

# IMPACT OF AUTOMATIC SENTENCE SEGMENTATION ON MEETING SUMMARIZATION

Yang Liu, Shasha Xie

The University of Texas at Dallas, Richardson, TX, USA

{yangl, shasha}@hlt.utdallas.edu

## ABSTRACT

This paper investigates the impact of automatic sentence segmentation on speech summarization using the ICSI meeting corpus. We use a hidden Markov model (HMM) for sentence segmentation that integrates the N-gram language model and pause information, and a maximum marginal relevance (MMR) based extractive summarization method. The system-generated summaries are compared to multiple human summaries using the ROUGE scores. The decision thresholds from the segmentation system are varied to examine the impact of different segments on summarization. We find that (1) using system generated utterance segments degrades summarization performance compared to using human annotated sentences; (2) segmentation needs to be optimized for summarization instead of the segmentation task itself, however, the patterns are slightly different from prior work for other tasks such as parsing; and (3) there are effects from different summarization evaluation metrics as well as speech recognition errors.

*Index Terms*— meeting summarization, sentence segmentation, ROUGE, MMR.

## 1. INTRODUCTION

Speech summarization is an effective technique to help browse the large volume of audio data. Even though significant progress has been made in text summarization in the past decades, when applying text summarization techniques to spoken language, in particular the multiparty meeting domain, many problems arise because of the significant style differences in written text and spoken language, for example, the lack of punctuation marks, the presence of disfluencies and many speakers with possible speech overlap. In this paper, our focus is on one kind of structural information in speech, that is, sentences. The sentence structure is essential to enrich speech recognition output, making it easier to use for downstream language processing tasks, such as machine translation and summarization.

Different approaches and rich feature sets have been recently developed for meeting summarization (such as [1, 2]). However, most of the approaches used human annotated sentence segments in human transcriptions, or aligned them to recognition output, instead of using automatic segmentation. These are not the real scenario. Mrozinski et al. [3] studied the effect of automatic sentence segmentation on speech summarization in the domain of broadcast news and lecture, and found that system-generated sentence segmentation degrades summarization performance; yet they simply generated a hard decision in sentence segmentation and used that in speech summarization. Whether such segments are optimal for the subsequent speech summarization task is the question this paper is aiming to address.

Recently automatic speech segmentation has gained more attention as more language applications start to deal with speech input. The effect of speech segmentation has been examined for several tasks, such as translation and parsing [4, 5]. In [5], experimental results showed that the best decision when optimizing sentence segmentation itself is different from that optimized for the downstream language processing tasks (parsing in that case). To our knowledge, this kind of study has not been performed for the speech summarization task. In addition, meetings differ from other domains such as broadcast news or lectures used in some prior studies in several dimensions (e.g., multiple speakers, conversational style), raising questions about whether sentence segmentation has a different impact on summarization. The goal of this paper is thus to better link automatic sentence segmentation and summarization of meetings.

## 2. DATA

We used the ICSI meeting data [6], which contains 75 naturally-occurring meetings, each about an hour in length. These meetings have been transcribed, and annotated with dialog acts (DAs) [7], topics, and abstractive and extractive summaries [1]. Note that the definition of sentences is not clear for conversational speech like in the meeting domain. We use the DA information in the corpus as the sentence boundary annotation.

Following the set up in [1, 2], we used the same 6 meetings as the test set. These meetings have multiple human summaries, 3 of which (the three common annotators for these meetings) are used as the reference summaries for each meeting in our experiments. The Kappa statistics [8] among the 3 annotators varies from 0.211 to 0.345 for different meetings (humans generally do not have a high agreement when creating summaries, even for text summarization). The length of the human summaries varies among the annotators and the meetings as well. The average percentage of the DAs and the words selected in the human generated summaries is 6.5% and 13.5% respectively. There was no strict rule on the summary length when these annotators created the extractive summaries. See [1] for more information on summary annotation.

## 3. SENTENCE SEGMENTATION

Sentence segmentation is an important component in spoken language processing, and thus has received increasing attention, such as the sentence-like unit detection task in Rich Transcription (RT) evaluation in the recent DARPA EARS program, and the study of using automatic sentence segmentation for spoken language translation in the DARPA GALE program. Many approaches have been developed for sentence boundary detection, including HMM, maximum entropy, conditional random fields, and Boostexter. These have been tested in various domains, e.g., conversational telephone speech, broadcast news speech, as well as meetings [9, 10]. Unlike

the prior work on sentence segmentation, this paper focuses on its effect on a downstream task, rather than evaluating segmentation by itself.

In this paper, we use the HMM segmentation approach. For a given word sequence  $W$ , the task is to determine whether there is a DA boundary after each word. The most likely DA sequence  $\hat{E}$  can be obtained as follows,

$$\begin{aligned}\hat{E} &= \operatorname{argmax}_E P(E|W, P) \\ &= \operatorname{argmax}_E P(W, E)P(P|W, E)\end{aligned}\quad (1)$$

where  $P$  represents the prosodic features. The transition probability in the HMM (resulting in  $P(W, E)$  in Eq (1)) is obtained from an N-gram language model (LM), which is trained by explicitly including the DA boundary as a token in the vocabulary. The observation probability  $P(P|W, E)$  is approximated as  $P(P|E)$ , and obtained from a classifier based on the prosodic features, which can be properly normalized if the classifier’s output is posterior probabilities (i.e.,  $P(E|P)$ ). In practice, we apply a weighting factor when combining the transition and the observation probabilities. We only use pause for the prosodic information, instead of using all the prosodic features as in [10], since our previous work has shown that adding additional features yields limited gain. The pause duration is obtained using a state-of-the-art recognizer [11]. We split the transcriptions into chunks based on speaker turns, and used them as the sequences in the HMM. This assumption of known speaker information is reasonable since it is readily available for the separate channel recording condition, or can be obtained from a speaker segmentation system for the single microphone setup.

The HMM approach is implemented using the SRILM toolkit [12]. We first split the training data (69 meetings) and used 10% as the development set to optimize parameters (e.g., the interpolation weight for the LM and the pause model). The LM is a 4-gram LM with Kneser-Ney interpolation smoothing, and the pause model is a decision tree classifier. Then we retrained the LM and the pause prosody model using the entire training set. Note that for the speech recognition (ASR) condition, we trained the models using the ASR output of the training set. During testing, using a forward-backward decoding approach in the HMM, we obtained a posterior probability (or confidence) of having a segment boundary at each interword boundary. This allows us to vary the decision threshold to examine the impact of different segmentation on summarization.

#### 4. SUMMARIZATION APPROACH

Our task is generic extractive summarization, that is, the system selects the important sentences in the transcripts to include in the summary, without any compaction or rewriting. Unlike text-based summarization where dealing with redundant or conflicting information is quite important, we believe that extractive speech summarization is appropriate for applications such as meeting browsing, where the goal is to identify the salient segments for easy information access.

There has been a great effort on text summarization for both single document and multi-document summarization (for example, the annual Document Understanding Conferences [13]). Different approaches have also been developed for speech summarization [1, 2, 14, 15]. We choose to use the maximal marginal relevance (MMR) [16] approach because of its simplicity and effectiveness. It is not a statistical learning approach and does not require any training data. It extracts the most relevant sentences and at the same time avoids

redundancy in the summaries. The MMR score for a sentence  $S_i$  is:

$$\text{Score}(S_i) = \lambda * \text{sim}(S_i, D) - (1 - \lambda) * \text{sim}(S_i, \text{SUM}) \quad (2)$$

where  $D$  is the document vector,  $\text{SUM}$  contains those sentences chosen in the current summary, and  $\lambda$  balances the relevance and redundancy. The MMR approach iteratively selects summary sentences. The units used in the MMR summarization system ( $S_i$  in the formula above) are either based on human annotation or automatic DA segmentation.

We use cosine similarity for the second similarity measure in Eq (2), defined as follows for two documents  $D_1$  and  $D_2$ :

$$\text{sim}(D_1, D_2) = \frac{\sum_i t_{1i} t_{2i}}{\sqrt{\sum_i t_{1i}^2} \times \sqrt{\sum_i t_{2i}^2}} \quad (3)$$

where  $t_i$  is the term weight for a word  $w_i$ , for which we use TF-IDF (term frequency and inverse document frequency). One widely used method for IDF is  $\log(N/n_i)$ , where  $n_i$  is the number of documents containing  $w_i$  in a collection of  $N$  documents. The first similarity score in Eq (2) is based on the centroid value of a sentence obtained in MEAD [17], which only considers those words with a high TFIDF score. We calculated the IDF for each word using the 69 training meetings, based on the human transcripts or ASR output respectively for the two different transcript conditions.

To select summaries, we used word-based selection (16% of the words), or sentence-based selection (4.2% of the sentences). Note that both methods use sentences as selection units in the MMR approach—the only difference is the stopping criteria (i.e., whether it is based on the number of sentences, or words, respectively). Word-based selection yields similar number of words in the summary (except the difference due to the last chosen sentence), no matter it uses human DA annotations or automatic segmentation; whereas, for sentence-based selection, the total number of the segments is different when using different DA segments; therefore, there is no guarantee of a similar summary length.

## 5. EXPERIMENTS

### 5.1. Performance Measurement

The sentence segmentation performance is evaluated using the segmentation error rate, defined as the number of inserted and deleted sentence boundaries divided by the total number of DAs in the test set. This is similar to that used in the NIST Rich Transcription evaluation for metadata (sentence boundary, disfluency) extraction.<sup>1</sup>

We use the ROUGE [18] package for summarization evaluation. It compares the system generated summary to the reference summaries, and reports recall, precision, and F-measure results for various matches (e.g., N-gram, skip bigram). Multiple reference summaries are allowed in ROUGE. We use the same options as those in the DUC text summarization evaluations (i.e., with the porter stemming, no stop words) [13]. In a preliminary study, we observed better correlation of ROUGE scores with human evaluation when using the R-SU4 score (skip bigram plus unigram), therefore, we will report the unigram match scores, R-1, as well as R-SU4 in this study.

<sup>1</sup>See <http://www.nist.gov/speech/tests/rt/rt2004/fall/index.htm> for more information.

decision threshold	DA error rate %	average DA len.	number of DA seg.	Summarization results									
				word-based				sentence-based					
				R-1		R-SU4		R-1		R-SU4		avg. len.	% of words
				recall	F-measure	F-measure		recall	F-measure	F-measure			
0.05	68.63	4.7	12735	70.25	67.89	39.95	69.55	67.62	40.15	16.6	15.1		
0.1	51.31	5.4	10921	71.33	68.39	41.3	73.28	68.48*	41.7	20.9	16.3		
0.2	39.35	6.3	9353	71.78	68.44	40.16	74.81	68.19	40.96	25.4	17.0		
0.3	34.22	7.0	8454	72.68	69.08*	41.5	76.47	68.19	41.4	29.2	17.7		
0.4	32.19	7.5	7884	72.92	68.85	41.7*	78.31	68.03	42.16	32.8	18.5		
0.5	31.20*	8.0	7371	72.47	68.59	41.7*	78.85	67.62	42.3*	35.9	18.9		
0.6	31.35	8.6	6909	72.42	68.16	40.3	79.26	66.60	40.93	39.2	19.5		
0.7	32.19	9.2	6444	72.41	68.10	39.98	80.66	65.85	40.78	44.5	20.6		
0.8	34.36	10.0	5907	72.54	67.65	39.08	81.12	65.26	40.54	50.0	21.2		
0.9	39.17	11.5	5152	72.59	67.56	38.51	82.29	64.05	39.53	60.6	22.4		
ref. DA	0	7.4	7966	74.19	70.67	45.68	75.47	70.68	46.3	28.7	16.3		

**Table 1.** Results of DA segmentation and meeting summarization when varying the decision threshold from the automatic DA segmentation output using human transcriptions. The last row corresponds to using the human annotated DA segments. The best results are marked with \* associated with the scores.

## 5.2. Segmentation and Summarization Results

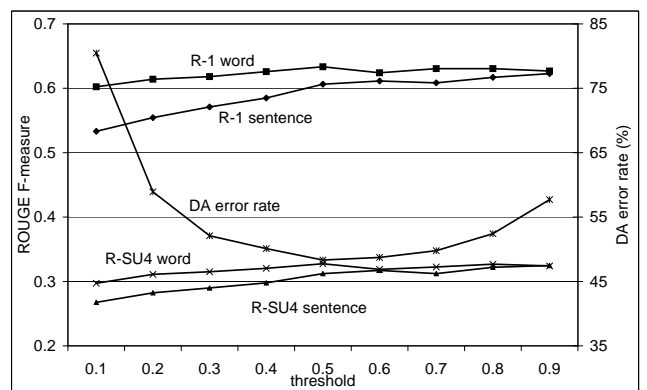
We first evaluate the impact of DA segmentation on summarization using human transcripts to avoid the effect from recognition errors. Table 1 shows the segmentation performance and summarization results obtained by varying the threshold from the HMM-based DA segmentation output. The ROUGE summarization results are shown for both word- and sentence-based selection. We also show the average length of the DAs yielded from different segmentation threshold for the entire test set, as well as in the selected summaries. The best results for segmentation and summarization (using F-measure) are marked with \* in the table. For comparison, results using human annotated DA segments are also included in Table 1 (the last row).

We observe that using automatic DA segmentation (no matter what the thresholds are) degrades summarization performance compared to using human annotations. This difference is more noticeable when using R-SU4 score. In [5], it is clear that the optimal thresholds for parsing and sentence segmentation are different — generally shorter segments are preferred for parsing. However, even though overall shorter segments are preferred for summarization (than those optimized for the segmentation task), the trend is not as clear as for the parsing task. In addition, the pattern is different for different evaluation metrics. We can see from Table 1 that based on the R-1 scores, the best segmentation threshold is not necessarily the best for the summarization task, especially for the sentence-based selection scenario, but that the best threshold for sentence segmentation is about the same as for summarization using the R-SU4 measure.

There is a difference between the results for word- and sentence-based selection. The larger threshold results in longer segments; however, because of the frequent short DAs in meetings (e.g., backchannels), the total number of segments obtained from different segmentation does not change that significantly. Therefore, sentence-based selection (choosing a certain percent of the segments) tends to generate more sentences and words (last column in Table 1) in the summaries than word-based selection. This typically yields better recall rate. In addition, comparing the average length of the DAs in the entire set versus those selected in the summaries in Table 1, we also notice that they are not proportional—the sentences in the summaries are relatively longer. Different metrics also have an impact on word- and sentence-based summarization. We observe

that sentence-based selection yields higher R-SU4 skip-bigram score but lower R-1 unigram scores.

Next we evaluate the impact of DA segmentation using ASR output. Figure 1 shows the segmentation and summarization results when varying the decision threshold from the DA segmentation output. We notice that there is degradation due to the use of ASR output, and that the patterns are in general similar to those in Table 1 in the sense that the best segmentation output may not be the best for summarization. In addition, there is some difference compared to using human transcripts. For example, sentence-based extraction is slightly worse than the word-based selection using both metrics. This might be because DA segmentation yields inappropriate segments using ASR output which affects more sentence-based summary selection.



**Fig. 1.** DA segmentation error rate (right Y-axis) and ROUGE F-measure scores (left Y-axis) for word- and sentence-based summary selection when varying the DA segmentation decision threshold using ASR output.

## 5.3. Discussions

Most of the previous work on meeting summarization used human annotated segments, therefore we cannot compare our results to

those. An investigation most related to this is [3], which used a different domain and is not directly comparable to our study.

The MMR summarization approach achieves reasonable performance—the ROUGE score using human transcripts and DA annotation is only slightly worse than the best results reported in [2]. Therefore, we think it is an acceptable system to use for evaluating the impact of DA segmentation. In the MMR approach, the different DA boundaries affect the similarity measure of a sentence to the entire document and to the selected summaries. It is likely that DA segmentation has a different impact on other summarization approaches, in which more lexical and prosodic features are used.

One reference point we use to evaluate the impact of DA segmentation on summarization is human annotated DA boundaries. However, the best unit in the MMR-based summarization system (or other approaches) may not be the DA units. For example, a question-answer pair might be combined as a unit for extractive summarization, or smaller units such as prosodic phrases may be more appropriate. We will investigate these in our future study.

Since ROUGE results have been reported for meeting summarization in other previous studies [1, 2], we chose to use it as the summarization evaluation metric in this work, where the focus is on the impact of sentence segmentation. However, the automatic ROUGE unigram and skip bigram scores may not properly reward or penalize those sentences with a wrong boundary. In addition, there are questions in general about whether ROUGE is a good evaluation metric for meeting summarization. [1] showed that ROUGE does not correlate well with human evaluations for the meeting domain. We also found that meeting characteristics (e.g., disfluencies and multiple speakers) affect the correlation of ROUGE score and human evaluation. Yet all the prior studies have only used reference sentence boundaries, and the impact of automatic segmentation on human evaluation of summarization is unclear and thus a further study is still needed.

## 6. CONCLUSION AND FUTURE WORK

Automatic sentence segmentation is a crucial first step for sentence-based extractive summarization systems. In this paper we used an HMM for sentence segmentation, and evaluated the effect of different segments on an MMR-based summarization system. By varying the decision threshold in the automatic segmentation output, different granularity of segments can be generated. We find that using automatic sentence segmentation degrades summarization performance compared to using human annotated segments, for both the human transcripts and ASR output. In addition, automatic segmentation needs to be optimized for the subsequent language processing task; however, the patterns of the segmentation effect on summarization are affected by factors such as summarization performance measures and ASR. This study will help us better understand the impact of structural information on speech summarization, in particular for the meeting domain.

In our future work, we will investigate the effect of sentence segmentation using other summarization approaches, especially when more lexical and acoustic features are used. We also plan to develop segmentation algorithms with the goal of generating appropriate segments for summarization, such as prosodic phrases or longer units containing question-answer pairs. Finally, we need to investigate the interaction of different summarization evaluation metrics and the impact of segmentation on summarization.

## 7. ACKNOWLEDGMENT

The authors thank Michel Galley for sharing his data processing tool, University of Edinburgh for sharing the meeting annotations, and Feifan Liu for useful discussions. This work is supported by NSF grant IIS-0714132. Any opinions expressed in this work are those of the authors and do not necessarily reflect the views of NSF.

## 8. REFERENCES

- [1] G. Murray, S. Renals, J. Carletta, and J. Moore, “Evaluating automatic summaries of meeting recordings,” in *Proc. of ACL 2005 MTSE Workshop*, 2005.
- [2] M. Galley, “A skip-chain conditional random field for ranking meeting utterances by importance,” in *Proc. of EMNLP*, 2006.
- [3] J. Mrozinski, E. Whittaker, P. Chatain, and S. Furui, “Automatic sentence segmentation of speech for automatic summarization,” in *Proc. of ICASSP*, 2006.
- [4] E. Matusov, D. Hillard, M. Magimai-Doss, D. Hakkani-Tur, M. Ostendorf, and H. Ney, “Improving speech translation with automatic prediction,” in *Proc. of Interspeech*, 2007.
- [5] M. Harper, B. Dorr, B. Roark, J. Hale, Z. Shafran, Y. Liu, M. Lease, M. Snover, L. Young, R. Stewart, and A. Krasnyanskaya, “Final report: parsing speech and structural event detection,” <http://www.clsp.jhu.edu/ws2005/groups/eventdetect/documents/finalreport.pdf>, 2005.
- [6] A. Janin, D. Baron, J. Edwards, D. Ellis, G. Gelbart, N. Morgan, B. Peshkin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, “The ICSI meeting corpus,” in *Proc. of ICASSP*, 2003.
- [7] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, “The ICSI meeting recorder dialog act (MRDA) corpus,” in *Proc. of 5th SIGDAL Workshop*, 2004.
- [8] J. Carletta, “Assessing agreement on classification tasks: The kappa statistic,” *Computational Linguistics*, vol. 22(2), pp. 249–254, 1996.
- [9] S. Cuendet, D. Tur, and G. Tur, “Modal adaptation for sentence segmentation from speech,” in *Proc. of IEEE Workshop on Spoken Language Technology*, 2006.
- [10] J. Kolar, E. Shriberg, and Y. Liu, “On speaker-specific prosodic models for automatic dialog act segmentation of multiparty meetings,” in *Proc. of Interspeech*, 2006.
- [11] A. Janin, A. Stolcke, J. Frankel, O. Cetin, K. Boakye, X. Anguera, and C. Wooters, “The ICSI-SRI spring 2006 meeting speech-to-text system,” in *Proc. of Workshop on MLMI*, 2006.
- [12] A. Stolcke, “SRILM - an extensible language modeling toolkit,” in *Proc. of ICSLP*, 2002.
- [13] National Institute of Standards and Technology, “Document understanding conferences (DUC),” <http://duc.nist.gov>.
- [14] S. Maskey and J. Hirschberg, “Comparing lexical, acoustic/prosodic, discourse, and structural features for speech summarization,” in *Proc. of Interspeech*, 2005.
- [15] S.Y. Kong and L.S. Lee, “Improved spoken document summarization using probabilistic latent semantic analysis (PLSA),” in *Proc. of ICASSP*, 2006.
- [16] J. Carbonell and J. Goldstein, “The use of MMR, diversity-based reranking for reordering documents and producing summaries,” in *Proc. of ACM SIGIR*, 1998.
- [17] D. Radev, S. Blair-Goldensohn, and Z. Zhang, “Experiments in single and multi-document summarization using MEAD,” in *Proc. of The First Document Understanding Conference*, 2001.
- [18] C. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Proc. of Workshop on Text Summarization Branches Out*, at *ACL*, 2004.