LEVERAGING EVALUATION METRIC-RELATED TRAINING CRITERIA FOR SPEECH SUMMARIZATION

Shih-Hsiang Lin, Yu-Mei Chang, Jia-Wen Liu, and Berlin Chen
National Taiwan Normal University, Taiwan
{896470017, 697470133, 697470171, berlin}@ntnu.edu.tw

ABSTRACT

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 variety of summarization tasks. However, the imbalanced-data problem sometimes results in a trained speech summarizer with unsatisfactory performance. On the other hand, training the summarizer by improving the associated classification accuracy does not always lead to better summarization evaluation performance. In view of such phenomena, we hence investigate two different training criteria to alleviate the negative effects caused by them, as well as to boost the summarizer’s performance. One is to learn the classification capability of a summarizer on the basis of the pair-wise ordering information of sentences in a training document according to a degree of importance. The other is to train the summarizer by directly maximizing the associated evaluation score. Experimental results on the broadcast news summarization task show that these two training criteria can give substantial improvements over the baseline SVM summarization system.

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

1. INTRODUCTION

Speech summarization is anticipated to distill important information and remove redundant and incorrect information from spoken documents, enabling user to efficiently review spoken documents and understand the associated topics quickly [1-6]. 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.”

Aside from traditional ad-hoc summarization methods [1], 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 [2-6]. 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; and 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) [5].

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 document usually are in a small percent 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 the document to be summarized to the majority class (i.e., the class of non-summary sentences). Several methods have been proposed to relieve this problem, like re-sampling (up-sampling, down-sampling, or both) or re-weighting of the training exemplars [6-7]. On the other hand, higher sentence classification accuracy does not always imply better summarization quality. This is mainly because that the summarizer usually classifies each sentence individually with little consideration of relationships among the sentences of the document to be summarized.

Building on these observations, in this paper, we explore to leverage two different training criteria so as to mitigate the potential defects of the existing speech summarizers, which have the ability to connect the decision of a summarizer to the evaluation metric. Our first attempt is inspired from the notion of “learning to rank,” which trains a summarizer in a pair-wise rank-sensitive manner [8-12]. 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. Nevertheless, it turns out that this attempt in essence would be loosely related to the final evaluation metric. In this regard, the other attempt is instead to train the summarizer by directly optimizing the evaluation score of the summarizer [13-15]. The rest of this paper is structured as follows. Section 2 elucidates the principles the two training criteria are built upon, and how they can be exploited for speech summarization. Then, the experimental settings and a series of summarization experiments are presented in Sections 3 and 4, respectively. Finally, conclusions are drawn in Section 5.

2. SUMMARIZATION APPROACHES

In general, the sentence ranking strategy for extractive speech summarization can be stated as follows. Each sentence $s_i$ in a spoken document to be summarized is associated with a set of $M$
representative features $X_i = \{x_{i1}, \ldots, x_{im}, \ldots, x_{iM}\}$ (cf. Section 3.2) and a summarizer (or a ranking function) $f: X_i \rightarrow R$ is employed to classify and assign a importance (or relevance) score to each sentence $S_i$ according to its associated features $X_i$. Then, sentences of the document can be ranked and iteratively selected into the summary based on their scores until the length limitation or a desired summarization ratio is reached. During the training phase, a set of training spoken documents $D = \{d_1, \ldots, d_m, \ldots, d_N\}$, consisting of $N$ documents and the corresponding handcrafted summary information, is given. The summarizer is trained in the sense of reducing the classification errors of the summarizer made on the sentences of these training spoken document exemplars. It is expected that minimizing the classification errors caused by the summarizer would be equivalent to maximizing the lower bound of the summarization evaluation score (Usually, the higher the score, the better the performance.). In what follows, we describe the baseline summarizer, i.e., support vector machines (SVM), and two different training criteria that are leveraged to cope with the aforementioned fundamental problems facing the existing classification-based summarizers.

2.1. Support Vector Machine (SVM)

A SVM classifier is built upon the principle of structural risk minimization (SRM) in the statistical learning theory. If the dataset is linear separable, SVM attempts to find an optimal hyper-plane by utilizing a decision function that can correctly separate the positive and negative samples, and ensure the margin is maximal. In the nonlinear separable case, SVM uses kernel functions or defines slack variables to transform the problem into a linear discrimination problem. In this paper, we construct a binary SVM classifier where the radial basis function (RBF) is chosen as the kernel function. The posterior probability of a sentence $S_i$ being in the summary class can be approximated by a sigmoid operation where the weights are estimated from the development set by minimizing a negative log-likelihood function [5, 17].

2.2. Learning to Rank

The concept of "learning to rank" is to create a rank- or preference-sensitive ranking function. It assumes there exists a set of ranks (or preferences) $L = \{l_1, l_2, \ldots, l_M\}$ in the output space, while in the context of speech summarization, the value of $M$ , for example, can be simply set to 2 representing that a sentence can have either a summary ($l_1$) or a non-summary ($l_2$) sentence label. The elements in the rank set have a total ordering relationship $l_1 \succ l_2 \succ \cdots \succ l_M$ where $\succ$ denotes a preference relationship. In this paper, we explore the use of the so-called "pair-wise training" strategy for speech summarization, which considers not only the importance (or relevance) of sentences to a training spoken document but also the order of each sentence pair on the ideal ranked list [10]. Several approaches have been implemented to fulfill the "pair-wise training" strategy for various information retrieval (IR) related tasks in the past decade. Typical techniques include Ranking SVM [8, 10], RankBoost [11] and RankNet [12]. Each of these methods has its own merits and limitations. Here, we take Ranking SVM [10] as an example to implement this strategy for speech summarization, since it has shown to offer consistent improvements over traditional SVM in many IR-related tasks [8-10]. In extractive speech summarization, the training objective of Ranking SVM is to find a ranking function that can correctly determine the preference relation between any pair of sentences. For a more thorough discussion of Ranking SVM, interested readers can refer to [10].

2.3. Direct Optimization

Although reducing the sentence classification (e.g., SVM) or ranking (e.g., Ranking SVM) errors would be equivalent to maximizing the lower bound of the performance evaluation score of a given summarization system, it is still not closely related enough to the final evaluation metric for speech summarization. Recently, quite a few approaches have been proposed to train an IR system by directly maximizing the associated evaluation score. For instance, Joachims [13] presented an SVM-based method for directly optimizing multivariate nonlinear performance measures like the F1-score or Precision/Recall Breakeven Point (PRBEP) adopted in document classification. On the other hand, Cossock and Zhang [14] discussed the issue of learning ranking with preference to the top scoring documents. More recently, Xu and Li [15] proposed an ensemble-based algorithm that can iteratively optimize an exponential loss function based on various kinds of IR evaluation metrics, referred to as AdaRank.

In this paper, we try to adopt such a "direct optimization" training strategy for speech summarization, and AdaRank is taken as the initial attempt. The success of AdaRank lies in that ensemble-based systems may produce more favorable results than their single-classifier counterparts [16]. AdaRank is one variation of the AdaBoost algorithm that generates a set of weak rankers (or ranking functions) and integrates them through a linear combination to form the final ranking model [15, 16]. A weak ranker can be constructed in several ways by using, for example, different subsets of training exemplars. In implementation, we follow the original definition of AdaRank [15] by using single summarization features (cf. Section 3.2) as weak rankers. Conceptually, AdaRank learns a weight for each weak ranker from an iteratively updated distribution of the training document exemplars. At each round, the updated distribution will emphasize those training spoken documents having more sentences incorrectly ranked by the previously selected weak rankers, which actually is evidenced by the corresponding summarization evaluation scores of the training spoken documents. Hence, consecutive rankers are concentrated on dealing with those "hard-to-summarize" training spoken documents [16].

3. EXPERIMENTAL SETUP

3.1. Corpus and Evaluation Metric

All the summarization experiments were conducted on a set of 205 broadcast news documents compiled from the MATBN corpus [5]. Three subjects with professional background were asked to create summaries of the 205 spoken documents for the summarization experiments as references for evaluation: A development set consisting of 100 documents were defined for tuning the summarization models while the remaining 105 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 used the ROUGE measure [17] as the evaluation
metric, which is computed based on n-grams co-occurrences statistics between automatic summary and a set of reference (or manual) summaries. 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.

### 3.2. Features for Summarizers

Several features have been designed and widely used in the supervised summarization approaches [4-5]. In this paper, we use a set of 29 features to characterize a spoken sentence, including the structural feature, the lexical features, the acoustic features and the relevance features [5]. For each kind of acoustic features, the minimum, maximum, mean, difference value and mean difference value (indexed from 1 to 5) of a spoken sentence are extracted. The difference value is defined as the difference between the minimum and maximum values of the spoken sentence, while the mean difference value is defined as the mean difference between a sentence and its previous sentence. The features are outlined in Table 1, where WTM (Word Topic Model) [5], VSM (Vector Space Model) [18], LSA (Latent Semantic Analysis) [18] and MRW (Markov Random Walk) [19] are different unsupervised summarizers, respectively, producing single summarization features. Notice that the positional feature is excluded because it is not general enough and would highly depend on the epochs and genres of spoken documents [5].

<table>
<thead>
<tr>
<th>Types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural Feature</td>
<td>1. Duration of the current sentence (S1)</td>
</tr>
<tr>
<td>Lexical Features</td>
<td>1. Number of named entities (L1)</td>
</tr>
<tr>
<td></td>
<td>2. Number of stop words (L2)</td>
</tr>
<tr>
<td></td>
<td>3. Bigram language model scores (L3)</td>
</tr>
<tr>
<td></td>
<td>4. Normalized bigram scores (L4)</td>
</tr>
<tr>
<td>Acoustic Features</td>
<td>1. The 1&quot; formant (F1-1—F1-5)</td>
</tr>
<tr>
<td></td>
<td>2. The 2&quot; formant (F2-1—F2-5)</td>
</tr>
<tr>
<td></td>
<td>3. The pitch value (P-1—P-5)</td>
</tr>
<tr>
<td></td>
<td>4. The peak normalized cross-correlation of pitch (C-1—C-5)</td>
</tr>
<tr>
<td>Relevance Features</td>
<td>1. Relevance score obtained by WTM</td>
</tr>
<tr>
<td></td>
<td>2. Relevance score obtained by VSM</td>
</tr>
<tr>
<td></td>
<td>3. Relevance score obtained by LSA</td>
</tr>
<tr>
<td></td>
<td>4. Relevance score obtained by MRW</td>
</tr>
</tbody>
</table>

Table 1: Features used in the summarizers.

Based on the observations made on the above experimental results, we further group the structural, lexical and acoustic features together to make them more competitive (denoted by SET 1). The remaining four relevance features are also grouped together to form another feature set (denoted by SET 2). In the next set of experiments, we assess the performance of SVM with respect to different feature sets and different amounts of labeled data. SD case. This might explained by the fact that the ROUGE measure is based on counting the number of overlapping units (e.g., word bigrams) between the automatic summary and the reference summary. Even thought the summary sentences can be correctly selected (or identified), the evaluation will inevitably be strongly affected by the recognition errors. On the other hand, the relevance features seem to be more effective than the other features. This is because the relevance features, to some extent, are designed for capturing the relevance of a sentence to the whole document. They thus might be more closely related to the notion of identifying important or relevant sentences from a spoken document.

Based on the observations made on the above experimental results, we further group the structural, lexical and acoustic features together to make them more competitive (denoted by SET 1). The remaining four relevance features are also grouped together to form another feature set (denoted by SET 2). In the next set of experiments, we assess the performance of SVM with respect to different feature sets and different amounts of labeled data being used. More concretely, the proportions of summary sentences in a training spoken document being used are set in accordance with different ratios (i.e., 10%, 20% and 30%) of all the sentences in the document. The corresponding ROUGE-2 results are presented in Table 2. SVM will perform better when the numbers of labeled summary and non-summary sentences become more balanced (e.g., 30% summary labels), but its performance will degrade significantly when the numbers of labeled summary and non-summary sentences become more unbalanced (e.g., 10% summary labels). Meanwhile, it is interesting to mention that combining the structural, lexical and acoustic features together (SET 1) seems to provide more indicative cues than combining relevance features together (SET 2) for important sentence selection using SVM.

### 4. EXPERIMENTAL RESULTS

We first examine the summarization performance of different kinds of features (cf. Table 1) and the associated results are graphically illustrated in Figure 1. The ROUGE-2 results based on the manual transcripts of spoken documents (denoted by TD, text documents) are also depicted in Figure 1 for reference, in addition to those results based on the recognition transcripts (denoted by SD, spoken documents). For the TD case, the acoustic features were obtained by performing word-level forced alignment of the spoken documents to their corresponding manual transcripts. Observing Figure 1 we notice two particularities. On one hand, the performance of the TD case is significantly better than that of the
In the third set of experiments, we evaluate the utility of Ranking SVM and AdaRank, as a function of different feature sets and evaluation metrics [17] being used. The results for the SD case are shown in Table 3. Notice here that SVM, Ranking SVM and AdaRank are all learned from the 10% summary labels. As can be seen, both Ranking SVM and AdaRank provide substantial improvements over SVM, while AdaRank outperforms Ranking SVM when using all features or the features of SET 2. The values shown in the parentheses of Table 3 are the best results that can be achieved by AdaRank. The gaps between the actual and the best results are mainly due to that the final ranking model for AdaRank is optimized by using the development set rather than the evaluation set. Such performance mismatch (in ROUGE-1) of AdaRank with all features, for the first ten training iterations, is also illustrated in Figure 2.

In words, both the “pair-wise training” strategy and the “direct optimization” training strategy turn out to demonstrate good potential for extractive speech summarization. They also have the side effect of mitigating the imbalanced-data problem as compared to the traditional SVM approach.

5. CONCLUSIONS

In this paper, we have investigated two different training criteria for training a speech summarizer, which can not only deal with the imbalanced-data problem but also can optimize the summarizer’s performance according to the final evaluation metric. The experimental results indeed justify our expectation. Our future research directions include: 1) investigating more elaborate prosodic features that can be used for speech summarization, 2) seeking other alternative approaches to optimizing a summarizer’s performance, and 3) exploring better ways to represent the recognition hypotheses of spoken documents beyond the top scoring ones [20].

6. ACKNOWLEDGEMENT

This work was supported in part by the National Science Council, Taiwan, under Grants: 98-2221-E-003-011-MY3, 96-2628-E-003-015-MY3, and 98-2631-S-003-002.

7. REFERENCES


### Table 3: The summarization results for the SD case achieved by different summarization approaches.

<table>
<thead>
<tr>
<th></th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.427</td>
<td>0.269</td>
<td>0.398</td>
</tr>
<tr>
<td>Ranking SVM</td>
<td>0.449</td>
<td>0.283</td>
<td>0.418</td>
</tr>
<tr>
<td>AdaRank</td>
<td>0.459 (0.462)</td>
<td>0.303 (0.303)</td>
<td>0.432 (0.432)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.376</td>
<td>0.228</td>
<td>0.353</td>
</tr>
<tr>
<td>SET 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ranking SVM</td>
<td>0.407</td>
<td>0.243</td>
<td>0.380</td>
</tr>
<tr>
<td>AdaRank</td>
<td>0.378 (0.409)</td>
<td>0.237 (0.237)</td>
<td>0.362 (0.409)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.346</td>
<td>0.180</td>
<td>0.316</td>
</tr>
<tr>
<td>SET 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ranking SVM</td>
<td>0.417</td>
<td>0.255</td>
<td>0.380</td>
</tr>
<tr>
<td>AdaRank</td>
<td>0.438 (0.438)</td>
<td>0.273 (0.273)</td>
<td>0.403 (0.403)</td>
</tr>
</tbody>
</table>

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