INTERSPEECH 2009 BRIGHTON

Hybrids of Supervised and Unsupervised Models for Extractive Speech Summarization

Shih-Hsiang Lin, Yueng-Tien Lo,Yao-Ming Yeh, Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University, Taipei, Taiwan
{shlin, berlin}@csie.ntnu.edu.tw

ABSTRACT

Speech summarization, distilling important information and removing redundant and incorrect information from spoken documents, has become an active area of intensive research in the recent past. In this paper, we consider hybrids of supervised and unsupervised models for extractive speech summarization. Moreover, we investigate the use of the unsupervised summarizer to improve the performance of the supervised summarizer when manual labels are not available for training the latter. A novel training data selection and relabeling approach designed to leverage the inter-document or/and the inter-sentence similarity information is explored as well. Encouraging results were initially demonstrated.

Index Terms— Speech summarization, hybrid summarizer, unsupervised training

10.21437/Interspeech.2009-458

1. INTRODUCTION

Speech summarization, which aims at extracting important information and removing redundant and incorrect information from spoken documents, enables us to efficiently review spoken documents and understand their associated topics quickly [1-2]. It hence has the side effect of improving the efficiency for searching or organizing large volumes of spoken documents. The research of text summarization dates back to the late 1950s. In the past decade, the focus of the summarization research has been extended from text to spoken documents. Generally, the summarization techniques can be classified as either extractive or abstractive. Extractive summarization produces a summary by selecting salient sentences or paragraphs from an original document according to a predefined target summarization ratio. Abstractive summarization, on the other hand, provides a fluent and concise abstract of a certain length that reflects the key concepts of the document. This requires highly sophisticated techniques, including semantic representation and inference, as well as natural language generation [3]. Thus, in recent years, researchers have tended to focus on extractive summarization.

Aside from traditional ad-hoc methods, such as those based on document structure and style information [4], linguistic information [5], proximity [6] or significance measures [7] to identify salient sentences or paragraphs, machine-learning approaches with either supervised or unsupervised learning strategies have gained much attention and been applied with empirical success to many extractive summarization tasks [8]. For supervised machine-learning approaches, the summarization task is usually 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 [8]. Representative supervised machine-learning summarizers include, but not limited to, Bayesian classifier, support vector machine (SVM), and conditional random fields (CRF). The major shortcoming of these summarizers is that a set of handcrafted document-reference summary exemplars are required for training the summarizers; however, manual annotation is expensive in terms of time and personnel. Moreover, such summarizers trained on a specific domain might not be directly applicable to another one. The other potential problem is the bag-of-instances assumption implicitly made by most of these summarizers. That is, sentences are classified independently of each other, with little consideration of the dependence relationships among the sentences or the global structure of the document.

Another school of thought attempts to conduct document summarization using unsupervised machine-learning approaches, getting rid of the demand for manually labeled training data. For example, the graph-based methods, such as TextRank [9] and LexRank [10], 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). Put simply, sentences more similar to others are deemed more salient to the main theme of the document. Moreover, we have recently proposed a probabilistic generative framework for speech summarization, which can perform the summarization task in a purely unsupervised manner [11]. Each sentence of a spoken document to be summarized is treated as a probabilistic generative model or a language model for generating the document, and sentences are selected according to their likelihoods.

Even though the performance of unsupervised summarizers is usually worse than that of supervised summarizers, their domain-independent property still makes them attractive. Therefore, we expect that researches conducted along the aforementioned 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 also investigate the use of unsupervised summarizer to improve the performance of supervised summarizer when manual labels are not available for training the latter. A novel training data selection and relabeling approach designed to leverage the inter-document or/and the inter-sentence similarity information is explored as well.

2. EVALUATION CORPUS

All the experiments were conducted on a set of 205 broadcast news documents compiled from the MATBN corpus [8, 11]. For each news article, three manual summaries are provided as references. A development set consisting of 100 documents were defined for tuning the parameters or settings while the remaining documents were taken as the held-out evaluation set. The average Chinese character error rate obtained for the spoken documents is about 30% and sentence boundaries are determined by speech pauses.

To evaluate the quality of the automatic generated summaries, we used the ROUGE evaluation approach [12], which is based on *N*-grams co-occurrences statistics between automatic summary and a set of reference (or manual) summaries. More precisely, we

adopted the ROUGE_2 measure, which uses word bigrams as matching units. The levels of agreement on the ROUGE_2 measure between the three subjects are about 0.646, 0.668 and 0.684 respectively, for summarization ratios of 10%, 20% and 30%.

3. SUMMARIZERS

In this paper we investigate speech summarization based on two different probabilistic ranking models: support vector machine (SVM) and word topic model (WTM). The two different summarizers associated with their baseline results are also briefly described in the section.

3.1. Supervised Summarizer - SVM

SVM attempts to find an optimal hyper-plane by utilizing a decision function that can correctly separate the positive and negative samples, and ensure that the margin is maximal when the dataset is linearly separable. In a non-linearly separable case, SVM uses kernel functions or defines slack variables to transform the problem into a linear discrimination problem. In this paper, we constructed a binary SVM summarizer with the radial basis function (RBF) as the kernel function. The posterior probability of a sentence S_i being included in the summary can be approximated by following sigmoid operation:

$$P_{\mathrm{SVM}}(S_i \in \mathbf{S} \mid X_i) \approx \frac{1}{1+\exp(\alpha \cdot g(X_i)+\beta)}, \quad (1)$$

where X_i is a set of features used to characterize S_i; α and β are weights that are estimated from the development set by minimizing a negative log-likelihood function; and $g(X_i)$ is the decision value of X_i provided by the SVM summarizer.

For the SVM summarizer, we use a set of 19 features to characterize a spoken sentence, including the structural features (St), the lexical features (Le), the acoustic features (Ac), and the relevance features (Re). The features are outlined in Table 1, where each of them is further normalized to zero mean and unit variance.

3.2. Unsupervised Summarizer – WTM

We presented an unsupervised probabilistic generative framework for speech summarization recently [8, 11]. Each sentence S_i of a spoken document D is treated as a language model for generating D, and the sentences are ranked and selected according to their posterior probability $P(S_i|D)$, which can be expressed by

$$P(S_i|D) = \frac{P(D|S_i)P(S_i)}{P(D)}, \quad (2)$$

where $P(D|S_i)$ is the sentence generative probability, i.e., the likelihood that D is generated by S_i; $P(S_i)$ is the prior probability of S_i being important, which was set to uniform in our previous study [8]; and $P(D)$ is the probability of D. The sentence generative probability $P(D|S_i)$ can be taken as a relevance measure between the document and its sentences.

In order to estimate the sentence generative probability, we exploited a concept matching strategy [1], called word topic model (WTM) [13] for such purpose. Each word w_j of the language is treated as a model M_{w_j} consisting of a set of latent topics for predicting the occurrences of the other word w:

$$P_{\mathrm{WTM}}(w \mid \mathrm{M}_{w_j}) = \sum_{k=1}^{K} P(w \mid T_k) P(T_k \mid \mathrm{M}_{w_j}), \quad (2)$$

where $P(w \mid T_k)$ and $P(T_k \mid \mathrm{M}_{w_j})$ are, respectively, the probability of a word w occurring in a specific latent topic T_k and the probability of a topic T_k conditioned on M_{w_j}. During the summarization process, we can linearly combine the associated WTM models of the words involved in a sentence S_i to form a composite word topic model for S_i, and the likelihood of the document D being generated by S_i can be expressed as:

$$P_{\mathrm{WTM}}(D \mid S_i) = \prod_{w \in D} \left[\sum_{w_j \in S_i} P_{\mathrm{WTM}}(w \mid \mathrm{M}_{w_j}) P_{\mathrm{ML}}(w_j \mid S_i) \right]^{n(w,D)}, \quad (3)$$

where $n(w,D)$ is the number of times that a specific word w occurring in D; $P_{\mathrm{ML}}(w_j \mid S_i)$ is estimated according the frequency of w_j in S_i. In this paper, we investigated an unsupervised approach for training $P_{\mathrm{WTM}}(w \mid \mathrm{M}_{w_j})$, which is accomplished by concatenating the words occurring within a context window of size *m* around each occurrence of w_j in the collection [13]. We postulate that these contextual words are relevant to w_j, and can therefore be used as an observation for training M_{w_j} with the expectation-maximization (EM) algorithm.

Structural features (St)	*POSITION*: Sentence position *DURATION*: Duration of the preceding/current/ following sentence
Lexical Features (Le)	*BIGRAM_SCORE*: Normalized bigram language model scores *SIMILARITY*: Similarity scores between a sentence and its preceding/following sentence *NUM_NAME_ENTITIES*: Number of named entities (NEs) in a sentence
Acoustic Features (Ac)	*PITCH*: Min/max/mean/difference pitch values of a spoken sentence *ENERGY*: Min/max/mean/difference value of energy features of a spoken sentence *CONFIDENCE*: Posterior probabilities
Relevance Features (Re)	*R-VSM*: Relevance score obtained by using the VSM summarizer [1] *R-LSA*: Relevance score obtained by using the LSA summarizer [1]

Table 1: The features used for SVM summarizer.

3.3. Baseline Performance

Table 2 shows the baseline performance of SVM and WTM for different summarization ratios. For WTM, the sentence prior probability $P(S_i)$ is assumed to be uniform, whereas an account on the impact of using a non-uniform sentence prior probability will be given later. The results based on manual transcripts of the spoken documents (denoted by TD, text documents) are also listed in Table 2 for reference, in addition to the results based on the recognition transcripts (denoted by SD, spoken documents). As can be seen, SVM significantly outperforms WTM, and yields the results that are comparable to those obtained by the human subjects for the TD case, except for the case of a very low summarization ratio, e.g., 10%. The superiority of SVM over WTM might be explained by two factors. The first is that SVM makes use of the handcrafted (or supervised) document-summary information for model training, whereas WTM does not utilize such information. The second is that the WTM relies merely on word and topic unigram probabilities, whereas SVM fuses more indicative features besides the lexical features (including bigrams) to fulfill spoken document summarization. Nevertheless, almost all kinds of these features are more or less vulnerable to speech recognition errors. Thus, SVM shows larger performance difference between the TD and SD cases than WTM that merely use lexical features.

		Summarization Ratio		
		10%	20%	30%
TD	SVM	0.548	0.625	0.632
	WTM	0.359	0.483	0.517
SD	SVM	0.329	0.361	0.350
	WTM	0.205	0.247	0.282

Table 2: The results achieved by different summarizers under different summarization ratios.

		Summarization Ratio		
		10%	20%	30%
TD	Augmented	0.596	0.646	0.635
	Combined	0.599	0.652	0.649
SD	Augmented	0.338	0.366	0.354
	Combined	0.344	0.368	0.356

Table 3: The results achieved by combing SVM and WTM.

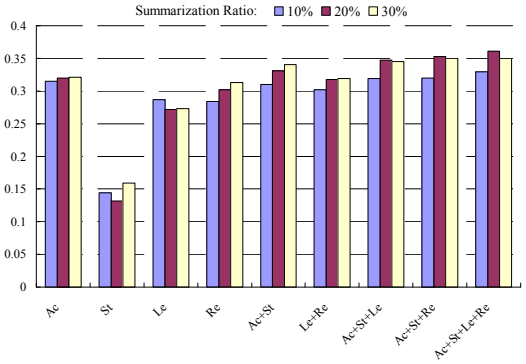

Figure 1: The summarization results of the SD case achieved by SVM with different features and their combinations.

In the next set of experiments, we examine the contributions that different kinds of features (cf. Table 1) made to the performance of SVM for the SD case. As illustrated by Figure 1, the summarization performance can be improved steadily by including a substantial number of indicative features. The acoustic features make more substantial contributions to the performance than the lexical features and the relevance features. Combining the acoustic features and structural features together outperforms combing lexical features and relevance features together. These results show that lexical cues might not be the dominating predictors when summarization is carried out with erroneous recognition transcripts. The results also reveal that relevance features are more effective than lexical features.

In the third set of experiments, we attempt to combine SVM with WTM. The combination can be conducted in two alternative ways. The first is to directly take the score obtained by WTM as an additional feature to augment the feature set defined in Table 1 (denoted by "Augmented"). The second is to compute the sentence prior probability $P(S_i)$ of WTM on the basis of the decision probability $P(S_i \in \mathbf{S} \mid X_i)$ provided by SVM (denoted by "Combined"). Therefore, $P(S_i)$ can be estimated using a variety of features instead of being simply set to uniform. As reported in Table 3, both these two combinations significantly boost the summarization performance especially at the summarization ratio of 10%, as compared to the baseline performance of SVM or WTM shown in Table 2. These results seem to justify our postulation that supervised and unsupervised summarizers may complement each other.

4. TRAINING SVM WITH LABELS DERIVED BY WTM

As stated in the previous sections, the supervised summarizer such as SVM suffers from several limitations, including the need of document-reference summary pairs for either model training or task migration. In contrast, the unsupervised summarizer such as WTM usually considers the relevance of a sentence to the whole document, which might be more robust across different summarization tasks. In this paper, we investigate the use of WTM to improve the performance of SVM under the condition that the handcrafted document-reference summary pairs are not available for training the latter. Nevertheless, as will be discussed in Section 4.3, it was experimentally observed that the performance of SVM trained simply with the document-summary labels derived automatically from WTM would be worse than the original performance of WTM. Therefore, how to filter out unreliable automatic labels or collect more reliable automatic labels for training SVM without supervision is deemed to be an important issue for reducing the performance gap. To this end, we propose a training data selection and relabeling approach, which leverages either the inter-document or the inter-sentence similarity information, to filter out unreliable labels or collect more reliable automatic labels for training SVM without supervision.

4.1. Inter-Document Similarity Information

The inter-document similarity (IDS) of a sentence S_i is defined by the average similarity of documents in the relevant news document set $\mathbf{R}$ of S_i, where $\mathbf{R}$ is obtained by taking S_i as a query and posing it to an information retrieval (IR) system to obtained a list of M most relevant documents from a contemporaneous news document repository [11]. Our assumption is that the relevant news documents retrieved for a summary sentence might have the same or similar topics because a summary sentence is usually indicative for some specific topic related to the document. In contrast, the relevant text documents retrieved for a non-summary sentence might cover diverse topics. In other words, the IDS information estimated based on the similarity of documents in the relevant news document set might be a good indicator for determining the importance of a sentence. Consequently, we can select or collect more reliable summary/non-summary sentences for training the supervised summarizers based on such information. The average similarity of documents in the relevant news document set $\mathbf{R}$ for a sentence S_i is computed by

$$avgSim(S_i) = \frac{\sum_{D_l \in \mathbf{R}} \sum_{\substack{D_u \in \mathbf{R} \\ D_l \neq D_u}} \frac{\vec{D}_l \cdot \vec{D}_u}{\|\vec{D}_l\| \cdot \|\vec{D}_u\|}}{M \cdot (M-1)}, \quad (5)$$

where $\vec{D}_l$ is the TF-IDF vector representation of a document D_l, and M is the number of documents in the retrieved relevant news document set $\mathbf{R}$.

4.2. Inter-Sentence Similarity Information

As opposed to the IDS information, the inter-sentence similarity (ISS) of a sentence S_i is derived from the concept of the centrality among all sentences in a document. To be specific, if a sentence S_i is more similar to other sentences in a document, it might be a representative sentence and can be used to depict the main theme of

		Summarization Ratio		
		10%	20%	30%
IDS	Summary	0.059	0.057	0.055
	Non-Summary	0.047	0.046	0.045
ISS	Summary	0.826	0.822	0.821
	Non-Summary	0.475	0.473	0.473

Table 4: The averages of the inter-document and inter-sentence similarity for the manual summary/non-summary sentences of the development set at different summarization ratios.

		Summarization Ratio		
Labeling		10%	20%	30%
TD	WTM	0.288	0.476	0.507
	+IDS	0.350	0.499	0.529
	+ISS	0.380	0.502	0.551
SD	WTM	0.171	0.259	0.287
	+IDS	0.214	0.265	0.295
	+ISS	0.190	0.267	0.313

Table 5: The summarization results achieved by SVM trained without supervision.

the document. The notion is similar to that of the graph-based summarization methods [9, 10], which try to find prestige sentences in the conceptualized network of a document where each node represents a sentence and the associated weight of each link represents the similarity relationship between a pair of nodes. Here, the LexRank algorithm [10] was modified for the purpose of deriving ISS. We adopted WTM to compute the topical similarity between sentences, which differs from the literal term matching strategy (the cosine measure) exploited by LexRank. The main reason is that since sentences in a spoken document usually involve only few words, the literal term matching strategy will result in a zero score when there is no common word shared between a pair of sentences, even though they describe the same concept. Another reason is the similarity between a pair of sentences estimated by WTM can be asymmetric; that is, the network will have directed links. After the modified LexRank algorithm has been conducted on the conceptualized network of a document, the associated normalized similarity score of each sentence can be taken as its ISS.

4.3. Experimental Results and Discussions

The averages of the IDS and ISS scores for the handcrafted summary and non-summary sentences of the development set are shown in Table 4. The IDS and ISS scores for summary sentences are higher than those of non-summary sentences, respectively. These observations indeed support our postulation that either IDS or ISS might be helpful for filtering out unreliable labels or collecting more reliable automatic labels. In this paper, if a sentence is labeled by WTM as a summary/non-summary sentence and has the IDS or ISS score higher/lower than threshold τ_S / τ_{NS}, it will be marked as a reliable summary/non-summary sentence. The reliable sentences will be ultimately selected for training SVM, while the remaining ones are discarded instead. Along a similar vein, if a sentence is labeled by WTM as a summary/non-summary sentence and has the IDS or ISS score lower/higher than a threshold ρ_S / ρ_{NS}, its label will be changed. Our empirical observations reveal that IDS can accommodate itself to training data selection, while ISS is superior for relabeling of training data. One possible explanation is that IDS considers the topic convergence of a sentence, a generic property of a summary sentence, which does not always imply the sentence is the closest sentence to the main theme of a document. On the contrary, ISS captures the similarity of a sentence to other sentences in the document to be summarized; no doubt, it would work properly on relabeling the training sentences. As evident in Table 5, both IDS and ISS can substantially enhance the performance of SVM trained without supervision.

5. CONCLUSIONS

In this paper, we have presented two alternatives ways to combine supervised summarizers with unsupervised summarizers. The experimental results seem to reveal that they both can significantly boost the summarization performance. In addition, we have also proposed the use of the inter-document and the inter-sentence similarity for training data selection and relabeling, respectively. Encouraging results were initially demonstrated. We believe that this initial attempt could provide a new avenue for future research on speech summarization. Our future research directions include: 1) incorporating more syntactic, semantic, and prosodic features for supervised summarizers 2) investigating efficient feature selection algorithms to select more indicative features and 3) seeking other ways to filter out the unreliable (or noisy) labels.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Science Council, Taiwan, under Grants: NSC96-2628-E-003-015-MY3, NSC95-2221-E-003-014-MY3, and NSC97-2631-S-003-003.

7. REFERENCES

[1] L.S. Lee, B. Chen., "Spoken document understanding and Organization," *IEEE Signal Processing Magazine* 22(5), 2005.

[2] C. Chelba et al, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine* 25(3), 2008.

[3] M. Witbrock and V. Mittal, "Ultra summarization: a statistical approach to generating highly condensed non-extractive summaries," *in Proc. SIGIR 1999*.

[4] M. Hirohata et al., "Sentence extraction-based presentation summarization techniques and evaluation metrics", *in Proc. ICASSP 2005*.

[5] Y. Gong, X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," *in Proc. ACM SIGIR 2001*.

[6] G. Murray et al., "Extractive summarization of meeting recordings," *in Proc. INTERSPEECH 2005*.

[7] S. Furui et al., "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Trans. on Speech Audio Processing* 12(4), 2004.

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

[9] M. Rada and T. Paul, "TextRank: bringing order into texts," *in Proc. EMNLP 2004.*

[10] G. Erkan and D. R. Radev, "LexRank: graph-based lexical centrality as salience in text summarization," *Journal or Artificial Intelligence Research* 22, 2004.

[11] Y. T. Chen et al., "A probabilistic generative framework for extractive broadcast news speech summarization," *IEEE Trans. on Audio, Speech and Language Processing* 17(1), 2009.

[12] C.Y. Lin, "ROUGE: recall-oriented understudy for gisting evaluation," http://www.isi.edu/~cyl/ROUGE/, 2003.

[13] B. Chen, "Word topic models for spoken document retrieval and transcription" *ACM Trans. on Asian Language Information Processing*, 8(1), 2009.