Procedia Computer Science

Volume 51, 2015, Pages 620–629

ICCS 2015 International Conference On Computational Science

Blending Sentence Optimization Weights of Unsupervised Approaches for Extractive Speech Summarization

Noraini Seman and Nursuriati Jamil

Computer Science Department, Faculty of Computer & Mathematical Sciences
Digital Image, Audio and Speech Technology (DIAST) Research Group,
Advanced Computing and Communication Communities of Research,
Universiti Teknologi MARA (UiTM), 40450 Shah Alam,
Selangor Darul Ehsan, Malaysia
{aini, liza}@tmsk.uitm.edu.my

Abstract

This paper evaluates the performance of two unsupervised approaches, Maximum Marginal Relevance (MMR) and concept-based global optimization framework for speech summarization. Automatic summarization is very useful techniques that can help the users browse a large amount of data. This study focuses on automatic extractive summarization on multi-dialogue speech corpus. We propose improved methods by blending each unsupervised approach at sentence level. Sentence level information is leveraged to improve the linguistic quality of selected summaries. First, these scores are used to filter sentences for concept extraction and concept weight computation. Second, we pre-select a subset of candidate summary sentences according to their sentence weights. Last, we extend the optimization function to a joint optimization of concept and sentence weights to cover both important concepts and sentences. Our experimental results show that these methods can improve the system performance comparing to the concept-based optimization baseline for both human transcripts and ASR output. The best scores are achieved by combining all three approaches, which are significantly better than the baseline system.

Keywords: Maximum Marginal Relevance (MMR), Global optimization, Extractive summarization, Multi-dialogue

1 Introduction

With massive amounts of speech recordings and multimedia data available, an important problem is how to efficiently process these data to meet user's information need. There has been increasing

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015

doi:10.1016/j.procs.2015.05.330

interest recently in automatically processing the speech data, including summarization, and other understanding tasks in the research community (Hain et al., 2008; Mostefa et al., 2007). Automatic summarization is a very useful technique to facilitate users to browse large amount of data efficiently. Efficient speech summarization saves time for reviewing speech documents and improves the efficiency of document retrieval. Summarization can be divided into different categories along several different dimensions (Mani & Maybury, 1999). Based on whether or not there is an input query, the generated summary can be query-oriented or generic; based on the number of input documents, summarization can use a single document or multiple documents; in terms of how sentences in the summary are formed, summarization can be conducted using either extraction or abstraction — the former only selects sentences from the original documents, whereas the latter involves natural language generation (Furui et al., 2003). Overall, automatic summarization systems aim to generate a good summary, which is expected to be concise, informative, and relevant to the original input.

A lot of techniques and approaches have been proposed for automatic text summarization the past decades to summarize and process the speech data, a natural solution is to transcribe the speech recordings to texts, and apply some well studied text summarization approaches. However, usually when the traditional Natural Language Processing (NLP) methods are directly applied to speech transcripts, the performance is not as good as for text processing. There are several issues that make the speech transcripts different from the written texts. First, the speech transcripts often have a lot of disfluencies, while written texts are normally well formed and organized. This is especially the case for spontaneous speech domains, such as multiparty meetings, conversational telephone speech. Second, if we use the output of the automatic speech recognition (ASR) system for summarization, the output is actually only word sequence, and there are no punctuation marks or sentence segments associated with it. Therefore we first need to segment the word sequence into pieces to use as summarization units. However, these automatically segmented sentences are still different from the human annotated linguistic sentences. This often affects the down-streaming language understanding performance. Lastly, the speech transcripts contain word errors, especially for the conversational speech. For example, for meeting recordings, the word error rate could be as high as around 40%. With such lower accuracy, it is hard to read the transcripts and generate a summary even for human annotators. If the cue words or cue phrases are not correctly recognized, it will have great impact on the selection of important sentences (Xie, 2010).

All of these issues make speech summarization different from text summarization, and the traditional text summarization approaches generally do not perform well in speech summarization task. This paper focuses on the task of extractive summarization using the multi-dialogue speech data. Given the multi-dialogue speech document (recording together with its transcript), our aim is to select the most important and representative parts, and concatenate them together to form a summary. Automatic summarization on the multi-dialogue speech domain is more challenging comparing to summarization of other speech genres, such as broadcast news, lectures, and voice mail. Different from broadcast news and lectures, which are read or pre-planned speech, multi-dialogue speech are more spontaneous, so the transcripts contain a lot of disfluencies, such as filled pauses, repetitions, revisions, and etc. Although some dialogues have pre-defined topics, in general the content of dialogue is less coherent than other speech genres. For speech recognition, the ASR performance in spontaneous speech such as in meeting and dialogue domain is also worse than other speech domains. Such noisy data has great impact on the performance of traditional summarization approaches.

Automatic speech summarization has received a lot of attentions in recent years, and different approaches have been explored. Maximum Marginal Relevance (MMR) and Latent semantic analysis (LSA) are examples of unsupervised approaches, which are relatively simple, and robust to different corpora (Furui et al., 2003). The summary sentences are usually selected according to their own importance, and their relationship to other sentences in the document. Another line of work for extractive speech summarization is based on supervised methods, where all the utterances in a document are divided into two classes, in summary or not, then the summarization task can be

considered as a binary classification problem (Sameer & Hirschberg, 2005). Although a large amount of labelled data is necessary for training the classifier, supervised approaches usually achieve better performance comparing to the unsupervised ones. Various models have been investigated for this classification task, such as Bayesian network (Zhu & Penn, 2006), maximum entropy (Anne et al., 2005), support vector machines (SVM) (Zhang et al., 2014), hidden Markov model (HMM) (Sameer & Hirschberg, 2005), and conditional random fields (CRF) (Galley, 2006).

In this research, we study two unsupervised approaches for extractive multi-dialogue speech summarization, maximum marginal relevance (MMR) and a concept-based global optimization framework. Among all the approaches for summarization, MMR is one of the simplest techniques, and has been effectively used for text and speech summarization (Carbonell & Goldstein, 1998). The extractive summarization problem can also be modeled using a global optimization framework based on the assumption that sentences contain independent concepts of information, and that the quality of a summary can be measured by the total value of unique concepts it contains (Garg et al., 2009). In this work, we propose to blend these two unsupervised methods to obtain better summarization performance.

The direction of this work is composed into several sections. Section 2 provides an overview of the speech data collection. The details of the methods and implementation are described in section 3, broken into two parts which is described the experimental setup and the evaluation measure. In section 4, the experimental result are reported and discussed. Finally, the conclusion is drawn in section 5.

2 Multi-dialogue Speech Data Collection

The speech data used in this research is gathered from multi-dialogue *hansard* documents of Malaysian House of Parliament. The *hansard* documents consists of Dewan Rakyat (DR) Parliamentary debates session for the year 2008 (Seman, 2012 & Seman et al., 2010). It contains spontaneous and formal speeches and it is the daily records of words spoken by 222 elected members of DR. The *hansard* documents comprises of 51 live videos and audio recording files (.avi form) of daily parliamentary session and 42 text files (.pdf form). Each part of parliamentary session contains six to eight hours spoken speeches surrounded with medium noise condition or environment ($\geq$ 30 dB), speakers interruption (Malay, Chinese and Indian races) and different speaking styles (low, medium and high intonation or shouting). The reason of choosing this kind of data is due to its naturalness and spontaneous speaking styles during each session. For the purpose of this study, eight hours of one day Parliament session document was selected as our experimental data. The document collection consists of 12 topics, 120 speakers and a total of 148,411 spoken words.

All the data collection will be resample at 16 kHz for further preprocessing stage purposes. The multi-dialogue speech contains filled pauses, restarts, interjections, unknown or mispronounced words, ellipsis and also ungrammatical construction were spoken by three different speakers in one session of debates. Many phenomena contains in the spontaneous type of speech that will causes problem for current system that are listed previously. All of the phenomena violate constraints to increase performance currently used by the speech recognizer.

3 Methods And Implementation

3.1 Maximum Marginal Relevance (MMR)

MMR is a greedy algorithm and was introduced in (Carbonell & Goldstein, 1998) for text summarization, and has been applied to speech summarization. This algorithm can select the most relevant sentences, and at the same time avoid the redundancy by removing the sentences that are too similar to the already selected ones. The summary sentences with the highest scores are selected iteratively and the weight for each sentence is calculated using two similarity functions (Sim_1) and (Sim_2) as shown in Equation (1), representing the similarity of a sentence to the entire document and to the selected summary, respectively.

$$MMR(S_i) = \lambda \times Sim_1(S_i, D) - (1-\lambda) \times Sim_2(S_i, Summ) \quad (1)$$

where D is the document vector, *Summ* represents the sentences that have been extracted into the summary, and λ is used to adjust the combined score to emphasize the relevance or to avoid redundancy.

For each sentence, we calculate its similarity to all the other sentences that have a higher similarity score to the document (according to the results of (Sim_1)), and use it as an approximation for (Sim_2). Therefore, the summary selection process only needs to find the top sentences that have high combined scores, which is an offline processing. Another approximation we use is not to consider all the sentences in the document, but rather only a small percent of sentences (based on a predefined percentage) that have a high similarity score to the entire document. Our hypothesis is that the sentences that are closely related to the document are worth being selected. An important part in MMR is how we can appropriately represent the similarity of two text segments. We adopted cosine similarity measure, which we use as our baseline in this study.

In this approach, each document (or a sentence) is represented using a vector space model. The cosine similarity between two vectors (D_1, D_2) is:

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

where t_i is the term weight for a word w_i, for which we use the TF-IDF (term frequency, inverse document frequency) value, as widely used in information retrieval. The IDF weighting is used to represent the specificity of a word: a higher weight means a word is specific to a document, and a lower weight means a word is common across many documents. IDF values are generally obtained from a large corpus as follows:

$$IDF(w_i) = \log(N / N_i) \quad (3)$$

where N_i is the number of documents containing w_i in a collection of N documents. Murray and Renals (2007) compared different term weighting approaches to rank the importance of the sentences (simply based on the sum of all the term weights in a sentence) for meeting summarization, and showed that TF-IDF weighting is competitive.

3.2 Sentence Weights in Global Optimization Framework

The MMR method we used in the previous section is local optimal because the decision is made based on the sentences' scores in the current iteration. There was a studied of modeling the multi-document summarization problem using a global inference algorithm with some definition of relevance and redundancy (Ryan, 2006). The Integer Linear Programming (ILP) solver was used to efficiently search a large space of possible summaries for an optimal solution. Other researchers adopted the global optimization framework assuming that concepts are the minimum units for summarization, and summary sentences are selected to cover as many concepts as possible (Gillick et al., 2009). They showed better performance than MMR. However, this method tends to select short sentences with fewer concepts in order to increase the number of concepts covered, instead of selecting sentences rich in concepts even if they overlap. According to manual examination, this seems to result in the degradation of the linguistic quality of the summary. In this section, we try to blend the sentence importance weights in the concept-based optimization method, and we propose different ways to use sentence weights.

First, we use a similar framework to build the baseline system and the global optimization function is adopted (Gillick et al., 2009) as in Equations (4) to (5).

$$\max imize \sum_i w_i c_i \tag{4}$$

$$subject\ to \sum_j l_j s_j < L \tag{5}$$

where w_i is the weight of concept i, c_i is a binary variable indicating the presence of that concept in the summary, l_j is the length of sentence j, L is the desired summary length, and s_j represents whether a sentence is selected for inclusion in the summary. Integer linear programming method was used to select sentences that maximize the objective function under the length constraint, L. In our research, we use the following procedure for concept extraction (Xie, 2010), which is slightly different from the previous work (Gillick et al., 2009), where they used the rule-based algorithm for concept selection.

- Extract all content word n-grams for $n = 1,2,3$.
- Remove the n-grams appearing only once.
- Remove the n-grams if one of its word's IDF value is lower than a predefined threshold.
- Remove the n-grams enclosed by other higher-order n-grams, if they have the same frequency.
- For example, we remove /`kasih/` if its frequency is the same as /`terima kasih/`.
- Weight each n-gram k_i as:

$$w_i = frequency(k_i) * n * \max_j idf(word_j) \tag{6}$$

where n is the n-gram length, and $word_j$ goes through all the words in the n-gram.

The IDF values are also calculated using the new "documents" split according to the topic segmentation using Equation (3). Unlike Gillick (2009) and Xie (2010), we use the IDF values to remove less informative words instead of using a manually generated stopword list, and also use IDF

information to compute the final weights of the extracted concepts. Furthermore, we do not use WordNet or part-of-speech tag constraints during the extraction. Therefore, using this new algorithm, the concepts are created automatically, without requiring much human knowledge. We use this method as the baseline for our research in the section.

Since the global optimization model is unsupervised, the sentence weights that can also be obtained in an unsupervised fashion. The cosine similarity scores between each sentence and the entire document, which is also adopted as the baseline in the MMR method calculated using Equation (2). We propose different ways to blend these sentence scores in the concept-based optimization framework.

3.2.1 Filtering Sentences for Concept Generation

First we use sentence weights to select important sentences, and then extract concepts from the selected sentences only. The only difference is that they are generated based on this subset of sentences, instead of the entire document. Once the concepts are extracted, the optimization framework is the same as before.

3.2.2 Pruning Sentences from the Selection

Sentence weights can also be used to filter unlikely summary sentences and pre-select a subset of candidate sentences for summarization, rather than considering all the sentences in the document. We use the same method to generate the summary as in Section 3.1, but only using preserved candidate sentences.

Finally, we extend the optimization function in Equation (4) to consider sentence importance weights, as shown in Equation (7) below.

$$\max imize(1-\lambda)*\sum_{i} w_i c_i + \lambda * \sum_{j} u_j s_j \tag{7}$$

where u_j is the weight for sentence j, λ is used to balance the weights for concepts and sentences, and all the other notations are the same as in Equation (4). The summary length constraint is the same as Equation (5). After adding the sentence weights in the optimization function, this model will select a subset of relevant sentences which can cover the important concepts, as well as the important sentences.

4 Experimental Results

We first use the development set to evaluate the effectiveness of our proposed approaches and the impact of various parameters in those methods, and then provide the final results on the test set. All experiments.

4.1 Baseline Results

Several baseline results are provided in Table 2 using different word compression ratios for both human transcripts and ASR output on the development set. The first one (long sentence) is to construct the summary by selecting the longest sentences, which has been shown to provide competitive results for multi-dialogue summarization task. The second one (MMR) is using cosine similarity as the similarity measure on the MMR framework. The last result (concept-based) is from the concept-based

algorithm as previously mentioned. These scores are comparable with those presented in Gillick (2009). For both human transcripts and ASR output, the longest-sentence baseline is worse than the greedy MMR approach, which, in turn, is worse than the concept-based algorithm. The performance on human transcripts is consistently better than on ASR output because of the high WER. In the following experiments, we will use the concept-based summarization results as the baseline, and a 16% word compression ratio.

Table 2. ROUGE-1 F-measure results (%) of three baselines on the development set for both human transcripts (REF) and ASR output.

compression		14%	15%	16%	17%	18%
REF	long sentence	65.50	67.16	68.47	69.58	69.23
	MMR	77.81	77.06	77.90	77.64	77.09
	concept-based	78.20	78.98	78.30	78.82	78.51
ASR	long sentence	74.11	75.01	75.72	75.65	75.89
	MMR	74.60	75.32	75.80	76.03	76.14
	concept-based	74.99	76.04	76.45	76.44	76.30

4.2 Filtering Sentences for Concept Generation

In Figure 1, we show the results on the development set using different percentages of important sentences for concept extraction. When the percentage of the sentences is 100%, the result is the same as the baseline using all the sentences. We observe that using a subset of important sentences outperforms using all the sentences for both human transcripts and ASR output. For human transcripts, using 30% sentences yields the best ROUGE score 0.6996, while for ASR output, the best result, 0.6604, is obtained using 70% sentences.

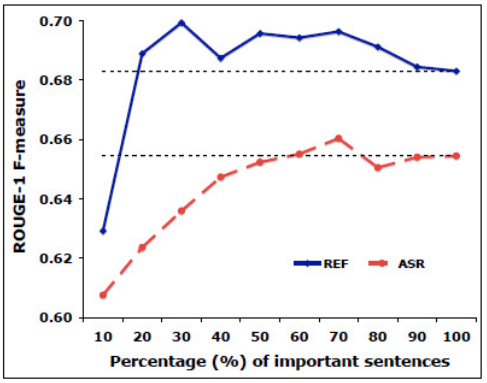

Figure 1. ROUGE-1 F-measure results (%) using different percentage of important sentences during concept extraction on the development set for both human transcripts (REF) and ASR output. The horizontal dashed lines represent the scores of the baselines using all the sentences.

4.3 Pruning Sentences from the Selection

This experiment evaluates the impact of using sentence weights to prune sentences and preselect summary candidates. Figure 2 shows the results of preserving different percentages of candidate sentences in the concept-based optimization model. For this experiment, we use the concepts extracted

from the original document. For both human transcripts and ASR output, using a subset of candidate sentences can significantly improve the performance, where the best results are obtained using 20% candidate sentences for human transcripts and 30% for ASR output. We also evaluate a length-based sentence selection and find that it is inferior to sentence score based pruning.

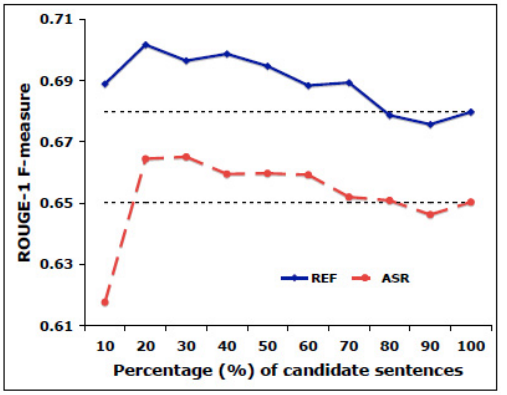

Figure 2. ROUGE-1 F-measure results (%) using pruning to preserve different percentage of candidate summary sentences on the dev set for both human transcripts (REF) and ASR output. The horizontal dashed lines represent the scores of the baselines using all the sentences.

4.4 Concept and Sentence Optimization Weights

Finally we evaluate the impact of incorporating sentence scores in the global optimization framework using Equation (7). We use all the sentences from the documents for concept extraction and sentence selection. All sentences are weighted according to their cosine scores, and the λ parameter is used to balance them with concept weights. Our experimental results show that sentence-level scores did not improve performance for most of the values of λ and sometimes hurt performance. An explanation for this disappointing result is that raw sentence weights do not seem to be suitable in a global model because sentences of very different length can have similar scores. In particular, the cosine score is normalized by the total TF-IDF weight of the words of a sentence, which gives high scores to short sentences containing high-weight words. For example, if two one-word sentences with a score of 0:9 are in the summary, they contribute 1:8 to the objective function while one two-word sentence with a better score of 1:0 only contributes 1:0 to the summary. To eliminate this problem, raw cosine scores need to be rescaled to ensure a fair comparison of sentences of different length. Therefore, in addition to using raw cosine similarity scores as the weights for sentences, we consider two variations: multiplying the cosine scores by the number of concepts and the number of words in that sentence, respectively.

Table 3. ROUGE-1 F-measure results (%) on the development set for both human transcripts (REF) and ASR output.

	baseline	*raw cosine*	*#concept norm*	*#words norm*
REF	77.20	77.32	77.32	77.40
ASR	74.35	70.02	74.08	75.19

Table 3 presents results for these three methods together with the baseline scores. We can see that for human transcripts when adding cosine similarity sentence weights, the result is slightly better than the

baseline. For the ASR condition, adding the cosine similarity sentence weights significantly degrades performance compared to the baseline. Reweighting the sentence scores using the number of concepts does not improve the performance; however, we observe better results by reweighting the scores based on the number of words, with more improvement on the ASR condition.

Table 4. ROUGE-1 F-measure results (%) of incorporating sentence importance weights on the dev set using both human transcripts (REF) and ASR output.

	baseline	*sentence weight for*			*all*
		concept	*pruning*	*joint*	
REF	78.40	79.96	82.50	79.60	**80.15**
ASR	75.55	76.14	76.51	76.29	**76.13**

Table 4 summarizes the results for various approaches. In addition to using each method alone, we also combine them, that is, we use sentence weights for concept extraction, and use a pre-selected set of sentences in the global optimization framework in combination with the concept scores. The best scores are obtained by combining all the proposed approaches for incorporating sentence importance weights. Among them, pruning contributes the highest scores using this approach alone can achieve very similar results to the best scores.

4.5 Results on Test Set

The experimental results on the test set using all the approaches proposed in this section are shown in Table 5. The parameter values are selected according to the performance on the development set. The baseline results are calculated using the concept-based summarization model, obtaining comparable results to the ones presented in Gillick (2009). ROUGE-1 scores are improved using our proposed three approaches for blending sentence importance weights: for concept extraction, selecting the candidate summary sentences, and extending the global optimization function with reweighted sentence weights. The best results are obtained by a combination of these methods, which is consistent with our findings on the development set. The improvement is consistent for both human transcripts and ASR output. We also verified that the results are significantly better than the baseline according to a paired t-test ($p < 0.05$).

Table 5. ROUGE-1 F-measure results (%) for different word compression ratios on test set for both human transcripts (REF) and ASR output.

compression		14%	15%	16%	17%	18%
REF	baseline	77.08	77.84	78.35	78.82	79.00
	concept	78.75	79.80	80.07	80.24	79.77
	pruning	78.85	79.30	80.10	80.33	80.43
	joint	77.48	78.40	78.97	79.19	79.16
	all	**79.35**	**80.29**	**80.87**	**80.72**	**80.30**
ASR	baseline	73.30	74.51	75.31	75.27	75.84
	concept	74.00	75.44	76.15	76.52	76.39
	pruning	75.83	76.78	76.63	76.79	76.48
	joint	73.82	74.76	75.80	76.11	75.77
	all	**75.87**	**76.67**	**77.07**	**77.20**	**76.91**

5 Conclusions

In this work, we studied and blend two unsupervised learning approaches for extractive meeting summarization. Under the baseline framework of MMR, cosine similarity measurement is used to better measure the semantic level information. Another unsupervised approach we evaluated is a global optimization framework. Sentence level information is combined to improve the linguistic quality of selected summaries. First, these scores are used to filter sentences for concept extraction and concept weight computation. Second, we pre-select a subset of candidate summary sentences according to their sentence weights. Finally, we extend the optimization function to a joint optimization of concept and sentence weights to cover both important concepts and sentences. Our experimental results show that these methods can improve the system performance comparing to the concept-based optimization baseline for both human transcripts and ASR output. The best scores are achieved by combining all three approaches, which are significantly better than the baseline.

References

Anne, H. B; Wessel, K. and Stephan, R. (2005). *Automatic summarization of meeting data: A feasibility study*, in Proc. of CLIN.

Chin, Y. L. (2004). *ROUGE: A package for automatic evaluation of summaries*, in Proc. Of The Workshop on Text Summarization Branches Out.

Defense Advanced Research Project Agency, "Translingual information detection, extraction and summarization (TIDES)," http://projects.ldc.upenn.edu/TIDES/ index.html.

Djamel, M.; Nicolas, M.; Khalid, C.; Potamianos, G.; Stephen M. C.; Ambrish, T.; Josep, R. C.; Jordi, T.; Luca, C.; Francesco, T.; Aristodemos, P.; Vassilis, M.; Fotios, T.; Susanne, B.; Rainer, S.; Keni, B. and Cedrick, R. (2007). *The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms*, in Journal of Language Resources and Evaluation, vol. 41, pp. 389–407.

Dragomir, R.; Hongyan, J. and Malgorzata, B. (2000). *Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies*, in NAACL-ANLP 2000 Workshop on Automatic summarization.

Hain, T.; Lukas, B.; John, D.; Giulia, G.; Martin, K; Leeuwen, D.; Lincoln, M. and Vincent, W. (2008). *The 2007 AMI(DA) system for meeting transcription*, in Journal of Multimodel Technologies for Perception of Humans, 2008, vol. 4625, pp. 414–428.

Inderjeet, M. and Mark, T. M. (1999). Advances in Automatic Text Summarization, MIT Press, 1999.

Jaime, C. and Jade, G. (1998). *The use of MMR, diversity-based reranking for reordering documents and producing summaries*, in Proc. of SIGIR.

Michel, G. (2006). *A skip-chain conditional random field for ranking meeting utterances by importance*, in Proc. of EMNLP.

Peter, T. (2001). *Mining the web for synonyms: PMI–IR versus LSA on TOEFL*, in Proc. of the 12th European Conference on Machine Learning.

Rada, M.; Courtney, C. and Carl, S. (2006). *Corpus-based and knowledgebased measures of text semantic similarity*, in Proc. of the American Association for Artificial Intelligence.

Sameer, M. and Hirschberg, J. (2006). *Summarizing speech without text using Hidden Markov Models*, in Proc. of HLT-NAACL.

Sameer, M. and Hirschberg, J. (2003). *Automatic summarization of broadcast news using structural features*, in Proc. of Eurospeech.

Zhu, X. and Penn, G. (2006). *Summarization of spontaneous conversations*, in Proc. of Interspeech, 2006.