A COMPARATIVE STUDY OF PROBABILISTIC RANKING MODELS FOR SPOKEN DOCUMENT SUMMARIZATION

Shih-Hsiang Lin[1], Yi-Ting Chen[1,2], Hsin-Min Wang[2], Berlin Chen[1]

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

ABSTRACT

The purpose of extractive document 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 present a comparative study of various supervised and unsupervised probabilistic ranking models for spoken document summarization on the Chinese broadcast news. Moreover, we also investigate the possibility of using unsupervised summarizers to boost the performance of supervised summarizers when manual labels are not available for the training of supervised summarizers. Encouraging results were initially demonstrated.

Index Terms **—** spoken document summarization, extractive summarization, probabilistic ranking models, unsupervised summarizers

1. INTRODUCTION

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 [1-11]. Generally, spoken document summarization can be either extractive or abstractive. In this paper, we focus on extractive spoken document summarization. Extractive spoken document summarization may roughly 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 sentence classification. In [2, 3], the authors suggested that important sentences can be selected from the significant parts of a document, e.g., sentences can be selected form the introductory and/or concluding parts. However, such approaches can be only applied to some specific domains or document structures. On the other hand, 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, language model probability, 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, 4], latent semantic analysis (LSA) method [4], maximum marginal relevance (MMR) method [5], sentence significant score method [6] are the most popular for spoken document summarization. Besides, a bulk of classification-based methods using statistical features and/or sentence structure (or position) information also have been developed, such as the Gaussian mixture models (GMM) [5], hidden Markov models (HMM) [7-8], Bayesian classifier (BC) [9], support vector machine (SVM) [10], conditional random fields (CRFs) [11]. In these methods, sentence selection is usually formulated as a binary classification problem. A sentence can either be included in a summary or not. These methods need a set of training documents together with their corresponding handcrafted summaries (or labeled data) for training the classifiers. However, manual labeling is expensive in terms of time and personnel. In order to overcome this shortcoming, we have proposed a probabilistic generative framework for spoken document summarization, which performed the summarization task in a purely unsupervised manner [12-13]. In such a framework, each sentence of a spoken document to be summarized is treated as a probabilistic generative model for generating the document, and sentences are ranked and selected according to their likelihoods.

In this paper, we present a comparative study of various probabilistic ranking models for spoken document summarization including supervised classifier-based approaches and unsupervised probabilistic generative approaches. Moreover, we investigate the possibility of using unsupervised summarizers to boost the performance of supervised summarizers when manual labeling is not available for model training. The remainder of this paper is organized as follows. Section 2 describes three popular supervised classifiers used in this paper for document summarization, namely, the Bayesian classifier, support vector machine classifier and conditional random fields. Section 3 elucidates the theoretical foundations of the probabilistic generative framework. Then, the experimental settings and a series of summarization results are presented in Sections 4 and 5. Finally, conclusions are drawn in Section 6.

2. SUPERVISED SUMMARIZERS

Extractive spoken document summarization can be treated as a two-class classification problem. Each sentence S_i with a set of M representing features $X_i = \{x_{i1}, \cdots, x_{im}, \cdots, x_{iM}\}$ is being fed into the classifier and will be selected as a part of summary if it belongs to the positive class. On the contrary, it will be excluded from the summary if it belongs to the negative class. By doing so, quite a few popular classifiers can be utilized for this purpose. In this paper, we exploit three different classification-based classifiers for spoken document summarization, including the Bayesian classifier, support vector machine classifier and conditional random fields. In order to summarize the document in different summary ratios, importance sentences S_i of a spoken document D can be selected (or ranked) based on the posterior probability of the sentence being included in the summary $\boldsymbol{S}$.

2.1. Bayesian Classifier (BC)

BC is a simple but powerful supervised classification technique based on Bayes' theorem. The posterior probability of a sentence S_i being in the summary class $\boldsymbol{S}$ can be computed as follows [9]:

$$P(S_i \in \boldsymbol{S} \mid X_i) = \frac{p(X_i \mid S_i \in \boldsymbol{S}) P(S_i \in \boldsymbol{S})}{P(X_i)}. \quad (1)$$

where the evidence $P(X_i)$ is the marginal probability that the set of the representing features of a sentence is seen, regardless of whether it belongs to the summary class or the non-summary class. The evidence $P(X_i)$ can be further expressed as follows:

$$P(X_i) = P(S_i \in \mathbf{S} \mid X_i) P(S_i \in \mathbf{S}) + P(S_i \in \widetilde{\mathbf{S}} \mid X_i) P(S_i \in \widetilde{\mathbf{S}}), \quad (2)$$

where $P(X_i \mid S_i \in \mathbf{S})$ and $P(X_i \mid S_i \in \widetilde{\mathbf{S}})$ are the likelihood of X_i generated by the summary class and the non-summary class, respectively; and the prior probability of S_i belonging to the summary class $P(S_i \in \mathbf{S})$ or the non-summary class $P(S_i \in \widetilde{\mathbf{S}})$ is set to equal in this paper.

2.2. Support Vector Machine (SVM)

A SVM classifier is based on the principle of structural risk minimization (SRM) in the statistical learning theory. If the dataset is linear separable, SVM attempts to find an optimal hyper-plane by utilizing a decision function that can correctly separate the positive and negative samples and it also ensures the margin is maximal. In the nonlinear separable case, SVM uses kernel functions or defines slack variables to transform the problem into a linear discrimination problem. In this paper, we use LIBSVM [14] for constructing the binary SVM classifier, where the radial basis function (RBF) is chosen as the kernel function. The posterior probability of a sentence S_i being in the summary class can be approximated by a sigmoid operation [15]:

$$P(S_i \in \boldsymbol{S} \mid X_i) \approx \frac{1}{1 + \exp(Af(X_i) + B)}, \quad (3)$$

where A and B are estimated from the training data by minimizing a negative log-likelihood function, and $f(X_i)$ is the decision value of X_i provided by the SVM classifier.

2.3. Conditional Random Fields (CRFs)

Thought BC and SVM have shown their effectiveness in many classification problems, one of the main defects of them however is the bag-of-instances assumption (or the bag-of-sentences assumption in the paper). In more precise terms, they classify each instance independently without considering the relationship among instances. In contrast, CRFs can effectively capture the dependence relationship among instances. Therefore, we also investigate CRFs here for the summarization task. CRFs are undirected discriminative graphical models that combine the merits of maximum entropy Markov model (MEMM) and hidden Markov models (HMM) to calculate the probability of a state sequence $\boldsymbol{Y} = \{Y_1, \ldots, Y_i, \ldots, Y_I\}$ globally conditioned on the entire instance sequence $\boldsymbol{X} = \{X_1, \ldots, X_i, \ldots, X_I\}$ [16]:

$$p(\boldsymbol{Y} \mid \boldsymbol{X}) = \frac{1}{Z_{\boldsymbol{X}}} \exp\left(\sum_{i=1}^{I} \sum_{k} \lambda_k f_k(y_{i-1}, y_i, X_i) + \sum_{i=1}^{I} \sum_{l} \mu_l g_l(y_i, X_i) \right), \quad (4)$$

where $Z_{\boldsymbol{X}}$ is a normalization factor which is computed by summing up over all possible state sequences to ensure the probability of all state sequences being summed to one; I is the number of sentences in a document D; $f_k(y_{i-1}, y_i, X_i)$ captures co-occurrences between sentences i and $i-1$; $g_l(y_i, X_i)$ captures the co-occurrences between the sentence S_i to be selected as a summary sentence (or a non-summary sentence) and it associated observation features X_i; λ_k and μ_l are the weights for each feature function that are learnt from the training data. In this paper, we adapt a liner-chain CRFs to summarize a spoken document and we simply apply the forward-backward algorithm to rank the posterior probability of each sentence S_i being chosen as a summary sentence given the whole sentence sequences.

3. UNSUPERVISED SUMMARIZERS

We also address the issue of extractive summarization under an unsupervised probabilistic generative framework [12-13]. Each sentence S_i of a spoken document D to be summarized is treated as a probabilistic generative model for generating the document, and sentences are ranked and selected according to the posterior probabilities $P(S_i \mid D)$ of the sentences, which can be expressed as:

$$P(S_i|D) = \frac{P(D|S_i) P(S_i)}{P(D)}, \quad (5)$$

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. (5) can be eliminated because it is identical for all sentences and will not affect the ranking of them. Furthermore, since there still has great difficulty to estimate the sentence prior probability $P(S_i)$, we may simply assume that $P(S_i)$ is uniformly distributed. The sentence generative probability $P(D|S_i)$ can be taken as a relevance measure between the document and sentences. Therefore, the sentences of the spoken document to be summarized can be ranked by the sentence generative probability $P(D|S_i)$.

3.1. Language Modeling Approach (LM)

In the language modeling approach, each sentence of a document to be summarized was treated as a probabilistic generative model consisting of *N*-gram distributions for predicting the document [12], 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 investigated for the LM approach:

$$P_{LM}(D|S_i) = \prod_{w_n \in D} \left[\lambda \cdot P(w_n|S_i) + (1-\lambda) P(w_n|Corpus) \right]^{c(w_n, D)}, \quad (6)$$

where λ is a weighting parameter and $c(w, D)$ is the occurrence count of a term w in D. The sentence model $P(w \mid S_i)$ and the collection model $P(w \mid C)$ 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. (6) can be further optimized by using the expectation-maximum (EM) training algorithm [12].

3.2. Topical Mixture Model (TMM)

In the topical mixture model, a set of K latent topical distributions characterized by unigram language models are used to predict the document terms, and each of the latent topics is associated with a sentence-specific weight. The sentence generative probability therefore can be expressed as [13]:

$$P_{TMM}(D|S_i) = \prod_{w_n \in D} \left[\sum_{k=1}^{K} P(w_n|T_k) P(T_k|S_i) \right]^{c(w_n, D)}, \quad (7)$$

where $P(w_n|T_k)$ and $P(T_k|S_i)$, 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 sentence S_i. More precisely, the topical unigram distributions, e.g., $P(w_n|T_k)$, are tied among the sentences, which can be estimated by using a set of contemporary (or in-domain) text news collection and then by maximizing the collection likelihood. On the other hand, each sentence S_i of the spoken document to be summarized has its own probability distribution over the latent topics, e.g., $P(T_k|S_i)$, which can be estimated on the fly [13].

4. EXPERIMENTAL SETUP

4.1. Speech and Text Corpora

The speech data set consists of about 200 hours of MATBN Mandarin broadcast news, which were collected by Academia Sinica and Public Television Service Foundation of Taiwan [17]. From them, a set of 205 documents (7.5 hours), collected during the period of November 2001 to August 2002 and covering a wide range of topics, was reserved for the document summarization experiments. Each document contains approximately 600 words. The remainder of the speech data was used to train an acoustic model for speech recognition. The Chinese character error rate (CER) for the 205 documents reserved for the summarization experiments was 30.30%. Furthermore, 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. 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 are used for estimating the parameters used in Eqs. (6) and (7).

4.2. Features for Supervised Summarizers

Quite a few features have been designed and widely used in the supervised summarization approaches [10-11]. In this paper, we use a set of 19 features to characterize a spoken sentence, including the structure features, the lexical features, the acoustic features, and the relevance features. We use the sentence location and the sentence length features, as well as those of its preceding and following sentences as the structure features. Normalized bigram-based language model scores, similarity scores between a sentence and its preceding/following neighbors, and number of name entities in a sentence are taken as the lexical features. The acoustic features consist of confidence scores and min/max/mean/difference values of F0 and energy features. The relevance features are the similarity scores between a sentence and the whole document, which were obtained by using the LSA and VSM method [4]. Furthermore, each feature x_m is normalized by using the following equation:

$$\hat{x}_m = \frac{x_m - \mu_m}{\sigma_m} \tag{8}$$

where μ_m and σ_m are the mean and standard deviation of the m-th feature.

4.3. Evaluation Metric

Three human subjects were instructed to do human summarization on the 205 broadcast news documents, to be taken as the reference for development and evaluation. These spoken documents were divided into two parts: the first part consisting of 100 documents is taken as the development set, while the remaining part consisting of 105 documents as the held-out test set. The supervised summarizers are trained with the development set, and then evaluated on the test set, while the unsupervised summarizers are directly evaluated on the test set. The *ROUGE* measure [18] is used for performance evaluation. 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:

$$ROUGE-N = \frac{\sum_{S \in \mathbf{S}_R} \sum_{\text{gram}_N \in \mathbf{S}} Count_{match}(gram_N)}{\sum_{S \in \mathbf{S}_R} \sum_{\text{gram}_N \in \mathbf{S}} Count(gram_N)}, \tag{9}$$

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. The levels of agreement on the *ROUGE-2* measure between the three subjects are about 0.64, 0.66, 0.68 and 0.71 respectively, for summarization ratios of 10%, 20%, 30% and 50%.

	Basic Features			Complex Features		
	BC	SVM	CRFs	BC	SVM	CRFs
10%	0.3209	0.3327	0.3456	0.3283	0.3445	0.3507
20%	0.3307	0.3631	0.3710	0.3340	0.3669	0.3716
30%	0.3166	0.3526	0.3637	0.3262	0.3529	0.3686
50%	0.3374	0.3482	0.3646	0.3390	0.3511	0.3846

Table 1: The results achieved by supervised summarizers under different summarization ratios.

5. EXPERIMENTAL RESULTS AND DISSCUSSIONS

We first evaluate the summarization performance of the three different supervised summarizers. The associated results are shown in Table 1, where each column illustrates the *ROUGE-2* recall rates for different supervised summarizers at different summarization ratios. It is worth mentioning that the amount of labels used for training a summarizer is in accordance with the target summarization ratio it wants to achieve in the evaluation. More specifically, the summarizers trained with the manual summaries at a given summarization ratio will be also tested at the same summarization ratio. As can been seen from the left part of Table 1, the discriminative summarizers (CRFs and SVM) outperform the generative summarizer (BC). Moreover, the performance of CRFs is considerably better than SVM, which may be probably explained by the fact that CRFs have the ability to model the relationship among sentences.

In the next set of experiments, we attempt to augment the basic features with two additional generative scores obtained from the LM and TMM approaches, as those defined in Eqs. (6) and (7), respectively, to improve the performance of the supervised summarizers. As evidenced by the results shown in the right part of Table 1, additionally incorporating the generative scores obtained from LM and TMM indeed boosts the performance of the supervised summarizers. Therefore, how to effectively define or

	BC	SVM	CRFs	VSM	LSA	MMR	SIG	LM	TMM	RND
10%	0.3283	0.3445	0.3507	0.2044	0.1866	0.2037	0.1790	0.2008	0.2110	0.1626
20%	0.3340	0.3669	0.3716	0.2385	0.2398	0.2412	0.2129	0.2501	0.2618	0.2231
30%	0.3262	0.3529	0.3686	0.2818	0.2762	0.2801	0.2475	0.2820	0.2861	0.2302
50%	0.3390	0.3511	0.3846	0.3660	0.3519	0.3590	0.3102	0.3622	0.3659	0.2138

Table 2: The results achieved by different summarizers under different summarization ratios.

	LM Labeling			TMM Labeling			Random Labeling		
	BC	SVM	CRFs	BC	SVM	CRFs	BC	SVM	CRFs
10%	0.1372	0.2004	0.2031	0.1342	0.1652	0.1935	0.0977	0.1186	0.1194
20%	0.2103	0.2397	0.2375	0.2134	0.2530	0.2619	0.1440	0.1516	0.1364
30%	0.2740	0.2796	0.2793	0.2761	0.2906	0.2959	0.1717	0.1811	0.1745
50%	0.3523	0.3543	0.3512	0.3390	0.3511	0.3846	0.3228	0.3102	0.3256

Table 3: The results achieved by different supervised summarizers trained in an unsupervised manner.

select additional salient features to improve the performance of supervised summarizers might be an important research issue.

In the third set of experiments, we compare the performance of the supervised summarizers with those of the unsupervised summarizers, including VSM [1], MMR [5], LSA [4], and sentence significance score (SIG) [6], as well as our previously proposed LM and TMM models. The results for these unsupervised summarizers are shown in Table 2, where the results for the supervised summarizers are directly copied from the right part of Table 1 and the results obtained by random selection (RND) are also listed for comparison. It can be found that the supervised summarizers significantly outperform all the unsupervised ones. Although LM and TMM do not perform better than the supervised summarizers, they both have competitive performance as compared with the other unsupervised summarizers, especially at lower summarization ratios.

Finally, we investigate the possibility of using unsupervised summarizers to boost the performance of supervised summarizers when manual labels are not available for the training of supervised summarizers. Table 3 shows the results of the supervised summarizers trained with the labels provided by different unsupervised summarizers and random selection. As the results indicated, CRFs trained with TMM labeling in most cases can achieve slightly better performance than that obtained using TMM alone. Therefore, how to filter out the unreliable labels or to collect more reliable labels for training the summarizer might be an important issue for further studies. We believe that this initial attempt opens a new direction for future research on spoken document summarization.

6. CONCLUSIONS

In this paper, we have studied the use of probabilistic ranking models for extractive spoken document summarization. Various kinds of modeling and learning approaches have been extensively investigated. As shown by the experimental results, CRFs can lead to significant performance improvements as compared to the other summarizers. In addition, we have also pointed out a possible future research direction for training supervised classifiers without manual labels. Encouraging results were initially demonstrated.

7. REFERENCES

[1] L.S. Lee, 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*, 1958.

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

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

[5] G. Murray et al., "Extractive summarization of meeting recordings," *in Proc. EUROSPEECH 2005*.

[6] S. Furui et al. "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Trans. on Speech Audio Processing* 12 (4), 2004.

[7] J. Conroy and D. P. O'Leary, "Text summarization via hidden Markov models, " *Research and Development in Information Retrieval,* 2001.

[8] S. Maskey and J. Hischberg, "Summarizing speech without text using hidden Markov models," *in Proc. HLT-NAACL* 2006.

[9] J. Kupiec et al., "A trainable Document Summarizer," *in Proc. ACM SIGIR* 1999.

[10] J. Zhang et al. "A Comparative Study on Speech Summarization of Broadcast News and Lecture Speech," *in Proc. EUROSPEECH 2007*.

[11] D. Shen et al., "Document Summarization using Conditional Random Fields," *in Proc. IJCAI 2007*.

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

[13] Y. T. Chen et al., "A Unified Probabilistic Generative Framework for Extractive Spoken Document Summarization", *in Proc. EUROSPEECH 2007*.

[14] C. C. Chang et al., "LIBSVM : a library for support vector machines," Software available at http://www.csie.ntu.edu.tw/ ~cjlin/libsvm, 2001.

[15] H. T. Lin et al., "A note on Platt's probabilistic outputs for support vector machines," *Machine Learning* 68, 2007.

[16] J. Lafferty et al, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", *in Proc. ICML 2001*.

[17] H. M. Wang et al., "MATBN: A Mandarin Chinese Broadcast News Corpus," *International Journal of Computational Linguistics & Chinese Language Processing* 10(1), 2005.

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