# MODELING MULTIWORD PHRASES WITH CONSTRAINED PHRASE TREES FOR IMPROVED TOPIC MODELING OF CONVERSATIONAL SPEECH[1]

*Timothy J. Hazen and Fred Richardson*

MIT Lincoln Laboratory
Lexington, Massachusetts, USA

## ABSTRACT

Latent topic modeling has proven to be an effective means for learning the underlying semantic content within document collections. Latent topic modeling has traditionally been applied to bag-of-words representations that ignore word sequence information that can aid in semantic understanding. In this work we introduce a method for efficiently incorporating arbitrarily long word sequences into a topic modeling approach. This method iteratively constructs a constrained set of phrase trees in an unsupervised fashion from a document collection using weighted pointwise mutual information statistics to guide the process. In experiments on the Fisher Corpus of conversational speech, the incorporation of learned phrases into a latent topic model yielded significant improvements in the unsupervised discovery of the known topics present within the data.

***Index Terms***— topic modeling, phrases, conversational speech

## 1. INTRODUCTION

Latent topic modeling techniques can provide an effective means for improving a wide variety of text and speech applications including document clustering, document link detection, query-by-example document retrieval, document summarization, corpus summarization, and automatic speech recognition. Modeling approaches such as probabilistic latent semantic analysis (PLSA) [1] and latent Dirichlet allocation (LDA) [2] can learned a low dimension space of latent topics in a completely unsupervised fashion. For a variety of different corpora, these techniques have demonstrated the capability to learn latent topic spaces that efficiently describe the underlying semantic concepts contained in the data [3].

Traditionally, latent topic models operate on document collections that are represented using a *bag-of-words* modeling assumption, i.e. the observation of each word is deemed conditionally independent of the observations of all other words in a document. Under this assumption, word ordering information is completely ignored, and each document is simply represented by the counts of the individual words present within it. Despite its simplicity, the bag-of-words approach to latent topic modeling has been shown to work well for many tasks.

Though the bag-of-words approach has proven effective, multi-word sequences often convey additional information that is not available when individual words are viewed in isolation. Thus, recent research efforts have attempted to move beyond unigram words to identify topically relevant word sequences. One approach is to label each word in each document with its most dominant topic label as determined by a bag-of-words topic model. From this labeling of the document collection, common multi-word sequences that share the same topic label can be extracted and used for the summarization of the topics in a document or document collection [4, 5]. Though effective, this approach still retains the bag-of-words assumption during training, and it is reasonable to assume that topic modeling improvements could be attained if the models directly incorporated knowledge of informative multi-word sequences or phrases.

There are two potential mechanisms for incorporating multi-word sequences into latent topic models. The first (and simpler) way is to pre-learn a collection of multi-word phrases and incorporate these sequences as single units into a standard bag-of-words modeling approach. The difficulty in this approach is determining a relevant set of multi-word sequences without any prior topical knowledge. Prior approaches to this problem include adding multi-word expressions present in WordNet [6], or from gloss look-ups of sequences present in Wikipedia titles or search engine query logs [7].

The second (and more complicated) method is to learn relevant multi-word sequences and topics jointly. This has been attempted through the addition of dependency links within an LDA model. The first example of this approach is the bigram topic model [8], which adds word dependency links into a standard LDA model such that each word is modeled in the context of the previous word. The added dependency structure dramatically increases the number of parameters in the topic language model by a factor equal to the size of the model's unigram word vocabulary. The LDA collocation model [9] builds upon the topical bigram model by incorporating a hidden Boolean collocation variable for each bigram that allows the training process to incorporate only the specific bigram pairs that improve the model. The topical $n$-gram model [10] extends the LDA collocation model further by allowing each hidden bigram collocation variable to be further dependent upon the latent topic variable. See [10] for a detailed comparison of these models.

While there may be advantages to the joint learning employed by the three methods listed above, there are also disadvantages to these methods as they were implemented. First, the models explicitly only learn bigram sequences and not longer sequences. Second, the incorporation of the added dependency structure increases the complexity of the model which can increase both the training time as well the potential for training instability or over-fitting. To avoid these issues, we explore an iterative tree-growing method for pre-determining the phrasal units to be incorporated into the model vocabulary based on mutual information statistics. Once the vocabulary of phrasal units is determined, standard bag-of-words latent modeling approaches such as PLSA and LDA can be applied without explicitly requiring the inclusion of any additional dependency structure into the model.

[1] This work was sponsored by the Department of Defense 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

978-1-4673-5126-3/12/$31.00 ©2012 IEEE

## 2. LATENT TOPIC MODELING OF DOCUMENTS

### 2.1. Document Representation

Before explaining our new approach, we first define the basic elements of traditional latent topic models. These approaches operate upon a collection $D$ of $N_D$ different documents:

$$D = \{d_1, \ldots, d_{N_D}\} \tag{1}$$

Each document $d_i$ is comprised of an ordered string of the $N_i$ words present in the document, as expressed as:

$$W_i = \{w_1, \ldots, w_{N_i}\} \tag{2}$$

The collection of all words present across all documents defines a vocabulary $V$ of $N_V$ unique words, as expressed as:

$$V = \{w_1, \ldots, w_{N_V}\} \tag{3}$$

When using a bag-of-words independence assumption, the word string $W_i$ for a document $d_i$ can be alternatively represented by the counts of each of the $N_V$ vocabulary words present in the document:

$$C_i = \{c_1, \ldots, c_{N_V}\} \tag{4}$$

The counts contained in $C_i$ provide a sparse feature vector within a feature space of dimension $N_V$, with most of the dimensions for each document typically having a count value of zero.

### 2.2. Stop-Listing

Many common words (e.g., articles, conjunctions, prepositions) provide little value for topic modeling. It is has become common practice in text-based topic modeling to create a hand crafted *stop list* of these common words which are then removed from the vocabulary $V$ prior to model training. For speech data the standard stop list can be expanded to include common conversational artifacts (e.g, filled pauses, back-channels, etc.). Statistical stop-listing can also be applied to remove any additional words that are exceedingly common or rare in the specific document collection being modeled.

### 2.3. Probabilistic Latent Semantic Analysis

Within latent topic models, documents in a collection are modeled using a weighted combination of $N_Z$ latent topics from a set $Z$:

$$Z = \{z_1, \ldots, z_{N_Z}\} \tag{5}$$

In the PLSA approach, each latent topic possesses a probabilistic unigram language model $P(w|z)$ representing the likelihood that word $w$ could be randomly generated by topic $z$. Each document $d_i$ in the document collection $D$ is then assumed to generate topics from a weighted mixture $P(z|d_i)$ over the latent topics in $Z$. The full PLSA likelihood function for observing the collection of word counts $C_i$ associated with document $d_i$ is expressed as:

$$P(C_i|d_i) = \prod_{w \in V} \left( \sum_{z \in Z} P(w|z) P(z|d_i) \right)^{c_w} \tag{6}$$

PLSA learns the unigram language models $P(w|z)$ and the document specific latent topic distributions $P(z|d_i)$ using the EM algorithm applied over the full document collection [1]. Though our work here uses PLSA, the LDA approach could also be employed.

## 3. CONSTRAINED PHRASE TREE MODELING

### 3.1. Goals

The primary focus of this paper is the improvement of latent topic models through the incorporation of multi-word sequences or phrases. In this work we liberally define a phrase to be any word sequence that follows constraints that we define below (i.e., we do not use the stricter definition of a phrase as applied in syntactic linguistics). When creating a method for phrase incorporation, there are several design goals that we wish to achieve:

1. The learned phrases should be as informative as possible.
2. Phrases of variable length should be allowed.
3. The incorporation of phrases should not require changes to the standard bag-of-words latent topic training algorithm.

We discuss these goals and our design choices for achieving these goals below.

### 3.2. Learning Informative Phrases

While some multi-word sequences provide more information when viewed in sequence than when treated independently, not all multi-word sequences are informative. Word bigram sequences that are informative can be determined by examining their weighted pointwise mutual information (WPMI) score within the document collection. The WPMI score for a bigram sequence $\{w_i, w_j\}$ is expressed as:

$$\text{wpmi}(w_i, w_j) = p(w_i, w_j) \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{7}$$

In this expression, the log term represents the pointwise mutual information (PMI) between $w_i$ and $w_j$ as estimated from the document collection. For the PMI term, we note that:

$$\text{pmi}(w_i, w_j) > 0 \iff p(w_i, w_j) > p(w_i)p(w_j) \tag{8}$$

Thus, bigrams with a positive PMI measure are statistically more informative than the bigram's constituent words themselves. By ranking all bigram sequences (excluding those that contain stop words) using the full WPMI measure, the top ranked bigrams will be those that are both frequent (as based on the initial $p(w_i, w_j)$ term) and informative (as based on the PMI measure).

Using the approach above, standard bigram sequences such as *minimum wage* or *affirmative action* can be learned. While these bigrams are informative, it would also be useful to learn longer phrases that can potentially contain words from the stop list, e.g., *cost of living* or *rest of the world*. This can be achieved by allowing phrases of the constrained form $w_i, s_j, w_k$ or $w_i, s_j, s_k, w_l$ to be considered, where $s_j$ and $s_k$ represent stop words that would usually be ignored during topic modeling. The WPMI scoring function can be generalized to incorporate inter-bigram stop word sequences as follows:

$$\text{wpmi}(w_i, s*, w_j) = p(w_i, s*, w_j) \log \frac{p(w_i, s*, w_j)}{p(w_i)p(s*, w_j)} \tag{9}$$

Here $s*$ is used to represent a stop word sequence containing zero, one, or two stop words and each of the probability expressions $p(\cdot)$ is a maximum likelihood estimate obtained directly from counts obtained from the document collection.

### 3.3. Growing Variable Length Phrases

Variable length phrases can be learned through an iterative process in which learned phrases are added as new independent units into the system vocabulary. These learned phrase units replace their constituent units within the documents and their associated count vectors. For example, if the learning process selects the phrase *minimum wage*, then the new phrase unit *minimum_wage* is incorporated into the system vocabulary and replaces the bigram sequence *minimum wage* across the document collection. All associated counts for the units *minimum*, *wage*, and *minimum_wage* are also adjusted appropriately. In essence the phrase learning process produces a rewrite rule for each learned phrase, such as:

| | | |
|---|---|---|
| middle east | → | middle_east |
| minimum wage | → | minimum_wage |
| new york | → | new_york |

By iteratively adding new phrases, the system can grow longer phrases from shorter phrases. This learning process generates a series of phrase rules that can be applied sequentially. For example, in the following phrase rules, it can be seen that a phrase rule covering the 5-word sequence *peace in the middle east* can be constructed if the phrase unit *middle_east* has been previously learned:

| | | |
|---|---|---|
| peace in the middle_east | → | peace_in_the_middle_east |
| earning minimum_wage | → | earning_minimum_wage |
| new_york city | → | new_york_city |

By learning an ordered sequence of phrase rules, the conversion of unigram word sequences into phrase unit sequences is deterministic and can be efficiently represented using finite state transducers. For phrases constructed from multiple rules, the rule sequence can also be viewed as producing a hierarchical parse tree. For example, Fig. 1 shows the phrase parse tree for the phrase *peace in the middle east*. The dashed lines in the tree represent the *stop words* while solid lines represent the *content words*. Because any word sequence can be deterministically parsed into a sequence of words and/or constrained tree-structured phrases, we refer to this modeling representation as *constrained phrase trees*. The iterative tree growing process continues until a set of pre-determined stopping criteria are met.

### 3.4. Latent Topic Modeling with Phrases

Standard latent topic modeling approaches such as PLSA or LDA do not need any modification to operate on the output of the phrase discovery algorithm. Because the phrase discovery algorithm simply replaces word sequences with multiword phrase units, documents can still be represented using simple count vectors. The only difference is that the active vocabulary for topic modeling is expanded to include all of the learned phrase units.

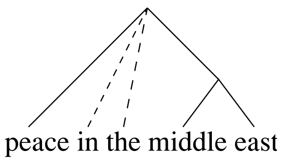

**Fig. 1**. Example of a learned hierarchical phrase tree.

## 4. EXPERIMENTS AND RESULTS

### 4.1. Corpus

For our experiments we have used the text transcripts of a collection of 5850 conversations extracted from the English Phase 1 portion of the Fisher Corpus [11]. 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. "Family", "Family Values", "Life Partners"). Although instructed to discuss a specific topic, participants occasionally strayed off-topic and discussed other non-prompted topics.

### 4.2. Phrase Learning

Using the process described in Section 3, we apply iterative phrase tree learning to the full collection of Fisher text transcripts. Before any phrases are learned, the system begins with a unigram vocabulary of 19,725 unique words. The initial unigram vocabulary excludes words from our own manually crafted stop-list of 594 words. It also excludes other statistically common words that appear in greater than 25% of the documents and rare words that occur less than 3 times total in the data.

Starting from the initial vocabulary, phrase rules are sequentially added to the rule list using the WPMI rankings. After each new phrase is added, the counts of the words and phrases affected by the inclusion of the new phrase are updated and the WPMI rankings are readjusted. Rules are only added if (1) the observed phrase occurs at least 3 times in the corpus and (2) the WPMI score is positive. Phrase learning ends when there are no potential phrases left that satisfy these two constraints. For the Fisher Corpus this resulted in 20,122 phrases being added into the system vocabulary. Thus the vocabulary of the system after all learned phrases are added is roughly double the size of the original unigram vocabulary.

Table 1 shows the first 20 basic bigram phrases (i.e., those multiword sequences with two content words separated by zero, one or two common stop words). The table contains a mix of phrases relevant to the Fisher Corpus topics (*minimum wage*, *million dollars*, *affirmative action*), phrases related to common geographic locations (*new york*, *los angeles*, *san francisco*), common conversational phrases (*pretty good*, *long time*), and phrases related to the data collection process (*supposed to talk*, *ten minutes*).

Table 2 shows the first 10 hierarchical phrases that were constructed using previously learned bigram phrases. Table 3 shows some of the longest phrases learned by the process. In this table, phrases (1) through (6) are directly related to specific prompted topics. Phrases (7) through (9) are related to the discussion of the Fisher data collection process. At 10 words long, phrase (9) is the longest learned phrase and it corresponds to the telephone number that subjects called to initiate new Fisher conversations. Phrase (10) is commonly used when callers introduce themselves to each other.

### 4.3. PLSA Training

For PLSA training we use the same approach described in [12] to train both our unigram and phrase-based systems. For all systems, PLSA models are trained using a fixed setting of 40 latent topics matching the known number of prompted topics in the corpus. In reality it is unclear what the appropriate number of latent topics

**Table 1**. Top twenty ranked learned bigram phrases in Fisher.

| | |
|---|---|
| new york | long time |
| minimum wage | nine eleven |
| high school | five fifteen |
| ten minutes | los angeles |
| united states | middle east |
| years ago | affirmative action |
| million dollars | san francisco |
| september eleventh | kind of thing |
| supposed to talk | life partner |
| pretty good | new jersey |

**Table 2**. First ten hierarchical phrases learned in Fisher.

| |
|---|
| new_york city |
| five_fifteen an hour |
| weapons of mass_destruction |
| long_time ago |
| couple of years_ago |
| world trade_center |
| osama bin_laden |
| public_school system |
| ten years_ago |
| stay_at_home mom |

**Table 3**. A sample of ten of the longest phrases learned in Fisher.

| | |
|---|---|
| (1) | five_dollars_and_fifteen_cents an hour |
| (2) | million_dollars_to_leave the united_states |
| (3) | draw_the_line_between_acceptable_humor and humor |
| (4) | commit_perjury for a close_friend_or_family_member |
| (5) | guess_we're_supposed_to_talk about comedy |
| (6) | guess_we're_supposed_to_talk about minimum_wage |
| (7) | guess_we're_supposed_to_talk for ten_minutes |
| (8) | opportunity to leave_feedback_about_the_call |
| (9) | eight_six_six_six_eight_seven four_seven_five_eight |
| (10) | part_of_the_country do you live |

should be, as the Fisher conversations also contain off-topic diversions. In fact, automatically determining the number of latent topics to be used remains an open research question [12]. While the selection of the number of topics will undoubtedly affect the absolute performance of a topic model, in this work we are primarily concerned with relative comparisons between different feature sets. Thus, keeping a fixed number of topics for all models is appropriate for our experimental comparisons.

### 4.4. Corpus Summarization

To understand qualitatively how the incorporation of phrases affects topic modeling, we can examine and compare automatically generated summaries created from different PLSA models. In our summaries, the latent topics are automatically ranked by a topical importance score and a collection of *signature* words/phrases are presented to describe the semantic content for each latent topic. The mechanisms for ranking the topics and selecting the signature words/phrases are described in detail in [12]. Table 4 shows the summary generated from the PLSA model learned from purely unigram features, while Table 5 shows the summary generated after the full set of 20,122 learned phrases were incorporated into feature set.

The third column in the summaries shows the best matching Fisher topic $t$ for a given latent topic $z$ along with the topical match overlap score $P(t|z)$. This score measures the proportion of data modeled by latent topic $z$ that is derived from documents initiated with the prompted topic $t$ and is computed using this expression:

$$P(t|z) = \frac{\sum_{\forall d \in D_t} |d| \cdot P(z|d)}{\sum_{\forall d \in D} |d| \cdot P(z|d)} \qquad (10)$$

Here, $D_t$ is the subset of documents in the document collection $D$ associated with the known Fisher topic $t$ and $|d|$ is the number of modeled words contained in document $d$. In essence, $P(t|z)$ provides a measure of overlap between $t$ and $z$

Qualitatively, the models learned when using the learned phrases improve the latent topic models in two observable ways. First, improvement is generally observed in the overlap between the learned latent topics and the prompted topics with which they are most closely associated. For example, the second ranked topic in both tables is associated with the Fisher "Minimum Wage" topic. The overlap of the best associated latent topic with the "Minimum Wage" topic is .825 in the unigram-based model but the overlap increases to .865 for the phrase-based model. Similar overlap improvements are generally observed across the range of other latent topics.

Next, we can observe that the topic summaries generated from the phrase-based PLSA model provide a more coherent description of various topics that those generated using the unigram-based PLSA model. This is evident in the "Minimum Wage" topic where the phrase-based model includes the signature phrases *minimum wage* and *five fifteen an hour* (which was the national minimum wage at the time of the data collection). By contrast, the unigram model summary for the same topic contains the individual signature words *minimum*, *wage*, *hour* and *fifteen* but the coherence of these is less evident when not used in the sequential context of their parent phrases.

Perhaps the added value of phrases is most evident in the "Time Travel" topic. Conversations on this topic were initiated by the prompt: "If each of you had the opportunity to go back in time and change something that you had done, what would it be and why?" The unigram-based system has trouble discovering this topic. However the phrase-based system is able to discover this topic by latching onto signature phrases such as *back in time and change* and *time travel*. These phrases are very distinctive to the topic, whereas as the underlying unigrams *back*, *time*, *change* and *travel* are generic terms that are relatively frequent and don't strongly signify any particular topic by themselves.

### 4.5. Quantitative Topical Evaluation

The primary concern of our work is to determine if latent topic modeling can be improved by adding phrases into a model's feature set. To measure this, a method for quantitatively assessing a latent topic model is required. Most corpora are not annotated with accurate topical information, making comparison between learned latent topics and reference topics impossible. However, because each of the conversations in the Fisher Corpus is initiated with a prompted topic, we can assess the agreement between the latent topics ascribed to documents with their known prompted topics.

One metric for comparing a latent topic labeling against a set of reference topic labels is the *erroneous information ratio* (EIR) [13], which is defined as:

$$EIR(Z,T) = \frac{H(Z|T) + H(T|Z)}{H(T)} \qquad (11)$$

**Table 4**. An abridged version of the automatic summary of the Fisher Corpus PLSA model learned using only word unigram features.

| Topic Rank | Top 10 Ranked Signature Words | Matching Fisher Topics ($P(t\|z)$) |
|---|---|---|
| 1 | dog, cat, pets, animals, fish, bird, puppy, feed, yard, cute | Pets (.880) |
| 2 | minimum, wage, hour, fifteen, jobs, raise, cost, paying, higher, fifty | Minimum Wage (.831) |
| 3 | sports, football, basketball, team, baseball, game, watching, hockey, play, fan | Sports on TV (.787) |
| 4 | reality, show, watched, survivor, millionaire, joe, bachelor, idol, factor, fear | Reality TV (.824) |
| 5 | security, airport, plane, fly, flight, shoes, airplane, flew, flown, check | Airport Security (.694)<br>September $11^{th}$ (.128) |
| : | : | : |
| 20 | married, marriage, divorced, church, young, wife, regret, parents, changed, wedding | Time Travel (.296)<br>Life Partners (.255) |
| : | : | : |
| 40 | drugs, test, truck, driver, company, privacy, random, driving, marijuana, urine | Drug Testing (.469) |

**Table 5**. An abridged version of the automatically generated summary of the Fisher Corpus PLSA model learned when the full set of 20,122 of learned phrases are incorporated in the feature set.

| Topic Rank | Top 10 Ranked Signature Words and Phrases | Matching Fisher Topics ($P(t\|z)$) |
|---|---|---|
| 1 | dog, cat, pets, animals, fish, bird, feed, puppy, cute, cage | Pets (.900) |
| 2 | minimum wage, pay, jobs, five fifteen an hour, paid, making, tips, cost of living, higher, low | Minimum Wage (.864) |
| 3 | sports, football, basketball, baseball, game, team, watching, hockey, t.v., soccer | Sports on TV (.849) |
| 4 | airport security, plane, fly, september eleventh, flight, airplane, flown, shoes, travel, terrorists | Airport Security (.523)<br>September $11^{th}$ (.351) |
| 5 | show, watched, survivor, reality t.v., reality shows, bachelor, joe millionaire, real world, fear factor, american idol | Reality TV (.901) |
| : | : | : |
| 23 | back and change, back in time and change, regret, time travel, future, past, degree, would've, differently, high school | Time Travel (.767) |
| : | : | : |
| 27 | married, marriage, divorced, church, wife, wedding, parents, christian, family values, met | Life Partners (.404)<br>Time Travel (.133) |
| : | : | : |
| 39 | drugs, drug test, company, hair, marijuana, invasion of privacy, urine, illegal, hired, alcohol | Drug Testing (.516) |
| 40 | linguistics, o.k., study, phone number, u. penn, speech recognition, boston, nice talking, program | Life Partners (.148) |

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 estimated from the joint distribution $P(z, t|d)$ estimated over all documents $d \in D$. The EIR measure compares the sum of the erroneous information captured by $H(Z|T)$ and $H(T|Z)$ against the total information $H(T)$ of the reference labels, with smaller EIR values representing greater similarity between the latent topics and the reference topics.

Another commonly used metric is *normalized mutual information* (NMI) which is expressed as:

$$NMI(Z, T) = \frac{2 * I(Z; T)}{H(Z) + H(T)} \qquad (12)$$

The NMI measure normalizes the mutual information $I(Z; T)$ by the average of $H(Z)$ and $H(T)$ such that perfect correlation between $Z$ and $T$ will yield the maximal NMI value of 1, with smaller values in NMI representing decreasing similarity between the latent topics and the reference topics.

Figure 2 shows the EIR and NMI scores for the PLSA model as the number of multi-word phrases used in modeling is increased from 0 to the full set of 20,122 learned phrases. When incorporating all learned phrases into the model the EIR is reduced from 1.009 to 0.801, which corresponds to a 21% relative decrease in erroneous information between $Z$ and $T$. Similarly the NMI measure is increased from 0.501 to 0.604 when the phrases are used, corresponding to a 21% relative increase in mutual information between $Z$ and $T$.

It is also important to note that the sizeable improvements obtained in Figure 2 using constrained phrase trees result from only a doubling of the feature space dimensionality. This is a relatively small increase in the feature set relative to the potential polynomial increase in features that can result from unconstrained $n$-gram modeling techniques.

For further comparative study, we examined the performance of David Blei's variational EM implementation of LDA [14] used in conjunction with our phrasal features. Table 6 compares our PLSA implementation against Blei's LDA system using both unigram features and our new phrase features. Because Blei's LDA system employs random initialization, we report the median result from 25 random training trials. The LDA system sees similar improvement in EIR and NMI to the PLSA system when our new phrase features are

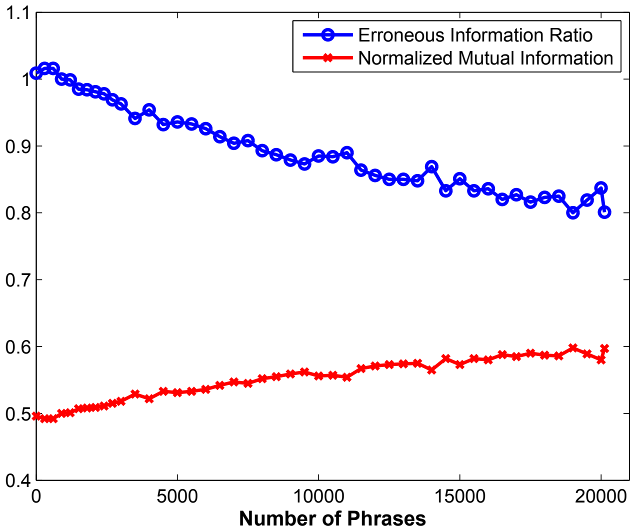

**Fig. 2**. PLSA modeling results on the Fisher Corpus using constrained phrase trees as the number learned phrases is increased from zero up the full set of 20,122 phrases.

**Table 6**. A comparison of unsupervised topic discovery results on the Fisher Corpus using unigram and phrase features within topic models trained using our own PLSA implementation and Blei's open source LDA implementation.

| Model | EIR | NMI |
|---|---|---|
| PLSA w/ Unigrams | 1.009 | 0.504 |
| PLSA w/ Phrases | 0.801 | 0.604 |
| LDA w/ Unigrams | 0.955 | 0.528 |
| LDA w/ Phrases | **0.779** | **0.615** |

used, confirming that our phrase learning approach is general and should be effective for use within other latent topic model implementations employing a bag-of-words assumption.

Our best results for unsupervised topic discovery resulted from the use of our new phrase features within an LDA topic model trained using variational EM. These results are highlighted in bold in Table 6. Though the LDA implementation outperformed PLSA, we should note that this improvement did come at the cost of significantly higher computational requirements for variational EM LDA topic modeling training.

We also performed preliminary evaluations of open source packages to perform Collocation LDA [15] and Topical $N$-gram Modeling [16] on the same Fisher data set. Unfortunately, our initial results using these packages were significantly worse than all results reported in Table 6. However, further study is needed to understand the reasons for the performance deficiencies observed in these preliminary evaluations before we will be able to make definitive claims about our new approach relative to these prior approaches.

## 5. SUMMARY

In this paper we have introduced a new method for learning phrases from text or speech data based on constrained phrase trees. This method uses weighted pointwise mutual information statistics to guide an iterative learning process that generates a sequential collection of automatically learned rules that can deterministically parse a sequence of words into a sequence of informative phrases. By applying latent topic modeling to the output of our phrase discovery algorithm, we have shown significant improvement in the unsupervised discovery of topics in the Fisher Corpus. This improvement is observed using both the PLSA and the variational EM LDA topic modeling approaches thus demonstrating the flexibility of our new phrase discovery algorithm.

## 6. REFERENCES

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

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

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

[4] D. Blei and J. Lafferty, "Visualizing topics with multi-word expressions", Technical Report in arXiv.org, arXiv:0907.1013, July 2009.

[5] B.-J. Hsu and J. Glass, "Style & topic language model adaptation using HMM-LDA" in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Sydney, July 2006.

[6] J. Boyd-Graber, D. Blei, and X. Zhu, "A topic model for word sense disambiguation", in Proc. Empirical Methods in Natural Language Processing (EMNLP), Prague, June 2007.

[7] D. Ramage, S. Dumais and D. Liebling, "Characterizing microblogs with topic models", in *Proc. AAAI Conf. on Weblogs and Social Media*, Washington DC, May 2010.

[8] H. Wallach, "Topic modeling: Beyond bag-of-words", in *Proc. International Conf. Machine Learning*, Pittsburgh, PA, June 2006.

[9] T. Griffiths, M. Steyvers and J. Tenenbaum, "Topics in semantic representation", *Psychological Review*, vol. 114, no. 2, pp. 211-244, April 2007.

[10] X. Wang, A. McCallum and X. Wei, "Topical N-grams: Phrase and topic discovery, with an application to information retrieval", in *Proc. International Conf. on Data Mining (ICDM)*, Omaha, NE, Oct. 2007.

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

[12] T. Hazen, "Latent topic modeling for audio corpus summarization", in *Proc. of Interspeech*, Florence, Italy, August 2011.

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

[14] D. Blei, *Latent Dirichlet Allocation in C*, available at: `http://www.cs.princeton.edu/~blei/lda-c/`

[15] D. Steyvers and T. Griffiths, *Matlab Topic Modeling Toolbox*, available at: `http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm/`

[16] A. McCallum *et al*, *MALLET: Machine Learning for Language Toolkit*, available at: `http://mallet.cs.umass.edu/`