Topic Identification and Discovery on Text and Speech

Chandler May, Francis Ferraro, Alan McCree, Jonathan Wintrode,
Daniel Garcia-Romero, and **Benjamin Van Durme**
Human Language Technology Center of Excellence
Johns Hopkins University
cjmay@jhu.edu, ferraro@cs.jhu.edu, alan.mccree@jhu.edu, jcwintr@cs.jhu.edu,
dgromero@jhu.edu, vandurme@cs.jhu.edu

Abstract

We compare the multinomial i-vector framework from the speech community with LDA, SAGE, and LSA as feature learners for topic ID on multinomial speech and text data. We also compare the learned representations in their ability to discover topics, quantified by distributional similarity to gold-standard topics and by human interpretability. We find that topic ID and topic discovery are competing objectives. We argue that LSA and i-vectors should be more widely considered by the text processing community as pre-processing steps for downstream tasks, and also speculate about speech processing tasks that could benefit from more interpretable representations like SAGE.

1 Introduction

The text processing and speech processing research communities have similar problems and goals, but the technical approaches in these two communities develop largely independently. In this paper we compare dimensionality reduction techniques on multinomial language data from the text and speech communities. We consider a multinomial formulation of the i-vector model (hereafter "mi-vector" model) from the speech community (Soufifar et al., 2011), the sparse additive generative (SAGE) (Eisenstein et al., 2011) and latent Dirichlet allocation (LDA) (Blei et al., 2003b) topic models from the text community, and latent semantic analysis (LSA) (Deerwester et al., 1990). Both the mi-vector model and the SAGE topic model represent a multinomial parameter vector as the softmax of a sum of vectors, one of which is a background vector representing overall word usage in the corpus, and so we might expect mi-vectors and SAGE to produce similar results on real-world data. We evaluate these two recent models and two more conventional models, LDA and LSA (a term describing a class of methods based on the singular value decomposition, or SVD, which is used broadly in both research communities). We assess the similarity of mi-vectors and SAGE and expose the strengths and weaknesses of all four learned representations by evaluating them on the supervised task of topic identification (topic ID), depicted in Figure 1. We also evaluate the representations on the unsupervised, less easily-measurable task of topic discovery. As a proxy for controlled human annotations, we quantify topic discovery performance by distributional similarity to gold-standard topics.

We use the bag-of-words multinomial representation of text data, i.e., each document is represented by a vector of counts over the word vocabulary. For speech data, we use a modern automatic speech recognition (ASR) system to produce frame-wise triphone state cluster posteriors and we take the sum of these posteriors across all frames in a document to obtain a document-level vector of triphone state cluster soft counts. Modern topic ID systems for speech use ASR output instead of a lower-resource representation like these soft counts to improve performance (Hazen et al., 2007). ASR word counts are high-resource and can be viewed as a noisy version of word counts from text. We wish to assess the relative performance of our learned representations, not the quality of the data pre-processing scheme, and we desire to strengthen our results by evaluating performance on two distinct views of a corpus. Hence we break from convention and use triphone state cluster soft counts as speech data.

While previous work has juxtaposed the mi-vector model against LDA (Chen et al., 2014; Morchid et al., 2014), the current study is the first to provide *cross-community* evaluations of mi-vectors and a contemporaneous model from

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2377–2387,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

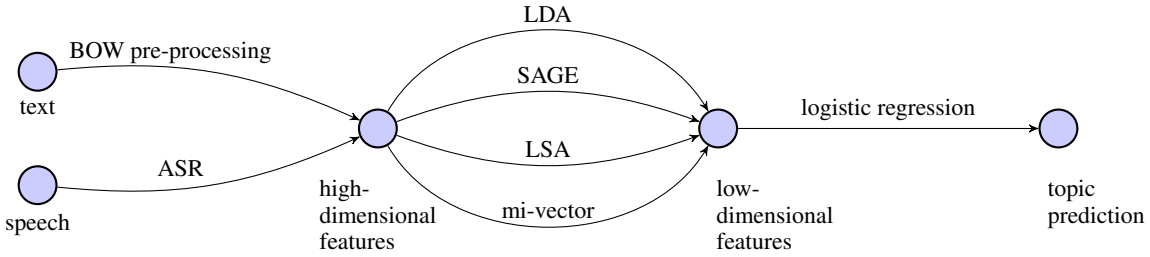

Figure 1: Depiction of the topic ID pipeline. Raw text or speech data is processed into multinomial counts, which are then transformed into a learned representation, and a classifier then predicts the topic of each document based on its representation.

the text community on both text and speech data. This study is also novel in its direct application of the mi-vector model to topic ID and topic discovery, two separate tasks with different motivations and preferring different types of models, and in its use of low-resource triphone state cluster soft counts as speech data for topic ID. The low-resource setting reflects constraints often faced in real-world applications, and we report topic ID performance under limited supervision to better illuminate the practical strengths and weaknesses of the learned representations. Finally, we believe that the centralized comparison herein of several prominent learned representations on two complementary tasks on both text and speech will provide a useful point of reference for future research.

2 Background

Previous work has compared and composed the mi-vector model with older dimensionality reduction techniques, including LDA. Chen et al. (2014) compared a mi-vector language model against LDA and other models on the task of spoken document retrieval, and found the mi-vector model to significantly outperform the other models on words, but not on subwords (syllable pairs), derived from ASR. The syllable pairs are similar in granularity to the triphone state clusters used as multinomial speech data in the current work.

Morchid et al. (2014) improved conversation theme identification by employing LDA and a *Gaussian* i-vector model in a pipeline. They learn LDA models of varying dimensions (numbers of topics) on ASR output and use them to generate a suite of feature vectors. The feature vector for each document-dimension pair is created by marginalizing over topics according to the document's inferred topic proportions. A Gaussian i-vector model is then learned on those feature vectors; the i-vectors are normalized and used to identify document themes via the Bayes decision rule.

Note that we have fundamentally different approaches, goals and methodology from that of Morchid et al. (2014). First, in an effort to provide a scientific comparison of independently created models, we use *multinomial* i-vectors, whereas Morchid et al., focusing on a particular task setting, used traditional *Gaussian* i-vectors. Similarly, while we treat multiple types of topic models as goals in their own rights, directly comparing SAGE and LDA, Morchid et al. use LDA as a pre-processing step to Gaussian i-vectors. Second, we use triphone state cluster soft counts instead of ASR word counts, hence our representation of speech data is significantly lower-resource. Third, we also evaluate performance on text data, and where Morchid et al. limit their vocabulary (from ASR) to 166 task-specific words, we use all 26,606 words present in our training data.

3 Input Representations

Our data is drawn from Part 1 of the Fisher English speech corpus (Cieri et al., 2004c), which contains audio recordings (Cieri et al., 2004a) and manual transcriptions (Cieri et al., 2004b) of telephone conversations. Specifically, we use the topic ID training and evaluation test subsets defined in prior work (Hazen et al., 2007). In each conversation in these subsets of the data, two study participants are prompted to speak on one of a predefined set of forty topics. There are 1374 training conversations and 686 test conversations. We represent each conversation by two documents, one for each side (speaker), resulting in a training set of 2748 documents and a test set of 1372 documents. The deep neural network (DNN) used to infer the triphone state cluster posteriors forming the basis of our speech data was trained on Parts 1 and 2 of the Fisher English speech corpus (Cieri et al., 2004a; Cieri et al., 2005); see the supplement for further

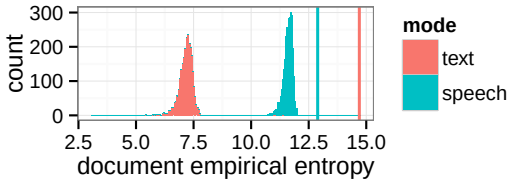

Figure 2: Distributions of the empirical entropy (in bits) of documents under the two multinomial views of our corpus. The vertical lines are the respective upper bounds (entropy of the uniform distributions). The distribution of the entropy of the text documents has median 7.2, over seven bits away from the upper bound of 14.7, thus the text representation is approximately sparse. The speech distribution has median 11.6, within two bits of the upper bound of 12.9, thus the speech representation is nearly uniform.

details about our dataset and ASR system.

To quantify the sparsity of the raw text (word count) and speech (triphone state cluster soft count) representations, we consider the representation density (number of non-zero entries) on our training set. The text representation is sparse, with median density 292 and maximum 500 (out of 26,606 dimensions); the speech representation is dense, with median density 7586 and maximum 7591 (out of 7591 dimensions).

To assess approximate sparsity, we plot histograms of the entropy of the normalized multinomial views of our training set in Figure 2. The median entropy for speech is less than two bits away from the uniform entropy, so the speech data is neither sparse nor approximately sparse.

Finally, we note that topic occurrence in the Fisher English training set is unbalanced, with quartiles (including minimum and maximum) of 6, 18.75, 29.5, 50.25, and 87.

4 Learned Representations

We consider four main dimensionality reduction models: the mi-vector model from the speech community, the SAGE and LDA topic models from the text community, and LSA. The learned representations we consider explain which words appear in a document d via a latent, lower-dimensional representation $\boldsymbol{\theta}^{(d)}$. All representations operate under a bag-of-words assumption. To compare mi-vectors, topic models and LSA, we find it useful to formulate each learned representation as operating on different "areas" or "contexts" a of a document; such a formulation does not negate the fundamental bag-of-words assumption. The four models represent the words that appear in an area a—either the entire document or each token—via multinomial-style parameters $\boldsymbol{\phi}^{(a)}$.[1,2] Each model consists of K *components* (e.g., a K-dimensional affine subspace), and shared parameters $H_{k,v}$ prescribe the amount of weight each component k places on each vocabulary word v. The models construct $\boldsymbol{\phi}^{(a)}$ by combining $\boldsymbol{H}$ and $\boldsymbol{\theta}^{(d)}$; often empirical word statistics $\mathbf{m}$ are also used to stabilize the representations.

4.1 LSA

LSA (Deerwester et al., 1990) factorizes a term-document matrix by truncated SVD, learning the projection of the data onto a linear subspace of fixed rank such that the approximation error of the reconstructed term-document matrix (as measured by the Frobenius norm) is minimized. In the basic version of LSA, SVD is applied to the raw term counts, giving the low-dimensional representation

$$\boldsymbol{\phi}^{(d)} = \boldsymbol{H}\boldsymbol{\theta}^{(d)}, \tag{1}$$

where $\boldsymbol{\phi}^{(d)}$ is the vector of observed multinomial counts in document d, $\boldsymbol{H}$ is the matrix of left singular vectors of the term-document count matrix, and $\boldsymbol{\theta}^{(d)}$ is the inferred representation of $\boldsymbol{\phi}^{(d)}$. In practice, LSA is often applied instead to the term-document matrix weighted by term frequency–inverse document frequency (tf-idf) in order to normalize terms by importance. We can also apply further pre-processing steps, such as term-wise centering by subtracting the column-wise mean $\mathbf{m}$ of the data, in which case LSA finds an affine subspace that approximates the data.

4.2 Mi-vector Model

The original acoustic i-vector model represents continuous, high-dimensional ASR system state (namely, Gaussian mixture model supervectors) in an affine subspace (Dehak et al., 2011). Prior work has found this dense, low-dimensional representation to be effective for a number of tasks, including language recognition (Martínez et al., 2011) and speaker recognition (Dehak et al., 2011; Garcia-Romero and Espy-Wilson, 2011).

Recently the i-vector model was augmented for multinomial observations (Soufifar et al., 2011)

[1] Other efforts have modeled documents with intermediate granularity, e.g., sentence-level (Titov and McDonald, 2008) or entity-level (Newman et al., 2006) granularity.

[2] For brevity, we use the multinomial distribution and its parameter interchangeably throughout.

and applied accordingly to language recognition (Soufifar et al., 2011; McCree and Garcia-Romero, 2015), speaker recognition (Kockmann et al., 2010), and spoken document retrieval (Chen et al., 2014). In this version of the i-vector model the observations are draws from a multinomial and the (unnormalized) natural parameters of that distribution are represented in an affine subspace:

$$\begin{aligned} \boldsymbol{\phi}^{(d)} &= \text{softmax}\left(\mathbf{m} + \boldsymbol{H}\boldsymbol{\theta}^{(d)}\right) \\ \boldsymbol{\theta}^{(d)} &\sim \mathcal{N}(0, \mathbf{I}). \end{aligned} \tag{2}$$

We call this multinomial version of the i-vector model the *mi-vector* model. The latent variable $\boldsymbol{\theta}^{(d)}$ is the *multinomial i-vector*, or *mi-vector*. $\boldsymbol{H}$ is an unconstrained linear transformation. The bias term $\mathbf{m}$ is computed as the log of the l_1-normalized background word count vector. The Gaussian prior on the mi-vector $\boldsymbol{\theta}^{(d)}$ is effectively an l_2 regularizer; mi-vectors are neither non-negative nor sparse in general.

Unlike many Bayesian topic models, word occurrences in the mi-vector model are i.i.d. draws from a document-level multinomial $\boldsymbol{\phi}^{(d)}$; as in LSA, each latent component contributes equally to each word within a given document. Specifically, in the mi-vector model, the natural parameter vector of the multinomial for *all* words in a given document is determined by an additive offset from a background parameter vector.

4.3 Bayesian Discrete Topic Models

Bayesian topic models explain word occurrences via K latent components $\boldsymbol{H}_k$ (topics) each drawn from some prior distribution G. Unlike mi-vectors and LSA, multinomial topic models are admixture models: each token n is drawn from a particular distribution $\boldsymbol{H}_k$. Latent token assignment variables $z_n^{(d)}$, taking integral values between 1 and K, dictate the token's topic choice. A document d controls how often each topic is chosen via the K-dimension multinomial distribution $\boldsymbol{\theta}^{(d)}$. In the parametric settings we consider, Dirichlet priors are often placed on $\boldsymbol{\theta}^{(d)}$, allowing experimentation with the topic representation $\boldsymbol{H}$.[3] A mapping $Q(\boldsymbol{H}_k)$, possibly the identity, ensures $\boldsymbol{\phi}^{(d,n)}$ are probability vectors. A general formulation is then

$$\begin{aligned} \boldsymbol{\phi}^{(d,n)} &= Q\left(\boldsymbol{H}_{z_n^{(d)}}\right) \\ \boldsymbol{H}_k &\sim G\left(\boldsymbol{\eta}\right) \\ z_n^{(d)} &\sim \text{Discrete}\left(\boldsymbol{\theta}^{(d)}\right) \\ \boldsymbol{\theta}^{(d)} &\sim \text{Dirichlet}\left(\boldsymbol{\alpha}\right). \end{aligned} \tag{3}$$

The hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ dictate the informativeness of the priors over $\boldsymbol{H}_k$ and $\boldsymbol{\theta}^{(d)}$: often (empirically optimized) symmetric hyperparameters are employed, resulting in a form of Laplace smoothing during topic estimation. In the current work, we follow this strategy, noting that there have been concerted efforts to encode domain or expert knowledge via the hyperparameters (Gormley et al., 2012; Paul and Dredze, 2015).

SAGE Topic Model The Sparse Additive Generative (SAGE) model (Eisenstein et al., 2011) is a generative Bayesian modeling framework in which $\boldsymbol{\phi}^{(d,n)}$ are formed by summing a background vector and one or more sparse vectors generated from appropriate priors. The additive components can reflect the contributions of documents, aspects, topics, or other factors chosen by the modeler. A basic SAGE topic model sets $\boldsymbol{\phi}^{(d,n)} = \text{softmax}\left(\mathbf{m} + \boldsymbol{H}_{z_n^{(d)}}\right)$, and draws $\boldsymbol{H}_k$ from some sparsity-inducing distribution G, e.g., the Laplace distribution. As $\mathbf{m}$ is a shared background frequency vector, $\boldsymbol{H}_k$ is the learned residual frequency vector of topic k.

Replacing the topic assignment in SAGE by its conditional expectation gives

$$\begin{aligned} \tilde{\boldsymbol{\phi}}^{(d,n)} &= \text{softmax}\left(\mathbf{m} + \mathbb{E}_{z_n^{(d)}}\left[\boldsymbol{H}_{z_n^{(d)}} \middle| \boldsymbol{\theta}^{(d)}, \boldsymbol{H}\right]\right) \\ &= \text{softmax}\left(\mathbf{m} + \boldsymbol{H}\boldsymbol{\theta}^{(d)}\right). \end{aligned} \tag{4}$$

This modification of the SAGE topic model is the same as the mi-vector model but with different regularization on the representation vector $\boldsymbol{\theta}^{(d)}$ and l_1 regularization on the basis vectors $\boldsymbol{H}_k$. This "marginal SAGE" model could be useful in future work: the marginalization may mitigate the problem of topic-switching, yielding a more identifiable (but perhaps less interpretable) model and lending to downstream tasks such as topic ID.

LDA Latent Dirichlet Allocation (LDA) (Blei et al., 2003b) is a generative Bayesian topic model similar to SAGE, but in which each topic is drawn

[3] There have been many efforts to provide or induce latent structure among the topics (Blei et al., 2003a; Li and McCallum, 2006; Wallach et al., 2009; Paul and Girju, 2010), but most models ground out to Dirichlet and discrete random variables.

from a Dirichlet prior G rather than a sparsity-inducing distribution. LDA does not explicitly account for the background distribution; to account for this, it is common practice to threshold the vocabulary *a priori* to remove very common and very rare words (though in our experiments, we do not do this). Therefore, $\boldsymbol{\phi}^{(d,n)}$ is exactly $\boldsymbol{H}_{z_n^{(d)}}$, and $\boldsymbol{H}_k \sim \text{Dirichlet}\,(\boldsymbol{\eta})$.

5 Experiments

We compare these four models of learned representations empirically on two distinct tasks, topic ID and topic discovery. The essential implementation details of the models are as follows; further details are provided in the supplement. We learn the **mi-vector** model in a maximum a posteriori framework as in McCree and Garcia-Romero (2015). Our own C++ implementation of **SAGE**, available online,[4] uses approximate mean-field variational inference, as in Eisenstein et al. (2011). We learn the **LDA** model using Gibbs sampling, implemented in MALLET (McCallum, 2002).[5] We perform **LSA** using centered tf-idf–weighted word counts and centered l_2-normalized triphone state cluster soft counts. We implement tf-idf by scaling the raw term count by the log inverse document frequency. We apply l_2 normalization rather than tf-idf weighting to the speech data because it is dense and tf-idf is thus inappropriate. On both text and speech, mean-centering is performed *after* the respective normalization, as this pre-processing recipe performed best of all the variants we tried.[6]

For each of the four models, the low-dimensional real vector $\boldsymbol{\theta}^{(d)}$ represents a given document d in our experiments. We also consider two high-dimensional baseline representations: **raw** (soft) counts on both the text and speech data, and, only on the text data, **tf-idf**–weighted word counts. These tf-idf weights constitute a high-dimensional *learned* representation.

5.1 Topic ID

In our first topic ID experiment we evaluate topic ID error on raw multinomial views of the data. To our knowledge, we are the first to adopt a multinomial view of triphone state clusters and apply it to topic ID. In subsequent experiments we explore the interaction of representation dimension with each model and dataset, and evaluate relative performance when the classifier is only given a fraction of the available data for training. This latter configuration is the most interesting, as it reflects the cost of obtaining supervised data in practice.

Given feature vectors for some representation of the documents in a corpus, topic ID is performed in a one-versus-all framework. We use logistic regression as the per-class binary classifier, implemented using LIBLINEAR (Fan et al., 2008). Results were similar when logistic regression was replaced by support vector machines. All document representations are length-normalized (divided by their l_2 norm) before they are input to the classifier. Performance is measured by topic ID error, the error of multi-class prediction where the class predicted for each document is that of the per-class classifier that gave it the highest weight. Baseline performance on the test set (where the baseline classifier chooses the most prevalent topic in the training set for all test examples) is 96.2% error. Note that this error rate differs from the uniform-at-random classification error rate of 97.5% because of the uneven distribution of topics.

Document Construction Prior work (Hazen et al., 2007; Wintrode and Khudanpur, 2014) treated whole conversations as documents in addition to separating each conversation into its two sides. We perform a small topic ID experiment in this configuration to probe the impact of this design choice. Ten-fold cross-validation (CV) is used to tune the logistic regression regularizers. On the test set, the classifier achieves topic ID error of 12.4% and 15.6% for whole-conversation and individual-side text data, respectively, and 20.1% and 29.5% for whole-conversation and individual-side speech data, respectively. These results correspond roughly to results listed in Table 3 of Hazen et al. (2007), specifically, the topic ID error of 8.2% and 12.4% for whole-conversation and individual-side transcriptions, respectively, and 22.9% and 35.3% for whole-conversation and individual-side triphones derived from ASR lattices, respectively (Hazen et al., 2007). However, we use logistic regression without feature selection instead of Naïve Bayes with feature selection, and we apply our classifier to triphone state cluster soft counts inferred by a DNN instead of triphone

[4] https://github.com/fmof/sagepp

[5] For Gibbs sampling, fractional counts are truncated.

[6] Results for other versions of LSA are provided in the supplement. We did not present the conventional, uncentered tf-idf weighting scheme here because although it performs best in topic ID, it yields extremely variable V-measure.

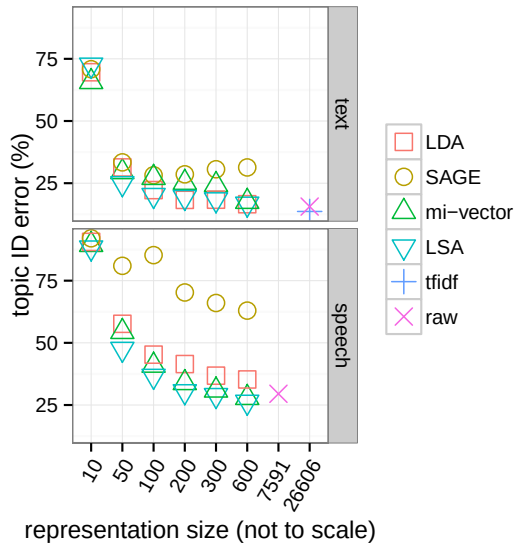

Figure 3: Topic ID error (%) on the test set for raw and tf-idf representations and lower-dimensional learned representations at dimensions $K \in \{10, 50, 100, 200, 300, 600\}$. We see many of the learned representations approach the error rate of the raw representation, but at much lower dimensionality.

counts from ASR lattices. We believe that the discrepancies in performance with respect to prior work are due to these differences in experimental configuration. Our results and those of prior work show that using whole-conversation documents instead of individual-side documents make the topic ID task easier. As a result, we expect that differences in performance between the different learned representations will be more clearly pronounced on individual conversation sides and we restrict the rest of our study to that setting.

Dimensionality Study We perform topic ID on learned representations at dimensions $K \in \{10, 50, 100, 200, 300, 600\}$ on individual conversation sides, using ten-fold cross-validation to tune the logistic regression regularizers. Figure 3 gives topic ID error results on the test set, varying K. (Selected values are listed in Table 1.) In both datasets, as the dimension K increases, topic ID error decreases, approaching (approximately) the raw baseline. On text, tf-idf performs slightly better than the raw representation. LSA is marginally the best-performing lower-dimensional learned representation; LDA and mi-vectors perform well at some representation sizes, depending on the data source, but their performance is less consistent. SAGE performs poorly overall.

view	model	dimension	error
text	LDA	600	16.5
text	SAGE	600	31.3
text	mi-vector	600	17.6
text	LSA	600	16.7
text	tf-idf	26,606	13.6
text	raw	26,606	15.6
speech	LDA	600	35.3
speech	SAGE	600	63.0
speech	mi-vector	600	27.9
speech	LSA	600	26.2
speech	raw	7591	29.5

Table 1: Selected topic ID error (%) values from Figure 3.

Limited Data Study The raw text and speech representations (multinomial observations) are very high-dimensional, and the classifier is likely to overfit to specific components (words or triphone state clusters) in these representations. To measure this effect and attempt to separate the predictive power of logistic regression from the quality of the learned representations in our analysis, we experiment with reducing the number of *labeled* training examples the *classifier* can use; we still learn representations on the full (unlabeled) training set. This experiment represents the limited-supervision setting in which supervised data is costly to obtain but unlabeled data abounds.

We run this experiment twice, using $\ell = 2$ and $\ell = 6$ labeled examples per topic, for a total of 80 and 240 classifier training examples, respectively. Ten-fold cross-validation is used to fit the regularizer; per-class loss coefficients are set according to the class prior in the original training set in order to counteract the artificial balancing of the classes in the limited-supervision dataset. We report cross-validation estimates of the topic ID error on the training set for $K = 10$ (Figure 4), $K = 100$ (Figure 5), and $K = 600$ (Figure 6). For $K = 100$ and $K = 600$, LSA dominates in the limited-supervision setting. Mi-vectors perform as well as or better than other low-dimensional learned representations at $K = 10$, and exhibit mixed performance for larger K. SAGE performs poorly overall.[7] LDA performs significantly bet-

[7] We believe that approximately sparse posterior $\boldsymbol{\theta}^{(d)}$ values result in a kind of topic switching, contributing to the poor performance of SAGE. To examine this we "tested on train" and analyzed the top topics inferred for each document: while the highest-weighted topic tended to be consistent, SAGE infers approximately sparse $\boldsymbol{\theta}^{(d)}$ with large variation in the next four highest-weighted topics (the remaining topics are assigned trace mass). Second, a phenomenon known as conversation drift, explained in Section 3 of the supplement, is so pronounced in Fisher that the first 25% percent of words of each conversation side are nearly as predictive as the entire

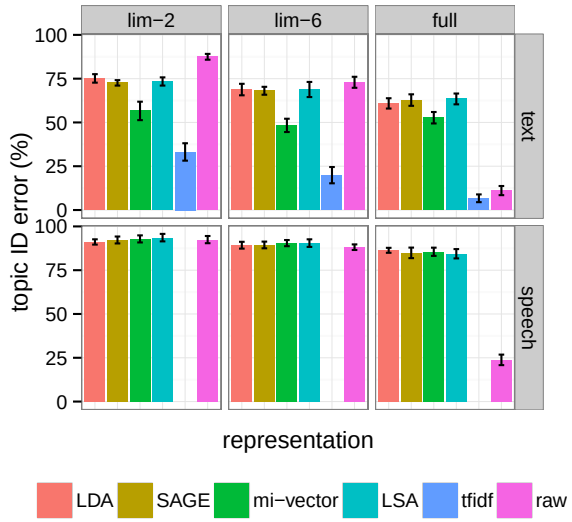

Figure 4: CV topic ID error (%) for raw and tf-idf representations and lower-dimensional learned representations of size $K = 10$. Error bars denote plus and minus one standard deviation according to the CV empirical distribution. We see underparametrized mi-vectors excel at compressing the topic label information for text, particularly in the limited-supervision settings.

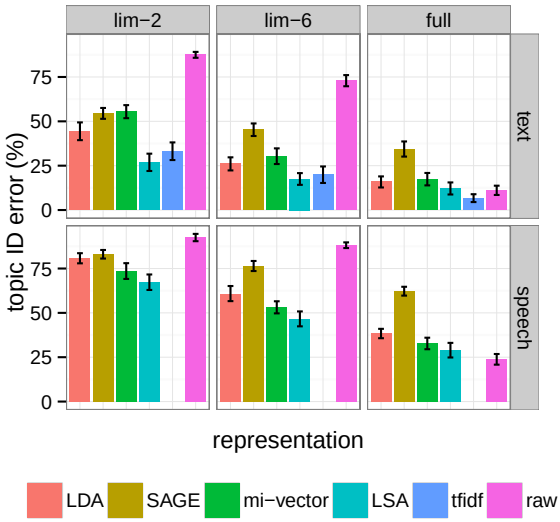

Figure 5: CV topic ID error (%) for raw and tf-idf representations and lower-dimensional learned representations of size $K = 100$. Error bars denote plus and minus one standard deviation according to the CV empirical distribution.

view	model	dimension	V-measure
text	LDA	600	0.517
text	SAGE	600	0.663
text	mi-vector	600	0.507
text	LSA	600	0.525
text	tf-idf	26,606	0.626
text	raw	26,606	0.134
speech	LDA	600	0.358
speech	SAGE	600	0.511
speech	mi-vector	600	0.468
speech	LSA	600	0.190
speech	raw	7591	0.132

Table 2: Selected V-measure values from Figure 7.

ter than SAGE, but not as well as mi-vectors. Finally, tf-idf–weighted word counts perform very well on text, often achieving the best performance of all representations, even under limited supervision (but at the same dimension as the raw data).

5.2 Topic Discovery

To quantitatively assess representations' potential for topic discovery we compute their V-measure against the gold-standard labels. V-measure is an unsupervised measure of similarity between two partitions (Rosenberg and Hirschberg, 2007) and is equivalent to the mutual information normalized by the sum of the entropy (Becker, 2011).

For all representations, we compute V-measure between a partition induced by that representation and the gold-standard topic labels on the test set. A partition is induced on a representation by assigning each document d to the cluster indexed by the coordinate of $\boldsymbol{\theta}^{(d)}$ with highest value (the argmax). Results of this analysis are displayed in Figure 7. (Selected values are listed in Table 2.) On the text data, SAGE dominates the lower-dimensional representations, LSA is next best overall, and LDA and mi-vectors exhibit relatively low performance; the high-dimensional tf-idf weights are surpassed by SAGE for $K > 10$ but beat other representations by a significant margin. On speech, SAGE is best overall, mi-vectors exhibit similar but generally lower performance, LDA performs worse, and LSA is worst.

We also measure the topic discovery potential of the mi-vector and SAGE representations more directly. First, we provide a manual inspection of the learned topics: in Table 3 we show the top-five word lists for five random topics from the 600-dimensional mi-vector and SAGE models (respectively) learned on the text data. In both models, the top five words in a topic are selected according to the five largest positive values in the corresponding vector $\boldsymbol{H}_k$. Qualitatively, the SAGE topics are considerably more interpretable than the mi-vector topics: the SAGE topics represent issues of censorship, foreign relations, coffee fran-

document (Wintrode, 2013). All representations must contend with this drift, but $\boldsymbol{\theta}^{(d)}$ sparsity may make SAGE particularly susceptible. These two issues may make the classification we use much less robust.

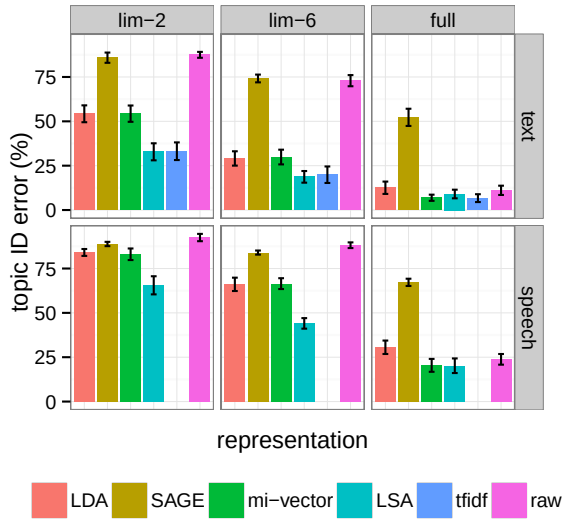

Figure 6: CV topic ID error (%) for raw and tf-idf representations and lower-dimensional learned representations of size $K = 600$. Error bars denote plus and minus one standard deviation according to the CV empirical distribution.

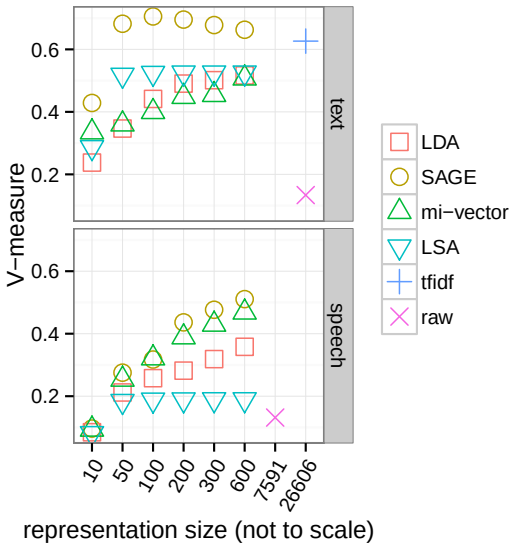

Figure 7: V-measure on the Fisher English text and speech data, respectively, for raw and tf-idf representations and lower-dimensional learned representations at selected dimensions. As in topic ID, we see underparametrized mi-vectors perform well on the text data.

chises, welfare, and professional basketball, while the mi-vector topics are less succinctly characterizable and more polluted by uninformative words.

We complement this qualitative analysis with Mimno et al. (2011)'s intrinsic coherence measure, a standard quantitative method. This scoring function, which correlates well with human quality judgments, averages estimates of the conditional log-likelihoods of each topic's M highest-weighted words across all topics. Using $K = 600$ models on text as before and picking $M = 20$, we compute mi-vector coherence as -453.34 and SAGE coherence (averaged over three runs) as -407.52, indicating that SAGE is more amenable to topic discovery and human interaction.

mi-vector
you've, florida, each, a-, bit
hours, never, couldn't, check, communicate
pregnant, water, lifestyle, awful, called
forgot, ran, social, topics, unique
tough, way, let's, fifties, hand

SAGE
censor, books, censorship, neat, agree
sanctions, siblings, democratic, rely, u._n.
starbucks, franchise, coffee, franchising, studio
welfare, wage, minimum, cents, tips
team, role, professional, blazers, basketball

Table 3: Top five words in five random mi-vector and SAGE topics learned on text data at $K = 600$.

6 Discussion

We have theoretically and empirically compared several content-bearing representations of text and speech from prior work. We have measured the relative performance of these representations on topic ID, an easy-to-evaluate task familiar to both text and speech research communities. We have also assessed the representations in their ability to discover the topics inherent in a corpus, a task that is more prominent in the text community and more difficult to evaluate. On our subset of the Fisher English data, these tasks appear to have competing objectives: the best representations in one task are not necessarily the best in the other. In particular, while SAGE yields the worst performance as a feature learner for topic ID, it is demonstrably superior to other low-dimensional learned representations in topic discovery. We have evaluated performance in topic discovery by distributional similarity to gold-standard topics as a proxy for human-annotated judgments of topic quality, and briefly compared the interpretability of mi-vectors and SAGE; future work could pursue expert or crowd-sourced human evaluations.

In the full-supervision setting of topic ID, the lower-dimensional learned representations converge in performance to the raw representation as the dimension K increases. However, if only a couple of labeled examples per class are available,

reflecting the expense of obtaining labels in practice, then learned representations generally outperform the raw representation, which is more prone to overfitting. It is surprising that tf-idf performs so well in the limited supervision setting; it is learned from the data, but it should be prone to overfitting due to its high dimensionality. It is also surprising that SAGE performance on text degrades significantly at high dimensions; we suspect this is due to topic switching, but further investigation is warranted. Overall, though, for topic ID on word counts or triphone state cluster soft counts, if labeled data is scarce, we benefit from training on unsupervised learned representations.

In the V-measure experiment, the documents were partitioned according to the heaviest coordinate in their representations. This choice of experimental protocol is a nuisance variable in our results; other partition constructions may yield different conclusions. In particular, the heaviest-coordinate partition may favor topic models, whose representations are probability vectors, and disfavor mi-vectors and LSA, whose representations may have positive and negative coordinates encoding general linear combinations.

Within each task, the ranking of the representations (by performance) is generally consistent between the text and speech data; however, mi-vectors often outperform LDA on the speech data, while LDA often outperforms mi-vectors on the text data. This may be evidence that the two communities have already independently identified appropriate dimensionality reduction techniques for their respective data sources. However, our results support that the speech community can benefit from broader use of sparsity-inducing graphical models such as SAGE in tasks like spoken topic discovery and recommendation, in which human-interpretable representations are desired. The text community may similarly benefit from parsimonious models such as LSA or mi-vectors in downstream tasks; underparametrized mi-vectors perform particularly well on text, and future work may benefit from investigating this setting.

Word counts and triphone state cluster soft counts provide only one view of text and speech (respectively), and other input representations may yield different conclusions. The particular LSA approach we used for text, based on tf-idf weighting, is not as appropriate for our speech data, which is dense. Future work could evaluate other implementations of LSA or use a higher-level view of speech, such as triphone state cluster n-grams, that more naturally exhibits sparsity and lends to tf-idf weighting. In particular, weighting by a likelihood ratio test statistic and applying a log transform has generated better performance in several other tasks (Lapesa and Evert, 2014). Future work could also test our conclusions on higher-resource views of speech, such as ASR word counts, or lower-resource views such as mel-frequency cepstral coefficients (MFCCs).

We have provided a brief cross-community evaluation of learned representations on multinomial text and speech data. Some prior work has evaluated related learned representations on text data alone, surveying parameters and tasks at greater breadth (Lapesa and Evert, 2014; Levy et al., 2015). A similarly comprehensive evaluation spanning the text and speech research communities would demand great effort but provide a large and versatile resource. In complement, a detailed, case-by-case analysis of errors made by the models in our study could illuminate future modeling efforts by exposing exactly how and why each model errs or excels in each task.

7 Conclusion

Topic ID and topic discovery are competing objectives in our setting: we found that the best-performing representations per task were the same whether considering text- or speech-based communications. By evaluating learned representations from both the text and speech communities on a common set of data and tasks, we have provided a framework for better understanding the topic ID and topic discovery objectives, among others. More generally, we hope to encourage cross-community collaboration to accelerate convergence toward comprehensive models of language.

Acknowledgments

We would like to thank the three anonymous reviewers for their feedback. A National Science Foundation Graduate Research Fellowship, under Grant No. DGE-1232825, supported the second author. We would like to thank the Johns Hopkins HLTCOE for providing support. Any opinions expressed in this work are those of the authors.

References

Hila Becker. 2011. *Identification and Characterization of Events in Social Media*. Ph.D. thesis, Columbia University.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2003a. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems 16 (NIPS)*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003b. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Kuan-Yu Chen, Hung-Shin Lee, Hsin-Min Wang, Berlin Chen, and Hsin-Hsi Chen. 2014. I-vector based language modeling for spoken document retrieval. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 7083–7088.

Christopher Cieri, David Graff, Owen Kimball, Dave Miller, and Kevin Walker. 2004a. Fisher english training speech part 1 speech LDC2004S13. DVD.

Christopher Cieri, David Graff, Owen Kimball, Dave Miller, and Kevin Walker. 2004b. Fisher english training speech part 1 transcripts LDC2004T19. Web Download.

Christopher Cieri, David Miller, and Kevin Walker. 2004c. The fisher corpus: a resource for the next generations of speech-to-text. In *International Conference on Language Resources and Evaluation (LREC)*, pages 69–71.

Christopher Cieri, David Graff, Owen Kimball, Dave Miller, and Kevin Walker. 2005. Fisher english training part 2, speech LDC2005S13. DVD.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society For Information Science*, 41(6):391–407.

Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2011. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 1041–1048.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Daniel Garcia-Romero and Carol Y. Espy-Wilson. 2011. Analysis of i-vector length normalization in speaker recognition systems. In *Interspeech*, pages 249–252.

Matthew R Gormley, Mark Dredze, Benjamin Van Durme, and Jason Eisner. 2012. Shared components topic models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pages 783–792.

Timothy J. Hazen, Fred Richardson, and Anna Margolis. 2007. Topic identification from audio recordings using word and phone recognition lattices. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 659–664.

Marcel Kockmann, Lukáš Burget, Ondřej Glembek, Luciana Ferrer, and Jan Černocký. 2010. Prosodic speaker verification using subspace multinomial models with intersession compensation. In *Interspeech*, pages 1061–1064.

Gabriella Lapesa and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2:531–545.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23th International Conference on Machine Learning (ICML)*, pages 577–584.

David Martínez, Oldřich Plchot, Lukáš Burget, Ondřej Glembek, and Pavel Matějka. 2011. Language recognition in ivectors space. In *Interspeech*, pages 861–864.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Alan McCree and Daniel Garcia-Romero. 2015. DNN senone MAP multinomial i-vectors for phonotactic language recognition. In *Interspeech*. To appear.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 262–272.

Mohamed Morchid, Mohamed Bouallegue, Richard Dufour, Georges Linarès, Driss Matrouf, and Renato de Mori. 2014. An i-vector based approach to compact multi-granularity topic spaces representation of textual documents. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 443–454.

David Newman, Chaitanya Chemudugunta, and Padhraic Smyth. 2006. Statistical entity-topic models. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 680–686.

Michael Paul and Mark Dredze. 2015. Sprite: Generalizing topic models with structured priors. *Transactions of the Association for Computational Linguistics*, 3:43–57.

Michael Paul and Roxana Girju. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pages 545–550.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 EMNLP-CoNLL Joint Conference*, pages 410–420.

Mehdi Soufifar, Marcel Kockmann, Lukáš Burget, Oldřich Plchot, Ondřej Glembek, and Torbjørn Svendsen. 2011. iVector approach to phonotactic language recognition. In *Interspeech*, pages 2913–2916.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International World Wide Web Conference (WWW)*, pages 111–120.

Hanna M. Wallach, David Mimno, and Andrew McCallum. 2009. Rethinking lda: Why priors matter. In *Advances in Neural Information Processing Systems 22 (NIPS)*.

Jonathan Wintrode and Sanjeev Khudanpur. 2014. Limited resource term detection for effective topic identification of speech. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pages 7118–7122.

Jonathan Wintrode. 2013. Leveraging locality for topic identification of conversational speech. In *Interspeech*, pages 1579–1583.