Evaluating Automatic Summaries of Meeting Recordings

Gabriel Murray
Centre for Speech Technology Research
University of Edinburgh
Edinburgh, United Kingdom

Steve Renals
Centre for Speech Technology Research
University of Edinburgh
Edinburgh, United Kingdom

Jean Carletta
Human Communication Research Centre
University of Edinburgh
Edinburgh, United Kingdom

Johanna Moore
Human Communication Research Centre
University of Edinburgh
Edinburgh, United Kingdom

Abstract

The research below explores schemes for evaluating automatic summaries of business meetings, using the ICSI Meeting Corpus (Janin et al., 2003). Both automatic and subjective evaluations were carried out, with a central interest being whether or not the two types of evaluations correlate with each other. The evaluation metrics were used to compare and contrast differing approaches to automatic summarization, the deterioration of summary quality on ASR output versus manual transcripts, and to determine whether manual extracts are rated significantly higher than automatic extracts.

1 Introduction

In the field of automatic summarization, it is widely agreed upon that more attention needs to be paid to the development of standardized approaches to summarization evaluation. For example, the current incarnation of the Document Understanding Conference is putting its main focus on the development of evaluation schemes, including semi-automatic approaches to evaluation. One semi-automatic approach to evaluation is ROUGE (Lin and Hovy, 2003), which is primarily based on n-gram co-occurrence between automatic and human summaries. A key question of the research contained herein is how well ROUGE correlates with human judgments of summaries within the domain of meeting speech. If it is determined that the two types of evaluations correlate strongly, then ROUGE will likely be a valuable and robust evaluation tool in the development stage of a summarization system, when the cost of frequent human evaluations would be prohibitive.

Three basic approaches to summarization are evaluated and compared below: Maximal Marginal Relevance, Latent Semantic Analysis, and feature-based classification. The other major comparisons in this paper are between summaries on ASR versus manual transcripts, and between manual and automatic extracts. For example, regarding the former, it might be expected that summaries on ASR transcripts would be rated lower than summaries on manual transcripts, due to speech recognition errors. Regarding the comparison of manual and automatic extracts, the manual extracts can be thought of as a gold standard for the extraction task, representing the performance ceiling that the automatic approaches are aiming for.

More detailed descriptions of the summarization approaches and experimental setup can be found in (Murray et al., 2005). That work relied solely on ROUGE as an evaluation metric, and this paper proceeds to investigate whether ROUGE alone is a reliable metric for our summarization domain, by comparing the automatic scores with recently-gathered human evaluations. Also, it should be noted that while we are at the moment only utilizing intrinsic evaluation methods, our ultimate plan is to evaluate these meeting summaries extrinsically within the context of a meeting browser (Wellner et al., 2005).

Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 33–40, Ann Arbor, June 2005.

2 Description of the Summarization Approaches

2.1 Maximal Marginal Relevance (MMR)

MMR (Carbonell and Goldstein, 1998) uses the vector-space model of text retrieval and is particularly applicable to query-based and multi-document summarization. The MMR algorithm chooses sentences via a weighted combination of query-relevance and redundancy scores, both derived using cosine similarity. 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.

This implementation of MMR uses lambda annealing so that relevance is emphasized while the summary is still short and minimizing redundancy is prioritized more highly as the summary lengthens.

2.2 Latent Semantic Analysis (LSA)

LSA is a vector-space approach which involves projecting the original term-document matrix to a reduced dimension representation. It 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 (Gong and Liu, 2001), 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.

Two drawbacks of this method are that dimensionality is tied to summary length and that good sentence candidates may not be chosen if they do not "win" in any dimension (Steinberger and Ježek, 2004). The authors in (Steinberger and Ježek, 2004) found one solution, by 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 classification approaches have been widely used in text and speech summarization, with positive results (Kupiec et al., 1995). In this work we combined textual and 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 our second feature-based approach, we derived single LSA-based sentence scores (Steinberger and Ježek, 2004) to complement the six features described above, to determine whether such an LSA sentence score is beneficial in determining sentence importance. We reduced the original term-document matrix to 300 dimensions; however, Steinberger and Ježek found the greatest success in their work by reducing to a single dimension (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. Our summaries 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. 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.

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 (Shriberg et al., 2004) (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.

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 (Hain et al., 2005). A trigram language model was employed. Fair recognition output for the whole corpus was obtained by dividing the 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.

3.1 Description of the Evaluation Schemes

A particular interest in our research is how automatic measures of informativeness correlate with human judgments on the same criteria. During the development stage of a summarization system it is not feasible to employ many hours of manual evaluations, and so a critical issue is whether or not software packages such as ROUGE are able to measure informativeness in a way that correlates with subjective summarization evaluations.

3.1.1 ROUGE

Gauging informativeness has been the focus of automatic summarization evaluation research. We used the ROUGE evaluation approach (Lin and Hovy, 2003), 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 (Lin and Hovy, 2003) has found that ROUGE-1 and ROUGE-2 correlate well with human judgments.

3.1.2 Human Evalautions

The subjective evaluation portion of our research utilized 5 judges who had little or no familiarity with the content of the ICSI meetings. Each judge evaluated 10 summaries per meeting, for a total of sixty summaries. In order to familiarize themselves with a given meeting, they were provided with a human abstract of the meeting and the full transcript of the meeting with links to the audio. The human judges were instructed to read the abstract, and to consult the full transcript and audio as needed, with the entire familiarization stage not to exceed 20 minutes.

The judges were presented with 12 questions at the end of each summary, and were instructed that upon beginning the questionnaire they should not reconsult the summary itself. 6 of the questions regarded informativeness and 6 involved readability and coherence, though our current research concentrates on the informativeness evaluations. The evaluations used a Likert scale based on agreement or disagreement with statements, such as the following Informativeness statements:

1. The important points of the meeting are represented in the summary.
2. The summary avoids redundancy.
3. The summary sentences on average seem relevant.
4. The relationship between the importance of each topic and the amount of summary space given to that topic seems appropriate.
5. The summary is repetitive.
6. The summary contains unnecessary information.

Statements such as 2 and 5 above are measuring the same impressions, with the polarity of the statements merely reversed, in order to better gauge the reliability of the answers. The readability/coherence portion consisted of the following statements:

1. It is generally easy to tell whom or what is being referred to in the summary.
2. The summary has good continuity, i.e. the sentences seem to join smoothly from one to another.
3. The individual sentences on average are clear and well-formed.
4. The summary seems disjointed.
5. The summary is incoherent.
6. On average, individual sentences are poorly constructed.

It was not possible in this paper to gauge how responses to these readability statements correlate with automatic metrics, for the reason that automatic metrics of readability and coherence have not been widely discussed in the field of summarization. Though subjective evaluations of summaries are often divided into informativeness and readability questions, only automatic metrics of informativeness have been investigated in-depth by the summarization community. We believe that the development of automatic metrics for coherence and readability should be a high priority for researchers in summarization evaluation and plan on pursuing this avenue of research. For example, work on coherence in NLG (Lapata, 2003) could potentially inform summarization evaluation. Mani (Mani et al.,

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

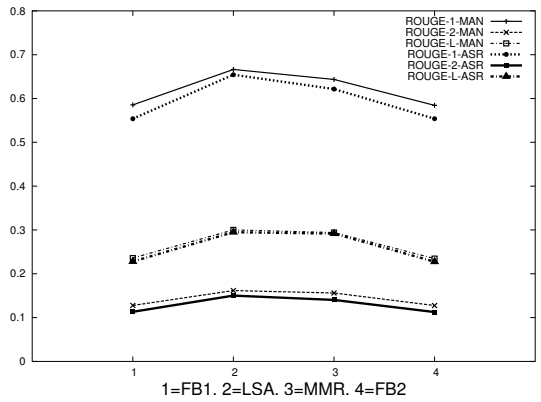

Figure 1: *ROUGE Scores for the Summarization Approaches*

1999) is one of the few papers to have discussed measuring summary readability automatically.

4 Results

The results of these experiments can be analyzed in various ways: significant differences of ROUGE results across summarization approaches, deterioration of ROUGE results on ASR versus manual transcripts, significant differences of human evaluations across summarization approaches, deterioration of human evaluations on ASR versus manual transcripts, and finally, the correlation between ROUGE and human evaluations.

4.1 ROUGE results across summarization approaches

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.

4.1.1 ASR versus Manual

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 (Valenza et al., 1999) and Zechner and Waibel (Zechner and Waibel, 2000) 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% for Summaries and 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.

The similarity between the MMR and LSA approaches here mirrors Gong and Liu's findings, giving credence to the claim that LSA maximizes relevance and minimizes redundancy, in a different and more opaque manner then MMR, but with similar

STATEMENT	FB1	LSA	MMR	FB2
IMPORT. POINTS	5.03	4.53	4.67	4.83
NO REDUN.	**4.33**	2.60	3.00	3.77
RELEVANT	4.83	4.07	4.33	4.53
TOPIC SPACE	4.43	3.83	3.87	4.30
REPETITIVE	**3.37**	4.70	4.60	3.83
UNNEC. INFO.	**4.70**	6.00	5.83	5.00

Table 1: Human Scores for 4 Approaches on Manual Transcripts

STATEMENT	FB1	LSA	MMR	FB2
IMPORT. POINTS	3.53	**4.13**	3.73	3.50
NO REDUN.	3.40	2.97	2.63	3.57
RELEVANT	3.47	3.57	3.00	3.47
TOPIC SPACE	3.27	3.33	3.00	3.20
REPETITIVE	4.43	4.73	4.70	4.20
UNNEC. INFO.	5.37	6.00	6.00	5.33

Table 2: Human Scores for 4 Approaches on ASR Transcripts

results. Regardless of whether or not the singular vectors of V^T can rightly be thought of as topics or concepts (a seemingly strong claim), the LSA approach was as successful as the more popular MMR algorithm.

4.2 Human results across summarization approaches

Table 1 presents average ratings for the six statements across four summarization approaches on manual transcripts. Interestingly, the first feature-based approach is given the highest marks on each criterion. For statements 2, 5 and 6 FB1 is significantly better than the other approaches. It is particularly surprising that FB1 would score well on statement 2, which concerns redundancy, given that MMR and LSA explicitly aim to reduce redundancy while the feature-based approaches are merely classifying utterances as relevant or not. The second feature-based approach was not significantly worse than the first on this score.

Considering the difficult task of evaluating ten extractive summaries per meeting, we are quite satisfied with the consistency of the human judges. For example, statements that were merely reworded versions of other statements were given consistent ratings. It was also the case that, with the exception of evaluating the sixth statement, judges were able to tell that the manual extracts were superior to the automatic approaches.

4.2.1 ASR versus Manual

Table 2 presents average ratings for the six statements across four summarization approaches on ASR transcripts. The LSA and MMR approaches performed better in terms of having less deterioration of scores when used on ASR output instead of manual transcripts. LSA-ASR was not significantly worse than LSA on any of the 6 ratings. MMR-ASR was significantly worse than MMR on only 3 of the 6. In contrast, FB1-ASR was significantly worse than FB1 for 5 of the 6 approaches, reinforcing the point that MMR and LSA seem to favor extracting utterances with fewer errors. Figures 2, 3 and 4 depict the how the ASR and manual approaches affect the INFORMATIVENESS-1, INFORMATIVENESS-4 and INFORMATIVENESS-6 ratings, respectively. Note that for Figure 6, a higher score is a worse rating.

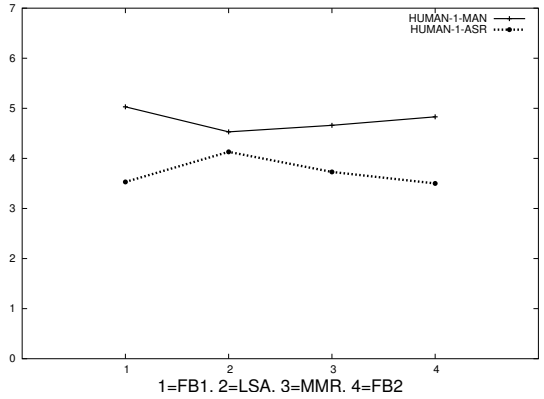

Figure 2: *INFORMATIVENESS-1 Scores for the Summarization Approaches*

4.3 ROUGE and Human correlations

According to (Lin and Hovy, 2003), ROUGE-1 correlates particularly well with human judgments of informativeness. In the human evaluation survey discussed here, the first statement (INFORMATIVENESS-1) would be expected to correlate most highly with ROUGE-1, as it is ask-

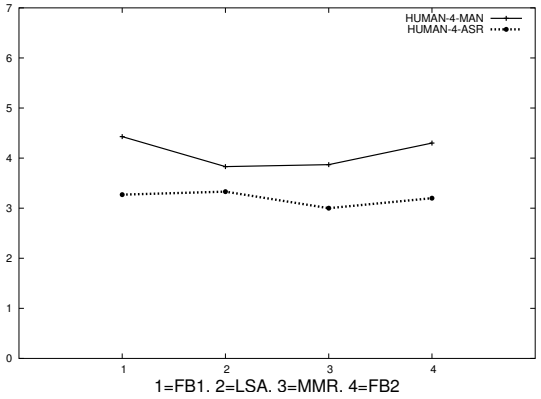

Figure 3: *INFORMATIVENESS-4 Scores for the Summarization Approaches*

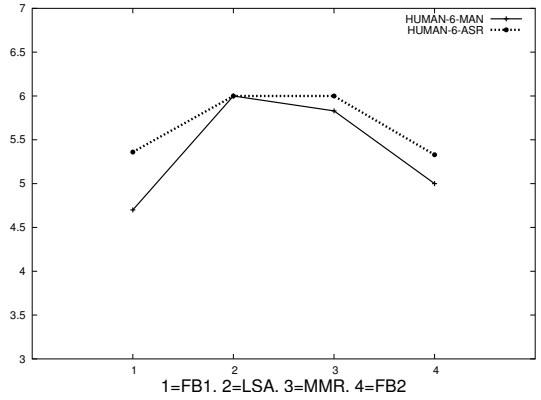

Figure 4: *INFORMATIVENESS-6 Scores for the Summarization Approaches*

ing whether the summary contains the important points of the meeting. As could be guessed from the discussion above, there is no significant correlation between ROUGE-1 and human evaluations when analyzing only the 4 summarization approaches on manual transcripts. However, when looking at the 4 approaches on ASR output, ROUGE-1 and INFORMATIVENESS-1 have a moderate and significant positive correlation (Spearman's rho = 0.500, $p < 0.05$). This correlation on ASR output is strong enough that when ROUGE-1 and INFORMATIVENESS-1 scores are tested for correlation across all 8 summarization approaches, there is a significant positive correlation (Spearman's rho = 0.388, $p < 0.05$).

The other significant correlations for ROUGE-1 across all 8 summarization approaches are with INFORMATIVENESS-2, INFORMATIVENESS-5 and INFORMATIVENESS-6. However, these are negative correlations. For example, with regard to INFORMATIVENESS-2, summaries that are rated as having a high level of redundancy are given high ROUGE-1 scores, and summaries with little redundancy are given low ROUGE-1 scores. Similary, with regard to INFORMATIVENESS-6, summaries that are said to have a great deal of unnecessary information are given high ROUGE-1 scores. It is difficult to interpret some of these negative correlations, as ROUGE does not measure redundancy and would not necessarily be expected to correlate with redundancy evaluations.

5 Discussion

In general, ROUGE did not correlate well with the human evaluations for this data. The MMR and LSA approaches were deemed to be significantly better than the feature-based approaches according to ROUGE, while these findings were reversed according to the human evaluations. An area of agreement, however, is that the LSA-ASR and MMR-ASR approaches have a small and insignificant decline in scores compared with the decline of scores for the feature-based approaches. One of the most interesting findings of this research is that MMR and LSA approaches used on ASR tend to select utterances with fewer ASR errors.

ROUGE has been shown to correlate well with human evaluations in DUC, when used on news corpora, but the summarization task here – using conversational speech from meetings – is quite different from summarizing news articles. ROUGE may simply be less applicable to this domain.

6 Future Work

It remains to be determined through further experimentation by researchers using various corpora whether or not ROUGE truly correlates well with human judgments. The results presented above are mixed in nature, but do not present ROUGE as being sufficient in itself to robustly evaluate a summarization system under development.

We are also interested in developing automatic metrics of coherence and readability. We now have human evaluations of these criteria and are ready to

begin testing for correlations between these subjective judgments and potential automatic metrics.

7 Acknowledgements

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

References

J. Carbonell and J. Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. ACM SIGIR*, pages 335–336.

Y. Gong and X. Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proc. ACM SIGIR*, pages 19–25.

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

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

J. Kupiec, J. Pederson, and F. Chen. 1995. A trainable document summarizer. In *ACM SIGIR '95*, pages 68–73.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *ACL*, pages 545–552.

C.-Y. Lin and E. H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. HLT-NAACL*.

Inderjeet Mani, Barbara Gates, and Eric Bloedorn. 1999. Improving summaries by revising them. In *Proceedings of the 37th conference on Association for Computational Linguistics*, pages 558–565, Morristown, NJ, USA. Association for Computational Linguistics.

G. Murray, S. Renals, and J. Carletta. 2005. Extractive summarization of meeting recordings. *Submitted to Eurospeech*.

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

J. Steinberger and K. Ježek. 2004. Using latent semantic analysis in text summarization and summary evaluation. In *Proc. ISIM '04*, pages 93–100.

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

Pierre Wellner, Mike Flynn, Simon Tucker, and Steve Whittaker. 2005. A meeting browser evaluation test. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 2021–2024, New York, NY, USA. ACM Press.

K. Zechner and A. Waibel. 2000. Minimizing word error rate in textual summaries of spoken language. In *Proc. NAACL-2000*.