The Use of Semantic and Acoustic Features for Open-Domain TED Talk Summarization

Fajri Koto*†, Sakriani Sakti*, Graham Neubig*, Tomoki Toda*, Mirna Adriani†, and Satoshi Nakamura*
*Graduate School of Information Science, Nara Institute of Science and Technology, Japan
E-mail: {ssakti,neubig,tomoki,s-nakamura}@is.naist.jp Tel: +81-743-725264
†Faculty of Computer Science, University of Indonesia, Indonesia
E-mail: fajri91@ui.ac.id, mirna@cs.ui.ac.id Tel: +62-21-7863419

***Abstract*—In this paper, we address the problem of automatic speech summarization on open-domain TED talks. The large vocabulary and diversity of topics from speaker-to-speaker presents significant difficulties. The challenges increase not only how to handle disfluencies and fillers, but also how to extract topic-related meaningful messages within the free talks. Here, we propose to incorporate semantic and acoustic features within the speech summarization technique. In addition, we also propose a new evaluation method for speech summarization by checking semantic similarity between system and human summarization. Experiments results reveal that the proposed methods are effective in spontaneous speech summarization.**

I. Introduction

Recently, information in Internet is available with various data such as text, images, sound, and also video. Consequently, many researchers start to study how to retrieve information from these various data. Automatic speech summarization has been also actively investigated. By using audio and video of speech data, many researchers have investigated summarization based on the output of automatic speech recognition (ASR) [2][3][5]. Here, the summarization process is performed over the text output of ASR system without involving audio features information. For example, the study by Hori et al., extract and calculate the word significance score and the linguistic likelihood from the ASR output [2]. Furthermore, some other techniques like random walk, words and sentence extraction, weighted finite-state transducers, and Hidden Markov Model have been also studied by some researchers to improve speech summarization technique [3][14][15][16].

However, despite a lot of progress in speech summarization, most works focused primarily only on news content, news broadcast, and other non spontaneous speech data. On the other hand, there are many spontaneous speech exist where people are willing to have a summarization of the talks but difficult to obtain. One of the cases is open-domain talk like TED talks[1] that are still limited to be used. TED is a nonprofit devoted to Ideas Worth Spreading. It started out in 1984 as a conference bringing together people from three worlds: Technology, Entertainment, Design. TED talks bring together the world's most fascinating thinkers and doers, who are challenged to give the talk of their lives in 18 minutes or less. Here, we initiate to address the problem of automatic speech summarization (ASR) on open-domain TED talks.

It is obvious that spontaneous speech in TED talks is very different from speech in broadcast news in which speakers do not have any text guidance in their hand. This resulted in output of ASR system will have higher error than broadcast news speech recognition. Furthermore, The large vocabulary and diversity of topics from speaker-to-speaker presents a significant difficulties. The challenges increase not only how to handle disfluencies and fillers, but also how to extract topic-related meaningful messages within the free talks. In this study, we propose to incorporate semantic features in automatic speech summarization. In this way, the topic related sentences are scored higher than unrelated sentences. As preliminary study, we start with incorporating the proposed methods within the widely-used MMR summarization technique [1].

In addition, we also include acoustic features to the summarization framework, since the acoustic features are also one of significant factors in speech summarization [17]. The MMR technique is done by processing the output of ASR in term frequency (TF) and term frequency-inverse document frequency (TF-IDF) model. Then, various combination with acoustic features, semantic features, as well as acoustic and semantic features together are investigated.

We also propose a new evaluation method for speech summarization by checking semantic similarity between system and human summarization. We argue that Common evaluation like Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and Longest Common Subsequence (LCS) have limitation for spontaneous speech because they are only based on the number of overlapping units such as n-gram and word sequences [13]. Whereas, public speeches like TED are more unstructured and rich with dictionary. Therefor, performing evaluation with semantic similarity is more promising.

II. Overview of MMR-Based Sumarization Techniques

MMR has been widely used for text summarization. MMR is a measure where the retrieval status value (RSV) of a document is influenced by other already retrieved documents: documents similar to retrieved documents have their RSV lowered, thus boosting dissimilar documents [1]. Carbonell

[1]http://www.ted.com/

978-616-361-823-8 © 2014 APSIPA

and Goldstein proposed this formula as follow:

$$MMR(S_i) = \alpha * Sim_1(S_i, D) + (1-\alpha) * Sim_2(S_i, Summ) \quad (1)$$

where S_i is *i-th* sentence in document *D*, *Summ* is summary result that is being built according to highest MMR score for every iteration, and Sim_1 and Sim_2 are similarity measures, which can be the same, or can be set to different similarity metrics. In this study, we use cosine similarity to calculate Sim_1 and Sim_2 as follows:

$$sim(D_1, D_2) = \frac{\sum_i t_{1i}, t_{2i}}{\sqrt{\sum_i t_{1i}^2} * \sqrt{\sum_i t_{2i}^2}} \quad (2)$$

The value of α allows a readjustment of the behavior of MMR to control diversity ranking between unselected sentence with selected summary sentences. Here, TF and TF-IDF model to MMR are performed.

III. The Proposed Summarization Techniques

A. Acoustic and Semantic Feature

1) Acoustic Feature: Acoustic feature that is used in this study is based on INTERSPEECH 2010 paralinguistic challenge configuration (IS10 Paraling features) [18]. It consists of 1582 features described in Table 1, which are obtained in three steps: (1) the 38 low-level descriptors are extracted and smoothed by simple moving average low-pass filtering; (2) their first order regression coefficients are added in full HTK compliance; (3) 21 functionals are applied. However, 16 zero-information features (e. g. minimum F0, which is always zero) are discarded. Finally, the 2 single features F0 number of onsets and turn duration are added. More details of description of each feature can be found in [18].

2) Semantic Feature: Semantic similarity feature is a similarity score that describes the similarity between a sentence and document. We proposed this formula to re-rank the sentences according to similarity score between sentence and whole document.

$$Sim_{sem}(s_i, D) = \frac{\sum_{j=1 \wedge j \neq i}^{j=|S|} Sim_{sem}(s_i, s_j)}{|S| - 1} \quad (3)$$

where s_i is the *i-th* sentence in document *D*, and $|S|$ represents the number of all sentences in document *D*. That formula calculates all semantic similarity score between one sentence with other sentences. We take the mean score as our final score to make the sentence rank. The Sim_{sem} is calculated according to [6], in which one sentence is divided into noun set and verb set and the similarity score between two sentence is then calculated based on the similarity of those noun and verb set described in Eq. 6.

$$S_1 = \{V_1, N_1\} \text{ and } S_2 = \{V_2, N_2\} \quad (4)$$

$$Nb = N_1 \cup N_2 \text{ and } Vb = V_1 \cup V_2 \quad (5)$$

$$Sim_{Sem}(S_1, S_2) = \beta_1 * Sim_v(v_1, v_2) + \beta_2 * Sim_n(n_1, n_2) \quad (6)$$

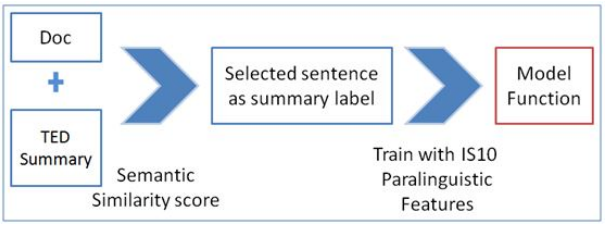

Fig. 1. Build summary label by performing Semantic Similarity.

Eq. 6 above uses two kinds of vector: noun and verb. These vectors are described simply below:

$$Vv_{1k} = Max_{i=1}^{|V_1|}(Sim(Vv_{1i}, Vb_k)) \quad (7)$$

$$Vv_{2k} = Max_{i=1}^{|V_2|}(Sim(Vv_{2i}, Vb_k)) \quad (8)$$

$$Nn_{1k} = Max_{i=1}^{|N_1|}(Sim(Nn_{1i}, Nb_k)) \quad (9)$$

$$Nn_{1k} = Max_{i=1}^{|N_2|}(Sim(Nn_{2i}, Nb_k)) \quad (10)$$

To calculate semantic similarity score between two words above, we also use words-relationship-Tree based on Wu and Palmer's Algorithm [7]. This function utilized online lexical database *WordNet* [8][9]. While, the similarity score (Sim_v and Sim_n) between this two vector can be calculated easily by using cosine similarity formula.

B. Summarization Method

1) Incorporating Semantic Features: In summarization process, MMR re-ranks every sentence by calculating cosine similarity score between two term vectors: sentence and documents. To boost the system accuracy we propose modified MMR by replacing $Sim_1(S_i, D)$ in Eq. 1 with Eq. 11. Our modified MMR incorporates Semantic Similarity in similarity calculation which also considers the cosine similarity. In this study, we use β equals to 0.5.

$$\beta * Sim_{Sem}(S_i, D) + (1-\beta) * Sim_1(S_i, D) \quad (11)$$

2) Incorporating Acoustic Features: The motivation of incorporating acoustic features within summarization framework is to give more score into the sentences that are considered as candidate summary based on acoustic characteristic of the sentences. This is done by naive bayes (NB) classifier, implemented based on this formula below, where it output "1" if sentence Si is considered as candidate summary for that document, and 0 otherwise:

$$MMR_{speech}(S_i) = 0.5 * MMR(S_i) + 0.5 * NBModel(S_i) \quad (12)$$

We train the classifier by training acoustic feature of each sentences as described in Fig. 1. We create labels by calculating semantic score between sentences of every document with existing summary provided by TED website. We built the label by selecting top 10 sentences with highest semantic similarity score for each document.

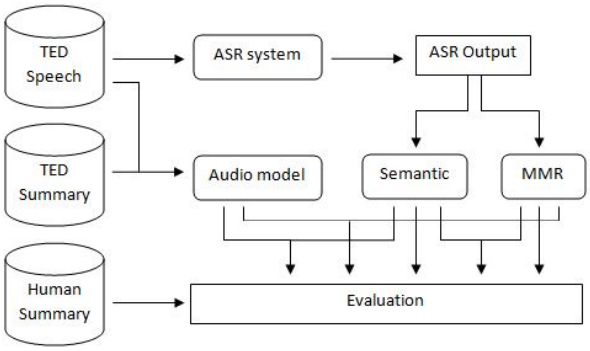

Fig. 2. Automatic Speech Summarization stage

3) Incorporating Acoustic and Semantic Features: To get more elaboration we also implement Naive-bayes-based acoustic classifier to our new method in speech summarization. We formulate the new score of similarity by adding the score with *NBModel* function like Eq. 12. In this study, we use γ equals to 0.1.

$$Sim'_{sem}(s_i, D) = (1-\gamma)*Sim_{sem}(s_i, D)+\gamma*NBModel(S_i) \quad (13)$$

C. Evaluation Metric: Semantic Similarity Checking (SSC)

The well-known automatic evaluation method for Summarizer: ROUGE and LCS have been introduced by Lin [13]. Formally, ROUGE-N is an n-gram recall between a candidate summary and a set of reference summary. Whereas, LCS does not require consecutive matches but in-sequence matches that reflect sentence level order as n-grams [12]. Since we focused on unstructured document like Spontaneous Speech, in this study we propose semantic similarity checking (SSC) in Eq. 14 as a new automatic evaluation method for summarizer which utilizes semantic similarity calculation. Intuitively, this method will be more powerful than ROUGE and LCS because the resulting score is not just built by counting matched words or sequences.

$$SSC(D_1, D_2) = \frac{\sum_{i=1}^{i=|D_1|} Sim_{sem}(s_i, D_2)}{|D_1|} \quad (14)$$

In Eq. 14, the D_1 and D_2 represents document of resulting summary and document of reference summary consecutively. The equation simply calculates the average of all semantic similarity score between every sentence s_i in D_1 and D_2. The similarity score $Sim_{sem}(s_i, D_2)$ is also calculated by averaging the semantic similarity score between s_i and all sentences in D_2.

IV. Overall Architecture of Summarrizer

Fig. 2 above shows our summarization experiment stage. We use output of ASR system to build sentence-based summary by doing some techniques: MMR, Semantic Similarity and their incorporation with the audio model. This model is built by training acoustic feature. Then we do evaluation by calculating similarity score between the resulting summary and human summarization. The ASR system that is used here was trained using 157 hours of TED talks released before the cut-off date of 31 December 2010, downloaded from the TED websites with the corresponding subtitles.

V. Experimental set-up

A. The TED Data

The TED talks that are used in summarization are the same data which were used to evaluate the speech recognition system. There are 20 TED talks in total. The reference summarization is obtained from human summarization. In this study, five native speakers are required to pick ten sentences that were considered as most representative sentences to the speaker topic for each speech document.

B. Preprocessing

To build the vector space models (TF and TF-IDF) we did preprocessing to all TED speech data. We replace all capital letters of transcription file with lowercase and eliminate all punctuations that exist in the transcription file. We also remove some of the unimportant word or segment like laugh and applause. We use all TED documents to build *idf(t,D)* score in calculating TF-IDF.

For acoustic features, we perform segmentation based on time sequences obtained from the *srt* file and our ASR system. It aims to get the valid timing of every sentence in document. Segmented audio file will be extracted by *openSMILE* toolkit [11]. *openSMILE* is a feature extraction toolkit, which unites feature extraction algorithms from the speech processing and the Music Information Retrieval communities [10].

The noun and verb vector that will be used to calculate semantic similarity are processed by implementing python code with NLTK library. The similarity checking will be performed for the same tagging of words (only between two nouns or two verbs).

VI. Experiment Result

As our baseline system, we perform MMR in TF and TF-IDF model. Various value of alpha parameters in MMR formula are investigated. The results are then compare with two proposed methods: incorporating semantic similarity and acoustic features. We do evaluation with SSC and take top 30% highest MMR score of sentences in document as its summary.

Our first experiment is comparing SSC and ROUGE to look their tendency in evaluating system summary. In Fig. 3 we use ROUGE-3 gram and SSC to perform evaluation of MMR-TF for various alpha value. The results reveal that our proposed metric evaluation has in line performance with ROUGE. However, SSC is still better to be used because it is calculated based on semantic.

In MMR experiment we use some alpha parameters: 0.1, 0.3, 0.5, 0.7, and 0.9. Fig. 4 and Fig. 5 show that Semantic Similarity can boost the accuracy of MMR for all alpha parameters. The incorporation of MMR, semantic and acoustic feature is shown by the top line for both graph. It affirms that semantic and acoustic have important role for optimizing

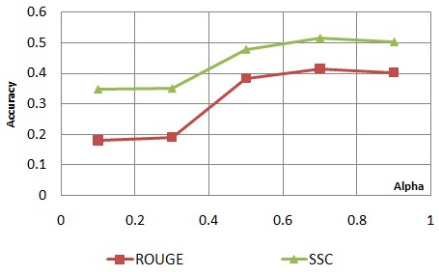

Fig. 3. SSC and ROUGE performance for MMR-TF

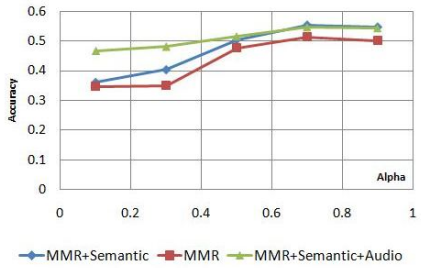

Fig. 4. MMR-TF Experiment result.

automatic speech summarization. According to both line graph (Fig. 4 and Fig. 5), the highest accuracy is achieved by MMR and Semantic at alpha equals with 0.7. They are 55.29% and 55.30% for TF model and TF-IDF model consecutively.

In Table. I, we present the best performance of each methods for vary parameters. Here we compare MMR, MMR+Audio, and MMR+Audio+Semantic. And the result reveals that the combination of MMR, Acoustic and Semantic feature always give better performance than standard technique for both vector space model.

TABLE I
Highest accuracy performance of MMR and its incorporation

Incorporation	TF	TFIDF
MMR	51.38%	50.71%
MMR+Audio	51.77%	53.18%
MMR+Audio+Semantic	**54.64%**	**55.10%**

VII. Conclusion

In this study, we attempt to incorporate both semantic and acoustic features in automatic speech summarization for open-domain TED talks. The experimental results reveal that they can improve textual speech summarization. In short, our study reveals that semantic similarity can be used in speech summarization: 1) as summarization feature and 2) evaluation method. Our experiments also show that the incorporation of MMR, Semantic and Acoustic feature can achieve best performance. It affirms the both features have important role in speech summarization. In our future work, we will further investigate various incorporation approaches of semantic and acoustic features into MMR as well as the combination with other summarization techniques

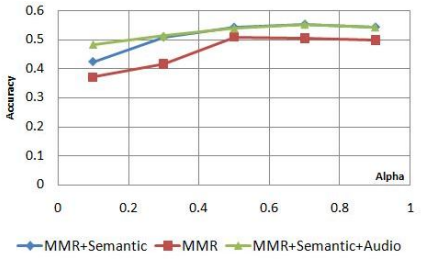

Fig. 5. MMR-TFIDF Experiment result.

VIII. Acknowledgment

Part of this work was supported by JSPS KAKENHI Grant Number 26870371.

References

[1] Carbonell, J., and Goldstein, J. "The Use of MMR, dicersity-based reranking for reordering documents and producing summaries". SIGIR 1998, pp. 335-336. ACM, New York, 1998.
[2] Hori, C., and Furui, S. "Automatic speech summarization based on word significance and linguistic likelihood". Proc. ICASSP2000, Istanbul, Vol.3, pp.1579-1582, 2000.
[3] Hori, C., and Furui, S. "Improvements in automatic speech summarization and evaluation methods". ICSLP2000 4: 326-329, 2000
[4] Maskey, S., and Hirschberg, J. "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization". In INTERSPEECH, pp. 621-624, 2005
[5] Xie, S., and Yang L. "Using confusion networks for speech summarization in Human Language Technologies". The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 46-54. Association for Computational Linguistics, 2010.
[6] Liu, D., Liu, Z., and Dong, Q. "A dependency grammar and WordNet based sentence similarity measure". Journal of Computational Information Systems 8:3 1027-1035, 2012.
[7] Wu, Z., and P. Martha. "Verbs semantics and lexical selection". In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp. 133-138. Association for Computational Linguistics, 1994.
[8] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. "Introduction to wordnet: An on-line lexical database". International journal of lexicography 3, no. 4: 235-244, 1990
[9] Available: http://wordnet.princeton.edu/
[10] Eyben, F., Woellmer, M., and Schuller, B. "openSMILE the Munich open Speech and Music Interpretation by Large Space Extraction toolkit". Institute for Human-Machine Communication, version 1.0.1, 2010.
[11] Available: hhtp://opensmile.sourceforge.net/
[12] Eyben, F., Martin W., and Bjrn S. "Opensmile: the munich versatile and fast open-source audio feature extractor". In Proceedings of the international conference on Multimedia, pp. 1459-1462. ACM, 2010.
[13] Lin, C.-Y. "ROUGE: A Package for Automatic Evaluation of Summaries". In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pp. 74-81, 2004.
[14] Chen, Y.-N., Huang, Y., Yeh C.-F., and Lee. L.-S. "Spoken Lecture Summarization by Random Walk over a Graph Constructed with Automatically Extracted Key Terms." In INTERSPEECH, pp. 933-936, 2011.
[15] Furui, S., Hirohata, M., Shinnaka, Y., and Iwano, K. "Sentence extraction-based automatic speech summarization and evaluation techniques." In Proceedings of the Symposium on Large-scale Knowledge Resources, pp. 33-38, 2005.
[16] Maskey, S. and Hirschberg, J. "Summarizing speech without text using hidden markov models." In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, pp. 89-92. Association for Computational Linguistics, 2006.
[17] Inoue, A., Mikami, T. and Yamashita Y. "Improvement of Speech Summarization Using Prosodic Information." In Speech Prosody 2004, International Conference, 2004.
[18] Schuller, B., Steild, S., Batliner, A., Burkardt, F., Devillers, L., Muller, C., Narayanan, S. "The INTERSPEECH 2010 Paralinguistic Challenge". In INTERSPEECH, pp. 2794-2797, 2010.