USING CORPUS AND KNOWLEDGE-BASED SIMILARITY MEASURE IN MAXIMUM MARGINAL RELEVANCE FOR MEETING SUMMARIZATION

Shasha Xie, Yang Liu

The University of Texas at Dallas, Richardson, TX, USA
{shasha, yangl}@hlt.utdallas.edu

ABSTRACT

MMR (Maximum Marginal Relevance) is widely used in summarization for its simplicity and efficacy, and has been demonstrated to achieve comparable performance to other approaches for meeting summarization. How to appropriately represent the similarity of two text segments is crucial in MMR. In this paper, we evaluate different similarity measures in the MMR framework for meeting summarization on the ICSI meeting corpus. We introduce a corpus-based measure to capture the similarity at the semantic level, and compare this method with cosine similarity and centroid score that only considers the salient words in the segments. Our experimental results evaluated by the ROUGE summarization metrics show that both the centroid score and the corpus-based similarity measure yield better performance than the commonly used cosine similarity. In addition, adding part-of-speech information in the corpus-based approach helps for the human transcripts condition, but not when using ASR output.

Index Terms— meeting summarization, MMR, centroid score, corpus-based similarity

1. INTRODUCTION

Recently there has been an increasing interest in automatically processing the large amount of meeting speech, including recognition, browsing, and summarization. Extractive meeting summarization selects salient parts from the original recordings and presents them together as a summary. This will facilitate users to search and browse the meeting recordings. Many techniques have been proposed for meeting summarization. Some rely on textual information, such as Maximum Marginal Relevance (MMR) and Latent Semantic Analysis (LSA) [1]; others incorporate acoustic/prosodic cues in the statistical learning approaches, for example, Hidden Markov Model (HMM), Maximum Entropy, Conditional Random Fields (CRF), and Support Vector Machines (SVM) [2, 3, 4, 5]. Among these, MMR is one of the simplest techniques for summarization, and has been effectively used for text summarization [6]. In [1], Murray et al. compared three approaches (MMR, LSA, and feature-based methods) and showed that MMR achieved comparable performance to other methods for meeting summarization.

In MMR, a function is needed to measure the similarity between two text segments. Cosine similarity has been widely used for the similarity measurement between two documents, each of which is typically represented using a vector of term weights. However, the simple lexical matching in cosine similarity may not appropriately represent the distance between two documents. To address this problem, we will evaluate other similarity measures in the MMR framework for meeting summarization in this paper. First, a centroid score is used to measure the similarity of a sentence to the entire document by only counting the salient words. The second one is a corpus-based measure, which has been proposed to capture the semantic similarity of texts, and has been shown to outperform the vector-based approach in text processing [7]. Our experiments on the ICSI meeting data have shown that these approaches achieve significantly better summarization performance than using the cosine similarity in the MMR framework, both on the manual transcripts and the ASR output.

The rest of this paper is organized as follows. In Section 2, we introduce the MMR summarization approaches, and different methods for similarity measures we use in MMR for meeting summarization. The experimental results are shown in Section 3. Conclusion and future work are given in Section 4.

2. SUMMARIZATION APPROACHES

2.1. Maximum Marginal Relevance (MMR)

MMR [6] has been widely used in text summarization because of its simplicity and efficacy. It selects the most relevant sentences at the same time avoiding redundancy. In extractive summarization, the final score of a given sentence S_i in MMR is calculated as follows:

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

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. The two similarity functions (Sim_1 and Sim_2) represent the similarity of a sentence to the entire document and to the selected summary, respectively. The sentences with the highest MMR scores will be iteratively chosen into the summary until the summary reaches a predefined proper size.

For meeting summarization, Murray et al. [1] showed that MMR is comparable with other summarization methods. However, to our knowledge, no prior studies have examined the similarity measures in MMR for speech summarization, which is our focus in this paper.

2.2. Cosine Similarity

One most commonly used similarity measure is cosine similarity, 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

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on May 07,2010 at 21:42:26 UTC from IEEE Xplore. Restrictions apply.

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. One widely used method for the IDF value for a word w_i is

$$IDF(w_i) = log(N/N_i) \tag{3}$$

where N_i is the number of documents containing w_i in a collection of N documents.

In [8], Murray and Renals 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. Therefore in this study, we will use TF-IDF for term weighting and focus on the problem of how to calculate the similarity between two documents in the MMR framework.

In our experiments, we also found that different normalization methods for the cosine similarity have a great effect on the system performance. The method we adopt in this paper is to first calculate the dot product score (i.e., without the denominator in Eq 2) for Sim_1, then scaling it to [0,1] based on the maximum scores among all the sentences. We use the original cosine score for Sim_2.

2.3. Centroid Score

Another distance measure we evaluate is the centroid score [9], which only considers the salient words for the distance between a sentence and the entire document. The same vector representation is used as in cosine similarity. In this approach, each word in a sentence S_i is checked to see if it occurs in the text segment T and if the term weight (TF-IDF value) of this word is greater than a predefined threshold. If these requirements are met, the term weight of this word is added to the centroid score for the sentence.

$$\begin{aligned} Score_{centroid}(i) = \sum_{w_j \in S_i} & bool(w_j \in T) * \\ & bool(tw(w_j) > v) * tw(w_j) \end{aligned} \tag{4}$$

where $tw(w_j)$ represents the term weight for the word w_j, and the functions $bool(w_j \in T)$ and $bool(tw(w_j) > v)$ check the two conditions mentioned above.

In the MMR system, we use the centroid score as the first similarity function (Sim_1 in Eq 1). The second similarity measure Sim_2 is still the cosine distance.

2.4. Corpus-based Semantic Similarity

The cosine and centroid scores between a sentence and a document are all based on simple lexical matching, that is, only the words that occur in both contribute to the similarity. Such literal comparison can not always capture the semantic similarity of text. Therefore we use the following function to compute the similarity score between two text segments [7].

$$\begin{aligned} sim(T_1, T_2) = \frac{1}{2}(& \frac{\sum_{w \in \{T_1\}} (maxSim(w, T_2) * idf(w))}{\sum_{w \in \{T_1\}} idf(w)} \\ & + \frac{\sum_{w \in \{T_2\}} (maxSim(w, T_1) * idf(w))}{\sum_{w \in \{T_2\}} idf(w)}) \end{aligned} \tag{5}$$

$$maxSim(w, T_i) = \max_{w_i \in \{T_i\}} \{sim(w, w_i)\} \tag{6}$$

For each word w in segment T_1, we find a word in segment T_2 that has the highest semantic similarity to w ($maxSim(w, T_2)$). Similarly, for the words in T_2, we identify the corresponding words in segment T_1. The similarity score of the two text segments is then calculated by combining the similarity of the words in each segment, weighted by their word specificity (i.e., IDF values).

To calculate the semantic similarity between two words w_1 and w_2, we use a corpus-based approach and measure the pointwise mutual information (PMI) [7, 10]:

$$PMI(w_1, w_2) = log_2 \frac{c(w_1 \; near \; w_2)}{c(w_1) * c(w_2)} \tag{7}$$

This indicates the statistical dependency between w_1 and w_2, and can be used as a measure of the semantic similarity of two words. $c(w_1 \; near \; w_2)$ represents the number of times that word w_1 appears near word w_2. For this co-occurrence count, a window of length l is used, that is, we only count when w_1 and w_2 co-occur within this window. For a word, we define $PMI(w, w) = 1$, therefore, $maxSim(w, T)$ is 1 if w appears in T.

The part-of-speech (POS) information of each word can also be taken into consideration when calculating the similarity of two text segments [7]. Eq 6 can be modified as

$$maxSim(w, T_i) = \max_{\substack{w_i \in \{T_i\} \\ pos(w_i) = pos(w)}} \{sim(w, w_i)\} \tag{8}$$

This means when finding the $maxSim$ between a word w and a text segment T_i, we will only consider the words in T_i with the same POS as word w. The reason behind this is that it is more meaningful to calculate the similarity of two words with the same POS. For example, it is hard to tell the relationship between word *but* and *dog*.

Note that two different words in the two segments also contribute to the similarity score using this corpus-based approach, unlike in the cosine similarity. We call this approach corpus-based similarity following [7], even though in the cosine and centroid scores, the IDF values are also generated based on a corpus. For the MMR score, we use the corpus-based similarity for the two similarity functions (Sim_1, Sim_2) in Eq 1, since it is more comparable than using a corpus-based similarity for Sim_1 and a cosine similarity for Sim_2.

2.5. Approximation in MMR Computation

In the MMR approach, for each sentence in a test document, its similarity score to the whole document (Sim_1 in Eq 1) can be calculated off-line. However, when extracting the summary sentences using Eq 1, computation is needed on the fly since the second similarity function is with respect to the currently selected summary, which changes in every iteration. The speed of the system is especially a problem for the corpus-based similarity. It is more complex and time-consuming than cosine similarity since we need to compare every word pair in the two text segments. To speed up the process, we adopt an approximated method [9]. 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 in Eq 1). This is approximated as Sim_2 in Eq 1, and can be computed off-line. Therefore, the summary selection process only needs to find the top sentences that have high combined scores.

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on May 07,2010 at 21:42:26 UTC from IEEE Xplore. Restrictions apply.

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. These approximations significantly speed up the extraction process.

3. EXPERIMENTS

3.1. Data and Experimental Setup

We use the ICSI meeting corpus [11], which contains 75 recordings from natural meetings. Each meeting is about an hour long. These meetings have been transcribed and annotated with topic information and extractive summaries [12]. The ASR output is obtained from a state-of-the-art SRI conversational telephone speech (CTS) system [13], which was trained using no acoustic data or transcripts from the meeting corpus. The word error rate on the entire corpus is about 38.2%. Annotated dialog acts (DA) in the corpus [14] are used as the sentence units for extractive summarization in the human transcripts case. For the ASR condition, sentences are obtained by aligning human annotated DA boundaries to the ASR words.

We use the same 6 meetings as in [1] to form the test set, and the other 69 meetings as the training set. Furthermore, we randomly select 6 meetings from the training set as the development set, then the rest is used to compose the training corpus for mutual information measure (i.e., Eq 7). The development set is used to optimize the λ value in Eq 1. Each of the 6 test meetings has 3 human annotated summaries, which we use as references.

For the term weights in the vector representation, IDF values are obtained from the 69 training meetings. For the human transcripts condition where the annotated topic information is available, we split each of the 69 training meetings into multiple topics, and then use these new "documents" to calculate the IDF values. This generates more robust estimation for IDF, compared with simply using the original 69 meetings as the documents (a concept similar to language model smoothing). The PMI information is generated using the training meetings for the human transcripts and ASR outputs respectively.

We tagged all the meetings using the TnT POS tagger [15]. The POS model is retrained using the Penn Treebank-3 Switchboard data, which is expected to be more similar to the meeting style than domains such as Wall Street Journal.

3.2. Evaluation Measurement

We use ROUGE [16] to evaluate summarization performance. ROUGE compares the system generated summary with the reference summaries, and measures different matches such as N-gram, longest common sequence, skip bigrams. It can accept multiple reference summaries. ROUGE has been used in previous studies of meeting summarization [1, 4, 17], therefore we believe it is a reasonable method for performance measure in our study.

3.3. Experimental Results

We evaluate the different approaches for similarity measure under the MMR framework for meeting summarization. The top 4.2% sentences are selected into the summary using the reference transcripts according to the combined MMR score. Table 1 shows the summarization results (ROUGE unigram match R-1) on the dev set using human transcripts. The columns Sim_1 and Sim_2 are the similarity measures we used for the two similarity functions in Eq 1, which represent the similarity of a sentence S_i to the whole document, and the similarity of the sentence S_i to the currently selected summary, respectively. $approx\_1$ and $approx\_2$ represent whether the two approximations of MMR introduced in Section 2.5 are adopted: $approx\_1$ approximates Sim_2 in Eq 1 using the similarity of the sentence with those that have a higher similarity score to the entire document; $approx\_2$ considers a small percent of sentences that have a high similarity score to the entire document, with the percentage of the candidates shown in the table (where perc is the compression rate of the summary).

Sim_1	Sim_2	approx_1	approx_2	R-1
cosine	cosine	no	no	0.60465
cosine	cosine	yes	2*perc	0.65255
centroid	cosine	no	no	0.68011
centroid	cosine	yes	no	0.68104
centroid	cosine	yes	2*perc	0.68274
corpus	corpus	yes	2*perc	0.68910
corpus	corpus	yes	3*perc	0.68443
corpus_pos	corpus_pos	yes	2*perc	0.69316

Table 1. Summarization results (ROUGE R-1 F-measure) using different similarity approaches on dev data using human transcripts.

For the cosine scores and centroid scores, applying the two approximation in MMR does not hurt the system performance, instead it yields slight improvement. Among the different similarity measures, both the centroid and the corpus-based similarity measures outperform the cosine similarity. Adding POS constraint for word similarity is also helpful, achieving the best performance among all the approaches. We also considered allowing more candidate sentences, for example, using 3 times of the target percent, however, there is a slight degradation (the result is shown in Table 1 for the corpus-based approach).

The results on the test set using human transcripts are shown in Table 2. Consistent with the dev set, we observe that the similarity measures we introduced improve the system performance. When POS information is considered in the corpus-based similarity measure, there is a further improvement.

Sim_1	Sim_2	approx_1	approx_2	R-1
cosine	cosine	no	no	0.58843
cosine	cosine	yes	2*perc	0.65300
centroid	cosine	no	no	0.68938
centroid	cosine	yes	no	0.68688
centroid	cosine	yes	2*perc	0.69103
corpus	corpus	yes	2*perc	0.69274
corpus_pos	corpus_pos	yes	2*perc	0.71243

Table 2. Summarization results (ROUGE R-1 F-measure) using different similarity approaches on test data using human transcripts.

Table 3 shows the results for a few selected approaches using ASR output on the test set. We notice that there is a performance degradation compared to using reference transcripts, but the new proposed similarity measure still outperforms the baseline. In the corpus-based method, considering POS information does not improve the system performance, different from what have observed on the human transcript condition. This is probably because the POS

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on May 07,2010 at 21:42:26 UTC from IEEE Xplore. Restrictions apply.

tagging accuracy for the ASR transcripts is relatively low[1], which impacts the word similarity in Eq 6.

Sim_1	Sim_2	approx_1	approx_2	R-1
cosine	cosine	no	no	0.51425
cosine	cosine	yes	2*perc	0.60621
centroid	cosine	yes	2*perc	0.65024
corpus	corpus	yes	2*perc	0.65129
corpus_pos	corpus_pos	yes	2*perc	0.61733

Table 3. Summarization results (ROUGE R-1 F-measure) using different similarity approaches on ASR output.

4. CONCLUSION AND FUTURE WORK

In this paper, we have evaluated different similarity measures under the MMR framework for meeting summarization. The centroid score focuses on the salient words of a text segment, ignoring words with lower TF-IDF values. The corpus-based semantic approach estimates the similarity of two segments based on their word distribution on a large corpus. Our experimental results have shown that these methods outperform the commonly used cosine similarity both on manual and ASR transcripts. In addition, we also found that using approximation in MMR does not hurt performance, while significantly increasing the speed.

The proper measurement of text similarity is an important topic in information retrieval and text summarization. We will continue to leverage the improvement in those domains for speech summarization using MMR as well as other modeling approaches. In addition, currently we use the human annotated sentences, therefore we will evaluate the effect from automatic sentence segmentation in our future work. Finally, different from text summarization, meeting recordings contain rich information such as multiple speakers and prosody. We will investigate incorporating these information into the MMR framework.

5. ACKNOWLEDGE

This work is supported by NSF grant IIS-0714132. Any opinions expressed in this work are those of the authors and do not necessarily reflect the views of NSF.

6. REFERENCES

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

[2] Sameer Maskey and Julia Hirschberg, "Summarizing speech without text using Hidden Markov Models," in *Proceedings of HLT-NAACL*, 2006.

[3] Anne Hendrik Buist, Wessel Kraaij, and Stephan Raaijmakers, "Automatic summarization of meeting data: A feasibility study," in *Proceedings of the 15th CLIN conference*, 2005.

[4] Michel Galley, "A skip-chain conditional random field for ranking meeting utterances by importance," in *Proceedings of EMNLP*, July 2006.

[1]The meeting corpus is not annotated with POS information, therefore we cannot evaluate the POS tagging performance.

[5] Jian Zhang and Pascale Fung, "Speech summarization without lexical features for mandarin broadcast news," in *Proceedings of HLT-NAACL*, April 2007.

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

[7] Rada Mihalcea, Courtney Corley, and Carlo Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity.," in *Proceedings of the American Association for Artificial Intelligence*, September 2006.

[8] Gabriel Murray and Steve Renals, "Term–weighting for summarization of multi-party spoken dialogues," in *Proceedings of Interspeech*, September 2007.

[9] Dragomir Radev, Hongyan Jing, and Malgorzata Budzikowska, "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies," in *NAACL-ANLP 2000 Workshop on Automatic summarization*, April 2000.

[10] Peter Turney, "Mining the web for synonyms: PMI–IR versus LSA on TOEFL," in *Proceedings of the 12th European Conference on Machine Learning*, 2001.

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

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

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

[14] Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proceedings of 5th SIGDAL Workshop*, 2004.

[15] Thorsten Brants, "TnT – a statistical part-of-speech tagger," in *Proceedings of the 6th Applied NLP Conference*, April 2000.

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

[17] Xiaodan Zhu and Gerald Penn, "Summarization of spontaneous conversations," in *Proceedings of Interspeech*, September 2006.

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on May 07,2010 at 21:42:26 UTC from IEEE Xplore. Restrictions apply.