DIRECT AND LATENT MODELING TECHNIQUES FOR COMPUTING SPOKEN DOCUMENT SIMILARITY[1]

Timothy J. Hazen

MIT Lincoln Laboratory
Lexington, Massachusetts, USA

ABSTRACT

Document similarity measures are required for a variety of data organization and retrieval tasks including document clustering, document link detection, and query-by-example document retrieval. In this paper we examine existing and novel document similarity measures for use with spoken document collections processed with automatic speech recognition (ASR) technology. We compare direct vector space approaches using the cosine similarity measure applied to feature vectors constructed with various forms of term frequency inverse document frequency (TF-IDF) normalization against latent topic modeling approaches based on latent Dirichlet allocation (LDA). In document link detection experiments on the Fisher Corpus, we find that an approach that applies bagging to models derived from LDA substantially outperforms the direct vector space approach.

Index Terms— document similarity, document link detection, latent topic modeling

1. INTRODUCTION

There are a variety of useful data organization and retrieval techniques that can be applied to collections of audio documents. These include document clustering (where groups of topically related documents are clustered together), document link detection (where document pairs which are topically related or *linked* are identified), and query-by-example document retrieval (where documents which are topically related to an example or *query* document are listed and ranked). All of these applications require an accurate method for computing the similarity between pairs of documents.

In order to compare two audio documents, the first step is to construct a vector of content bearing features extracted from each document. Generally, the feature vector contains simple counts of observed words or short word n-grams. In fact, the basic *bag-of-words* approach using only counts of individual word unigrams (devoid of any local contextual information) has proven surprisingly effective for a variety of document modeling tasks [9, 11]. After observed features have been extracted, there are two primary methods in which documents are commonly compared during the similarity computation process: (1) with direct comparison of the observed features extracted the documents, or (2) via comparison of latent topic variables inferred from the observed features.

In the direct modeling approach, the bag-of-words feature vectors typically lie in a high-dimensional but generally sparse space containing one dimension for every word in the known vocabulary. Vectors are typically normalized in some fashion (e.g. using stoplisting and/or inverse document frequency weighting) to reduce the impact of common words that possess little or no content bearing information. The feature vectors for two documents are often compared in the vector space using the cosine similarity measure. This widely-used method was the primary approach employed by systems performing the link detection task in NIST's Topic Detection and Tracking (TDT) Evaluations conducted from 1999 to 2004 [1, 13].

In the latent topic modeling approach, a set of hidden topics are learned from the data collection using a technique such as latent semantic analysis (LSA) [5], probabilistic latent semantic analysis (PLSA) [8], or latent Dirichlet allocation (LDA) [2]. Given a learned set of latent topics, a weighting or probabilistic distribution of the latent topics is inferred from the observed features for each document in the collection. The similarity of the vectors of latent topic weights for two documents can then be compared in the latent vector space.

This paper explores the use of both direct and latent approaches to the document modeling problem. In addition to examining traditional methods for both approaches, several variations on these techniques which can further improve their performance are also presented. In particular, this paper shows that the use of bootstrap aggregation (or *bagging* [3]) of multiple randomly initialized probabilistic latent models can yield substantial improvements in the accuracy of latent topic similarity measures. Evaluations of the explored techniques are conducted for the document link detection task using a collection of spoken conversations from the Fisher Corpus that have been processed using an automatic speech recognition (ASR) system.

2. DIRECT MODELING OF DOCUMENTS

2.1. The Cosine Similarity Measure

In the direct modeling approach, a high dimension feature vector, represented as $\vec{x}_i$, is constructed to describe each document d_i in a collection D. Within this vector space, the similarity between the two documents d_i and d_j, represented by vectors, $\vec{x}_i$ and $\vec{x}_j$, is commonly computed using the cosine similarity (CS) measure as follows:

$$S_{CS}(d_i, d_j) = \cos\phi = \frac{\vec{x}_i \cdot \vec{x}_j}{\|\vec{x}_i\| \, \|\vec{x}_j\|} \quad (1)$$

Here ϕ is the angle between the vectors $\vec{x}_i$ and $\vec{x}_j$. If we assume that every element of $\vec{x}_i$ and $\vec{x}_j$ is non-negative then $0 \leq \cos\phi \leq 1$ with values closer to 1 representing document pairs with greater similarity.

[1]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

2.2. The TF-IDF Representation

Each element x_w of a feature vector $\vec{x}$ represents the contribution of a particular feature w. These features are typically unique words, though any type of countable feature is possible including word n-grams, local word co-occurrences, etc. Each word w is an element of the vocabulary V, which is comprised of N_V pre-determined words. This vector space representation of a document based on the counts of the individual words, independent of their ordering in the document, is commonly referred to as the *bag-of-words* representation. Using this representation, each element x_w is derived from the underlying observed count c_w of word w in the document as follows:

$$x_w = \lambda_w c_w \tag{2}$$

Here, λ_w is a weighting term which can be determined in a variety of ways. In general λ_w is intended to boost the contribution of word w when it contains a large amount of content bearing information or reduce its contributions when w possesses little or no content bearing information. If no weighting function were applied to the collection of counts, the computation of the cosine similarity measure would be dominated by common words (e.g., articles, conjunctions, auxiliary verbs, etc.) which are devoid of content information. It is also worth noting that the true value of c_w is generally unknown for audio documents and c_w is typically estimated from hypothesis lattices generated by an automatic speech recognition (ASR) system.

The most commonly used weighting function is inverse document frequency (IDF) weighting. Standard IDF weighting is expressed as:

$$\lambda_w = \text{idf}(w) = \log \frac{N_D}{N_{D \cap w}} \tag{3}$$

Here N_D is the total number of total documents in the collection and $N_{D \cap w}$ is the number of those documents containing the word w. Using this expression, words that appear in fewer documents are presumed to carry more content information than more common words and hence receive more weight. The use of the $\log$ function prevents extremely rare words (i.e., those appearing in only one or a few documents) from getting an excessively large weight. The application of IDF weighting to feature counts is commonly referred to as the *term frequency - inverse document frequency* (TF-IDF) representation.

When applying document modeling techniques to spoken documents, it is important to note that the actual words in each document are not known and as such $N_{D \cap w}$ is also not known and must be estimated [14]. In our work, $N_{D \cap w}$ is estimated from the document collection as:

$$N_{D \cap w} = \max \left(\kappa, \sum_{\forall d \in D} \min(c_w, 1) \right) \tag{4}$$

Here, c_w is an estimated count of word w in document d as predicted using an ASR system. It is important to note that, for any given document, the estimate of c_w can have a positive value less than one. The parameter κ sets a floor on $N_{D \cap w}$ thus providing an upper bound on $\text{idf}(w)$. In our work, we set $\kappa = 0.01$.

2.3. Stop-Listing

While TF-IDF weighting has proven effective for many tasks, it has its flaws. In particular, it still allows large numbers of common function words to receive non-zero weight despite their apparent lack of content information. Towards this end many systems employ the use of an explicit *stop list*, i.e., a list of words that are pre-selected to receive a weight of zero and hence to be ignored. While these stop lists are often manually crafted, they can also be automatically generated based on the document frequency of words. Similarly, extremely rare words, by their sporadic nature, may not contain enough trustworthy information to be useful for topical comparisons either.

To incorporate these constraints, Equation 3 can be re-expressed as follows:

$$\lambda_w = \begin{cases} 0 & \text{if } \frac{N_{D \cap w}}{N_D} > t_d \\ 0 & \text{if } N_w < t_c \\ \text{idf}(w) & \text{otherwise} \end{cases} \tag{5}$$

Here, the weight λ_w for word w is set to zero if the document frequency for w exceeds the threshold t_d or if the total count N_w of word w over all documents does not exceed the threshold t_c. This weighting scheme will be referred to as TF-IDF with *hard stop-listing*.

2.4. Soft Stop-Listing

As an alternative to hard stop-listing, we introduce a new softer de-emphasis of the very common and very rare words which we will call *soft stop-listing*. This is expressed as:

$$\lambda_w = \text{idf}(w) \left(1 - .5^{\frac{\log \frac{N_{D \cap w}}{N_D}}{\log t_d}} \right) \left(1 - .5^{\frac{N_w}{t_c}} \right) \tag{6}$$

In this expression, hard cut-offs on the document frequency and total word count are replaced with functions that vary smoothly between 0 and 1.

3. LATENT MODELING OF DOCUMENTS

3.1. Probabilistic Latent Semantic Analysis

When using latent modeling, documents in a collection are modeled as a weighted combination of latent topics from a set $Z = \{z_1, \ldots, z_{N_Z}\}$. In the probabilistic latent semantic analysis (PLSA) approach, each latent topic possesses a probabilistic unigram language model $P(w|z)$ representing the probability that word w could be generated by topic z. Each document d_i in the document collection D is then assumed to have been generated by a weighted mixture of the latent topic unigram models. If a document is modeled by a collection of word counts $C = \{c_1, \ldots, c_{N_V}\}$, the generative PLSA model for observing the word count collection C for document d_i is:

$$P(C|d_i) = \prod_{w \in V} \left(\sum_{z \in Z} P(w|z) P(z|d_i) \right)^{c_w} \tag{7}$$

3.2. Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a generalization of PLSA in which the point estimate of $P(z|d_i)$ for document d_i in PLSA is replaced by a prior probabilistic Dirichlet distribution over all possible latent topic distributions within Z. The specific details of incorporation of the Dirichlet distribution into the formulation are available in [2], so we will not cover them here. It is important, however, to note that the Dirichlet distribution in LDA constrains the topic distribution over $z \in Z$ through a shared α variable in which $\alpha \ll 1$

places a strongly preference on distributions which are dominated by one single topic, while $\alpha \gg 1$ gives preference to distributions which spread the weight evenly across all topics. In practice, the α value is typically estimated during the LDA learning process, and because documents often discuss only one specific topic (or at most a few topics), it is usually the case that $\alpha < 1$. When a fixed value of $\alpha = 1$ is used, LDA assumes all topic distributions over Z are equally likely.

When LDA uses the $\alpha = 1$ constraint in conjunction with the application of MAP estimation in its inference stage, it has been shown that LDA is equivalent to PLSA [6]. In practice though, LDA typically uses either a variational technique or a sampling technique in its inference stage, so the equivalence between LDA and PLSA when $\alpha = 1$ is only loosely approximate.

3.3. Latent Topic Representations of Documents

Using latent topic modeling (either PLSA or LDA), documents are represented by a set of estimated latent topic probabilities $P(z|d_i)$ over the set of N_Z different latent topics $z \in Z$, with the $P(z|d_i)$ values for each document being inferred using some form of the EM algorithm. These weights can equivalently be represented in the vector form $\vec{z}_i$ such that:

$$\vec{z}_i = \begin{bmatrix} P(z_1|d_i) \\ \vdots \\ P(z_{N_Z}|d_i) \end{bmatrix} \tag{8}$$

3.4. Similarity Measures for Latent Topic Representations

Given a set of estimated latent topic probabilities for each document, there are several measures that can be used to compare the similarity of two documents [12]. To begin with, the cosine similarity (CS) measure can be applied to two latent vectors $\vec{z}_i$ and $\vec{z}_j$, as was done in the direct modeling case, as follows:

$$S_{CS}(d_i, d_j) = \frac{\vec{z}_i \cdot \vec{z}_j}{\|\vec{z}_i\| \, \|\vec{z}_j\|} \tag{9}$$

An alternative to the cosine similarity measure is the unnormalized *dot product* (DP) similarity measure:

$$S_{DP}(d_i, d_i) = \vec{z}_i \cdot \vec{z}_j \tag{10}$$

The dot product of $\vec{z}_i$ and $\vec{z}_j$ is equivalent to the estimated probability that the two documents were derived from the same underlying latent topic, under the assumption that only one topic was used to derive each document.

The symmetric Kullback-Leibler (KL) divergence measure can also be used to compare latent topic distributions as follows:

$$D_{KL}(d_i, d_j) = \sum_{z \in Z} P(z|d_i) \log \frac{P(z|d_i)}{P(z|d_j)} + P(z|d_j) \log \frac{P(z|d_j)}{P(z|d_i)} \tag{11}$$

The KL divergence is 0 for documents with identical topic distributions and gets larger as the latent topic distributions get more dissimilar. The KL divergence can be converted to a similarity measure (ranging from 0 to 1) using this expression:

$$S_{KL}(d_i, d_j) = \exp(-D_{KL}(d_i, d_j)) \tag{12}$$

3.5. Bagging of Latent Topic Representations

In PLSA or LDA, the EM training algorithm requires an initial estimate of either $P(w|z)$ for each z or $P(z|d)$ for each d. The most common practice is to randomly initialize a $P(w|z)$ for each z. The EM algorithm is only guaranteed to converge to a locally optimal maximum likelihood solution and not the globally optimal solution. Thus, in practice, different random initializations of PLSA or LDA models will yield different final models.

Because different initializations lead to different final models, latent models derived from PLSA or LDA are ideally suited for the application of *bootstrap aggregation* or *bagging*, in which multiple models generated with different bootstrapped initializations are aggregated [3]. Bagging of PLSA models has previously been successfully applied to the task of image scene recognition [10].

When using the bagging technique, we begin by assuming that K different latent topic models M_1 through M_K are trained from K different random model initializations. From each model M_k, an estimate of the topic distribution $P(z|d_i, M_k)$ is produced for each document d_i. The aggregation stage is most easily performed by averaging the document similarity measures between two documents d_i and d_j produced by each of the K models. In this case, we represent a similarity score produced using model M_k generically as $S(d_i, d_j, M_k)$. In our work we examine two different averaging methods, the arithmetic mean and the geometric mean. The arithmetic mean is realized as:

$$S_{AM}(d_i, d_j) = \frac{1}{K} \sum_{k=1}^{K} S(d_i, d_j, M_k) \tag{13}$$

The geometric mean is realized as:

$$S_{GM}(d_i, d_j) = \exp\left(\frac{1}{K} \sum_{k=1}^{K} \log S(d_i, d_j, M_k) \right) \tag{14}$$

4. EXPERIMENTAL CONDITIONS

4.1. Corpus

For our experiments we have used a collection of 1374 calls extracted from the English Phase 1 portion of the Fisher Corpus [4]. The corpus consists of 10-minute long recorded conversations between two people connected over the telephone network. At the start of each conversation, the two participants were given prompted instructions to discuss a specific topic. Data was collected from a set of 40 different topics. The topics were varied and included relatively distinct topics (e.g. "Movies", "Hobbies", "Education", etc.) as well as topics covering similar subject areas (e.g. "Issues in Middle East", "Arms Inspections in Iraq", "Foreign Relations").

4.2. Link Detection Evaluation

To assess the various approaches to measuring document similarity proposed in this paper, we use a document link detection evaluation. In this evaluation, we attempt to detect whether a given pair of documents are topically related or not. For the Fisher Corpus, two documents are considered topically related when the topic the participants in the conversation were prompted to discuss is the same for both documents. Over our 1374 conversation experimental data set, the evaluation paradigm yields 943,251 unique document pairs of which 30,921 pairs are deemed topically linked.

To evaluate link detection performance, we compute a ranked list of document pairs from our corpus for each document similarity

measure. From the ranked list, we generate a detection/error tradeoff (DET) curve which measures the miss rate (i.e., the fraction of topically linked document pairs we fail to detect) against the false alarm rate (i.e., the fraction of document pairs that are not topically related that we falsely detect) as the detection threshold is swept through all valid values. We report performance using the equal error rate (EER) of the DET curve (i.e., the point on the curve where the miss rate and the false alarm rate are equal). We choose to use this evaluation paradigm because it provides a mechanism for assessing document similarity measures which is straightforward to implement and easy to interpret.

4.3. Automatic Speech Recognition

In our experiments, word-based automatic speech recognition (ASR) is applied to each audio segment of each conversation. The ASR system generates a network, or *lattice*, of speech recognition hypotheses for each audio segment. Within each lattice, the posterior probability of all hypothesized word arcs in the lattice is estimated. From these lattices an *expected count* for each word within each conversation is computed by summing the posterior scores over all hypothesized instances of that word over all lattices from that conversation.

We use the MIT SUMMIT speech recognizer as our ASR system [7]. The system's acoustic models were trained using a standard maximum-likelihood approach on a separate 553 hour set of Fisher Corpus data. For language modeling, the system uses a basic trigram language model with a 31.5K word vocabulary trained using the transcripts of the recognizer's training set. Because this recognizer applies very basic modeling techniques with no adaptation, the system performs recognition faster than real time (on a current workstation) but word error rates can be high (typically over 40%).

4.4. Latent Dirichlet Allocation Implementation

For our latent modeling experiments we use David Blei's C implementation of latent Dirichlet allocation (LDA).[2] This implementation was modified slightly to accommodate the use of floating point estimated counts of words instead of the integer counts typically used in text processing. We have explored using both the standard LDA approach where the α Dirichlet parameter is estimated, and also the constrained form of LDA where $\alpha = 1$ and remains fixed (which can be viewed as a loose approximation of PLSA). For all of our experiments the latent topic unigram models are initialized by seeding each individual unigram model with the statistics from a randomly selected document from the data collection.

Because common function words are also known to impede the performance of LDA, automatic stop word selection is employed using the minimum word count threshold t_c and maximum document frequency threshold t_d parameters as previously applied in the direct modeling approach. In all of the LDA experiments these parameters are set to $t_c = 5$ and $t_d = .5$. These settings result in the exclusion of 268 words that are estimated to appear in more than half of the documents and over 22K words that are estimated to occur 5 times or less over all documents. This leaves the LDA model with an active vocabulary of 7,615 words in these experiments.

The final collection of latent topic posteriors $P(z|d_i)$ for each d_i in the corpus are extracted from the inferred latent Dirichlet γ parameters produced during the variational EM process in Blei's LDA implementation. A MAP estimate of the topic distribution $P(z|d_i)$ is derived by performing an L1 norm over the collection of N_Z different γ parameters produced by the LDA process for document d_i.

[2]David Blei's open source LDA implementation is available from: `http://www.cs.princeton.edu/~blei/lda-c/`

Term Weighting	t_c	t_d	EER (%)
Standard TF-IDF	N/A	N/A	12.8
TF-IDF w/ hard stop-listing	150	0.5	11.5
TF-IDF w/ soft stop-listing	300	0.05	10.4

Table 1. Document link detection EER results for the direct modeling approach using the cosine similarity measure with three different term weighting schemes.

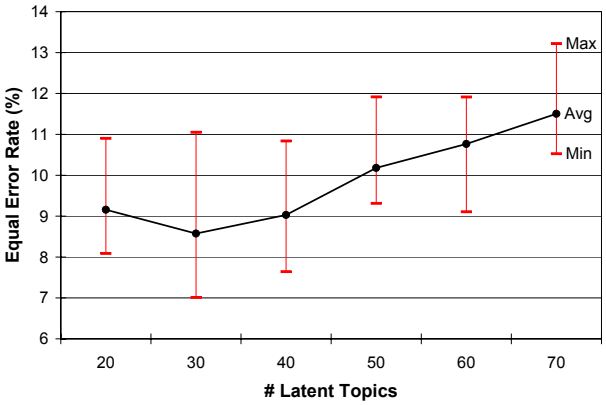

Fig. 1. Link detection EER using individual randomly initialized LDA training trials with $\alpha = 1$ over latent topic sets varying from 20 topics to 70 topics. The comparison of latent vectors of document pairs was performed with the dot product similarity measure.

5. EXPERIMENTAL RESULTS

5.1. Direct Modeling Experiments

In our first experiment, we examine the link detection EER performance of the three direct modeling approaches proposed in Section 2, i.e., the use of the cosine similarity measure in conjunction with (1) standard TF-IDF weighting, (2) TF-IDF weighting with hard stop-listing, and (3) TF-IDF weighting with soft stop-listing. Results for these three approaches are provided in Table 1. For the stop-listing approaches, results are shown for the optimal word count threshold t_c and document frequency threshold t_d as determined by a grid search over the range of appropriate values for these terms. Thus, these results provide a lower bound on the EER achievable by these techniques on this data set. The best result of 10.4% EER was acheived using TF-IDF with soft stop-listing.

5.2. Baseline Latent Modeling Experiments

Our initial set of latent modeling experiments are shown in Figure 1. In this experiment, LDA is applied to the data set using a fixed Dirichlet parameter of $\alpha = 1$. The number of latent topics is varied by 10 from 20 to 70. For each specific number of latent topics,

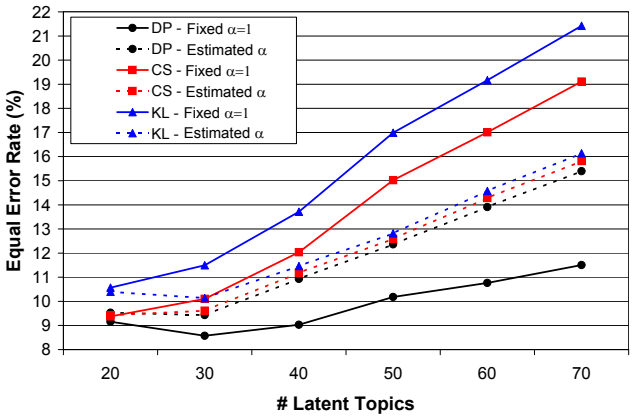

Fig. 2. Average link detection EER for LDA training trials over latent topic sets varying from 20 topics to 70 topics. Results are compared using a fixed Dirichlet parameter value of $\alpha = 1$ versus an estimated value of α when using three different document similarity measures: the dot product (DP) measure, the cosine similarity (CS) measure, and the KL divergence measure.

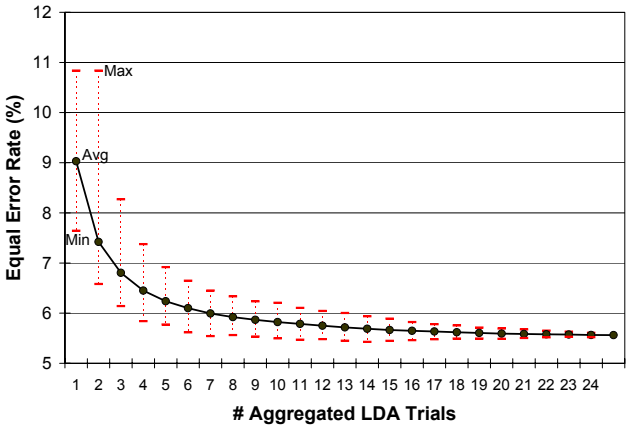

Fig. 3. Minimum, average and maximum EER performance of geometric mean bagging as the number of aggregated LDA training trials is varied from 1 to 25 when the number of latent topics is fixed at 40.

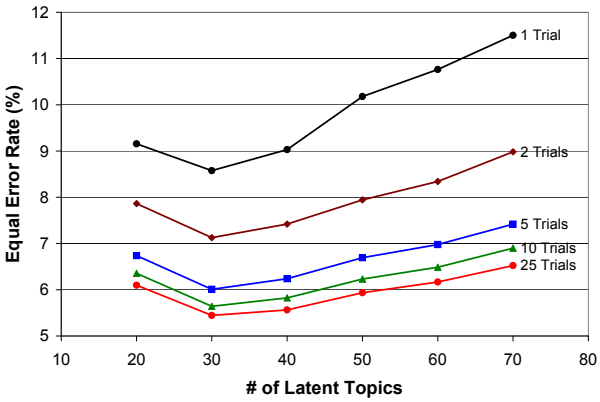

Fig. 4. Average EER performance of bagging with a geometric mean as the number of aggregated LDA training trials is varied from 1 to 25 and the number of latent topics is varied from 20 to 70.

25 different training trials are run with a different random initialization of the models for each trial. The documents are compared using the latent vector dot product similarity measure.

In the figure, we observe average link detection EER measures ranging from 8.6% using 30 latent topic to 11.5% for 70 topics. The average performance of this baseline LDA system outperforms the best TF-IDF direct modeling approach when the number of latent topics is between 20 and 50, but the worst performing LDA training trials across all latent topic set sizes are all worse than the best TF-IDF direct model system.

Figure 2 explores the effect of varying two aspects of the modeling process. First, results are examined when the Dirichlet α parameter is fixed to a value of $\alpha = 1$ versus the standard LDA approach in which α is estimated from the data. In our experiments the estimated α values range from $\alpha \approx .05$ for 20 latent topics to $\alpha \approx .02$ for 70 latent topics. In general, these small α values force the estimated latent topic distributions to be heavily skewed towards one topic. The figure also explores the effect on performance of the three different similarity measures: the dot product (DP) measure, the cosine similarity (CS) measure, and the symmetric KL divergence measure. In Figure 2, it is observed that the three similarity measure all performed similarly when the LDA α parameter is estimated, though the DP measure yields a consistently better EER than the CS and KL measures.

The most interesting result in Figure 2 is the effect of fixing $\alpha = 1$ during LDA estimation. This forces the LDA algorithm to learn smoother distributions over the latent topics than when α is estimated. When $\alpha = 1$, obvious improvements in link detection accuracy are observed for the DP measure, while the CS and KL measures see obvious degradations in performance.

Intuitively, it makes sense that the DP measure would outperform the CS and KL measures on this data, because the Fisher Corpus explicitly contains one topic per audio document and the link detection evaluation used in these experiments assumes links exist only between documents discussing the same labeled topic. If there is a close association between the latent topics learned via LDA and the actual topics in the Fisher Corpus, then the DP measure would explicitly capture the probability of two documents discussing the same underlying topic. The CS and KL measures may be more appropriate for data in which multiple topics are present in each document, because they explicitly measure the similarity of the topic distributions and make no assumptions about the actual number of topics that may be relevant to the document.

5.3. Bagged Latent Modeling Experiments

As observed in Figure 1, the performance of an LDA estimated model on the link detection evaluation is highly dependent on the initialization of the model, leading us explore the use of bootstrap aggregation or bagging. In our experiments we generated 25 different models from 25 different randomly initialized training trials for each condition we explored. Figure 3 shows the EER performance as the number of aggregated models is varied from $K = 1$ to $K = 25$. For each K, 25 pseudo-randomly selected subsets of K

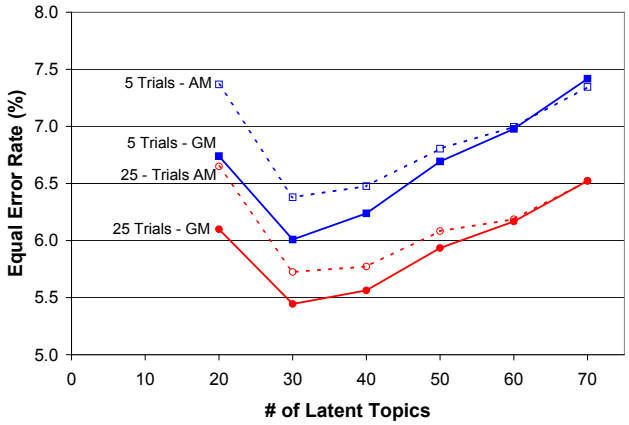

Fig. 5. Average EER performance of bagging using the arithmetic mean (AM) versus the geometric mean (GM) during averaging. Results are shown when using 5 or 25 aggregated training trials as the number of latent topics is varied from 20 to 70.

models were created and the minimum, average and maximum EER results using these 25 subsets of K models are presented. For all results in this figure, 40 latent topics are trained using a fixed $\alpha = 1$, the DP measure is used for document similarity, and the geometric mean is used for similarity score averaging. The figure shows that the average link detection EER can be reduced from 9.0% to 5.6% simply by aggregating the scores produced by 25 different randomly initialized LDA models.

Figure 4 shows results using the bagged LDA approach with geometric averaging, as the number of latent topics is varied from 20 to 70. Bagging substantially improves performance regardless of the number of latent topics used. With 30 latent topics, bagging reduces the EER by a relative 37% (from 8.6% to 5.4%). For 70 latent topics, the error rate reduction is 43% (from 11.5% to 6.5%). The results obtained from 25 bagged LDA models for the range of latent topics from 20 to 70 are all significantly better than the best direct modeling system which only achieved an EER of 10.4%. The best LDA system performance of 5.4% EER represents a relative 48% reduction in link detection EER over the best direct modeling performance of 10.4% EER.

Figure 5 compares the use of arithmetic mean averaging and geometric mean averaging during bagging. Results are shown for the aggregation of 5 training trials and 25 training trials as the number of latent topics is varied from 20 to 70. Here geometric averaging performs better than arithmetic averaging when the number of latent topics is 50 or less, but the two averaging techniques perform similarly when the number of latent topics grows larger than 50.

6. CONCLUSION

In this paper we have explored the use of both direct and latent modeling techniques for the purpose of computing similarity measures for comparing spoken audio documents. On experiments conducted using the Fisher Corpus of human-human conversations, we have shown that similarity measures based on the bagging of similarity scores derived from randomly initialized latent Dirichlet allocation (LDA) models dramatically outperformed the direct modeling approach using various measures based on TF-IDF. On a link detection task, the EER of the best LDA-based system was a relative 48% lower than the EER of the best TF-IDF system.

Though we have focused this work on the task of document link detection, the methods explored in the paper are applicable to any task that requires the use of a document similarity measure. In future work we plan to apply these techniques to the tasks of spoken audio document clustering and query-by-example document retrieval.

7. REFERENCES

[1] J. Allan, editor, *Topic Detection and Tracking: Event-based Information Organization*, Kluwer Academic Publishers, Boston, 2002.

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

[3] L. Breiman "Bagging predictors", *Machine Learning* vol. 24, no. 2, pp. 123-140, August 1996.

[4] C. Cieri, D. Miller, and K. Walker, "From Switchboard to Fisher: Telephone collection protocols, their uses and yields," in *Proc. Interspeech*, Geneva, Sep. 2003.

[5] S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman "Indexing by latent semantic analysis" *Journal of the American Society for Information Science* vol. 11, no. 6, pp. 391–407, September 1990.

[6] M. Girolami and A. Kaban, "On an equivalence between PLSI and LDA", in *Proc. of ACM SIGIR Conf. on Research and Development in Information Retrieval*, Toronto, July 2003.

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

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

[9] D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval", in *Proc. European Conf. of Machine Learning (ECML)*, Chemnitz, April 1998.

[10] E. Rodner and J. Denzler, "Randomized probabilistic latent semantic analysis for scene recognition", in *Proc. of Iberoamerican Conf. on Pattern Recognition*, Guadalajara, November 2009.

[11] F. Sebastiani, "Machine learning in automated text categorization", *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, March 2002.

[12] M. Steyvers and T. Griffiths, "Probabilistic topic models", chapter in *Handbook of Latent Semantic Analysis* T.Landauer, et al, (eds.), Psychology Press, London, 2007.

[13] C. Wayne, "Multilingual topic detection and tracking: successful research enabled by corpora and evaluation", in *Proc. of Language Resources and Evaluation Conf. (LREC)*, Athens, June 2000.

[14] J. Wintrode and S. Kulp, "Confidence-based techniques for rapid and robust topic identification of conversational telephone speech", in *Proc. Interspeech*, Brighton, England, 2009.