A Latent Dirichlet Framework for Relevance Modeling*

Viet Ha-Thuc and Padmini Srinivasan
Computer Science Department – The University of Iowa, Iowa City, IA 52242, US
hviet@cs.uiowa.edu and padmini-srinivasan@uiowa.edu

Abstract. Relevance-based language models operate by estimating the probabilities of observing words in documents relevant (or pseudo relevant) to a topic. However, these models assume that if a document is relevant to a topic, then all tokens in the document are relevant to that topic. This could limit model robustness and effectiveness. In this study, we propose a *Latent Dirichlet relevance model*, which relaxes this assumption. Our approach derives from current research on Latent Dirichlet Allocation (LDA) topic models. LDA has been extensively explored, especially for generating a set of topics from a corpus. A key attraction is that in LDA a document may be about several topics. LDA itself, however, has a limitation that is also addressed in our work. Topics generated by LDA from a corpus are synthetic, i.e., they do not necessarily correspond to topics identified by humans for the same corpus. In contrast, our model explicitly considers the relevance relationships between documents and given topics (queries). Thus unlike standard LDA, our model is directly applicable to goals such as relevance feedback for query modification and text classification, where topics (classes and queries) are provided upfront. Thus although the focus of our paper is on improving relevance-based language models, in effect our approach bridges relevance-based language models and LDA addressing limitations of both. Finally, we propose an idea that takes advantage of "bag-of-words" assumption to reduce the complexity of Gibbs sampling based learning algorithm.

Keywords: LDA, topic models, relevance-based language models.

1 Introduction

Relevance is a key concept in retrieval theory [7][14]. Among the formal relevance models that have been proposed, relevance-based language models is perhaps the most popular one [9][10][11]. Given a topic of interest t, relevance-based language models estimate the probability distribution $p(w| R_t)$ of observing a word w in documents relevant to topic t. The distribution is estimated by using a set of relevant (or pseudo relevant) documents as training data. Good results have been obtained with these approaches [9][10][11]. Nonetheless, these relevance-based language models have a limitation; they make an overly-strict assumption that *all* tokens in each training document are generated by a *single topic* to which the document belongs. This assumption is obviously not true in many practical cases. The example below is the first paragraph of a Wall Street Journal article judged relevant to the topic "machine translation" (TREC topic 63). As we see, many portions of it are non-relevant to the topic.

Buried among the many trade issues that bedevil the U.S. and Japan is the $1 billion of translation work done every year in Japan that could be done better and more efficiently in the U.S. And in the next two years, the dollar value of Japanese-to-English translations is expected to double. Think about it. Every car, videocassette recorder, boom box or stereo imported into the U.S. from Japan has operating and assembly instructions. And every dealer and repair shop gets parts catalogs and repair guides. In almost every instance, the translations have been done in Japan -- a fact often obvious upon reading

* In Proceedings of the 5th Asia Information Retrieval Symposium (AIRS' 09, Hokkaido, Japan) - Lecture Notes in Computer Science, Springer.

them. The Japanese have been slow to realize that it would be in everyone's best interest to have the translations done in the U.S.

Alongside the development of relevance-based models, we observe a strong strand of research on Latent Dirichlet Allocation (LDA) or probabilistic topic models, that has been shown to be effective in many text-related applications [2][5][6][15]. LDA offers a strong theoretical framework within which we may consider each document as generated by a *mixture of multiple topics*. However, LDA typically used to generate K topics from a corpus also has a limitation. Specifically, the K topics generated from a corpus are synthetic and do not explicitly correspond to the prior knowledge of human beings regarding topics in the corpus. In other words, if experts identified K topics manually for a corpus, then these may have little or no correspondence with the K synthetic topics identified by LDA. From a different perspective, we may say that probabilistic topic models are unable to model the concept of relevance to given topics of interest. Thus, not surprisingly LDA has not found use in applications such as relevance feedback based query modification. Our work shows how this can be done.

In this paper, we propose an approach that bridges relevance-based language models and LDA. Our approach allows us to address the limitation of relevance-based language models, specifically their assumption that *all* tokens of a relevant document are equally relevant to a topic. We do this by estimating the relevance model using the multiple-topic framework of LDA. In essence, we consider that although a document d may be relevant to a given topic t, it could still have non-relevant portions. Some portions could pertain to background information shared by many documents. Other non relevant portions while specific to d may be on themes other than t. Specifically, each document d is hypothesized to be generated by a combination of three topics: the topic t to which it is relevant, a background topic b representing the general language in the document set, and a third topic $t\_o(d)$ responsible for generating themes that though specific to d are neither b nor t. Because we consider this mixture of three topics, our model is able to identify just those portions of the document that are truly relevant to the topic t. In our work, these selected portions are the ones that contribute to the estimation of the relevance model $p(w| R_t)$. In this way, we utilize the Latent Dirichlet framework to solve for a limitation in relevance-based language models.

As in previous work in standard probabilistic topic models [5][6][15], we also implement the inference process using Gibbs sampling [1][3]. A secondary contribution of this paper is that, we exploit the "bag-of-words" assumption in order to reduce the computational complexity of the inference algorithm. Since token order in a document is not considered, we can re-arrange the tokens in any order that is convenient for the learning algorithm. In our case, we group tokens with the same stem into continuous segments because the topics of the tokens are sampled from the same distribution. That helps to reduce the running time of the sampling process. The proposed idea is also applicable for standard probabilistic topic models.

2 A Latent Dirichlet Relevance Model

2.1 Notation

A *Vocabulary set* (dictionary) V is a set of W possible *words* (*terms*) $V = \{word_1, word_2 \ldots word_W\}$. A *token* is a specific occurrence of one of the W words in a document. *Document* d is a sequence of N_d tokens. A *training set* D_t of a topic of interest t is a set of $|D_t|$ relevant (or

pseudo relevant) documents: $D_t = \{(w_1, d_1), (w_2, d_2) \ldots (w_{Nt}, d_{Nt})\}$, where $N_t = \sum_{d \in D_t} N_d$, w_i and d_i are word index and document index of the i^{th} token. A *topic* is a multinomial distribution over the vocabulary set.

As we mentioned above, each document d in the training set of a topic of interest t is generated by a mixture of three topics $\boldsymbol{x_d} = \{b, t, t\_o(d)\}$, where b denotes the background topic and $t\_o(d)$ denotes a document-leveled topic covering other themes rather t also mentioned in d. The topic mixing proportion of the three topics in d is represented by $\theta_{d,z} = p(z|\, d)$ where $z \in \mathbf{x_d}$. Each topic z is represented by a distribution over the vocabulary set denoted by: $\Phi_{z,w} = p(w|\, z)$ where $1 \le w \le W$. In this study, vector variables are denoted by bold letters such as $\boldsymbol{\Phi}_z = \{\, \Phi_{z,w} |\ 1 \le w \le W\}$, single variables are denoted by un-bold letters such as $\Phi_{z,w}$.

2.2 Model Description

The proposed Latent Dirichlet relevance model is a generative model describing the process of generating relevant documents for K_0 given topics of interest.

In this model, the language used to generate a document relevant to a topic of interest t is a combination of (1) the language reflecting the meaning of t itself, (2) the language of a general background topic, (3) the language reflecting themes other than t that are also mentioned in the document. For example, in the domain of computer science research papers, suppose that the training set for the topic *machine learning* (ML) includes d_1 a document about applying ML to information retrieval (IR) and d_2 a document about ML tools for the banking industry. The general background topic would be responsible for common words in English and common words in the domain such as "paper", "propose", "approach"... The distribution for topic ML, representing the meaning of ML, would likely give high probabilities to words like "learning", "training", "test" … Topic $t\_o(d_1)$ responsible for other themes in document d_1 would likely generate words relating to the IR aspects mentioned in d_1, while for d_2, $t\_o(d_2)$ would likely generate words such as "bank", "sales", "marketing" that are related to the banking industry emphasis in d_2.

The process of generating relevant documents for K_0 topics of interest is formally described as follows:

1) Pick a multinomial distribution $\boldsymbol{\Phi_b}$ for the background topic (b) from a W-dimensional Dirichlet distribution $Dir(\beta)$.
2) For each topic t in K_0 topics of interest:
   a) Pick a multinomial distribution $\boldsymbol{\Phi_t}$ for t from the W-dimensional $Dir(\beta)$.
   b) For each document d relevant to t:
      i) Pick a multinomial distribution $\boldsymbol{\Phi_{t\_o(d)}}$ for the topic covering themes other than t that are also mentioned in d from the W-dimensional $Dir(\beta)$.
      ii) Pick a multinomial distribution $\boldsymbol{\theta_d}$ from a 3-dimensional $Dir(\alpha)$, each element of $\boldsymbol{\theta_d}$ corresponds to a topic in $\boldsymbol{x_d} = \{b, t, t\_o(d)\}$.
      iii) For each token in document d:
         (1) Pick a topic z among the three topics in $\boldsymbol{x_d}$ from multinomial $\boldsymbol{\theta_d}$.
         (2) Then, pick a word from the corresponding multinomial distribution $\boldsymbol{\Phi_z}$.

The graphical model using plate notation in Fig. 1 describes this process. Numbers in the right-lower corner of the plates (boxes) indicate the number of repetitions of the corresponding plates. In the Figure, w_i (word identity of a token i^{th}) and $\boldsymbol{x_d} = \{b, t, t\_o(d)\}$ (topics generate document d) are observable variables and denoted by shaded circles; z_i (latent topic of a token

i^{th}), $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ are hidden variables and denoted by un-shaded circles; α, β are hyper-parameters of Dirichlet distributions. In our model, values of α, β are pre-defined as in [15][16].

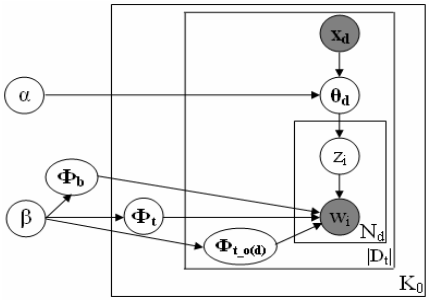

Fig. 1. Latent Dirichlet relevance model

Observe that unlike standard LDA describing how all documents in a corpus are generated, our model describes how *relevant* documents for a set of given topics are composed. Consequentially, each given topic of interest is explicitly associated with a multinomial distribution over the vocabulary. Therefore, we are able to explicitly model relevance.

Again, our advantage is that compared to relevance-based language models which assume all tokens of each training document d of a topic of interest t are generated only by that topic, our model considers two more topics b and $t\_o(d)$. The purpose of the background topic b is to explain words commonly appearing in training documents of all topics. That allows the distribution of t the topic of interest to be more discriminative. The purpose of $t\_o(d)$ is to explain words frequently appearing in the particular document d, but not in other training documents of topic t. That prevents the distribution of topic t from wasting its probability mass on these extra document-specific features. Thus, the consideration of document-specific $t\_o(d)$ topics minimizes the risk of t over-fitting the given set of training documents.

We model the topic mixing proportion $\boldsymbol{\theta_d}$ and topic-word distribution $\boldsymbol{\Phi_z}$ by latent variables which are assumed to be sampled from prior probability distributions of 3-dimensional $Dir(\alpha)$ and W-dimensional $Dir(\beta)$, respectively. The explicit assumption about the prior sources of these variables provides complete generative semantics for the model [2][6][16]. Moreover, the mathematical property that the Dirichlet priors of $p(\boldsymbol{\theta_d}|\ \alpha)$ and $p(\boldsymbol{\Phi_z}|\ \beta)$ are conjugate to their likelihoods (multinomial distributions) $p(z|\ \boldsymbol{\theta_d})$ and $p(w|\ \boldsymbol{\Phi_z})$ results in the fact that their posteriors $p(\boldsymbol{\theta_d}|\ \alpha,\ \{z_i|$ for all tokens in doc $d\})$ and $p(\boldsymbol{\Phi_z}|\ \beta,\ \{w_i|$ for all tokens generated by $z\})$ are also Dirichlet distributions. Mathematically that makes the inference feasible.

2.3 Inference

As in previous work on LDA, we also apply Gibbs sampling to infer latent variables. Formally, Gibbs sampling estimates the conditional distribution of latent variables given observable ones:

$$p(\{\boldsymbol{\Phi_z}|\text{ all topics } z\},\ \{\boldsymbol{\theta_d}|\text{ all docs } d\}\ \{z_i|\text{ all tokens } i\}|\ \{\mathbf{x_d}|\text{ all docs d}\},\ \{w_i|\text{ all tokens } i\}) \quad (1)$$

by generating sequence of $(S+1)$ samples, where each sample contains values for all latent variables. The sampling algorithm is presented in Fig. 2.

For the first sample, $\boldsymbol{\Phi_b}^{(0)}$ is initialized by smoothed term frequencies on all relevant sets, $\boldsymbol{\Phi_t}^{(0)}$ for each topic of interest t is initialized by smoothed term frequencies on its relevant set, $\boldsymbol{\Phi_{t\_o(d)}}^{(0)}$ for each document d is initialized by smoothed term frequencies in that document. $\boldsymbol{\theta_d}^{(0)}$ for each document d is the uniform distribution i.e. $\boldsymbol{\theta_d}^{(0)}=\{1/3,\ 1/3,\ 1/3\}$.

For each of the following S samples (Step 2, Fig. 2), each latent variable is randomly sampled from its posterior distribution given current values of all other variables. Specifically, latent topic of token i^{th} is sampled from its posterior distribution that is estimated by using values of $\boldsymbol{\Phi}$ and $\boldsymbol{\theta}$ in the previous sample (Step 2.1), where w_i and d_i are word index and document index of token i^{th}. After sampling z_i for every token, we update the values for $\boldsymbol{\Phi}$ and $\boldsymbol{\theta}$. For simplicity, instead of truly sampling the values for these variables from their posteriors, which are also Dirichlet distributions as explained above, we deterministically assign them to the expected values w.r.t. these posteriors. In Step 2.2, the numerator is the number of times word w is assigned to topic z in the current sample (i.e. sample $(s+1)$), and the denominator is the number of times topic z appears in the current sample, smoothed by the factor β. In Step 2.3, the numerator is the number of times topic z is assigned in document d, the denominator is document length, smoothed by the factor α.

Given the $(S+1)$ samples, we ignore the first S' samples (samples in the burn-in period), then select every P^{th} samples (i.e. samples S', $(S'+P)$, $(S'+2P)$ …) to approximate the distribution in (1) and to estimate expected values of latent variables w.r.t. this distribution. In our relevance model, the eventual goal is to estimate $\Phi_{z,w}^{*} = p(word=w|\ topic=z)$ for all topics. Those distributions are estimated by averaging over $\Phi_{z,w}^{(s)}$ in these selected samples.

1. Initialize variables: $\boldsymbol{\Phi}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$
2. For s = 0 to (S-1):
   2.1 For i = 1 to N: (N is the number of tokens)
   Sample $z_i^{(s+1)}$ from:

$$p(z_i=z|\ w_i, d_i, others) = p(z_i=z|\ w_i, d_i, \boldsymbol{\theta}^{(s)}, \boldsymbol{\Phi}^{(s)})$$
$$= p(w_i|\ z_i=z, \boldsymbol{\Phi}^{(s)})\ p(z_i=z|\ d_i, \boldsymbol{\theta}^{(s)}) = \phi_{z,w_i}^{(s)}\ \theta_{d_i,z}^{(s)}$$

   2.2 For z = 1 to K: (K is the total number of topics)
   Estimate $\boldsymbol{\Phi}_z^{(s+1)}$:
   For w = 1 to W: (W is the number of words)

$$\Phi_{z,w}^{(s+1)} = p(w|\ z) = \frac{m_{z,w}^{(s+1)} + \beta}{\sum_{w'=1}^{W}(m_{z,w'}^{(s+1)} + \beta)}$$

   2.3 For d = 1 to |D|: ($|D|$ is the number of documents)
   Estimate $\boldsymbol{\theta}_d^{(s+1)}$:
   For each z in $\boldsymbol{x}_d = \{b, t, t\_o(d)\}$:

$$\theta_{d,z}^{(s+1)} = p(z|\ d) = \frac{n_{d,z}^{(s+1)} + \alpha}{\sum_{z' \in X_d}(n_{d,z'}^{(s+1)} + \alpha)}$$

Fig. 2. Gibbs sampling-based inference algorithm

Reducing Complexity: In the algorithm (Fig. 2), the conditional distribution of each $z_i^{(s+1)}$ is independent of any other $z_k^{(s+1)}$ (k≠i) (Step 2.1). So, we can re-arrange the sampling order in Step 2.1 in any order without affecting the final results. Exploiting this observation, in the preprocessing step, we re-arrange tokens in each document such that tokens from the same word (or stem in the case we use stemming) are consecutive. Since the latent topics for the tokens are sampled from the same posterior distribution, the re-arranging could reduce the complexity by a factor of $r=W_d/N_d$, where W_d is the average number of distinct words (or stems), and N_d is the average number of distinct tokens in a document.

3 Experiments

3.1 Pseudo-Relevance Feedback

In this section, we evaluate the effectiveness of our Latent Dirichlet relevance model (abbreviated Dir Rel) on the task of pseudo-relevance feedback based retrieval. We compare the performance of our model against the performance of a standard relevance-based language model (Rel LM). Our implementation of the Rel LM follows the description given in [7]. We also compare performance against a simple no feedback Lucene retrieval baseline [19].

Our experiments are done using four corpora (Table 1). AP and WSJ contain newswire articles. For these corpora, we use 100 topics (title only) and partial judgments for these topics provided by TREC. 20 Newsgroup contains discussion posts. Each of the posts is labeled by one of 20 topics. Cora contains computer science abstract research papers. These papers are also manually assigned to topics. We use 20 topics for this corpus. For 20 Newsgroup and Cora, we have complete relevance judgments.

Table 1. Corpora

Corpus	# of documents	# of topics (queries)
TREC AP	242 918	100 (051-100, 151-200)
TREC WSJ	173 252	100 (051-100, 151-200)
20 Newsgroup	19 956	20
Cora	25 705	20

All documents are stemmed using the Porter stemmer [12] and indexed using Lucene. We do not remove stop words in this experiment. For a simple retrieval baseline, we use all Lucene default parameter settings. From the results returned by Lucene, the top *50* documents for each query are used to train relevance-based language and Latent Dirichlet relevance models. In each case the top ranked *50* words, with the highest probabilities estimated by each model, are used to expand the original query. These parameter values (50x50) have been tuned for Rel LM in previous work. We also use these values for our model. Tuning these values specifically for our model could result in a better performance. We leave this for future work. The expanded query is rerun using Lucene. The performances of the baseline retrieval and pseudo-relevance feedback by the two models are shown for each dataset in Tables 2-5. We measure averages across topics of precision at top 10, top 100 and top 1000 ranked documents. We also measure average precision (averaged across topics to yield MAP) and the total number of relevant documents retrieved (#_rel_ret).

As expected with only a single exception both feedback models are consistently better than the no feedback Lucene baseline (row 1 of the tables) for all measures. The only exception is for 20 Newsgroup for P@10 w.r.t. our model. Focusing just on MAP (the fifth column), notations α and β indicate statistically significant over the baseline and Rel LM (*p-value<0.05* by the paired t-test). The improvements against baseline for the Rel LM are generally in the range of 10% to 43%, while for Dir Rel are in 23% to 100%. In *3* of the *4* cases, the MAP improvements for Dir Rel against Rel LM are around 10%. The best improvement is observed in the 20 Newsgroup dataset (40%). In terms of precision, for example the P@100 score, we find that Dir Rel is consistently better than Baseline and Rel LM in all cases. Thus on the

whole, we find that Dir Rel is successful at achieving improvements over the Rel LM, and both feedback models are, as expected, better than the no feedback baseline. These results support our contention that a) relevant documents may contain portions that are not relevant to the topic of interest and b) it is possible to build more robust relevance models using the Latent Dirichlet framework.

Tables 2. Cora

	P@10	P@100	P@1000	MAP	MAP-Impr	#_rel_ret
Baseline	0.625	0.485	0.1795	0.2307	---	5010
Rel LM	0.65	0.5015	0.1967	0.2549^{α}	10%	6667
Dir Rel	0.665	0.532	0.2124	$0.2844^{\alpha,\beta}$	23%	7130

Tables 3. 20 Newsgroup

	P@10	P@100	P@1000	MAP	MAP-Impr	#_rel_ret
Baseline	0.715	0.5905	0.273	0.1783	---	7875
Rel LM	0.73	0.612	0.3223	0.2548^{α}	43%	15170
Dir Rel	0.67	0.625	0.3933	$0.3568^{\alpha,\beta}$	100%	17621

Tables 4. AP

	P@10	P@100	P@1000	MAP	MAP-Impr	#_rel_ret
Baseline	0.326	0.232	0.0778	0.1948	---	7783
Rel LM	0.372	0.2701	0.0887	0.2409^{α}	23.7%	8864
Dir Rel	0.385	0.2895	0.0945	$0.2650^{\alpha,\beta}$	36.0%	9444

Tables 5. WSJ

	P@10	P@100	P@1000	MAP	MAP-Impr	#_rel_ret
Baseline	0.371	0.2311	0.0618	0.2340	---	6179
Rel LM	0.451	0.2689	0.0678	0.2817^{α}	20.4%	6780
Dir Rel	0.482	0.2904	0.0713	$0.3118^{\alpha,\beta}$	33.2%	7124

3.2 Perplexity

The goal of both relevance-based language models and our Latent Dirichlet relevance model is to estimate the unknown true relevance distribution $p(w| t)$ of some topic of interest t. A traditional measure for comparing the two estimations is perplexity. Perplexity indicates how well estimated distributions predict a new sequence of tokens drawn from the true distribution. Better estimations of the true distribution tend to give higher probabilities to test tokens. As a result, they have lower perplexity, which means they are less surprised by these tokens.

In our experiment such ideal test data is not available. Instead, for each topic (query) t, we approximate the new sequence of relevant tokens by using a held out set of *50* actual relevant documents that do not appear in the top *50* retrieved documents (i.e. training set). We remove stop words from a standard list and also rare words in these relevant documents. Then, we use the remaining tokens as test data. Given estimated distributions $p_{RelLM}(w| t)$ and $p_{DirRel}(w| t)$ obtained from the previous experiment, we compute Perplexity (PPX) for each topic as follows:

$$\text{PPX (Test datal t)} = \exp\{\frac{-1}{N}\sum_{w_i \in Test\ Data}\log(p(w_i \mid t))\}$$

where N is the number of tokens in the test data. Table 6 shows the average perplexity over *20* topics of Cora and 20 Newsgroup. We experiment on Cora and 20 Newsgroup since each topic of these corpora has hundreds of relevant documents. As we see, the perplexity of relevance distributions estimated by the proposed model is significantly lower than distributions estimated by relevance-based language models. The asterisk symbol (*) means that the difference between the two results is statistically significant (i.e. *p-value<0.05* by the paired t-test). This indicates that our Latent Dirichlet relevance model is better able to predict unseen test data from the true distribution as compared to relevance-based language model. Again, the key difference here is that our model considers each document to be generated by a mixture of topics and not just the relevant topic alone.

Table 6. Average Perplexity

	Cora	20 Newsgroup
Rel LM	1364	4976
Dir Rel	942*	3134*

4 Deeper Analyses

In this section, we further analyze the key feature of our proposed model, i.e., the important fact that a document relevant to a given topic could also talk about other non-relevant themes and also have uninformative background terms. Our model's strength is that it automatically extracts relevant terms and rules out non-relevant background terms and terms belonging to other themes in each document. We illustrate this ability with the example below.

The following is a relevant document in the training set for the topic of *information retrieval (IR).* The document seems to be about image retrieval in the medical domain. (Note: to make it more readable, we restore the stemmed words to the original forms.) After running the inference algorithm described in Section 3.3, our model determines the latent topic of each token as shown in the example. Bold tokens are inferred to be generated by *IR* topic (i.e. are relevant terms), italicized tokens are inferred to be generated by the background topic (i.e. are non-relevant terms), underlined tokens are inferred to be generated by *t_o(d)* (and so also non-relevant to *t*).

> *We present a* **principled** *method of obtaining a weighted* **similarity metric** *for* 3D **image retrieval**, firmly rooted *in* **Bayes** *decision theory. The basic idea is to determine a set of most* discriminative *features by evaluating how well they perform on the task of classifying* **images** *according to* **predefined semantic categories**. *We propose this* indirect *method as a* rigorous *way to solve the difficult feature selection problem that comes up in most* **content** *based* **image retrieval** *tasks. The method is applied to* **normal** *and* pathological neuroradiological CT **images**, *where we take advantage of the fact that* **normal** *human* brains *present an approximate* bilateral symmetry *which is often* absent *in* pathological brains. *The* quantitative *evaluation of the* **retrieval** *system shows* **promising** *results.*

As we see all stop words as well as words popular in the domain such as "present", "method" "obtain" are inferred as background terms (recall that Cora contains computer science research papers). Most of the bold are really relevant to *IR* such as "similarity" "retrieval" "semantics". The *t_o(d)* terms identified by the model reflect the specific context of

the document and contain almost nothing about the topic of *IR*. Table 7 shows the top ranked words in the distribution of topic *t_o(d)* for the example document *d* above. This distribution is estimated by averaging over *50* samples. As we see in the table, the topic *t_o(d)* focuses on words representing the specific context of document *d.* We remind the reader that in contrast to our model, relevance-based language models would consider all of the terms in the document as generated by the topic of *IR*. The inability to identify non-relevant terms in training documents results in wasting important probability mass on these non-informative features. So, the example above re-affirms the observation in experiment 3.2 that perplexity achieved by our model is significantly lower than by relevance–based language models.

Table 7. Top ranking words in the t_o(d) distribution specific to our example document

word	**p(word| t_o(d))**
pathology	0.0159
brain	0.0134
3d	0.0104
normal	0.0089
neuroradiolog	0.0074
indirect	0.007
absent	0.007
quantit	0.0065
bilat	0.0065
metric	0.0065

A secondary hypothesis that we now explore is that the proportion of relevant (on topic) tokens in top retrieved documents is likely to be higher than in lower ranked ones. Analogously, the contributions of *t_o(d)* topics in lower ranked documents are likely to be more serious than in top ranked ones. To test this, we explore the contributions, in percentages, of the relevant topical component and the non-relevant component generated by *t_o(d)* over top *100* retrieved documents. We group the results by bins. Each bin contains *10* documents (i.e. the first bin in Fig. 3 includes the top *10* documents, the last bin includes documents from ranks *91-100*). Fig. 3 shows the result averaged over *20* topics on Cora. We see that proportion of relevant tokens in the first bin is *19%* higher than in the 10th bin. Similarly, the contribution of *t_o(d)* topics in the last bin is *27%* higher than in the first bin. The results on other datasets also have the same trend (not shown due to lack of space). Recall that the contribution proportions of topics in documents are modeled as a latent variable in our model, and are determined automatically by the inference algorithm.

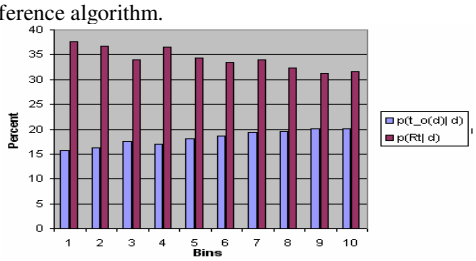

Fig. 3. Topic Contribution Proportion

5 Related Work

Our work proposed in this paper is related to two separate existing directions: relevance models, and probabilistic topic models.

Relevance-based language models [9], a popular approach for relevance modeling, expand possibly human-generated topic (query) to a multinomial distribution over a finite set of words. Specifically, given a topic *t*, the models determine the probability $p(w|R_t)$ of observing a word *w* in documents relevant to topic *t*. The probabilities are estimated by using a set of relevant documents as training data.

$$p(w \mid R_t) = \sum_{D_t \in R_t} p(w \mid D_t)\, p(D_t \mid R_t)$$

where $p(w|D)$ is a language model, and $p(D|R_t) = 1/|R_t|$ as assumed in previous work of Hiemstra et al. [7]. In our experiment, we use the same assumption for implementing relevance-based language models. A limitation of the relevance-based language models is that they are based on a strict assumption that if a document D_t is relevant to a topic, all tokens in the document are equally relevant to that topic.

In [7], Hiemstra et al. propose a three-component mixture relevance model. Besides the relevance component (R_t), the authors introduce two additional components to capture the background (b) and local features (d) in documents. However, the model assumption that the mixing proportions of the three components (λ_b, λ_{Rt}, λ_d) are known in advance and the same for all documents is not reasonable. For instance, in the case where we use top *50* retrieved documents for the query *t* as the training set, the contribution of relevance component in the first document is likely to be higher than in the 50^{th} document. Also in relevance modeling literature, Zhang et al. [17] use a similar three-component mixture model for detecting novelty and redundancy in adaptive filtering. In this framework, inference is done separately by each relevant document. So theoretically, if the mixing proportion for each document were trained instead of being assigned to pre-defined values, we get the result that $\lambda_b = \lambda_{Rt} = 0$ and $\lambda_d = 1$ since the last component best fits the content of the document [17]. On the contrary, in our model, we take both intra- and inter-document statistics into account and associate three types of topics: b, t and t_o(d) to different scopes (Fig. 1), so our model do not have the problem above.

Another approach to alleviate the problem of noises in training documents is to build relevance model on passages (usually windows of text) instead of the whole documents (Liu et al. [11]). However, the way that documents are broken into passages is rather ad-hoc and corpus specific. Moreover, all tokens in each passage are still considered equally relevant. As in the WSJ and Cora example documents we show above, relevant and non-relevant terms appear together even within a sentence.

In topic model literature, Hofmann [8] proposes probabilistic Latent Semantic Indexing (pLSI) modeling each document as a mixture of topics, where a topic is a multinomial distribution. Each word in a document is generated by a topic, and different words in the same document may be generated by different topics. Topics are automatically discovered from the corpus. One limitation of pLSI is that it is not clear how the mixing proportions for topics in a document are generated [2].

To overcome the limitation, Blei et al. [2] propose Latent Dirichlet Allocation (LDA). In LDA, topic proportion of every document is a *K*-dimensional hidden variable randomly drawn from the same Dirichlet distribution, where *K* is the number of topics. Thus, generative semantics of LDA are complete [16]. LDA and its variants have been applied in many applications such as finding scientific topics [6], E-community discovery [18], mixed-

membership analysis [5] and ad-hoc retrieval for representing document language model [4][16]. However, a common problem of both pLSI and LDA is their inability to model the concept of *relevance*, which is key in information retrieval [7][13][14]. Consequently, there is no explicit mapping between the resulting topics generated by pLSI or LDA and the topics in the prior knowledge of human beings. Therefore, the approach could not be applied directly for problems, such as relevance feedback for query modification and text classification, where topics (classes and queries) are provided upfront.

Compared to these two sets of approaches, our Latent Dirichlet relevance model has the following advantages. First, our model explicitly takes the key concept of relevance in account, as in the relevance models [9]. Second, our model could be able to identify relevant and non-relevant terms in training documents. Only relevant terms contribute to the estimation of relevance models. Third, our model possesses complete generative semantics by treating document-topic mixing proportion ($\boldsymbol{\theta}_d$) and topic-word distribution ($\boldsymbol{\Phi}_z$) as hidden random variables sampled from Dirichlet distributions as in the original LDA [2]. As a result, we could exploit the Latent Dirichlet theoretical framework to automatically infer both these variables by taking semantics of topics and content of each relevant document into account.

6 Conclusions

This paper presents a Latent Dirichlet relevance model that combines the advantages of both relevance-based language models [9] and probabilistic topic models [2][15]. Crucially, our model relaxes the strict assumption of relevance-based language models that if a document is relevant to a topic, the entire document is relevant to that topic. This is done by automatically identifying the non-relevant parts in the document. Second, in the context of research on probabilistic topic models, our model explicitly considers the notion of relevance by starting with given topics and estimating their distributions over the corpus vocabulary. We also propose the idea of exploiting the assumption of exchangeability for the tokens in a document ("bag-of-words" assumption) to reduce the computational complexity of the learning algorithm. This idea is not only applicable to our Latent Dirichlet relevance models, but also to conventional LDA.

Our preliminary experiments on pseudo-relevance feedback show the effectiveness the proposed model. The results obtained by the model are consistently better across all of the four corpora than the results of the baseline retrieval (23%-100% improvement in terms MAP) and relevance-based language models (10%-40%). Our work on perplexity re-affirms the advantages of our model over relevance-based language models for the task of estimating the true unknown relevance model.

For future directions, we plan to apply the model for some other applications such as text classification without any human-labeled training data. Instead, we will use as training sets documents returned from a global search engine (e.g. Google) or an intranet search engine, retrieved by the topics themselves. The challenge of this approach is that there is a lot of noise (non-relevant portions) in the returned sets. The ability to automatically detect non-relevant parts in documents of our model is the key to tackling this challenge. Moreover, the background topic in our model could cover common word features of all given classes (topics of interest), so that each of these classes could spend its probability mass on its discriminative features that distinguish itself from the rest of the classes. The background topic could, therefore, increase the margins among the distributions of the classes. This idea is similar to SVM classification technique, but in our model it is not only applicable to case of two classes but also naturally applicable any set of cases.

References

[1] Adrieu, C., Freitas, N., Doucet, A., Jordan, M., *An Introduction to Markov Chain Monte Carlo for Machine Learning*, Machine Learning, 50, (2003).

[2] Blei, M., Ng, A., Jordan, M., *Latent Dirichlet Allocation*, Journal of Machine Learning Research, 3, (2003).

[3] Casella, G., George, E., *Explaining the Gibbs Sampler*, The American Statistician, 46(3), (1992).

[4] Chemudugunta, C., Smyth P., Steyvers, M., *Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model*, In Proceedings of the 20th NIPS, 2006.

[5] Erosheva, E., Fienberg, S., Lafferty, J., *Mixed-membership Models of Scientific Publication*, In Proceedings of National Academy of Science (PNAS), 2004.

[6] Griffiths, T., Steyvers, M., *Finding Scientific Topics*, In Proceedings of National Academy of Science (PNAS), 2004.

[7] Hiemstra, D., Robertson, S., Zaragoza, H., *Parsimonious Language Models for Information Retrieval*, In Proceedings of the 27th ACM SIGIR, 2004.

[8] Hofmann, T., *Probabilistic Latent Semantic Indexing*, In Proceedings of the 15th UAI, 1999.

[9] Lavrenko, V., Croft W. B., *Relevance-based Language Models*, In Proceedings of the 24th ACM SIGIR, 2001.

[10] Lavrenko, V., Croft W. B., *Relevance Models in Information Retrieval*, In Croft, B. and Lafferty, J. (eds.) Language Modeling for Information Retrieval, Kluwer Academics, 2003

[11] Liu, X., Croft, B., *Passage Retrieval Based on Language Models*, In Proceedings of the 11th ACM CIKM, 2002.

[12] Rijsbergen, C., Robertson, S., Porter, M., *New Models in Probabilistic Information Retrieval*, British Library Research and Development Report, 5587, 1980.

[13] Robertson, S., Sparck-Jones, K., *Relevance Weighting of Search Terms*, Journal of American Society for Information Science, 27, 1988.

[14] Sparck-Jones, A., Robertson, S., Hiemstra, D., Zaragoza, H., *Language Modelling and Relevance*, In Croft, B., and Lafferty, J. (eds.) Language Modeling for Information Retrieval, Kluwer Academics, 2003.

[15] Steyvers, M., Griffiths, T., *Probabilistic Topic Models*, In Landauer, T. *et al.* (eds.) Latent Semantic Analysis: A Road to Meaning, Laurence Erlbaum, 2006.

[16] Wei, X., Croft, B., *LDA-based Document Models for Ad-hoc Retrieval*, In Proceedings of the 29th ACM SIGIR, 2006.

[17] Zhang, Y., Callan, J., Minka, T., *Novelty and Redundancy Detection in Adaptive Filtering*, In Proceedings of the 25th ACM SIGIR, 2002.

[18] Zhou, D., Manavoglu, E. Li, J., Giles, L., Zha, H., *Probabilistic Models for Discovering E-Communities,* In Proceedings of the 15th ACM WWW, 2006.

[19] http://lucene.apache.org/