A Comparative Study of Probabilistic Ranking Models for Chinese Spoken Document Summarization

SHIH-HSIANG LIN and BERLIN CHEN
National Taiwan Normal University
and
HSIN-MIN WANG
Institute of Information Science, Academia Sinica

Extractive document summarization automatically selects a number of indicative sentences, passages, or paragraphs from an original document according to a target summarization ratio, and sequences them to form a concise summary. In this article, we present a comparative study of various probabilistic ranking models for spoken document summarization, including supervised classification-based summarizers and unsupervised probabilistic generative summarizers. We also investigate the use of unsupervised summarizers to improve the performance of supervised summarizers when manual labels are not available for training the latter. A novel training data selection approach that leverages the relevance information of spoken sentences to select reliable document-summary pairs derived by the probabilistic generative summarizers is explored for training the classification-based summarizers. Encouraging initial results on Mandarin Chinese broadcast news data are demonstrated.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms: Algorithms, Performance, Theory

Additional Key Words and Phrases: Spoken document summarization, extractive summarization, probabilistic ranking models, relevance information

ACM Reference Format:
Lin S.-H., Chen, B., and Wang, H.-M. 2009. A comparative study of probabilistic ranking models for Chinese spoken document summarization. ACM Trans. Asian Lang. Inform. Process. 8, 1, Article 3 (March 2009), 23 pages. DOI = 10.1145/1482343.1482346.
http://doi.acm.org/10.1145/1482343.1482346.

This work was supported in part by the National Science Council of Taiwan, under Grants. NSC 96-2628-E-003-015-MY3, NSC 97-2631-S-003-003, and NSC 96-3113-H-001-012.
Corresponding author's address: B. Chen, Department of Computer Science and Information Engineering, National Taiwan Normal University, No. 88, Sec. 4, Ting-Chow Rd., Taipei 116, Taiwan, R.O.C.; email: berlin@csie.ntnu.edu.tw.

1. INTRODUCTION

Because of the rapid development and maturity of multimedia technologies, huge quantities of multimedia content, including audio and video materials, are already available and databases continue to expand. In recent years, a great deal of research has focused on using speech content to understand and organize multimedia data [Lee and Chen 2005; Koumpis and Renals 2005; Gilbert and Feng 2008; Chelba et al. 2008; Christensen et al. 2008]. Spoken document summarization, which tries to distill salient information and remove redundant or incorrect information from spoken documents, can help users review spoken documents efficiently and understand associated topics quickly. Generally speaking, spoken document summarization techniques can be classified as either extractive or abstractive. Extractive summarization produces a summary by selecting indicative sentences, passages, or paragraphs from an original document according to a predefined target summarization ratio. Abstractive summarization, on the other hand, provides a fluent and concise abstract of a certain length that reflects the key concepts of the document [Paice 1990; Witbrock and Mittal 1999]. This requires highly sophisticated techniques, including semantic representation and inference, as well as natural language generation. Thus, in recent years, researchers have tended to focus on extractive summarization.

Extractive spoken document summarization methods generally fall into three broad categories: 1) approaches based on sentence structure or location information, 2) approaches based on proximity or significance measures, and 3) approaches based on sentence classification. Baxendale [1958] and Hirohata et al. [2005] suggested that important sentences can be selected from the significant parts of a document, for example, the introduction and the conclusion. However, such approaches can only be applied to documents in some specific domains or structures. In contrast, approaches based on proximity or significance measures attempt to select salient sentences based on the statistical features of the sentences or the words in the sentences, such as the term frequency (TF), inverse document frequency (IDF), *N*-gram scores, and the topic or concept information. Associated methods based on these features have attracted much attention in recent years. For example, the vector space model (VSM) [Gong and Liu 2001; Lee and Chen 2005] and the maximum marginal relevance (MMR) method [Murray et al. 2005] represent the whole document and each of its sentences in vector form consisting of statistical features, and then select important sentences based on the proximity measure between the vector representations of the document and its sentences; the latent semantic analysis (LSA) method [Gong and Liu 2001] estimates the significance of a sentence by projecting the vector representation of the sentence into the latent semantic space of the document; and the sentence significance score method (SIG) [Furui et al. 2004] estimates the significance of a sentence by linearly combining a set of statistical features of the sentence. In addition, a number of classification-based methods that use statistical features and/or sentence structure information have been developed. Examples of such methods include the Bayesian classifier (BC) [Kupiec et al. 1999], the Gaussian

mixture model (GMM) [Murray et al. 2005], the hidden Markov model (HMM) [Conroy and O'Leary 2001; Maskey and Hirschberg 2006], the support vector machine (SVM) [Zhang et al. 2007], and the conditional random fields (CRF) [Shen et al. 2007]. These methods usually formulate sentence selection as a binary classification problem, that is, a sentence can be included in or excluded from a summary. A training set, comprised of documents and their corresponding handcrafted summaries (or labeled data), is needed to train the classifiers. However, manual labeling is expensive in terms of time and personnel. To overcome this shortcoming, we proposed a probabilistic generative framework that can perform spoken document summarization tasks in a completely unsupervised manner [Chen et al. 2006a; Chen and Chen 2008]. The framework treats each sentence of a spoken document to be summarized as a probabilistic generative model for generating the document, and ranks and selects sentences according to their likelihoods.

In this article, we present a comparative study of various probabilistic ranking models for extractive spoken document summarization, including supervised classification-based summarizers and unsupervised probabilistic generative summarizers. Furthermore, we also investigate the use of unsupervised summarizers to improve the performance of supervised summarizers when manual labels are not available for training the latter. A novel training data selection approach that leverages the relevance information of spoken sentences to select reliable document-summary pairs derived by the unsupervised summarizers is explored for training the supervised summarizers.

The remainder of the article is organized as follows. Section 2 briefly describes three popular supervised classification-based methods for spoken document summarization, namely, the Bayesian classifier, the support vector machine, and the conditional random fields. Section 3 sheds light on the theoretical foundations of two unsupervised probabilistic generative summarizers, namely, the language modeling approach and the sentence topic model. Then we elucidate the training data selection approach for unsupervised training of the classification-based summarizers in Section 4. The experimental setup and a series of summarization experiments are presented in Sections 5 and 6, respectively. We then draw our conclusions and directions for future work in Section 7.

2. SUPERVISED CLASSIFICATION-BASED SUMMARIZERS

Extractive spoken document summarization can be treated as a two-class (positive/negative) classification problem. A sentence S_i with a set of J representative features $X_i = \{x_{i1}, \cdots, x_{ij}, \cdots, x_{iJ}\}$ is input to the classifier. If it belongs to the positive class, it will be selected as part of the summary; otherwise, it will be excluded. Several popular classifiers can be used for this purpose. In this article, we exploit three such classifiers, namely the Bayesian classifier, the support vector machine, and the conditional random fields. To summarize spoken documents with different summary ratios, the important sentences of a document D can be selected (or ranked) based on $P(S_i \in \mathbf{S}|X_i)$,

the posterior probability of a sentence S_i being included in the summary $\mathbf{S}$ given the feature set X_i.

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 included in the summary class $\mathbf{S}$ can be computed as follows [Kupiec et al. 1999]:

$$P\left(S_i \in \mathbf{S}|X_i\right) = \frac{p\left(X_i|S_i \in \mathbf{S}\right) P\left(S_i \in \mathbf{S}\right)}{P\left(X_i\right)}, \tag{1}$$

where the evidence $P(X_i)$ is the marginal probability that the set of representative features of a sentence is observed, regardless of whether the sentence belongs to the summary (positive) class or the nonsummary (negative) class. The evidence $P(X_i)$ can be expressed as:

$$P(X_i) = P\left(X_i|S_i \in \mathbf{S}\right) P\left(S_i \in \mathbf{S}\right) + P\left(X_i|S_i \in \tilde{\mathbf{S}}\right) P\left(S_i \in \tilde{\mathbf{S}}\right), \tag{2}$$

where $P\left(X_i|S_i \in \mathbf{S}\right)$ and $P\left(X_i|S_i \in \tilde{\mathbf{S}}\right)$ are the likelihoods that X_i is generated by the summary class and the nonsummary class, respectively. In this article, the prior probability of S_i belonging to the summary class $P\left(S_i \in \mathbf{S}\right)$ or the nonsummary class $P\left(S_i \in \tilde{\mathbf{S}}\right)$ is set to be equal.

2.2 Support Vector Machine (SVM)

The concept of SVM is based on the principle of structural risk minimization (SRM) in statistical learning theory [Vapnik 1998]. If a dataset is linearly separable, SVM attempts to find an optimal hyper-plane by utilizing a decision function that can correctly separate the positive and negative instances, and ensure that the margin is maximal. In a nonlinearly separable case, SVM uses kernel functions or defines slack variables to transform the problem into a linear discrimination problem. In this article, we use the LIBSVM toolkit [Chang and Lin 2001] to construct a binary SVM summarizer and adopt the radial basis function (RBF) as the kernel function. The posterior probability of a sentence S_i being included in the summary class can be approximated by the following sigmoid operation [Lin et al. 2003]:

$$P\left(S_i \in \mathbf{S}|X_i\right) \approx \frac{1}{1 + exp\left(\alpha \cdot g\left(X_i\right) + \beta\right)}, \tag{3}$$

where the weights α and β are estimated from the development set by minimizing a negative log-likelihood function, and $g(X_i)$ is the decision value of X_i provided by the SVM summarizer.

2.3 Conditional Random Fields (CRF)

Although BC and SVM have proved effective for many classification problems, the bag-of-instances assumption (or the bag-of-sentences assumption when applied in extractive spoken document summarization) is a major shortcoming. More precisely, BC and SVM classify each sentence independently

without considering the dependent relationships among sentences. The CRF model, on the other hand, can effectively capture the dependent relationships among sentences. It is an undirected discriminative graphical model that combines the advantages of the maximum entropy Markov model (MEMM) and the hidden Markov model (HMM). The probability of a state sequence $\mathbf{Y} = \{y_1, \ldots, y_i, \ldots, y_I\}$ globally conditioned on the entire instance (or sentence) sequence $\mathbf{X} = \{X_1, \ldots, X_i, \ldots, X_I\}$ [Lafferty et al. 2001] is computed by

$$p(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z_{\mathbf{X}}} \exp\left(\sum_{i=1}^{I} \sum_{k} \lambda_k f_k(y_i, X_i)\right), \tag{4}$$

where each y_i can be a summary or nonsummary state; $Z_{\mathbf{X}}$ is a normalization factor computed by summing all possible state sequences to ensure that the summation of the probabilities of all state sequences is equal to one; I is the number of sentences in a document D; $f_k(y_i, X_i)$ is a function that measures a feature relating the state y_i for sentence i with the input features X_i; and λ_k is the weight of each feature function learned from the development set. In this article, we adopt a linear-chain CRF model to summarize a spoken document, and simply apply the forward-backward algorithm to obtain the posterior probability of each sentence S_i being a summary sentence given the whole sentence sequence for sentence ranking.

3. UNSUPERVISED PROBABILISTIC GENERATIVE SUMMARIZERS

In our recent works, we addressed the issue of extractive summarization under an unsupervised probabilistic generative framework [Chen et al. 2006a; Chen and Chen 2008]. Each sentence S_i of a spoken document D is treated as a probabilistic generative model for generating the document, and the sentences are ranked and selected according to their posterior probabilities $P(S_i|D)$, which can be expressed as:

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

where $P(D|S_i)$ is the sentence generative probability, that is, the likelihood that D is generated by S_i; $P(S_i)$ is the prior probability of S_i being important; and $P(D)$ is the probability of D. Note that $P(D)$ in Equation (5) can be eliminated because it is identical for all sentences; hence, it will not affect their ranking. Furthermore, since the way to estimate the prior probability of a sentence is still under active study, we assume that the prior probability is uniformly distributed in this article. The sentence generative probability $P(D|S_i)$ can be taken as a relevance measure between the document and its sentences. Therefore, the sentences of a spoken document D can be ranked by means of the probability $P(D|S_i)$, instead of the probability $P(S_i|D)$.

3.1 Language Modeling Approach (LM)

In the language modeling approach, each sentence in a document is regarded as a probabilistic generative model consisting of N-gram distributions for predicting the document [Chen et al. 2006a]. The N-gram distributions are

directly estimated from each sentence and smoothed by the N-gram distributions estimated from a large text corpus. In this article, we only investigate the unigram (bag-of-words) modeling for the LM approach:

$$P_{\mathrm{LM}}(D|S_i) = \prod_{w_n \in D} [\gamma \cdot P(w_n|S_i) + (1-\gamma) P(w_n|Collection)]^{c(w_n,D)}, \quad (6)$$

where γ is a weighting parameter, and $c(w_n, D)$ is the occurrence count of a specific type of word (or term) w_n in D, reflecting that w_n will contribute more in the calculation of $P_{LM}(D|S_i)$ if it occurs more frequently in D [Croft and Lafferty 2003]. The sentence model $P(w_n|S_i)$ and the collection model $P(w_n|Collection)$ are estimated from the sentence itself and a large general text collection (see Section 5.1), respectively, using maximum likelihood estimation (MLE). Note that the process defined in Equation (6) is similar to interpolation-based language model smoothing or adaptation for speech recognition [Zhai and Lafferty 2001; Bellegarda 2004]. The weighting parameter γ can be further optimized by using the expectation-maximization (EM) training algorithm [Dempster et al. 1977]. This relevance measure is computed based on the frequency that the document words occur in the sentence, which is actually a kind of literal term matching [Lee and Chen 2005].

3.2 Sentence Topic Model (STM)

In the sentence topic model, a set of K latent topical distributions characterized by unigram language models are used to predict document terms, and each latent topic is associated with a sentence-specific weight. In other words, each term can belong to several topics. Therefore, the sentence generative probability can be expressed as follows:

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

where $P(w_n|T_k)$ and $P(T_k|S_i)$ denote, respectively, the probability of a specific type of word w_n occurring in a latent topic T_k and the posterior probability (or weight) of the topic T_k conditioned on the sentence S_i. The topical unigram $P(w_n|T_k)$ is shared by all sentences and can be estimated by maximizing the collection likelihood on a set of contemporaneous (or in-domain) text news documents (see Section 5.1). In contrast, each sentence S_i of a spoken document has its own probability distribution $P(T_k|S_i)$ over the latent topics, which is unknown in advance but can be estimated on the fly during the summarization process by maximizing the log-likelihood of the document D generated by the STM model of the sentence S_i, using the EM training algorithm [Chen et al. 2006a].

Note that this relevance measure is not computed directly according to the frequency that the document words occur in the sentence. Instead, it is derived from the frequency of the document words in the latent topics as well as the likelihood that the sentence will generate the respective topics. Hence,

STM is actually a type of concept matching approach [Lee and Chen 2005]. In recent years, structures similar to the presented topic model have also been extensively investigated for information retrieval (IR) tasks [Hofmann 2001; Blei et al. 2003; Chen 2006].

4. UNSUPERVISED TRAINING OF CLASSIFICATION-BASED SUMMARIZERS

One major disadvantage of classification-based summarizers is the need for a certain amount of labeled data for model training. However, manual labeling of the document-reference summary information is time-consuming and impractical for many summarization tasks. The issue of semi-supervised (or purely unsupervised) training of various classification-based models by utilizing a small amount of labeled data together with a large amount of unlabeled data has attracted a great deal of interest in many speech and language processing tasks over the decade [Duda et al. 2001; Nomoto and Matsumoto 2001; Chen et al. 2004; Zhu 2005; Wong et al. 2008]. In this article, we investigate the use of unsupervised probabilistic generative summarizers to improve the performance of supervised classification-based summarizers under the condition that the handcrafted document-reference summary pairs are not available for training the latter. Nevertheless, as will be discussed in Section 6.3, it was experimentally observed that the performance of a classification-based summarizer (e.g., CRF) trained with the document-summary labels derived automatically by a probabilistic generative summarizer (e.g., STM) would be slightly worse than the original performance of the probabilistic generative summarizer. This observation seems to reflect the fact that there exists a significant performance gap between the results obtained by the classification-based summarizer trained with manually labeled data and automatically labeled data. Thus, how to filter out unreliable automatic labels or collect more reliable automatic labels for training the classification-based summarizers without supervision is deemed to be an important issue for reducing the performance gap.

To this end, this article presents an initial attempt to exploit a training data selection approach, which leverages the relevance information of the sentences in a training spoken document, to filter out unreliable summary or nonsummary sentence labels for training the classification-based summarizers without supervision. The relevance information of a spoken sentence S_i is defined by the average similarity of documents in the relevant text news document set $\mathbf{D}^i_{\text{top}M}$ of S_i, where $\mathbf{D}^i_{\text{top}M}$ is obtained by taking S_i as a query and posing it to a information retrieval (IR) system to obtained a list of M most relevant documents from a contemporaneous text news repository (see Section 5.1). Our assumption is that the relevant text documents retrieved for a summary sentence might have the same or similar topics because a summary sentence is usually indicative for some specific topic related to the document. In contrast, the relevant text documents retrieved for a nonsummary sentence might cover diverse topics. In other words, the relevance information estimated based on the similarity of documents in the relevant text document set might be a good indicator for determining the importance of a spoken sentence. Consequently, we can

select reliable summary or nonsummary sentences derived by the probabilistic generative summarizers for training the classification-based summarizers based on such relevance information. The average similarity of documents in the relevant text document set $\mathbf{D}^i_{\text{top}M}$ for a spoken sentence S_i is computed by

$$avgSim(S_i) = \frac{\sum_{D_l \in \mathbf{D}^i_{\text{top}M}} \sum_{\substack{D_u \in \mathbf{D}^i_{\text{top}M} \\ D_l \neq D_u}} \frac{\vec{D}_l \cdot \vec{D}_u}{\left\|\vec{D}_l\right\| \cdot \left\|\vec{D}_u\right\|}}{M \cdot (M-1)} \tag{8}$$

where $\vec{D}_l$ is the TF-IDF vector representation of the document D_l, and M is the number of documents in the retrieved relevant text document set $\mathbf{D}^i_{\text{top}M}$. The practical implementation issue of such data selection for unsupervised training of the classification-based summarizers will be further detailed in Section 6.3.

5. EXPERIMENTAL SETUP

5.1 Speech and Text Corpora

The speech data set used in this research is the MATBN corpus [Wang et al. 2005], which is different from the set of broadcast news documents used in our previous studies [Chen et al. 2006a; Chen and Chen 2008]. It contains approximately 200 hours of Mandarin Chinese TV broadcast news collected by Academia Sinica and the Public Television Service Foundation of Taiwan between November 2001 and April 2003. The content has been segmented into separate stories and transcribed manually. Each story contains the speech of one studio anchor person, as well as several field reporters and interviewees. A subset of 205 broadcast news documents (spoken documents that covered a wide range of topics) compiled between November 2001 and August 2002 was reserved for the summarization experiments. Twenty-five hours of gender-balanced speech from the remaining speech data were used to train the acoustic models for speech recognition. The data was first used to bootstrap the acoustic model training with the MLE criterion. Then, the acoustic models were further optimized by the minimum phone error (MPE) discriminative training algorithm [Povey and Woodland 2002; Liu et al. 2007].

A large number of text news documents collected by the Central News Agency (CNA) between 1991 and 2002 (the Chinese Gigaword Corpus released by LDC) were also used. The documents collected in 2000 and 2001 were used to train N-gram language models for speech recognition with the SRI Language Modeling Toolkit [Stolcke 2005]. In addition, a subset of about 14,000 text news documents, collected during the same period as the broadcast news documents to be summarized, was used to estimate the collection model $P(w_n|Collection)$ in Equation (6) for the LM approach and the topical unigram $P(w_n|T_k)$ in Equation (7) for the STM approach. This subset of text news documents was also used as the basis for estimating the model parameters for the VSM, LSA, MMR, and SIG approaches.

Table I. The Statistical Information of the Broadcast News Documents Used for the Summarization Experiments

	Development Set	Evaluation Set
Recording Period	November 07, 2001 – January 22, 2002	January 23, 2002 – August 22, 2002
Number of Documents	100	105
Average Duration per Document (in sec.)	129	135
Average Number of words per Document	326	340
Average Number of Chinese Characters per Document	578	599
Average Number of Sentences per Document	21	21

5.2 Automatic Broadcast News Transcription System

Front-end processing was implemented with the HLDA (Heteroscedastic Linear Discriminant Analysis)-based, data-driven Mel-frequency feature extraction approach [Kumar 1997], and processed by MLLT (Maximum Likelihood Linear Transformation) for feature decorrelation [Saon et al. 2000]. We then applied utterance-based feature mean subtraction and variance normalization.

The speech recognizer was implemented with a left-to-right frame-synchronous Viterbi tree-copy search and a lexical prefix tree organization of the lexicon [Aubert 2002]. To select the most promising path hypotheses for each speech frame, we used a beam pruning technique, which considers the decoding scores of the path hypotheses together with their corresponding unigram language model look-ahead scores and syllable-level acoustic model look-ahead scores [Chen et al. 2004]. If the scores of the word hypotheses at the end of each speech frame were higher than a predefined threshold, their associated decoding information, such as the words' start and end frames, the identities of the current and preceding words, and the acoustic score, were kept to build a word graph for further language model rescoring [Ortmanns et al. 1997]. 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.

5.3 Spoken Documents for the Summarization Experiments

Three subjects were asked to create summaries of the 205 spoken documents for the summarization experiments as references (the gold standard) for evaluation. The summaries were compiled by selecting 50% of the most important sentences in the reference transcript of a spoken document, and ranking them by importance without assigning a score to each sentence. The average Chinese character error rate (CER) obtained for the 205 spoken documents was about 30%. Detailed statistics of the 205 spoken documents are given in Table I.

We divided the 205 spoken documents into two parts. The first part, consisting of 100 documents, was taken as the development set, which formed the basis for tuning the parameters or settings. The second part, consisting of the remaining 105 documents, was taken as the held-out test set. That is, all the summarization experiments conducted on the test set followed the same training (or parameter) settings and model complexities that were

Table II. The Levels of Agreement on the ROUGE_2 Measure Between the Three Subjects for Important Sentence Ranking for the Test Set

	Summarization Ratio		
	10%	20%	30%
Agreement	0.646	0.668	0.684

optimized based on the development set.[1] Therefore, the experimental results can be used to estimate the effectiveness of the summarizers on comparable real-world data.

5.4 Performance Evaluation

For the performance evaluation, we used the ROUGE measure [Lin 2003]. It evaluates the quality of the summarization by counting the number of overlapping units, such as N-grams and word sequences, in the automatic summary and a set of reference (or manual) summaries. The ROUGE_N is an N-gram recall measure defined as follows:

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

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 cooccurring in the automatic summary and the reference summary; and $Count(gram_N)$ is the number of N-grams in the reference summary. Since ROUGE_N is a recall measure, increasing the summary length (or the summarization ratio) tends to increase the chances of getting higher scores. In this article, we adopt the widely used ROUGE_2 measure [Maskey and Hirschberg 2005; Murray et al. 2005; Shen et al. 2007], which uses word bigrams as the matching units.

The summarization results were evaluated by using several summarization ratios (10%, 20%, and 30%), defined as the ratio of the number of sentences in the automatic (or manual) summary to that in the reference transcript of a spoken document. As shown in Table II, the levels of agreement on the ROUGE_2 measure between the three subjects for important sentence ranking are about 0.65, 0.67, and 0.68 for the summarization ratios of 10%, 20%, and 30%, respectively. Each of these values was obtained by calculating the ROUGE_2 recall rate, using the summary created by one of the three subjects as the reference summary, in turn for each subject, while those of the other two subjects as the test summaries, and then taking their average. These observations seem to reflect the fact that people may not always agree with each other in selecting the important sentences for representing a given document.

[1]Note that the document-relevance summary information of the 100 spoken documents in the development set was not used for training the unsupervised summarizers.

Table III. The Features Used in the Supervised Summarizers

Structural features	*POSITION*: Sentence position *DURATION*: Duration of the preceding/current/following sentence
Lexical Features	*BIGRAM_SCORE*: Normalized bigram language model scores *SIMILARITY*: Similarity scores between a sentence and its preceding/following neighbor sentence *NUM_NAME_ENTITIES*: Number of named entities (NEs) in a sentence
Acoustic Features	*PITCH*: Min/max/mean/difference pitch values of a spoken sentence *ENERGY*: Min/max/mean/difference value of energy features of a spoken sentence *CONFIDENCE*: Posterior probabilities
Relevance Features	*R-VSM*: Relevance score obtained by using the VSM summarizer *R-LSA*: Relevance score obtained by using the LSA summarizer

5.5 Features for Supervised Summarizers

Several features have been designed and widely used in the supervised summarization approaches [Shen et al. 2007; Zhang et al. 2007; Wong et al. 2008]. In this article, we use a set of 19 features to characterize a spoken sentence, including the structural features, the lexical features, the acoustic features, and the relevance features. The features are outlined in Table III. The structural features (or the so-called surface features) basically stem from two kinds of information: *POSITION* and *DURATION*. The structure of a broadcast news story is usually rather regular, and may include an introductory remark, event description, conclusion, and a footnote by the reporter. Obviously, the summary often appears in the introduction and conclusion. Therefore, if a spoken document contains K sentences, the position feature of the j-th sentence in the document is defined as $|j - (K/2)|$. The duration features of a sentence consist of the information about the duration of the sentence itself, as well as the duration of the preceding and subsequent sentences [Shen et al. 2007]. The lexical features (or the so-called content features) we consider are *BIGRAM_SCORE*, *SIMILARITY*, and *NUM_NAME_ENTITIES*. The bigram language model score is computed as the product of the bigram probabilities of words occurring in a sentence, and then normalized by the sentence length (number of words). The similarity score between a sentence and its neighboring sentences is computed using the cosine measure. To do this, we represent each sentence in vector form, where each dimension specifies a weighted statistic, that is, the product of the term frequency (TF) and the inverse document frequency (IDF) associated with an indexing term (or word) in the sentence [Baeza-Yates and Ribeiro-Neto 1999]. We only use the similarity of a sentence to those that immediately precede or follow it. The number of named entities (NEs) in a sentence is also taken as a predictor of the lexical cues [Lee and Chen 2005], which is based on the idea that a sentence containing more NEs can very often give an overview of the spoken documents and thus is more likely to be included in the summary [Maskey and Hirschberg 2005]. The acoustic features are *PITCH*, *ENERGY*, and *CONFIDENCE*. The pitch and energy features are extracted from the broadcast news speech using the Snack toolkit [Sjölander 2001]. The confidence feature is computed as the average word posterior probability of a spoken sentence [Wessel et al. 2001], which to some extent quantifies the degree of correctness of the recognition transcripts.

Table IV. The Summarization Results Achieved by Supervised Summarizer under Different Summarization Ratios

		Summarization Ratio		
		10%	20%	30%
BC	TD	0.490	0.583	0.589
	SD	0.321	0.331	0.317
SVM	TD	0.545	0.625	0.637
	SD	0.333	0.363	0.353
CRF	TD	0.547	0.654	0.637
	SD	0.346	0.371	0.364

Note: TD = text documents, SD = spoken documents.

For each of these acoustic features, the minimum, maximum, mean, and difference values of a spoken sentence are extracted. The difference value is defined as the difference between the minimum and maximum values of the spoken sentence. Finally, the relevance features consists of *R-LSA* and *R-VSM,* which are the relevance scores of a sentence to the whole document obtained by using the LSA and VSM summarizers [Gong and Liu 2001], respectively. However, in a sense they are still derived from the lexical information. Each of the above features is further normalized by the following equation:

$$\hat{x}_m = \frac{x_m - \mu_m}{\sigma_m}, \tag{10}$$

where μ_m and σ_m are, respectively, the mean and standard deviation of a feature x_m estimated from the development set.

6. EXPERIMENTAL RESULTS AND DISCUSSIONS

6.1 Results of Experiments on the Supervised Summarizers

In the first set of experiments, we evaluate the BC, SVM, and CRF supervised summarizers discussed in Section 2. The experimental results are detailed in Table IV, where each column lists the ROUGE_2 recall rates for different summarizers using different summarization ratios. The results based on manual transcripts of the spoken documents (denoted as TD, text documents) are also shown for reference. For the TD case, the *PITCH* and *ENERGY* features were obtained by aligning the manual transcripts to their spoken documents counterpart by performing word-level forced alignment, while the *CONFIDENCE* feature was set to 1. Unlike the TD case, in which there are no recognition errors and sentence boundary errors, the recognition transcripts used in the SD (spoken documents) case may contain both recognition errors and sentence boundary errors, which inevitably degrade the performance. In this research, sentence boundaries were determined by speech pauses. It is worth mentioning that the number of labels used for training a summarizer was in accordance with the target summarization ratio in the evaluation; that is, the summarizers trained with the manual summaries at a given summarization ratio were tested at the same summarization ratio. Table IV shows that the discriminative summarizers (CRF and SVM) outperform the generative summarizer (BC). Moreover, the performance of CRF is considerably better than that of SVM, especially at lower summarization ratios (10% and 20%). This may be

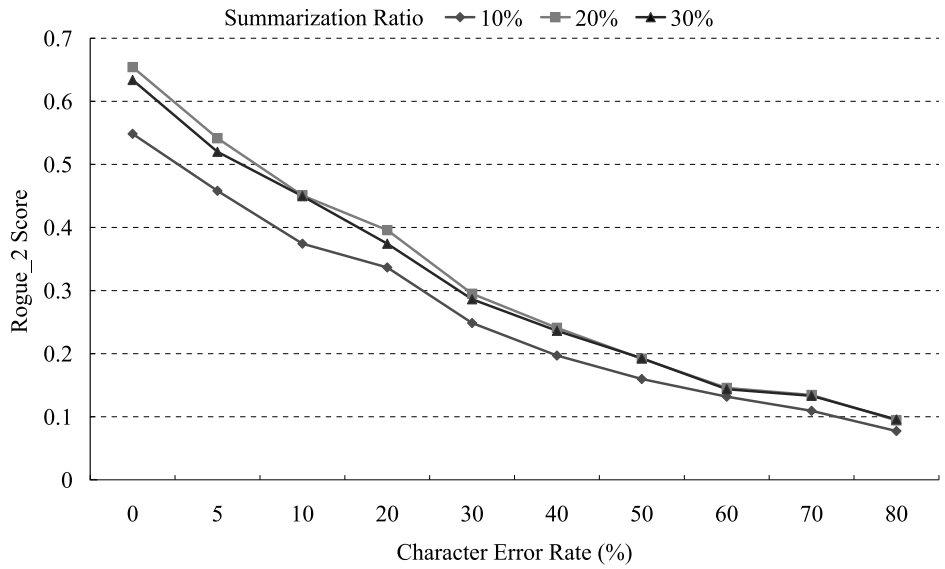

Fig. 1. The summarization results achieved by using CRF and simulated recognition transcripts of different Chinese character error rates.

because the CRF method can model the relationships among sentences. In brief, the ROUGE_2 recall rates obtained by CRF for the TD case are approximately 0.55, 0.65, and 0.64, respectively, for summarization ratios of 10%, 20%, and 30%; while they are 0.35, 0.37, and 0.36, respectively, for the same summarization ratios in the SD case. Compared with the results in Table II, both SVM and CRF yield results that are comparable to those obtained by the human subjects for the TD case, except for the case of a very low summarization ratio (e.g., 10%). However, when SVM and CRF are applied to erroneous recognition transcripts of spoken documents (i.e., the SD case), the performance degrades severely. Compared to manual transcripts, recognition transcripts are always created in the face of speech recognition errors and sentence boundary detection errors. It has been shown that speech recognition errors are the dominating factor for the performance degradation of spoken document summarization when using recognition transcripts instead of manual transcripts, whereas erroneous sentence boundaries cause relatively minor problems [Christensen et al. 2008; Liu and Xie 2008].

In order to evaluate the impact of speech recognition errors on the spoken document summarization task studied in this article, we further conducted a set of summarization experiments using CRF and simulated recognition transcripts of different Chinese character error rates (CER), which were obtained by randomly replacing a certain percentage of Chinese characters in the manual transcripts with other Chinese characters or inserting/deleting a certain percentage of Chinese characters into/from the manual transcripts. As the results show in Figure 1, the performance at all summarization ratios appears to degrade severely as CER increases; for example, when the CER is in the range between 20% and 30%, a performance drop of about 50% is encountered for the

SD case at all summarization ratios, as compared to the TD case. It should be also noted that the performance achieved by using the simulated recognition transcripts with a CER of 30% is somehow worse than that achieved by using the "real" recognition transcripts with the same CER (cf. Table IV). One possible reason is that the character errors of the real recognition transcripts tend to aggregate in some spoken words that are actually confusing and difficult to recognize, whereas the character errors of the "simulated" recognition transcripts instead are almost uniformly distributed (or scattered) over the words of the manual transcripts, which would lead to a higher word error rate and thus have relatively lower ROGUE_2 recall rates for all summarization ratios. Though the summarization methods, together with the associated experiments and evaluations, presented in this article are not intended to focus on dealing with the problems caused by speech recognition errors, they still remain worthy of further investigation, especially when summarizing spontaneous spoken documents such as voice mails, lectures, and meeting recordings [Koumpis and Renals 2005]. A straightforward remedy, apart from the many approaches improving recognition accuracy, might be to develop more robust representations for spoken documents. For example, multiple recognition hypotheses, beyond the top scoring ones, obtained from *N*-best lists, word lattices, or confusion networks, can provide alternative (or soft) representations for the confusing portions of the spoken documents [Chelba et al. 2008]. Moreover, the use of subword units (for example, syllables or segments of them), as well as the combination of words and subword units, for representing the spoken documents has also been proven beneficial for Chinese spoken document summarization [Chen et al. 2006b].

In the next set of experiments, we examine the contributions that different kinds of features (cf. Table III) make to the performance of a supervised summarizer. We, again, take the CRF model as an example because it achieved the best performance among the three supervised summarizers studied in this article. The results are shown in Table V, where the first four rows detail the ROUGE_2 recall rates obtained by CRF trained with only one set of features. Interestingly, the acoustic features (Ac) make more substantial contributions to the summarizer than the lexical features (Le) and the relevance features (Re). In the SD case, the performance of the structural features (St) is severely degraded probably because of poor sentence boundary detection. Since both the *POSITION* feature and the *DURATION* feature are closely related to the number of sentences extracted from a spoken document, the wrong number of sentences given by sentence boundary detection might result in poor estimation of them. It has also been shown that the structural (or stylistic) features tend to be affected by story or sentence segmentation inaccuracies; interested readers can also refer to the work of Christensen et al. [2008] for more comprehensive analysis of such an issue. From Table V, we observe that the combination of two kinds of features leads to a more consistent improvement than using the features separately. Combining the structural features, acoustic features, and relevance features further improves the performance, but combining the structural features, acoustic features, and lexical features does not. These results show that lexical cues are not the dominating predictors when

Table V. The Summarization Results Achieved by CRF with Different Features and Their Combinations

		Summarization Ratio		
		10%	20%	30%
Ac	TD	0.425	0.567	0.574
	SD	0.315	0.336	0.321
St	TD	0.369	0.458	0.490
	SD	0.144	0.132	0.159
Le	TD	0.324	0.464	0.494
	SD	0.287	0.272	0.273
Re	TD	0.391	0.486	0.529
	SD	0.284	0.302	0.313
Ac + St	TD	0.501	0.609	0.621
	SD	0.327	0.350	0.345
Le + Re	TD	0.510	0.555	0.577
	SD	0.302	0.318	0.319
Ac + St + Le	TD	0.495	0.634	0.622
	SD	0.319	0.368	0.343
Ac + St + Re	TD	0.545	0.631	0.634
	SD	0.346	0.362	0.350
Ac + St + Le + Re	TD	0.547	0.654	0.637
	SD	0.346	0.371	0.364
Ac + St + Le + Re + Ge	TD	0.595	0.657	0.644
	SD	0.351	0.372	0.369

Note: St = structural features, Le = lexical features, Ac = acoustic features, Re = relevance features, Ge = generative scores.

recognition transcripts contain recognition errors. The results in Figure 1 also show that the performance degrades severely as the recognition error rate increases. Therefore, exploring more nonlexical features might be beneficial for spoken document summarization, especially when the speech recognition accuracy is not perfect. Similar observations were also made by other groups [Koumpis and Renals 2005; Maskey and Hirschberg 2005, 2006; Zhang and Fung 2007]. The results in Table V also show that incorporating more indicative features can improve the performance of the summarizer.

For the CRF summarizer, relevance features (Re) are more effective than lexical features (Le), as shown by the results in Table V. This highlights the importance of capturing the relevance of a sentence to the whole document. Therefore, to further improve the performance of the CRF summarizer, we augment the feature sets with two additional generative scores (Ge) obtained from the LM and STM approaches defined in Equations (6) and (7), respectively. The results in the last row of Table V show that incorporating additional generative scores improves the performance of the supervised summarizer slightly but consistently in most cases. These results again justify our postulation that complicated features (e.g., Re or Ge) provide more useful information to the supervised summarizer than simple features extracted directly from the spoken document or sentences (e.g., Le, Ac, and St).

6.2 Results of Experiments on the Unsupervised Summarizers

In the fourth set of experiments, we compare the performance of the following unsupervised summarizers: VSM [Gong and Liu 2001; Lee and Chen 2005],

MMR [Murray et al. 2005], LSA [Gong and Liu 2001], and SIG [Furui et al. 2004], as well as our previously proposed LM and STM models [Chen et al. 2006a]. The VSM approach represents each sentence of a document and the document itself in vector form, and computes the relevance score between each sentence and the document (i.e., the cosine measure of the similarity between two vectors). Then, the sentences with the highest relevance scores are included in the summary [Gong and Liu 2001]. In contrast, LSA represents each sentence of a document as a vector in the latent semantic space of the document, which is constructed by performing singular value decomposition (SVD) on the "term-sentence" matrix of the document. The right-singular vectors with larger singular values represent the dimensions of the more important latent semantic concepts in the document. Therefore, the sentences with the largest index values in each of the top L right-singular vectors are included in the summary [Gong and Liu 2001]. The difference between VSM and LSA is that VSM performs spoken document summarization based on literal term matching, while LSA is based on concept matching [Lee and Chen 2005]. MMR is close in spirit to VSM because it also represents each sentence of a document and the document itself in vector form, and then uses the cosine measure for sentence selection; however, MMR selects sentences iteratively. In each round, the important sentence is selected according to two criteria: 1) whether it is more similar to the whole document than the other sentences, and 2) whether it is less similar to the set of sentences that have been selected so far. Consequently, MMR not only selects relevant sentences for the summary, but also ensures that the summary covers more concepts [Murray et al. 2005]. SIG, on the other hand, simply selects indicative sentences from a spoken document based on a weighted sum of the lexical, grammar, and confidence scores for each sentence [Furui et al. 2004].

In this article, when STM is employed in evaluating the relevance between a document D and each one of its sentences S_i for spoken document summarization, we additionally incorporate the unigram probabilities of a document word occurring in the sentence $P(w_n|S_i)$ and a general text collection $P(w_n|Collection)$ into STM, for probability smoothing and better performance. The probability of the document D generated by the STM model of a sentence S_i [i.e., $P_{\mathrm{STM}}(D|S_i)$ in Equation (7)] is therefore modified as follows:

$$\hat{P}_{\mathrm{STM}}(D|S_i) = \prod_{w_n \in D} \left[(1-\rho_1-\rho_2) \cdot \left(\sum_{k=1}^{K} P(w_n|T_k) P(T_k|S_i) \right) + \rho_1 \cdot P(w_n|S) + \rho_2 \cdot P(w_n|Collection) \right]^{c(w_n,D)}, \quad (11)$$

where ρ_1 and ρ_2 are weighting parameters ($0 < \rho_1, \rho_2 < 1$ and $\rho_1 + \rho_2 < 1$). Similar treatments also have been studied for the IR tasks [Hoffmann 1999; Wei and Croft 2006; Chen 2009]. $P(w_n|S_i)$ and $P(w_n|Collection)$ actually are the two constituent probability terms for the LM model, as stated earlier in Section 3.1, which can be estimated simply based on the MLE criterion;

Table VI. The Summarization Results Achieved by Various Unsupervised Summarizers under Different Summarization Ratios

		Summarization Ratio		
		10%	20%	30%
VSM	TD	0.286	0.427	0.492
	SD	0.204	0.239	0.282
LSA	TD	0.213	0.325	0.418
	SD	0.187	0.240	0.276
MMR	TD	0.292	0.433	0.492
	SD	0.204	0.241	0.280
SIG	TD	0.248	0.408	0.450
	SD	0.179	0.213	0.248
LM	TD	0.328	0.450	0.501
	SD	0.201	0.250	0.282
STM	TD	0.335	0.453	0.494
	SD	0.211	0.262	0.286
RND	TD	0.110	0.188	0.289
	SD	0.163	0.223	0.230

the weighting parameters ρ_1 and ρ_2 can be further optimized using the EM algorithm and the development set.

Table VI shows the results derived by the above unsupervised summarizers. The results obtained by random selection (denoted as RND) are also listed for comparison. All the unsupervised summarizers were tuned based on experiments on the development set. However, the document-reference summary information of the development set was not utilized in the construction of the models. Comparing the results with those in Table IV, we observe that the supervised summarizers significantly outperform all the unsupervised summarizers. Although LM and STM do not perform as well as the supervised summarizers, they clearly outperform MMR, LSA, and SIG, and their performance is comparable to that of VSM at lower summarization ratios. It is interesting that MMR almost has the same performance as VSM at all summarization ratios, despite that MMR is expected to outperform VSM because it is designed to allow the summary to cover more topics. This, in a sense, reflects that the issue of topic redundancy seems to have only a very limited impact on the accuracy of the automatic summarization studied here, probably due to the reason that each of the broadcast news documents to be summarized is short in its nature and centers on some specific topic or concept [Wang et al. 2005]. Moreover, the summarization results indicate that STM is slightly more effective than LM. It is also worth mentioning that VSM, LSA, and MMR simply use TF-IDF features for the representations of a spoken document and its associated sentences, while only word or topical unigrams (multinomial distributions) are employed for modeling the sentence generative probability in LM or STM. Finally, the superiority of the supervised summarizers over the unsupervised summarizers can be explained by two factors. The first is that the supervised summarizers make use of the handcrafted document-reference summary information for model training, whereas the unsupervised summarizers do not utilize such information. The second is that most

Table VII. The Summarization Results Achieved by SVM and CRF Trained in a Data-Selection-Based Unsupervised Manner for the SD Case

	STM Labeling + Data Selection		STM Labeling		Manual Labeling	
	SVM	CRF	SVM	CRF	SVM	C RF
10%	0.232	0.283	0.165	0.194	0.333	0.346
20%	0.262	0.275	0.253	0.262	0.363	0.371
30%	0.291	0.295	0.291	0.296	0.353	0.364

of the unsupervised summarizers rely merely on lexical features (e.g., TF-IDF and word or topic unigrams), whereas the supervised summarizers fuse more indicative features besides the lexical features to fulfill spoken document summarization. Nevertheless, almost all kinds of these features are more or less vulnerable to speech recognition errors. Therefore, the supervised summarizers show larger performance difference between the TD and SD cases than the unsupervised summarizers (except SIG) that merely use lexical features.

6.3 Training the Supervised Summarizers with the Automatic Summarization Results

In the last set of experiments, we take SVM and CRF as examples for studying data selection for unsupervised training of the supervised summarizers. The implementation of training data selection is conducted at two levels: the sentence level and the document level. For the sentence-level data selection, the sentences of each spoken document in the development set labeled by STM as summary sentences and having the average similarity defined in Equation (8) higher than a threshold τ_s will be marked as reliable summary sentences, while those sentences labeled by STM as nonsummary sentences and having the average similarity lower than τ_{ns} will be marked as reliable nonsummary sentences. The document-level data selection is then executed according to the ratio of the number of the reliable summary and nonsummary sentences to the total number of sentences in a spoken document. More specifically, the spoken documents having the ratio of reliable summary and nonsummary sentences exceeding a threshold τ_{D} will be ultimately selected for training the supervised summarizers. The number M and thresholds τ_s, τ_{ns}, and τ_{D} were tuned based on experiments on the development set. The summarization results of SVM and CRF trained in the above data-selection-based unsupervised manner (STM Labeling + Data Selection) are shown in Table VII, where the results of SVM and CRF trained without supervision and data selection (STM Labeling) and trained with supervision (Manual Labeling) are also listed for comparison. It can be found that the performance of SVM and CRF is substantially enhanced at the summarization ratio of 10% as compared to that of SVM and CRF without using data selection in unsupervised model training; however, no apparent performance improvement at higher summarization ratios (20% and 30%) is obtained. Table VIII presents the average of the average similarity among the retrieved relevant text documents for the manual summary and nonsummary sentences of the development set at different summarization

Table VIII. The Average of the Average Similarity among the Retrieved Text Documents for the Manual Summary and Nonsummary Sentences of the Development Set at Different Summarization Ratios

	10%	20%	30%
Summary sentences	0.059	0.057	0.055
Nonsummary sentences	0.047	0.046	0.045

ratios. It is observed that the retrieved relevant text documents for a manual summary sentence of a spoken document have a higher similarity than the retrieved relevant text documents for a manual nonsummary sentence, and the difference becomes smaller as the summarization ratio increases. These observations seem to explain why training data selection based on the average similarity among the retrieved text documents for the spoken sentences can improve the performance of SVM and CRF at the lower summarization ratios when they are trained in an unsupervised manner.

6.4 Discussions

The results shown in Tables IV to VIII allow us to draw several conclusions. First, for the SD case, the unsupervised summarizers (except SIG) are trained or constructed solely on the basis of erroneous lexical information. In other words, other structural or acoustic clues are not considered. To be fair, we should compare the results obtained by the unsupervised summarizers to those of the supervised summarizers (e.g., CRF) trained with lexical features only [cf. the third row (Le) in Table V]. From this perspective, the performance of the unsupervised summarizers is comparable to that of the supervised summarizer (CRF).

Second, the advantage of the supervised summarizers is that it is easy to modify and augment the feature set with more indicative features. The summarization performance can be improved steadily by including a substantial number of indicative features in the supervised summarizer, as evidenced by the results in Table V.

Third, the supervised summarizer usually learns its summarization capability by using a set of handcrafted document-reference summary exemplars; however, such information might not always be available because the summarization task changes over time. In contrast, the unsupervised summarizers (except SIG) usually consider the relevance of a sentence to the whole document, which might be more robust across different summarization tasks.

Fourth, even though supervised summarizers achieve higher accuracy for summary/nonsummary classification tasks, a considerable amount of human effort is usually required to label the training data. Though we have proposed a data selection approach for unsupervised training of the supervised summarizers, the issue of how to reduce the amount of human effort still needs further investigation.

Finally, Table IX provides a comprehensive comparison of the abovementioned classification-based and probabilistic generative summarizers from several aspects, such as feature set augmentation and sentence selection criterion.

Table IX. A Comprehensive Comparison of the Classification-Based and Probabilistic Generative Summarizers

	Classification-based Summarizers			Probabilistic Generative Summarizers	
	BN	SVM	CRF	LM	STM
Probabilistic Modeling	Generative	Discriminative	Discriminative	Generative	Generative
Manual Labeling	Yes	Yes	Yes	No	No
Feature Set Augmentation	Easy	Easy	Easy	N/A	N/A
Important Sentence Selection	Individual	Individual	Global	Individual	Individual
Use of Sentence-Document Relevance	Dependent on the Feature Set	Dependent on the Feature Set	Dependent on the Feature Set	Yes	Yes
Online Model Estimation	No	No	No	Yes	Yes
Applicable to New Tasks	Hard	Hard	Hard	Easy	Easy

7. CONCLUSIONS

We have studied the use of probabilistic ranking models for extractive spoken document summarization, and evaluated various modeling and learning approaches. The experimental results on Mandarin Chinese broadcast news show that CRF can achieve significant performance improvements compared to other summarizers. In addition, various kinds of feature representations and their effectiveness have also been investigated as well. Our future research directions include the following: 1) utilizing effective feature selection algorithms to select the features automatically, 2) exploiting more representative features for supervised summarizers, 3) exploring more elaborate data selection approaches for training supervised summarizers without manual labels, and 4) seeking other ways to combine discriminative summarizers and generative summarizers.

REFERENCES

AUBERT, X. L. 2002. An overview of decoding techniques for large vocabulary continuous speech recognition. *Comput. Speech Lang. 16*, 1, 89–114.

BAEZA-YATES, R. AND RIBEIRO-NETO, B. 1999. Modern information retrieval. Addison-Wesley.

BAXENDALE, P. B. 1958. Machine-made index for technical literature—An experiment. *IBM J. Res. Dev. 2*, 4, 354–361.

BELLEGARDA, J. R. 2004. Statistical language model adaptation: Review and perspectives. *Speech Comm. 42*, 1, 93–108.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res. 3*, 993–1022.

CHANG, C. C. AND LIN, C. J. 2001. LIBSVM: A library for support vector machines. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/libsvm.

CHELBA, C., HAZEN, T. J., AND SARAÇLAR, M. 2008. Retrieval and browsing of spoken content. *IEEE Signal Process. Mag. 25*, 3, 39–49.

CHEN, B., KUO, J. W., AND TSAI, W. H. 2004. Lightly supervised and data-driven approaches to mandarin broadcast news transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, 777–780.

CHEN, B. 2009. Word topic models for spoken document retrieval and transcription. *ACM Trans. Asian Lang. Inform. Process. 8*, 1.

CHEN, B. 2006. Exploring the use of latent topical information for statistical Chinese spoken document retrieval. *Pattern Recogn. Lett. 27*, 1, 9–18.

CHEN, B. AND CHEN, Y. T. 2008. Extractive spoken document summarization for information retrieval. *Pattern Recogn. Lett. 29*, 4, 426–437.

CHEN, B., YEH, Y. M., HUANG, Y. M., AND CHEN, Y. T. 2006a. Chinese spoken document summarization using probabilistic latent topical information. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, 969–972.

CHEN, Y. T., YU, S., WANG, H. M., AND CHEN, B. 2006b. Extractive Chinese spoken document summarization using probabilistic ranking models. In *Proceedings of the International Symposium on Chinese Spoken Language Processing (ISCSLP'06)*, 660–671.

CHRISTENSEN, H., GOTOH, Y., AND RENALS, S. 2008. A cascaded broadcast news highlighter. *IEEE Trans. Audio Speech Lang. Process. 16*, 1, 151–161.

CONROY, J. M. AND O'LEARY, D. P. 2001. Text summarization via hidden markov models. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR'01)*, 406–407.

CROFT, W. B. AND LAFFERTY, J. EDS. 2003. *Language Modeling for Information Retrieval*. Kluwer-Academic Publishers.

DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. Series B 39*, 1, 1–38.

DUDA, R. O., HART, P. E., AND STORK, D. G. 2001. *Pattern Classification*. Wiley-Interscience.

FURUI, S., KIKUCHI, T., SHINNAKA, Y., AND HORI, C. 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Trans. Speech Audio Process. 12*, 4, 401–408.

GILBERT, M. AND FENG, J. 2008. Speech and language processing over the Web. *IEEE Signal Process. Mag. 25*, 3, 42–60.

GONG, Y. AND LIU, X. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR'01)*, 19–25.

HIROHATA, M., SHINNAKA, Y., IWANO, K., AND FURUI, S. 2005. Sentence extraction-based presentation summarization techniques and evaluation metrics. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, 1065–1068.

HOFMANN, T. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn. 42*, 177–196.

KOUMPIS, K. AND RENALS S. 2005. Automatic summarization of voicemail messages using lexical and prosodic features. *ACM Trans. Speech Lang. Process. 2*, 1, 1–24.

KUMAR, N. 1997. Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition. Ph.D. Thesis. John Hopkins University.

KUPIEC, J., PEDERSEN, J., AND CHEN, F. 1999. A trainable document summarizer. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR'99)*, 68–73.

LAFFERTY, J., MCCALLUM, A., AND PEREIRA, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML'01)*, 282–289.

LEE, L. S. AND CHEN, B. 2005. Spoken document understanding and organization. *IEEE Signal Process. Mag. 22*, 5, 42–60.

LIN, C. Y. 2003. ROUGE: Recall-oriented understudy for gisting evaluation. Retrieved from http://www.isi.edu/~cyl/ROUGE/.

LIN, H. T., LIN, C. J., AND WENG, R. C. 2003. A note on Platt's probabilistic outputs for support vector machines. *Mach. Learn. 68*, 3, 267–276.

LIU, S. H., CHU, F. H., LIN, S. H., LEE, H. S., AND CHEN, B. 2007. Training data selection for improving discriminative training of acoustic models. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'07)*, 284–289.

LIU, Y. AND XIE, S. 2008. Impact of automatic sentence segmentation on meeting summarization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08)*, 5009–5012.

MASKEY, S. AND HIRSCHBERG, J. 2005. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *Proceedings of the European Conference Speech Communication and Technology (EUROSPEECH'05)*, 621–624.

MASKEY, S. AND HIRSCHBERG, J. 2006. Summarizing speech without text using hidden markov models. In *Proceedings of Human Language Technology Conference – North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL'06)*.

MURRAY, G., RENALS, S., AND CARLETTA, J. 2005. Extractive summarization of meeting recordings. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'05)*, 593–596.

NOMOTO, T. AND MATSUMOTO, Y. 2001. A new approach to unsupervised text summarization. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR'99)*, 26–34.

ORTMANNS, S., NEY, H., AND AUBERT, X. 1997. A word graph algorithm for large vocabulary continuous speech recognition. *Comput. Speech Lang. 11*, 1, 43–72.

PAICE, C. D. 1990. Constructing literature abstracts by computer: Techniques and prospects. *Inform. Process. Manag. 26*, 1, 171–186.

POVEY, D. AND WOODLAND, P. C. 2002. Minimum phone error and I-smoothing for improved discriminative training. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, 105–108.

SAON, G., PADMANABHAN, M., GOPINATH, R., AND CHEN, S. 2000. Maximum likelihood discriminant feature spaces. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, 1129–1132.

SHEN, D., SUN, J. T., LI, H., YANG, Q., AND CHEN, Z. 2007. Document summarization using conditional random fields. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'07)*, 2862–2867.

SJÖLANDER, K. 2001. Snack Sound Toolkit. Retrieved from http://www.speech.kth.se/snack/.

STOLCKE, A. 2005. SRI Language Modeling Toolkit. http://www.speech.sri.com/projects/srilm/.

VAPNIK, V. N. 1998. *Statistical Learning Theory*. Wiley-Interscience.

WANG, H. M., CHEN, B., KUO, J. W., AND CHENG, S. S. 2005. MATBN: A Mandarin Chinese broadcast news corpus. *Int. J. Comput. Ling. Chinese Lang. Process. 10*, 2, 219–236.

WEI, X. AND CROFT, W. B. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR'06)*, 178–185.

WESSEL, F., SCHLUTER, R., MACHEREY, K. AND NEY, H. 2001. Confidence measures for large vocabulary continuous speech recognition. *IEEE Trans. Speech Audio Process. 9*, 3, 288–298.

WITBROCK, M. AND MITTAL, V. 1999. Ultra summarization: A statistical approach to generating highly condensed non-extractive summaries. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, 315–316.

WONG, K. F., WU, M., AND LI, W. 2008. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the International Conference on Computational Linguistics (COLING'08)*, 985–992.

ZHAI, C. AND LAFFERTY, J. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR'01)*, 334-342.

ZHANG, J., CHAN, H. Y., FUNG, P., AND CAO, L. 2007. A comparative study on speech summarization of broadcast news and lecture speech. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'07)*, 2781–2784.

ZHANG, J. AND FUNG, P. 2007. Speech summarization without lexical features for mandarin broadcast news. In *Proceedings of Human Language Technology Conference – North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL'07)*, 213–216.

ZHU, X. 2005. Semi-supervised learning literature survey. Tech. rep. Computer Sciences, University of Wisconsin-Madison.

Received January 2008; revised August 2008; accepted November 2008