A Unified Probabilistic Generative Framework for Extractive Spoken Document Summarization

Yi-Ting Chen[1, 2], Hsuan-Sheng Chiu[1], Hsin-Min Wang[2] and Berlin Chen[1]

[1] National Taiwan Normal University, Taiwan
[2] Institute of Information Science, Academia Sinica, Taiwan
{g93470070, g93470240, berlin}@csie.ntnu.edu.tw, whm@iis,sinica.edu.tw

Abstract

In this paper, we consider extractive summarization of Chinese broadcast news speech. A unified probabilistic generative framework that combined the sentence generative probability and the sentence prior probability for sentence ranking was proposed. Each sentence of a spoken document to be summarized was treated as a probabilistic generative model for predicting the document. Two different matching strategies, i.e., literal term matching and concept matching, were extensively investigated. We explored the use of the hidden Markov model (HMM) and relevance model (RM) for literal term matching, while the word topical mixture model (WTMM) for concept matching. On the other hand, the confidence scores, structural features, and a set of prosodic features were properly incorporated together using the whole sentence maximum entropy model (WSME) for the estimation of the sentence prior probability. The experiments were performed on the Chinese broadcast news collected in Taiwan. Very promising and encouraging results were initially obtained.

Index Terms: spoken document summarization, hidden Markov model, relevance model, word topical mixture model, whole sentence maximum entropy model

1. Introduction

Huge quantities of multimedia contents including audio and video are continuously growing and filling networks and our lives. Speech information is one of the most important sources for multimedia contents, and usually represents the concepts and topics. Therefore, multimedia access based on associated spoken documents has received a great of attention in recent years. Spoken documents are often automatically transcribed into words, while incorrect speech recognition results and redundant acoustic effects prevent them from being accessed easily. Spoken document summarization, which aims at distilling the important information and remove redundant and incorrect information from a spoken document, can help to efficiently review spoken documents and understand associated topics quickly [1].

This paper investigated extractive summarization approaches that are to select a number of indicative sentences from the original document according to a target summarization ratio, and then sequence them to form a summary. In general, the approaches can fall into three main categories: 1) approaches based on the sentence structure or location information, 2) approaches based on statistical measures, and 3) approaches based on the sentence generative probability. In [2, 3], the authors suggested that important sentences can be selected from the significant parts of a document, e.g., the introductory and concluding parts. However, such approaches can only be applied to some specific domains or document structures. Statistical approaches 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) [4], latent semantic analysis (LSA) method [5], maximum marginal relevance (MMR) method [6], and sentence significant score method [3, 7] are most popular for spoken document summarization. Besides, a bulk of classification-based methods using statistical features have also been developed, such as the Bayesian network classifier [8], support vector machine (SVM), and logistic regression [9]. In these methods, sentence selection is formulated as a binary classification problem. However, these methods need a training set consisting of documents and their corresponding handcrafted summaries (or labeled data) for training the classifiers. More recently, several approaches based on the sentence generative probability have also been proposed. The hidden Markov model (HMM), relevance model [10], sentence topical mixture model (STMM) [11], and word topical mixture model (WTMM) [12, 13] all have demonstrated competitive results in the Chinese spoken document summarization task.

In this paper, a unified probabilistic generative framework that combined the sentence generative probability and the sentence prior probability for sentence ranking was investigated. Each sentence of a spoken document to be summarized was treated as a probabilistic generative model for predicting the document. Two different matching strategies, i.e., literal term matching and concept matching, were extensively investigated. We explored the use of HMM and relevance model (RM) for literal term matching, while WTMM for concept matching. In addition, a set of extra features extracted from the spoken sentences were properly incorporated together using the whole sentence maximum entropy model (WSME) for estimating the sentence prior probability.

2. Spoken Document Summarization

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 expressed as:

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

where $P(D|S_i)$ is the sentence generative probability, i.e., the likelihood of D being generated by S_i; $P(S_i)$ is the prior probability of S_i being important; and $P(D)$ is the prior probability of D. $P(D)$ in Eq.(1) can be eliminated because it is identical for all sentences and will not affect the ranking of them. The sentence generative probability $P(D|S_i)$ can be taken as a relevance measure between the document and sentences, while the sentence prior probability to some extent is a measure of importance of the sentences themselves. Therefore, the sentences of the spoken document to be summarized can be ranked by the product of the sentence generative probability $P(D|S_i)$ and the sentence prior probability $P(S_i)$. Fig. 1 illustrates a schematic depiction of extractive spoken document summarization using the probabilistic generative framework.

2.2. HMM-RM-based Sentence Generative Model

In our early work [11], HMM can be applied to extractive spoken document summarization, where each sentence S_i of the document D to be summarized is represented as a probabilistic generative model consisting of n-gram distributions for predicting D, and the terms (or words) in D are taken as an input observation sequence. For example, the sentence HMM of unigram modeling can be expressed as:

$$P_{HMM}(D|S_i) = \prod_{w \in D} \left[\lambda \cdot P(w|S_i) + (1-\lambda) \cdot P(w|C)\right]^{n(w,D)}, \quad (2)$$

where λ is a weighting parameter and $n(w,D)$ is the occurrence count of a term w in D. The sentence model $P(w|S_i)$ and the collection model $P(w|C)$ are simply estimated from the sentence itself and a large external text collection, respectively, using the maximum likelihood estimation (MLE). The weighting parameter λ in Eq. (2) can be further optimized by using the expectation-maximum (EM) training algorithm [10].

However, in HMM, the true sentence model $P(w|S_i)$ might not be accurately estimated by MLE, since the sentence only consists of a few terms, and the portions of the terms in the sentence are not the same as the probabilities of those terms in the true model. Therefore, we explored the use of the relevance model (RM), $P(w|R_{S_i})$, to derive a more accurate estimation of the sentence model [10]. Each sentence of the document to be summarized has its own associated relevant class, which can be approximated by the set of documents retrieved from a large text collection that are relevant to the sentence S_i. Therefore, $P(w|R_{S_i})$ is defined as the probability that we would observe a term if we randomly select some document form the relevant document set and then pick up a random term from the document. Once the relevance model of the sentence is constructed, it can be used to linearly combine with the original sentence model, denoted as HMM-RM, for better estimation of the sentence generative probability. Both HMM and HMM-RM belong to the literal term matching strategy [1].

2.3. WTMM-based Sentence Generative Model

We also exploited a concept matching strategy [1], called the word topical mixture model (WTMM), for representing the sentence generative probability. Each word w_j of the language is treated as a WTMM M_{w_j} for predicting the occurrences of the other word w [12]:

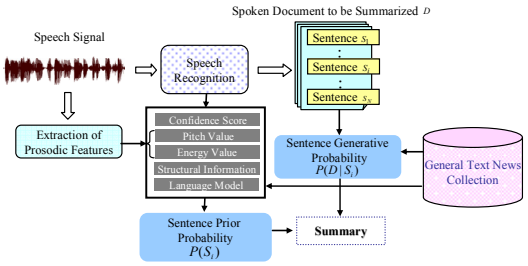

Figure 1: A schematic depiction of extractive spoken document summarization using the probabilistic generative framework.

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

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 WTMM 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 ($\sum_{w_j \in S_i} \alpha_{j,i} = 1$). In this paper, we investigated an unsupervised approach for training the WTMM models used in the sentence model. Each WTMM M_{w_j} of word w_j was trained by concatenating those words occurring within a context window of size N (for simplicity, N is set to 1 in this study) around each occurrence of w_j, which are postulated to be relevant to w_j, in the broadcast news document collection to form the observation for training M_{w_j} [13].

2.4. Sentence Prior Probability

Normally, because the way to estimate the prior probability of the sentences is still an open issue, we might simply assume that the prior probability is uniformly distributed, as reported in [10, 11, 13]. However, the importance of the sentences in the spoken document to be summarized should not be made equal, which depends on a wide variety of factors, such as the structural (positional and linguistic) information, recognition accuracy, and inherent prosodic properties, of the sentences. Therefore, in this paper, we attempted to model the sentence prior probability (or importance) based on a set of features extracted from the spoken sentences. Features to be considered include the positional information, language model score, confidence score, and prosodic information of the spoken sentences. These features have also been utilized to calculate the sentences significant scores [3, 7], or taken as the features of the classifiers [6, 8], for spoken document summarization.

It has been shown that the leading few sentences of an article are important and can provide a good summary of the article [14], which means that the positional information (denoted as F1 below) can be used to model the sentence prior probability according to the position of the sentence in the broadcast news (The front the sentence, the higher prior probability it has). The language model score (denoted as F2

below), such as the word *n*-gram score, can be used to judge the appropriateness (in syntax) of the recognized word string of the spoken sentence, while the confidence score (denoted as F3 below), expressed by the posterior probability of each transcribed word, can be used to evaluate the reliability of the recognized word string. In addition, the prosodic information, like the average pitch value (denoted as F4 below) and average energy value (denoted as F5) of a spoken sentence, are commonly used to observe the stress or emphasis on the sentence.

Each of the features mentioned above can be either used alone or incorporated together for modeling the sentence prior probability. In this paper, we used the whole sentence maximum entropy (WSME) model to efficiently combine the multiple phenomena of a sentence [15, 16]. The prior probability of sentence S_i is modeled as:

$$P(S_i) = \frac{1}{Z} P_0(S_i) \exp\left(\sum_j \lambda_{j,i} f_{j,i}(S_i) \right), \quad (5)$$

where Z is a normalization constant; $P_0(S_i)$ is an arbitrary initial probability distribution of the sentence S_i; $f_{j,i}(S_i)$ is the *j*-th predefined feature of S_i; $\lambda_{j,i}$ is the parameter corresponding to $f_{j,i}(S_i)$.

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 subset of 200 documents (1.6 hours) collected in August 2001 was reserved for the summarization experiments [4], and the Chinese character error rate (CER) 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 [17]. 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 construct RM 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. These documents were divided into two parts, each of which contained 100 spoken documents. The first part of spoken documents was taken as the development set, which formed the basis for tuning parameters or settings. The rest part of spoken documents was taken as the evaluation set; i.e., all the summarization experiments on it were conducted following the same training (or parameter) settings that were optimized based on the development set. In addition, the ROUGE measure [18] 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:

Table 1: The results achieved by the different summarization models under different summarization ratios.

	VSM	MMR	LSA	HMM	HMM-RM	WTMM
10%	0.3073	0.3073	0.3034	0.2932	0.3182	0.3248
20%	0.3188	0.3214	0.2926	0.3191	0.3264	0.3324
30%	0.3593	0.3678	0.3286	0.3705	0.3671	0.3816
50%	0.4485	0.4501	0.3906	0.4732	0.4774	0.4581

Table 2: The results achieved by the HMM-RM when the sentence prior probability was modeled by using different kinds of extra sentence features.

	F1	F2	F3	F4	F5
10%	0.4864	0.3029	0.3146	0.3172	0.3211
20%	0.4724	0.3151	0.3228	0.3216	0.3261
30%	0.4687	0.3517	0.3639	0.3590	0.3627
50%	0.4761	0.4625	0.4761	0.4763	0.4773

Table 3: The results achieved by the WTMM when the sentence prior probability was modeled by using different kinds of extra sentence features.

	F1	F2	F3	F4	F5
10%	0.4692	0.3128	0.3276	0.3179	0.3260
20%	0.4507	0.3179	0.3352	0.3250	0.3331
30%	0.4203	0.3753	0.3821	0.3725	0.3801
50%	0.4709	0.4583	0.4619	0.4602	0.4598

$$ROUGE-N = \frac{\sum_{M \in \mathbf{M}_R} \sum_{gram_n \in M} Count_{match}(gram_n)}{\sum_{M \in \mathbf{M}_R} \sum_{gram_n \in M} Count(gram_n)}, \quad (6)$$

where N denotes the length of the *n*-gram; M is an individual reference (or manual) summary; $\mathbf{M}_R$ 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 the matching units.

4. Experimental Results

4.1. Baseline Experimental Results

We first evaluate the summarization performance of the sentence generative approaches (HMM, HMM-RM and WTMM) and the conventional approaches (VSM [4], MMR [6] and LSA [5]). For the sentence generative approaches, the sentence prior distribution was assumed to be uniform. The results for these models on the evaluation set are shown in Table 1. We observe that the sentence generative approaches (especially, HMM-RM and WTMM) significantly outperform the statistical approaches. Moreover, the results achieved by WTMM are substantially better than that achieved by the other two sentence generative approaches.

4.2. Non-uniform Sentence Prior Probability

As mentioned in Section 2.4, the importance (or prior probability) of the sentences of a spoken document to be summarized should not be identical. Therefore, we tried to model the sentence prior probability $P(S_i)$ by using a set of extra features extracted from the sentences. The measure or score of each feature can be normalized and then taken as the sentence prior probability, satisfying the constraint $\sum_{S_i \in D} P(S_i) = 1$. The summarization results of two probabilistic

Table 4: The results obtained by using HMM-RM and WSME with different initial probability distributions $P_0(S_i)$.

	F1	F2	F3	F4	F5
10%	0.4907	0.3196	0.3760	0.3736	0.3763
20%	0.4749	0.3194	0.3757	0.3717	0.3759
30%	0.4641	0.3441	0.3814	0.3761	0.3769
50%	0.4822	0.4581	0.4822	0.4741	0.4757

Table 5: The results obtained by using WTMM and WSME with different initial probability distributions $P_0(S_i)$.

	F1	F2	F3	F4	F5
10%	0.4615	0.3464	0.3574	0.3526	0.3451
20%	0.4445	0.3344	0.3580	0.3519	0.3457
30%	0.4252	0.3581	0.3838	0.3816	0.3816
50%	0.4723	0.4608	0.4615	0.4631	0.4618

generative approaches, HMM-RM and WTMM, integrated with different kinds of sentence prior probabilities under different summarization ratios are shown in Tables 2 and 3, respectively. Comparing Tables 2 and 3 with Table 1, it can be found that the performance of both approaches is significantly boosted by utilizing F1. For HMM-RM, modeling the sentence prior using F5 can achieve slightly better summarization results when the summarization ratio is set to 10%, while for WTMM, using F3 and F5 to model the sentence prior can also achieve slightly better summarization results under summarization ratios of 10% and 20%. Tables 2 and 3 show that consistently better results at lower summarization ratios can be obtained if F1 or F5 was used to model the sentence prior.

4.3. Fusion of Extra Sentence Features using WSME

Different extra sentence features might bring different information about the sentences of the spoken document to be summarized. We further tried to integrate these features together into our proposed probabilistic generative framework, and therefore WSME was used as the vehicle for the fusion of these features. In order to train WSME, the manually transcribed document-summary pairs (20% summarization ratio) of the development set were used. All five features are used to train WSME, and one of them was taken as the initial sentence prior distribution $P_0(S_i)$, while the others for tuning the initial distribution. The summarization results on the evaluation set obtained by using different sentence generative models and WSME with different initial probability distributions $P_0(S_i)$ are shown in Tables 4 and 5, where each column denotes the source of the initial probability adopted in the WSME training. Compared to the results in Tables 2 and 3, we found that the fusion of five features can achieve slightly better performance than a single feature in most cases. We believe that a tight combination of these extra sentence features is necessary and WSME can satisfy this purpose.

5. Conclusions

We have presented a unified probabilistic generative framework for extractive spoken document summarization that successfully combined the sentence generative probability and the sentence prior probability for sentence ranking. Two different matching strategies, i.e., literal term matching and concept matching, have been extensively investigated for modeling the sentence generative probability, while a set of extra features extracted from the spoken sentences have been exploited for modeling the sentence prior probability. The experiments have demonstrated promising results on the broadcast news summarization task.

6. Acknowledgements

This work was supported in part by the National Science Council, Taiwan, under Grants: NSC95-2221-E-003-014-MY3 and NSC95-2422-H-001-031. The authors also would like to thank the NTU Speech Processing Lab for providing the necessary speech and language data.

7. References

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

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

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

[4] Y. Ho, An initial study on automatic summarization of Chinese spoken documents, Master Thesis, National Taiwan University, July 2003.

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

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

[9] X. Zhu and 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] H. S. Chiu and B. Chen, "Word Topical Mixture Models for Dynamic Language Model Adaptation," in *Proc. ICASSP 2007*.

[13] B. Chen and Y.T. Chen, "Word Topical Mixture Models for Extractive Spoken Document Summarization," in *Proc. ICME 2007*.

[14] R. Brandow et al, "Automatic Condensation of Electronic Publications by Sentence Selection," *Information Processing & Management*, 31(5), 1995.

[15] R. Rosenfeld, S. F. Chen, and X. Zhu, "Whole-sentence exponential language models: A vehicle for linguistic-statistical integration," *Computer Speech and Language* 15(1), 2001.

[16] O. Chan and R. Togneri, "Prosodic Features for a Maximum Entropy Language Model", in *Proc. ICSLP 2006*.

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

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