INTERSPEECH 2009 BRIGHTON

Improved Speech Summarization with Multiple-Hypothesis Representations and Kullback-Leibler Divergence Measures

Shih-Hsiang Lin and Berlin Chen

Department of Computer Science & Information Engineering
National Taiwan Normal University, Taipei, Taiwan
{shlin, berlin}@csie.ntnu.edu.tw

Abstract

Imperfect speech recognition often leads to degraded performance when leveraging existing text-based methods for speech summarization. To alleviate this problem, this paper investigates various ways to robustly represent the recognition hypotheses of spoken documents beyond the top scoring ones. Moreover, a new summarization method stemming from the Kullback-Leibler (KL) divergence measure and exploring both the sentence and document relevance information is proposed to work with such robust representations. Experiments on broadcast news speech summarization seem to demonstrate the utility of the presented approaches.

Index Terms: speech summarization, multiple recognition hypotheses, KL divergence, relevance information

1. Introduction

Extractive summarization produces a summary by selecting salient sentences from an original document according to a predefined target summarization ratio. The wide spectrum of extractive summarization approaches that have been developed so far may roughly fall into three main categories [1-2]: 1) approaches based on the sentence structure or location information, 2) approaches based on proximity or significance measures, and 3) approaches based on sentence classification.

For the first category, the important sentences can be selected from the significant parts of a document, e.g., sentences can be selected form the introductory and/or concluding parts. However, such approaches can be only applied to some specific domains or document structures. In contrast, approaches based on proximity or significance measures attempt to select salient sentences based on the statistical features of the sentences or the words in the sentences, such as the term frequency (TF), the inverse document frequency (IDF), the *N*-gram scores, and the topic or semantic information. The associated methods based on these features have gained much attention of research. Besides, a number of classification-based methods using statistical features and/or sentence structure (or position) information also have been developed, such as the Gaussian mixture models (GMM), the Bayesian classifier (BC), the support vector machine (SVM) and the conditional random fields (CRFs). In these methods, important sentence selection is usually formulated as a binary classification problem. A sentence can either be included in a summary or not. These classification-based methods need a set of training documents together with their corresponding handcrafted summaries (or labeled data) for training the classifiers (or summarizers). However, manual annotation is expensive in terms of time and personnel. Even if the performance of unsupervised summarizers is not always comparable to that of supervised summarizers, their easy-to-implement and portable property still makes them attractive [3].

Although most of the above approaches can be equally applied to both text and spoken documents, the latter presents unique difficulties, such as speech recognition errors, problems with spontaneous speech, and the lack of correct sentence or paragraph boundaries. It has been shown that speech recognition errors are the dominating factor for the performance degradation of speech summarization when using recognition transcripts instead of manual transcripts, whereas erroneous sentence boundaries cause relatively minor problems [4, 5]. A straightforward remedy, apart from the many approaches improving recognition accuracy, might be to develop more robust representations for spoken documents. For example, multiple recognition hypotheses, beyond the top scoring ones, are expected to provide alternative representations for the confusing portions of the spoken documents [6, 7]. Moreover, the use of subword units, as well as the combination of words and subword units, for representing the spoken documents should be beneficial for speech summarization.

In this paper, we investigate various ways to robustly represent the recognition hypotheses of spoken documents, including the use of the confusion network (CN) [8] and the position specific posterior lattice (PSPL) [6], for the summarization purpose. Moreover, a new summarization method originating from the Kullback-Leibler (KL) divergence measure [9] and exploring both the sentence and document relevance information is proposed to work with such robust representations.

2. Summarization Method

Extractive summarization produces a concise summary by selecting salient sentences or paragraphs from an original document according to a predefined target summarization ratio. Conceptually, it could be cast as an ad hoc information retrieval (IR) problem, where the document is treated as an information need and each sentence of the document is regarded as a candidate information unit to be retrieved according to its relevance (or importance) to the information need. Therefore, the ultimate goal of extractive summarization could be stated as the selection of the most representative sentences that can succinctly describe the main concepts of the document. In the past several years, the language modeling approaches have been introduced to IR problems and demonstrated with good empirical success [9]; this modeling paradigm has also been adopted for speech summarization recently [2].

In this paper, we present a novel summarization model, stemming from the KL-divergence measure, for important sentence selection, which models the relationship between the sentences of a document to be summarized and the document itself from an information-theoretic perspective. To this end, two different language models are estimated: one for the whole document and the other for each sentence. We assume that words in the document are simple random draws from a

Copyright © 2009 ISCA

language distribution describing some topics of interest and words in the sentences that belong to the summary should also be drawn from the same distribution. Therefore, we can use KL-divergence to measure how close the document and its sentences are: the closer the sentence model $P(w|\theta_S)$ to the document model $P(w|\theta_D)$, the more likely the sentence would be part of the summary. The KL-divergence of the sentence model with respect to the document model is defined by

$$KL(\theta_D \| \theta_S) = \sum_{w \in V} P(w|\theta_D) \log \frac{P(w|\theta_D)}{P(w|\theta_S)} \quad (1)$$

where w denotes a specific word in the vocabulary set V; and a sentence S has a smaller value of $KL(\theta_D \| \theta_S)$ is deemed to be more important. The important sentence ranking problem has now been reduced the problem of estimating the two unigram (or multinomial) models $P(w|\theta_D)$ and $P(w|\theta_S)$. The simplest way is to estimate these two models on the basis of the frequency of words occurring in the document D and in each sentence S, respectively, with the maximum likelihood estimator (MLE). It is noteworthy that the KL-divergence model expressed in Eq. (1) can be viewed as a generalization of our previously proposed document-likelihood based scoring method for speech summarization [2]. This method has the merit of being able to accommodate relevance feedback information to improve summarization accuracy in a systematic way.

For instance, the true document or sentence model might not be accurately estimated by MLE, since either the document or the sentence consists of only a few words and the portions of words present are not the same as the probabilities of words in the true models. Therefore, we can explore the use of the relevance model (RM) to get a more accurate estimation of the document or sentence model [9]. To illustrate, we take the sentence model as an example. Each sentence S of the spoken document D to be summarized has its own associated relevance class R_S. This class is defined as the subset of documents in the collection that are relevant to the sentence S. The relevance model of the sentence S is defined to be the probability distribution $P(w|\theta_{R_S})$, which gives the probability that we would observe a word w, if we were to randomly select a document from the relevance class R_S and then pick up a random word from that document. Once the relevance model of the sentence S is constructed, it can be used to replace the original sentence model or to be combined with the original sentence model to produce a more accurate estimate. Because there is no prior knowledge about the subset of relevant documents for each sentence S, a local relevance feedback-like procedure can be employed by taking S as a query and posing it to an IR system to obtain a ranked list of documents. The top L documents returned from the IR system are assumed to be the ones relevant to S, and the relevance model of S can be therefore constructed through the following equation:

$$P(w|\theta_{R_S}) = \sum_{D_l \in \mathbf{D}_{\text{Top } L}} P(D_l|S) P(w|\theta_{D_l}) \quad (2)$$

where $\mathbf{D}_{\text{Top } L}$ is the set of the top L retrieved documents; and the probability $P(D_l|S)$ can be approximated by the following equation using Bayes' rule:

$$P(D_l|S) = \frac{P(D_l) \cdot P(S|\theta_{D_l})}{\sum_{d_u \in \mathbf{D}_{\text{Top } L}} P(D_u) \cdot P(S|\theta_{D_u})} \quad (3)$$

A uniform prior probability $P(D_l)$ can be further assumed for the top L retrieved documents, and $P(S|\theta_{D_l})$ is the sentence likelihood (or retrieval score). After obtaining the relevance model, we adapt a two-stage smoothing method to form the final sentence model:

$$\tilde{P}(w|\theta_S) = \lambda \cdot \left(\frac{c(w,S) + \mu \cdot P(w|\theta_{R_S})}{|S| + \mu} \right) + (1-\lambda) \cdot P(w|\theta_C) \quad (4)$$

where $|S|$ is the length of S; $c(w,S)$ is the frequency count of w in S; $P(w|\theta_C)$ is a collection (or background) language model estimated from the a large document collection for reflecting the general word frequencies in the language; μ is a smoothing parameter that can be further estimated by maximizing the leave-one-out log likelihood of the retrieved document set [10]. Finally, a bit of terminology: Eq. (4) can be regarded as a combination of the Bayesian smoothing (with a Dirichlet prior) and the Jelinek-Mercer smoothing. The Dirichlet prior gives more emphasis on the discriminative (or informative) word and the Jelinek-Mercer is used to absorb the common and non-informative words. Along a similar vein, the relevance model $P(w|\theta_{R_D})$ for the spoken document D can be constructed as well.

3. Robust Document Representations

A word lattice is usually served as an intermediate representation of the automatic speech recognition (ASR) output. It is a connected, directed acyclic graph where each arc includes a word hypothesis along with a posterior probability (combining acoustic and language model scores) as well as the time alignment information. It provides a large set of alternative recognition hypotheses, and each path from the start node to the exit node stands for one hypothesis of the spoken word sequence. However, since a word lattice often contains many confusing word hypotheses (e.g., word arcs with very low posterior probabilities) and costs huge storage space, various compact representations of the word lattice have been developed [6, 8]. In this paper, we investigate the use of CN and PSPL for representing spoken documents and sentences, for each of which word arcs of the word lattice are binned into several strictly linear clusters based on the time span of the arcs, and each cluster consists of a list of competing word hypotheses along with their corresponding posterior probabilities, to represent the confusing portions. The sentence boundaries are first determined with the 1-best ASR transcript of the spoken document, and each sentence is then treated as a speech segment $\mathbf{o}$ for generating its own word lattice.

3.1. Confusion Network (CN)

A confusion network [8] is a multiple string alignment of the speech recognition results, which transforms all hypotheses in a word lattice into a sequence of linear clusters. The original purpose of CN is used to minimize the expected word errors by concatenating those words having the highest posterior probability in each cluster (or confusion set) to form the recognition output, where the posterior probability of each word hypothesis in a cluster can be also thought as the expected word count. In implementation, the transformation of a CN from a word lattice is fulfilled by a two-stage clustering procedure. The first stage is *intra-word clustering* where word arcs with the same word identity are grouped into together based on their time overlaps and word posterior probabilities. The second stage then performs *inter-word clustering* where several heterogeneous clusters are grouped together according to their phonetic similarity [8].

3.2. Position Specific Posterior Lattice (PSPL)

The basic idea of PSPL is to calculate the posterior probability of a word w occurring at a specific position l in a word

lattice. Since it is likely that more than one path contains the same word, to compute the expected count of a word w at a specific position l in the lattice, one would need to sum over all possible paths in a lattice that contain w at l. This computation can be accomplished by employing a modified forward-backward algorithm. For the forward search, the forward probability $\alpha(w,t)$ is split into several more subtle probability masses $\alpha(w,t,l)$ according to the length of partial paths that start from the start node and end at w; while the procedure of the backward search remains unchanged. Finally, the posterior probability of a given word w occurring at a given position l in a lattice can be easily calculated [6].

3.3. Pruning and Expected Count Computation

After the construction of CN or PSPL, a simple pruning procedure is adopted to remove the unlikely word hypotheses (i.e., words with lower posterior probabilities) [6]. For each cluster (or position) l, the pruning procedure first finds the most likely word entry in it. Then, those word entries that have log posterior probabilities lower than that of the most likely one minus a predefined threshold τ are then removed from l. Finally, we can compute the expected frequency count of each word w in a given speech segment $\mathbf{o}$:

$$\mathrm{E}\left[c(w,\mathbf{o})\right] = \sum_l \sum_{w_l} P(w_l = w \mid \mathrm{LAT}) \quad (5)$$

where w_l is an arbitrary word that occurs in cluster (or at position) l; LAT denotes CN (or PSPL); $P(w_l = w \mid \mathrm{LAT})$ denotes the posterior probability of word w in cluster (or at position) l.

4. Experiments

4.1. Experimental Setup

All the summarization experiments were conducted on a set of 205 broadcast news documents compiled from the MATBN corpus [3]. Three subjects 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 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%.

To assess the goodness of the automatic generated summaries, we use the ROUGE evaluation [11], 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 the matching units. The summarization results were evaluated by using several summarization ratios (10%, 20%, and 30%), 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.

4.2. Experimental Results

We first evaluate the utility of using CN and PSPL for representing spoken documents. The vector space method (VSM) is employed as the default summarization method, since it usually achieves quite comparative results as compared to other unsupervised methods [2-3]. VSM represents each sentence and the whole document in vector form, where each dimension specifies the product of the TF and IDF scores associated with a word in the sentence (or document). Sentences that have the highest cosine or proximity scores to the whole document will be included in the summary.

VSM	Summarization Ratio		
	10%	20%	30%
Text	0.313	0.405	0.464
1-best	0.175 (0.252)	0.212 (0.310)	0.261 (0.374)
CN	0.198 (0.285)	0.232 (0.327)	0.259 (0.371)
PSPL	0.234 (0.313)	0.260 (0.368)	0.285 (0.401)

Table 1: The Rogue_2 summarization results achieved by the vector space method under different summarization ratios.

KL-divergence	Summarization Ratio		
	10%	20%	30%
Text	0.360	0.407	0.454
1-best	0.163 (0.242)	0.209 (0.307)	0.251 (0.363)
CN	0.215 (0.316)	0.237 (0.340)	0.265 (0.382)
PSPL	0.246 (0.339)	0.263 (0.370)	0.293 (0.415)

Table 2: The Rogue_2 summarization results achieved by the KL-divergence method under different summarization ratios.

In Table 1, Row "Text" shows the results obtained by using manual transcripts of the spoken document and sentences for sentence ranking, while Rows "1-best", "CN" and "PSPL" are the results obtained by using the 1-best ASR transcripts, CN and PSPL representations, respectively. Also noteworthy is that when performing the ROUGE evaluation, for both CN and PSPL, only the top scoring word sequence derived from them was used to compare to the manual summaries. As can be seen, there are significant performance gaps between summarization using the manual transcripts and the 1-best ASR transcripts; however, summarization using either CN or PSPL indeed can provide substantial performance boosts over the 1-best ASR transcripts. Moreover, PSPL seems to outperform CN for the purpose of speech summarization. On the other hand, the resulting summary sentences can also be present in speech form (besides text form) to bypass the problem caused by speech recognition errors [12]. In order to simulate such a scenario as well as to assess the performance of the proposed approaches on it, we therefore align the ASR transcripts of the summary sentences to their respective waveform segments to obtain the correct (manual) transcripts for evaluation. The corresponding results are shown in the parentheses of Table 1, which reveal that with aid of PSPL, we can achieve almost the same performance level as that using manual transcripts, when the summarization ratio is lower (e.g., 10%) and the resulting summary is present in speech form.

In the next set of experiments, we evaluate the performance of the proposed KL-divergence summarization method, as well as its integration with various representations of spoken documents; the corresponding results are illustrated in Table 2. The KL-divergence method dose not seem to outperform VSM (cf. Table 1) when the sentence and document models were estimated merely based on the 1-best ASR transcripts. One possible explanation is that the recognition errors contained in the 1-best ASR transcripts would seriously hurt the accuracy of model estimation. On the contrary, the KL-divergence method can yield superior results when compared to VSM, if either CN or PSPL is adopted for representing the spoken sentences and the spoken document. From the results shown in Tables 1 and 2, we can confirm that speech summarization can benefit greatly by the introduction of CN and PSPL for robust spoken sentence and document representations.

KL-divergence	Summarization Ratio		
	10%	20%	30%
PSPL	0.246 (0.339)	0.263 (0.370)	0.293 (0.415)
+ RM-SEN	0.255 (0.343)	0.271 (0.381)	0.294 (0.415)
+ RM-DOC	0.246 (0.339)	0.266 (0.378)	0.297 (0.420)
+ RM-SEN + RM-DOC	0.257 (0.344)	0.277 (0.386)	0.304 (0.429)

Table 3: The Rogue_2 summarization results achieved by combing the KL-divergence method with the sentence or/and document relevance information.

To go a step further, we explore the use of relevance feedback (cf. Section 2) for more accurate estimation of the sentence models (denoted by RM-SEN) and the document model (denoted by RM-DOC) in the KL-divergence method. Here we use PSPL for document and sentence representations since it achieved the best performance in the previous experiments. As reported in Table 3, the summarization performance is consistently improved at lower summarization ratios when the sentence relevance information (RM-SEN) is used for model estimation; however, only moderate improvements are observed for using the document relevance information (RM-DOC). This can be explained by the fact that the spoken sentences are quite short when compared to the spoken document, and thus require more statistical evidence contributed from the relevant documents for better sentence model estimation. Moreover, integration of RM-SEN and RM-DOC together into in the KL-divergence method can provide additional gains (cf. the last row in Table 3), which leads to absolute improvements of about 9%, 7% and 5%, respectively, for summarization ratios of 10%, 20%, and 30% as compared to the results by using the 1-best ASR transcripts (cf. the second row in Table 2).

Finally, we consider using subword-level and topical information to improve the summarization performance. The corresponding results are shown in Table 4. Here we use overlapping syllable bigrams as the subword units. The reason for fusion of word- and subword-level information (denoted by SUB) for representing the spoken document and sentences is that, incorrectly recognized spoken words often include several subword units correctly recognized, and important sentence selection based on subword-level representations hence may take advantage of partial matching. On the other hand, there probably would be word usage mismatch between the spoken document and a spoken sentence even if they are topically related to each other. Consequently, we can exploit the probabilistic topic models [9, 13] to represent the spoken document and sentences, in addition to the existing document and sentence models that are constructed based on literal term information, as described in Section 2, in the KL-divergence method (denoted by TOP). As can be seen, further inclusion of either subword-level (SUB) or topical (TOP) information can provide additional performance gains when summarization is conducted on the 1-best ASR transcripts. However, combining subword-level information with PSPL provides almost negligible improvements over that using PSPL alone. This may be probably due to the fact that PSPL, to some extent, contains sufficient lexical information for estimating the document and sentence models.

KL-divergence	Summarization Ratio		
	10%	20%	30%
1-best	0.163 (0.242)	0.209 (0.307)	0.251 (0.363)
+ SUB	0.170 (0.250)	0.213 (0.310)	0.253 (0.370)
+ TOP	0.178 (0.255)	0.216 (0.318)	0.255 (0.370)
PSPL	0.246 (0.339)	0.263 (0.370)	0.293 (0.415)
+ SUB	0.247 (0.341)	0.266 (0.375)	0.293 (0.418)
+ TOP	0.250 (0.348)	0.268 (0.380)	0.296 (0.420)

Table 4: The Rogue_2 summarization results achieved by integrating subword-level or topical information into the KL-divergence method.

5. Conclusions

In this paper, we have investigated various ways to robustly represent spoken documents and sentences for speech summarization. We have also proposed a KL-divergence-based summarization method and conducted a series of experiments to test its capability. The experimental results indeed confirm our expectation. Our future research directions include: 1) investigating more elaborate approaches to estimate the document and sentence models, 2) seeking other ways to represent the ASR output more robustly, and 3) incorporating the summarization results into audio indexing for better retrieval and browsing of spoken documents.

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] Furui, S., "Recent Advances in Automatic Speech Summarization," *in Proc. SLT 2006.*

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

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

[4] Christensen, H. et al., "A Cascaded Broadcast News Highlighter," *IEEE Transactions on Audio, Speech, and Language Processing* 16(1), 2008.

[5] Liu, Y. and Xie, S., "Impact of Automatic Sentence Segmentation on Meeting Summarization," *in Proc. ICASSP 2008*.

[6] Chelba, C. et al., "Soft Indexing of Speech Content for Search in Spoken Documents," *Computer Speech and Language* 21, 2007.

[7] Chia, T. K., et al., "A Lattice-based Approach to Query-by Example Spoken Document Retrieval," *in Proc. ACM SIGIR 2008*.

[8] Mangu, L., "Finding consensus in speech recognition: word error minimization and other applications of confusion network," *Computer Speech and Language* 14, 2000.

[9] Zhai, C. X., Statistical Language Models for Information Retrieval (Synthesis Lectures Series on Human Language Technologies), Morgan & Claypool Publishers, 2008.

[10] Zhai, C. X., et al., "A study of Smoothing Methods for Language Model Applied to Information Retrieval," *ACM Trans. on Information Systems* 22(2), 2004.

[11] Lin, C.Y., "ROUGE: Recall-oriented Understudy for Gisting Evaluation," http://www.isi.edu/~cyl/ROUGE/, 2003.

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

[13] Chen, B., "Latent topic modeling of word co-occurrence information for spoken document retrieval," *in Proc. ICASSP 2009*.