Experts' Retrieval with Multiword-Enhanced Author Topic Model

Nikhil Johri Dan Roth Yuancheng Tu
Dept. of Computer Science Dept. of Linguistics
University of Illinois at Urbana-Champaign
{njohri2,danr,ytu}@illinois.edu

Abstract

In this paper, we propose a multiword-enhanced author topic model that clusters authors with similar interests and expertise, and apply it to an information retrieval system that returns a ranked list of authors related to a keyword. For example, we can retrieve *Eugene Charniak* via search for *statistical parsing*.

The existing works on author topic modeling assume a "bag-of-words" representation. However, many semantic atomic concepts are represented by multiwords in text documents. This paper presents a pre-computation step as a way to discover these multiwords in the corpus automatically and tags them in the term-document matrix. The key advantage of this method is that it retains the simplicity and the computational efficiency of the unigram model. In addition to a qualitative evaluation, we evaluate the results by using the topic models as a component in a search engine. We exhibit improved retrieval scores when the documents are represented via sets of latent topics and authors.

1 Introduction

This paper addresses the problem of searching people with similar interests and expertise without inputting personal names as queries. Many existing people search engines need people's names to do a "keyword" style search, using a person's name as a query. However, in many situations, such information is impossible to know beforehand. Imagine a scenario where the statistics department of a university invited a world-wide known expert in Bayesian statistics and machine learning to give a keynote speech; how can the department head notify all the people on campus who are interested without spamming those who are not? Our paper proposes a solution to the aforementioned scenario by providing a search engine which goes beyond "keyword" search and can retrieve such information semantically. The department head would only need to input the domain keyword of the keynote speaker, i.e. *Bayesian statistics, machine learning*, and all professors and students who are interested in this topic will be retrieved. Specifically, we propose a **M**ultiword-enhanced **A**uthor-**T**opic **M**odel (MATM), a probabilistic generative model which assumes two steps of generation process when producing a document.

Statistical topical modeling (Blei and Lafferty, 2009a) has attracted much attention recently due to its broad applications in machine learning, text mining and information retrieval. In these models, semantic topics are represented by multinomial distribution over words. Typically, the content of each topic is visualized by simply listing the words in order of decreasing probability and the "meaning" of each topic is reflected by the top 10 to 20 words in that list. The Author-Topic Model (ATM) (Steyvers et al., 2004; Rosen-Zvi et al., 2004) extends the basic topical models to include author information in which topics and authors are modeled jointly. Each author is a multinomial distribution over topics and each topic is a multinomial distribution over words.

Our contribution to this paper is two-fold. First of all, our model, MATM, extends the original ATM by adding semantically coherent multiwords into the term-document matrix to relax the model's "bag-of-

Proceedings of the NAACL HLT 2010 Workshop on Semantic Search, pages 10–18,
Los Angeles, California, June 2010.

words" assumption. Each multiword is discovered via statistical measurement and filtered by its part of speech pattern via an off-line way. One key advantage of tagging these semantic atomic units off-line, is the retention of the flexibility and computational efficiency in using the simpler word exchangeable model, while providing better interpretation of the topics author distribution.

Secondly, to the best of our knowledge, this is the first proposal to apply the enhanced author topic modeling in a semantic retrieval scenario, where searching people is associated with a set of hidden semantically meaningful topics instead of their names. While current search engines cannot support interactive and exploratory search effectively, search based on our model serves very well to answer a range of exploratory queries about the document collections by semantically linking the interests of the authors to the topics of the collection, and ultimately to the distribution of the words in the documents.

The rest of the paper is organized as follows. We present some related work on topic modeling, the original author-topic model and automatic phrase discovery methods in Sec. 2. Then our model is described in Sec. 3. Sec. 4 presents our experiments and the evaluation of our method on expert search. We conclude this paper in Sec. 5 with some discussion and several further developments.

2 Related Work

Author topic modeling, originally proposed in (Steyvers et al., 2004; Rosen-Zvi et al., 2004), is an extension of another popular topic model, Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a probabilistic generative model that can be used to estimate the properties of multinomial observations via unsupervised learning. LDA represents each document as a mixture of probabilistic topics and each topic as a multinomial distribution over words. The Author topic model adds an author layer over LDA and assumes that the topic proportion of a given document is generated by the chosen author.

Both LDA and the author topic model assume *bag-of-words* representation. As shown by many previous works (Blei et al., 2003; Steyvers et al., 2004), even such unrealistic assumption can actually lead to a reasonable topic distribution with relatively simple and computationally efficient inference algorithm. However, this unigram representation also poses major handicap when interpreting and applying the hidden topic distributions. The proposed MATM is an effort to try to leverage this problem in author topic modeling. There have been some works on Ngram topic modeling over the original LDA model (Wallach, 2006; Wang and McCallum, 2005; Wang et al., 2007; Griffiths et al., 2007). However, to the best of our knowledge, this paper is the first to embed multiword expressions into the author topic model.

Many of these Ngram topic models (Wang and McCallum, 2005; Wang et al., 2007; Griffiths et al., 2007) improves the base model by adding a new indicator variable x_i to signify if a bigram should be generated. If $x_i = 1$, the word w_i is generated from a distribution that depends only on the previous word to form an Ngram. Otherwise, it is generated from a distribution only on the topic proportion (Griffiths et al., 2007) or both the previous words and the latent topic (Wang and McCallum, 2005; Wang et al., 2007). However, these complex models not only increase the parameter size to $\mathcal{V}$ times larger than the size of the original LDA model parameters ($\mathcal{V}$ is the size of the vocabulary of the document collection) [1], it also faces the problem of choosing which word to be the topic of the potential Ngram. In many text retrieval tasks, the humongous size of data may prevent us using such complicated computation on-line. However, our model retains the computational efficiency by adding a simple tagging process via pre-computation.

Another effort in the current literature to interpret the meaning of the topics is to label the topics via a post-processing way (Mei et al., 2007; Blei and Lafferty, 2009b; Magatti et al., 2009). For example, Probabilistic topic labeling (Mei et al., 2007) first extracts a set of candidate label phrases from a reference collection and represents each candidate labeling phrase with a multinomial distribution of words. Then KL divergence is used to rank the most probable labels for a given topic. This method needs not only extra reference text collection, but also facing

[1] LDA collocation models and topic Ngram models also have parameters for the binomial distribution of the indicator variable x_i for each word in the vocabulary.

the problem of finding discriminative and high coverage candidate labels. Blei and Lafferty (Blei and Lafferty, 2009b) proposed a method to annotate each word of the corpus by its posterior word topic distribution and then cast a statistical co-occurrence analysis to extract the most significant Ngrams for each topic and visualize the topic with these Ngrams. However, they only applied their method to basic LDA model.

In this paper, we applied our multiword extension to the author topic modeling and no extra reference corpora are needed. The MATM, with an extra precomputing step to add meaningful multiwords into the term-document matrix, enables us to retain the flexibility and computational efficiency to use the simpler word exchangeable model, while providing better interpretation of the topics and author distribution.

3 Multiword-enhanced Author-Topic Model

The MATM is an extension of the original ATM (Rosen-Zvi et al., 2004; Steyvers et al., 2004) by semantically tagging collocations or multiword expressions, which represent atomic concepts in documents in the term-document matrix of the model. Such tagging procedure enables us to retain computational efficiency of the word-level exchangeability of the orginal ATM while provides more sensible topic distributions and better author topic coherence. The details of our model are presented in Algorithm 1.

3.1 Beyond Bag-of-Words Tagging

The first *for* loop in Algorithm 1 is the procedure of our multiword tagging. Commonly used ngrams, or statistically short phrases in text retrieval, or so-called collocations in natural language processing have long been studied by linguistics in various ways. Traditional collocation discovery methods range from frequency to mean and variance, from statistical hypothesis testing, to mutual information (Manning and Schtze, 1999). In this paper, we use a simple statistical hypothesis testing method, namely Pearson's chi-square test implemented in Ngram Statistic Package (Banerjee and Pedersen, 2003), enhanced by passing the candidate phrases through some pre-defined part of speech patterns that are likely to be true phrases. This very simple heuristic has been shown to improve the counting based methods significantly (Justenson and Katz, 1995).

The χ^2 test is chosen since it does not assume any normally distributed probabilities and the essence of this test is to compare the observed frequencies with the frequencies expected for independence. We choose this simple statistic method since in many text retrieval tasks the volume of data we see always makes it impractical to use very sophisticated statistical computations. We also focus on nominal phrases, such as bigram and trigram noun phrases since they are most likely to function as semantic atomic unit to directly represent the concepts in text documents.

3.2 Author Topic Modeling

The last three generative procedures described in Algorithm 1 jointly model the author and topic information. This generative model is adapted directly from (Steyvers et al., 2004). Graphically, it can be visualized as shown in Figure 1.

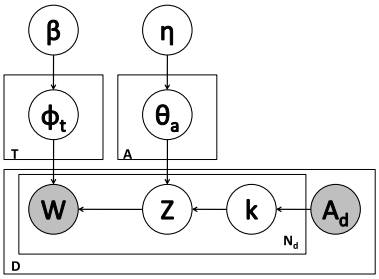

Figure 1: Plate notation of our model: MATM

The four plates in Fiture 1 represent topic (T), author (A), document (D) and Words in each document (N_d) respectively. Each author is associated with a multinomial distribution over all topics, $\vec{\theta_a}$ and each topic is a multinomial distribution over all words, $\vec{\phi_t}$. Each of these distribution has a symmetric Dirichlet prior over it, $\vec{\eta}$ and $\vec{\beta}$ respectively. When generating a document, an author *k* is first chosen according to a uniform distribution. Then this author chooses the topic from his/her associated multinomial distribution over topics and then generates a word from the multinomial distribution of that topic over the

words.

Algorithm 1: MATM: $\mathcal{A}, \mathcal{T}, \mathcal{D}, \mathcal{N}$ are four plates as shown in Fig. 1. The first *for* loop is the off-line process of multiword expressions. The rest of the algorithm is the generative process of the author topic modeling.

Data: $\mathcal{A}, \mathcal{T}, \mathcal{D}, \mathcal{N}$
for *all documents* $d \in \mathcal{D}$ **do**
  Part-of-Speech tagging ;
  Bigram extraction ;
  Part-of Speech Pattern Filtering ;
  Add discovered bigrams into $\mathcal{N}$;
for *each author* $a \in \mathcal{A}$ **do**
  draw a distribution over topics:
  $\vec{\theta_a} \sim Dir_{\mathcal{T}}(\vec{\eta})$;
for *each topic* $t \in \mathcal{T}$ **do**
  draw a distribution over words:
  $\vec{\phi_t} \sim Dir_{\mathcal{N}}(\vec{\beta})$;
for *each document* $d \in \mathcal{D}$ *and* k *authors* $\in d$ **do**
  for *each word* $w \in d$ **do**
    choose an author $k \sim$ uniformly;
    draw a topic assignment i given the author: $z_{k,i}|k \sim Multinomial(\theta_a)$;
    draw a word from the chosen topic: $w_{d,k,i}|z_{k,i} \sim Multinomial(\phi_{z_{k,i}})$;

MATM includes two sets of parameters. The $\mathcal{T}$ topic distribution over words, ϕ_t which is similar to that in LDA. However, instead of a document-topic distribution, author topic modeling has the author-topic distribution, θ_a. Using a matrix factorization interpretation, similar to what Steyvers, Griffiths and Hofmann have pointed out for LDA (Steyvers and Griffiths, 2007) and PLSI (Hofmann, 1999), a word-author co-occurrence matrix in author topic model can be split into two parts: a word-topic matrix ϕ and a topic-author matrix θ. And the hidden topic serves as the low dimensional representation for the content of the document.

Although the MATM is a relatively simple model, finding its posterior distribution over these hidden variables is still intractable. Many efficient approximate inference algorithms have been used to solve this problem including Gibbs sampling (Griffiths and Steyvers, 2004; Steyvers and Griffiths, 2007; Griffiths et al., 2007) and mean-field variational methods (Blei et al., 2003). Gibbs sampling is a special case of Markov-Chain Monte Carlo (MCMC) sampling and often yields relatively simple algorithms for approximate inference in high dimensional models.

In our MATM, we use a collapsed Gibbs sampler for our parameter estimation. In this Gibbs sampler, we integrated out the hidden variables θ and ϕ as shown by the delta function in equation 2. This Dirichlet delta function with a M dimentional symmetric Dirichlet prior is defined in Equation 1. For the current state j, the conditional probability of drawing the k^{th} author K_j^k and the i^{th} topic Z_j^i pair, given all the hyperparameters and all the obeserved documents and authors except the current assignment (the exception is denoted by the symbol $\neg j$), is defined in Equation 2.

$$\Delta_M(\lambda) = \frac{\Gamma\left(\lambda^M\right)}{\Gamma(M\lambda)} \tag{1}$$

$$\begin{aligned} & P(Z_j^i, K_j^k | W_j = w, Z_{\neg j}, K_{\neg j}, W_{\neg j}, A_d, \vec{\beta}, \vec{\eta}) \\ \propto \quad & \frac{\Delta(n_Z + \vec{\beta})}{\Delta(n_{Z,\neg j} + \vec{\beta})} \frac{\Delta(n_K + \vec{\eta})}{\Delta(n_{K,\neg j} + \vec{\eta})} \\ = \quad & \frac{n^w_{i,\neg j} + \vec{\beta_w}}{\sum_{w=1}^V n^w_{i,\neg j} + V\vec{\beta_w}} \frac{n^i_{k,\neg j} + \vec{\eta_i}}{\sum_{i=1}^T n^i_{k,\neg j} + T\vec{\eta_i}} \end{aligned} \tag{2}$$

And the parameter sets ϕ and θ can be interpreted as sufficient statistics on the state variables of the Markov Chain due to the Dirichlet conjugate priors we used for the multinomial distributions. The two formulars are shown in Equation 3 and Equation 4 in which n_i^w is defined as the number of times that the word w is generated by topic i and n_k^i is defined as the number of times that topic i is generated by author k. The Gibbs sampler used in our experiments is from the Matlab Topic Modeling Toolbox [2].

$$\phi_{w,i} = \frac{n_i^w + \vec{\beta_w}}{\sum_{w=1}^V n_i^w + V\vec{\beta_w}} \tag{3}$$

$$\theta_{k,i} = \frac{n_k^i + \vec{\eta_i}}{\sum_{i=1}^T n_k^i + T\vec{\eta_i}} \tag{4}$$

[2] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

4 Experiments and Analysis

In this section, we describe the empirical evaluation of our model qualitatively and quantitatively by applying our model to a text retrieval system we call *Expert Search*. This search engine is intended to retrieve groups of experts with similar interests and expertise by inputting only general domain key words, such as *syntactic parsing, information retrieval*.

We first describe the data set, the retrieval system and the evaluation metrics. Then we present the empirical results both qualitatively and quantitatively.

4.1 Data

We crawled from ACL anthology website and collected seven years of annual ACL conference papers as our corpus. The reference section is deleted from each paper to reduce some noisy vocabulary, such as idiosyncratic proper names, and some coding errors caused during the file format conversion process. We applied a part of speech tagger[3] to tag the files and retain in our vocabulary only content words, i.e., nouns, verbs, adjectives and adverbs.

The ACL anthology website explicitly lists each paper together with its title and author information. Therefore, the author information of each paper can be obtained accurately without extracting from the original paper. We transformed all pdf files to text files and normalized all author names by eliminating their middle name initials if they are present in the listed names. There is a total of 1,326 papers in the collected corpus with $2,084$ authors. Then multiwords (in our current experiments, the bigram collocations) are discovered via the χ^2 statistics and part of speech pattern filtering. These multiwords are then added into the vocabulary to build our model. Some basic statistics about this corpus is summarized in Table 1.

Two sets of results are evaluated use the retrieval system in our experiments: one set is based on unigram vocabulary and the other with the vocabulary expanded by the multiwords.

4.2 Evaluation on Expert Search

We designed a preliminary retrieval system to evaluate our model. The functionality of this search is to associate words with individual authors, i.e., we rank the joint probability of the query words and the target author $P(W, a)$. This probability is marginalized over all topics in the model to rank all authors in our corpus. In addition, the model assumes that the word and the author is conditionally independent given the topic. Formally, we define the ranking function of our retrieval system in Equation 5:

ACL Corpus Statistics	
Year range	2003-2009
Total number of papers	1,326
Total number of authors	2,084
Total unigrams	34,012
Total unigram and multiwords	205,260

Table 1: Description of the ACL seven-year collection in our experiments

$$P(W, a) = \sum_{w_i} \alpha_i \sum_t P(w_i, a|t)P(t)$$
$$= \sum_{w_i} \alpha_i \sum_t P(w_i|t)P(a|t)P(t) \quad (5)$$

W is the input query, which may contain one or more words. If a multiword is detected within the query, it is added into the query. The final score is the sum of all words in this query weighted by their inverse document frequency α_i The inverse document frequency is defined as Equation 6.

$$\alpha_i = \frac{1}{DF(w_i)} \quad (6)$$

In our experiments, we chose ten queries which covers several most popular research areas in computational linguistics and natural language processing. In our unigram model, query words are treated token by token. However, in our multiword model, if the query contains a multiword inside our vocabulary, it is treated as an additional token to expand the query. For each query, top 10 authors are returned from the system. We manually label the relevance of these 10 authors based on the papers they submitted to these seven-year ACL conferences collected in our corpus. Two evaluation metrics are used to measure the precision of the retrieving results. First we evaluate the precision at a given cut-off rank, namely precision at K with K ranging from 1 to 10.

[3]The tagger is from:
http://l2r.cs.uiuc.edu/∼cogcomp/software.php

We also calculate the average precision (AP) for each query and the mean average precision (MAP) for all the 10 queries. Average precision not only takes ranking as consideration but also emphasizes ranking relevant documents higher. Different from precision at K, it is sensitive to the ranking and captures some recall information since it assumes the precision of the non-retrieved documents to be zero. It is defined as the average of precisions computed at the point of each of the relevant documents in the ranked list as shown in equation 7.

$$AP = \frac{\sum_{r=1}^{n}(Precision(r) \times rel(r))}{\sum_{relevant\,documents}} \quad (7)$$

Currently in our experiments, we do not have a pool of labeled authors to do a good evaluation of recall of our system. However, as in the web browsing activity, many users only care about the first several hits of the retrieving results and precision at K and MAP measurements are robust measurements for this purpose.

4.3 Results and Analysis

In this section, we first examine the qualitative results from our model and then report the evaluation on the external expert search.

4.3.1 Qualitative Coherence Analysis

As have shown by other works on Ngram topic modeling (Wallach, 2006; Wang et al., 2007; Griffiths et al., 2007), our model also demonstrated that embedding multiword tokens into the simple author topic model can always achieve more coherent and better interpretable topics. We list top 15 words from two topics of the multiword model and unigram model respectively in Table 2. Unigram topics contain more general words which can occur in every topic and are usually less discriminative among topics.

Our experiments also show that embedding the multiword tokens into the model achieves better clustering of the authors and the coherence between authors and topics. We demonstrate this qualitatively by listing two examples respectively from the multiword models and the unigram model in Table 3. For example, for the topic on dependency parsing, unigram model missed *Ryan-McDonald* and the ranking of the authors are also questionable. Further quantitative measurement is listed in our quantitative evaluation section. However, qualitatively, multiword model seems less problematic.

MultiWord Model	**Unigram Model**
TOPIC 4	Topic 51
coreference-resolution	resolution
antecedent	antecedent
treesubstitution-grammars	pronoun
completely	pronouns
pronoun	is
resolution	information
angry	antecedents
candidate	anaphor
extracted	syntactic
feature	semantic
pronouns	coreference
model	anaphora
perceptual-cooccurrence	definite
certain-time	model
anaphora-resolution	only
TOPIC 49	Topic 95
sense	sense
senses	senses
word-sense	disambiguation
target-word	word
word-senses	context
sense-disambiguation	ontext
nouns	ambiguous
automatically	accuracy
semantic-relatedness	nouns
disambiguation	unsupervised
provided	target
ambiguous-word	predominant
concepts	sample
lexical-sample	automatically
nouns-verbs	meaning

Table 2: Comparison of the topic interpretation from the multiword-enhanced and the unigram models. Qualitatively, topics with multiwords are more interpretable.

Some of the unfamiliar author may not be easy to make a relevance judgment. However, if we trace all the papers the author wrote in our collected corpus, many of the authors are coherently related to the topic. We list all the papers in our corpus for three authors from the machine translation topic derived from the multiword model in Table 4 to demonstrate the coherence between the author and the related topic. However, it is also obvious that our model missed some *real* experts in the corresponding field.

MultiWord Model		Unigram Model	
Topic 63	Topic 145	Topic 23	Topic 78
Word	**Word**	**Word**	**Word**
translation	dependency-parsing	translation	dependency
machine-translation	dependency-tree	translations	head
language-model	dependency-trees	bilingual	dependencies
statistical-machine	dependency	pairs	structure
translations	dependency-structures	language	structures
phrases	dependency-graph	machine	dependent
translation-model	dependency-relation	parallel	order
decoding	dependency-relations	translated	word
score	order	monolingual	left
decoder	does	quality	does
Author	**Author**	**Author**	**Author**
Shouxun-Lin	Joakim-Nivre	Hua-Wu	Christopher-Manning
David-Chiang	Jens-Nilsson	Philipp-Koehn	Hisami-Suzuk
Qun-Liu	David-Temperley	Ming-Zhou	Kenji-Sagae
Philipp-Koehn	Wei-He	Shouxun-Lin	Jens-Nilsson
Chi-Ho-Li	Elijah-Mayfield	David-Chiang	Jinxi-Xu
Christoph-Tillmann	Valentin-Jijkoun	Yajuan-Lu	Joakim-Nivre
Chris-Dyer	Christopher-Manning	Haifeng-Wang	Valentin-Jijkoun
G-Haffari	Jiri-Havelka	Aiti-Aw	Elijah-Mayfield
Taro-Watanabe	Ryan-McDonald	Chris-Callison-Burch	David-Temperley
Aiti-Aw	Andre-Martins	Franz-Och	Julia-Hockenmaier

Table 3: Two examples for topic and author coherece from multiword-enhanced model and unigram model. Top 10 words and authors are listed accordingly for each model.

For example, we did not get *Kevin Knight* for the *machine translation* topic. This may be due to the limitation of our corpus since we only collected papers from one conference in a limited time, or because usually these *experts* write more divergent on various topics.

Another observation in our experiment is that some experts with many papers may not be ranked at the very top by our system. However, they have pretty high probability to associate with several topics. Intuitively this makes sense, since many of these famous experts write papers with their students in various topics. Their scores may therefore not be as high as authors who have fewer papers in the corpus which are concentrated in one topic.

4.3.2 Results from Expert Search

One annotator labeled the relevance of the retrieval results from our expert search system. The annotator was also given all the paper titles of each corresponding retrieved author to help make the binary judgment. We experimented with ten queries and retrieved the top ten authors for each query.

We first used the precision at K for evaluation. we calculate the precision at K for both of our multiword model and the unigram model and the results are listed in Table 5. It is obvious that at every rank position, the multiword model works better than the unigram model. In order to focus more on relevant retrieval results, we then calculate the average precision for each query and mean average precision for both models. The results are in Table 6.

When only comparing the mean average precision (MAP), the multiword model works better. However, when examining the average precision of each query within these two models, the unigram model also works pretty well with some queries. How the query words may interact with our model deserves further investigation.

5 Discussion and Further Development

In this paper, we extended the existing author topic model with multiword term-document input and applied it to the domain of expert retrieval. Although our study is preliminary, our experiments do return

Author	Papers from ACL(03-09)
Shouxun-Lin	Log-linear Models for Word Alignment Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation Tree-to-String Alignment Template for Statistical Machine Translation Forest-to-String Statistical Translation Rules Partial Matching Strategy for Phrase-based Statistical Machine Translation
David-Chiang	A Hierarchical Phrase-Based Model for Statistical Machine Translation Word Sense Disambiguation Improves Statistical Machine Translation Forest Rescoring: Faster Decoding with Integrated Language Models Fast Consensus Decoding over Translation Forests
Philipp-Koehn	Feature-Rich Statistical Translation of Noun Phrases Clause Restructuring for Statistical Machine Translation Moses: Open Source Toolkit for Statistical Machine Translation Enriching Morphologically Poor Languages for Statistical Machine Translation A Web-Based Interactive Computer Aided Translation Tool Topics in Statistical Machine Translation

Table 4: Papers in our ACL corpus for three authors related to the "machine translation" topic in Table 3.

Precision@K		
K	Multiword Model	Unigram Model
1	0.90	0.80
2	0.80	0.80
3	0.73	0.67
4	0.70	0.65
5	0.70	0.64
6	0.72	0.65
7	0.71	0.64
8	0.71	0.66
9	0.71	0.66
10	0.70	0.64

Table 5: Precision at K evaluation of the multiword-enhanced model and the unigram model.

Average Precision (AP)		
Query	Multi. Mod.	Uni. Mod.
Language Model	0.79	0.58
Unsupervised Learning	1.0	0.78
Supervised Learning	0.84	0.74
Machine Translation	0.95	1.0
Semantic Role Labeling	0.81	0.57
Coreference Resolution	0.59	0.72
Hidden Markov Model	0.93	0.37
Dependency Parsing	0.75	0.94
Parsing	0.81	0.98
Transliteration	0.62	0.85
MAP:	**0.81**	0.75

Table 6: Average Precision (AP) for each query and Mean Average Precision (MAP) of the multiword-enhanced model and the unigram model.

promising results, demonstrating the effectiveness of our model in improving coherence in topic clusters. In addition, the use of the MATM for expert retrieval returned some useful preliminary results, which can be further improved in a number of ways.

One immediate improvement would be an extension of our corpus. In our experiments, we considered only ACL papers from the last 7 years. If we extend our data to cover papers from additional conferences, we will be able to strengthen author-topic associations for authors who submit papers on the same topics to different conferences. This will also allow more prominent authors to come to the forefront in our search application. Such a modification would require us to further increase the model's computational efficiency to handle huge volumes of data encountered in real retrieval systems.

Another further development of this paper is the addition of citation information to the model as a layer of supervision for the retrieval system. For instance, an author who is cited frequently could have a higher weight in our system than one who isn't, and could occur more prominently in query results.

Finally, we can provide a better evaluation of our system through a measure of recall and a simple baseline system founded on keyword search of paper titles. Recall can be computed via comparison to a set of expected prominent authors for each query.

Acknowledgments

The research in this paper was supported by the Multimodal Information Access & Synthesis Center at UIUC, part of CCICADA, a DHS Science and Technology Center of Excellence.

References

S. Banerjee and T. Pedersen. 2003. The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381.

D. Blei and J. Lafferty. 2009a. Topic models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Francis.

D. Blei and J. Lafferty. 2009b. Visualizing topics with multi-word expressions. In *http://arxiv.org/abs/0907.1013*.

D. Blei, A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*.

T. Griffiths and M. Steyvers. 2004. Finding scientific topic. In *Proceedings of the National Academy of Science*.

T. Griffiths, M. Steyvers, and J. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*.

T. Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of SIGIR*.

J. Justenson and S. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for indentification in text. *Natural Language Engineering*.

D. Magatti, S. Calegari, D. Ciucci, and F. Stella. 2009. Automatic labeling of topics. In *ISDA*, pages 1227–1232.

Christopher D. Manning and Hinrich Schtze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts.

Q. Mei, X. Shen, and C. Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. 2004. the author-topic model for authors and documents. In *Proceedings of UAI*.

M. Steyvers and T. Griffiths. 2007. Probabilistic topic models. In *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates.

M. Steyvers, P. Smyth, and T. Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of KDD*.

H. Wallach. 2006. Topic modeling; beyond bag of words. In *International Conference on Machine Learning*.

X. Wang and A. McCallum. 2005. A note on topical n-grams. Technical report, University of Massachusetts.

X. Wang, A. McCallum, and X. Wei. 2007. Topical n-grams: Phrase and topic discoery with an application to information retrieval. In *Proceedings of ICDM*.