WORD TOPICAL MIXTURE MODELS FOR EXTRACTIVE SPOKEN DOCUMENT SUMMARIZATION

Berlin Chen and Yi-Ting Chen

Department of Computer Science & Information Engineering,
National Taiwan Normal University, Taipei, Taiwan
{berlin, g93470070}@csie.ntnu.edu.tw

ABSTRACT

This paper considers extractive summarization of Chinese spoken documents. In contrast to conventional approaches, we attempt to deal with the extractive summarization problem under a probabilistic generative framework. A word topical mixture model (w-TMM) was proposed to explore the co-occurrence relationship between words of the language. Each sentence of the spoken document to be summarized was treated as a composite word TMM model for generating the document, and sentences were ranked and selected according to their likelihoods. Various kinds of modeling structures and learning approaches were extensively investigated. In addition, the summarization capabilities were verified by comparison with the other conventional summarization approaches. The experiments were performed on the Chinese broadcast news collected in Taiwan. Noticeable performance gains were obtained. The proposed summarization technique has also been properly integrated into our prototype system for voice retrieval of broadcast news via mobile devices.

1. INTRODUCTION

Due to the rapid development and maturity of multimedia technology, large volumes of information content have been represented as audio-visual multimedia instead of static texts. Clearly, speech is one of the most important sources of information about multimedia content. However, unlike text documents, which are structured with titles and paragraphs and are thus easier to retrieve and browse, associated spoken documents of multimedia content are only presented with video or audio signals; hence, they are difficult to browse from beginning to end. Even though spoken documents are automatically transcribed into words, incorrect information (resulting from recognition errors and inaccurate sentence or paragraph boundaries) and redundant information (generated by disfluencies, fillers, and repetitions) prevent them from being accessed easily. Spoken document summarization, which attempts to distill important information and remove redundant and incorrect content from spoken documents, can help users review spoken documents efficiently and understand associated topics quickly [1, 2].

Extractive spoken document summarization is to automatically select a number of indicative sentences from the original document according to a target summarization ratio and then sequence them to form a concise summary. Quite several approaches have been developed for this task, and they in general can fall into three main categories: 1) approaches based on the sentence structure or location information, 2) approaches based on statistical features, and 3) approaches based on a probabilistic generative framework.

In [3, 4], the authors suggested that important sentences can be selected from the significant parts of a document. For example, sentences can be selected form the introductory and concluding parts. However, such approaches can be only applied to some specific domains or document structures.

Statistical approaches for extractive spoken document summarization attempt to select salient sentences based on statistical features of the sentences or of the words in the sentences. Statistical features, for example, can be the term (word) frequency, linguistic score and recognition confidence measure, as well as the prosodic information. The associated methods based on these features have gained much attention of research; among them, the vector space model (VSM) [1], latent semantic analysis (LSA) method [5], maximum marginal relevance (MMR) method [6], sentence significant score method [4, 7] are the most popular for spoken document summarization. Besides, a bulk of classification-based methods using statistical features also have been developed, such as the Gaussian mixture models (GMM) [6], Bayesian network classifier [8], support vector machine (SVM) and logistic regression [9]. In these methods, sentence selection is formulated as a binary classification problem. A sentence can either be included in a summary or not. However, these methods need a set of training documents together with their corresponding handcrafted summaries (or labeled data) for training the classifiers.

Recently, yet another set of approaches based on a probabilistic generative framework also have been proposed. In such approaches, each sentence of a document is treated as a probabilistic generative model for predicting the document, and the sentences are ranked and selected according to their likelihoods. The hidden Markov model (HMM) [10] and sentence topical mixture model (S-TMM) [11] both have demonstrated competitive results in the Chinese spoken document summarization task.

In this paper, we propose the use of a word topical mixture model (w-TMM) exploring the co-occurrence relationship between words for extractive spoken document summarization. Each sentence of the spoken document was treated as a composite word TMM model for generating the document, and sentences were ranked and selected according to their likelihoods. Various kinds of modeling structures and learning approaches were extensively investigated. In addition, the summarization capabilities were verified by comparison with the other conventional summarization approaches. The proposed summarization technique has also been properly integrated into our prototype system for voice retrieval of Mandarin broadcast news via mobile devices [12].

The remainder of this paper is organized as follows. In Section 2, we introduce the probabilistic generative framework for extractive spoken document summarization and elucidate the proposed word topical mixture model. Then, the experimental settings and a series of summarization experiments are presented in Sections 3 and 4, respectively. Finally, conclusions and future work are given in Section 5.

2. PROPOSED SUMMARIZATION MODEL

2.1. Probabilistic Generative Framework

In the probabilistic generative framework for extractive spoken document summarization, important sentences S_i of a document D can be selected (or ranked) based the posterior probability of the sentence given the document $P(S_i|D)$, which can be transformed to the following equation by applying Bayes' rule:

$$P(S_i|D) = \frac{P(D|S_i)P(S_i)}{P(D)}, \quad (1)$$

where $P(D|S_i)$ is the likelihood of the document D being generated by a sentence S_i, $P(S_i)$ is the prior probability of the sentence S_i, and $P(D)$ is the prior probability of the document. $P(D)$ in Eq.(1) can be eliminated because it is identical for all sentences and will not affect the ranking of the sentences. Furthermore, because the way to estimate the probability $P(S_i)$ is still unknown, we may simply assume that $P(S_i)$ is uniformly distributed, or identical for all sentences. In this way, the sentences of the spoken document to be summarized can be ranked by means of the probability $P(D|S_i)$ instead of using the probability $P(S_i|D)$.

In our previously proposed sentence topical mixture model (S-TMM) [11], each sentence of the document to be summarized is represented as a probabilistic generative model consisting of a set of K latent topical distributions for predicting the document, such that the likelihood of the document D being generated by a sentence S_i can be expressed as:

$$P(D|S_i) = \prod_{w \in D} \left[\sum_{k=1}^{K} P(w|T_k) P(T_k|S_i) \right]^{n(w,D)}, \quad (2)$$

where $P(w|T_k)$ and $P(T_k|S_i)$ respectively denote the probability of a word w occurring in a specific latent topic T_k and the weight of a topic T_k conditioned on the sentence S_i; $n(w,D)$ is the number of times a word w occurring in D. The words in D are assumed to be conditionally independent given S_i. The probability $P(w|T_k)$ can be estimated beforehand using a set of contemporary (or in-domain) text news documents. However, because the sentences of the spoken document to be summarized are not known in advance, the sentence's probability distribution over the latent topics $P(T_k|S_i)$ has to be estimated on the fly. For example, during summarization, we can keep the topic factors $P(w|T_k)$ unchanged, but let the sentence's probability distribution over the latent topics $P(T_k|S_i)$ be estimated in an on-line manner [11].

2.2. Word Topic Mixture Model (w-TMM)

In this paper, we present an alternative probabilistic latent topic approach by treating each word w_j of the language as a word topical mixture model (w-TMM) M_{w_j} for predicting the occurrences of the other word w:

$$P(w|M_{w_j}) = \sum_{k=1}^{K} P(w|T_k) P(T_k|M_{w_j}), \quad (3)$$

where $P(w|T_k)$ and $P(T_k|M_{w_j})$ are respectively the probability of a word w occurring in a specific latent topic T_k and the probability of a topic T_k conditioned on M_{w_j}. During the summarization process, we can linearly combine the associated TMM models of the words involved in a sentence S_i to form a composite word TMM model for S_i, and the likelihood of the document D being generated by S_i can be expressed as:

$$P(D|S_i) = \prod_{w \in D} \left[\sum_{w_j \in S_i} \alpha_{j,i} \sum_{k=1}^{K} P(w|T_k) P(T_k|M_{w_j}) \right]^{n(w,D)}, \quad (4)$$

where the weighting coefficient $\alpha_{j,i}$ is set to be in proportion to the frequency of w_j occurring in S_i and summed to 1 ($\sum_{w_j \in S_i} \alpha_{j,i} = 1$). Then, the sentences with the highest likelihoods can be thus selected and sequenced to form the final summary according to different summarization ratios.

When the word TMM modeling approach is applied to extractive summarization of broadcast news, we can use 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) to train the word topical mixture models. For each training document D_c, its human-generated title H_c is instead treated here as a composite word TMM model used to generate the document itself:

$$P(D_c|H_c) = \prod_{w \in D_c} \left[\sum_{w_j \in H_c} \alpha_{j,c} \sum_{k=1}^{K} P(w|T_k) P(T_k|M_{w_j}) \right]^{n(w,D_c)}. \quad (5)$$

The parameters of the word TMM models can be estimated by the expectation-maximization (EM) algorithm [13], using the following formulae:

$$\hat{P}(w|T_k) = \frac{\sum_{D_c} n(w,D_c) P(T_k|w,H_c)}{\sum_{D_{c'}} \sum_{w_n \in D_{c'}} n(w_n,D_{c'}) P(T_k|w_n,H_{c'})}, \quad (6)$$

$$\hat{P}(T_k|M_{w_j}) = \frac{\sum_{D_c} \sum_{w \in D_c} n(w,D_c) P(M_{w_j}|w,M_{H_c}) P(T_k|w,M_{w_j})}{\sum_{D_{c'}} \sum_{w' \in D_{c'}} n(w',D_{c'}) P(M_{w_j}|w',M_{H_{c'}})}, \quad (7)$$

where $P(T_k|w,M_{H_c})$ can be expressed as:

$$P(T_k|w,M_{H_c}) = \frac{P(w|T_k) \left[\sum_{w_j \in H_c} \alpha_{j,c} P(T_k|M_{w_j}) \right]}{\sum_{l=1}^{K} P(w|T_l) \left[\sum_{w_j \in H_c} \alpha_{j,c} P(T_l|M_{w_j}) \right]}, \quad (8)$$

and $P(T_k|w,M_{w_j})$ can be expressed as:

$$P(T_k|w,M_{w_j}) = \frac{P(w|T_k) P(T_k|M_{w_j})}{\sum_{l=1}^{K} P(w|T_l) P(T_l|M_{w_j})}, \quad (9)$$

and $P(M_{w_j}|w,M_{H_c})$ can be expressed as:

$$P(M_{w_j}|w,M_{H_c}) = \frac{\alpha_{j,c} P(w|M_{w_j})}{\sum_{w_l \in H_c} \alpha_{l,c} P(w|M_{w_l})}. \quad (10)$$

Our postulation is that the co-occurrence relationship between the words in the titles and the words in the documents might

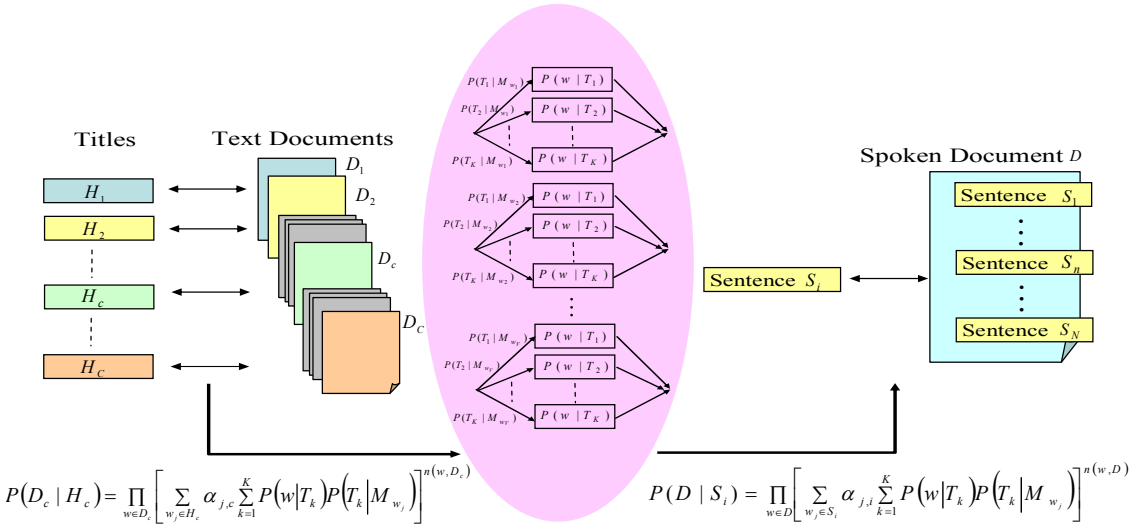

Figure 1: A schematic depiction of extractive spoken document summarization using word TMM models.

provide very helpful clues for the subsequent broadcast news summarization task.

It is also noteworthy that unlike the sentence topical mixture model where the topic mixture weights trained with text news documents are entirely discarded during the summarization process [12], the topic mixture weights of word TMM models are instead retained and exploited. Figure 1 shows a schematic depiction of extractive spoken document summarization using word TMM models.

3. EXPERIMENTAL SETUP

3.1. Speech and Text Corpora

The speech data set was comprised of approximately 176 hours of radio and TV broadcast news documents collected from several radio and TV stations in Taipei between 1998 and 2004. From them, a set of 200 documents (1.6 hours) collected in August 2001, was reserved for the summarization experiments [1]. The remainder of the speech data was used to train an acoustic model for speech recognition, of which about 4.0 hours of data with corresponding orthographic transcripts was used to bootstrap the acoustic model training, while 104.3 hours of the remaining un-transcribed speech data was reserved for unsupervised acoustic model training [14]. The acoustic models were further optimized by the minimum phone error (MPE) training algorithm. The Chinese character error rate (CER) for the 200 broadcast news documents reserved for summarization was 14.17%.

A large number of text news documents collected from the Central News Agency (CNA) between 1991 and 2002 (the Chinese Gigaword Corpus released by LDC) was also used [15]. The text news documents collected in 2000 and 2001 were used to train *n*-gram language models for speech recognition; and a subset of about 14,000 text news documents collected in the same period as that of the broadcast news documents to be summarized (August 2001) was used to training the word TMM models.

3.2. Evaluation Metric

Three subjects were asked to summarize the 200 broadcast news documents (testing corpus), which were to be used as references for evaluation [1]. In addition, the ROUGE measure [16] was used to evaluate the performance levels of the proposed models and the conventional models. The measure evaluates the quality of the summarization by counting the number of overlapping units, such as *n*-grams and word sequences, between the automatic summary and a set of reference (or manual) summaries. ROUGE-N is an *n*-gram recall measure defined as follows:

$$ROUGE-N = \frac{\sum_{S \in \mathbf{S}_R} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \mathbf{S}_R} \sum_{gram_n \in S} Count(gram_n)}, \quad (11)$$

where N denotes the length of the *n*-gram; S is an individual reference (or manual) summary; $\mathbf{s}_R$ is a set of reference summaries; $Count_{match}(gram_n)$ is the maximum number of *n*-grams co-occurring in the automatic summary and the reference summary; and $Count(gram_n)$ is the number of *n*-grams in the reference summary. In this paper, we adopted the ROUGE-2 measure, which uses word bigrams as matching units.

4. EXPERIMENTAL RESULTS

4.1. Comparison of w-TMM and Other Summarization Models

We first evaluate the summarization performance of the word topical mixture models (w-TMM) trained using a set of contemporary text news documents and varying model complexities. The summarization results obtained by the w-TMM 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 best summarization results, especially for low summarization ratios ($\leq 20\%$), were obtained with 32 topics. The summarization accuracies for the word-TMMs with 32 topics are about 0.32, 0.34, 0.37 and 0.47 for summarization ratios of 10%, 20%, 30% and 50%, respectively.

Then, we attempt to compare the w-TMM with the conventional VSM [1], MMR [6], LSA [5], and sentence significance score (SIG) [4, 10] models, as well as our previously proposed HMM [10] and S-TMM [11] models. The results for these models trained or tuned with optimum settings are shown in Table 2, and the results obtained by random selection (Random) were also listed for comparison. As can be seen, the probabilistic generative models (w-TMM, HMM and S-HMM) significantly outperform the statistical approaches (VSM, MMR, LSA and SIG). Moreover, the best results achieved by the w-TMM are also substantially better than that achieved by the other two probabilistic generative models (HMM and S-HMM).

4.2. w-TMM Trained in an Unsupervised Manner

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 with corresponding human-generated titles for model training. Thus, in this paper we investigated an unsupervised approach for the training of the w-TMM. Each w-TMM M_{w_j} was instead trained by concatenating those words occurring within a context window of size N (for simplicity, N is set to 2 in this study) around each occurrence of w_j, which are postulated to be relevant to w_j, in the news document collection to form the observation for training M_{w_j}. The results for the w-TMM trained in such an unsupervised manner are shown in Tables 3. Compared to the results shown in Table 1, it can be found that the results obtained by the w-TMM trained without supervision are quite similar to those of the w-TMM trained with supervision.

5. CONCLUSIONS AND FUTURE WORK

In the paper, we have studied the use of the word topical mixture model for extractive Chinese 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. Exploration of using extra structural and prosodic information for modeling the sentence prior distributions for the word-TMM summarization approach is currently undertaken. The word-TMM models also have been applied to dynamic language model adaptation for speech recognition with very promising results initially demonstrated [17].

6. REFERENCES

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

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

[3] P.B. Baxendale, "Machine-Made Index for Technical Literature-An Experiment", *IBM Journal*, pp. 354-361, October 1958.

[4] M. Hirohata et al., "Sentence Extraction-Based Presentation Summarization Techniques and Evaluation Metrics", in Proc. *ICASSP 2005*.

	2	4	8	16	32	64
10%	0.3013	0.2997	0.3053	0.3152	0.3193	0.2986
20%	0.3282	0.3273	0.3302	0.3379	0.3437	0.3209
30%	0.3731	0.3641	0.3703	0.3694	0.3716	0.3713
50%	0.4732	0.4741	0.4730	0.4700	0.4676	0.4759

Table 1: The results achieved by the word TMM models (w-TMM) that were trained using a set of contemporary news documents, and using different mixture numbers and under different summarization ratios.

	VSM	MMR	LSA	SIG	HMM	S-TMM	Random
10%	0.2845	0.2875	0.2755	0.2760	0.2989	0.3043	0.1122
20%	0.3110	0.3218	0.2911	0.3190	0.3295	0.3345	0.1263
30%	0.3435	0.3493	0.3081	0.3491	0.3670	0.3688	0.1834
50%	0.4565	0.4668	0.4070	0.4804	0.4743	0.4753	0.3096

Table 2: The results achieved by the other summarization models under different summarization ratios.

	2	4	8	16	32	64
10%	0.3108	0.3064	0.3088	0.3090	0.3130	0.3114
20%	0.3345	0.3387	0.3378	0.3380	0.3405	0.3334
30%	0.3692	0.3749	0.3731	0.3729	0.3695	0.3659
50%	0.4736	0.4750	0.4761	0.4750	0.4719	0.4708

Table 3: The results achieved by the word TMM models that were trained in an unsupervised manner and using different mixture numbers and under different summarization ratios.

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

[6] Gabriel Murray et al., "Extractive Summarization of Meeting Recordings", in Proc. *Eurospeech 2005*.

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

[8] S. Maskey, J. Hirschberg, "Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization", in Proc. *Eurospeech 2005*.

[9] X. Zhu, G. Penn, "Evaluation of Sentence Selection for Speech Summarization", in Proc *RANLP 2005*.

[10] Y.T. Chen et al., "Extractive Chinese Spoken Document Summarization Using Probabilistic Ranking Models," in Proc. *ISCSLP 2006*.

[11] B. Chen et al., "Chinese Spoken Document Summarization Using Probabilistic Latent Topical Information," in Proc. *ICASSP 2006*.

[12] B. Chen et al., "Spoken Document Retrieval and Summarization," in the book "Advances in Chinese Spoken Language Processing," World Scientific Publisher, December 30, 2006.

[13] A. P. Dempster et al., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society B* 39(1), 1977.

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

[15] Central News Agency (CNA) http://210.69.89.224/search/hypage.cgi?HYPAGE=login.htm

[16] C.Y. Lin, "ROUGE: Recall-oriented Understudy for Gisting Evaluation," 2003, http://www.isi.edu/~cyl/ROUGE/.

[17] H.S. Chiu, B. Chen, "Word Topical Mixture Models for Dynamic Language Model Adaptation," to appear in Proc. *ICASSP 2007*.