Latent Topic Modeling for Audio Corpus Summarization

Timothy J. Hazen

MIT Lincoln Laboratory, Lexington, Massachusetts, USA

Abstract

This work presents techniques for automatically summarizing the topical content of an audio corpus. Probabilistic latent semantic analysis (PLSA) is used to learn a set of latent topics in an unsupervised fashion. These latent topics are ranked by their relative importance in the corpus and a summary of each topic is generated from signature words that aptly describe the content of that topic. This paper presents techniques for producing a high quality summarization. An example summarization of conversational data from the Fisher corpus that demonstrates the effectiveness of our approach is presented and evaluated.
Index Terms: latent topic modeling, speech summarization

1. Introduction

Recently, there has been increased interest in the area of summarization of speech documents. Most research has focused on single document summarization where a system generates a short summary, typically a few sentences/utterances in length, to succinctly describe the content of a single document [11, 13]. A related area is multi-document summarization where multiple documents discussing the same topic are processed to generate one joint summarization [3, 8]. Our work focuses on a slightly different problem: summarizing the contents of an entire collection of audio documents spanning many topics. There are various applications where such a summary could be useful. For example, the collection of all news broadcasts aired during some period of time could be analyzed to produce a summary of the most dominant topics present at that time.

Research in this area has often focused on the clustering of documents into trees or disjoint groupings [12]. However, these clustering techniques make hard assignments of documents to clusters based on similarity measures and are only appropriate when the documents are largely homogeneous in their topical content (e.g., new stories, scientific articles, etc.). This approach may not be suited for data sets containing documents that exhibit topical variation over their duration. Automatic topic segmentation can be applied before clustering [1], but this generally only works accurately on data containing long, topically homogeneous segments such as news broadcasts. However, not all data is as well-structured as news broadcasts.

Recent efforts to characterize document collections have shifted towards probabilistic latent topic model approaches which allow individual documents to be modeled as a probabilistic mixture of the latent topics learned in an unsupervised fashion [2, 9]. Statistics extracted from latent topic models are used to ascertain the dominant topical themes in the corpus, and signature words from the topics are used to summarize these topical themes. This approach has been widely used in the text processing field in recent years, but has typically been applied to collections of carefully-prepared topically-homogeneous documents such as scientific articles or news articles [7].

In this paper we focus on the summarization of a collection of conversational speech. Spoken conversations are largely produced in a spontaneous and unplanned manner. As such their topical content can be less focused than prepared audio such as news broadcasts, and diversions from the primary topic of discussion can be relatively common. Experiments in this paper specifically use conversations extracted from the Fisher corpus [4]. Within this problem space, this paper examines several technical challenges related to the task including determining the appropriate number of latent topic models to train, ranking the relative importance of the individual learned topics, and extracting appropriate signature keywords for summarizing the individual topics.

2. Experimental Conditions

2.1. Corpus

Our experiments use a collection of 1374 calls extracted from the English Phase 1 portion of the Fisher Corpus [4]. This corpus consists of audio from 10-minute-long telephone conversations between two people. Before the start of each conversation, the participants were prompted to discuss a specific topic. Data was collected from a set of 40 different prompted topics including relatively distinct topics (e.g. “Pets”, “Movies”, “Hobbies”, etc.) as well as topics covering similar subject areas (e.g. “Family”, “Life Partners”, “Family Values”). The topical content in the data is generally dominated by discussion of these 40 prompted topics, but it is not unusual for conversations to stray off-topic. Various topics that fall outside of the domain of prompted topics, such as personal information proffered by the participants or discussion of the data collection process, can be routinely observed in the data.

2.2. ASR

In our experiments, word-based automatic speech recognition (ASR) is applied to each conversation using the MIT SUMMIT system [5]. Specific details for this recognizer can be found in [6]. The ASR system generates a word hypothesis lattice for every audio segment. From the posterior word probabilities contained in the lattices an estimated occurrence count for each word in the vocabulary is computed for each conversation.

2.3. PLSA

Probabilistic latent semantic analysis (PLSA) is used in our experiments to learn a set of latent topics within a document collection in an unsupervised fashion [9]. We have used PLSA primarily for computational reasons, though all of our techniques could also be applied to other probabilistic latent topic modeling approaches such as latent Dirichlet allocation (LDA) [2].

This work was sponsored by the Air Force Research Laboratory under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

In PLSA, a probabilistic framework is used to learn the relationship between a latent topic space and the observed feature space. The basic form of PLSA seeks to learn a model which optimally represents a collection of documents $D = \{d_1, \ldots, d_{N_D}\}$, each of which contains a collection of observed words $d_i = \{w_1, \ldots, w_{N_{d_i}}\}$. PLSA learns a probability model for observing a word w within a document d via a latent topic space Z, as expressed as:

$$P(w|d) = \sum_{\forall z \in Z} P(w|z)P(z|d) \quad (1)$$

Here, the latent variable z lies in a latent topic space Z with N_Z hidden topics. Using PLSA, each document d is represented as a mixture of latent topics, $P(z|d)$, and each latent topic possesses its own generative model for producing word features, $P(w|z)$. In our work, a document is not represented by its exact sequence of words (which is unknown), but instead by the estimated counts of the words produced by the ASR system.

The PLSA model is trained over the collection D using an iterative EM algorithm which learns a model which (locally) maximizes the likelihood of the data collection. The total log likelihood of the model over the data can be expressed as:

$$\mathcal{L}(D) = \sum_{\forall d \in D} \sum_{\forall w \in V} c_{w,d} \log P(w|d) \quad (2)$$

Here, V is the full vocabulary of the ASR system, $c_{w,d}$ is the estimated count of word w appearing in document d, and $P(w|d)$ is determined from Equation 1.

To initialize the PLSA models prior to EM training we perform agglomerative clustering of the audio documents until the data is collapsed into N_Z clusters. This initial clustering uses a standard cosine similarity measure between vectors of word counts to compare documents. The documents in these N_Z clusters are then used to seed the N_Z different $P(w|z)$ models. The $P(z|d)$ models are all initialized to $P(z)$ as determined from the initial agglomerative clustering.

2.4. Stop Listing

To reduce the effect of noise contributed by non-content-bearing words (e.g., articles, conjunctions, auxiliary verbs, etc.), a *stop-list* of words to be ignored by the PLSA model is often used. Although stop-lists are often manually crafted, a simple yet effective method for creating a stop list is to ignore words with a high document frequency, as defined by:

$$\text{df}(w) = \frac{N_{D \cap w}}{N_D} \quad (3)$$

Here, $N_{D \cap w}$ is the number of documents containing the word w. This value is not known for audio data, but can be estimated by the expression:

$$N_{D \cap w} = \sum_{\forall d \in D} \min(c_{w,d}, 1) \quad (4)$$

In this work, we assign any word with $\text{df}(w) > .25$ to the stop list. We also assign to the stop list any infrequently occurring word, i.e. words estimated to appear less than 3 times in the corpus. After stop-listing with these constraints the size of the vocabulary used by the PLSA model in the experiments on our 1374 document data set is 9757 unique words.

2.5. Determining the Number of Topics

To determine the appropriate number of latent topics to use, we have implemented a method which examines the total likelihood $\mathcal{L}(D)$ as the number of latent topics N_Z is varied over an appropriately large enough range and attempts to find the *knee-in-the-curve*, e.g. the point where increasing the number of latent topics begins to show an obvious retardation in likelihood gains as N_Z is increased. We assume the knee-in-the-curve represents the point where all primary topics are being adequately modeled. Beyond this point we presume likelihood improvements in $\mathcal{L}(D)$ are being achieved from modeling topically related sub-topics, superficial or spurious similarities between smaller sets of documents, or even individual documents. To find the knee-in-the-curve we used a linear spline modeling approach detailed in [14]. PLSA models were generated for latent topic sizes ranging from 5 to 125 by increments of 5, and the knee-in-the-curve method selected the appropriate number of latent topics to be 35.

3. Summarization From PLSA Models

Our system's primary goal is to examine a collection of audio documents and provide a summarization of its topical content. From a learned PLSA model, our system ranks the relative importance of the latent topics and provides a concise summary of each topic that can be easily interpreted by a human user.

3.1. Ranking the Latent Topics

To rank the relative importance of topics there are two primary considerations. First, topics that occur more frequently are likely to be more important. The distribution of the topics across the document collection is easily represented as:

$$P(z) = \frac{1}{N_W} \sum_{\forall d \in D} N_{W|d} P(z|d) \quad (5)$$

Here, N_W is the total number of words in the entire collection D and $N_{W|d}$ is the total number of words in document d.

Second, if we assume that most conversations are dominated by a few or only one actual topic, we would expect that learned topics that dominate the documents in which they appear are more semantically important than learned topics that represent only a small portion of many documents. For example, common conversational elements such as greetings and farewells may occur frequently across many conversations in a corpus, but these small portions of the conversations carry little or no topical information. To measure this notion, we introduce the *Z→D purity measure*, which is expressed as:

$$\mathcal{P}_{Z \to D}(z) = \exp\left(\frac{\sum_{\forall d} P(z|d) \log P(z|d)}{\sum_{\forall d} P(z|d)}\right) \quad (6)$$

For any topic $z \in Z$, this measure will have a maximum purity value of 1 when $P(z|d) = 1$ for some subset of documents $d \in D$ and $P(z|d) = 0$ for the remaining documents. Small purity values for z indicate the topic is weakly spread across many documents.

Assuming that the most important learned topics are those that are strongly present in the document collection while also possessing a large *Z→D* purity measure, we create a latent topic quality score which is expressed as:

$$\mathcal{Q}(z) = 100 * P(z) * \mathcal{P}_{Z \to D}(z) \quad (7)$$

We use this metric to rank order the list of learned latent topics.

3.2. Summarizing the Latent Topics

For summarization, our system needs to provide a short but informative description of the topics present in the corpus. A common approach is to produce a short list of *signature words* for each topic. A common signature word selection method is to rank order the words for each topic based on their likelihood in the latent topic unigram model $P(w|z)$ [7]. This approach is informative about the most common words observed in each z, but this listing can include words that are also commonly found in one or more other topics and are not distinctive to that topic. Another approach is to rank the words in each z by the *a posteriori* probability $P(z|w)$, thus ensuring that the list is dominated by words that are highly distinctive of the topic. However, these topically distinctive words may not be commonly used throughout all of the documents discussing this topic and may thus be less descriptive of the topic as a whole.

A useful compromise between commonality and distinctiveness is to rank the words $w \in W$ for each particular topic z using a weighted point-wise mutual information scoring metric:

$$I(w,z) = P(w,z) \log \frac{P(w,z)}{P(w)P(z)} \tag{8}$$

This function represents the contribution of the specific elements w and z to the full mutual information measure $I(W;Z)$. This function can equivalently be written as:

$$I(w,z) = P(w|z)P(z) \log \frac{P(z|w)}{P(z)} \tag{9}$$

Here it can be seen that this measure combines the commonality property of $P(w|z)$ with the distinctiveness property of $P(z|w)$ when ranking each word $w \in V$ for any fixed value of z.

To avoid redundancy in the summaries, our system also applies word stemming when selecting signature words and omits any word that shares the same root word with a word presented higher in a topic's word list.

3.3. Summarization of the Fisher Corpus

Table 1 shows an abridged version of the summary produced by our system for the Fisher Corpus data set. Several key observations that can be made about this summary. First, the system is doing an excellent job of discovering and summarizing the actual prompted topics in the Fisher Corpus. This is evident by examining the $P(t|z)$ value for the best matching Fisher topic t associated with each learned topic z. The majority of the Fisher topics can be manually matched one-to-one to automatically learned latent topics. In a handful of cases, two similar Fisher topics were merged into a single latent topic. For example, latent topic 3 in Table 1, though dominated by the "Life Partners" topic, also subsumes the rarer "Family Values" topic.

Only two of the 35 latent topics could not be manually matched to actual Fisher topics. One of these topics (topic 35 in Table 1) was ranked last (i.e., least important) by our latent topic ranking mechanism and accounted for only .65% of the corpus. In the other case, the PLSA models identified a hidden topic which we call the "Mystery Shopping" topic (topic 29 in Table 1). Examination of the data associated with this topic reveals that many of the participants in the Fisher Corpus learned about the corpus collection effort from advertisements placed on mystery shopping websites, and off-topic discussions about mystery shopping often ensued when this information was proffered by one of the participants.

The effects of the topic ranking mechanism are also observable in the table. The ranked list is clearly correlated with the prevalence of the latent topics in the corpus, but the use of the $Z \rightarrow D$ purity measure also plays a role. For topics representing the same portion of the corpus, a higher $Z \rightarrow D$ purity measure indicates that the latent topic played a more dominant role in a smaller number of conversations, while a lower purity score indicates the topic played a weaker role across more conversations. An example where a higher purity measure plays a role in the topic rankings is evident when comparing topic 2 and topic 3 in the ranked list. Latent topic 2 (corresponding to the "Minimum Wage" Fisher topic) frequently dominated the conversations it appeared making it a purer topic than latent topic 3 (corresponding to the "Life Partners" Fisher topic). This is largely because discussion related to spouses or partners contributed not only to the conversations about "Life Partners", but also contributed in smaller amounts to many other Fisher topics such as "Family" and "Family Values"

Another positive aspect of the system's output is that the signature words are very clearly strong indicators of the underlying topic. However, there is room to improve these summaries even further. For example, the discovery of predictive n-gram units can be used to present the user with salient multi-word sequences such as *minimum wage*, *september eleventh*, or *drug testing*. We leave this improvement for future work.

3.4. Evaluation Metrics

While the quality of the Fisher Corpus summarization can be anecdotally confirmed through visual inspection, quantitative evaluation metrics are also available for comparative purposes. In our experiments, we have evaluated both the quality of the latent topic model and the quality its corresponding summarization as the number of latent topics is varied.

To assess the similarity between our PLSA model and the reference topic labels we use a measure we refer to as the *erroneous information ratio* (EIR) [10], which is defined as:

$$EIR(Z,T) = \frac{H(Z|T) + H(T|Z)}{H(T)} \tag{10}$$

Here, $T = \{t_1, \ldots, t_{N_T}\}$ is the set of N_T reference topic labels associated with the document collection. The entropy measures $H(T)$, $H(Z|T)$ and $H(T|Z)$ can be computed in the standard fashion from the joint distribution $P(z,t|d)$ estimated over all documents $d \in D$. This ratio compares the sum of the erroneous information captured by $H(Z|T)$ and $H(T|Z)$ with the total information $H(T)$ in the labeled reference data. Values closer to 0 represent greater similarity between the PLSA model and the labeled reference data.

To evaluate the summarizations, the signature word lists automatically generated from the latent topics can be compared against the reference word lists generated from the reference distributions $P(w|t)$ and $P(t|w)$. When comparing the summary word list with the reference word list we can compute a summary error ratio (SER) as:

$$SER(Z,T) = (F + M)/R \tag{11}$$

Here, F is the number of signature words in the automatic summary that don't appear in the reference, M is the total number of signature words in the reference summary that don't appear in the automatic summary, and R is the total number of signature words contained in the reference summary. When computing this we utilize only the collection of unique word stems present in the lists.

Rank	Topic Score	$Z \rightarrow D$ Purity	% of Corpus	Highest Ranked Signature Words	Matching Fisher Topic ($P(t\|z)$)
1	2.33	0.408	5.70	dog cats pets fish animals german apartment door shepherd	Pets (.705)
2	2.31	0.540	4.27	wage minimum fifteen jobs higher fifty welfare cost california	Minimum Wage (.855)
3	2.14	0.386	5.54	important relationship partner marriage together divorced	Life Partners (.625)
4	2.10	0.397	5.28	september eleventh changes scary trade terrorist travel military	September 11^{th} (.680)
5	2.01	0.462	4.36	security airport plane check terrorists fly travel flight airplane	Airport Security (.751)
:	:	:	:	:	
29	0.69	0.404	1.70	shopping mystery surveys dot email husband com internet	None
:	:	:	:	:	
33	0.55	0.347	1.59	games computer played video internet laptop solitaire playstation	Computer Games (.601)
34	0.43	0.459	0.94	drug test company medical military certainly excellent privacy	Drug Testing (.365)
35	0.28	0.433	0.65	shh lost challenge texas salad insurance church special alabama	None

Table 1: A portion an automatically generated summary of a collection of 1374 Fisher Corpus conversations. The far right column shows the closest matching Fisher Corpus topic $t \in T$ for each automatically learned topic determined for cases where $P(t|z) > 0.25$.

Evaluation Metric	Number of Latent Topics in Z										
	20	25	30	35	40	45	50	55	60	65	70
EIR	.968	.932	.897	.867	.847	**.840**	.846	.841	.850	.857	.857
SER	.503	.419	**.320**	.345	.357	.394	.460	.519	.575	.646	.720

Table 2: Evaluation of learned PLSA topic models and their corresponding automatic summarizations over varying sizes of Z in comparison to the reference topic models and their corresponding automatically generated summarizations for the known topics in T.

3.5. Evaluation Results

Table 2 shows the evaluation of the PLSA models learned from the Fisher Corpus data using the erroneous information ratio (EIR) metric (discussed in Section 3.4) as the number of latent topics N_Z is varied from 20 to 70. The EIR metric trades-off decreases in $H(T|Z)$ with increases in $H(Z|T)$ as N_Z increases. The EIR metric achieves its minimum value at $N_Z = 45$ before slowly increasing for $N_Z > 45$. Because the number of labeled topics in the Fisher Corpus is 40, the EIR scores align well with our expectations about the overlap between Z and T.

Table 2 also shows the results comparing the signature word summaries generated from the PLSA model against the reference summaries generated from the known topic models using the summary error ratio (SER) metric described in Section 3.4. This metric trades off signature word false alarms F against signature word misses M as N_Z increases. For our selected value of $N_Z = 35$, the SER value of .345 is only slightly worse than the optimal value of .320 at $N_Z = 30$. At $N_Z = 35$, 85% of the automatic summary words appear in the reference, and 80% of the reference words appear in the automatic summary. This demonstrates that our approach is accurately discovering the most topically relevant words in the data collection.

4. Summary

In this paper, we have presented an approach for automatically summarizing the topical contents of an audio corpus of conversational speech. Standard document clustering techniques are not appropriate for this task because conversations are spontaneous and unplanned with off-topic diversions a common occurrence. Instead probabilistic latent semantic analysis (PLSA) was used to learn a set of latent topics in an unsupervised fashion. Techniques were presented for ranking learned latent topics by their relative importance in the corpus and selecting appropriate signature words for succinctly summarizing the content of each topic. An example summarization demonstrating the effectiveness of this technique was generated using the output of an ASR system applied to data from the Fisher corpus of conversational speech.

5. References

[1] J. Allan, editor, *Topic detection and tracking: Event-based information organization*, Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[2] D. Blei, A. Ng and M. Jordan "Latent Dirichlet allocation," *Journal of Machine Learning Research* vol. 3, pp. 993–1022, 2003.

[3] A. Celikyilmaz and D. Hakkani-Tür, "Extractive summarization using a latent variable model," in *Proc. Interspeech*, Makuhari, 2010.

[4] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generation of speech-to-text," in *Proc. Int. Conf. on Lang. Resources and Eval.*, Lisbon, 2004.

[5] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, no. 2-3, pp. 137-152, 2003.

[6] T. Hazen, F. Richardson and A. Margolis, "Topic identification from audio recordings using word and phone recognition lattices", in *Proc. ASRU*, Kyoto, 2007.

[7] T. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. of National Academy of Sciences*, vol. 101, pp. 5228-5235, 2004.

[8] S. Harabagiu and F. Lacatusu, "Topic themes for multi-document summarization,", in *Proc. of SIGIR*, Salvador, Brazil, 2005.

[9] T. Hofmann, "Probabilistic latent semantic analysis,", in *Proc. of Conf. on Uncertainty in Artificial Intelligence*, Stockholm, 1999.

[10] R. Holt, *et al*, "Information theoretic approach for performance evaluation of multi-class assignment systems," Proc. of SPIE, vol. 7697, April 2010.

[11] C. Hori, *et al*, "Automatic speech summarization applied to English broadcast news," in *Proc. ICASSP*, Orlando, 2002.

[12] K. Kummamuru, *et al*, "A hierarchical monothetic document clustering algorithm for summarization and browsing search results," in *Proc. Int. Conf on World Wide Web*, New York, 2004.

[13] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *Proc. Interspeech*, Lisbon, 2005.

[14] S. Salvador and P. Chan, "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms," in *Proc. Int. Conf on Tools with Artificial Intel.*, Boca Raton, 2004.