Title Language Model for Information Retrieval

Rong Jin

Language Technologies Institute
School of Computer Science
Carnegie Mellon University

Alex G. Hauptmann

Computer Science Department
School of Computer Science
Carnegie Mellon University

ChengXiang Zhai

Language Technologies Institute
School of Computer Science
Carnegie Mellon University

ABSTRACT

In this paper, we propose a new language model, namely, a title language model, for information retrieval. Different from the traditional language model used for retrieval, we define the conditional probability *P*(*Q*|*D*) as the probability of using query *Q* as the title for document *D*. We adopted the statistical translation model learned from the title and document pairs in the collection to compute the probability *P*(*Q*|*D*). To avoid the sparse data problem, we propose two new smoothing methods. In the experiments with four different TREC document collections, the title language model for information retrieval with the new smoothing method outperforms both the traditional language model and the vector space model for IR significantly.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models — *language model; machine learning for IR*

General Terms

Algorithms

Keywords

title language model, statistical translation model, smoothing, machine learning

1. INTRODUCTION

Using language models for information retrieval has been studied extensively recently [1,3,7,8,10]. The basic idea is to compute the conditional probability *P*(*Q*|*D*), i.e. the probability of generating a query *Q* given the observation of a document *D*. Several different methods have been applied to compute this conditional probability. In most approaches, the computation is conceptually decomposed into two distinct steps: (1) Estimating a document language model; (2) Computing the query likelihood using the estimated document model based on some query model. For example, Ponte and Croft [8] emphasized the first step, and used several heuristics to smooth the Maximum Likelihood Estimate (MLE) of the document language model, and assumed that the query is generated under a multivariate Bernoulli model. The BBN method [7] emphasized the second step and used a two-state hidden Markov model as the basis for generating queries, which, in effect, is to smooth the MLE with linear interpolation, a strategy also adopted in Hiemstra and Kraaij [3]. In Zhai and Lafferty [11], it has been found that the retrieval performance is affected by both the estimation accuracy of document language models and the appropriate modeling of the query, and a two-stage smoothing method was suggested to explicitly address these two distinct steps.

A common deficiency in these approaches is that they all apply an estimated document language model directly to generating queries, but presumably queries and documents should be generated through different stochastic processes, since they have quite different characteristics. Therefore, there exists a “gap” between a document language model and a query language model. Indeed, such a gap has been well-recognized in [4], where separate models are proposed to model queries and documents respectively. The gap has also been recognized in [6], where a document model is estimated based on a query through averaging over document models based on how well they explain the query. In most existing approaches using query likelihood for scoring, this gap has been implicitly addressed through smoothing. Indeed, in [11] it has been found that the optimal setting of smoothing parameters is actually query-dependent , which suggests that smoothing may have helped bridge this gap.

Although filling the gap by simple smoothing has been shown to be empirically effective, ideally we should estimate a *query language model* directly based on the observation of a *document,* and apply the estimated query language model, instead of the document language model, to generate queries. The question then is, “What evidence do we have for estimating a query language model given a document?”. This is a very challenging question, since the information available to us in a typical ad hoc retrieval setting includes no more than a database of documents and queries.

In this paper, we propose to use the titles of documents as the evidence for estimating a query language model for a given document -- essentially to approximate the query language model given a document by the title language model for that document, which is easier to estimate. The motivation of this work is based on the observation that queries are more like titles than documents in many aspects. For example, both titles and queries tend to be very short and concise description of information. The reasoning process in author’s mind when making up the title for a document is similar to what is in a user’s mind when formulating a query

SIGIR ’02, August 11-15, 2002, Tempere, Finland.
Copyright 2002 ACM 1-58113-561-0/02/0008…$5.00.

based on some “ideal document” -- both would be trying to capture what the document is *about*. Therefore, it is reasonable to assume that the titles and queries are created through a similar generation process. The title information has been exploited previously for improving information retrieval, but, so far, only heuristic methods, such as increasing the weight of title words have been tried (e.g., [5,10]). Here we use the title information in a more principled way by treating a title as an observation from a document-title statistical translation model.

Technically, the title language model approach falls into the general source-channel framework proposed in Berger and Lafferty [1], where the difference between a query and a document is explicitly addressed by treating query formulation as a “corruption” of the “ideal document” in the information theoretic sense. Conceptually, however, the title language model is different from the synthetic query translation model explored in [1]. The use of synthesized queries provides an interesting way to train a statistical translation model that can address important issues such as synonymy and polysemy, whereas the title language model is meant to directly approximate queries with titles. Moreover, training with the titles poses special difficulties due to data sparseness, which we discuss below.

A document can potentially have many different titles, but the author only provides one title for each document. Thus, if we estimate title language models only based on the observation of the author-given titles, it will suffer severely from the problem of sparse data. The use of a statistical translation model can alleviate this problem. The basic idea is to treat the document-title pairs as ‘translation’ pairs observed from some translation model that captures the intrinsic document to query translation patterns. This means, we would train the statistical ‘translation’ model based on the document-title pairs in the whole collection. Once we have this general translation model in hand, we can estimate the title language model for a particular document by applying the learned translation model to the document.

Even if we pool all the document-title pairs together, the training data is still quite sparse given the large number of parameters involved. Since titles are typically much shorter than documents, we would expect that most words in a document would never occur in any of the titles in the collection. To address this problem, we extend the standard learning algorithms of the translation models by adding special parameters to model the “self-translation” probabilities of words. We propose two such techniques: One assumes that all words have the same self-translation probability and the other assumes that each title has an extra unobserved null word slot that can only be filled by a word generated through self-translation.

The proposed title language model and the two self-translation smoothing methods are evaluated with four different TREC databases. The results show that the title language model approach consistently performs better than both the simple language modeling approach and the Okapi retrieval function. We also observe that the smoothing of self-translation probabilities has a significant impact on the retrieval performance. Both smoothing methods improve the performance significantly over the non-smoothed version of the title language model. The null word based smoothing method consistently performs better than the method of tying self-translation probabilities. The rest of the paper is organized as follows: We first present the title language model approach in Section 2, describing the two self-translation smoothing methods. We then present the experiments and results in Section 3. Section 4 gives the conclusions and future work.

2. A TITLE LANGUAGE MODEL FOR IR

The basic idea of the title language model approach is to estimate the title language model for a document and then to compute the likelihood that the query would have been generated from the estimated model. Therefore, the key issue is how to estimate the title language model for a document based on the observation of a collection of documents.

A simple approach would be to estimate the title language model for a document using only the title of that document. However, because of the flexibility in choosing different titles and the fact that each document has only one title given by the author(s), it would be almost impossible to obtain a good estimation of title language model directly from the titles.

Our approach is to exploit statistical translation models to find the title language model based on the observation of a document. More specifically, we use a statistical translation model to “convert” the language model of a document to the title language model for that document. To accomplish this conversion process, we need to answer two questions:

1. How to estimate such a statistical translation model?
2. How to apply the estimated statistical translation model to convert a document language model to a title language model and use the estimated title language model to score documents with respect to a query?

Sections 2.1 and 2.2 address these two questions respectively.

2.1 Learning a Statistical Title Translation Model

The key component in a statistical title translation model is the word translation probability $P(tw|dw)$, i.e. the probability of using word tw in the title, given that word dw appears in the document. Once we have the set of word translation probabilities $P(tw|dw)$, we can easily calculate the title language model for a document based on the observation of that document.

To learn the set of word translation probabilities, we can take advantage of the document-title pairs in the collection. By viewing documents as samples of a ‘verbose’ language and titles as samples of a ‘concise’ language, we can treat each document-title pair as a translation pair, i.e. a pair of texts written in the ‘verbose’ language and the ‘concise’ language respectively.

Formally, let $\{<t_i, d_i>, i = 1, 2, \ldots, N\}$ be the title-document pairs in the collection. According to the standard statistical translation model [2], we can find the optimal model M^* by maximizing the probability of generating titles from documents, or

$$M^* = \arg\max_M \prod_{i=1}^{N} P(t_i \mid d_i, M) \quad (1)$$

Based on the model 1 for the statistical translation model [2], Equation (1) can be expanded as

$$
\begin{aligned}
M^* &= \underset{M}{\operatorname{argmax}} \prod_{i=1}^{N} P(t_i \mid d_i, M) \\
&\approx \underset{M}{\operatorname{argmax}} \prod_{i=1}^{N} \prod_{tw \in t_i} \left\{ \frac{\varepsilon}{|d_i|+1} \left(P(tw \mid \phi, M) + \sum_{dw \in d_i} P(tw \mid dw, M) c(dw, d_i) \right) \right\} \\
&\approx \underset{M}{\operatorname{argmax}} \prod_{i=1}^{N} \prod_{tw \in t_i} \left\{ \frac{P(tw \mid \phi, M)}{|d_i|+1} + \sum_{dw \in d_i} P(tw \mid dw, M) P(dw \mid d_i) \right\}
\end{aligned} \qquad (2)
$$

where ε is a constant, ϕ stands for the null word, $|d_i|$ is the length of document d_i, $c(dw, d_i)$ is the number of times that word dw appears in document d. In the last step of Equation (2), we throw out the constant ε and use the approximation that $P(dw \mid d) \approx c(dw, d)/(|d|+1)$. To find the optimal word translation probabilities $P(tw|dw, M^*)$, we can use the EM algorithm. The details of the algorithm can be found in the literature for statistical translation models, such as [2]. We call this model "model 1" for easy reference.

2.1.1 The problem of under-estimating self-translation probabilities

There is a serious problem with using model 1 described above directly to learn the correlation between the words in documents and titles. In particular, the self-translation probability of a word (i.e., $P(w'=w|w)$) will be under-estimated significantly. A document can potentially have many different titles, but authors generally only give one title for every document. Because titles are usually much shorter than documents, only an extremely small portion of the words in a document can be expected to actually appear in the title. We measured the vocabulary overlapping between titles and documents on three different TREC collections: AP(1988), WSJ(1990-1992) and SJM(1991), and found that, on average, only 5% of the words in a document also appear in its title. This means that, most of the document words would never appear in any title, which will result in a zero self-translation probability for most of the words. Therefore, if we follow the learning algorithm for the statistical translation model directly, the following scenario may occur: For some documents, even though they contain every single query word, the probability $P(Q|D)$ can still be very low due to the zero self-translation probability. In the following subsections, we propose two different learning algorithms that can address this problem. As will be shown later, both algorithms improve the retrieval performance significantly over the model 1, indicating that the proposed methods for modeling the self-translation probabilities are effective.

2.1.2 Tying self-translation probabilities (Model 2)

One way to avoid the problem of zero self translation probability is to tie all the self translation probabilities $P(w'=w|w)$ with a single parameter P_{self}. Essentially, we assume that all the self-translation probabilities have approximately the same value, and so can be replace with a single parameter. Since there are always some title words actually coming from the body of documents, the unified self-translation probability P_{self} will not be zero. We call the corresponding model Model 2.

We can also apply the EM algorithm to estimate all the word translation probabilities, including the smoothing parameter P_{self}. The updating Equations are as follows:

Let $P(w'|w)$ and P_{self} stand for the parameters obtained from the previous iteration, $P'(w|w)$ and P'_{self} stand for the updated values of the parameters in the current iteration. According to the EM algorithm, the updating equation for the self-translation probability P'_{self}, will be

$$
P'_{self} = \frac{1}{Z_{self}} \sum_i \sum_w \frac{P_{self} C(w, d_i) C(w, t_i)}{P_{self} C(w, d_i) + \sum_{w' \in d_i \wedge w \neq w'} P(w \mid w') C(w', d_i)} \qquad (3)
$$

where variable Z_{self} is the normalization constant and is defined as

$$
Z_{self} = \sum_i \left(\begin{array}{c} \displaystyle\sum_w \sum_{w' \neq w} \frac{P(w \mid w') C(w, t_i) C(w', d_i)}{P_{self} C(w, d_i) + \sum_{w'' \in d_i \wedge w'' \neq w} P(w \mid w'') C(w'', d_i)} + \\ \displaystyle\sum_w \frac{P_{self} C(w, d_i) C(w, t_i)}{P_{self} C(w, d_i) + \sum_{w' \in d_i \wedge w \neq w'} P(w \mid w') C(w', d_i)} \end{array} \right) \qquad (4)
$$

For those non-self-translation probabilities, i.e. $P(w'\neq w|w)$, the EM updating equations are identical to the ones used for the standard learning algorithm of a statistical translation model except that in the normalization equations, the self-translation probability should be replaced with P_{self}, or

$$
\forall w \quad \sum_{w' \neq w} P'(w' \mid w) = 1 - P'_{self} \qquad (5)
$$

2.1.3 Adding a Null Title Word Slot (Model 3)

One problem with tying all the self-translation probabilities for different words with a single unified self-translation probability is that we lose some information about the relative importance of words. Specifically, those words with a higher probability in the titles should have a higher self-translation probability than those with a lower probability in the titles. Tying them would cause under-estimation of the former and over-estimation of the latter. As a result, the self-translation probability may be less than the translation probability for other words, which is not desirable.

In this subsection, we propose a better smoothing model that is able to discriminate the self-translation probabilities for different document words. It is based on the idea of introducing an extra NULL word slot in the title. An interesting property of this model is that the self-translation probability is guaranteed to be no less than the translation probability for any other word, i.e. $P(w|w) \geq P(w'\neq w|w)$. We call this model Model 3.

Titles are typically very short and therefore only provide us with very limited data. Now, suppose we had sampled more title words from the title language model of a given document, what kinds of words would we expect to have seen? Given no other information, it would be reasonable to assume that we will more likely observe a word that occurs in the document. To capture this intuition, we assume that there is an extra NULL, unobserved, word slot in each title, that can only be filled in by self-translating any word in the body of the document. Use e_t to stand for the extra word slot in the title *t*. With the count of this extra word slot, the standard statistical translation model between the document *d* and title *t* will be modified as

$$
\begin{aligned}
P(t \mid d, M) &\approx P(e_t \mid d, M) \prod_{tw \in t} P(tw \mid d, M) \\
&\approx \left(\sum_{dw \in d} P(dw \mid dw, M) P(dw \mid d) \right) \times \\
&\prod_{tw \in t} \left(\frac{P(tw \mid \phi, M)}{\mid d \mid +1} + \sum_{dw \in d} P(tw \mid dw, M) P(dw \mid d) \right)
\end{aligned}
\qquad (6)
$$

To find the optimal statistical translation model, we will still maximize the translation probability from documents to titles. Substituting the document-title translation probability *P*(*t*|*d*,*M*) with equation (6), the optimization goal (Equation (1)) can be written as

$$
M^* = \underset{M}{\operatorname{argmax}} \prod_{i=1}^{N} \left\{ \begin{aligned} & \sum_{dw \in d_i} P(dw \mid dw, M) P(dw \mid d_i) \times \\ & \prod_{tw \in t_i} \left(\frac{P(tw \mid \phi, M)}{\mid d_i \mid +1} + \sum_{dw \in d_i} P(tw \mid dw, M) P(dw \mid d_i) \right) \end{aligned} \right\} \qquad (7)
$$

Because the extra word slot in every title provides a chance for any word in the document to appear in the title through the self-translation process, it is not difficult to prove that, this model will ensure that the self-translation probability *P*(*w*|*w*) will be no less than *P*(*w'*≠*w*|*w*) for any word *w*. The EM algorithm can again be applied to maximize Equation (7) and learn the word translation probabilities. The updating equations for the word translation probabilities are essentially the same as what are used for the standard learning algorithm for statistical translation models, except for the inclusion of the extra counts due to the null word slot.

2.2 Computing Document Query Similarity

In this section, we discuss how to apply the learned statistical translation model to find the title language model for a document and use the estimated title language model to compute the relevance value of a document with respect to a query. To accomplish this, we define the conditional probability *P*(*Q*|*D*) as the probability of using query *Q* as the title for document *D*, or, the probability of translating document *D* into query *Q* using the statistical title translation model, which is given below.

$$
\begin{aligned}
& P(Q \mid D, M) = \\
& \prod_{qw \in Q} \left\{ \frac{\varepsilon}{\mid d \mid +1} \left(P(qw \mid \phi, M) + \sum_{dw \in d} P(qw \mid dw, M) c(dw, D) \right) \right\} \\
& \approx \varepsilon \prod_{qw \in Q} \left\{ \frac{P(qw \mid \phi, M)}{\mid D \mid +1} + \sum_{dw \in D} P(qw \mid dw, M) P(dw \mid D) \right\}
\end{aligned}
\qquad (8)
$$

As can be seen from Equation (8), the document language model *P*(*dw*|*D*) is not directly used to compute the probability of a query term. Instead, it is "converted" into a title language model through using word translation probabilities *P*(*qw*|*dw*). Such conversion also happens in the model proposed in [1], but there the translation model is meant to capture synonym and polysemy relations, and is trained with synthetic queries. Similar to the traditional language modeling approach, to deal with the query words that can't be generated from title language model, we need to do further smoothing, i.e.

$$
\begin{aligned}
& P(Q \mid D, M) = \\
& \prod_{qw \in Q} \left\{ \begin{aligned} & \frac{\lambda \varepsilon}{\mid d \mid +1} \left(P(qw \mid \phi, M) + \sum_{dw \in d} P(qw \mid dw, M) c(dw, D) \right) + \\ & (1-\lambda) P(qw \mid GE) \end{aligned} \right\} \\
& \approx \varepsilon \prod_{qw \in Q} \left\{ \begin{aligned} & \lambda \left(\frac{P(qw \mid \phi, M)}{\mid D \mid +1} + \sum_{dw \in D} P(qw \mid dw, M) P(dw \mid D) \right) + \\ & (1-\lambda) P(qw \mid GE) \end{aligned} \right\}
\end{aligned}
\qquad (8')
$$

where constant λ is the smoothing constant and *P*(*qw*|*GE*) is the general English language model which can be easily estimated from the collection [1]. In our experiment, we set the smoothing constant λ to be 0.5 for all different models and all different collections.

Equation (8') is the general formula that can be used to score a document with respect to a query with any specific translation model. A different translation model would thus result in a different retrieval formula. In the next section, we will compare the retrieval performance using different statistical title translation models, including Model 1, Model 2 and Model 3.

3. EXPERIMENT

3.1 Experiment Design

The goal of our experiments is to answer the following three questions:

1. *Will the title language model be effective for information retrieval?* To answer this question, we will compare the performance of title language model with that of the state-of-art information retrieval methods, including the Okapi method and the traditional language model for information retrieval.
2. *How general is the trained statistical title translation model*? Can a model estimated on one collection be applied to another? To answer this question, we conduct an experiment that applies the statistical title translation model learned from one collection to other collections. We then compare the performance of using a "foreign" translation model with that of using no translation model.
3. *How important is the smoothing of self-translation in the title language model approach for information retrieval?* To answer this question, we can compare the results for title language model 1 with model 2 and model 3.

We used three different TREC testing collections for evaluation: AP88 (Associated Press, 1988), WSJ90-92 (wall street journal from 1990 to 1992) and SJM (San Jose Mercury News, 1991). We used TREC4 queries (201-250) and their relevance judgments for evaluation. The average length of the titles in these collections is four to five words. The different characteristics of the three databases allow us to check the robustness of our models.

4.2 Baseline Methods

The two baseline methods are the Okapi method[9] and the traditional language modeling approach. The exact formula for the Okapi method is shown in Equation (9)

$$Sim(Q,D)=\sum_{qw\in Q}\left\{\frac{tf(qw,D)\log(\frac{N-df(qw)+0.5}{df(qw)+0.5})}{0.5+1.5\frac{|D|}{avg\_dl}+tf(qw,D)}\right\} \tag{9}$$

where *tf*(*qw*,*D*) is the term frequency of word *qw* in document *D*, *df*(*qw*) is the document frequency for the word *qw* and *avg_dl* is the average document length for all the documents in the collection. The exact equation used for the traditional language modeling approach is shown in Equation (10).

$$P(Q\,|\,D)=\prod_{qw\in Q}((1-\lambda)P(qw\,|\,GE)+\lambda P(dw\,|\,D)) \tag{10}$$

The constant λ is the smoothing constant (similar to the λ in Equation (8')), and *P*(*qw*|*GE*) is the general English language model estimated from the collection. To make the comparison fair, the smoothing constant for the traditional language model is set to be 0.5, which is same as for the title language model.

3.2 Experiment Results

The results on AP88, WSJ and SJM are shown in Table 1, Table 2, and Table 3, respectively. In each table, we include the precisions at different recall points and the average precision. Several interesting observations can be made on these results:

Table 1: Results for AP88 Collection 'LM' stands for traditional language model, 'Okapi' stands for Okapi formula and model-1, model-2 and model-3 stand for title language model 1, model 2 and model 3.

Collection	LM	Okapi	Model 1	Model 2	Model 3
Recall 0.1	0.4398	0.4798	0.2061	0.4885	0.5062
Recall 0.2	0.3490	0.3789	0.1409	0.4082	0.4024
Recall 0.3	0.3035	0.3286	0.1154	0.3417	0.3572
Recall 0.4	0.2492	0.2889	0.0680	0.2830	0.3133
Recall 0.5	0.2114	0.2352	0.0525	0.2399	0.2668
Recall 0.6	0.1689	0.2011	0.0277	0.1856	0.2107
Recall 0.7	0.1369	0.1596	0.0174	0.1460	0.1742
Recall 0.8	0.0811	0.0833	0.0174	0.0897	0.1184
Recall 0.9	0.0617	0.0611	0.0115	0.0651	0.0738
Recall 1.0	0.0580	0.0582	0.0115	0.0618	0.0639
Avg. Prec.	**0.2238**	**0.2463**	**0.2108**	**0.2516**	**0.2677**

First, let us compare the results between different title language models, namely model 1, model 2 and model 3. As seen from Table 1, 2 and 3, for all the three collections, model 1 is inferior to model 2, which is inferior to model 3, in terms of both average precision and precisions at different recall points. In particular, on the WSJ collection, title language model 1 performs extremely poorly compared with the other two methods. This result indicates that title language model 1 may fail to find relevant documents in some cases due to the problem of zero self-translation probability, as we discussed in Section 2. Indeed, we computed the percentage of title words that cannot be found in their documents. This number is 25% for AP88 collection, 34% for SJM collection and 45% for WSJ collection. This high percentage of "missing" title words strongly suggests that the smoothing of self-translation probability will be critical. Indeed, for the WSJ collection, which has the highest percentage of missing title words, title language model 1, without any smoothing of self-translation probability, degrades the performance more dramatically than for collections AP88 and SJM, where more title words can be found in the documents, and the smoothing of self-translation probability is not as critical.

Table 2: Results for WSJ collection. 'LM' stands for traditional language model, 'Okapi' stands for Okapi formula and model-1, model-2 and model-3 stand for title language model 1, model 2 and model 3.

Collection	LM	Okapi	Model 1	Model 2	Model 3
Recall 0.1	0.4308	0.4539	0.2061	0.4055	0.4271
Recall 0.2	0.3587	0.3546	0.1409	0.3449	0.3681
Recall 0.3	0.2721	0.2724	0.1154	0.2674	0.2878
Recall 0.4	0.2272	0.1817	0.0680	0.2305	0.2432
Recall 0.5	0.1812	0.1265	0.0525	0.1723	0.1874
Recall 0.6	0.1133	0.0840	0.0277	0.1172	0.1369
Recall 0.7	0.0525	0.0308	0.0174	0.0764	0.0652
Recall 0.8	0.0328	0.0218	0.0174	0.0528	0.0465
Recall 0.9	0.0153	0.0106	0.0115	0.0350	0.0204
Recall 1.0	0.0153	0.0106	0.0115	0.0321	0.0204
Avg. Prec.	**0.1844**	**0.1719**	**0.0761**	**0.1851**	**0.1950**

Table 3: Results for SJM Collection. 'LM' stands for traditional language model, 'Okapi' stands for Okapi formula and model-1, model-2 and model-3 stand for title language model 1, model 2 and model 3.

Collection	LM	Okapi	Model 1	Model 2	Model 3
Recall 0.1	0.4009	0.4054	0.4226	0.4249	0.4339
Recall 0.2	0.3345	0.3232	0.3281	0.3650	0.3638
Recall 0.3	0.2813	0.2348	0.2712	0.2890	0.3019
Recall 0.4	0.2076	0.1692	0.1991	0.2236	0.2296
Recall 0.5	0.1815	0.1378	0.1670	0.1874	0.1919
Recall 0.6	0.1046	0.0986	0.1095	0.1393	0.1431
Recall 0.7	0.0816	0.0571	0.0782	0.0862	0.0974
Recall 0.8	0.0460	0.0312	0.0688	0.0591	0.0788
Recall 0.9	0.0375	0.0312	0.0524	0.0386	0.0456
Recall 1.0	0.0375	0.0312	0.0524	0.0386	0.0456
Avg. Prec.	**0.1845**	**0.1727**	**0.1910**	**0.1983**	**0.2081**

The second dimension of comparison is to compare title language models with traditional language model. As already pointed out by Berger and Lafferty [1], the traditional language model can be viewed as a special case of translation language model, i.e. all the translation probability $P(w'|w)$ become delta functions $\delta(w,w')$. Therefore, the comparison along this dimension can indicate if the translation probabilities learned from the correlation between titles and documents are effective in improving retrieval accuracy. As seen from Table 1, Table 2, and Table 3, title language model 3 performances significantly better than the traditional language model over all the three collections in terms of all the performance measures. Thus, we can conclude that the translation probability learned from title-document pairs appears to be helpful for finding relevant documents.

Lastly, we can also compare the performance of the title language model approach with the Okapi method [8]. For all the three collections the title language model 3 outperforms Okapi significantly in terms of all the performance measures, except in one case -- The precision at 0.1 recall on the WSJ collection is slightly worse than both the traditional language model approach and Okapi.

To test the generality of the estimated translation model, we applied the statistical title translation model leaned from the AP88 collection to the AP90 collection. We hypothesize that, if two collections are 'similar', the statistical title translation model learned from one collection should be able to give a good approximation of the correlation between documents and titles of the other collection. Therefore, it would make sense to apply the translation model learned from one collection to another 'similar' collection.

Table 4: Results for AP90. 'LM' stands for traditional language model, 'Okapi' stands for Okapi formula and model-3 stand for title language model 3. Different from the previous experiments in which the translation model is learned from the retrieved collection itself, this experiment applies the translation model learned from AP88 to retrieve relevant document in AP90 collection.

Collection	LM	Okapi	Model3
Recall 0.1	0.4775	0.4951	0.5137
Recall 0.2	0.4118	0.4308	0.4454
Recall 0.3	0.3124	0.3374	0.3628
Recall 0.4	0.2700	0.2894	0.3248
Recall 0.5	0.2280	0.2567	0.2665
Recall 0.6	0.1733	0.2123	0.2222
Recall 0.7	0.1294	0.1230	0.1372
Recall 0.8	0.0991	0.0969	0.1136
Recall 0.9	0.0782	0.0659	0.0963
Recall 1.0	0.0614	0.0550	0.0733
Avg. Prec.	**0.2411**	**0.2511**	**0.2771**

Table 4 gives the results of applying the translation model learned from AP88 to AP90. Since title language model 3 already demonstrated its superiority to model 1 and model 2, we only considered model 3 in this experiment. From Table 3, we see that title generation model 3 outperforms the traditional language model and Okapi method significantly in terms of all measures.

We also applied the statistical title translation model learned from AP88 to WSJ to further examine the generality of the model and our learning method. This time, the performance of title language model 3 with the statistical title translation model learned from AP88 is only about the same as the traditional language model and Okapi method for the collection WSJ. Since the statistical title translation model learned from AP88 can be expected to be a much better approximation of the correlation between documents and titles for AP90 than for WSJ, these results suggest that applying the translation model learned from a "foreign" database is helpful only when the "foreign" database is similar to the "native" one. But, it is interesting to note that it has never resulted in any degradation of performance.

4. CONCLUSIONS

Bridging the "gap" between a query language model and document language model is an important issue when applying language models to information retrieval. In this paper, we propose bridging this gap by exploiting document titles to estimate a title language model, which can be regarded as an approximate query language model. The essence of our work is to approximate the query language model for a document with the title language model for the document. Operationally, we first estimate such a translation model by using all the document-title pairs in a collection. The translation model can then be used to "convert" a regular document language model to a title language model. Finally, the title language model estimated for each document is used to compute the query likelihood. Intuitively, the scoring is based on the likelihood that the query could have been a title for a document.

Based on the experiment results, we can draw the following conclusions:

- Based on the comparison between the title language models and the traditional language model and the Okapi method, we can conclude that the title language model for information retrieval is an effective retrieval method. In all our experiments, the title language model gives a better performance than both the traditional language model and the Okapi method.
- Based on the comparison between three different title language models for information retrieval, we can conclude that title generation model 2 and 3 are superior to model 1, and model 3 is superior to model 2. Since the difference between the three different title language models is on how to handle the self-translation probability, we can conclude that, first, it is crucial to smooth the self-translation probability to avoid the zero self-translation probability. Second, a better smoothing method for self-translation probability can improve the performance. Results show that adding an extra null word slot to the title is a reasonable smoothing method for the self-translation probabilities.
- The success of applying the title language model learned from AP88 to AP90 appears to indicate that, in the case when the two collections are similar, the correlation between documents and titles in one collection also tend to be similar to that in the other. Therefore, it would seem to be appropriate to apply the statistical title translation model

learned from one collection to the retrieval task of another similar collection. Even if the collections are not similar, applying a learned statistical title translation model from a foreign database does not seem to degrade the performance either. Thus, the statistical title translation model learned from title-document pairs may be used as a "general" resource that can be applied to retrieval task for different collections.

There are several directions for the future work. First, it would be interesting to see how the style or quality of titles would affect the effectiveness of our model. One possibility is to use the collections where the quality of titles has high variances (e.g., the Web data). Second, we have assumed that queries and titles are similar, but there may be queries (e.g., long and verbose queries) that are quite different from titles. So, it would be interesting to further evaluate the robustness of our model by using many different types of queries. Finally, using title information is only one way to bridge the query-document gap; it would be very interesting to further explore other effective methods that can generate an appropriate query language model for a document.

5. ACKNOWLEDGEMENTS

We thank Jamie Callan, Yiming Yang, Luo Si, and the anonymous reviewers for their helpful comments on this work. This material is based in part on work supported by National Science Foundation under Cooperative Agreement No. IRI-9817496. Partial support for this work was provided by the National Science Foundation's National Science, Mathematics, Engineering, and Technology Education Digital Library Program under grant DUE-0085834. This work was also supported in part by the Advanced Research and Development Activity (ARDA) under contract number MDA908-00-C-0037. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or ARDA.

6. REFERENCES

[1] A. Berger and J. Laffety (1999). Information retrieval as statistical translation. In *Proceedings of SIGIR '99.* pp. 222-229.

[2] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2), pp. 263--311.

[3] D. Hiemstra and W. Kraaij (1999), Twenty-One at TREC-7: ad-hoc and cross-language track, In *Proceedings of the seventh Text Retrieval Conference TREC-7*, NIST Special Publication 500-242, pages 227-238, 1999.

[4] J. Lafferty and C. Zhai (2001), Document language models, query models, and risk minimization for information retrieval, In *Proceedings of SIGIR 2001*, pp. 111-119.

[5] A. M. Lam-Adesina, G. J. F. Jones, Applying summarization techniques for term selection in relevance feedback , In *Proceedings of SIGIR 2001*, pp. 1-9.

[6] V. Lavrenko and W. B. Croft (2001), Relevance-based Language Models, In *Proceedings of SIGIR 2001*, pp. 120-127.

[7] D. Miller, T. Leek and R. M. Schwartz (1999). A hidden Markov model information retrieval system. Proceedings of SIGIR'1999, pp. 214-222. .

[8] J. Ponte and W. B. Croft (1998). A language modeling approach to information retrieval. In *Proceedings of SIGIR' 1998*, pp. 275-281.

[9] S.E. Robertson et al.(1993). Okapi at TREC-4. In *The Fourth Text Retrieval Conference (TREC-4)*. 1993

[10] E. Voorhees and D. Harman (ed.) (1996), The Fifth Text REtrieval Conference (TREC-5), NIST Special Publication 500-238.

[11] C. Zhai and J. Lafferty (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceeding of SIGIR'01*, 2001, pp. 334-342.