Evaluation of Sentence Selection for Speech Summarization

Xiaodan Zhu and **Gerald Penn**
Department of Computer Science
University of Toronto
{xzhu, gpenn}@cs.toronto.edu

Abstract

In the last several years, a number of papers have addressed the area of automatic speech summarization. Many of them have applied evaluation metrics adapted from those used in speech recognition research, rather than from those used in text summarization. We consider whether ASR-inspired evaluation metrics produce different results than those taken from text summarization, and why. We evaluate various standard summarizers as well as our own systems on a subset of the SWITCHBOARD spoken dialogue dataset with both kinds of metrics. We find a statistically significant departure between the two classes in their relative rank of these systems. Our preliminary conclusion is that considerably greater caution must be exercised when using ASR-based measures than we have witnessed to date in the speech summarization literature.

Keywords: speech summarization, evaluation, sentence selection, extractive summarization.

1. Introduction

The goal of speech summarization is to distill important information from speech data. Similar to text summarization, most current speech summarizers are extractive rather than abstractive, that is, they extract and present pieces of original speech transcripts or audio data as the output rather than rephrase or rewrite them. The pieces of audio to be extracted could correspond to words. (Koumpis 02), for example, extracts important words from transcribed voicemail messages using classification algorithms. (Hori & Furui 03) extracts words from broadcast news by selecting a path that maximizes a predefined score. (Valenza et al. 99) extracts N-grams, as well as keywords.

The extracts could be sentences, too. Sentence selection is useful. First, it could be a preliminary stage applied before word extraction, as proposed by (Kikuchi et al. 03) in their two-stage summarizer. Second, with sentence-level extracts, one can play the corresponding audio to users, as with the speech-to-speech summarizer discussed in (Furui et al. 03). In this paper, we will focus on sentence-level extraction, which at present appears to be the only way to ensure comprehensibility if the summaries are to be delivered as excerpts of audio themselves. There are various methods proposed for sentence selection from speech transcripts, which will be discussed below.

Many of these summarizers are evaluated on precision/recall metrics, where a 0/1 value is assigned to each sentence by human judges, indicating whether this sentence should occur in the summary or not. This could be problematic, according to (Radev & Tam 03) and (Radev et al. 04), because, while there generally is a relatively high amount of interjudge agreement on which sentences are important, the selection of the top N% important sentences can still differ widely with respect to binary judgements. (Radev & Tam 03) proposes using *relative utility* (RU) instead to evaluate text summarizers. In this paper, we are interested in whether different evaluation metrics affect speech summarizers in the same way, and if so, why. We evaluate several summarizers as well as our own system on a spoken dialogue dataset with different evaluation metrics. We find a statistically significant departure between the two classes in their relative rankings of these systems.

(Hori et al. 03) evaluates speech summarizers using several variations of the well-known word error rate (WER) and word accuracy measures from speech recognition, as well as BLEU scores, but comparisons to metrics used in text summarization are conspicuously absent. Indeed, the best metric reported there, weighted summarization accuracy, is the only one that incorporates posterior weights that combine annotator preferences, in a manner reminiscent of RU (although the existence of RU itself may not have been widely known yet). Several of the methods they compare, moreover, including weighted summarization accuracy, crucially rely on word-level extraction, which carries with it its own problems with comprehension, particularly if the summaries are to be delivered in audio. (Zechner 00) considers sentence-level extraction from spoken language transcripts, but also proposes a score (also

called summarization accuracy) based on word accuracy.

Our study also bears a certain similarity to (Radev et al. 03) in its motivation. They introduce the relevance *correlation* metric, which is based on intersubstitutability of summaries in an information retrieval task. Again, because we are focusing on speech rather than text, search and retrieval take on an altogether different complexity, and so we have not as yet adapted this metric to our purposes.

The remainder of this paper is structured as follows: Section 2 discusses sentence-selection-based summarization of speech. Section 3 introduces the evaluations metrics used in this paper. Section 4 presents our experimental results. Section 5 discusses the results.

2. Speech Summarization by Sentence Selection

Having identified sentence boundaries, the most straightforward approach to sentence selection is to select the first N% of sentences from the beginning of the transcript. We refer to this strategy here as LEAD. The performance of LEAD is good on some datasets. However, LEAD serves more often as the baseline for evaluation, together with a summarizer that randomly guesses (RAND).

As for state-of-the-art research, (Zechner 01) applies maximum marginal relevance (MMR) to select sentences for open domain spoken dialogue transcripts. (Kikuchi et al. 03) selects sentences to maximize a predefined score that linearly combines linguistic, significance and confidence scores. (Maskey & Hirschberg 03) proposes using Bayesian networks to capture structural features so as to select important segments of speakers' turns. (Gurevych & Strube 04) developed a shallow knowledge-based approach to extract the essential utterances from dialogue data also. (Christensen et al. 04) applies multilayer perceptron networks to one-sentence extractive summarization of broadcast news data.

In our experiments, we reimplemented two of these approaches: MMR, as described in (Zechner 01), and the shallow knowledge-based approach described in (Gurevych & Strube 04). We refer to them as MMR and SEM respectively. In addition, we have implemented our own summarizers. We use SVM and logistic regression to include more features for sentence selection, which are referred to as SVM and LOG respectively in this paper. The remainder of this section discusses these summarizers briefly.

2.1 Knowledge-based Approach

(Gurevych & Strube 04) developed a shallow knowledge-based approach to extract essential utterances from spoken dialogue transcription. To calculate semantic similarity between a given utterance and the dialogue, the noun portion of WordNet is used as a knowledge source, with semantic distance between senses computed using (Leacock & Chodorow 98) normalized path length. The performance of the system is reported as better than LEAD, RAND and TF*IDF based methods. However, the noun senses were manually disambiguated rather than automatically. In our reimplementation, we simply use the most frequent sense of each noun. We applied Brill's POS tagger to acquire the nouns. According to (Gurevych & Strube 04), several other widely used measures perform close to Leacock-Chodorow for summarization on SWITCHBOARD data. The experiments reported below use (Pedersen 02) semantic similarity package.

2.2 MMR-based Approach

(Zechner 01) applies maximum marginal relevance (MMR) to select sentences for open domain spoken dialogue transcripts. MMR selects sentences with the following formulae:

$$nextsent = \arg\max_{t_{nr,j}} (\lambda sim1(t_{nr,j}, query) - (1-\lambda) \max_{t_{r,k}} sim2(t_{nr,j}, t_{r,k}))$$

MMR ranks sentences by their relevance. It selects the next unranked sentence into the rank according to two criteria: (1) whether it is more similar to the whole dialogue (Sim1 in the formula), and (2) whether it is less similar (Sim2) to the sentences that have so far been selected. Parameter . linearly combines these two properties. In our experiment, the "query" is a vector for the content words of the spoken dialogue to be summarized. In (Zechner 01), MMR is combined with sentence boundary detection, false start detection, repetition filtering, detection of question-answering pairs, and topic segmentation.

2.3 Classification-Based Approaches

In our own methods, we also formulate sentence selection as a binary classification problem. A sentence can either be included in a summary or not. We exploit more features, however, such as those shown in the following table:

Feature Types	Features	Descriptions
Content features	Similarities	Similarity (relevance) scores output from MMR
	Redundancy	Redundancy (dissimilarity) scores output from MMR
	Named Entity	Indicate how many named entities are contained in this sentence.
	Question	Indicate whether the sentence is a question sentence or not.
Structural features	Position	Indicate a sentence's position: in the first, mid-, or last one-third of the file
	Length	Length of a sentence
Spoken language features	Disfluency	Indicate disfluencies contained in the sentence: UH-word, discourse markers, editing term, etc.
	Repetition	Indicate the number of repetitions in a sentence

Table 1. Features used in classifiers

Of the many classification methods we have experimented with, the best two have consistently been SVM and logistic regression.

2.3.1 SVM

A support vector machine (SVM) is a supervised learning technique based on the principle of structural risk minimization. SVM seeks an optimal separating hyperplane, where the margin is maximal. For linearly non-separable samples, SVMs employ the "kernel trick" to implicitly transform the problem to a high-dimensional feature space. The training of SVM solves a quadratic programming problem. In the testing phase, for an input example x, the decision function is:

$$f(x) = \mathrm{sgn}(\sum_{j=1}^{N_s} a_j y_j K(s_j, x))$$

In our experiment, we use OSU-SVM (Ma et al.) package.

2.3.2 Logistic Regression

Logistic regression strives to model the posterior probabilities of the class label with linear functions:

$$\log \frac{p(Y=k \mid X=x)}{p(Y=K \mid X=x)} = \beta_{k0} + \beta_k{}^T x$$

X are feature sets and Y are class labels. For the binary classification that we require in our experiments, the model is especially simple:

$$p(Y=1 \mid X=x) = \frac{\exp(\beta_{10} + \beta_0{}^T x)}{1+\exp(\beta_{10} + \beta_1{}^T x)}$$

$$p(Y=2 \mid X=x) = \frac{1}{1+\exp(\beta_{10} + \beta_1{}^T x)}$$

The detailed discussion can be found in (Hastie et al. 01).

3. Evaluation Metrics

3.1 Precision /Recall

Precision/recall and F-measure are standard evaluation metrics for many NLP tasks. However, as pointed out in (Radev & Tam 03) and (Radev et al. 04), they are not satisfactory. As an example (taken from (Radev et al. 04)), consider two hypothetical summarizers, sys1 and sys2, which select two sentences from four S1-S4, as shown in Table 2. The importance of these four sentences is annotated in both binary values and integers between 0 and 9.

	Utility and binary annotation	Sys1	Sys2
S1	10 (+)	+	-
S2	9 (+)	-	-
S3	8 (-)	+	+
S4	7 (-)	-	+

Table 2. Examples for calculating precision/recall and relative utility score

When evaluated on binary annotations and using precision/recall metrics, sys1 and sys2 achieve 50% and 0%, respectively; this is unintuitive.

3.2 Relative Utility

Relative Utility was first proposed by (Radev & Tam 03), and aims to more closely match our intuition on such examples. For the above example, if using relative utility, sys1 gets 18/19 and sys2 gets 15/19. These results are more reasonable. Relative utility is calculated based on the formulae in (Radev & Tam 03)

While relative utility is, in our view, a very intuitive idea, some of the interpretations of its performance by (Radev & Tam 03) seem less than convincing. The fact that J (inter-judge performance) exceeds the average inter-annotator agreement that they witness in their P/R evaluation, for example, certainly means that RU is measuring something different, but does not necessarily mean that RU is actually a better measure; the values obtained are higher than with P/R, but they are higher for all of the systems evaluated, including the random baselines. In addition, J is not necessarily an upper

bound on system performance, as J and S have been defined in that paper. In our evaluation below, we find that RU makes roughly the same predictions as binary P/R measures. This is also true of (Radev & Tam 03) evaluation. As with our re-interpretation of their results, furthermore, we can only say that the two appear to be correlated - not that one is better than the other. In our view, the only ways to support such a claim would be to conduct an independent human evaluation of the summaries themselves and compare, or likewise to compare them with some other extrinsic evaluation such as a task-based retrieval study.

3.3 Word Error Rate

The length of sentences in speech data could be very short or very long. In addition to evaluating the summarizers by regarding each sentence as a single unit, we also compare their performance at the word level. Word error rate as used in the experiments is defined as the sum of insertion error, substitution error and deletion error of words, divided by the number of all these errors plus the number of correct words.

3.4 Zechner's Summarization Accuracy

(Zechner & Waibel 00) proposes *summarization accuracy* scores to evaluate the summarizers on both manual and automatic transcripts. The judges' annotations are averaged together to produce a single relevance score for words. The summarization accuracy is defined as the sum of the relevance scores of all the words in the automatic summary, divided by the maximum achievable relevance score with the same number of words.

3.5 ROUGE

ROUGE (Lin 04) is a widely used evaluation package for text summarization. It evaluates a summary against gold standards by measuring overlapping units such as n-grams, word sequences, and word pairs. In this paper, we report results on those ROUGE metrics that have been used in Document Understanding Conferences (DUC, 2004), i.e. ROUGE-N and ROUGE-L.

4. Experiments

4.1 Precision/Recall

The corpus we use for our experiments is the SWITCHBOARD dataset. SWITCHBOARD is a corpus of open-domain spoken dialogue; many extractive speech summarizers report experimental results on it (Zechner, 2000; Gurevych & Strube 04). We randomly select 27 spoken dialogues from SWITCHBOARD. Three annotators are asked to assign 0/1 labels to indicate whether a sentence is in the summary or not. The annotators are required to select around 10% of the sentences into the summary. The P/R agreement is shown in Table 3. The values shown in the cells are the F-scores obtained when we evaluate one judge's annotation relative to another.

	Judge1	Judge2	Judge3
Judge1	-	0.51	0.45
Judge2	0.51	-	0.42
Judge3	0.45	0.42	-

Table 3. Agreement between annotators

We can obtain several gold standards by combining these three annotations. One standard marks a sentence as in the summary only when all three annotators agree. We evaluate the summarizers discussed above relative to this dataset in Table 4, which shows the F-measures obtained by varying the summary length. We only present the results on lengths that make realistic sense (5-40%). SEM stands for the knowledge-based approach with the semantic distance measure discussed above. LOG is logistic regression.

%	RAND	LEAD	SEM	MMR	LOG	SVM
5	.03	.07	.12	.24	.31	.30
10	.04	.15	.14	.23	.25	.25
20	.04	.11	.15	.19	.18	.18
30	.05	.09	.13	.15	.15	.14
40	.05	.08	.12	.12	.12	.11

Table 4. F-measure of summarizers on P/R dataset 1

LOG and SVM have similar performance and outperform the others, with MMR following, and then SEM and LEAD. The performance of SEM is a worse than what (Gurevych & Strube 04) reports, probably because we did not spend the effort to manually disambiguate the nouns (which in our view cannot really count as an automatic method anyway).

The second P/R evaluation standard is acquired by including those sentences that at least two of the three judges include in the summary. The performance of the summarizers on this standard is shown in Table 5.

%	RAND	LEAD	SEM	MMR	LOG	SVM
5	.06	.17	.18	.35	.44	.44
10	.09	.22	.27	.44	.46	.46
20	.11	.22	.35	.44	.44	.44
30	.13	.20	.34	.40	.39	.39
40	.13	.19	.31	.33	.33	.32

Table 5. F-measure of summarizers on P/R dataset 2

The F-measures in Table 5 are higher than those in Table 4 because there are more sentences in the gold standard of dataset 2, and therefore a higher random chance for a true positive. However, the performance rank of the summarizers is still the same as in the first standard.

As a third standard, we take a summary to consist of sentences annotated by any of the three annotators as belonging. The performances of the summarizers are shown in Table 6. Again, we witness the same rank.

%	RAND	LEAD	SEM	MMR	LOG	SVM
5	.08	.12	.27	.33	.34	.33
10	.13	.20	.41	.49	.52	.52
20	.18	.25	.55	.66	.67	.67
30	.22	.28	.60	.67	.67	.67
40	.24	.29	.56	.60	.59	.59

Table 6. F-measure of summarizers on P/R data set 3

4.2 Relative Utility

For the same SWITCHBOARD subset, but for three different human judges, we also obtained an assignment of a number between 0 and 9 to each sentence, to indicate the confidence that this sentence should be included in the summary. We calculate the inter-judge performance J, random performance R, system performance S and normalized Relative Utility D, in the same way proposed in (Radev & Tam 03) and (Radev et al. 04). The results are shown in Table 7.

%	J	R	LEAD		SEM		MMR		LOG		SVM	
			S	D	S	D	S	D	S	D	S	D
5	.60	.09	.29	.39	.34	.49	.50	.80	.63	1.06	.64	1.08
10	.61	.11	.30	.38	.40	.58	.55	.88	.64	1.06	.64	1.06
20	.67	.20	.37	.36	.62	.89	.75	1.17	.78	1.23	.79	1.26
30	.71	.30	.49	.46	.82	1.27	.92	1.51	.93	1.54	.94	1.56
40	.75	.40	.57	.49	.94	1.54	.99	1.69	.98	1.66	.99	1.69

Table 7. Relative Utility

We first observe that the performance ranks of the five summarizers are the same here as they are in the three P/R evaluations. This might be due to several reasons. First, the P/R agreement among annotators is not low, ranging between 42% and 51%. Actually it is much higher than the data used by (Radev & Tam 03) and (Radev et al. 04), where the P/R agreement is between 25% and 29% when 10% is selected for the summary length. Higher P/R agreement decreases the usefulness of relative utility. Second, the redundancy in the data is much less than in the multi-document summarization tasks used in (Radev & Tam 03) and (Radev et al. 04). If there are more redundant sentences in the data, the summarizers have more chances to choose different sentences without loss. From this point of view, RUs might be more suitable for multi-document summarization. Third, the summarizers we compare might tend to select the same sentences.

We may also observe that these trainable classifiers improve the performance of sentence selection under RU and the three P/R evaluation metrics, because of their ability to avail themselves of more features, including spoken-language features. This agrees with our intuition.

4.3 Word Error Rate and Summarization Accuracy

Turning to the classical WER measure from speech recognition, we see in each of the P/R gold standards a remarkably better performance for both LEAD and RAND (this is an error rate so smaller numbers are better). LEAD overtakes the real systems, beginning in dataset 1 (Table 8) at 10% summary length, and in dataset 2 (Table 9) at 30%. In dataset 3 (Table 10), the inflexion point is beyond 40%, although it is already close. Here, the difference is mainly a reflection of the overall difference in magnitude among the three datasets, with SEM, MMR, LOG and SVM steadily decreasing in WER as the threshold for judicial agreement decreases – and thus the number of positive sentences to choose from increases.

%	RAND	LEAD	SEM	MMR	LOG	SVM
5	.96	.84	.90	.81	.76	.74
10	.93	.76	.90	.85	.84	.83
20	.92	.80	.90	.88	.88	.88
30	.92	.85	.90	.90	.90	.90
40	.92	.87	.91	.91	.91	.91

Table 8. WER on Dataset 1.

%	RAND	LEAD	SEM	MMR	LOG	SVM
5	.95	.88	.79	.64	.60	.61
10	.90	.77	.73	.60	.58	.59
20	.82	.67	.71	.67	.67	.66
30	.78	.66	.72	.71	.71	.70
40	.76	.64	.73	.73	.73	.73

Table 9. WER on Dataset 2.

%	RAND	LEAD	SEM	MMR	LOG	SVM
5	.95	.93	.74	.70	.69	.68
10	.90	.86	.58	.52	.49	.48
20	.80	.74	.42	.35	.35	.34
30	.70	.64	.42	.40	.40	.41
40	.62	.56	.44	.45	.45	.45

Table 10. WER on Dataset 3.

In the case of Zechner's summarization accuracy score, the score is computed by averaging the judges' annotations together to produce a single weight.

Given this single gold standard, there is no clear best system with this metric. One may note, however, the prominence of Zechner's MMR system with respect to summarization accuracy. (Zechner & Waibel 00) does not mention the use of separate development and evaluation test sets, so it is possible that the metric itself evolved to work well with MMR. In any case, the clear preference for LOG and SVM observed above is not in evidence here.

%	RAND	LEAD	SEM	MMR	LOG	SVM
5	.26	.56	.45	.63	.66	.67
10	.31	.55	.54	.72	.70	.71
20	.36	.54	.72	.87	.88	.87
30	.42	.54	.89	.97	.96	.96
40	.47	.60	.97	1.00	.99	.99

Table 11. Zechner's SA scores on the averaged judgements.

4.4 ROUGE

The following tables provide the results for the ROUGE metrics.

%	LEAD	SEM	MMR	LOG	SVM
5	.22	.26	.27	.29	.28
10	.37	.40	.45	.46	.46
20	.51	.57	.61	.63	.63
30	.57	.63	.66	.67	.68
40	.58	.62	.65	.65	.66

Table 12. Results for ROUGE-1 metric.

%	LEAD	SEM	MMR	LOG	SVM
5	.09	.15	.19	.23	.21
10	.16	.21	.31	.32	.33
20	.24	.33	.41	.44	.44
30	.29	.42	.47	.48	.50
40	.35	.43	.49	.50	.51

Table 13. Results for ROUGE-2 metric.

%	LEAD	SEM	MMR	LOG	SVM
5	.06	.12	.17	.21	.19
10	.10	.16	.27	.28	.29
20	.17	.26	.36	.39	.38
30	.21	.36	.41	.43	.45
40	.27	.37	.44	.46	.46

Table 14. Results for ROUGE-3 metric.

%	LEAD	SEM	MMR	LOG	SVM
5	.05	.11	.16	.20	.18
10	.09	.14	.25	.26	.27
20	.15	.24	.34	.37	.36
30	.19	.34	.39	.41	.43
40	.25	.35	.42	.44	.44

Table 15. Results for ROUGE-4 metric.

%	LEAD	SEM	MMR	LOG	SVM
5	.21	.25	.26	.28	.27
10	.35	.38	.43	.44	.44
20	.49	.54	.59	.62	.61
30	.54	.62	.65	.65	.66
40	.56	.60	.64	.64	.64

Table 16. Results for ROUGE-L metric.

The results on ROUGE agree with the other two text-summarization-inspired metrics, RU and P/R: SVM and LOG are the best in all ROUGE metrics listed above, followed by MMR, and then SEM.

5. Conclusion

In this paper, we considered whether ASR-inspired evaluation metrics produce different results than those taken from text summarization. We evaluated five summarizers on three text-summarization-inspired metrics: precision/recall (P/R), relative utility (RU), and ROUGE; as well as on two ASR-inspired evaluation metrics: word error rate (WER) and summarization accuracy (SA). We observe that the performance ranks of the five summarizers are consistent in the three text-summarization-inspired metrics. The more complicated metrics such as RU and ROUGE do not produce different results on our SWITCHBOARD sample, compared with the simple P/R metric. For ASR-inspired evaluation metrics, we find obvious differences in the relative rankings of these systems. We have not yet done the subjective experiments to extrinsically validate our belief that P/R-based, RU and ROUGE metrics are better for speech summarization than WER and summarization accuracy scores. However, our preliminary conclusion is that considerably greater caution must be exercised when using ASR-based measures than we have witnessed to date in the speech summarization literature.

References

(Christensen et al. 04) Christensen, H., Kolluru, B., Gotoh, Y., Renals, S., 2004. *From text summarisation to style-specific summarisation for broadcast news*, In Proc. ECIR-2004.

(Furui et al. 03) Furui, S., Kikuichi T. Shinnaka Y., and Hori C. 2003. *Speech-to-speech and speech to text summarization*,. First International workshop on Language Understanding and Agents for Real World Interaction, 2003.

(Gurevych & Strube 04) Gurevych I. and Strube M.. 2004. *Semantic Similarity Applied to Spoken Dialogue Summarization*. In Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, 23-27 August 2004, p.p. 764-770.

(Hastie et al. 01) Hastie, T., Tibshirani, R., Friedman, J., 2001. *THE ELEMENTS OF STATISTICAL LEARNING: Data Mining, Inference and Prediction*, Springer Series in Statistics, 2001

(Hori & Furui 03) Hori C. and Furui S., 2003. *A New Approach to Automatic Speech Summarization IEEE Transactions on Multimedia*, Vol. 5, NO. 3, SEPTEMBER 2003, pp. 368-378.

(Hori et al. 03) Hori C, Takaaki Hori and Sadaoki Furui, 2003b, *Evaluation Methods for Automatic Speech Summarization*, Proc. Eurospeech2003, Geneva. 2003.

(Kikuchi et al. 03) Kikuchi T., Furui S. and Hori C., 2003. *Automatic Speech Summarization Based on Sentence Extraction and Compaction*, Proc. ICASSP2003, Hongkong, Vol. I, pp 384-387

(Koumpis 02) Koumpis K., 2002. *Automatic Voicemail Summarisation for Mobile* Messaging Ph.D. Thesis in Computer Science, University of Sheffield, UK, 2002.

(Leacock & Chodorow 98) Leacock C. and Chodorow M.. 1998. *Combining local context with WordNet similarity for word sense identification*. In C. Fellbaum, editor, WordNet: A Lexical Reference System and its Application. MIT Press, Cambridge.

(Lin 04) Lin C., 2004. *Rouge: a package for automatic evaluation of summaries*. In Proceedings of 42st Annual Meeting of the Association for Computational Linguistics, Workshop on Text Summarization Branches Out.

(Ma et al.) Ma, J., Zhao, Y., Ahalt, S. OSU SVM Classifier Matlab Toolbox. http://www.ece.osu.edu/~maj/osu_svm/

(Maskey & Hirschberg 03) Maskey S., Hirschberg J., 2003. *Automatic Summarization of Broadcast News using Structural Features*, Eurospeech 2003

(Pedersen 02) Pedersen, T. 2002. Semantic Similarity Package. http://www.d.umn.edu/~tpederse/similarity.html.

(Radev et al. 04) Radev D., Jing H., Stys M., and Tam D., 2004. *Centroid-based summarization of multiple documents*. Information Processing and Management, 40:919-938, December 2004.

(Radev & Tam 03) Radev D. and Tam D., 2003. *Single-document and multi-document summary evaluation via relative utility*. In CIKM 2003 poster session, New Orleans, LA, November 2003.

(Radev et al. 03) Radev, D., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Qi, H., Celebi, A., Liu, D., and Drabek, E. 2003b. *Evaluation challenges in large-scale document summarization*. Proc. Of 41st ACL, pp. 375-382.

(Valenza et al. 99) Valenza B., Robinson T., Hickey M., Tucker R., 1999. *Summarisation of Spoken Audio Through Information Extraction*, Proceedings of the ESCA ETRW workshop, 1999.

(Zechner & Waibel 00) Zechner K. and Waibel A., 2000. *Minimizing word error rate in textual summaries of spoken language*. In Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics, NAACL-2000, Seattle, WA, April/May, pages 186-193, 2000.

(Zechner 01) Zechner K., 2001. *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*. Ph.D. thesis, Carnegie Mellon University, School of Computer Science, Language Technologies Institute, November 2001.