Summarizing Speech Without Text Using Hidden Markov Models

Sameer Maskey, Julia Hirschberg
Dept. of Computer Science
Columbia University
New York, NY
{smaskey, julia}@cs.columbia.edu

Abstract

We present a method for summarizing speech documents without using any type of transcript/text in a Hidden Markov Model framework. The hidden variables or states in the model represent whether a sentence is to be included in a summary or not, and the acoustic/prosodic features are the observation vectors. The model predicts the optimal sequence of segments that best summarize the document. We evaluate our method by comparing the predicted summary with one generated by a human summarizer. Our results indicate that we can generate ’good’ summaries even when using only acoustic/prosodic information, which points toward the possibility of text-independent summarization for spoken documents.

1 Introduction

The goal of single document text or speech summarization is to identify information from a text or spoken document that summarizes, or conveys the essence of a document. EXTRACTIVE SUMMARIZATION identifies portions of the original document and concatenates these segments to form a summary. How these segments are selected is thus critical to the summarization adequacy.

Many classifier-based methods have been examined for extractive summarization of text and of speech (Maskey and Hirschberg, 2005; Christensen et. al., 2004; Kupiec et. al., 1995). These approaches attempt to classify segments as to whether they should or should not be included in a summary. However, the classifiers used in these methods implicitly assume that the posterior probability for the inclusion of a sentence in the summary is only dependent on the observations for that sentence, and is not affected by previous decisions. Some of these (Kupiec et. al., 1995; Maskey and Hirschberg, 2005) also assume that the features themselves are independent. Such an independence assumption simplifies the training procedure of the models, but it does not appear to model the factors human beings appear to use in generating summaries. In particular, human summarizers seem to take previous decisions into account when deciding if a sentence in the source document should be in the document’s summary.

In this paper, we examine a Hidden Markov Model (HMM) approach to the selection of segments to be included in a summary that we believe better models the interaction between extracted segments and their features, for the domain of Broadcast News (BN). In Section 2 we describe related work on the use of HMMs in summarization. We present our own approach in Section 3 and discuss our results in Section 3.1. We conclude in Section 5 and discuss future research.

2 Related Work

Most speech summarization systems (Christensen et. al., 2004; Hori et. al., 2002; Zechner, 2001) use lexical features derived from human or Automatic Speech Recognition (ASR) transcripts as features to select words or sentences to be included in a summary. However, human transcripts are not generally available for spoken documents, and ASR transcripts are errorful. So, lexical features have practical limits as a means of choosing important segments for summarization. Other research efforts have focussed on text-independent approaches to extractive summarization (Ohtake et. al., 2003), which rely upon acoustic/prosodic cues. However, none of these efforts allow for the context-dependence of extractive summarization, such that the inclusion of

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 89–92,
New York, June 2006. ©2006 Association for Computational Linguistics

one word or sentence in a summary depends upon prior selection decisions. While HMMs are used in many language processing tasks, they have not been employed frequently in summarization. A significant exception is the work of Conroy and O'Leary (2001), which employs an HMM model with pivoted QR decomposition for text summarization. However, the structure of their model is constrained by identifying a fixed number of 'lead' sentences to be extracted for a summary. In the work we present below, we introduce a new HMM approach to extractive summarization which addresses some of the deficiencies of work done to date.

3 Using Continuous HMM for Speech Summarization

We define our HMM by the following parameters: $\Omega = 1..N$: The state space, representing a set of states where N is the total number of states in the model; $O = o_{1k}, o_{2k}, o_{3k}, ...o_{Mk}$: The set of observation vectors, where each vector is of size k; $A = \{a_{ij}\}$: The transition probability matrix, where a_{ij} is the probability of transition from state i to state j; $b_j(o_{jk})$: The observation probability density function, estimated by $\Sigma_{k=1}^{M} c_{jk} N(o_{jk}, \mu_{jk}, \Sigma_{jk})$, where o_{jk} denotes the feature vector; $N(o_{jk}, \mu_{jk}, \Sigma_{jk})$ denotes a single Gaussian density function with mean of μ_{jk} and covariance matrix Σ_{jk} for the state j, with M the number of mixture components and c_{jk} the weight of the k^{th} mixture component; $\Pi = \pi_i$: The initial state probability distribution. For convenience, we define the parameters for our HMM by a set λ that represents A, B and Π. We can use the parameter set λ to evaluate $P(O|\lambda)$, i.e. to measure the maximum likelihood performance of the output observables O. In order to evaluate $P(O|\lambda)$, however, we first need to compute the probabilities in the matrices in the parameter set λ

The Markov assumption that state durations have a geometric distribution defined by the probability of self transitions makes it difficult to model durations in an HMM. If we introduce an explicit duration probability to replace self transition probabilities, the Markov assumption no longer holds. Yet, HMMs have been extended by defining state duration distributions called Hidden Semi-Markov Model (HSMM) that has been succesfully used (Tweed et. al., 2005). Similar to (Tweed et. al., 2005)'s use of HSMMs, we want to model the position of a sentence in the source document explicitly. But instead of building an HSMM, we model this positional information by building our position-sensitive HMM in the following way:

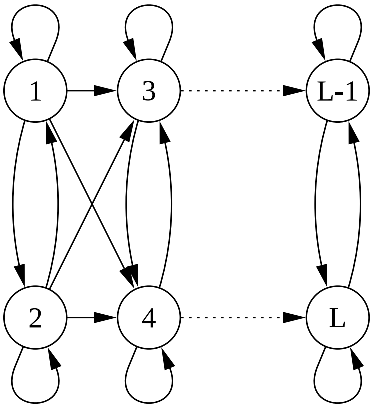

Figure 1: L state position-sensitive HMM

We first discretize the position feature into L number of bins, where the number of sentences in each bin is proportional to the length of the document. We build 2 states for each bin where the second state models the probability of the sentence being included in the document's summary and the other models the exclusion probability. Hence, for L bins we have $2L$ states. For any bin lth where $2l$ and $2l-1$ are the corresponding states, we remove all transitions from these states to other states except $2(l+1)$ and $2(l+1)-1$. This converts our ergodic L state HMM to an almost Left-to-Right HMM though l states can go back to $l-1$. This models sentence position in that decisions at the lth state can be arrived at only after decisions at the $(l-1)th$ state have been made. For example, if we discretize sentence position in document into 10 bins, such that 10% of sentences in the document fall into each bin, then states 13 and 14, corresponding to the seventh bin (.i.e. all positions between 0.6 to 0.7 of the text) can be reached only from states 11, 12, 13 and 14.

The topology of our HMM is shown in Figure 1.

3.1 Features and Training

We trained and tested our model on a portion of the TDT-2 corpus previously used in (Maskey and Hirschberg, 2005). This subset includes 216 stories from 20 CNN shows, comprising 10 hours of audio data and corresponding manual transcript. An annotator generated a summary for each story by extracting sentences. While we thus rely upon humanidentified sentence boundaries, automatic sentence detection procedures have been found to perform with reasonable accuracy compared to human performance (Shriberg et. al., 2000).

For these experiments, we extracted only acoustic/prosodic features from the corpus. The intuition behind using acoustic/prosodic features for speech summarization is based on research in speech prosody (Hirschberg, 2002) that humans use acoustic/prosodic variation — expanded pitch range, greater intensity, and timing variation — to indicate the importance of particular segments of their speech. In BN, we note that a change in pitch, amplitude or speaking rate may signal differences in the relative importance of the speech segments produced by anchors and reporters — the professional speakers in our corpus. There is also considerable evidence that topic shift is marked by changes in pitch, intensity, speaking rate and duration of pause (Shriberg et. al., 2000), and new topics or stories in BN are often introduced with content-laden sentences which, in turn, often are included in story summaries.

Our acoustic feature-set consists of 12 features, similar to those used in (Inoue et. al., 2004; Christensen et. al., 2004; Maskey and Hirschberg, 2005). It includes **speaking rate** (the ratio of voiced/total frames); **F0 minimum**, **maximum**, and **mean**; **F0 range** and **slope**; **minimum, maximum**, and **mean RMS energy** (minDB, maxDB, meanDB); **RMS slope** (slopeDB); **sentence duration** (timeLen = endtime - starttime). We extract these features by automatically aligning the annotated manual transcripts with the audio source. We then employ Praat (Boersma, 2001) to extract the features from the audio and produce normalized and raw versions of each. Normalized features were produced by dividing each feature by the average of the feature values for each speaker, where speaker identify was determined from the Dragon speaker segmentation of the TDT-2 corpus. In general, the normalized acoustic features performed better than the raw values.

We used 197 stories from this labeled corpus to train our HMM. We computed the transition probabilities for the matrix A_{NXN} by computing the relative frequency of the transitions made from each state to the other valid states. We had to compute four transition probabilities for each state, i.e. a_{ij} where $j = i, i+1, i+2, i+3$ if i is odd and $j = i-1, i, i+1, i+2$ if i is even. Odd states signify that the sentence should not be included in the summary, while even states signify sentence inclusion. Observation probabilities were estimated using a mixture of Gaussians where the number of mixtures was 12. We computed a $12X1$ matrix for the mean μ and $12X12$ matrices for the covariance matrix Σ for each state. We then computed the maximum likelihood estimates and found the optimal sequence of states to predict the selection of document summaries using the Viterbi algorithm. This approach maximizes the probability of inclusion of sentences at each stage incrementally.

4 Results and Evaluation

We tested our resulting model on a held-out test set of 19 stories. For each sentence in the test set we extracted the 12 acoustic/prosodic features. We built a $12XN$ matrix using these features for N sentences in the story where N was the total length of the story. We then computed the optimal sequence of sentences to include in the summary by decoding our sentence state lattice using the Viterbi algorithm. For all the even states in this sequence we extracted the corresponding segments and concatenated them to produce the summary.

Evaluating summarizers is a difficult problem, since there is great disagreement between humans over what should be included in a summary. Speech summaries are even harder to evaluate because most objective evaluation metrics are based on word overlap. The metric we will use here is the standard information retrieval measure of Precision, Recall and F-measure on sentences. This is a strict metric, since it requires exact matching with sentences in the human summary; we are penalized if we identify sentences similar in meaning but not identical to the gold standard.

We first computed the F-measure of a baseline system which randomly extracts sentences for the summary; this method produces an F-measure of 0.24. To determine whether the positional information captured in our position-sensitive HMM model was useful, we first built a 2-state HMM that models only inclusion/exclusion of sentences from a summary, without modeling sentence position in the document. We trained this HMM on the training corpus described above. We then trained a position-sensitive HMM by first discretizing position into 4 bins, such that each bin includes one-quarter of the sentences in the story. We built an 8-state HMM that captures this positional information. We tested both on our held-out test set. Results are shown in Table 1. Note that recall for the 8-state position-sensitive HMM is 16% better than recall for the 2-state HMM, although precision for the 2-state model is slightly (1%) better than for the 8-state model. The F-measure for the 8-state position-sensitive model represents a slight improvement over the 2-state model, of 1%. These results are encouraging, since, in skewed datasets like documents with their summaries, only a few sentences from a document are usually included in the summary; thus, recall is generally more important than precision in extractive summarization. And, compared to the baseline, the position-sensitive 8-state HMM obtains an F-measure of 0.41, which is 17% higher than the baseline.

ModelType	Precision	Recall	F-Meas
HMM-8state	0.26	0.95	0.41
HMM-2state	0.27	0.79	0.40
Baseline	0.23	0.24	0.24

Table 1: Speech Summarization Results

5 Conclusion

We have shown a novel way of using continuous HMMs for summarizing speech documents without using any lexical information. Our model generates an optimal summary by decoding the state lattice, where states represent whether a sentence should be included in the summary or not. This model is able to take the context and the previous decisions into account generating better summaries. Our results also show that speech can be summarized fairly well using acoustic/prosodic features alone, without lexical features, suggesting that the effect of ASR transcription errors on summarization may be minimized by techniques such as ours.

6 Acknowledgement

We would like to thank Yang Liu, Michel Galley and Fadi Biadsy for helpful comments. This work was funded in part by the DARPA GALE program under a subcontract to SRI International.

References

Boersma P. *Praat, a system for doing phonetics by computer* Glot International 5:9/10, 341-345. 2001.

Christensen H., Kolluru B., Gotoh Y., Renals S. *From text summarisation to style-specific summarisation for broadcast news* Proc. ECIR-2004, 2004

Conroy J. and Leary D.O *Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition* Technical report, University of Maryland, March 2001

Hirschberg J *Communication and Prosody: Functional Aspects of Prosody* Speech Communication, Vol 36, pp 31-43, 2002.

Hori C., Furui S., Malkin R., Yu H., Waibel A.. *Automatic Speech Summarization Applied to English Broadcast News Speech* Proc. of ICASSP 2002, pp. 9-12 .

Inoue A., Mikami T., Yamashita Y. *Improvement of Speech Summarization Using Prosodic Information* Proc. of Speech Prosody 2004, Japan

Kupiec J., Pedersen J.O., Chen F. *A Trainable Document Summarizer* Proc. of SIGIR 1995

Language Data Consortium *"TDT-2 Corpus* Univ. of Pennsylvania.

Maskey S. and Hirschberg J. 2005. *Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features* Proc. of ICSLP, Lisbon, Portugal.

Ohtake K., Yamamoto K., Toma y., Sado S., Masuyama S. *Newscast Speech Summarization via Sentence Shortening Based on Prosodic Features* Proc. of SSPR pp.167-170. 2003

Shriberg E., Stolcke A., Hakkani-Tur D., Tur G. *Prosody Based Automatic Segmentation of Speech into Sentences and Topics"* Speech Communication 32(1-2) September 2000

Tweed D., Fisher R., Bins J., List T, *Efficient Hidden Semi-Markov Model Inference for Structured Video Sequences* Proc. of (VS-PETS), pp 247-254, Beijing, Oct 2005.

Witbrock M.J. and Mittal V.O. *Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries* Proc. of SIGIR 1999

Zechner K. *Automatic Generation of Concise Summaries of Spoken Dialogues in Unrestricted Domains* Research and Development in Information Retrieval, 199-207, 2001.