A Risk Minimization Framework for Extractive Speech Summarization

Shih-Hsiang Lin and Berlin Chen
National Taiwan Normal University
Taipei, Taiwan
{shlin, berlin}@csie.ntnu.edu.tw

Abstract

In this paper, we formulate extractive summarization as a risk minimization problem and propose a unified probabilistic framework that naturally combines supervised and unsupervised summarization models to inherit their individual merits as well as to overcome their inherent limitations. In addition, the introduction of various loss functions also provides the summarization framework with a flexible but systematic way to render the redundancy and coherence relationships among sentences and between sentences and the whole document, respectively. Experiments on speech summarization show that the methods deduced from our framework are very competitive with existing summarization approaches.

1 Introduction

Automated summarization systems which enable user to quickly digest the important information conveyed by either a single or a cluster of documents are indispensible for managing the rapidly growing amount of textual information and multimedia content (Mani and Maybury, 1999). On the other hand, due to the maturity of text summarization, the research paradigm has been extended to speech summarization over the years (Furui et al., 2004; McKeown et al., 2005). Speech summarization is expected to distill important information and remove redundant and incorrect information caused by recognition errors from spoken documents, enabling user to efficiently review spoken documents and understand the associated topics quickly. It would also be useful for improving the efficiency of a number of potential applications like retrieval and mining of large volumes of spoken documents.

A summary can be either abstractive or extractive. In abstractive summarization, a fluent and concise abstract that reflects the key concepts of a document is generated, whereas in extractive summarization, the summary is usually formed by selecting salient sentences from the original document (Mani and Maybury, 1999). The former requires highly sophisticated natural language processing techniques, including semantic representation and inference, as well as natural language generation, while this would make abstractive approaches difficult to replicate or extend from constrained domains to more general domains. In addition to being extractive or abstractive, a summary may also be generated by considering several other aspects like being generic or query-oriented summarization, single-document or multi-document summarization, and so forth. The readers may refer to (Mani and Maybury, 1999) for a comprehensive overview of automatic text summarization. In this paper, we focus exclusively on generic, single-document extractive summarization which forms the building block for many other summarization tasks.

Aside from traditional ad-hoc extractive summarization methods (Mani and Maybury, 1999), machine-learning approaches with either supervised or unsupervised learning strategies have gained much attention and been applied with empirical success to many summarization tasks (Kupiec et al., 1999; Lin et al., 2009). For supervised learning strategies, the summarization task is usually cast as a two-class (summary and non-summary) sentence-classification problem: A sentence with a set of indicative features is input to the classifier (or summarizer) and a decision is then returned from it on the basis of these features. In general, they usually require a training set, comprised of several documents and their corresponding handcrafted summaries (or labeled data), to train the classifiers. However, manual labeling is expensive in terms of time and personnel. The other potential problem is the so-called "*bag-of-sentences*" assumption implicitly made by most of these summarizers. That is, sentences are classified independently of each other,

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 79–87,
Uppsala, Sweden, 11-16 July 2010.

without leveraging the dependence relationships among the sentences or the global structure of the document (Shen et al., 2007).

Another line of thought attempts to conduct document summarization using unsupervised machine-learning approaches, getting around the need for manually labeled training data. Most previous studies conducted along this line have their roots in the concept of sentence *centrality* (Gong and Liu, 2001; Erkan and Radev, 2004; Radev et al., 2004; Mihalcea and Tarau, 2005). Put simply, sentences more similar to others are deemed more salient to the main theme of the document; such sentences thus will be selected as part of the summary. Even though the performance of unsupervised summarizers is usually worse than that of supervised summarizers, their domain-independent and easy-to-implement properties still make them attractive.

Building on these observations, we expect that researches conducted along the above-mentioned two directions could complement each other, and it might be possible to inherit their individual merits to overcome their inherent limitations. In this paper, we present a probabilistic summarization framework stemming from Bayes decision theory (Berger, 1985) for speech summarization. This framework can not only naturally integrate the above-mentioned two modeling paradigms but also provide a flexible yet systematic way to render the redundancy and coherence relationships among sentences and between sentences and the whole document, respectively. Moreover, we also illustrate how the proposed framework can unify several existing summarization models.

The remainder of this paper is structured as follows. We start by reviewing related work on extractive summarization. In Section 3 we formulate the extractive summarization task as a risk minimization problem, followed by a detailed elucidation of the proposed methods in Section 4. Then, the experimental setup and a series of experiments and associated discussions are presented in Sections 5 and 6, respectively. Finally, Section 7 concludes our presentation and discusses avenues for future work.

2 Background

Speech summarization can be conducted using either supervised or unsupervised methods (Furui et al., 2004, McKeown et al., 2005, Lin et al., 2008). In the following, we briefly review a few celebrated methods that have been applied to extractive speech summarization tasks with good success.

2.1 Supervised summarizers

Extractive speech summarization can be treated as a two-class (positive/negative) classification problem. A spoken sentence S_i is characterized by set of T indicative features $X_i = \{x_{i1}, \cdots, x_{iT}\}$, and they may include lexical features (Koumpis and Renals, 2000), structural features (Maskey and Hirschberg, 2003), acoustic features (Inoue et al., 2004), discourse features (Zhang et al., 2007) and relevance features (Lin et al., 2009). Then, the corresponding feature vector X_i of S_i is taken as the input to the classifier. If the output (classification) score belongs to the positive class, S_i will be selected as part of the summary; otherwise, it will be excluded (Kupiec et al., 1999). Specifically, the problem can be formulated as follows: Construct a sentence ranking model that assigns a classification score (or a posterior probability) of being in the summary class to each sentence of a spoken document to be summarized; important sentences are subsequently ranked and selected according to these scores. To this end, several popular machine-learning methods could be utilized, like Bayesian classifier (BC) (Kupiec et al., 1999), Gaussian mixture model (GMM) (Fattah and Ren, 2009) , hidden Markov model (HMM) (Conroy and O'leary, 2001), support vector machine (SVM) (Kolcz et al., 2001), maximum entropy (ME) (Ferrier, 2001), conditional random field (CRF) (Galley, 2006; Shen et al., 2007), to name a few.

Although such supervised summarizers are effective, most of them (except CRF) usually implicitly assume that sentences are independent of each other (the so-called "*bag-of-sentences*" assumption) and classify each sentence individually without leveraging the relationship among the sentences (Shen et al., 2007). Another major shortcoming of these summarizers is that a set of handcrafted document-reference summary exemplars are required for training the summarizers; however, such summarizers tend to limit their generalization capability and might not be readily applicable for new tasks or domains.

2.2 Unsupervised summarizers

The related work conducted along this direction usually relies on some heuristic rules or statistical evidences between each sentence and the document, avoiding the need of manually labeled training data. For example, the vector space model (VSM) approach represents each sentence of a document and the document itself in vector space (Gong and Liu, 2001), and computes the relevance score between each sentence and the document (e.g., the cosine measure of the simi-

larity between two vectors). Then, the sentences with the highest relevance scores are included in the summary. A natural extension is to represent each document or each sentence vector in a latent semantic space (Gong and Liu, 2001), instead of simply using the literal term information as that done by VSM.

On the other hand, the graph-based methods, such as TextRank (Mihalcea and Tarau, 2005) and LexRank (Erkan and Radev, 2004), 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 or topical similarity relationship between a pair of nodes. Document summarization thus relies on the global structural information conveyed by such conceptualized network, rather than merely considering the local features of each node (sentence).

However, due to the lack of document-summary reference pairs, the performance of the unsupervised summarizers is usually worse than that of the supervised summarizers. Moreover, most of the unsupervised summarizers are constructed solely on the basis of the lexical information without considering other sources of information cues like discourse features, acoustic features, and so forth.

3 A risk minimization framework for extractive summarization

Extractive summarization can be viewed as a decision making process in which the summarizer attempts to select a representative subset of sentences or paragraphs from the original documents. Among the several analytical methods that can be employed for the decision process, the Bayes decision theory, which quantifies the tradeoff between various decisions and the potential cost that accompanies each decision, is perhaps the most suited one that can be used to guide the summarizer in choosing a course of action in the face of some uncertainties underlying the decision process (Berger, 1985). Stated formally, a decision problem may consist of four basic elements: 1) an observation O from a random variable $\mathbf{O}$, 2) a set of possible decisions (or actions) $a \in \mathbf{A}$, 3) the state of nature $\theta \in \mathbf{\Theta}$, and 4) a loss function $L(a_i, \theta)$ which specifies the cost associated with a chosen decision a_i given that θ is the true state of nature. The expected risk (or conditional risk) associated with taking decision a_i is given by

$$R(a_i \mid O) = \int_\theta L(a_i, \theta) p(\theta \mid O) d\theta, \quad (1)$$

where $p(\theta|O)$ is the posterior probability of the state of nature being θ given the observation O. Bayes decision theory states that the optimum decision can be made by contemplating each action a_i, and then choosing the action for which the expected risk is minimum:

$$a^* = \arg\min_{a_i} R(a_i \mid O). \quad (2)$$

The notion of minimizing the Bayes risk has gained much attention and been applied with success to many natural language processing (NLP) tasks, such as automatic speech recognition (Goel and Byrne, 2000), statistical machine translation (Kumar and Byrne, 2004) and statistical information retrieval (Zhai and Lafferty, 2006). Following the same spirit, we formulate the extractive summarization task as a Bayes risk minimization problem. Without loss of generality, let us denote $\pi \in \mathbf{\Pi}$ as one of possible selection strategies (or state of nature) which comprises a set of indicators used to address the importance of each sentence S_i in a document D to be summarized. A feasible selection strategy can be fairly arbitrary according to the underlying principle. For example, it could be a set of binary indicators denoting whether a sentence should be selected as part of summary or not. On the contrary, it may also be a ranked list used to address the significance of each individual sentence. Moreover, we refer to the k-th action a_k as choosing the k-th selection strategy π_k, and the observation O as the document D to be summarized. As a result, the expected risk of a certain selection strategy π_k is given by

$$R(\pi_k \mid D) = \int_\pi L(\pi_k, \pi) p(\pi \mid D) d\pi. \quad (3)$$

Consequently, the ultimate goal of extractive summarization could be stated as the search of the best selection strategy from the space of all possible selection strategies that minimizes the expected risk defined as follows:

$$\begin{aligned} \pi^* &= \arg\min_{\pi_k} R(\pi_k \mid D) \\ &= \arg\min_{\pi_k} \int_\pi L(\pi_k, \pi) p(\pi \mid D) d\pi. \end{aligned} \quad (4)$$

Although we have described a general formulation for the extractive summarization problem on the grounds of the Bayes decision theory, we consider hereafter a special case of it where the selection strategy is represented by a binary decision vector, of which each element corresponds to a specific sentence S_i in the document D and designates whether it should be selected as part of the summary or not, as the first such attempt. More concretely, we assume that the summary

sentences of a given document can be iteratively chosen (i.e., one at each iteration) from the document until the aggregated summary reaches a predefined target summarization ratio. It turns out that the binary vector for each possible action will have just one element equal to 1 and all others equal to zero (or the so-called "*one-of-n*" coding). For ease of notation, we denote the binary vector by S_i when the i-th element has a value of 1. Therefore, the risk minimization framework can be reduced to

$$\begin{aligned} S^* &= \arg\min_{S_i \in \widetilde{D}} R(S_i \mid \widetilde{D}) \\ &= \arg\min_{S_i \in \widetilde{D}} \sum_{S_j \in \widetilde{D}} L(S_i, S_j) P(S_j \mid \widetilde{D}), \end{aligned} \tag{5}$$

where $\widetilde{D}$ denotes the remaining sentences that have not been selected into the summary yet (i.e., the "*residual*" document); $P(S_j \mid \widetilde{D})$ is the posterior probability of a sentence S_j given $\widetilde{D}$. According to the Bayes' rule, we can further express $P(S_j \mid \widetilde{D})$ as (Chen et al., 2009)

$$P(S_j \mid \widetilde{D}) = \frac{P(\widetilde{D} \mid S_j) P(S_j)}{P(\widetilde{D})}, \tag{6}$$

where $P(\widetilde{D} \mid S_j)$ is the sentence generative probability, i.e., the likelihood of $\widetilde{D}$ being generated by S_j; $P(S_j)$ is the prior probability of S_j being important; and the evidence $P(\widetilde{D})$ is the marginal probability of $\widetilde{D}$, which can be approximated by

$$P(\widetilde{D}) \approx \sum_{S_m \in \widetilde{D}} P(\widetilde{D} \mid S_m) P(S_m). \tag{7}$$

By substituting (6) and (7) into (5), we obtain the following final selection strategy for extractive summarization:

$$S^* = \arg\min_{S_i \in \widetilde{D}} \sum_{S_j \in \widetilde{D}} L(S_i, S_j) \frac{P(\widetilde{D} \mid S_j) P(S_j)}{\sum_{S_m \in \widetilde{D}} P(\widetilde{D} \mid S_m) P(S_m)}. \tag{8}$$

A remarkable feature of this framework lies in that a sentence to be considered as part of the summary is actually evaluated by three different fundamental factors: (1) $P(S_j)$ is the sentence prior probability that addresses the importance of sentence S_j itself; (2) $P(\widetilde{D} \mid S_j)$ is the sentence generative probability that captures the degree of relevance of S_j to the residual document $\widetilde{D}$; and (3) $L(S_i, S_j)$ is the loss function that characterizes the relationship between sentence S_i and any other sentence S_j. As we will soon see, such a framework can be regarded as a generalization of several existing summarization methods. A detailed account on the construction of these three component models in the framework will be given in the following section.

4 Proposed Methods

There are many ways to construct the above mentioned three component models, i.e., the sentence generative model $P(\widetilde{D} \mid S_j)$, the sentence prior model $P(S_j)$, and the loss function $L(S_i, S_j)$. In what follows, we will shed light on one possible attempt that can accomplish this goal elegantly.

4.1 Sentence generative model

In order to estimate the sentence generative probability, we explore the language modeling (LM) approach, which has been introduced to a wide spectrum of IR tasks and demonstrated with good empirical success, to predict the sentence generative probability. In the LM approach, each sentence in a document can be simply regarded as a probabilistic generative model consisting of a unigram distribution (the so-called "*bag-of-words*" assumption) for generating the document (Chen et al., 2009):

$$P(\widetilde{D} \mid S_j) = \prod_{w \in \widetilde{D}} P(w \mid S_j)^{c(w, \widetilde{D})}, \tag{9}$$

where $c(w, \widetilde{D})$ is the number of times that index term (or word) w occurs in $\widetilde{D}$, reflecting that w will contribute more in the calculation of $P(\widetilde{D} \mid S_j)$ if it occurs more frequently in $\widetilde{D}$. Note that the sentence model $P(w \mid S_j)$ is simply estimated on the basis of the frequency of index term w occurring in the sentence S_j with the maximum likelihood (ML) criterion. In a sense, (9) belongs to a kind of literal term matching strategy (Chen, 2009) and may suffer the problem of unreliable model estimation owing particularly to only a few sampled index terms present in the sentence (Zhai, 2008). To mitigate this potential defect, a unigram probability estimated from a general collection, which models the general distribution of words in the target language, is often used to smooth the sentence model. Interested readers may refer to (Zhai, 2008; Chen et al., 2009) for a thorough discussion on various ways to construct the sentence generative model.

4.2 Sentence prior model

The sentence prior probability $P(S_j)$ can be regarded as the likelihood of a sentence being important without seeing the whole document. It could be assumed uniformly distributed over sentences or estimated from a wide variety of factors, such as the lexical information, the structural information or the inherent prosodic properties of a spoken sentence.

A straightforward way is to assume that the sentence prior probability $P(S_j)$ is in proportion to the posterior probability of a sentence S_j be-

ing included in the summary class when observing a set of indicative features X_j of S_j derived from such factors or other sentence importance measures (Kupiec et al., 1999). These features can be integrated in a systematic way into the proposed framework by taking the advantage of the learning capability of the supervised machine-learning methods. Specifically, the prior probability $P(S_j)$ can be approximated by:

$$P(S_j) \approx \frac{p(X_j \mid \mathbf{S})P(\mathbf{S})}{P(X_j \mid \mathbf{S})P(\mathbf{S}) + P(X_j \mid \overline{\mathbf{S}})P(\overline{\mathbf{S}})}, \quad (10)$$

where $P(X_j \mid \mathbf{S})$ and $P(X_j \mid \overline{\mathbf{S}})$ are the likelihoods that a sentence S_j with features X_j are generated by the summary class $\mathbf{S}$ and the non-summary class $\overline{\mathbf{S}}$, respectively; the prior probability $P(\mathbf{S})$ and $P(\overline{\mathbf{S}})$ are set to be equal in this research. To estimate $P(X_j \mid \mathbf{S})$ and $P(X_j \mid \overline{\mathbf{S}})$, several popular supervised classifiers (or summarizers), like BC or SVM, can be leveraged for this purpose.

4.3 Loss function

The loss function introduced in the proposed summarization framework is to measure the relationship between any pair of sentences. Intuitively, when a given sentence is more dissimilar from most of the other sentences, it may incur higher loss as it is taken as the representative sentence (or summary sentence) to represent the main theme embedded in the other ones. Consequently, the loss function can be built on the notion of the similarity measure. In this research, we adopt the cosine measure (Gong and Liu, 2001) to fulfill this goal. We first represent each sentence S_i in vector form where each dimension specifies the weighted statistic $z_{t,i}$, e.g., the product of the term frequency (TF) and inverse document frequency (IDF) scores, associated with an index term w_t in sentence S_i. Then, the cosine similarity between any given two sentences (S_i, S_j) is

$$Sim(S_i, S_j) = \frac{\sum_{t=1}^{T} z_{t,i} \times z_{t,j}}{\sqrt{\sum_{t=1}^{T} z_{t,i}^2} \times \sqrt{\sum_{t=1}^{T} z_{t,j}^2}}. \quad (10)$$

The loss function is thus defined by

$$L(S_i, S_j) = 1 - Sim(S_i, S_j). \quad (11)$$

Once the sentence generative model $P(\widetilde{D} \mid S_j)$, the sentence prior model $P(S_j)$ and the loss function $L(S_i, S_j)$ have been properly estimated, the summary sentences can be selected iteratively by (8) according to a predefined target summarization ratio. However, as can be seen from (8), a new summary sentence is selected without considering the redundant information that is also contained in the already selected summary sentences. To alleviate this problem, the concept of maximum marginal relevance (MMR) (Carbonell and Goldstein, 1998), which performs sentence selection iteratively by striking the balance between topic relevance and coverage, can be incorporated into the loss function:

$$L(S_i, S_j) = 1 - \left[\begin{array}{l} \beta \cdot Sim(S_i, S_j) \\ -(1-\beta) \cdot \max_{S' \in \mathbf{Summ}} Sim(S_i, S') \end{array} \right], \quad (12)$$

where **Summ** represents the set of sentences that have already been included into the summary and the novelty factor β is used to trade off between relevance and redundancy.

4.4 Relation to other summarization models

In this subsection, we briefly illustrate the relationship between our proposed summarization framework and a few existing summarization approaches. We start by considering a special case where a 0-1 loss function is used in (8), namely, the loss function will take value 0 if the two sentences are identical, and 1 otherwise. Then, (8) can be alternatively represented by

$$\begin{aligned} S^* &= \arg\min_{S_i \in \widetilde{D}} \sum_{S_j \in \widetilde{D}, S_j \neq S_i} \frac{P(\widetilde{D} \mid S_j)P(S_j)}{\sum_{S_m \in \widetilde{D}} P(\widetilde{D} \mid S_m)P(S_m)} \\ &= \arg\max_{S_i \in \widetilde{D}} \frac{P(\widetilde{D} \mid S_i)P(S_i)}{\sum_{S_m \in \widetilde{D}} P(\widetilde{D} \mid S_m)P(S_m)}, \end{aligned} \quad (13)$$

which actually provides a natural integration of the supervised and unsupervised summarizers (Lin et al., 2009), as mentioned previously.

If we further assume the prior probability $P(S_j)$ is uniformly distributed, the important (or summary) sentence selection problem has now been reduced to the problem of measuring the document-likelihood $P(\widetilde{D} \mid S_j)$, or the relevance between the document and the sentence. Alone a similar vein, the important sentences of a document can be selected (or ranked) solely based on the prior probability $P(S_j)$ with the assumption of an equal document-likelihood $P(\widetilde{D} \mid S_j)$.

5 Experimental setup

5.1 Data

The summarization dataset used in this research is a widely used broadcast news corpus collected by the Academia Sinica and the Public Television Service Foundation of Taiwan between November 2001 and April 2003 (Wang et al., 2005). Each story contains the speech of one studio anchor, as well as several field reporters and interviewees. A subset of 205 broadcast news doc-

Kappa	ROGUE-1	ROUGE-2	ROUGE-L
0.400	0.600	0.532	0.527

Table 1: The agreement among the subjects for important sentence ranking for the evaluation set.

Structural features	1.Duration of the current sentence 2.Position of the current sentence 3.Length of the current sentence
Lexical Features	1.Number of named entities 2.Number of stop words 3.Bigram language model scores 4.Normalized bigram scores
Acoustic Features	1.The 1st formant 2.The 2nd formant 3.The pitch value 4.The peak normalized cross-correlation of pitch
Relevance Feature	1.VSM score

Table 2: Basic sentence features used by BC.

uments compiled between November 2001 and August 2002 was reserved 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 generated by ranking the sentences in the reference transcript of a spoken document by importance without assigning a score to each sentence. The average Chinese character error rate (CER) obtained for the 205 spoken documents was about 35%.

Since broadcast news stories often follow a relatively regular structure as compared to other speech materials like conversations, the positional information would play an important (dominant) role in extractive summarization of broadcast news stories; we, hence, chose 20 documents for which the generation of reference summaries is less correlated with the positional information (or the position of sentences) as the held-out test set to evaluate the general performance of the proposed summarization framework, and 100 documents as the development set.

5.2 Performance evaluation

For the assessment of summarization performance, we adopted the widely used ROUGE measure (Lin, 2004) because of its higher correlation with human judgments. It 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 summaries. Three variants of the ROGUE measure were used to quantify the utility of the proposed method. They are, respectively, the ROUGE-1 (unigram) measure, the ROUGE-2 (bigram) measure and the ROUGE-L (longest common subsequence) measure (Lin, 2004).

The summarization ratio, defined as the ratio of the number of words in the automatic (or manual) summary to that in the reference transcript of a spoken document, was set to 10% in this research. Since increasing the summary length tends to increase the chance of getting higher scores in the recall rate of the various ROUGE measures and might not always select the right number of informative words in the automatic summary as compared to the reference summary, all the experimental results reported hereafter are obtained by calculating the F-scores of these ROUGE measures, respectively (Lin, 2004). Table 1 shows the levels of agreement (the Kappa statistic and ROUGE measures) between the three subjects for important sentence ranking. They seem to reflect the fact that people may not always agree with each other in selecting the important sentences for representing a given document.

5.3 Features for supervised summarizers

We take BC as the representative supervised summarizer to study in this paper. The input to BC consists of a set of 28 indicative features used to characterize a spoken sentence, including the structural features, the lexical features, the acoustic features and the relevance feature. For each kind of acoustic features, the minimum, maximum, mean, difference value and mean difference value 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. Finally, the relevance feature (VSM score) is use to measure the degree of relevance for a sentence to the whole document (Gong and Liu, 2001). These features are outlined in Table 2, where each of them was further normalized to zero mean and unit variance.

6 Experimental results and discussions

6.1 Baseline experiments

In the first set of experiments, we evaluate the baseline performance of the LM and BC summarizers (cf. Sections 4.1 and 4.2), respectively. The corresponding results are detailed in Table 3,

	Text Document (TD)			Spoken Document (SD)		
	ROGUE-1	ROUGE-2	ROUGE-L	ROGUE-1	ROUGE-2	ROUGE-L
BC	0.445 (0.390 - 0.504)	0.346 (0.201 - 0.415)	0.404 (0.348 - 0.468)	0.369 (0.316 - 0.426)	0.241 (0.183 - 0.302)	0.321 (0.268 - 0.378)
LM	0.387 (0.302 - 0.474)	0.264 (0.168 - 0.366)	0.334 (0.251 - 0.415)	0.319 (0.274 - 0.367)	0.164 (0.115 - 0.224)	0.253 (0.215 - 0.301)

Table 3: The results achieved by the BC and LM summarizers, respectively.

		Text Document (TD)			Spoken Document (SD)		
Prior	Loss	ROGUE-1	ROUGE-2	ROUGE-L	ROGUE-1	ROUGE-2	ROUGE-L
BC	0-1	0.501	0.401	0.459	0.417	0.281	0.356
	SIM	0.524	0.425	0.473	0.475	0.351	0.420
	MMR	0.529	0.426	0.479	0.475	0.351	0.420
Uniform	SIM	0.405	0.281	0.348	0.365	0.209	0.305
	MMR	0.417	0.282	0.359	0.391	0.236	0.338

Table 4: The results achieved by several methods derived from the proposed summarization framework.

where the values in the parentheses are the associated 95% confidence intervals. It is also worth mentioning that TD denotes the summarization results obtained based on manual transcripts of the spoken documents while SD denotes the results using the speech recognition transcripts which may contain speech recognition errors and sentence boundary detection errors. In this research, sentence boundaries were determined by speech pauses. For the TD case, the acoustic features were obtained by aligning the manual transcripts to their spoken documents counterpart by performing word-level forced alignment.

Furthermore, the ROGUE measures, in essence, are evaluated by counting the number of overlapping units between the automatic summary and the reference summary; the corresponding evaluation results, therefore, would be severely affected by speech recognition errors when applying the various ROUGE measures to quantify the performance of speech summarization. In order to get rid of the cofounding effect of this factor, it is assumed that the selected summary sentences can also be presented in speech form (besides text form) such that users can directly listen to the audio segments of the summary sentences to bypass the problem caused by speech recognition errors. Consequently, we can align the ASR transcripts of the summary sentences to their respective audio segments to obtain the correct (manual) transcripts for the summarization performance evaluation (i.e., for the SD case).

Observing Table 3 we notice two particularities. First, there are significant performance gaps between summarization using the manual transcripts and the erroneous speech recognition transcripts. The relative performance degradations are about 15%, 34% and 23%, respectively, for ROUGE-1, ROUGE2 and ROUGE-L measures. One possible explanation is that the erroneous speech recognition transcripts of spoken sentences would probably carry wrong information and thus deviate somewhat from representing the true theme of the spoken document. Second, the supervised summarizer (i.e., BC) outperforms the unsupervised summarizer (i.e., LM). The better performance of BC can be further explained by two reasons. One is that BC is trained with the handcrafted document-summary sentence labels in the development set while LM is instead conducted in a purely unsupervised manner. Another is that BC utilizes a rich set of features to characterize a given spoken sentence while LM is constructed solely on the basis of the lexical (unigram) information.

6.2 Experiments on the proposed methods

We then turn our attention to investigate the utility of several methods deduced from our proposed summarization framework. We first consider the case when a 0-1 loss function is used (cf. (13)), which just show a simple combination of BC and LM. As can be seen from the first row of Table 4, such a combination can give about 4% to 5% absolute improvements as compared to the results of BC illustrated in Table 3. It in some sense confirms the feasibility of combining the supervised and unsupervised summarizers. Moreover, we consider the use of the loss functions defined in (11) (denoted by SIM) and (12) (denoted by MMR), and the corresponding results are shown in the second and the third rows of Table 4, respectively. It can be found that

MMR delivers higher summarization performance than SIM (especially for the SD case), which in turn verifies the merit of incorporating the MMR concept into the proposed framework for extractive summarization. If we further compare the results achieved by MMR with those of BC and LM as shown in Table 3, we can find significant improvements both for the TD and SD cases. By and large, for the TD case, the proposed summarization method offers relative performance improvements of about 19%, 23% and 19%, respectively, in the ROUGE-1, ROUGE-2 and ROUGE-L measures as compared to the BC baseline; while the relative improvements are 29%, 46% and 31%, respectively, in the same measurements for the SD case. On the other hand, the performance gap between the TD and SD cases are reduced to a good extent by using the proposed summarization framework.

In the next set of experiments, we simply assume the sentence prior probability $P(S_j)$ defined in (8) is uniformly distributed, namely, we do not use any supervised information cue but use the lexical information only. The importance of a given sentence is thus considered from two angles: 1) the relationship between a sentence and the whole document, and 2) the relationship between the sentence and the other individual sentences. The corresponding results are illustrated in the lower part of Table 4 (denoted by Uniform). We can see that the additional consideration of the sentence-sentence relationship appears to be beneficial as compared to that only considering the document-sentence relevance information (cf. the second row of Table 3). It also gives competitive results as compared to the performance of BC (cf. the first row of Table 3) for the SD case.

6.3 Comparison with conventional summarization methods

In the final set of experiments, we compare our proposed summarization methods with a few existing summarization methods that have been widely used in various summarization tasks, including LEAD, VSM, LexRank and CRF; the corresponding results are shown in Table 5. It should be noted that the LEAD-based method simply extracts the first few sentences in a document as the summary. To our surprise, CRF does not provide superior results as compared to the other summarization methods. One possible explanation is that the structural evidence of the spoken documents in the test set is not strong enough for CRF to show its advantage of modeling the local structural information among sentences. On the other hand, LexRank gives a very promising performance in spite that it only utilizes lexical information in an unsupervised manner. This somewhat reflects the importance of capturing the global relationship for the sentences in the spoken document to be summarized. As compared to the results shown in the "BC" part of Table 4, we can see that our proposed methods significantly outperform all the conventional summarization methods compared in this paper, especially for the SD case.

		ROGUE-1	ROUGE-2	ROUGE-L
LEAD	TD	0.320	0.197	0.283
	SD	0.312	0.168	0.251
VSM	TD	0.345	0.220	0.287
	SD	0.337	0.189	0.277
LexRank	TD	0.435	0.314	0.377
	SD	0.348	0.204	0.294
CRF	TD	0.431	0.315	0.383
	SD	0.358	0.220	0.291

Table 5: The results achieved by four conventional summarization methods.

7 Conclusions and future work

We have proposed a risk minimization framework for extractive speech summarization, which enjoys several advantages. We have also presented a simple yet effective implementation that selects the summary sentences in an iterative manner. Experimental results demonstrate that the methods deduced from such a framework can yield substantial improvements over several popular summarization methods compared in this paper. We list below some possible future extensions: 1) integrating different selection strategies, e.g., the listwise strategy that defines the loss function on all the sentences associated with a document to be summarized, into this framework, 2) exploring different modeling approaches for this framework, 3) investigating discriminative training criteria for training the component models in this framework, and 4) extending and applying the proposed framework to multi-document summarization tasks.

References

James O. Berger *Statistical decision theory and Bayesian analysis*. Springer-Verlap, 1985.

Berlin Chen. 2009. Word topic models for spoken document retrieval and transcription. *ACM Transactions on Asian Language Information Processing*, 8, (1): 2:1 - 2:27.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proc. of Annual International ACM SIGIR Conference on*

Research and Development in Information Retrieval: 335 - 336.

Yi-Ting Chen, Berlin Chen and Hsin-Min Wang. 2009. A probabilistic generative framework for extractive broadcast news speech summarization. *IEEE Transactions on Audio, Speech and Language Processing*, 17, (1): 95 - 106.

John M. Conroy and Dianne P. O'Leary. 2001. Text summarization via hidden Markov models. In *Proc. of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*: 406 - 407.

Güneş Erkan and Dragomir R. Radev. 2004. LexRank: graph-based lexical centrality as salience in text summarization. *Journal or Artificial Intelligence Research*, 22: 457 - 479.

Mohamed Abdel Fattah and Fuji Ren. 2009. GA, MR, FFNN, PNN and GMM based models for automatic text summarization. *Computer Speech and Language*, 23, (1): 126 - 144.

Louisa Ferrier *A maximum entropy approach to text summarization*. School of Artificial Intelligence, University of Edinburgh, 2001.

Sadaoki Furui, Tomonori Kikuchi, Yousuke Shinnaka and Chiori Hori. 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing*, 12, (4): 401 - 408.

Michel Galley. 2006. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proc. of Conference on Empirical Methods in Natural Language Processing*: 364 - 372.

Vaibhava Goel and William Byrne. 2000. Minimum Bayes-risk automatic speech recognition. *Computer Speech and Language*, 14, (2): 115 - 135.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proc. of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*: 19 - 25.

Akira Inoue, Takayoshi Mikami and Yoichi Yamashita. 2004. Improvement of speech summarization using prosodic information, In *Proc. of Speech Prosody*: 599 - 602.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proc. of Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting*: 169 - 176.

Aleksander Kolcz, Vidya Prabakarmurthi and Jugal Kalita. 2001. Summarization as feature selection for text categorization. In *Proc. of Conference on Information and Knowledge Management*: 365 - 370.

Julian Kupiec, Jan Pedersen and Francine Chen. 1999. A trainable document summarizer. In *Proc. of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*: 68 - 73.

Konstantinos Koumpis and Steve Renals. 2000. Transcription And Summarization Of Voicemail Speech. In *Proc. of International Conference on Spoken Language Processing*: 688 - 691.

Chin-Yew Lin. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In *Proc. of Workshop on Text Summarization Branches Out*.

Shih-Hsiang Lin, Berlin Chen and Hsin-Min Wang. 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.

Shih-Hsiang Lin, Yueng-Tien Lo, Yao-Ming Yeh and Berlin Chen. 2009. Hybrids of supervised and unsupervised models for extractive speech summarization. In *Proc. of Annual Conference of the International Speech Communication Association*: 1507 - 1510.

Inderjeet Mani and Mark T. Maybury *Advances in automatic text summarization*. MIT Press, Cambridge, 1999.

Sameer R. Maskey and Julia Hirschberg. 2003. Automatic Summarization of Broadcast News using Structural Features. In Proc. of *the European Conf. Speech Communication and Technology*: 1173 - 1176.

Kathleen McKeown, Julia Hirschberg, Michel Galley and Sameer Maskey. 2005. From text to speech summarization. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*: 997 - 1000.

Rada Mihalcea and Paul Tarau. 2005. TextRank: bringing order into texts. In *Proc. of Conference on Empirical Methods in Natural Language Processing*: 404 - 411.

Dragomir R. Radev, Hongyan Jing, Małgorzata Stys and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40: 919 - 938.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang and Zheng Chen. 2007. Document summarization using conditional random fields. In *Proc. of International Joint Conference on Artificial Intelligence*: 2862 - 2867.

Hsin-Min Wang, Berlin Chen, Jen-Wei Kuo and Shih-Sian Cheng. 2005. MATBN: A Mandarin Chinese broadcast news corpus. *International Journal of Computational Linguistics and Chinese Language Processing*, 10, (2): 219 - 236.

ChengXiang Zhai and John Lafferty. 2006. A risk minimization framework for information retrieval. *Information Processing & Management*, 42, (1): 31 - 55.

ChengXiang Zhai. *Statistical language models for information retrieval*. Morgan & Claypool Publishers, 2008.

Justin Jian Zhang, Ho Yin Chan and Pascale Fung. 2007. Improving Lecture Speech Summarization Using Rhetorical Information. In *Proc. of Workshop of Automatic Speech Recognition Understanding*: 195 - 200.