Extractive Speech Summarization — From the View of Decision Theory

Shih-Hsiang Lin, Yao-Ming Yeh and Berlin Chen

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

{shlin, ymyeh, berlin}@csie.ntnu.edu.tw

Abstract

Extractive speech summarization can be thought of as a decision-making process where the summarizer attempts to select a subset of informative sentences from the original document. Meanwhile, a sentence being selected as part of a summary is typically determined by three primary factors: significance, relevance and redundancy. To meet these specifications, we recently presented a novel probabilistic framework stemming from the Bayes decision theory for extractive speech summarization. It not only inherits the merits of several existing summarization techniques but also provides a flexible mechanism to render the redundancy and coherence relationships among sentences and between sentences and the whole document, respectively. In this paper, we propose several new approaches to the ranking strategy and modeling paradigm involved in such a framework. All experiments reported were carried out on a broadcast news speech summarization task; very promising results were demonstrated.

Index Terms: extractive speech summarization, decision making, Bayes decision theory

1. Introduction

Speech summarization, enabling users to efficiently review or quickly spot the important information conveyed in a single spoken document or multiple spoken documents, is one of the most indispensable technologies for tackling the information overload problem in the multimedia era [1]. Typically, summarization tasks can be grouped into different categories by considering them from a wide range of aspects, for example, 1) input — single document or multiple documents, 2) purpose — generic or query-oriented and 3) output — extractive or abstractive. Interested readers may refer to [2] for a comprehensive and entertaining overview of document summarization. In this paper, we focus exclusively on generic, extractive speech summarization since it usually forms the building block for many other speech summarization tasks.

Further, a spoken sentence to be selected as part of a summary is usually being considered from the following three factors: 1) significance — the importance of the sentence itself, 2) relevance — the degree of the similarity between the sentence and other sentences in the document, and 3) redundancy — the information carried by the sentence and the already selected summary sentences should cover different topics or concepts of the document. Quite a few methods have been proposed to partially address the above three factors. For the significance factor, a typical example is to estimate the significance of each spoken sentence with supervised machine-learning techniques. It can be cast into a two-class (i.e., summary and non-summary) sentence-classification problem [3]. A spoken sentence equipped with a set of indicative features, such as acoustic cues [4], lexical cues [5], structural cues [6-7] or discourse cues [8], is input to the classifier (or summarizer) and a decision value (or score) is then returned from it in view of these features. Then, summary sentences are subsequently ranked and selected according to those scores. Although such supervised summarizers are effective, most of them usually explicitly assume that sentences are independent of each other and classify each sentence individually without leveraging the relationship among the sentences [3]. Another major shortcoming 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.

Another stream of thought attempts to conduct speech summarization based on some heuristic rules or statistical evidences between each sentence and the document, getting around the need of manually labeled training data. We may sometimes name them unsupervised summarizers. The central idea of these summarizers revolves around the notion of the relevance of a sentence to other sentences [9-11]. 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, these summarizers are usually constructed solely on the basis of the lexical information without considering other sources of information cues, and the imperfect speech recognition often leads to degraded performance. To address the latter issue, [12-13] investigated various ways to robustly represent the recognition hypotheses of spoken documents beyond the top one recognition hypothesis. For the issue of the redundancy, maximum marginal relevance (MMR) [14] is considered to be a good remedy. MMR performs sentence selection iteratively by striking the balance between topic relevance and coverage.

Building on these observations, we recently presented a novel probabilistic summarization framework stemming from the Bayes decision theory [15]. It not only has the capability to inherit the merits of most existing summarization methods but also provides a convenient and principled way to take into account the above-mentioned three factors. Our work in this paper continues this general framework of research in two significant aspects: 1) we extend the sentence selection strategy from the so-called "sentence-wise" to "list-wise," and 2) we explore various ways to construct the component models involved in such a framework.

2. Decision Making Process for Extractive Speech Summarization

Extractive speech summarization can be viewed as a decision making process in which the summarizer attempts to select a representative subset of spoken sentences from the original spoken document. 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 inherent in the decision process.
With loss of generality, let us denote \( \pi \in \mathcal{I} \) as one of possible selection strategies. 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. For ease of discussion, we denote by \( \pi_k \) and \( D \) the \( k \)-th selection strategy and the spoken document to be summarized, respectively. Then, the expected risk associated with a certain selection strategy \( \pi_k \) is expressed by
\[
R(\pi_k | D) = \mathbb{E}_\pi[L(\pi_k, \pi)]|p(\pi | D)|\pi,
\]
where \( L(\pi_k, \pi) \) is a loss function which specifies the cost associated with \( \pi_k \) given that \( \pi \) (another selection strategy) is the true state of nature; \( p(\pi | D) \) is the posterior probability of \( \pi \) given \( D \). Consequently, the ultimate goal of extractive speech 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:
\[
\pi^* = \arg \min_{\pi_k} R(\pi_k | D)
= \arg \min_{\pi_k} \mathbb{E}_\pi[L(\pi_k, \pi)]|p(\pi | D)|\pi.
\]
Although we have described a general formulation for the extractive summarization problem on the grounds of the Bayes decision theory in this section, we consider hereafter two special cases of it where the selection strategy is implemented in either a "sentence-wise" manner or a "list-wise" manner.

2.1. Sentence-wise Minimum Bayes Risk Decoding
We assume that summary sentences can be iteratively chosen (i.e., one at each iteration) from the document until the aggregated target summary reaches a predefined target summarization ratio. Therefore, the risk minimization framework can be reduced to
\[
S^* = \arg \min_{S \subseteq D} R(S | \bar{D})
= \arg \min_{S \subseteq D} \sum_{s \in S} L(s, \bar{D})|p(s | \bar{D}),
\]
where \( \bar{D} \) denotes the remaining sentences that have not been selected into the summary yet (i.e., the "residual" document); \( p(s | \bar{D}) \) is the posterior probability of a sentence \( s \) given \( D \). By applying the Bayes’ rule, we can have the selection strategy as
\[
S^* = \arg \min_{S \subseteq D} \sum_{s \in S} L(s, \bar{D})|p(s | \bar{D})|p(s | \bar{D}).
\]
A remarkable characteristic of this framework lies in that a sentence to be selected is actually evaluated by three fundamental factors: 1) \( P(S_j) \) is the sentence prior probability that addresses the importance of sentence \( S_j \) itself, 2) \( P(D | S_j) \) the sentence generative probability captures the degree of relevance of \( S_j \) to the residual document \( D \), and 3) \( L(S_j, \bar{D}) \) is the loss function that characterizes the relationship between sentence \( S_j \) and any other sentence \( \bar{D} \).

2.2. List-wise Minimum Bayes Risk Decoding
The iterative (or greedy) selection procedure described above may sometimes result in a suboptimal selection. For example, the information carried by a verbose summary sentence would be succinctly depicted by one or more other concise (short) sentences which are not included into the summary but cover more topics of interest. To avoid this potential defect, we may formulate the extractive summarization as a maximum convergence problem under a summary length constraint and to solve the problem by some global inference algorithms [16]. We, however, present here an alternative remedy to address the issue by leveraging the so-called “list-wise” selection strategy under the risk minimization framework. Specifically, we consider every possible combination (or subset) of sentences in a spoken document as a candidate summary \( \psi \) and then the best summary can be constructed through the following equation:
\[
\sum \arg \min_{\psi \subseteq \bar{D}} \sum_{\psi \subseteq \bar{D}} L(\psi, \bar{D})|p(\psi | \bar{D})|p(\psi | \bar{D}).
\]
where \( \psi \) denotes all possible combinations of part of sentences in a spoken document \( D \) (i.e., the set of candidate summaries). For practical implementation, it is impossible to enumerate all possible combinations of summary sentences for forming the summary of a spoken document, due to the reason that the number of possible combinations would grow exponentially as the number of sentences increases. To reduce the computational overhead, we can first use some prior knowledge, for example, the sentence-wise minimum Bayes risk decoding, to select a subset of possible summary sentences as the candidates for being considered to be included in the summary, and then enumerate all possible combinations (or samplings) of these sentences under a specific constraint of the length of the target summary.

3. Implementation
In this section, we shed light on the construction of the component models involved in the Eq. (4) or Eq. (5). Here, we take sentence-wise minimum the Bayes risk decoding as an illustrative example, while the models involved in the list-wise minimum Bayes risk decoding can be constructed along a similar vein.

3.1. Sentence Generative Model
We explore the language modeling (LM) approach [10], which has been introduced to a wide spectrum of information retrieval (IR) tasks and demonstrated with good empirical success, to estimate the sentence generative probability \( p(D | S_j) \). For example, each sentence in a document can be simply regarded as a probabilistic generative model consisting of a unigram distribution for generating the document [10]:
\[
P(D | S_j) = \prod_{w \in \bar{D}} P(w | S_j),
\]
where \( c(w, D) \) is the number of times that index term (or word) \( w \) occurs in \( D \) and \( P(w | S_j) \) is the probability of \( w \) being generated by \( S_j \). Due to space limitations, we ask the reader to refer to [10] for a thorough discussion on various ways to construct the sentence generative model.

3.2. Sentence Prior Model
The sentence prior probability \( P(S_j) \) (cf. Eq. (4)) can be regarded as the likelihood of a sentence \( S_j \) 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. Here, we take advantage of the learning capability of supervised machine-learning methods to estimate the sentence prior probability. We assume \( P(S_j) \) is in proportion to the posterior probability of sentence \( S_j \) being included in the summary class when observing a set of
indicative features $X_j$ of $S_j$. Specifically, the sentence prior probability $P(S_j)$ can be approximated by:

\[
P(S_j) \approx \frac{P(X_j | S)P(S)}{P(X_j | S)P(S) + P(X_j | S')P(S')}
\]

(7)

where $P(X_j | S)$ and $P(X_j | S')$ are the likelihoods that a sentence $S_j$ with features $X_j$ are generated by the summary class $S$ and the non-summary class $S'$, respectively; and the prior probability $P(S)$ and $P(S')$ are set to be equal in this paper. To estimate $P(X_j | S)$ and $P(X_j | S')$, several popular supervised classifiers (or summarizers), like Bayesian classifier (BC) or support vector machine (SVM), can be leveraged for this purpose. However, since the way to estimate the prior probability of each candidate summary $P(y)$ (cf. Eq. (5)) is still under active study, from here on, we will assume that the prior probability $P(y)$ involved in the list-wise minimum Bayes risk decoding is uniformly distributed.

3.3. Loss Function

The loss function introduced in the presented 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 summary sentence. Therefore, the loss function can be built on the notion of the similarity measure. In this paper, we define three different kinds of loss functions:

- **VSM Loss Function**: We use the “TF-IDF” weighting scheme to calculate the cosine similarity between the vector space model (VSM) representations of any given two sentences. The loss function is thus defined by

\[
L(S_j, S_i) = 1 - \text{Sim}(S_j, S_i)
\]

(8)

- **MMR Loss Function**: To avoid the newly selected summary sentence having the similar (redundant) information that is also contained in the already selected summary sentences, we borrow the idea from the MMR scheme to calculate the cosine similarity between the sentence-wise minimum Bayes risk decoding is uniformly distributed.

\[
L(S_j, S_i) = \begin{cases} 
\beta \cdot \text{Sim}(S_j, S_i) \\
(1 - \beta) \cdot \max_{S' \in \text{Summ}} \text{Sim}(S_j, S_i)
\end{cases}
\]

(9)

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

- **KL-Divergence Loss Function**: We assume that if sentences are similar to each other, they should be drawn from the same probability distribution. Therefore, we can use the KL-divergence measure to quantify how close any two sentences are [12]:

\[
L(S_j, S_i) = \sum_{w \in V} P(w | S_j) \log \frac{P(w | S_j)}{P(w | S_i)}
\]

(10)

where $w$ denotes a specific word in the vocabulary set $V$.

4. Experimental Setup

All the experiments were conducted on a set of 205 broadcast news documents compiled from the MATBN corpus [3]. We chose 20 documents as the held-out test set while the remaining documents as the development set [15]. The average Chinese character error rate obtained for the spoken documents is about 30% and sentence boundaries are determined by speech pauses.

The summarization ratio, defined as the ratio of the number of words in the automatic (or manual) summary to that

<table>
<thead>
<tr>
<th>Prior Loss</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1</td>
<td>0.417</td>
<td>0.281</td>
<td>0.356</td>
</tr>
<tr>
<td>SIM</td>
<td>0.475</td>
<td>0.351</td>
<td>0.420</td>
</tr>
<tr>
<td>MMR</td>
<td>0.475</td>
<td>0.351</td>
<td>0.420</td>
</tr>
<tr>
<td>KL</td>
<td>0.467</td>
<td>0.336</td>
<td>0.409</td>
</tr>
<tr>
<td>0-1</td>
<td>0.319</td>
<td>0.164</td>
<td>0.253</td>
</tr>
<tr>
<td>Uniform</td>
<td>0.365</td>
<td>0.209</td>
<td>0.305</td>
</tr>
<tr>
<td>MMR</td>
<td>0.391</td>
<td>0.236</td>
<td>0.338</td>
</tr>
<tr>
<td>KL</td>
<td>0.364</td>
<td>0.209</td>
<td>0.301</td>
</tr>
</tbody>
</table>

Table 2: The results achieved by several methods derived from the sentence-wise minimum Bayes risk decoding.

in the reference transcript of a spoken document, was set to 10% in this paper. To evaluate the quality of the automatic generated summaries, we used the ROUGE F-Score as the evaluation measurement [17]. The levels of agreement between the three subjects for important sentence ranking are 0.600, 0.532 and 0.527, respectively for ROUGE-1, ROUGE-2 and ROUGE-L F-score. Moreover, we align the speech recognition transcripts of the summary sentences to their respective waveform segments to obtain the correct (manual) transcripts for evaluation to get rid of the negative effect of recognition errors when applying various ROUGE evaluations on the automatic summaries obtained with the speech recognition transcripts [12].

For the sentence prior model, we take BC classifier as the example to study. The input to BC consists of a set of 28 indicative features used to characterize a spoken sentence [15], including the structural features (e.g., position, duration and length features, etc.), the lexical features (e.g., number of name entities or stop words and bigram scores, etc.), the acoustic features (e.g., formant and pitch, etc.) and the relevance feature [3]. For each kind of acoustic features, the minimum, maximum, mean, difference value and mean difference value of a spoken sentence are extracted.

5. Experimental Results and Discussion

In the first set of experiments, we report the baseline performance of the LM and BC summarizers (cf. Sections 3.1 and 3.2), respectively. The corresponding results are detailed in Table 1. We then assess the utility of several methods deduced from the proposed summarization framework. We start by considering a special case where a 0-1 loss function is used in Eq. (4), namely, the loss function will take value 0 if the two sentences are identical, and 1 otherwise. This attempt actually provides a natural integration of the supervised (i.e., BC) and unsupervised (i.e., LM) summarizers. As can be seen from the first row of Table 2, such a combination can give about 4% to 5% absolute improvements as compared to the results of BC illustrated in Table 1. It confirms the feasibility of combining the supervised and unsupervised summarizers. Moreover, we consider the use of the various loss functions defined in Eq. (8) (denoted by SIM), Eq. (9) (denoted by MMR) and Eq. (10) (denoted by KL); the corresponding results are shown in the second to the fourth rows of Table 2, respectively. Observing Table 2 we notice two particularities.
First, properly leveraging the loss function can greatly boost the summarization performance. Second, KL performs slightly worse than SIM and MMR. A possible explanation is that the speech recognition errors will result in an inaccurate estimation of Eq. (10) since there are only a few words present in the erroneous transcripts of spoken sentences. However, we believe that KL still has the merit of being able to accommodate more elaborate model estimation techniques to improve the performance in a systematic way [12]. It might be of special interest for future study.

In the next set of experiments, we simply assume the sentence prior probability \( P_S(j) \) defined in Eq. (4) is uniformly distributed; namely, we do not use any supervised information cue but use the lexical information only. The corresponding results are illustrated in the lower part of Table 2 (denoted by “Uniform”). It should be noted that the coupling of the uniform prior probability and the 0-1 loss function is equivalent to the baseline LM approach. The results seem to reflect that the additional consideration of the “sentence-sentence” relationship is beneficial as compared to that considering only the “document-sentence” relevance information (cf. the second row of Table 1). It also yields competitive results as compared to the performance of BC (cf. the first row of Table 1). Moreover, MMR delivers higher summarization performance than SIM, which clearly demonstrates the merit of incorporating the MMR concept into the proposed framework for extractive summarization.

We then evaluate the performance of the list-wise selection strategy, and the results are shown in Table 3. It can be found that the list-wise selection strategy consistently outperforms the sentence-wise selection strategy (cf. Table 2) even under the assumption of uniform distributed priors \( P(y) \). To better understand why it outperforms the sentence-wise selection strategy, we further analyze the average number of summary sentences respectively selected by these two strategies subject to the length constraint. We observe the list-wise selection strategy selected about 5.05 sentences on average while the sentence-wise selection strategy selected about 4.1 sentences into the summary under the same length constraint. These statistics might reveal that the list-wise selection strategy can, to some extent, avoid selecting verbose sentences into the summary.

We have also empirically compared our proposed summarization methods with a few popular summarization methods, including LEAD, LexRank [11] and CRF[18], where the LEAD-based method extracts the first few sentences in a document as the summary. The results shown in Table 4 demonstrate that our proposed methods outperform the conventional summarization approaches by a big margin.

6. Conclusions

In this paper, we have presented a novel probabilistic framework stemming from the Bayes decision theory for extractive speech summarization. Two variants of its implementation are thoroughly discussed as well. Meanwhile, we also demonstrate how to systematically integrate several existing summarization methods into the proposed framework. The empirical results show that our methods can significantly boost the summarization performance. We believe that this initial attempt provides a new avenue for future research on speech summarization.

7. Acknowledgements

This work was supported in part by the National Science Council, Taiwan, under Grants: NSC 98-2221-E-003-011-MY3, and NSC 99-2515-S-003-004, NSC 98-2631-S-003-002.

8. References