ARTICLE IN PRESS

Information Processing and Management xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Extractive speech summarization using evaluation metric-related training criteria

Berlin Chen *, Shih-Hsiang Lin, Yu-Mei Chang, Jia-Wen Liu

Department of Computer Science & Information Engineering, National Taiwan Normal University, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 15 September 2010
Received in revised form 14 December 2011
Accepted 16 December 2011
Available online xxxx

Keywords:
Speech summarization
Sentence ranking
Imbalanced-data
Evaluation metric
Discriminative training

ABSTRACT

The purpose of extractive speech summarization is to automatically select a number of indicative sentences or paragraphs (or audio segments) from the original spoken document according to a target summarization ratio and then concatenate them to form a concise summary. Much work on extractive summarization has been initiated for developing machine-learning approaches that usually cast important sentence selection as a two-class classification problem and have been applied with some success to a number of speech summarization tasks. However, the imbalanced-data problem sometimes results in a trained speech summarizer with unsatisfactory performance. Furthermore, training the summarizer by improving the associated classification accuracy does not always lead to better summarization evaluation performance. In view of such phenomena, we present in this paper an empirical investigation of the merits of two schools of training criteria to alleviate the negative effects caused by the aforementioned problems, as well as to boost the summarization performance. One is to learn the classification capability of a summarizer on the basis of the pair-wise ordering information of sentences in a training document according to a degree of importance. The other is to train the summarizer by directly maximizing the associated evaluation score or optimizing an objective that is linked to the ultimate evaluation. Experimental results on the broadcast news summarization task suggest that these training criteria can give substantial improvements over a few existing summarization methods.

1. Introduction

Due to the rapid development and maturity of multimedia technology, large volumes of information content have been represented as audio-visual multimedia instead of static texts. Clearly, speech is one of the most important sources of information about multimedia content. However, unlike text documents, which are structured with titles and paragraphs and are thus easier to retrieve and browse, the associated spoken documents of multimedia content are only presented with video or audio signals; hence, they are difficult to browse from beginning to end (Lee & Chen, 2005; Ostendorf, 2008). Even though spoken documents are automatically transcribed into words, incorrect information (resulting from recognition errors and inaccurate sentence or paragraph boundaries) and redundant information (generated by disfluencies, fillers, and repetitions) would prevent them from being accessed easily. Speech summarization, which attempts to distill important information and remove redundant and incorrect content from spoken documents, can facilitate users to review spoken documents efficiently and understand associated topics quickly (Chen, Chen, & Wang, 2009; Furui, Kikuchi, Shinnaka, & Hori, 2004).

* Corresponding author.
E-mail address: berlin@csie.ntnu.edu.tw (B. Chen).
URL: http://berlin.csie.ntnu.edu.tw (B. Chen).

0306-4573/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2011.12.002

Please cite this article in press as: Chen, B., et al. Extractive speech summarization using evaluation metric-related training criteria. *Information Processing and Management* (2012), doi:10.1016/j.ipm.2011.12.002

Automatic summarization of text documents dates back to the early 1950s (Baxendale, 1958). Nowadays, the research is extended to cover a wider range of tasks, including multidocument, multilingual, and multimedia summarization. Broadly speaking, summarization can be either extractive or abstractive. The former selects important sentences or paragraphs from an original document according to a target summarization ratio and concatenates them to form a summary; the latter, on the other hand, produces a concise abstract of a certain length that reflects the key concepts of the document, thus requiring highly sophisticated natural language processing techniques, like semantic inference and natural language generation, to name a few. Thus, in recent years, researchers have tended to focus on extractive summarization. In addition to being extractive or abstractive, a summary may also be generated by considering factors from other aspects like being generic or query-oriented. A generic summary highlights the most salient information in a document, whereas a query-oriented summary presents the information in a document that is most relevant to a user's query. Interested readers may refer to (Mani & Maybury, 1999) for a comprehensive overview of the principal trends and the classical approaches for text summarization.

This paper focuses exclusively on generic, extractive speech summarization since it usually constitutes the essential building block for many other speech summarization tasks. It should also be mentioned that speech summarization presents opportunities that do not exist for text summarization; for instance, information cues about prosody/acoustics and emotion/speakers can help the determination of the importance and structure of spoken documents (Christensen, Gotoh, & Renals, 2008; McKeown, Hirschberg, Galley, & Maskey, 2005). We thus set the goal at selecting the most representative sentences (or the associated audio segments) based on the speech recognition (erroneous) transcripts, as well as a rich set of lexical and non-lexical features, to form the summary for a given spoken document. In particular, we have recently introduced a new perspective on the problem of speech summarization, saying that some potential defects of the existing supervised speech summarizers (see Section 2) can be mitigated by leveraging training criteria that have the ability to connect the decision of a summarizer to the evaluation metric (Lin, Chang, Liu, & Chen, 2010).

Our work in this paper continues this general line of research, including exploring and comparing more speech summarizers developed along this line of research and providing more in-depth elucidations of their modeling characteristics and associated empirical evaluations. To this end, our first attempt, inspired from the notion of "*learning to rank*," is to train a summarizer in a pair-wise rank-sensitive manner (Burges et al., 2005; Cao et al., 2006; Herbrich, Graepel, & Obermayer, 2000, chap. 7; Joachims, 2002). This training objective is not only at the labeling correctness of each sentence of a training spoken document, but also at the correct ordering (preference) relationship of each sentence pair in accordance with their respective importance to the document. Nevertheless, it turns out that this attempt in essence would be loosely related to the evaluation metric. In this regard, the other attempt is instead to train the summarizer by directly maximizing the evaluation score of the summarizer (Joachims, 2005; Xu & Li, 2007) or optimizing an objective that is linked to the ultimate evaluation. Furthermore, we extensively study and evaluate the utility of augmenting the feature set of supervised summarizers with more indicative features derived from various unsupervised summarizers.

The remainder of this paper proceeds as follows. Section 2 reviews the conventional approaches to speech summarization. Section 3 sheds light on the principles that the evaluation metric-related training criteria are built upon, and explains how they can be exploited for speech summarization. Section 4 describes a variety of features that are generated to represent spoken documents and sentences. Then, the experimental settings and a series of summarization experiments are presented in Section 5. Finally, Section 6 concludes our presentation and discusses avenues for future work.

2. Related work

2.1. Supervised summarizers

As to the development of speech summarizers, quite several machine-learning methods have been explored with some success recently (Chen et al., 2009; Kupiec, Pedersen, & Chen, 1999; Ouyang, Li, Li, & Lu, 2011; Shen, Sun, Li, Yang, & Chen Z., 2007), and they may broadly fall into two main categories: supervised and unsupervised speech summarizers. Supervised summarizers usually formulize the speech summarization task as a two-class (summary/non-summary) sentence-classification problem: Each sentence S_i in a spoken document to be summarized is associated with a set of M indicative features $X_i = \{x_{i1}, \ldots, x_{im}, \ldots, x_{iM}\}$ (*cf.* Section 4) and a summarizer (or a ranking function) is employed to classify and assign a class-specific importance (or decision) score to each sentence S_i according to its associated features X_i. Then, sentences of the document can be iteratively selected into the summary based on their scores until the length limitation or a desired summarization ratio is reached. During the training phase, a set of training spoken documents $\mathbf{D} = \{d_1, \ldots, d_n, \ldots, d_N\}$, consisting of N documents and the corresponding handcrafted summary information, is given. The summarizer is trained in the sense of reducing the classification (labeling) errors of the summarizer made on the sentences of these training spoken document exemplars. It is expected that minimizing these errors caused by the summarizer would be equivalent to maximizing the lower bound of the summarization evaluation score (usually, the higher the score, the better the performance). Representative techniques include, but not limited to, Bayesian classifier (BC), support vector machine (SVM), and conditional random fields (CRF) (Lin, Chen, & Wang, 2009). Among them, support vector machines (SVM) has prominently used to formulate and crystallize the above conception for speech summarization (Penn & Zhu, 2008; Xie & Liu, 2010; Zhang, Chan, Fung, & Cuo, 2007). An SVM summarizer is developed under the basic 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

Please cite this article in press as: Chen, B., et al. Extractive speech summarization using evaluation metric-related training criteria. *Information Processing and Management* (2012), doi:10.1016/j.ipm.2011.12.002

function that can correctly separate the positive and negative samples, and ensure 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 the LIBSVM toolkit (Chang & 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 **S** can be approximated by the following sigmoid operation:

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

where the weights α and β are optimized by the development set, and $g(X_i)$ is the decision value of X_i provided by the SVM summarizer. Once the SVM summarizer has been properly constructed, the sentences of the spoken document to be summarized can be ranked by their posterior probabilities of being in the summary class. The sentences with the highest probabilities are then selected and sequenced to form the final summary according to different summarization ratios.

The imbalanced-data (or skewed-data) problem, however, might strongly affect the performance of a supervised speech summarizer. This problem stems from the fact that the summary sentences of a given training spoken document usually are in a smaller portion (e.g., 10%) as compared to non-summary ones. When training a supervised summarizer on the basis of such an imbalanced-data set, the resulting summarizer tends to assign sentences of the document to be summarized to the class of non-summary sentences (viz. the majority class), thereby leading to high classification accuracy over the class of non-summary sentences but poor accuracy over the class of summary sentences (viz. the minority class). Several heuristic methods have been proposed to relieve this problem, like re-sampling (up-sampling, down-sampling, or both) or re-weighting of the training exemplars, which demonstrate modest improvements (Maloof, 2003; Xie & Liu, 2010). On the other hand, higher sentence classification accuracy does not always imply better summarization quality. This is mainly because that the summarizer usually classifies each sentence individually with little consideration of relationships among the sentences of the document to be summarized.

2.2. Unsupervised summarizers

Another stream of thought attempts to conduct document summarization based on some statistical evidences between each sentence and the document, without recourse to manually labeled training data. We may name them unsupervised summarizers. For instance, the graph-based methods, including LexRank (Erkan & Radev, 2004), TextRank (Mihalcea & Tarau, 2005), Markov Random Walk (MRW) (Wan & Yang, 2008) and so on, conceptualize the document to be summarized as a network of sentences, where each node represents a sentence and the associated weight of each link represents the lexical similarity relationship between a pair of nodes. Document summarization thus relies on the global structural information embedded in such conceptualized network, rather than merely considering the local features of each node (sentence). Put simply, sentences more similar to others are deemed more salient to the main theme of the document. Some other studies investigate the use of probabilistic models to capture the relationship between sentences and the document (DauméIII & Marcu, 2006; Nenkova, Vanderwende, & McKeown, 2006). Yet, there is a recent attempt that employs a probabilistic ranking framework for speech summarization, where the summarization task is conducted in a purely unsupervised manner (Chen et al., 2009). In this framework, important sentences are selected on the basis of either the probability of a sentence model generating the document content or the probabilistic distance between each sentence model and the document model (Lin, Yeh, & Chen, 2011). Even though the performance of the abovementioned unsupervised summarizers is usually worse than that of supervised summarizers, their domain-independent and easy-to-implement properties still make them attractive.

3. Proposed speech summarization methods

3.1. Learning to rank with pair-wise preference information

The notion of "*learning to rank*" is to create a rank- or preference-sensitive ranking function. It assumes there exists a set of ranks (or preferences) $L = \{l_1, l_2, \cdots, l_M\}$ in the output space, while in the context of speech summarization, the value of M, for example, can be simply set to 2 representing that a sentence can have the label of being either a summary (l_1) or a non-summary (l_2) sentence. The elements in the rank set have a total ordering relationship $l_1 \succ l_2 \succ \cdots \succ l_M$ where $\succ$ denotes a preference relationship. In this paper, we explore the use of the so-called "*pair-wise training*" strategy for speech summarization, which considers not only the importance of sentences to a training spoken document but also the order of each sentence pair on the ideal ranked list (Cao et al., 2006). Several embodiments have been made to fulfill the "*pair-wise training*" strategy for various information retrieval (IR) related tasks in the past decade. Typical techniques include Ranking SVM (Cao et al., 2006; Herbrich et al., 2000, chap. 7), RankBoost (Freund, Iyer, Schapire, & Singer, 2003) and RankNet (Burges et al., 2005). Each of these methods has its own merits and limitations; however, to our knowledge, this criterion has not yet been extensively explored in the context of speech summarization. Thus, in this paper we take Ranking SVM (Cao et al., 2006) as an example to implement this strategy for speech summarization, since it has shown to offer consistent improvements over traditional SVM in many IR-related tasks (Cao et al., 2006; Herbrich et al., 2000, chap. 7; Joachims, 2005). In extractive speech

Please cite this article in press as: Chen, B., et al. Extractive speech summarization using evaluation metric-related training criteria. *Information Processing and Management* (2012), doi:10.1016/j.ipm.2011.12.002

summarization, the training objective of Ranking SVM is to find a ranking function that can correctly determine the preference relation between any pair of sentences:

$$l(S_i) \succ l(S_j) \Longleftrightarrow f(X_i) \succ f(X_j), \tag{2}$$

where $l(\cdot)$ denotes the label of a sentence and $f(\cdot)$ denotes the decision value of a sentence provided by Ranking SVM. For a more thorough and entertaining discussion of Ranking SVM, interested readers can refer to (Cao et al., 2006).

3.2. Training summarizers with objectives related the evaluation metric

Although reducing the sentence classification (e.g., SVM) or ranking (e.g., Ranking SVM) errors would be equivalent to maximizing the lower bound of the performance evaluation score of a given summarization system, it is still not closely related enough to the final evaluation metric for speech summarization. Recently, quite a few approaches have been proposed to train an IR system by directly maximizing the associated evaluation score. For instances, Joachims (2005) presented an SVM-based method for directly optimizing multivariate nonlinear performance measures like the F1-score or Precision/Recall Breakeven Point (PRBEP) adopted in document classification. On the other hand, Cossock and Zhang (2006) discussed the issue of learning to rank with preference to the top scoring documents of a given training query. More recently, Xu and Li (2007) proposed an ensemble-based algorithm that can iteratively optimize an exponential loss function based on various kinds of IR evaluation metrics, often referred to as AdaRank. However, as far as we are aware, there is little research exploring the evaluation metric-related training criteria for extractive speech summarization. As such, we try to adopt such notions for speech summarization, and the AdaRank training algorithm and two novel discriminative training objectives are taken as the initial attempts.

3.2.1. The AdaRank training algorithm

The fundamental premise of AdaRank basically lies in that ensemble-based systems may produce more favorable results than their single-classifier counterparts (Polikar, 2006). AdaRank is one variation of the AdaBoost algorithm that generates a set of weak rankers (or ranking functions) and integrates them through a linear combination to form the final ranking model (Polikar, 2006; Xu & Li, 2007). A weak ranker can be constructed in several ways by using, for example, different subsets of training exemplars. In implementation, we follow the original definition of AdaRank (Xu & Li, 2007) by using single summarization features (*cf.* Section 3.2) as weak rankers. Conceptually, AdaRank learns a weight for each weak ranker from an iteratively updated distribution of the training document exemplars (Polikar, 2006). At each round, the updated distribution will emphasize those training spoken documents having more sentences incorrectly ranked by the previously selected weak rankers, which actually is evidenced by the corresponding summarization performance of the training spoken documents. Consecutive rankers are concentrated on dealing with those "*hard-to-summarize*" training spoken documents. AdaRank, therefore, belongs to a kind of the "direct optimization" training algorithms.

A bit of terminology: Given a set of training spoken documents $H = \{(d_n, Y_n)\}_{n=1}^{N}$, where Y_n is the ideal importance ranking of sentences in a document d_n provided by human subjects, AdaRank will select, at each iteration t, a single summarization feature x_t (*cf.* Section 5) that has the best overall evaluation performance on the training documents:

$$\sum_{n=1}^{N} w_t(d_n) E(\Pi(d_n, x_t), Y_n), \tag{3}$$

where $\Pi(d_n, x_t)$ is the automatically generated summary with a specific ordering of selected sentences; $E(\Pi(d_n, x_t), Y_n)$ denotes the summarization evaluation performance on the document d_i evaluated using solely the feature x_t, usually ranging from 0 to 1 (the higher the value the better the performance); $w_t(d_n)$ denotes the contribution of d_n made to the training at iteration t, which can be further expressed by

$$w_t(d_n) = \frac{\exp\{-E(\hat{\Pi}(d_n, X_{t-1}), Y_n)\}}{\sum_{n'=1}^{N} \exp\{-E(\hat{\Pi}(d_{n'}, X_{t-1}), Y_{n'})\}}, \tag{4}$$

where $E(\hat{\Pi}(d_n, X_{t-1}), Y_n)$ is the summarization performance on the document d_i using all the summarization features selected from iterations 1 to $t-1$, i.e., $X_{t-1} = \{x_1, x_2, \ldots, x_{t-1}\}$; $\exp\{\cdot\}$ denotes the exponential function. Eq. (4) reveals that a document having higher summarization performance with the features selected so far (or during the previous $t-1$ iterations) will play a less pronounced role at the current iteration. Also noteworthy is that as a specific summarization feature x_t is being selected at iteration t, its corresponding weight α_t will be determined by AdaRank through the following equation:

$$\alpha_t = \frac{1}{2} \cdot \ln \frac{\sum_{n=1}^{N} w_t(d_n)\{1 + E(\Pi(d_n, x_t), Y_n)\}}{\sum_{n'=1}^{N} w_t(d_{n'})\{1 - E(\Pi(d_{n'}, x_t), Y_{n'})\}}. \tag{5}$$

At the end, with the completion of iteration t', we can rank a spoken sentence S_i according to its importance score $I(S_i)$ expressed by

Please cite this article in press as: Chen, B., et al. Extractive speech summarization using evaluation metric-related training criteria. *Information Processing and Management* (2012), doi:10.1016/j.ipm.2011.12.002

$$I(S_i) = \sum_{m=1}^{t'} \alpha_m g(S_i, x_m), \tag{6}$$

where $g(S_i, x_m)$ is the corresponding decision value of the selected feature x_m that is employed to represent S_i .

3.2.2. Discriminative training of speech summarizers

In recent years, there has been a growing interest in developing discriminative training algorithms for reranking of hypotheses output from a baseline speech recognition system in an attempt to optimize the final performance measure of speech recognition (Oba, Hori, & Nakamura, 2010; Roark, Saraclar, & Collins, 2007). These algorithms actually bear a close resemblance to Ranking SVM and AdaRank in their functionality, and are therefore anticipated to carry over well to extractive speech summarization. However, such a conception of discriminative training has never been extensively explored for speech summarization, as far as we know. Hence, in this paper, we investigate to leverage discriminative training to estimate speech summarizers. For this idea to work, we first adapt the global conditional log-linear model (GCLM) (Roark et al., 2007) to implement a speech summarizer for both its simplicity and effectiveness (Oba et al., 2010; Roark et al., 2007). GCLM will give a decision score to an arbitrary sentences S_i of a spoken document d_n to be summarized according to the posterior probability $P_{\mathrm{GCLM}}(S_i|d_n)$ which is approximated by

$$P_{\mathrm{GCLM}}(S_i|d_n) = \frac{\exp(X_i \cdot \zeta)}{\sum_{l=1}^{L_n} \exp(X_l \cdot \zeta)}, \tag{7}$$

where X_i is the M-dimensional feature vector X_i of S_i; ζ is the M -dimensional parameter vector of GCLM; $X_i \cdot \zeta$ is the inner product of X_i and ζ; and L_n is the total number of sentences in d_n. Further, as an instantiation of exploring the discriminative training paradigm for speech summarization, we define and optimize the following training objective so as to estimate the parameter vector ζ of GCLM:

$$F_{\mathrm{GCLM\_I}} = \sum_{n=1}^{N} \sum_{S_i \in \mathbf{Summ}_n} \log \frac{P_{\mathrm{GCLM}}(S_i|d_n)}{\sum_{l=1}^{L_n}(1 - e(S_l, \mathbf{Summ}_n))P_{\mathrm{GCLM}}(S_l|d_n)}, \tag{8}$$

where $\mathbf{Summ}_n$ is the reference summary of a training document d_n; $e(S_l, \mathbf{Summ}_n)$ is the summarization performance obtained by comparing a sentence S_l of d_n to $\mathbf{Summ}_n$ with a desired evaluation metric that will return a score ranging between 0 and 1 (again, the higher the value, the better the performance). The training objective defined in (8) seeks not to maximize the posterior probabilities of the summary sentences of all training spoken documents given the summarization model (viz. with the parameter vector ζ), but also to boost negative impact of those sentences that have inferior summarization performance (or are more dissimilar from the reference summary) on the training objective, thus generating more confusable data (or spoken documents) for discriminative training of GCLM. Note also that this training objective is intrinsically very similar to the other discriminative training objectives that have been studied and practiced in the acoustic modeling for speech recognition, such as boosted maximum mutual information estimation (boosted MMIE) (Povey et al., 2008) and conditional maximum likelihood estimation (CMLE) (Roark et al., 2007).

In this paper, we also explore the use of an alternative training objective for GCLM, which aims to maximize the expected summarization evaluation scores of all sentences of the training spoken documents:

$$F_{\mathrm{GCLM\_II}} = \sum_{n=1}^{N} \sum_{l=1}^{L_n} e(S_l, \mathbf{Summ}_n) P_{\mathrm{GCLM}}(S_l|d_n). \tag{9}$$

We can see from (9) that by training the GCLM model with the objective $F_{\mathrm{GCLM\text{-}II}}$, the summary sentences will tend to have higher posterior probabilities, and vice versa for the non-summary sentences (or those sentences that are more dissimilar from the reference summary). The training objective shown in (9) is close in spirit to those that had ever used in minimum phone error training (MPE) (Povey & Woodland, 2002) and minimum error rate training (MERT) (Och, 2003) in the fields of speech recognition and machine translation.

4. Features for speech summarization

Although the above approaches can be applied to both text and spoken documents, the latter presents unique difficulties, such as recognition errors, problems with spontaneous speech, and the lack of correct sentence or paragraph boundaries. To avoid redundant or incorrect content while selecting important and correct information, multiple recognition hypotheses, confidence scores, language model scores, and other grammatical knowledge can be utilized. In addition, acoustic features (e.g., intonation, pitch, energy, and pause duration) can provide important clues for summarization; although reliable and efficient ways to use these acoustic features are still under active research (Chen & Lin, 2012; Lin et al., 2009; Zhang et al., 2007).

In this paper, we use a set of 29 features to characterize a spoken sentence, including the structural feature, the lexical features, the acoustic features and the relevance features. Structural feature simply illustrates the duration or length information of a spoken sentence; lexical features represent the linguistic characteristics; acoustic features describe more about

Please cite this article in press as: Chen, B., et al. Extractive speech summarization using evaluation metric-related training criteria. *Information Processing and Management* (2012), doi:10.1016/j.ipm.2011.12.002

how things are said than what is said, and may provide additional important information for summarization; and relevance features evaluate the relevance between a document and each one of its sentences. For each kind of acoustic features, the minimum, maximum, mean, difference value and mean difference value (indexed from 1 to 5) of a spoken sentence are extracted. The difference value is defined as the difference between the minimum and maximum values of the spoken sentence, while the mean difference value is defined as the mean difference between a sentence and its previous sentence. The features are outlined in Table 1, where, in addition to MRW, VSM (Vector Space Model) (Gong & Liu, 2001), LSA (Latent Semantic Analysis) (Gong & Liu, 2001) and WTM (Word Topic Model) (Chen, 2009) are different unsupervised summarizers, respectively, producing single summarization (relevance) features. VSM represents each sentence of a document, and the whole document, in vector form. In this approach, each dimension specifies the weighted statistics, for example the product of the term frequency (TF) and inverse document frequency (IDF), associated with an indexing term (or word) in the spoken sentence or document. Sentences with the highest relevance scores to the whole document (usually calculated by the cosine score of two vectors) are included in the summary accordingly. VSM solely based on matching the literal words that are present in the sentences and the document would sometimes fail to include enough relevant sentences in the summary because of the word mismatch problem. LSA is a natural extension of VSM that represents each sentence of a document to be summarized in a latent semantic space. To accomplish this, singular value decomposition (SVD) is performed on the "term-sentence" matrix of the document, for which 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 R right singular vectors are included in the summary. LSA thus exhibits some sort of concept matching.

On the other hand, WTM regards each word w_j of the language as a generative M_{w_j} that can be used to predict the occurrence of another word. To get to this point, all words are assumed to share a same set of K latent topic distributions $\{T_1, \cdots, T_k, \cdots, T_K\}$, but have different weights over these topics $P(T_k|M_{w_j})$, while each topic offers a unigram (multinomial) distribution $P(w|T_k)$ for observing an arbitrary word w of the vocabulary:

$$P_{\mathrm{WTM}}(w|M_{w_j}) = \sum_{k=1}^{K} P(w|T_k)P(T_k|M_{w_j}). \tag{10}$$

Each sentence S of a document d (to be summarized) can be viewed as a composite WTM model for generating the document:

$$P_{\mathrm{WTM}}(d|S) = \prod_{w\in d}\sum_{w_j\in S} P_{\mathrm{WTM}}(w|M_{w_j})P(w_j|S), \tag{11}$$

where $P(w_j|S)$ is the probability of w_j occurring in S. The resulting composite WTM model for S, in a sense, can be thought of as a kind of language model for translating any word occurring in S to an arbitrary word of d. Important sentences are thus selected according to their associated document-likelihoods $P_{\mathrm{WTM}}(d|S)$. WTM, to some extent, can be viewed as a probabilistic counterpart of LSA for concept matching; due to limited space, we refer the reader to (Chen, 2009) for a more detailed introduction to the theoretical background (including the training) and some practical applications of WTM to speech recognition and information retrieval.

Each of the above features is further normalized by the following equation:

$$\hat{x}_m = \frac{x_m - \mu_m}{\sigma_m}, \tag{12}$$

where μ_m and σ_m are, respectively, the mean and standard deviation of a feature x_m estimated from the development set (cf. Section 5.1.2). Notice that the positional feature is excluded in this study because it is not general enough and would highly depend on the epoches and genres of spoken documents (Christensen et al., 2008; Lin et al., 2010).

Table 1
Features used in the summarizers.

Types	Description
Structural feature	1. Duration of the current sentence (S1)
Lexical features	1. Number of named entities (L1)
	2. Number of stop words (L2)
	3. Bigram language model scores (L3)
	4. Normalized bigram scores (L4)
Acoustic features	1. The 1st formant (F1-1 to F1-5)
	2. The 2nd formant (F2-1 to F2-5)
	3. The pitch value (P-1 to P-5)
	4. The peak normalized cross-correlation of pitch (C-1 to C-5)
Relevance features	1. Relevance score obtained by WTM
	2. Relevance score obtained by VSM
	3. Relevance score obtained by LSA
	4. Relevance score obtained by MRW

Please cite this article in press as: Chen, B., et al. Extractive speech summarization using evaluation metric-related training criteria. *Information Processing and Management* (2012), doi:10.1016/j.ipm.2011.12.002

5. Experiments

In this section, we will describe the experimental setup and then present a series of experiments conducted to assess summarization performance as a function of manual/recognition transcripts, features used for sentence ranking, and different supervised summarizers that are taken as the vehicle for combining features.

5.1. Experimental setup

5.1.1. Speech and text corpora

The speech data set used in this research is the MATBN corpus (Lin et al., 2010; Wang, Chen, Kuo, & Cheng, 2005), which contains approximately 200 h 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, 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 (Liu, Chu, Lin, Lee, & Chen, 2007; Povey & Woodland, 2002). The average Chinese character error rate (CER) obtained for the 205 spoken documents was about 35% (Liu et al., 2007). Some basic statistics of the 205 spoken documents are given in Table 2.

Additionally, 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 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). A subset of about 14,000 text news documents, compiled during the same period as the broadcast news documents to be summarized, was used to calculate the IDF statistics of VSM and estimate the parameters of WTM, as mentioned in Section 4.

5.1.2. Evaluation metric

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 generated 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. To assess the goodness of the automatically generated summaries, we used the ROUGE measure as the evaluation metric (Lin, 2003; Liu & Liu, 2010). The ROUGE measure evaluates the quality of the summarization by counting the number of overlapping units, such as *N*-grams, longest common subsequences or skip-bigram, between the automatic summary and a set of reference (manually-annotated) summaries. Three widely used variants of the ROGUE measure were adopted to assess the utility of the summarization methods presented in this paper. They are, respectively, the ROUGE-1 (unigram) measure, the ROUGE-2 (bigram) measure and the ROUGE-L (longest common subsequence) measure. Generally speaking, the ROUGE-1 measure is to evaluate the informativeness of automatic summaries while the ROUGE-2 measure is to estimate the fluency of automatic summaries. On the contrary, ROUGE-L does not reward for fixed-length *N*-grams but instead for a combination of the maximal substrings of words, which works well in general for evaluating both content and grammaticality. The summarization ratio, defined as the ratio of the number of sentences in the automatic (or manual) summary to that in the manual transcript of a spoken document, was set to 10% in this study.

Table 3 shows the levels of agreement between the three subjects for important sentence ranking. Each of these values was obtained by 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.

Table 2
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 s)	129.4	135.2
Avg. number of words per document	326	340
Avg. number of sentences per document	20	20
Avg. character error rate	34.4%	35.3%

Please cite this article in press as: Chen, B., et al. Extractive speech summarization using evaluation metric-related training criteria. *Information Processing and Management* (2012), doi:10.1016/j.ipm.2011.12.002

5.2. Experimental results

5.2.1. Baseline results by using single features

At the outset, we examine the summarization performance when sentence ranking acts on different single features (or unsupervised summarizers, *cf.* Table 1) that were derived based on the recognition transcripts along with their corresponding audio segments (denoted by SD, spoken documents). The associated results are graphically illustrated in Fig. 1. In addition, the results based on the manual transcripts of spoken documents (denoted by TD, text documents) are also sketched in Fig. 1 for reference. For the TD case, the acoustic features were obtained by performing word-level forced alignment of the audio segments of the spoken documents to their corresponding manual transcripts. Inspection of Fig. 1 reveals two particularities. On one hand, the performance of the TD case is significantly better than that of the SD case. This might explained by the fact that the various ROUGE measures are based on counting the number of overlapping units between the automatic summary and the reference summary. Even though the summary sentences can be correctly selected or identified, the evaluation will inevitably be strongly affected by the recognition errors. On the other hand, the relevance features generally seem to be more effective than the other simple (or raw) features. This is because the relevance features, to some extent, are designed for capturing the importance (or relevance) of a sentence to the whole document or/and the relevance between sentences. They thus might be more closely related to the notion of identifying important or relevant sentences from a spoken document.

To take a step further, WTM and MRW are competitive to each other. WTM performs slightly better than MRW when using manual transcripts (i.e., the TD case); however, an opposite phenomenon is witnessed when using recognition transcripts (i.e., the SD case). One possible speculation is that, unlike MRW, the model parameters of WTM are all estimated from an outside set of text news documents (cf. Section 5.1.1), which somewhat makes WTM unable to faithfully capture the topical relationship among words in the imperfect recognition transcripts (Chen, 2009).

Nevertheless, the performance of almost all the features compared here is more or less plagued by speech recognition 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). 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 *M*-best lists, word lattices, or confusion networks, can provide alternative (or soft) representations for the confusing portions of the spoken documents (Chelba, Silva, & Acero, 2007; Lin et al., 2011). Moreover, the use of subword units (for example, syllables or segments of them), as well as the pairing of words and subword units, for representing the spoken documents has also been proven beneficial for spoken document summarization (Chen, Yu, Wang, & Chen, 2006).

5.2.2. Summarization using sets of features

Building on the observations made on the above experimental results, in the next set of experiments, we attempt to group the simple (or raw) features, viz. the structural, lexical and acoustic features, together to make them more competitive (denoted by SET 1). The remaining four relevance features are also grouped together to form another feature set (denoted by SET 2). We take SVM as the vehicle to examine the utility of these two sets of features by respectively taking each set as the input (for characterizing a spoken sentence) to SVM, and the associated class-specific score output by SVM (cf. Section 2) is used for sentence ranking accordingly. Further, the proportions of summary sentences in a training spoken document being used are set in accordance with different ratios (viz. 10%, 20% and 30%) of all the sentences in the document. The corresponding results are presented in Table 4, in terms of ROUGE-1, ROUGE-2 and ROUGE-L measures. SVM appears to perform better when the numbers of labeled summary and non-summary sentences become more balanced (e.g., 30% summary labels), but its performance will degrade significantly when the numbers of labeled summary and non-summary sentences become more imbalanced (e.g., 10% summary labels). Meanwhile, it is interesting to mention that combining the structural, lexical and acoustic features together (SET 1) tends to provide more indicative cues than combining relevance features together (SET 2) for important sentence selection using SVM. As a final point, by consulting Table 4, we find that the marriage of these two sets of features (ALL) leads to substantial improvements than using each set of features separately. This evidence suggests that these two sets of features seem to be complementary to each other.

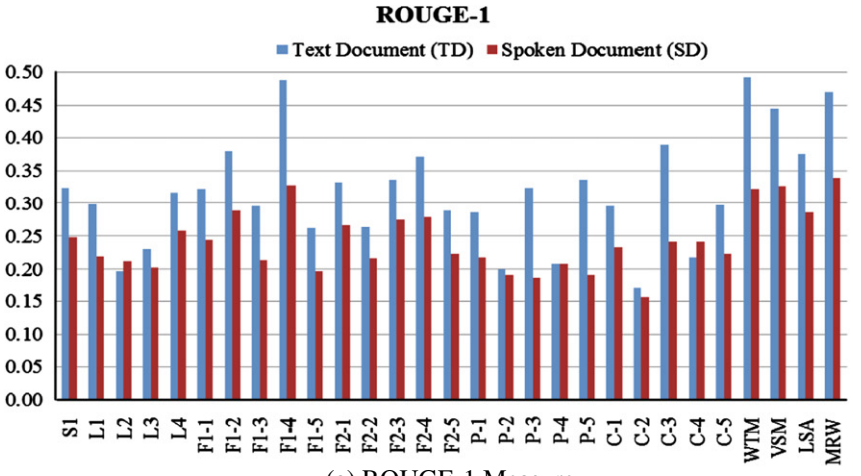
5.2.3. Summarization using evaluation metric-related training criteria

In the third set of experiments, we turn our attention on evaluating the utility of Ranking SVM, AdaRank and two variants of GCLM (viz. GCLM-I and GCLM-II), with respect to different feature sets and evaluation metrics being used. The results for

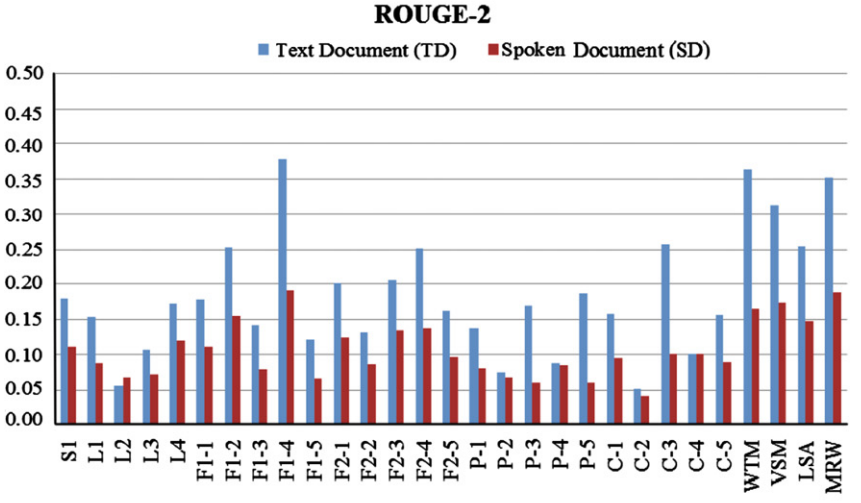
Table 3
The levels of agreement between the three subjects for important sentence ranking (10% summarization ratio) for the evaluation set.

	ROUGE-1	ROUGE-2	ROUGE-L
Agreement	0.675	0.645	0.631

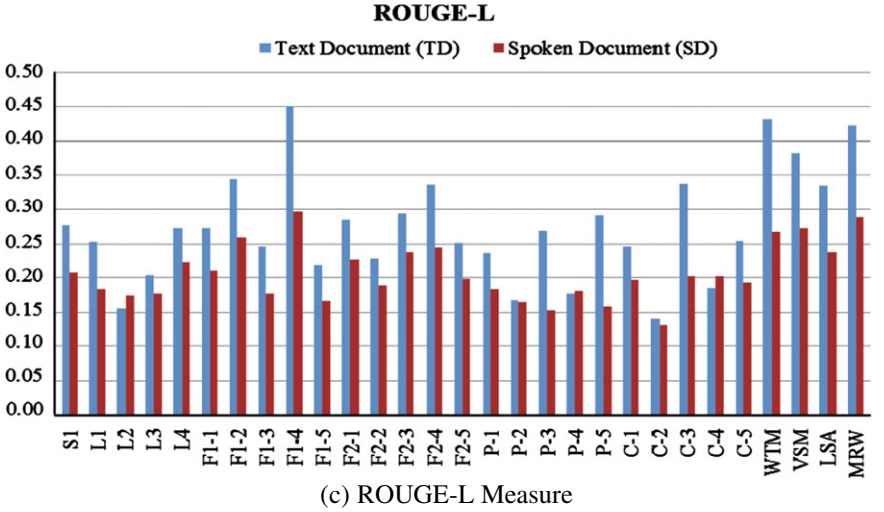
Please cite this article in press as: Chen, B., et al. Extractive speech summarization using evaluation metric-related training criteria. *Information Processing and Management* (2012), doi:10.1016/j.ipm.2011.12.002



(a) ROUGE-1 Measure



(b) ROUGE-2 Measure



(c) ROUGE-L Measure

Fig. 1. The summarization results achieved by using different single features.

Please cite this article in press as: Chen, B., et al. Extractive speech summarization using evaluation metric-related training criteria. *Information Processing and Management* (2012), doi:10.1016/j.ipm.2011.12.002

Table 4
The summarization results achieved by SVM with respect to different feature sets and different amounts of labeled data.

Features		10% Labels	20% Labels	30% Labels
(a) ROUGE-1 measure				
All	TD	0.611	0.676	0.672
	SD	0.427	0.468	0.490
SET 1	TD	0.584	0.599	0.638
	SD	0.376	0.431	0.455
SET 2	TD	0.413	0.454	0.528
	SD	0.346	0.391	0.412
(b) ROUGE-2 measure				
All	TD	0.500	0.587	0.590
	SD	0.269	0.309	0.332
SET 1	TD	0.474	0.546	0.531
	SD	0.228	0.276	0.299
SET 2	TD	0.256	0.399	0.436
	SD	0.180	0.225	0.246
(c) ROUGE-L measure				
All	TD	0.591	0.658	0.658
	SD	0.398	0.438	0.459
SET 1	TD	0.567	0.623	0.614
	SD	0.353	0.408	0.428
SET 2	TD	0.385	0.500	0.534
	SD	0.316	0.358	0.379

the SD case are shown in Table 5, in terms of ROUGE-1, ROUGE-2 and ROUGE-L measures (Lin, 2003); the corresponding results of SVM are also listed for comparison. Notice here that all these models are learned from the training spoken documents of the development set along with 10% summary labels and then tested on the spoken documents of the evaluation set. As can be seen, the two summarization models stemming from the IR community, viz. Ranking SVM and AdaRank, provide substantial improvements over SVM in the speech summarization task studied here, while AdaRank outperforms Ranking SVM when using all features or the features of SET 2. The values shown in the parentheses of Table 5 are the best results that can be achieved by AdaRank. The gaps between the actual and the best results are mainly due to that the final ranking model for AdaRank is optimized by using the development set rather than the evaluation set. Such performance mismatch (in ROUGE-1 for example) of AdaRank with all features, for the first ten training iterations, is also illustrated in Fig. 2. Further, we observe that GCLM-I (cf. (8)) and GCLM-II (cf. (9)) are quite comparable to each other and perform on par with AdaRank when using fewer features to represent the spoken sentences (viz. SET 1 or SET2). However, GCLM-I is substantially better than AdaRank and GCLM-II when more (all) features are being used (viz. ALL). This seems to confirm the merit of the GCLM-I training objective. GCLM-I aims not only to maximize the posterior probabilities of training summary sentences but also to emphasize the negative impact of the non-summary sentences that have higher posterior probabilities on the training objective, which can be interpreted as a kind of training data selection, viz. selecting (or focusing on) those training spoken documents that have more confusing non-summary sentences, for better model estimation and generalization.

Table 5
The summarization results for the SD case achieved by different supervised summarization approaches.

		ROUGE-1	ROUGE-2	ROUGE-L
All	SVM	0.427	0.269	0.398
	Ranking SVM	0.449	0.283	0.418
	AdaRank	0.459	0.303	0.432
		(0.462)	(0.303)	(0.432)
	GCLM-I	0.477	0.325	0.451
	GCLM-II	0.456	0.294	0.425
SET 1	SVM	0.376	0.228	0.353
	Ranking SVM	0.407	0.243	0.380
	AdaRank	0.378	0.237	0.362
		(0.409)	(0.237)	(0.409)
	GCLM-I	0.408	0.264	0.390
	GCLM-II	0.401	0.247	0.377
SET 2	SVM	0.346	0.180	0.316
	Ranking SVM	0.417	0.255	0.380
	AdaRank	0.438	0.273	0.403
		(0.438)	(0.273)	(0.403)
	GCLM-I	0.429	0.262	0.398
	GCLM-II	0.431	0.266	0.396

Please cite this article in press as: Chen, B., et al. Extractive speech summarization using evaluation metric-related training criteria. *Information Processing and Management* (2012), doi:10.1016/j.ipm.2011.12.002

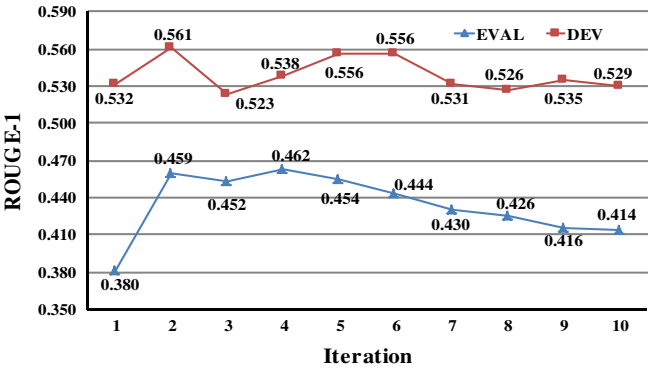


Fig. 2. Learning curves of AdaRank on the development set (DEV) and evaluation set (EVAL), respectively (for the SD case).

To recap, the superiority of the supervised summarizers (like SVM, Ranking SVM, AdaRank and GCLM) over the unsupervised summarizers (like VSM, LSA, WTM and MRW) stem from two factors. The first is that the supervised summarizers make use of the manually-annotated document-reference summary information for model training, whereas the unsupervised summarizers do not utilize such information. The second is that most of the unsupervised summarizers rely merely on lexical features (TF-IDF, word or topic unigrams, etc.), whereas the supervised summarizers integrate more indicative features besides the lexical features to realize spoken document summarization (Lin et al., 2009). On the other hand, the "*pairwise training*", "*direct optimization*" and "*discriminative training*" strategies turn out to show good promise for extractive speech summarization. They also have the side effect of mitigating the imbalanced-data problem as compared to the traditional SVM approach.

6. Conclusions

In this paper, we have investigated various kinds of summarization features and training criteria for training a speech summarizer; the evaluation metric-related training criteria not only can deal with the imbalanced-data problem but also can boost the summarizer's performance by maximizing the associated evaluation score or optimizing an objective that is linked to the ultimate evaluation. The experimental results indeed justify our expectation. Our future research directions include: (1) investigating more elaborate acoustic features that can be used for speech summarization, (2) seeking other alternative approaches to optimizing a summarizer's performance (Chen and Lin, 2012), (3) exploring better ways to represent the recognition hypotheses of spoken documents beyond the top scoring ones (Lin et al., 2011), (4) extending and applying the proposed model training paradigms to multi-document summarization tasks, and (5) incorporating the summarization results into audio indexing for better retrieval and browsing of spoken documents. Additionally, how to make effective use of semi-supervised (or even unsupervised) learning to improve the performance of supervised summarizers without recourse to manual annotation and specialized linguistic expertise might also be an important issue for spoken document summarization.

Acknowledgements

This work was sponsored in part by "Aim for the Top University Plan" of National Taiwan Normal University and Ministry of Education, Taiwan, and the National Science Council, Taiwan, under Grants NSC 99-2221-E-003-017-MY3, NSC 98-2221-E-003-011-MY3, NSC 100-2515-S-003-003, and NSC 99-2631-S-003-002.

References

Baxendale, P. (1958). Machine-made index for technical literature – an experiment. *IBM Journal of Research and Development*, 354–361.
Burges, C. J. C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., et al. (2005). Learning to rank using gradient descent. In *Proc. the international conference on machine learning* (pp. 89–96).
Cao, Y., Xu, J., Liu, T. Y., Li, H., Huang, Y., Hon, H. W. (2006). Adapting ranking SVM to document retrieval. In *Proc. the annual international ACM SIGIR conference on research and development in information retrieval* (pp. 186–193).
Chang, C. C., Lin, C. J. (2001). LIBSVM: A Library for Support Vector Machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
Chelba, C., Silva, J., & Acero, A. (2007). Soft indexing of speech content for search in spoken documents. *Computer Speech and Language, 21*, 458–478.
Chen, B. (2009). Word topic models for spoken document retrieval and transcription. *ACM Transactions on Asian Language Information Processing, 8*(1), 2:1–2:27.
Chen, Y. T., Chen, B., & Wang, H. M. (2009). A probabilistic generative framework for extractive broadcast news speech summarization. *IEEE Transactions on Audio, Speech and Language Processing, 17*, 95–106.
Chen, B., & Lin, S. -H. (2012). A risk-aware modeling framework for speech summarization. *IEEE Transactions on Audio, Speech and Language Processing, 20*(1), 199–210.

Please cite this article in press as: Chen, B., et al. Extractive speech summarization using evaluation metric-related training criteria. *Information Processing and Management* (2012), doi:10.1016/j.ipm.2011.12.002

Chen, Y. T., Yu, S., Wang, H. M., Chen, B. (2006). Extractive Chinese spoken document summarization using probabilistic ranking models. In *Proc. the international symposium on Chinese spoken language processing* (pp. 660–671).

Christensen, H., Gotoh, Y., & Renals, S. (2008). A cascaded broadcast news highlighter. *IEEE Transactions on Audio, Speech and Language Processing, 16*, 151–161.

Cossock, D., Zhang, T. 2006. Subset ranking using regression. In *proc. on learning theory* (pp. 605–619).

DauméIII, H., Marcu, D. (2006). Bayesian query focused summarization. In *Proc. annual meeting of the association for computational linguistics* (pp. 305–312).

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research, 22*, 457–479.

Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research, 4*, 933–969.

Furui, S., Kikuchi, T., Shinnaka, Y., & Hori, C. (2004). Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing, 12*, 401–408.

Gong Y., & Liu, X. (2010). Generic text summarization using relevance measure and latent semantic analysis. In *Proc. the Annual International ACM SIGIR conference on research and development in information retrieval* (pp. 19–25).

Herbrich, R., Graepel, T., Obermayer, K. (2000). *Large margin rank boundaries for ordinal regression. Advances in large margin classifier*, MIT Press.

Joachims, T. (2002). Optimizing search engines using clickthrough data, In *proc. the international conference on knowledge discovery and data mining* (pp. 133–142).

Joachims, T. (2005). A support vector method for multivariate performance measures. In *Proc. the international conference on machine learning* (pp. 377–384).

Kupiec, J., Pedersen, J., & Chen, F. (1999). A trainable document summarizer. In *Proc. the annual international ACM SIGIR conference on research and development in information retrieva*l (pp. 68–73).

Lee, L. S., & Chen, B. (2005). Spoken document understanding and organization. *IEEE Signal Processing Magazine, 22*, 42–60.

Lin, C. Y. (2003). ROUGE: Recall-oriented understudy for gisting evaluation. <http://haydn.isi.edu/ROUGE/>.

Lin, S. H., Chang, Y. M., Liu, J. W., & Chen, B. (2010). Leveraging evaluation metric-related training criteria for speech summarization. In *Proc. IEEE international conference on acoustics, speech, and signal processing* (pp. 5314–5317).

Lin, S. -H., Chen, B., & Wang, H. -M. (2009). A comparative study of probabilistic ranking models for Chinese spoken document summarization. *ACM Transactions on Asian Language Information Processing, 8*(1), 3:1–3:23.

Lin, S. H., Yeh, Y. M., & Chen, B. (2011). Leveraging Kullback-Leibler divergence measures and information-rich cues for speech summarization. *IEEE Transactions on Audio, Speech and Language Processing, 19*(4), 871–882.

Liu, S. H., Chu, F. H., Lin, S. H., Lee, H. S., & Chen, B. (2007). Training data selection for improving discriminative training of acoustic models. In *Proc. IEEE workshop on automatic speech recognition and understanding* (pp. 284–289).

Liu, F., & Liu, Y. (2010). Exploring correlation between ROUGE and human evaluation on meeting summaries. *IEEE Transactions on Audio, Speech and Language Processing, 18*(1), 187–196.

Maloof, M. (2003). Learning when data sets are imbalanced and when costs are unequal and unknown. In *Proc. the ICML'03 workshop on learning from imbalanced data sets*.

Mani, I., & Maybury, M. T. (1999). *Advances in automatic text summarization*. Cambridge: MIT Press.

McKeown, K., Hirschberg, J., Galley, M., & Maskey, S. (2005). From text to speech summarization. In *Proc. IEEE international conference on acoustics, speech, and signal processing* (pp. 997–1000).

Mihalcea, R., & Tarau, P. (2005). TextRank: Bringing order into texts. In *Proc. conference on empirical methods in natural language processing* (pp. 404–411).

Nenkova, A., Vanderwende, L., & McKeown, K. (2006). A compositional context sensitive multi-document summarizer: Exploring the factors that influence summarization. In *Proc. annual international ACM SIGIR conference on research and development in information retrieval* (pp. 573–580).

Oba, T., Hori, T, & Nakamura, A. (2010). A comparative study on methods of weighted language model training for reranking LVCSR *n*-best hypotheses. In *Proc. international conference on acoustics, speech and signal processing* (pp. 5126–5129).

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proc. annual meeting of the association for computational linguistics* (pp. 160–167).

Ostendorf, M. (2008). Speech technology and information access. *IEEE Signal Processing Magazine, 25*(3), 150–152.

Ouyang, Y., Li, W., Li, S., & Lu, Q. (2011). Applying regression models to query-focused multi-document summarization. *Information Processing & Management, 47*(2), 227–237.

Penn, G., & Zhu, X. (2008). A critical reassessment of evaluation baselines for speech summarization. In *Proc. the annual meeting of the association for computational linguistics* (pp. 470–478).

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine, 21*, 45.

Povey, D., & Woodland, P.C. (2002). Minimum phone error and I-smoothing for improved discriminative training. In *Proc. IEEE international conference on acoustics, speech, and signal processing* (pp. 105–108).

Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., & Visweswariah, K. (2008). Boosted MMI for model and feature-space discriminative training. In *Proc. IEEE international conference on acoustics, speech, and signal processing* (pp. 4057–4060).

Roark, B., Saraclar, M., & Collins, M. (2007). Discriminative *n*-gram language modeling. *Computer Speech and Language, 21*, 373–392.

Shen, D., Sun, J. T., Li, H., Yang, Q., & Chen Z. (2007). Document summarization using conditional random fields. In *Proc. international joint conference on artificial intelligence* (pp. 2862–2867).

Stolcke, A. (2005). SRILM – an extensible language modeling toolkit. In *Proc. the annual conference of the international speech communication association* (pp. 901–904).

Wan, X., & Yang, J. (2008). Multi-document summarization using cluster-based link analysis. In *Proc. the annual international ACM SIGIR conference on research and development in information retrieval* (pp. 299–306).

Wang, H. M., Chen, B., Kuo, J. W., & Cheng, S. S. (2005). MATBN: A Mandarin Chinese broadcast news corpus. *International Journal of Computational Linguistics and Chinese Language Processing, 10*, 219–236.

Xie, S., & Liu, Y. (2010). Improving supervised learning for meeting summarization using sampling and regression. *Computer Speech & Language, 24*(3), 495–514.

Xu, J., & Li, H. (2007). AdaRank: A boosting algorithm for information retrieval. In *Proc. the annual international ACM SIGIR conference on research and development in information retrieval* (pp. 391–398).

Zhang, J., Chan, H. Y., Fung, P., & Cuo, L. (2007). A comparative study on speech summarization of broadcast news and lecture Speech. In *Proc. the annual conference of the international speech communication association* (pp. 2781–2784).

Please cite this article in press as: Chen, B., et al. Extractive speech summarization using evaluation metric-related training criteria. *Information Processing and Management* (2012), doi:10.1016/j.ipm.2011.12.002