Using Confusion Networks for Speech Summarization

Shasha Xie and Yang Liu
Department of Computer Science
The University of Texas at Dallas
{shasha,yangl}@hlt.utdallas.edu

Abstract

For extractive meeting summarization, previous studies have shown performance degradation when using speech recognition transcripts because of the relatively high speech recognition errors on meeting recordings. In this paper we investigated using confusion networks to improve the summarization performance on the ASR condition under an unsupervised framework by considering more word candidates and their confidence scores. Our experimental results showed improved summarization performance using our proposed approach, with more contribution from leveraging the confidence scores. We also observed that using these rich speech recognition results can extract similar or even better summary segments than using human transcripts.

1 Introduction

Speech summarization has received increasing interest recently. It is a very useful technique that can help users to browse a large amount of speech recordings. The problem we study in this paper is extractive meeting summarization, which selects the most representative segments from the meeting transcripts to form a summary. Compared to text summarization, speech summarization is more challenging because of not only its more spontaneous style, but also word errors in automatic speech recognition (ASR) output. Intuitively the incorrect words have a negative impact on downstream summarization performance. Previous research has evaluated summarization using either the human transcripts or ASR output with word errors. Most of the prior work showed that performance using ASR output is consistently lower (to different extent) comparing to that using human transcripts no matter whether supervised or unsupervised approaches were used.

To address the problem caused by imperfect recognition transcripts, in this paper we investigate using rich speech recognition results for summarization. N-best hypotheses, word lattices, and confusion networks have been widely used as an interface between ASR and subsequent spoken language processing tasks, such as machine translation, spoken document retrieval (Chelba et al., 2007; Chia et al., 2008), and shown outperforming using 1-best hypotheses. However, studies using these rich speech recognition results for speech summarization are very limited. In this paper, we demonstrate the feasibility of using confusion networks under an unsupervised MMR (maximum marginal relevance) framework to improve summarization performance. Our experimental results show better performance over using 1-best hypotheses with more improvement observed from using confidence measure of the words. Moreover, we find that the selected summary segments are similar to or even better than those generated using human transcripts.

2 Related Work

Many techniques have been proposed for the meeting summarization task, including both unsupervised and supervised approaches. Since we use unsupervised methods in this study, we will not describe previous work using supervised approaches because of the space limit. Unsupervised meth-

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 46–54,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

ods are simple and robust to different corpora, and do not need any human labeled data for training. MMR was introduced in (Carbonell and Goldstein, 1998) for text summarization, and was used widely in meeting summarization (Murray et al., 2005a; Xie and Liu, 2008). Latent semantic analysis (LSA) approaches have also been used (Murray et al., 2005a), which can better measure document similarity at the semantic level rather than relying on literal word matching. In (Gillick et al., 2009), the authors introduced a concept-based global optimization framework using integer linear programming (ILP), where concepts were used as the minimum units, and the important sentences were extracted to cover as many concepts as possible. They showed better performance than MMR. In a follow-up study, (Xie et al., 2009) incorporated sentence information in this ILP framework. Graph-based methods, such as LexRank (Erkan and Radev, 2004), have been originally used for extractive text summarization, where the document is modeled as a graph and sentences as nodes, and sentences are ranked according to its similarity with other nodes. (Garg et al., 2009) proposed ClusterRank, a modified graph-based method in order to take into account the conversational speech style in meetings. Recently (Lin et al., 2009) suggested to formulate the summarization task as optimizing submodular functions defined on the document's semantic graph, and showed better performance comparing to other graph-based approaches.

Rich speech recognition results, such as N-best hypotheses and confusion networks, were first used in multi-pass ASR systems to improve speech recognition performance (Stolcke et al., 1997; Mangu et al., 2000). They have been widely used in many subsequent spoken language processing tasks, such as machine translation, spoken document understanding and retrieval. Confusion network decoding was applied to combine the outputs of multiple machine translation systems (Sim et al., 2007; Matusov et al., 2006). In the task of spoken document retrieval, (Chia et al., 2008) proposed to compute the expected word counts from document and query lattices, and estimate the statistical models from these counts, and reported better retrieval accuracy than using only 1-best transcripts. (Hakkani-Tur et al., 2006) investigated using confusion networks for name entity detection and extraction and user intent classification. They also obtained better performance than using ASR 1-best output.

There is very limited previous work using more than 1-best ASR output for speech summarization. Several studies used acoustic confidence scores in the 1-best ASR hypothesis in the summarization systems (Valenza et al., 1999; Zechner and Waibel, 2000; Hori and Furui, 2003). (Liu et al., 2010) evaluated using n-best hypotheses for meeting summarization, and showed improved performance with the gain coming mainly from the first few candidates. In (Lin and Chen, 2009), confusion networks and position specific posterior lattices were considered in a generative summarization framework for Chinese broadcast news summarization, and they showed promising results by using more ASR hypotheses. We investigate using confusion networks for meeting summarization in this study. This work differs from (Lin and Chen, 2009) in terms of the language and genre used in the summarization task, as well as the summarization approaches. We also perform more analysis on the impact of confidence scores, different pruning methods, and different ways to present system summaries.

3 Summarization Approach

In this section, we first describe the baseline summarization framework, and then how we apply it to confusion networks.

3.1 Maximum Marginal Relevance (MMR)

MMR is a widely used unsupervised approach in text and speech summarization, and has been shown perform well. We chose this method as the basic framework for summarization because of its simplicity and efficiency. We expect this is a good starting point for the study of feasibility of using confusion networks for summarization. For each sentence segment S_i in one document D, its score ($MMR(i)$) is calculated using Equation 1 according to its similarity to the entire document ($Sim_1(S_i, D)$) and the similarity to the already extracted summary ($Sim_2(S_i, Summ)$).

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

where parameter λ is used to balance the two factors to ensure the selected summary sentences are relevant to the entire document (thus important), and compact enough (by removing redundancy with the currently selected summary sentences). Cosine similarity can be used to compute the similarity of two text segments. If each segment is represented as a vector, cosine similarity between two vectors (V_1, V_2) is measured using the following equation:

$$sim(V_1, V_2) = \frac{\sum_i t_{1i} t_{2i}}{\sqrt{\sum_i t_{1i}^2} \times \sqrt{\sum_i t_{2i}^2}} \quad (2)$$

where t_i is the term weight for a word w_i, for which we can use the TFIDF (term frequency, inverse document frequency) value, as widely used in the field of information retrieval.

3.2 Using Confusion Networks for Summarization

Confusion networks (CNs) have been used in many natural language processing tasks. Figure 1 shows a CN example for a sentence segment. It is a directed word graph from the starting node to the end node. Each edge represents a word with its associated posterior probability. There are several word candidates for each position. "-" in the CN represents a NULL hypothesis. Each path in the graph is a sentence hypothesis. For the example in Figure 1, *"I HAVE IT VERY FINE"* is the best hypothesis consisting of words with the highest probabilities for each position. Compared to N-best lists, confusion networks are a more compact and powerful representation for word candidates. We expect the rich information contained in the confusion networks (i.e., more word candidates and associated posterior probabilities) can help to determine words' importance for summarization.

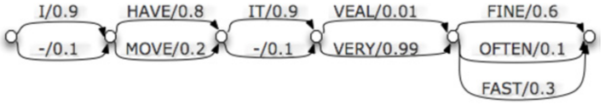

Figure 1: An example of confusion networks.

The core problems when using confusion networks under the MMR summarization framework are the definitions for S_i, D, and $Summ$, as shown in Equation 1. The extractive summary unit (for each S_i) we use is the segment provided by the recognizer. This is often different from syntactic or semantic meaningful unit (e.g., a sentence), but is a more realistic setup. Most of the previous studies for speech summarization used human labeled sentences as extraction units (for human transcripts, or map them to ASR output), which is not the real scenario when performing speech summarization on the ASR condition. In the future, we will use automatic sentence segmentation results, which we expect are better units than pause-based segmentation used in ASR. We still use a vector space model to represent each summarization unit S_i. The entire document (D) and the current selected summary ($Summ$) are formed by simply concatenating the corresponding segments S_i together. In the following, we describe different ways to represent the segments and how to present the final summary.

A. Segmentation representation

First, we construct the vector for each segment simply using all the word candidates in the CNs, without considering any confidence measure or posterior probability information. The same TFIDF computation is used as before, i.e., counting the number of times a word appears (TF) and how many documents it appears (used to calculate IDF).

Second, we leverage the confidence scores to build the vector. For the term frequency of word w_i, we calculate it by summing up its posterior probabilities $p(w_{ik})$ at each position k, that is,

$$TF(w_i) = \sum_k p(w_{ik}) \quad (3)$$

Similarly, the IDF values can also be computed using the confidence scores. The traditional method for calculating a word's IDF uses the ratio of the total number of documents (N) and the number of documents containing this word. Using the confidence scores, we calculate the IDF values as follows,

$$IDF(w_i) = log(\frac{N}{\sum_D (\max_k p(w_{ik}))}) \quad (4)$$

If a word w_i appears in the document, we find its maximum posterior probability among all the positions it occurs in the CNs, which is used to signal w_i's soft appearance in this document. We add these soft counts for all the documents as the denominator in Equation 4. Different from the traditional IDF calculation method, where the number of documents containing a word is an integer number, here the denominator can be any real number.

B. Confusion network pruning

The above vectors are constructed using the entire confusion networks. We may also use the pruned ones, in which the words with low posterior probabilities are removed beforehand. This can avoid the impact of noisy words, and increase the system speed as well. We investigate three different pruning methods, listed below.

- absolute pruning: In this method, we delete words if their posterior probabilities are lower than a predefined threshold, i.e., $p(w_i) < \theta$.
- max_diff pruning: First for each position k, we find the maximum probability among all the word candidates: $Pmax_k = \max_j p(w_{jk})$. Then we remove a word w_i in this position if the absolute difference of its probability with the maximum score is larger than a predefined threshold, i.e., $Pmax_k - p(w_{ik}) > \theta$.
- max_ratio pruning: This is similar to the above one, but instead of absolute difference, we use the ratio of their probabilities, i.e., $\frac{p(w_{ik})}{Pmax_k} < \theta$.

Again, for the last two pruning methods, the comparison is done for each position in the CNs.

C. Summary rendering

With a proper way of representing the text segments, we then extract the summary segments using the MMR method described in Section 3.1. Once the summary segments are selected using the confusion network input, another problem we need to address is how to present the final summary. When using the human transcripts or the 1-best ASR hypothesis for summarization, we can simply concatenate the corresponding transcripts of the selected sentence segments as the final summary for the users. However, when using the confusion networks as the representation of each sentence segment, we only know which segments are selected by the summarization system. To provide the final summary to the users, there are two choices. We can either use the best hypothesis from CNs of those selected segments as a text summary; or return the speech segments to the users to allow them to play it back. We will evaluate both methods in this paper. For the latter, in order to use similar word based performance measures, we will use the corresponding reference transcripts in order to focus on evaluation of the correctness of the selected summary segments.

4 Experiments

4.1 Corpus and Evaluation Measurement

We use the ICSI meeting corpus, which contains 75 recordings from natural meetings (most are research discussions) (Janin et al., 2003). Each meeting is about an hour long and has multiple speakers. These meetings have been transcribed, and annotated with extractive summaries (Murray et al., 2005b). The ASR output is obtained from a state-of-the-art SRI speech recognition system, including the confusion network for each sentence segment (Stolcke et al., 2006). The word error rate (WER) is about 38.2% on the entire corpus.

The same 6 meetings as in (Murray et al., 2005a; Xie and Liu, 2008; Gillick et al., 2009; Lin et al., 2009) are used as the test set in this study. Furthermore, 6 other meetings were randomly selected from the remaining 69 meetings in the corpus to form a development set. Each meeting in the development set has only one human-annotated summary; whereas for the test meetings, we use three summaries from different annotators as references for performance evaluation. The lengths of the reference summaries are not fixed and vary across annotators and meetings. The average word compression ratio for the test set is 14.3%, and the mean deviation is 2.9%. We generated summaries with the word compression ratio ranging from 13% to 18%, and only provide the best results in this paper.

To evaluate summarization performance, we use ROUGE (Lin, 2004), which has been widely used in previous studies of speech summarization (Zhang et al., 2007; Murray et al., 2005a; Zhu and Penn, 2006). ROUGE compares the system generated summary with reference summaries (there can be more than one reference summary), and measures different matches, such as N-gram, longest common sequence, and skip bigrams. In this paper, we present our results using both ROUGE-1 and

ROUGE-2 F-scores.

4.2 Characteristics of CNs

First we perform some analysis of the confusion networks using the development set data. We define two measurements:

- Word coverage. This is to verify that CNs contain more correct words than the 1-best hypotheses. It is defined as the percentage of the words in human transcripts (measured using word types) that appear in the CNs. We use word types in this measurement since we are using a vector space model and the multiple occurrence of a word only affects its term weights, not the dimension of the vector. Note that for this analysis, we do not perform alignment that is needed in word error rate measure — we do not care whether a word appears in the exact location; as long as a word appears in the segment, its effect on the vector space model is the same (since it is a bag-of-words model).
- Average node density. This is the average number of candidate words for each position in the confusion networks.

Figure 2 shows the analysis results for these two metrics, which are the average values on the development set. In this analysis we used absolute pruning method, and the results are presented for different pruning thresholds. For a comparison, we also include the results using the 1-best hypotheses (shown as the dotted line in the figure), which has an average node density of 1, and the word coverage of 71.55%. When the pruning threshold is 0, the results correspond to the original CNs without pruning.

We can see that the confusion networks include much more correct words than 1-best hypotheses (word coverage is 89.3% vs. 71.55%). When increasing the pruning thresholds, the word coverage decreases following roughly a linear pattern. When the pruning threshold is 0.45, the word coverage of the pruned CNs is 71.15%, lower than 1-best hypotheses. For node density, the non-pruned CNs have an average density of 11.04. With a very small pruning threshold of 0.01, the density decreases rapidly to 2.11. The density falls less than 2 when the threshold is 0.02, which means that for some nodes there is only one word candidate preserved after pruning (i.e., only one word has a posterior probability higher than 0.02). When the threshold increases to 0.4, the density is less than 1 (0.99), showing that on average there is less than one candidate left for each position. This is consistent with the word coverage results — when the pruning threshold is larger than 0.45, the confusion networks have less word coverage than 1-best hypotheses because even the top word hypotheses are deleted. Therefore, for our following experiments we only use the thresholds $\theta \leq 0.45$ for absolute pruning.

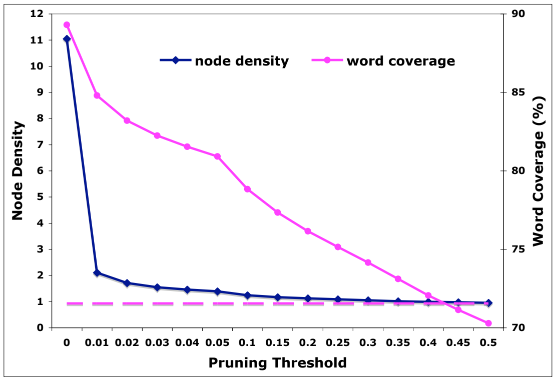

Figure 2: Average node density and word coverage of the confusion networks on the development set.

Note that the results in the figure are based on absolute pruning. We also performed analysis using the other two pruning methods described in Section 3.2. For those methods, because the decision is made by comparing each word's posterior probability with the maximum score for that position, we can guarantee that at least the best word candidate is included in the pruned CNs. We varied the pruning threshold from 0 to 0.95 for these pruning methods, and observed similar patterns as in absolute pruning for the word coverage and node density analysis. As expected, the fewer word candidates are pruned, the better word coverage and higher node density the pruned CNs have.

4.3 Summarization Results

4.3.1 Results on dev set using 1-best hypothesis and human transcripts

We generate the baseline summarization result using the best hypotheses from the confusion networks. The summary sentences are extracted using the MMR method introduced in Section 3.1. The term weighting is the traditional TFIDF value. The ROUGE-1 and ROUGE-2 scores for the baseline are listed in Table 1.

Because in this paper our task is to evaluate the summarization performance using ASR output, we generate an oracle result, where the summary extraction and IDF calculation are based on the human transcripts for each ASR segment. These results are also presented in Table 1. Comparing the results for the two testing conditions, ASR output and human transcripts, we can see the performance degradation due to recognition errors. The difference between them seems to be large enough to warrant investigation of using rich ASR output for improved summarization performance.

	ROUGE-1	ROUGE-2
Baseline: best hyp	65.60	26.83
Human transcript	69.98	33.21

Table 1: ROUGE results (%) using 1-best hypotheses and human transcripts on the development set.

4.3.2 Results on the dev set using CNs

A. Effect of segmentation representation

We evaluate the effect on summarization using different vector representations based on confusion networks. Table 2 shows the results on the development set using various input under the MMR framework. We also include the results using 1-best and human transcripts in the table as a comparison. The third row in the table uses the 1-best hypothesis, but the term weight for each word is calculated by considering its posterior probability in the CNs (denoted by "wp"). We calculate the TF and IDF values using Equation 3 and 4 introduced in Section 3.2. The other representations in the table are for the non-pruned and pruned CNs based on different pruning methods, and with or without using the posteriors to calculate term weights.

segment representation		ROUGE-1	ROUGE-2
Best hyp		65.60	26.83
Best hyp (wp)		66.83	29.84
Non-pruned CNs		66.58	28.22
Non-pruned CNs (wp)		66.47	29.27
Pruned CNs	Absolute	**67.44**	29.02
	Absolute (wp)	66.98	**29.99**
	Max_diff	67.29	28.97
	Max_diff (wp)	67.10	29.76
	Max_ratio	**67.43**	28.97
	Max_ratio (wp)	67.06	29.90
Human transcript		69.98	33.21

Table 2: ROUGE results (%) on the development set using different vector representations based on confusion networks: non-pruned and pruned, using posterior probabilities ("wp") and without using them.

In general, we find that using confusion networks improves the summarization performance comparing with the baseline. Since CNs contain more candidate words and posterior probabilities, a natural question to ask is, which factor contributes more to the improved performance? We can compare the results in Table 2 across different conditions that use the same candidate words, one with standard TFIDF, and the other with posteriors for TFIDF, or that use different candidate words and the same setup for TFIDF calculation. Our results show that there is more improvement using our proposed method for TFIDF calculation based on posterior probabilities, especially ROUGE-2 scores. Even when just using 1-best hypotheses, if we consider posteriors, we can obtain very competitive results. There is also a difference in the effect of using posterior probabilities. When using the top hypotheses representation, posteriors help both ROUGE-1 and ROUGE-2 scores; when using confusion networks, non-pruned or pruned, using posterior probabilities improves ROUGE-2 results, but not ROUGE-1.

Our results show that adding more candidates in the vector representation does not necessarily help summarization. Using the pruned CNs yields better performance than the non-pruned ones. There is not much difference among different pruning methods. Overall, the best results are achieved by using pruned CNs: best ROUGE-1 result without using posterior probabilities, and best ROUGE-2 scores when using posteriors.

B. Presenting summaries using human transcripts

segment representation		ROUGE-1	ROUGE-2
Best hyp		68.26	32.25
Best hyp (wp)		69.16	33.99
Non-pruned CNs		69.28	33.49
Non-pruned CNs (wp)		67.84	32.95
Pruned CNs	Absolute	69.66	34.06
	Absolute (wp)	69.37	34.25
	Max_diff	**69.88**	34.17
	Max_diff (wp)	69.38	33.94
	Max_ratio	69.76	34.06
	Max_ratio (wp)	69.44	**34.39**
Human transcript		69.98	33.21

Table 3: ROUGE results (%) on the development set using different segment representations, with the summaries constructed using the corresponding human transcripts for the selected segments.

In the above experiments, we construct the final summary using the best hypotheses from the confusion networks once the summary sentence segments are determined. Although we notice obvious improvement comparing with the baseline results, the ROUGE scores are still much lower than using the human transcripts. One reason for this is the speech recognition errors. Even if we select the correct utterance segment as in the reference summary segments, the system performance is still penalized when calculating the ROUGE scores. In order to avoid the impact of word errors and focus on evaluating whether we have selected the correct segments, next we use the corresponding human transcripts of the selected segments to obtain performance measures. The results from this experiment are shown in Table 3 for different segment representations.

We can see that the summaries formed using human transcripts are much better comparing with the results presented in Table 2. These two setups use the same utterance segments. The only difference lies in the construction of the final summary for performance measurement, using the top hypotheses or the corresponding human transcripts for the selected segments. We also notice that the difference between using 1-best hypothesis and human transcripts is greatly reduced using this new summary formulation. This suggests that the incorrect word hypotheses do not have a very negative impact in terms of selecting summary segments; however, word errors still account for a significant part of the performance degradation on ASR condition when using word-based metrics for evaluation. Using the best hypotheses with their posterior probabilities we can obtain similar ROUGE-1 score and a little higher ROUGE-2 score comparing to the results using human transcripts. The performance can be further improved using the pruned CNs.

Note that when using the non-pruned CNs and posterior probabilities for term weighting, the ROUGE scores are worse than most of other conditions. We performed some analysis and found that one reason for this is the selection of some poor segments. Most of the word candidates in the non-pruned CNs have very low confidence scores, resulting in high IDF values using our proposed methods. Since some top hypotheses are NULL words in the poorly selected summary segments, it did not affect the results when using the best hypothesis for evaluation, but when using human transcripts, it leads to lower precision and worse overall F-scores. This is not a problem for the pruned CNs since words with low probabilities have been pruned beforehand, and thus do not impact segment selection. We will investigate better methods for term weighting to address this issue in our future work.

These experimental results prove that using the confusion networks and confidence scores can help select the correct sentence segments. Even though the 1-best WER is quite high, if we can consider more word candidates and/or their confidence scores, this will not impact the process of selecting summary segments. We can achieve similar performance as using human transcripts, and sometimes even slightly better performance. This suggests using more word candidates and their confidence scores results in better term weighting and representation in the vector space model. Some previous work showed that using word confidence scores can help minimize the WER of the extracted summaries, which then lead to better summarization performance. However, we think the main reason for the improvement in our study is from selecting better utterances, as shown in Table 3. In our experiments, because different setups select different segments as the summary, we can not directly compare the WER of extracted summaries, and analyze whether lower WER is also helpful for better sum-

	output summary			
	best hypotheses		human transcripts	
	R-1	R-2	R-1	R-2
Best hyp	65.73	26.79	68.60	32.03
Best hyp (wp)	65.92	27.27	68.91	32.69
Pruned CNs	66.47	27.73	69.53	34.05
Human transcript	N/A	N/A	69.08	33.33

Table 4: ROUGE results (%) on the test set.

marization performance. In our future work, we will perform more analysis along this direction.

4.3.3 Experimental results on test set

The summarization results on the test set are presented in Table 4. We show four different evaluation conditions: baseline using the top hypotheses, best hypotheses with posterior probabilities, pruned CNs, and using human transcripts. For each condition, the final summary is evaluated using the best hypotheses or the corresponding human transcripts of the selected segments. The summarization system setups (the pruning method and threshold, λ value in MMR function, and word compression ratio) used for the test set are decided based on the results on the development set.

For the results on the test set, we observe similar trends as on the development set. Using the confidence scores and confusion networks can improve the summarization performance comparing with the baseline. The performance improvements from “Best hyp” to “Best hyp (wp)” and from “Best hyp (wp)” to “Pruned CNs” using both ROUGE-1 and ROUGE-2 measures are statistically significant according to the paired t-test ($p < 0.05$). When the final summary is presented using the human transcripts of the selected segments, we observe slightly better results using pruned CNs than using human transcripts as input for summarization, although the difference is not statistically significant. This shows that using confusion networks can compensate for the impact from recognition errors and still allow us to select correct summary segments.

5 Conclusion and Future Work

Previous research has shown performance degradation when using ASR output for meeting summarization because of word errors. To address this problem, in this paper we proposed to use confusion networks for speech summarization. Under the MMR framework, we introduced a vector representation for the segments by using more word candidates in CNs and their associated posterior probabilities. We evaluated the effectiveness of using different confusion networks, the non-pruned ones, and the ones pruned using three different methods, i.e., absolute, max_diff and max_ratio pruning. Our experimental results on the ICSI meeting corpus showed that even when we only use the top hypotheses from the CNs, considering the word posterior probabilities can improve the summarization performance on both ROUGE-1 and ROUGE-2 scores. By using the pruned CNs we can obtain further improvement. We found that more gain in ROUGE-2 results was yielded by our proposed soft term weighting method based on posterior probabilities. Our experiments also demonstrated that it is possible to use confusion networks to achieve similar or even better performance than using human transcripts if the goal is to select the right segments. This is important since one possible rendering of summarization results is to return the audio segments to the users, which does not suffer from recognition errors.

In our experiments, we observed less improvement from considering more word candidates than using the confidence scores. One possible reason is that the confusion networks we used are too confident. For example, on average 90.45% of the candidate words have a posterior probability lower than 0.01. Therefore, even though the correct words were included in the confusion networks, their contribution may not be significant enough because of low term weights. In addition, low probabilities also cause problems to our proposed soft IDF computation. In our future work, we will investigate probability normalization methods and other techniques for term weighting to cope with these problems.

6 Acknowledgment

This research is supported by NSF award IIS-0845484. Any opinions expressed in this work are those of the authors and do not necessarily reflect the views of NSF. The authors thank Shih-Hsiang Lin and Fei Liu for useful discussions.

References

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR*.

Ciprian Chelba, Jorge Silva, and Alex Acero. 2007. Soft indexing of speech content for search in spoken documents. In *Computer Speech and Language*, volume 21, pages 458–478.

Tee Kiah Chia, Khe Chai Sim, Haizhou Li, and Hwee Tou Ng. 2008. A lattice-based approach to query-by-example spoken document retrieval. In *Proceedings of SIGIR*.

Gunes Erkan and Dragomir R. Radev. 2004. LexRank: graph-based lexical centrality as salience in text summarization. *Artificial Intelligence Research*, 22:457–479.

Nikhil Garg, Benoit Favre, Korbinian Reidhammer, and Dilek Hakkani-Tur. 2009. ClusterRank: a graph based method for meeting summarization. In *Proceedings of Interspeech*.

Dan Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur. 2009. A global optimization framework for meeting summarization. In *Proceedings of ICASSP*.

Dilek Hakkani-Tur, Frederic Behet, Giuseppe Riccardi, and Gokhan Tur. 2006. Beyond ASR 1-best: using word confusion networks in spoken language understanding. *Computer Speech and Language*, 20(4):495 – 514.

Chiori Hori and Sadaoki Furui. 2003. A new approach to automatic speech summarization. *IEEE Transactions on Multimedia*, 5(3):368–378.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. In *Proceedings of ICASSP*.

Shih-Hsiang Lin and Berlin Chen. 2009. Improved speech summarization with multiple-hypothesis representations and Kullback-Leibler divergence measures. In *Proceedings of Interspeech*.

Hui Lin, Jeff Bilmes, and Shasha Xie. 2009. Graph-based submodular selection for extractive summarization. In *Proceedings of ASRU*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *the Workshop on Text Summarization Branches Out*.

Yang Liu, Shasha Xie, and Fei Liu. 2010. Using n-best recognition output for extractive summarization and keyword extraction in meeting speech. In *Proceedings of ICASSP*.

Lidia Mangu, Eric Brill, and Andreas Stolcke. 2000. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech and Language*, 14:373–400.

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proceedings of EACL*.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005a. Extractive summarization of meeting recordings. In *Proceedings of Interspeech*.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2005b. Evaluating automatic summaries of meeting recordings. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*.

Khe Chai Sim, William Byrne, Mark Gales, Hichem Sahbi, and Phil Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In *Proceedings of ICASSP*.

Andreas Stolcke, Yochai Konig, and Mitchel Weintraub. 1997. Explicit word error minimization in N-best list rescoring. In *Proceedings of Eurospeech*.

Andreas Stolcke, Barry Chen, Horacio Franco, Venkata Ra mana Rao Gadde, Martin Graciarena, Mei-Yuh Hwang, Katrin Kirchhoff, Arindam Mandal, Nelson Morgan, Xin Lei, Tim Ng, and et al. 2006. Recent innovations in speech-to-text transcription at SRI-ICSI-UW. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1729–1744.

Robin Valenza, Tony Robinson, Marianne Hickey, and Roger Tucker. 1999. Summarization of spoken audio through information extraction. In *Proceedings of the ESCA Workshop on Accessing Information in Spoken Audio*, pages 111–116.

Shasha Xie and Yang Liu. 2008. Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization. In *Proceedings of ICASSP*.

Shasha Xie, Benoit Favre, Dilek Hakkani-Tur, and Yang Liu. 2009. Leveraging sentence weights in concept-based optimization framework for extractive meeting summarization. In *Proceedings of Interspeech*.

Klaus Zechner and Alex Waibel. 2000. Minimizing word error rate in textual summaries of spoken language. In *Proceedings of NAACL*.

Jian Zhang, Ho Yin Chan, Pascale Fung, and Lu Cao. 2007. A comparative study on speech summarization of broadcast news and lecture speech. In *Proceedings of Interspeech*.

Xiaodan Zhu and Gerald Penn. 2006. Summarization of spontaneous conversations. In *Proceedings of Interspeech*.