INTERSPEECH 2009 BRIGHTON

Leveraging Sentence Weights in a Concept-based Optimization Framework for Extractive Meeting Summarization

Shasha Xie[1,2], *Benoit Favre*[2], *Dilek Hakkani-Tür*[2], *Yang Liu*[1]

[1]Department of Computer Science, The University of Texas at Dallas, Richardson, TX
[2]International Computer Science Institute, Berkeley, CA

{shasha,favre,dilek}@icsi.berkeley.edu, yangl@hlt.utdallas.edu

10.21437/Interspeech.2009-457

Abstract

We adopt an unsupervised concept-based global optimization framework for extractive meeting summarization, where a subset of sentences is selected to cover as many important concepts as possible. We propose to leverage sentence importance weights in this model. Three ways are introduced to combine the sentence weights within the concept-based optimization framework: selecting sentences for concept extraction, pruning unlikely candidate summary sentences, and using joint optimization of sentence and concept weights. Our experimental results on the ICSI meeting corpus show that our proposed methods can significantly improve the performance for both human transcripts and ASR output compared to the concept-based baseline approach, and this unsupervised approach achieves results comparable with those from supervised learning approaches presented in previous work.

Index Terms: global optimization, sentence weights, meeting summarization

1. INTRODUCTION

Extractive summarization selects salient sentences from the original documents (or recordings) and presents them as a summary. Meeting summarization has received increasing interest recently and many techniques have been proposed for extractive meeting summarization. Among them, an unsupervised learning approach, Maximum Marginal Relevance (MMR), achieved comparable performance to other methods for this task [1, 2]. MMR is a greedy algorithm, where at each step one sentence is selected for inclusion in the summary according to its weight. This algorithm can select the most relevant sentences, and at the same time avoid the redundancy by removing the sentences too similar to already selected ones. However, MMR is local optimal because the decision is made based on the sentences' scores in the current iteration.

In [3], the author studied modeling the multi-document summarization problem using a global inference algorithm with some definition of relevance and redundancy. The Integer Linear Programming (ILP) solver was used to efficiently search a large space of possible summaries for an optimal solution. In [4], the authors adopted this 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. They generated the summary by selecting the best set of sentences which can cover as many concepts as possible, and the redundancy was prevented indirectly by satisfying a length constraint. The authors showed that this global optimization approach outperformed MMR.

However, one problem with this concept-based global optimization approach is that it 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 [5]. According to manual examination, this results in the degradation of the linguistic quality of the summary. In this paper we propose to incorporate and leverage sentence importance weights in this concept-based optimization method. We investigate different ways to use sentence weights: use them to extract more indicative concepts, combine them with concept weights in the optimization function, and use them to prune sentence candidates. Our experimental results on the ICSI meeting corpus show consistent improvement over the original concept-based approach.

2. DATA

We use the ICSI meeting corpus [6], which contains 75 recordings from natural meetings. Each meeting is about an hour long and has multiple speakers. These meetings have been manually transcribed and annotated with dialog acts (DA) [7], topic segments, and extractive summaries [1]. The automatic speech recognition (ASR) output for this corpus is obtained from a state-of-the-art SRI conversational telephone speech system [8], with a word error rate of about 38.2% on the entire corpus. We align the human transcripts and ASR output, then map the human annotated DA boundaries and topic boundaries to the ASR words, such that we have human annotation for the ASR output.

The same 6 meetings as in [9] are used as the test set. We arbitrarily selected 6 other meetings from the corpus as the development set to evaluate the proposed methods and optimize parameters. We use three reference summaries from different annotators for each meeting in the test set. For the development set, we only have one reference summary for each meeting. 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%.

3. GLOBAL OPTIMIZATION FRAMEWORK FOR SUMMARIZATION

3.1. Concept-based Summarization

In [4], *concepts* were used as minimal independent pieces of information. Extractive summarization selects sentences to cover as many important concepts as possible, and at the same time satisfying a predefined length constraint. This idea can be modeled as seeking a summary that maximizes a global objective

function:

$$maximize \sum_i w_i c_i \quad (1)$$

$$subject\ to \sum_j l_j s_j < L \quad (2)$$

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 in [4] to select sentences that maximize the objective function under the length constraint.

We use simple n-grams as concepts following the setup in [4], and extract concepts using the following procedure:

- 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 words has an idf value 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 "manager" if its frequency is the same as "dialogue manager".
- Weight each n-gram k_i as
  $w_i = frequency(k_i) * n * \max_j idf(word_j)$
  where n is the n-gram length, and $word_j$ goes through all the words in the n-gram.

In the above computation, the IDF (inverse document frequency) values are calculated using transcripts of 75 meetings. For both human transcripts and ASR outputs, we split each of 75 meetings into multiple topics based on the manual topic segmentation, and then use these new "documents" to calculate the IDF values. Unlike [4], 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.

3.2. Using Sentence Importance Weight

We propose to incorporate sentence importance weights in the above summarization framework. Since the global optimization model is unsupervised, in this paper we choose to use sentence weights that can also be obtained in an unsupervised fashion. We use the cosine similarity between each sentence and the entire document, which can be calculated using the following equation:

$$sim(D_1, D_2) = \frac{\sum_i t_{1i} t_{2i}}{\sqrt{\sum_i t_{1i}^2} \times \sqrt{\sum_i t_{2i}^2}} \quad (3)$$

where t_i is the TF-IDF weight for a word. We investigate different ways to leverage these sentence scores in the concept-based optimization framework.

Filtering sentences for concept generation
First we use sentence weights to select important sentences, and then extract concepts from the selected sentences only. The concepts are obtained in the same way as described in Section 3.1. 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.

Pruning sentences from the selection
In this method, we use sentence weights 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. This approach is similar to the 'resampling' method in [10], where sentences with high salience scores are preserved as candidates in a supervised learning framework.

Joint optimization using sentence and concept weights
Lastly, we extend the optimization function (Equation 1) to consider sentence importance weights, i.e.,

$$maximize\ (1-\lambda) * \sum_i w_i c_i + \lambda * \sum_j u_j s_j \quad (4)$$

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 1. The summary length constraint is the same as Equation 2. 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

To evaluate summarization performance, we use ROUGE [11], which has been used in previous studies of speech summarization [9, 12, 13]. 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. To be consistent with previous work, we provide ROUGE-1 (unigram) F-measure results for all experiments. 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.

4.1. Evaluating the Approaches on the Development Set

4.1.1. Baseline results

We provide several baseline results in Table 1 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 meeting summarization task [14]. 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 introduced in Section 3.1. These scores are comparable with those presented in [4]. 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.

4.1.2. 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

compression		14%	15%	16%	17%	18%
REF	long sentence	54.50	56.16	57.47	58.58	59.23
	MMR	66.81	67.06	66.90	66.64	66.09
	concept-based	**67.20**	**67.98**	**68.30**	**67.82**	**67.51**
ASR	long sentence	63.11	64.01	64.72	64.65	64.89
	MMR	63.60	64.32	64.80	65.03	65.14
	concept-based	**63.99**	**65.04**	**65.45**	**65.44**	**65.30**

Table 1: *ROUGE-1 F results (%) of three baselines on the dev set for both human transcripts (REF) and ASR output: selecting the longest sentences, MMR, and concept-based optimization approach.*

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.

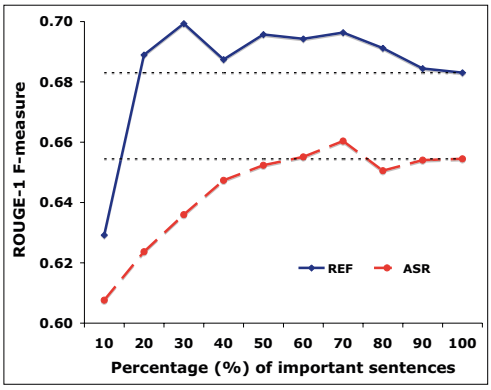

Figure 1: *ROUGE-1 F results using different percentage of important sentences during concept extraction on the dev set for both human transcripts (REF) and ASR output. The horizontal dashed lines represent the scores of the baselines using all the sentences.*

4.1.3. Pruning sentences from the selection

This experiment evaluates the impact of using sentence weights to prune sentences and pre-select 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. This is consistent with the result in [10]. We also evaluate a length-based sentence selection and find that it is inferior to sentence score based pruning.

4.1.4. Joint optimization using sentence and concept weights

Finally we evaluate the impact of incorporating sentence scores in the global optimization framework using Equation 4. 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

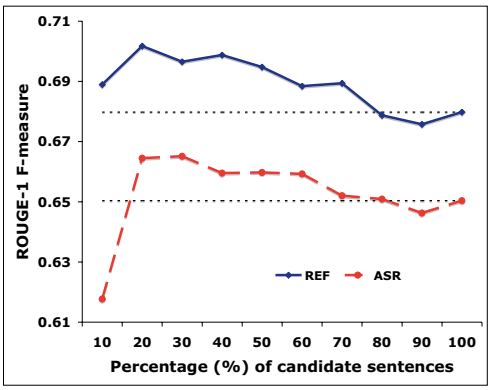

Figure 2: *ROUGE-1 F results using pruning to preserve different percentage of candidate summary sentences on the dev set for both human transcripts (REF) and ASR output. The horizontal dashed lines represent the scores of the baselines using all the sentences.*

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 2 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.

	baseline	raw cosine	#concept norm	**#words norm**
REF	68.30	68.42	68.42	**68.50**
ASR	65.45	61.12	65.08	**66.29**

Table 2: *ROUGE-1 F results (%) on the dev set for both human transcripts (REF) and ASR output. We compare: concept-weights alone, the combination of concept and sentence weights (at the best λ) for raw scores and scores rescaled by the number of concepts or the number of words in the sentence.*

Table 3 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 most — using this approach alone can

achieve very similar results to the best scores.

	baseline	sentence weights for			**all**
		concept	pruning	joint	
REF	68.30	69.96	70.17	68.50	**70.37**
ASR	65.45	66.04	66.51	66.29	**66.77**

Table 3: *ROUGE-1 F results (%) of incorporating sentence importance weights on the dev set using both human transcripts (REF) and ASR output. We compare the results of the baseline, using three proposed approaches by themselves, and their combination.*

4.2. Results on the Test Set

The experimental results on the test set using all the approaches proposed in this paper are shown in Table 4. The parameter values are selected according to the performance on the dev set. The baseline results are calculated using the concept-based summarization model, obtaining comparable results to the ones presented in [4]. ROUGE-1 scores are improved using our proposed three approaches for leveraging 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. Similar patterns also hold when evaluating using ROUGE-2 (bigram) and ROUGE-SU4 (skip-bigram) scores. We also verified that the results are significantly better than the baseline according to a paired t-test ($p < 0.05$).

Comparing with the results of using supervised learning approaches for meeting summarization presented in [10], (e.g., 70.38 for human transcripts, 65.98 for ASR output on the dev set with a 16% word compression ratio), we achieve comparable scores using our unsupervised framework. This is very important because the annotation of summary sentences is very expensive and time-consuming, especially for the meeting domain. Although our proposed method requires parameter tuning (based on some dev set), it is unsupervised and we do not need a large corpus annotated with summaries for training.

compression		14%	15%	16%	17%	18%
REF	baseline	67.08	67.84	68.35	68.82	69.00
	concept	68.75	69.80	70.07	70.24	69.77
	pruning	68.85	69.30	70.10	70.33	70.43
	joint	67.48	68.40	68.97	69.19	69.16
	all	**69.35**	**70.29**	**70.87**	**70.72**	**70.30**
ASR	baseline	63.30	64.51	65.31	65.27	65.84
	concept	64.00	65.44	66.15	66.52	66.39
	pruning	65.83	66.78	66.63	66.79	66.48
	joint	63.82	64.76	65.80	66.11	65.77
	all	**65.87**	**66.67**	**67.07**	**67.20**	**66.91**

Table 4: *ROUGE-1 F results (%) for different word compression ratios on test set for both human transcripts (REF) and ASR output, including the baseline using concept-based optimization framework, three proposed approaches of incorporating sentence weights, and their combination.*

5. CONCLUSION AND FUTURE WORK

In this paper, we have evaluated different approaches of leveraging the sentence importance weights in an unsupervised concept-based optimization framework. 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.

In our future work, we will investigate other sentence weighting methods, such as DICE coefficient or scores from supervised learning approaches. In this paper, we propose to use a linear combination for joint optimization of concept and sentence weights as shown in Equation 4, where the improvement is not very significant according to our experimental results. In our future work, we will evaluate different ways of extending the optimization function, or combine with forge sentences.

6. Acknowledgment

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811), and NSF grant IIS-0845484.

7. References

[1] Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore, "Evaluating automatic summaries of meeting recordings," in *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, 2005.

[2] Shasha Xie and Yang Liu, "Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization," in *Proc. of ICASSP*, 2008.

[3] Ryan McDonald, "A study of global inference algorithms in multi-document summarization," in *Proc. of 9th ICSLP*, 2006.

[4] Dan Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur, "A global optimization framework for meeting summarization," in *Proc. of ICASSP*, 2009.

[5] Dan Gillick, Benoit Favre, and Dilek Hakkani-Tur, "The ICSI Summarization System at TAC 2008," in *Proc. of the Text Analysis Conference workshop*, 2008.

[6] Adam Janin, Don Baron, and Jane Edwards et al., "The ICSI meeting corpus," in *Proc. of ICASSP*, 2003.

[7] Elizabeth Shriberg, Raj Dhillon, and Sonali Bhagat et al., "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. of 5th SIGDIAL Workshop*, 2004.

[8] Qifeng Zhu, Andreas Stolcke, Barry Chen, and Nelson Morgan, "Using MLP features in SRI's conversational speech recognition system," in *Proc. of Interspeech*, 2005.

[9] Gabriel Murray, Steve Renals, and Jean Carletta, "Extractive summarization of meeting recordings," in *Proc. of Interspeech*, 2005.

[10] Shasha Xie, Yang Liu, and Hui Lin, "Evaluating the effectiveness of features and sampling in extractive meeting summarization," in *Proc. of IEEE Spoken Language Technology (SLT)*, 2008.

[11] Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," in *the Workshop on Text Summarization Branches Out*, 2004.

[12] Justin Jian Zhang, Ho Yin Chan, and Pascale Fung, "Improving lecture speech summarization using rhetorical information," in *Proc. of ASRU*, 2007.

[13] Xiaodan Zhu and Gerald Penn, "Summarization of spontaneous conversations," in *Proc. of Interspeech*, 2006.

[14] Gerald Penn and Xiaodan Zhu, "A critical reassessment of evaluation baselines for speech summarization," in *Proc. of ACL*, 2008.