Extractive Summarization of Meeting Recordings

Gabriel Murray, Steve Renals, Jean Carletta

Centre for Speech Technology Research
University of Edinburgh, Edinburgh EH8 9LW, Scotland
g.murray-3@sms.ed.ac.uk, s.renals@ed.ac.uk, jeanc@inf.ed.ac.uk

Abstract

Several approaches to automatic speech summarization are discussed below, using the ICSI Meetings corpus. We contrast feature-based approaches using prosodic and lexical features with maximal marginal relevance and latent semantic analysis approaches to summarization. While the latter two techniques are borrowed directly from the field of text summarization, feature-based approaches using prosodic information are able to utilize characteristics unique to speech data. We also investigate how the summarization results might deteriorate when carried out on ASR output as opposed to manual transcripts. All of the summaries are of an extractive variety, and are compared using the software ROUGE.

1. Introduction

There is considerable research activity in text summarization (eg [1]); however, there has been less work in speech summarization. Most work in speech summarization has been in the domain of broadcast news [2, 3, 4]. It has been demonstrated that standard extractive text summarization techniques, using classifiers based on textual and structural features [5], work well on broadcast news transcripts [6].

Summarizing conversational speech is substantially different from text summarization. In addition to the problems that arise from speech recognition errors, the information density is quite different to textual documents, and information is also contained in the prosody of the speech signal. Christensen et al [7] provide evidence that more spontaneous parts of broadcast news (eg interviews) are less amenable to standard text summarization techniques. Zechner [8] reported experiments on the summarization of spoken multiparty dialogues, using an approach based on maximal marginal relevance (MMR) [9], with the addition of automatic speech disfluency removal, sentence boundary marking and question-answer pair detection.

These approaches to speech summarization are all based on processing speech recognition output. Hori et al [10] have developed an integrated speech summarization approach, based on finite state transducers, in which the recognition and summarization components are composed into a single finite state transducer, reporting results on a lecture summarization task.

In this paper we investigate extractive summarization of multiparty meetings, using the ICSI Meetings Corpus [11]. We have employed a number of summarization approaches, two of which are based on approaches to text summarization and two which are feature-based, including both prosodic and lexical features. The techniques borrowed from text summarization were MMR (which was used as a standard baseline) and latent semantic analysis (LSA). We also investigated whether an LSA-based sentence score could be used to supplement the prosodic and lexical features used in the feature-based approach. Our experiments were carried out using both human transcriptions and the output of an automatic speech recognizer, and we evaluated the quality of the summaries using ROUGE [12].

2. Summarization Approaches

2.1. Maximal Marginal Relevance (MMR)

MMR [9] is based on the vector-space model of text retrieval, and is well-suited to query-based and multi-document summarization. In MMR, sentences are chosen according to a weighted combination of their relevance to a query (or for generic summaries, their general relevance) and their redundancy with the sentences that have already been extracted. Both relevance and redundancy are measured using cosine similarity. Relevance would normally be the cosine similarity of the sentence and query vectors, but since this task consisted of generic rather than query-dependent summaries, relevance was determined by the cosine similarity of the sentence vector and a document vector representing the average of the sentence vectors for the complete meeting. The MMR score $Sc^{MMR}(i)$for a given sentence S_i in the document is given by

$$Sc^{MMR}(i) = \lambda(Sim(S_i, D)) - (1-\lambda)(Sim(S_i, Summ)),$$

where D is the average document vector, $Summ$ is the average vector from the set of sentences already selected, and λ trades off between relevance and redundancy. Sim is the cosine similarity between two documents.

In this implementation of MMR, the weight λ is annealed, so that relevance is emphasized when the summary is still short, and as the summary grows longer the emphasis is increasingly put on minimizing redundancy. For the first third of the summary, $\lambda = 0.7$, for the second third $\lambda = 0.5$, and for the final third of the summary $\lambda = 0.3$. It is likely that further experimentation is needed to determine the optimal annealing schedule.

2.2. Latent Semantic Analysis (LSA)

LSA is a vector-space approach which involves projection of the term-document matrix to a reduced dimension representation. It was originally applied to text retrieval [13], and has since been applied to a variety of other areas, including text summarization [14, 15]. LSA is based on the singular value decomposition (SVD) of an $m \times n$ term-document matrix A, whose elements A_{ij} represent the weighted term frequency of term i in document j. In SVD, the term-document matrix is decomposed as follows:

$$A = USV^T$$

where U is an $m \times n$ matrix of left-singular vectors, S is an $n \times n$ diagonal matrix of singular values, and V is the $n \times n$ matrix

of right-singular vectors. The rows of V^T may be regarded as defining topics, with the columns representing sentences from the document. Following Gong and Liu [14], summarization proceeds by choosing, for each row in V^T, the sentence with the highest value. This process continues until the desired summary length is reached.

Steinberger and Ježek [15] have offered two strong criticisms of the Gong and Liu approach. Firstly, the method described above ties the dimensionality reduction to the desired summary length. Secondly, a sentence may score highly but never "win" in any dimension, and thus will not be extracted despite being a good candidate. Steinberger and Ježek proposed a solution of extracting a single LSA-based sentence score, with variable dimensionality reduction.

We address the same concerns, following the Gong and Liu approach, but rather than extracting the best sentence for each topic, the n best sentences are extracted, with n determined by the corresponding singular values from matrix S. The number of sentences in the summary that will come from the first topic is determined by the percentage that the largest singular value represents out of the sum of all singular values, and so on for each topic. Thus, dimensionality reduction is no longer tied to summary length and more than one sentence per topic can be chosen. Using this method, the level of dimensionality reduction is essentially learned from the data.

2.3. Feature-Based Approaches

Feature-based approaches [5] have proven to be successful for both text and broadcast news summarization. In this work we augmented textual features with a set of prosodic features, using Gaussian mixture models for the extracted and non-extracted classes. The prosodic features were the mean and standard deviation of F0, energy, and duration, all estimated and normalized at the word-level, then averaged over the utterance. The two lexical features were both TFIDF-based: the average and the maximum TFIDF score for the utterance. For any word w in document j, the TFIDF score $Sc^{TFIDF}(w,j)$ is equal to $tf(w,j) \cdot \log(N/df(w))$, where N is the total number of documents, $df(w)$ is the number of documents containing w, and $tf(w)$ is the number of occurrences of w in j.

The second feature-based approach created single LSA-based sentence scores [15] which were used in addition to the six features above, in order to determine whether such a score is beneficial in determining sentence importance. We reduced the original term-document matrix to 300 dimensions, although Steinberger and Ježek found that reducing to a single dimension yielded the best results for their corpus (Steinberger, personal communication). The LSA sentence score was obtained using:

$$Sc_i^{LSA} = \sqrt{\sum_{k=1}^{n} v(i,k)^2 * \sigma(k)^2}\,,$$

where $v(i,k)$ is the kth element of the ith sentence vector and $\sigma(k)$ is the corresponding singular value.

3. Experimental setup

We used human summaries of the ICSI Meeting corpus for evaluation and for training the feature-based approaches. An evaluation set of six meetings was defined and multiple human summaries were created for these meetings, with each test meeting having either three or four manual summaries. The remaining meetings were regarded as training data and a single human summary was created for these. There is no standardized method for creating such summaries, and what they should look like depends on the uses to which they will be put. Ours were created as follows.

Annotators were given access to a graphical user interface (GUI) for browsing an individual meeting that included earlier human annotations: an orthographic transcription time-synchronized with the audio, and a topic segmentation based on a shallow hierarchical decomposition with keyword-based text labels describing each topic segment. Some of the summarization annotators had created the topic segmentation themselves in an earlier task; others had not seen the meetings before. The annotators were told to construct a textual summary of the meeting aimed at someone who is interested in the research being carried out, such as a researcher who does similar work elsewhere, using four headings:

- general abstract: "why are they meeting and what do they talk about?";
- decisions made by the group;
- progress and achievements;
- problems described

The annotators were given a 200 word limit for each heading, and told that there must be text for the general abstract, but that the other headings may have null annotations for some meetings. Annotators who were new to the data were encouraged to listen to a meeting straight through before beginning to author the summary.

Immediately after authoring a textual summary, annotators were asked to create an extractive summary, using a different GUI. This GUI showed both their textual summary and the orthographic transcription, without topic segmentation but with one line per dialogue act based on the pre-existing MRDA coding [16] (The dialogue act categories themselves were not displayed, just the segmentation). Annotators were told to extract dialogue acts that together would convey the information in the textual summary, and could be used to support the correctness of that summary. They were given no specific instructions about the number or percentage of acts to extract or about redundant dialogue act. For each dialogue act extracted, they were then required in a second pass to choose the sentences from the textual summary supported by the dialogue act, creating a many-to-many mapping between the recording and the textual summary. Although the expectation was that each extracted dialogue act and each summary sentence would be linked to something in the opposing resource, we told the annotators that under some circumstances dialogue acts and summary sentences could stand alone.

The MMR and LSA approaches are both unsupervised and do not require labelled training data. For both feature-based approaches, the GMM classifiers were trained on a subset of the training data representing approximately 20 hours of meetings.

We performed summarization using both the human transcripts and speech recognizer output. The speech recognizer output was created using baseline acoustic models created using a training set consisting of 300 hours of conversational telephone speech from the Switchboard and Callhome corpora. The resultant models (cross-word triphones trained on conversational side based cepstral mean normalised PLP features) were then MAP adapted to the meeting domain using the ICSI corpus [17]. A trigram language model was employed. Fair recognition output for the whole corpus was obtained by dividing the

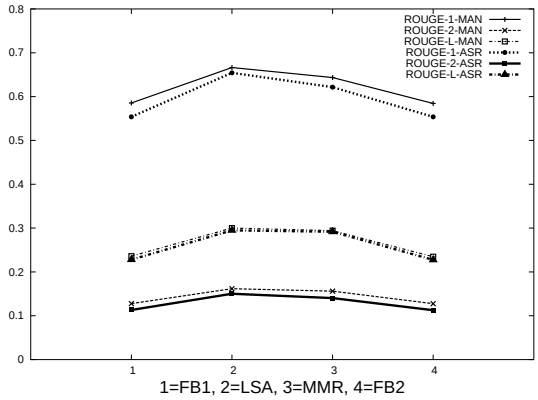

Figure 1: *ROUGE Scores for the Summarization Approaches*

corpus into four parts, and employing a leave one out procedure (training the acoustic and language models on three parts of the corpus and testing on the fourth, rotating to obtain recognition results for the full corpus). This resulted in an average word error rate (WER) of 29.5%. Automatic segmentation into dialogue acts or sentence boundaries was not performed: the dialogue act boundaries for the manual transcripts were mapped on to the speech recognition output.

Automatic evaluation of summarization is a very active research area, since subjective evaluation using human subjects is very time consuming. We used the ROUGE evaluation approach [12], which is based on n-gram co-occurrence between machine summaries and "ideal" human summaries. ROUGE is currently the standard objective evaluation measure for the Document Understanding Conference [1]; ROUGE does not assume that there is a single "gold standard" summary. Instead it operates by matching the target summary against a set of reference summaries. ROUGE-1 through ROUGE-4 are simple n-gram co-occurrence measures, which check whether each n-gram in the reference summary is contained in the machine summary. ROUGE-L and ROUGE-W are measures of common subsequences shared between two summaries, with ROUGE-W favoring contiguous common subsequences. Lin [12] has found that ROUGE-1 and ROUGE-2 correlate well with human judgments.

4. Results

All of the machine summaries were 10% of the original document length, in terms of the number of dialogue acts contained. Of the four approaches to summarization used herein, the latent semantic analysis method performed the best on every meeting tested for every ROUGE measure with the exception of ROUGE-3 and ROUGE-4. This approach was significantly better than either feature-based approach ($p<0.05$), but was not a significant improvement over MMR. For ROUGE-3 and ROUGE-4, none of the summarization approaches were significantly different from each other, owing to data sparsity. Figure 1 gives the ROUGE-1, ROUGE-2 and ROUGE-L results for each of the summarization approaches, on both manual and ASR transcripts.

The results of the four summarization approaches on ASR output were much the same, with LSA and MMR being comparable to each other, and each of them outperforming the feature-based approaches. On ASR output, LSA again consistently performed the best.

Interestingly, though the LSA approach scored higher when using manual transcripts than when using ASR transcripts, the difference was small and insignificant despite the nearly 30% WER of the ASR. All of the summarization approaches showed minimal deterioration when used on ASR output as compared to manual transcripts, but the LSA approach seemed particularly resilient, as evidenced by Figure 1. One reason for the relatively small impact of ASR output on summarization results is that for each of the 6 meetings, the WER of the summaries was lower than the WER of the meeting as a whole. Similarly, Valenza et al [2] and Zechner and Waibel [18] both observed that the WER of extracted summaries was significantly lower than the overall WER in the case of broadcast news. The table below demonstrates the discrepancy between summary WER and meeting WER for the six meetings used in this research.

Meeting	**Summary WER/%**	**Meeting WER/%**
Bed004	27.0	35.7
Bed009	28.3	39.8
Bed016	39.6	49.8
Bmr005	23.9	36.1
Bmr019	28.0	36.5
Bro018	25.9	35.6
WER Comparison for LSA Summaries and Whole Meetings		

There was no improvement in the second feature-based approach (adding an LSA sentence score) as compared with the first feature-based approach. The sentence score used here relied on a reduction to 300 dimensions, which may not have been ideal for this data.

In general, the comparable performance of LSA and MMR in this research reenforces some of Gong and Liu's key findings. In their work, implementations of LSA and MMR-style summarizers yielded very similar results, prompting the authors to claim that the relatively straightforward interpretation of the MMR algorithm is thus reflected in the more opaque LSA method. In other words, they make the strong claim that the singular vectors of V^T can be interpreted as topics or concepts, and that the LSA summarization method emphasizes relevance and minimizes redundancy.

5. Sample Summarization Output

Presented below are examples of part of a human summary and a corresponding part of an automatic summary using LSA, respectively. While they present roughly the same information, the automatic summary is relatively choppy and less clear.

- Snippet of Human Summary:

  > The experiment consisted of leading a subject to believe she were talking to a computer, then having the "computer" break down and be replaced with a human.

- Corresponding Snippet of LSA Summary:

  > I should say the system was supposed to break down and then these were the remaining three tasks that she was going to solve with a human. One time to pretending to be a human which is actually not pretending.

[1] http://duc.nist.gov/

- Corresponding Snippet of LSA-ASR Summary:

  > Reverse should so the system were supposed to break down and then this would be remaining three tasks that she was going to solve with a human.

As can be seen from the above examples, the ASR errors affect not only the readability of the summaries, but also which items are extracted in the first place, since the LSA and LSA-ASR approaches use different term/document matrices. The LSA-ASR summary does not contain the second utterance.

Ultimately, however, we want evaluations of the meeting summaries to be based on how useful human subjects determine them to be within the context of a meeting browser application. Reliance on such extrinisic measures is still critical for completely robust summarization evaluation.

6. Conclusion

Though the LSA method consistently performed the best, it was not a significant improvement over MMR and does not share some of the advantages of MMR. For example, MMR is ideal for query-based and multi-document summarization, and we eventually want users to be able to create query-based summaries of meetings they were unable to attend.

Though the feature-based approaches seemed to perform much worse than the others, it is unfortunately the case that finding the right features is not a trivial task, and the current work is preliminary in that it relies on a very small prosodic database. Adding pause and rate-of-speech information, for example, might prove very useful.

Surprisingly, extracting a single LSA score following the work of Steinberger and Ježek did not prove helpful. Further experimentation with the level of dimensionality reduction may yet replicate Steinberger and Ježek's success. A critical task is to lower the dimensionality without greatly biasing the major topics. 300 dimensions in this case was likely too high.

7. Future Work

The focus in the immediate future will be put on greatly expanding the prosodic database and on building various types of classifiers for the feature-based approach. An additional emphasis will be put on structural features to complement the prosodic and lexical features. A second avenue of research involves finding a method of automatic utterance detection, rather than relying on mapping the dialogue-act annotation to the ASR transcripts. Investigating disfluency removal and question-answer linking may also improve the extracted summaries.

Furthermore, extrinsic evaluation must be utilized in order to get a clearer picture of how useful these summaries will be within the context of a multimedia meeting browser.

8. Acknowledgements

Thanks to Thomas Hain and the AMI-ASR group for the speech recognition output. Thanks to Weiqun Xu for valuable assistance. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication).

9. References

[1] I. Mani, *Automatic Summarization*. John Benjamin, 2001.

[2] R. Valenza, T. Robinson, M. Hickey, and R. Tucker, "Summarization of spoken audio through information extraction," in *Proc. ESCA Workshop on Accessing Information in Spoken Audio*, 1999, pp. 111–116.

[3] M. J. Witbrock and V. O. Mittal, "Ultra-summarization: A statistical approach to generating highly condensed non-extractive summaries," in *Proc. ACM SIGIR '99*, 1999, pp. 315–316.

[4] C. Hori, S. Furui, R. malkin, H. Yu, and A. Waibel, "A statistical approach for automatic speech summarization," *EURASIP Journal on Applied Signal Processing*, vol. 2, pp. 128–139, 2003.

[5] J. Kupiec, J. Pederson, and F. Chen, "A trainable document summarizer," in *ACM SIGIR '95*, 1995, pp. 68–73.

[6] H. Christensen, Y. Gotoh, B. Kolluru, and S. Renals, "Are extractive text summarisation techniques portable to broadcast news?" in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, 2003.

[7] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals, "From text summarisation to style-specific summarisation for broadcast news," in *Proc. ECIR–2004*, 2004.

[8] K. Zechner, "Automatic summarization of open-domain multiparty dialogues in diverse genres," *Computational Linguistics*, vol. 28, no. 4, pp. 447–485, 2002.

[9] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proc. ACM SIGIR*, 1998, pp. 335–336.

[10] T. Hori, C. Hori, and Y. Minami, "Speech summarization using weighted finite-state transducers," in *Proc. Eurospeech*, 2003.

[11] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. IEEE ICASSP*, 2003.

[12] C.-Y. Lin and E. H. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proc. HLT-NAACL*, 2003.

[13] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391–407, 1990.

[14] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proc. ACM SIGIR*, 2001, pp. 19–25.

[15] J. Steinberger and K. Ježek, "Using latent semantic analysis in text summarization and summary evaluation," in *Proc. ISIM '04*, 2004, pp. 93–100.

[16] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, , and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. 5th SIGdial Workshop on Discourse and Dialogue*, 2004, pp. 97–100.

[17] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, I.Mc.Cowan, J.Vepa, and S.Renals, "An investigation into transcription of conference room meetings," *Submitted to Eurospeech*, 2005.

[18] K. Zechner and A. Waibel, "Minimizing word error rate in textual summaries of spoken language," in *Proc. NAACL-2000*, 2000.