CONSTRUCTING EFFECTIVE RANKING MODELS FOR SPEECH SUMMARIZATION

Yueng-Tien Lo[1], *Shih-Hsiang Lin*[2], *Berlin Chen*[1]

[1] National Taiwan Normal University, Taipei, Taiwan
[2] Delta Electronics, Inc., Taipei, Taiwan
{g96470198, berlin}@csie.ntnu.edu.tw; simon.sh.lin@delta.com.tw

ABSTRACT

Speech summarization, facilitating users to better browse through and understand speech information (especially, spoken documents), has become an active area of intensive research recently. Many of the existing machine-learning approaches to speech summarization cast important sentence selection as a two-class classification problem and have shown empirical success for a wide array of summarization tasks. One common deficiency of these approaches is that the corresponding learning criteria are loosely related to the final evaluation metric. To cater for this problem, we present a novel probabilistic framework to learn the summarization models, building on top of the Bayes decision theory. Two effective training criteria, viz. maximum relevance estimation (MRE) and minimum ranking loss estimation (MRLE), deduced from such a framework are introduced to characterize the pair-wise preference relationships between spoken sentences. Experiments on a broadcast news speech summarization task exhibit the performance merits of our summarization methods when compared to existing methods.

Index Terms— speech summarization, sentence-classification, imbalanced-data, ranking capability, evaluation metric

1. INTRODUCTION

In the recent past, speech summarization has received a growing amount of interest and activity in the speech processing community. This is due in large part to the advances in automatic speech recognition (ASR) and the ever-increasing volumes of multimedia associated with spoken documents made available to the public. Speech summarization is anticipated to distill important information and remove redundant and incorrect information from spoken documents, enabling user to efficiently browse through spoken documents and digest the associated topics quickly [1-3]. Broadly speaking, 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. In this paper, we focus exclusively on extractive speech summarization, even though we will typically omit the qualifier "extractive."

Apart from traditional ad-hoc summarization methods [4-5], such as those based on document structure, linguistic or prosodic information, and proximity or significance measures to identify salient sentences, the machine-learning approaches with supervised training have attracted much attention and been applied with good success in many summarization tasks [6-8]. In general, the summarization task is cast as a two-class (summary/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. Specifically, the problem of speech summarization can be formulated as follows: Construct a 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; then, important sentences are ranked and selected according to these scores. Representative techniques include, but not limited to, Bayesian classifier (BC), support vector machine (SVM) and conditional random fields (CRF) [6-8], to name but a few.

However, the imbalanced-data (or skewed-data) problem might strongly affect the performance of a speech summarizer since the summary sentences of a given training spoken document usually are a small percentage of the original document as compared to non-summary ones. When training a summarizer on the basis of such an imbalanced-data set, the resulting summarizer tends to assign sentences of a spoken document to be summarized to the majority class (i.e., the class of non-summary sentences). 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 [9]. 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 (viz. the so-called "*bag-of-sentences*" assumption) with little consideration of relationships among the sentences of the document to be summarized. Rather than treating speech summarization as a binary classification problem, there is a recent stream of thought attempting to adopt the so-called "*learning-to-rank*" conception, originating from the field of information retrieval (IR), to trains a summarizer [10]. For example, the Ranking SVM and AdaRank based summarization methods might be considered two basic representatives of this category. Ranking SVM trains a summarizer in a pair-wise rank-sensitive manner. Namely, the learning objective is not only at the labeling correctness of each sentence of a training spoken documents, but also at the correct ordering relationship of each sentence pair in accordance with their respective importance to the document. AdaRank, instead, is to train the summarizer by directly optimizing the ultimate evaluation score of the summarizer.

Building on these observations, we present in this paper a probabilistic framework for training speech summarization models, stemming from the Bayes decision theory [11]. It formulates speech summarization as a decision making process where a representative subset of sentences is selected from the original document to form a summary. In so doing, a summarization model will be trained with the aim at accurately quantifying the tradeoff between various decisions and the potential cost that accompanies

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

each decision. While such a notion has already set the foundations and been well practiced for many statistical pattern recognition and classification problems, it still remains under-explored in the context of speech summarization, as far as we know. In particular, two instantiations derived from this framework, viz. maximum relevance estimation (MRE) and minimum ranking loss estimation (MRLE), are introduced for training the summarization model.

2. MODELING FRAMEWORK

The Bayes decision theory, which quantifies the tradeoff between various decisions and the potential cost that accompanies each decision, is perhaps the one of the most prominent principles that can be used to guide the choice of a course of action in the face of some uncertainties underlying the decision process [11]. Without loss of generality, let us denote the input space $\mathbf{X}$ as all possible observations and the output space $\mathbf{Y}$ be equivalent to the space of all possible actions. Furthermore, we assume there exists a decision maker which is parameterized by a set of model parameters $\theta \in \Theta$. Therefore, the best decision could be expressed as the search (decoding) of the best candidate output y^* from the output space $\mathbf{Y}$ that minimizes the risk difined by

$$\begin{aligned} y^* &= \arg\min_{y} R(y \mid x;\theta) \\ &= \arg\min_{y} \int_{y' \in \mathbf{Y}} L(y,y') p(y' \mid x;\theta) dy', \end{aligned} \tag{1}$$

where $R(y \mid x;\theta)$ is the risk of choosing an output y given an input x and under the model set θ. $p(y' \mid x;\theta)$ is the posterior probability of x being assigned to the output y' under the model set θ, and $L(y,y')$ is used to measure the loss incurred by choosing the output y when the correct output is y'. On the other hand, for training the decision maker, if training instances, consisting of the ground-truth outputs $\hat{y}$ associated with all individual inputs x, are presented in the scenario of supervised training, then the optimum model parameters can be estimated by minimizing the overall expected risk defined as follows:

$$\begin{aligned} \theta^* &= \arg\min_{\theta} \int_{x \in \mathbf{X}} R(\hat{y} \mid x;\theta) P(x) dx \\ &= \arg\min_{\theta} \int_{x \in \mathbf{X}} \int_{y \in \mathbf{Y}} L(\hat{y}, y) P(y \mid x;\theta) P(x) dy dx. \end{aligned} \tag{2}$$

The notion of minimizing the Bayes risk has gained much attention and been applied with some success to a variety of NLP tasks like speech recognition [12], machine translation [13] and information retrieval [14]. Recently, this framework has been adapted to speech summarization [15] by simply using (1) to search for the best candidate summary output, while pulling together the existing supervised and unsupervised summarization models. Our work in this paper presents a continuation of this general line of thought by developing two novel discriminative summarization models that are deduced from (2).

3. DISCRIMINATIVE TRAINING OF SUMMARIZERS

3.1. Principle

Although we have described a general formulation for the model training (*cf.* (2)) and testing (*cf.* (1)) on the grounds of the Bayes decision theory in the previous section, we focus hereafter only on the aspect of model training. In the context of speech summarization, summary sentences that are presented to a user are usually ranked by the degree of importance of each sentence S_j of a spoken document D_i to be summarized. Hence, the problem of speech summarization could be stated as constructing an appropriate ranking function that assigns a preference (or rank) score to each sentence S_j of the spoken document D_i. Then, important sentences are selected according to these scores (or the corresponding ranks). Formally, we denote the sentences of the spoken document D_i to be summarized by $\mathbf{S}_i = \{S_1, \cdots, S_{|D_i|}\}$ and assume that each sentence S_j is associated with a set of M features $\mathbf{x}_j^{(D_i)} = \{x_{j1}^{(D_i)}, \cdots x_{jM}^{(D_i)}\}$ with respect to the document D_i. Furthermore, a ranking function (viz. a summarization model) $f : \mathbf{X} \rightarrow \mathbf{Y}$ with a parameter set θ is utilized to determine the preference score (or rank) of each sentence S_j of the document D_i based on such features $\mathbf{x}_j^{(D_i)}$.

In the training stage, we are given a set of training documents $\mathbf{D} = \{D_1, \cdots, D_{|\mathbf{D}|}\}$ and the information about their corresponding manually-labeled summary sentences. Additionally, we assume there exists a finite set of R ranks (or labels) $\mathbf{Y}=\{l_1, l_2, \cdots, l_R\}$, each of which can be assigned to the sentences of a spoken document, and the elements in the rank set have a total ordering relationship $l_1 \prec l_2 \prec \cdots \prec l_R$ where $\prec$ denotes a preference relationship. For example, R can be set to 3 representing that a given sentence can have the label of summary sentence (l_1), possible summary sentence (l_2) or non-summary sentence (l_3).

Building on the notion of Bayes risk minimization, the training procedure here could be stated as finding the best ranking function f^* that can minimize the overall expected risk defined in (2). It should be borne in mind that the integral in (2) will be conducted over the whole input and output spaces, which would be impossible to enumerate. In reality, we are only given a finite number of independent and identically distributed (i.i.d.) training instances (or documents). As such, we may instead try to estimate a ranking function f that can minimize the empirical expected risk $\widetilde{R}_{all}$ defined as follows:

$$\widetilde{R}_{all} = \sum_{i=1}^{|\mathbf{D}|} \sum_{j=1}^{|D_i|} R'\left(f\left(\mathbf{x}_j^{(D_i)};\theta\right)\right) P\left(\mathbf{x}_j^{(D_i)}\right), \tag{3}$$

where $|\mathbf{D}|$ is the number of training documents and $|D_i|$ is the number of sentences in the document D_i. We may further assume that the prior probability $P\left(\mathbf{x}_j^{(D_i)}\right)$ is uniformly distributed and properly expand the expected risk term $R'\left(f\left(x_j^{(D_i)};\theta\right)\right)$ in (3). Putting them together, we have

$$\widetilde{R}_{all} \approx \sum_{i=1}^{|\mathbf{D}|} \sum_{j=1}^{|D_i|} \sum_{k=1}^{|D_i|} L\left(y_j^{(D_i)}, y_k^{(D_i)}\right) P\left(y_k^{(D_i)} \mid \mathbf{x}_k^{(D_i)}, D_i;\theta\right), \tag{4}$$

where $P\left(y_k^{(D_i)} \mid \mathbf{x}_k^{(D_i)}, D_i;\theta\right)$ is the posterior probability of $y_k^{(D_i)}$ given that the sentence features $\mathbf{x}_k^{(D_i)}$ and the document D_i are observed; $L\left(y_j^{(D_i)}, y_k^{(D_i)}\right)$ is the loss function that characterizes the relationship between any pair of sentences. As a result, the optimum parameter set θ_{opt} of the ranking function f^* can be estimated by minimizing (4).

3.2. Maximum Relevance Estimation (MRE)

The most straightforward way of designing the loss function is to use a 0-1 loss function $L\left(y_j^{(D_i)}, y_k^{(D_i)}\right)$ where the loss function will take a value of 0 if the two sentences S_j and S_k have the identical label of belonging to the summary class and 1 otherwise. Given this assumption, it is easy to show that minimizing the empirical expected risk defined in (4) is approximately equivalent to maximizing the objective function defined as follows:

$$F_{\text{MRE}} = \sum_{i=1}^{|\mathbf{D}|} \sum_{S_j \in \mathbf{S}_{D_i}} P\left(y_j^{(D_i)} \mid \mathbf{x}_j^{(D_i)}, D_i; \theta\right), \tag{5}$$

where $\mathbf{S}_{D_i}$ denotes the set of reference (or true) summary sentences for the document D_i. Apparently, (5) states that if the ranking function has the capability to give higher scores (or preference labels) to reference summary sentences, then we can expect to have better summarization accuracy. We term (5) the maximum relevance estimation (MRE) hereafter. Also note that MRE intrinsically is very similar to other objective functions used in discriminative training of acoustic models in speech recognition, such as maximum mutual information estimation (MMIE) [16] and conditional maximum likelihood estimation (CMLE) [17].

3.3. Minimum Ranking Loss Estimation (MRLE)

A potential drawback of using the simple 0-1 loss function is that more elaborate ranking preference relationships have not been taken into account. To mitigate this problem, we thus define the loss function as:

$$L\left(y_j^{(D_i)}, y_k^{(D_i)}\right) = w_{jk}^{(D_i)} \begin{cases} 1, & S_k^{(D_i)} \succ S_j^{(D_i)} \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

where a loss will be incurred when $S_k^{(D_i)} \succ S_j^{(D_i)}$. The basic intuition is that we use (6) to render any spoken sentence pair that is incorrectly ranked (as opposed to their ideal preference order). Specifically, we can only consider risks incurred by those summary sentences which belong to the summary and ignore the risks caused by others. Consequently, (4) becomes

$$R_{\text{MRLE}} = \sum_{i=1}^{|\mathbf{D}|} \sum_{S_j \in \mathbf{S}_{D_i}} \sum_{k=1}^{|D_i|} P\left(y_k^{(D_i)} \mid \mathbf{x}_k^{(D_i)}, D_i; \theta\right) \times w_{jk}^{(D_i)}. \tag{7}$$

The last term (viz. the summation over the sentences in a document) in the right-hand side of (7) is equivalent to the calculation of the expected ranking error given the evidence that the sentence S_j is a summary sentence. We term (7) the minimum ranking loss estimation (MRLE).

MRLE has the ability to diminish ranking errors of any sentence pair with respect to a training document D_i to obtain a better ranking function (or summarization model). By taking advantage of the pair-wise learning strategy, MRLE, to some extent, can alleviate the problem caused by imbalanced-data problem. Essentially, MRLE is close in spirit to those that had ever been used in minimum phone error training (MPE) [18] and minimum error rate training (MERT) [19] in the field of speech recognition.

3.4. Model Implementation

As can be seen in (5) and (7), the calculation of the expected risk involves the estimation of the posterior probability $P\left(y_k^{(D_i)} \mid \mathbf{x}_k^{(D_i)}, D_i; \theta\right)$, which in fact is the key ingredient of the summarization model. A straightforward way is to use the so-called global conditional log-linear model (GCLM) to represent $P\left(y_k^{(D_i)} \mid \mathbf{x}_k^{(D_i)}, D_i; \theta\right)$:

$$P\left(y_k^{(D_i)} \mid \mathbf{x}_k^{(D_i)}, D_i; \theta\right) = \frac{1}{Z(\mathbf{x}_k^{(D_i)}, \theta)} \exp\left(\Phi\left(y_k^{(D_i)}, \mathbf{x}_k^{(D_i)}\right) \bullet \theta\right), \tag{8}$$

where $\Phi\left(y_k^{(D_i)}, \mathbf{x}_k^{(D_i)}\right)$ is an indicator vector used to describe the co-occurrence relationships between the label $y_k^{(D_i)}$ and the features $\mathbf{x}_k^{(D_i)}$ of the sentence S_k; θ is the corresponding parameter (weight) vector; $Z(\mathbf{x}_k^{(D_i)}, \theta)$ is a normalization factor. Whenever applying either the criteria (5) or (7) for estimating the parameter vector θ, we use stochastic gradient descent (SGD) to obtain an updated version of θ iteratively.

4. EXPERIMENTAL SETUP

All the summarization experiments were conducted on a set of broadcast news documents compiled from the MATBN corpus [15]. For each broadcast news document, three manual summaries are provided as references. A development set consisting of 100 documents were defined for training the model parameters while 20 documents were taken as the held-out evaluation set. The average Chinese character error rate obtained for the spoken documents is about 30% and the sentence boundaries were simply determined by speech pauses. To assess the goodness of the automatically generated summaries, we adopted the widely used recall oriented understudy for gisting evaluation (ROUGE) [20]. 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. The summarization results are evaluated at a default summarization ratio of 10%, defined as the ratio of the number of words in the automatic (or manual) summary to that of words in the manual transcript of the spoken document. The level of agreement on the ROUGE-2 measure between the three subjects for important sentence ranking is about 0.65.

In this paper, we use a set of 191 features to characterize a spoken sentence, including the structural feature, the lexical features, the acoustic features and the relevance features [8, 15]. 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.

5. EXPERIMENTAL RESULTS

At the outset, we report on the baseline summarization results obtained by three popular supervised summarization models compared in this paper, including SVM, RankSVM and AdaRank, which were all trained with 10% summary labels. They, respectively, belong to so-called the point-wise, pair-wise, and list-wise ranking strategies. The corresponding results are shown in the upper part of Table 1, where the results of CRF are also listed for reference. As can be seen, both Ranking SVM and AdaRank provide substantial improvements over SVM and CRF, while AdaRank performs slightly better than RankSVM. The experimental results reveal that RankSVM and AdaRank have good potential for extractive speech summarization. They also demonstrate the side benefit of mitigating the imbalanced-data problem as compared to the traditional SVM approach.

In the second set of experiments, we evaluate the utility of MRE and MRLE. Again, we used 10% summary labels to train MRE and MRLE. Consulting the corresponding results shown in the lower part of Table 1, we notice two particularities. First, both MRE and MRLE significantly outperform the SVM summarizer. It can be also seen that both MRE and MRLE perform on par with RankSVM and AdaRank. Second, there is no obvious difference between the performance of MRE and that of MRLE. This may be explained by the fact that using the binary (summary/non-summary sentence) labeling strategy will somehow weaken the learning ability of MRLE.

Taking a step further, we set three different labeling strategies for training the MRLE model so as to better understand the effect of labeling bias on MRLE. To do this, each sentence in the training documents was labeled as "summary sentence", "possible summary sentence", or "non-summary sentence". For example, we could label the top 10% important sentences in a document as summary sentences; 10%-20% important sentences as possible summary sentences and the remaining sentences as non-summary sentences (denoted by MRLE.1). Table 2 highlights the different labeling settings of MRLE that we investigated in this paper. As indicated in Table 3, we see that MRLE.3 outperforms MRLE.2, while MRLE.2 seems to perform slightly better than MRLE.1. These results, to some extent, demonstrate that MRLE has the capability to capture the intrinsic preference characteristics embodied in the training data, and is a good surrogate for the existing summarization methods. Compared to the results obtained by SVM, MRLE.3 achieves a relative improvement of about 8%-10% in the various ROUGE measures.

Table1: Summarization results achieved by various supervised summarization methods.

	ROUGE-1	ROUGE-2	ROUGE-L
SVM	0.406	0.270	0.341
CRF	0.365	0.221	0.300
RankSVM	0.416	0.288	0.353
AdaRank	0.427	0.286	0.356
MRE	0.427	0.288	0.363
MRLE	0.430	0.287	0.372

Table 2: Different labeling settings for summarization experiments.

	Summary	Possible Summary	Non-Summary
MRE	0 - 10%	-	10 - 100%
MRLE.1	0 - 10%	10 – 20%	20 - 100%
MRLE.2	0 - 10%	10 - 30%	30 - 100%
MRLE.3	0 - 20%	20 - 40%	40 - 100%

Table 3: Summarization results achieved by MRLE with respect to different labeling settings.

	ROUGE-1	ROUGE-2	ROUGE-L
MRLE.1	0.436	0.292	0.370
MRLE.2	0.439	0.296	0.374
MRLE.3	0.440	0.299	0.376

6. CONCLUSIONS

In this paper, we have presented two novel training methods for constructing a speech summarizer. The experimental results demonstrate the effectiveness of the proposed methods. It is worth emphasizing that MRLE seems to perform better than RankSVM and AdaRank; also, MRLE outperforms MRE. This can be reasonably inferred from the loss functions they utilized. The use of a 0-1 loss function in MRE does not fully consider the ranking preference between instances, while MRLE considers the loss incurred by any pair of training instances. As to future work, we envisage several directions, including 1) exploring more discriminative training algorithms [10, 19, 21], 2) leveraging different granularities of acoustic and lexical features for representing spoken documents, and 3) incorporating the summarization results into audio indexing for better retrieval and browsing of spoken documents [22].

7. REFERENCES

[1] S. Furui et al., "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Transactions on Speech and Audio Processing,* 12(4), pp. 401–408, 2004.

[2] K. McKeown et al., "From text to speech summarization," in *Proc. ICASSP 2005*.

[3] Y. Liu and D. Hakkani-Tür, "Speech summarization," Chapter 13 in *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, G. Tur and Renato D. Mori (eds.), Wiley, 2011.

[4] I. Mani and M. T. Maybury, *Advances in automatic text summarization*. Cambridge: MIT Press, 1999.

[5] A. Nenkova and K. McKeown, "Automatic summarization," *Foundations and Trends in Information Retrieval*, 5 (2-3), pp. 103–233.

[6] J. Zhang et al., "A Comparative study on speech summarization of broadcast news and lecture Speech," in *Proc. Interspeech 2007*.

[7] D. Shen et al., "Document summarization using conditional random fields," in *Proc. IJCAI 2007*.

[8] S.-H. Lin et al., "A comparative study of probabilistic ranking models for Chinese spoken document summarization," *ACM Transactions on Asian Language Information Processing*, 8(1), pp. 3:1–3:23, 2009.

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

[10] B. Chen et al., "Extractive speech summarization using evaluation metric-related training criteria," *Information Processing & Management*, available online 16 January 2012.

[11] J. Berger, *Statistical decision theory and Bayesian analysis*. Springer-Verlap, 1985.

[12] V. Goel and W. Byrne, "Minimum Bayes-risk automatic speech recognition," *Computer Speech and Language*, 14 (2), pp. 115-135, 2000.

[13] S. Kumar and W. Byrne, "Minimum Bayes-risk decoding for statistical machine translation," in *Proc. HLT-NAACL 2004*.

[14] C.X. Zhai and J. Lafferty, "A risk minimization framework for information retrieva," *Information Processing & Management*, 42 (1), pp. 31-55, 2006

[15] B. Chen and S-H Lin, "A risk-aware modeling framework for speech summarization," *IEEE Transactions on Audio, Speech and Language Processing*, 20(1), pp. 199-210, 2012.

[16] L. Bahl et al., "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," In *Proc. ICASSP 1986*.

[17] B. Roark et al., "Discriminative *n*-gram language modeling," *Computer Speech and Language*, 21 (2), pp. 373-392, 2007.

[18] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP2002*.

[19] F.J. Och, "Minimum error rate training in statistical machine translation," in *Proc. ACL 2003*.

[20] C.Y. Lin, "ROUGE: Recall-oriented Understudy for Gisting Evaluation." 2003.

[21] J.-W. Kuo, B. Chen, "Minimum word error based discriminative training of language models," in *Proc. Interspeech 2005*.

[22] B. Chen et al., "Query modeling for spoken document retrieval," in *Proc. ASRU 2011*.