CHINESE SPOKEN DOCUMENT SUMMARIZATION USING PROBABILISTIC LATENT TOPICAL INFORMATION

Berlin Chen, Yao-Ming Yeh, Yao-Min Huang, Yi-Ting Chen

Graduate Institute of Computer Science & Information Engineering,
National Taiwan Normal University, Taipei, Taiwan
berlin@csie.ntnu.edu.tw

ABSTRACT

The purpose of extractive summarization is to automatically select a number of indicative sentences, passages, or paragraphs from the original document according to a target summarization ratio and then sequence them to form a concise summary. In the paper, we proposed the use of probabilistic latent topical information for extractive summarization of spoken documents. Various kinds of modeling structures and learning approaches were extensively investigated. In addition, the summarization capabilities were verified by comparison with the conventional vector space model and latent semantic indexing model, as well as the HMM model. The experiments were performed on the Chinese broadcast news collected in Taiwan. Noticeable performance gains were obtained.

1. INTRODUCTION

Due to the successful development of much smaller electronic devices and the popularity of wireless communication and networking, it is widely believed that speech will play a more active role and will serve as the major human-machine interface for the interaction between people and different kinds of smart devices in the near future. On the other hand, huge quantities of multimedia contents including speech information, such as that in broadcast radio and television programs, lectures, voice mails, digital libraries, and so on, are continuously growing and filling our computers, networks and lives. It is obvious that speech is one of the most important sources of information for multimedia contents, and the understanding and organization of these contents using speech is now becoming more and more emphasized [1, 2]. For example, substantial efforts and very encouraging results on spoken document transcription, retrieval and summarization have been reported in the last few years [3].

Research work in automatic summarization of text documents can be dated back to the early work in the late fifties, and the efforts continued through decades. The World Wide Web not only led to a renaissance of this area, but extended it to cover a wider range of new tasks, including multi-document, multilingual and multi-media summarization [4]. The summarization can in general be either extractive or abstractive. The extractive summarization tries to select a number of indicative sentences, passages or paragraphs from the original document according to a target summarization ratio, and then sequence them together to form a summary. The abstractive summarization, on the other hand, tries to produce a concise abstract of desired length that can reflect the key concepts of the document. The latter seems to be more difficult, and recent approaches have focused more on the former. As one example, the vector space model (VSM) originally formulated for information retrieval (IR) can be used to respectively represent each sentence of the document, as well as the whole document, in a vector form, in which each dimension specifies the weighted statistics associated with an indexing term (or word) in the sentence or document, and the sentences that have the highest relevance scores (e.g., in the cosine measure) to the whole document are selected to be included in the summary. When it is desired to cover more important but different concepts in the summary, after the first sentence with the highest relevance score is selected, indexing terms in that sentence can be removed from the rest of sentences and the vectors are reconstructed, based on which the next sentence can be selected, and so on [5]. As another example, the latent semantic analysis (LSA) model for IR also can be used to represent each sentence of a document as a vector in the latent semantic space for that document, which is constructed by performing SVD on the "term-sentence" matrix for that document. The right singular vectors with larger singular values represent dimensions for more important latent semantic concepts in that document. Therefore the sentences that have the largest index values in each of the top m right singular vectors are included in the summary [5]. As still another example, each sentence in the document, represented as a sequence of terms, can be simply given a significance score which is evaluated using a weighted combination of statistical and linguistic measures, and the sentence selection can be performed based on this score [6]. These selected sentences in all the above cases can also be further condensed and shortened by removing some less important terms, if a higher compression ratio is desired. A survey on the use of the above approaches to extractive summarization and the other IR-related tasks, for the purpose of spoken document understanding and organization, can also be found in [2].

All the above equally applies to both text and spoken documents. However, the spoken documents bring extra difficulties such as the recognition errors, problems with spontaneous speech, and lack of correct sentence or paragraph boundaries. In order to avoid the redundant or incorrect parts while selecting the important and correct information, multiple recognition hypotheses, confidence scores, language model scores and other grammatical knowledge have been utilized [3, 7]. In addition, prosodic features (e.g., intonation, pitch, energy, pause duration) can be used as important clues for summarization as well; although reliable and efficient approaches to use these prosodic features are still under active research [8, 9]. The summary of spoken documents can be in either text or speech form. The text form has the advantages of easier browsing and further processing, but is inevitably subject to speech recognition errors, as well as the loss of the

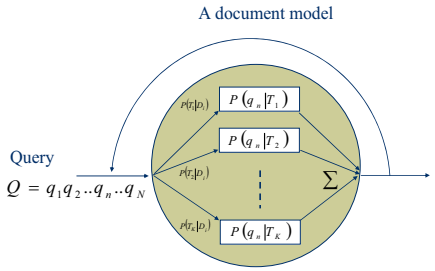

Figure 1: The TMM model for a specific document D_i.

speaker/emotional/prosodic information carried only by the speech signals.

In contrast to the above mentioned approaches, in the paper, we attempt to deal with the extractive summarization problem under a probabilistic framework by investigating the use of a topical mixture model for spoken document summarization, which is capable of exploring the probabilistic latent topical information conveyed in the spoken documents. Various kinds of modeling structures and training approaches are investigated. Moreover, the summarization capabilities are verified by comparison with the other summarization models. The proposed summarization model has also been successfully integrated into our prototype system for voice retrieval of Mandarin broadcast news via mobile devices [10].

2. Topical Mixture Model (TMM)

In IR, the relevance measure between a query Q and a document D_i can be expressed as $P(D_i|Q)$; i.e., the probability that the document D_i is relevant given that the query Q was posed. Based on Bayes' theorem and some independence assumptions, this measure can be approximated by $P(Q|D_i)$ and expressed using the following formula:

$$P(Q|D_i) = \prod_{w_n \in Q} P(w_n|D_i)^{c(w_n,Q)}, \tag{1}$$

where $c(w_n,Q)$ is the occurrence count of a term (or word) w_n in the query Q. Each individual document D_i can be interpreted as a probabilistic generative topical mixture model (TMM) [11], as depicted in Figure 1, which is just a special case of HMM. In this model, a set of K latent topical distributions characterized by unigram language models are used to predict the query terms, and each of the latent topics is associated with a document-specific weight. That is, each document can belong to many topics. The relevance measure therefore can be further expressed as:

$$P(Q|D_i) = \prod_{w_n \in Q} \left[\sum_{k=1}^{K} P(w_n|T_k) P(T_k|D_i) \right]^{c(w_n,Q)}, \tag{2}$$

where $P(w_n|T_k)$ and $P(T_k|D_i)$ respectively denote the probability of the term w_n occurring in a specific latent topic T_k and the posterior probability (or weight) of topic T_k conditioned on the document D_i. More precisely, the topical unigram distributions, e.g. $P(w_n|T_k)$, are tied among the entire document collection, while each document D_i has its own probability distribution over the latent topics, e.g., $P(T_k|D_i)$. Notice that such a relevance measure is not computed directly based on the frequency of the query terms occurring in the document, but instead through the frequency of the query terms in the latent topics as well as the likelihood that the document generates the respective topics, which in fact exhibits some sort of concept matching. The K-means algorithm can be first used to partition the entire document collection into K topical classes, and the initial topical unigram distribution for a cluster topic can be estimated according to the underlying statistical characteristics of the documents being assigned to it. While the probabilities for each document generating the topics are measured according to its proximity to the centroid of each respective cluster as well. The TMM model can be optimized by the expectation-maximization (EM) algorithm in either an unsupervised manner by using each individual document in the collection as a query exemplar to train its own TMM model, or in a supervised manner by using a training set of query exemplars with the corresponding query-document relevance information. A more detailed elucidation of the TMM model and its comparison to the other retrieval models, such as the probabilistic latent semantic analysis (PLSA) retrieval model, can be found in [11].

While the TMM modeling approach is applied to extractive summarization of broadcast news, a set of contemporary (or in-domain) text news documents with corresponding human-generated titles (a title can be viewed as an extremely short summary of a document) can be first collected to train their corresponding mixture models. For each document D_j, the human-generated title H_j is instead treated here as a TMM model used to generate the document itself:

$$P(D_j|H_j) = \prod_{w_n \in D_j} \left[\sum_{k=1}^{K} P(w_n|T_k) P(T_k|H_j) \right]^{c(w_n,D_j)}, \tag{3}$$

where $c(w_n,D_j)$ is the occurrence count of a term w_n in D_j. The title TMM models thus can be first trained by the K-means algorithm (i.e. by partitioning the entire titles of the document collection into K topical clusters) and then by the EM algorithms to optimize the probability that each title TMM model H_j generates its respective document D_j. Our postulation is that the latent topical factors $P(w_n|T_k)$ properly constructed based on the "title-document" relationships might provide very helpful clues for the subsequent spoken document summarization task. As a result, when performing extractive summarization of a broadcast news document D_g, we can treat each sentence $S_{g,l}$ of the document D_g as a TMM model for predicting the document D_g itself:

$$P(D_g|S_{g,l}) = \prod_{w_n \in D_g} \left[\sum_{k=1}^{K} P(w_n|T_k) P(T_k|S_{g,l}) \right]^{c(w_n,D_g)}. \tag{4}$$

In this way, we can keep the latent topical factors $P(w_n|T_k)$ unchanged, as those previously obtained by the information of the "title-document" pairs of the contemporary text news documents, but optimize the sentences' probability distributions over the latent topics, $P(T_k|S_{g,l})$, alone using the EM algorithm:

$$\hat{P}(T_k|S_{g,l}) = \frac{\sum_{w_n \in S_{g,l}} c(w_n,S_{g,l}) P(T_k|w_n,S_{g,l})}{\sum_{w_n \in S_{g,l}} c(w_n,S_{g,l})}, \tag{5}$$

$$P(T_k|w_n,S_{g,l}) = \frac{P(w_n|T_k) P(T_k|S_{g,l})}{\sum_{k=1}^{K} P(w_n|T_k) P(T_k|S_{g,l})}, \tag{6}$$

where $c(w_n,S_{g,l})$ is the occurrence count of a term w_n in the sentence $S_{g,l}$, and $P(T_k|w_n,S_{g,l})$ is the probability that the latent topic T_k occurs given the term w_n and the sentence $S_{g,l}$. Once

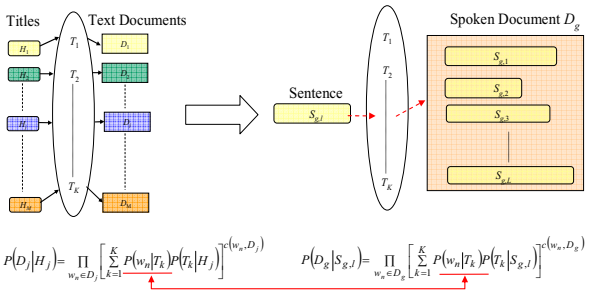

Figure 2: A schematic representation of extractive broadcast news summarization using the TMM models.

the TMM models for the sentences are estimated, they can thus be used to predict the occurrence probability of the terms in the spoken document, and the sentences with highest probabilities can be thus selected and sequenced to form the final summary according to different summarization ratios. Figure 2 depicts a schematic representation of extractive broadcast news summarization using the TMM models.

3. EXPERIMENTAL SETUP

3.1. Speech and Text Corpora

The speech data set consists of about 176 hours of radio/TV broadcast news, which were collected from several radio and TV stations located at Taipei during 1998 to 2004 [10]. Among them, a set of 200 broadcast news documents (1.6 hours) collected in August 2001 were reserved for the summarization experiments [2], and three human subjects were instructed to do the human summarization, to be taken as the references for evaluation, in two forms: the first simply to rank the importance of the sentences in the reference transcript of the broadcast news document from the top to the middle, and the second to write an abstract for the document by himself with a length being roughly 25% of the original broadcast news story. Several summarization ratios were tested, which are the ratios of summary length to the total length. Let E denote the extractive summary which was obtained from the concatenation of the top several important sentences selected by the human subject, and A the abstractive summary which was written by the subject. The summarization accuracy, R_g, of the g-th broadcast news document is then the averaged similarity score for the automatic summary, $\overline{E}$, with respect to E and A [12]:

$$R_g = \frac{1}{2}\left[sim(\overline{E}, E) + sim(\overline{E}, A)\right], \tag{7}$$

where the similarity scores $sim(\overline{E}, E)$ and $sim(\overline{E}, A)$ are calculated in the cosine measure based on the vector representations of the automatic and human-produced summaries. In this way, higher accuracy would be obtained if more sentences that are important in the broadcast news documents are included in the automatic summaries. The final summarization accuracy is defined as the average of R_g in (7) over all the broadcast news documents and all the three human subjects [2].

The rest of speech data was used for acoustic model training for speech recognition, in which only about 4.0 hours of data equipped with corresponding orthographic transcripts was used to bootstrap the acoustic model training, while 104.3 hours of the rest untranscribed speech data was reserved for unsupervised acoustic model training [13]. The acoustic models were further optimized by the minimum phone error (MPE) training algorithm [14]. On the other hand, a large volume of text news documents collected from Central News Agency (CNA) during 1991 to 2002 (the Chinese Gigaword Corpus released by LDC) were used. The text news documents collected in 2000 and 2001 were used to train the N-gram language models for speech recognition. While a subset of about 14,000 text news documents collected in the same time period as that of the broadcast news documents to be summarized (August 2001) were also used to train the latent topical distributions for the TMM models, as mentioned in Section 2.

3.2. Broadcast News Transcription

The front-end processing was conducted with the HLDA-based data-driven Mel-frequency feature extraction approach and then processed by MLLT transformation for feature de-correlation. On the other hand, the speech recognizer was implemented with a left-to-right frame-synchronous Viterbi tree search as well as a lexical prefix tree organization of the lexicon. The recognition hypotheses were organized into a word graph for further language model rescoring [13]. In this study, the word bigram language model was used in the tree search procedure while the trigram language model was used in the word graph rescoring procedure. The Chinese character error rate (CER) achieved for the 200 broadcast news documents to be summarized was 14.17%.

4. EXPERIMENTAL RESULTS

The summarization results (in the cosine measure) obtained by the TMM models are shown in Table 1, where each column illustrates the accuracies for different summarization ratios and different latent topics used. As can be seen, the summarization performance is almost the same for different model structures, and the accuracies are about 0.37, 0.40, 0.45 and 0.55 for summarization ratios of 10%, 20%, 30% and 50%, respectively. Then, we try to compare the TMM model with the conventional VSM [5] and LSA models. VSM is a typical example for literal term matching, while LSA for concept matching [2]. Two variants of LSA, i.e., the one mentioned in Section 1 [5] (LSA-1) and the one in [7] (LSA-2), were both evaluated here. We also proposed the use of HMM (Hidden Markov Model) model for the extractive summarization task. Each sentence of a document was treated as a probabilistic generative model (or an HMM) consisting of N-gram distributions for predicting the document, which were directly estimated from each sentence itself and smoothed by N-gram distributions estimated from a large text corpus. In this paper, only unigram modeling was initially investigated for HMM:

$$P(D_g|S_{g,l}) = \prod_{w_n \in D_g} \left[\lambda \cdot P(w_n|S_{g,l}) + (1-\lambda)P(w_n|Corpus)\right]^{c(w_n, D_g)}, \tag{8}$$

where λ is a weighting parameter. Notice that the HMM model can be similarly trained by the EM algorithm, and is also another example for literal term matching. The results for these models are shown in Table 2, and the results obtained by random selection (Random) were also listed for comparison. As can been seen, TMM is competitive with VSM and HMM, and is significant better than the two variants of LSA, which evidences that TMM is indeed a good candidate of concept matching for the summarization task. We also used the

	2	4	8	16	32	64
10%	0.3658	0.3658	0.3675	0.3658	0.3675	0.3662
20%	0.3952	0.3952	0.3967	0.3948	0.3958	0.3957
30%	0.4475	0.4477	0.4480	0.4482	0.4470	0.4450
50%	0.5470	0.5469	0.5467	0.5460	0.5463	0.5478

Table 1: The results (in the cosine measure) achieved by the TMM model using different mixture numbers and under different summarization ratios.

	VSM	LSA-1	LSA-2	HMM	Random
10%	0.3596	0.3339	0.3145	0.3647	0.2239
20%	0.3895	0.3566	0.3514	0.3929	0.2524
30%	0.4428	0.3986	0.4109	0.4447	0.3274
50%	0.5409	0.5034	0.5330	0.5453	0.4582

Table 2: The results (in the cosine measure) achieved by the VSM, LSA and HMM models and random selection under different summarization ratios.

	2	4	8	16	32	64
10%	0.2994	0.2994	0.3043	0.3014	0.2966	0.2934
20%	0.3296	0.3296	0.3345	0.3351	0.3274	0.3267
30%	0.3691	0.3693	0.3688	0.3663	0.3629	0.3609
50%	0.4763	0.4759	0.4753	0.4738	0.4757	0.4773

Table 3: The results (in the ROUGE-2 measure) achieved by the TMM model using different mixture numbers and under different summarization ratios.

	VSM	LSA-1	LSA-2	HMM	Random
10%	0.2845	0.2755	0.2498	0.2989	0.1122
20%	0.3110	0.2911	0.2917	0.3295	0.1263
30%	0.3435	0.3081	0.3378	0.3670	0.1834
50%	0.4565	0.4070	0.4666	0.4743	0.3096

Table 4: The results (in the ROUGE-2 measure) achieved by the VSM, LSA and HMM models and random selection under different summarization ratios.

	2	4	8	16	32	64
10%	0.3655	0.3667	0.3554	0.3640	0.3644	0.3748
20%	0.3907	0.3935	0.3805	0.3893	0.3913	0.4000
30%	0.4457	0.4447	0.4311	0.4339	0.4370	0.4428
50%	0.5493	0.5452	0.5415	0.5533	0.5456	0.5450

Table 5: The results (in the cosine measure) achieved by the TMM model trained in an unsupervised mode, and using different mixture numbers and under different summarization ratios.

	2	4	8	16	32	64
10%	0.3062	0.3081	0.3081	0.2932	0.2983	0.3175
20%	0.3290	0.3341	0.3341	0.3095	0.3215	0.3377
30%	0.3711	0.3648	0.3648	0.3455	0.3498	0.3545
50%	0.4781	0.4741	0.4741	0.4816	0.4716	0.4648

Table 6: The results (in the ROUGE-2 measure) achieved by the TMM model trained in an unsupervised mode, and using different mixture numbers and under different summarization ratios.

ROUGE-2 measure [15, 7, 9] to evaluate the performance levels of TMM and the other models. The results are shown in Tables 3 and 4, respectively (the larger the values the better the results). It can be found that TMM is substantially better than VSM and LSA, and again competitive with HMM.

On the other hand, in most real-world applications, it is not always the case that the spoken document summarization systems can have contemporary or in-domain text news documents for model training. Thus, we study here the use of unsupervised training for TMMs by merely using all the possible "sentence-document" pairs of the broadcast news to be summarized to construct the latent topical space and then the sentence TMM models. The results are shown in Tables 5 and 6 for different evaluation metrics. Compared to the results in Tables 1 and 3, it can be found that the results obtained by TMMs trained without supervision are quite similar to those of the TMMs trained with supervision.

5. CONCLUSIONS

In the paper, we have studied the use of topical mixture model for extractive spoken document summarization. Various kinds of modeling complexities and learning approaches were extensively investigated. In addition, the summarization capabilities were verified by comparison with the other summarization models. Noticeable and consistent performance gains were obtained. The proposed summarization technique has also been properly integrated into our prototype system for voice retrieval of Mandarin broadcast news via mobile devices.

6. REFERENCES

[1] B.H. Juang and S. Furui, "Automatic Recognition and Understanding of Spoken Language—A First Step Toward Natural Human–Machine Communication," *Proceedings of the IEEE*, 88(8), 2000.

[2] L.S. Lee and B. Chen, "Spoken Document Understanding and Organization," *IEEE Signal Processing Magazine*, 22(5), 2005.

[3] S. Furui et al., "Speech-to-Text and Speech-to-Speech Summarization of Spontaneous Speech," *IEEE Trans. on Speech and Audio Processing*, 12(4), 2004.

[4] I. Mani and M. T. Maybury, Eds., Advances in Automatic Text Summarization. Cambridge. MA: MIT Press, 1999.

[5] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," *SIGIR 2001*.

[6] J. Goldstein et al., "Summarizing text documents: sentence selection and evaluation metrics," *SIGIR 1999*.

[7] S. Hirohata et al., "Sentence Extraction-based Presentation Summarization Techniques and Evaluation Metrics," *ICASSP 2005*.

[8] K. Koumpis and S. Renals, "Automatic Summarization of Voicemail Message Using Lexical and Prosodic Features," *ACM Trans. on Speech and Language Processing*, 2(1), 2005.

[9] S. Maskey and J. Hirschberg, "Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization," *EUROSPEECH 2005*.

[10] B. Chen et al., "Speech Retrieval of Mandarin Broadcast News via Mobile Devices," *EUROSPEECH 2005*.

[11] B. Chen, "Exploring the Use of Latent Topical Information for Statistical Chinese Spoken Document Retrieval," *Pattern Recognition Letters*, 27(1), 2006.

[12] E. Hovy and D. Marcu, "Automated text summarization tutorial," *COLING/ACL 1998*.

[13] B. Chen et al., "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription," *ICASSP 2004*.

[14] J.W. Kuo, B. Chen, "Minimum Word Error Based Discriminative Training of Language Models," *EUROSPEECH 2005*.

[15] C.Y. Lin, "Looking for a Few Good Metrics: ROGUE and Its Evaluation," working notes of *NCTIR-4*, 2004.