary experiment described table 1. The evaluation item is shown as follows:

The Kyoto Protocol obligated the average reduction of 6% to Japan and of 8% to EU compared with the 1990 levels, in the five years from 2008 to 2012 as the first commitment period.

Fig.6 Sample scene with caption and keywords.

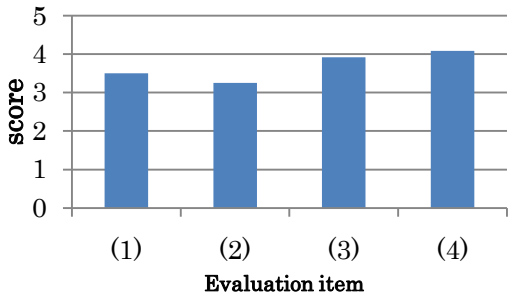

Fig.7 Evaluation for extracted keywords.

[Evaluation item]
(1) Number of keywords
(2) Appropriateness of keywords
(3) Helpfulness to understand
(4) Understanding the change in the topic

The evaluation result (item 4) in Fig.7 shows our proposed method is useful for user to understand topic changes. However, the experimental results (item 1 and item 2) are insufficient. The proposed system sometimes represents many keywords excessively since all keyword candidates belonging to latent topics of which topic mixture ratio is over a threshold were adopted.

There is a possibility that a color of keywords influences user's understanding level. For example, white texts on black background may be easy to read for users. Thus it is necessary to examine the influence of color of keywords in caption. As described in section II, representing caption texts in a balloon is one of the most important issues in our system. Properly inserting text breaks, summarizing transcribed texts, and simply displaying many technical terms should be investigated. Finally, further evaluations using the other contents are also required.

V. CONCLUSION

This paper has described usage of keywords and speech balloon captioning in one scene for video content. A method was proposed to extract keywords by topic model from the transcribed texts. They were tested from the viewpoint of helpfulness to understand the video content. It is found that adding appropriate keywords obtained favorable scores by subjective assessments. As future works, estimation of topic boundary and extraction of keywords need higher accuracy and should be more robust. When speech recognition results are used as input texts, recognition errors may be problematic to obtain appropriate keywords.

ACKNOWLEDGEMENTS

This research was partially supported by the Grant-in-Aid for Scientific Research (C) No. 22500151, promoted by Japan Society for the Promotion of Science.

REFERENCES

[1] K.Komine, H.Hoshino, H.Isono, T.Uchida and Y.Iwahana, "*Cognitive Experiments of News Captioning for Hearing Impaired Persons*", Technical Report of IEICE, HCS96-23, pp.7-12, in Japanese, 1996.
[2] A.Fujii, N.Hiroaki and Y.Takehiro, "*Speech balloon captioninging System for Information Support on Meetings"*, IPSJ SIG Technical Report, 2009-SLP-75-14, pp.75–82, in Japanese, 2009.
[3] J.Son, J.Kim, K.Kang and K.Bae, "*Application of speech recognition with closed caption for content-based video segmentation"*, Proc, IEEE DSP Workshop, 2000.
[4] T.Tanaka, K.Mori, S.Kobayashi and S.Nakagawa, "*Automatic Construction of CALL System from TV news Program with Captions"*, EUROSPEECH 2001, pp.2815-2818, 2001.
[5] T.Wakao, T.Ehara, E.Sawamura, I.Maruyama and K.Shirai, " *Project for production of closed-caption TV programs for the hearing impaired"*, Annual Meeting-Association for Computational Linguistics, pp.1340-1344, 1998.
[6] K.Knight and D.Marcu, *"Statistics-Based Summarization – Step One: Sentence Compression"*, Proc.17th- AAAI, pp.703-710, 2000.
[7] T.Hori and S.Furui, "*Summarized Sentence Generation Based on Word Extraction and Its Evaluation"*, Institute of Electronics, Information, and Communication Engineers, D-II, vol. J85-D-II, No,2, pp.200-209, in Japanese, 2002.
[8] T.Shibata and S.Kurohashi, "*Unsupervised Identification by Integrating Linguistic and Visual Information Based on Hidden Markov Models"*, Proceedings of the COLING/ACL pp.755-762, 2006.
[9] Y.Matsuo and M.Ishizuka, "*Keyword extraction from a single document using world co–occurrence statistical information*", International Journal on Artificial Intelligence Tools, 13(1):157–169, 2004.
[10] D.Blei, A.Ng and M.Jordan, "Latent Dirichlet Allocation", journal of Machine Learning Research, Vol.3, pp.993-1022, 2003.
[11] J.Lafferty, M.Andrew and P.Fernando, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", In proceeding of the 18th International Conference on Machine Learning, 2001.
[12] MeCab :Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/
[13] T.Hofman, "Probabilistic Latent Semantic Indexing", Proc. of 22nd Annual ACM Conference on Research and Development in Information Retrieval, pp.50-57, 1999.
[14] S.Sato, "*What can be said in 13 characters? Analysis of web news box titles"*, 14th, Annual Meeting NLP, in Japanese, C3-7, pp508-511, 2008.
[15] M.Okamoto, H.Ueji, S.Sobue, K.Yamamoto, S.Tamura and S.Hayamizu, "*Keywords and Caption Easy to View and helpful to Understand Video Content*", 15th, Annual Meeting NLP, in Japanese, P2-27, pp578-581, 2009.
[16] S.Sobue, K.Yamamoto, S.Tamura and S.Hayamizu, "Evaluation of Classification Models on Sentence Boundary Estimation from Speech Recognition Result", 15th, Annual Meeting NLP, in Japanese, P2-28, pp572-585, 2009.
[17] CaboCha :Yet Another Japanese Dependency Structure Analyzer, http://chasen.org/~taku/software/cabocha/
[18] Julius: Open-Source Large Vocabulary CSR Engine Julius, http://julius.sourceforge.jp/