Estimation of Statistical Translation Models Based on Mutual Information for Ad Hoc Information Retrieval

Maryam Karimzadehgan
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
mkarimz2@illinois.edu

ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
czhai@cs.illinois.edu

ABSTRACT

As a principled approach to capturing semantic relations of words in information retrieval, statistical translation models have been shown to outperform simple document language models which rely on exact matching of words in the query and documents. A main challenge in applying translation models to ad hoc information retrieval is to estimate a translation model without training data. Existing work has relied on training on synthetic queries generated based on a document collection. However, this method is computationally expensive and does not have a good coverage of query words. In this paper, we propose an alternative way to estimate a translation model based on normalized mutual information between words, which is less computationally expensive and has better coverage of query words than the synthetic query method of estimation. We also propose to regularize estimated translation probabilities to ensure sufficient probability mass for self-translation. Experiment results show that the proposed mutual information-based estimation method is not only more efficient, but also more effective than the synthetic query-based method, and it can be combined with pseudo-relevance feedback to further improve retrieval accuracy. The results also show that the proposed regularization strategy is effective and can improve retrieval accuracy for both synthetic query-based estimation and mutual information-based estimation.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Algorithms, Theory

Keywords

Statistical Machine Translation, Language Models, Estimation, Smoothing, Feedback

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
Copyright 2010 ACM 978-1-60558-896-4/10/07 ...$10.00.

1. INTRODUCTION

Designing effective retrieval models is central for information retrieval. In the past, many retrieval models such as vector space model [28, 29, 30] and probabilistic model [6, 22, 25, 27, 34] have been proposed and gained certain success. Recently, language modeling approaches have received considerable attentions because of its sound statistical foundation and good empirical performance [22, 42]. In language modeling approaches, documents are ranked according to how likely a query is generated from the corresponding document models. In basic language models, document models are estimated based on multinomial distribution and smoothing techniques are critical for document model estimation [42]. When ranking documents, the basic language modeling approach is primarily based on *exact* matching of terms between documents and queries. Since queries are generally succinct and relevant documents may use different vocabulary, such an approach can suffer from vocabulary gap problem.

As a principled approach to capturing semantic word relations, statistical translation language models have been proposed for information retrieval to reduce the gap between documents and queries [2, 8]. Based on statistical machine translation [3], the basic idea of translation language models is to estimate the likelihood of translating a document to a query. Since a term has certain probability to be translated into a different term, translation language models can alleviate the vocabulary gap problem in a direct manner. As a result, translation language models have been successfully applied to different tasks such as cross-lingual information retrieval [12, 20, 39], question answering [40], sentence retrieval [19], and tracking information flow [18].

Surprisingly, there has been little work on applying translation models to ad hoc retrieval. Indeed, the original paper [2] that proposed translation models for ad hoc retrieval appears to be the only study that we are aware of. One possible reason may be because of the difficulty in estimating translation models. In [2], authors solved the problem by generating synthetic queries. Unfortunately, this method has two deficiencies: (1) it is inefficient; (2) there is no guarantee that a query word is covered.

In this paper, we propose a simpler method for estimating a translation model, which is based on normalized mutual information between words. Our Contributions are as follows:

1. We propose an efficient and effective way of estimating word-to-word translation probabilities based on mutual information.

2. We propose regularization of self-translation probabilities, which can improve retrieval performance of translation models with both the existing estimation approach and the proposed mutual information-based approach.

3. We study the issue of smoothing in the context of translation language modeling and show that translation language models are less sensitive to the effect of smoothing.

4. We show that with mutual information, the translation language model can be combined with pseudo-relevance feedback to further improve the retrieval accuracy.

2. STATISTICAL TRANSLATION MODEL FOR RETRIEVAL

In this section, we review basic language modeling approach, statistical translation language model and smoothing methods for statistical translation model. Finally, we discuss the estimation of translation model.

2.1 Basic Language Modeling Approach

The language modeling approach to information retrieval was first introduced by Ponte and Croft [22]. The basic idea can be described as follows. We assume that a query q is generated by a probabilistic model based on a document d. Given a query $q = q_1, q_2, \ldots, q_m$, and a document d, we are interested in estimating $p(d|q)$, i.e. the probability that document d has been used to generate query q. By applying Bayes' formula, we have:

$$p(d|q) \propto p(q|d)p(d)$$

$p(d)$ on the right hand side of the above formula is our *prior* belief that document d is relevant to any query. $p(q|d)$ is the query likelihood for the given document d, which intuitively measures how well document d matches query q. $p(d)$ is often assumed to be uniform and thus can be ignored for ranking documents. Further assuming that each query word is generated independently, we can rewrite the above formula as (in the form of log likelihood):

$$\log p(d|q) \overset{\text{rank}}{=} \sum_{w \in V} c(w,q). \log p(w|d)$$

where $\overset{\text{rank}}{=}$ means equivalence for the purpose of ranking documents, $c(w,q)$ is count of word w in query q, and V is the vocabulary set. The challenging part is to estimate a document model $p(w|d)$. Based on multinomial distribution, the simplest way to estimate $p(w|d)$ is the *maximum likelihood estimator*:

$$p_{ml}(w|d) = \frac{c(w,d)}{\sum_{w'} c(w',d)}$$

Where $c(w,d)$ is count of word w in document d. Due to the data sparseness problem, maximum likelihood estimator under-estimates the probability of unseen words in a document. *Smoothing* techniques address this problem by assigning non-zero probabilities to the unseen words and thus improving the accuracy of probability estimation. Specifically, smoothing is to discount the probabilities of words seen in the text and then assign extra probability mass to the unseen words according to some fallback model. Usually, collection language model is used as fallback model [42]. Two commonly used methods are Jelinek-Mercer and Dirichlet Prior smoothing methods:

Jelinek-Mercer Method (JM Smoothing): This is a linear interpolation of maximum likelihood model with the collection model, using λ as a coefficient weight.

$$p(w|d) = (1-\lambda)p_{ml}(w|d) + \lambda p(w|C) \qquad (1)$$

Where $p(w|C)$ is probability of word w in collection C.

Bayesian Smoothing using Dirichlet Prior (Dirichlet Prior Smoothing): Since the conjugate prior of a multinomial distribution is the Dirichlet distribution, we can specify a Dirichlet prior distribution parameterized as

$$(\mu p(w_1|C), \mu p(w_2|C), \ldots, \mu p(w_n|C))$$

where μ is a parameter. The estimated document model based on the posterior mean is then:

$$p(w|d) = \frac{|d|}{|d|+\mu} p_{ml}(w|d) + \frac{\mu}{|d|+\mu} p(w|C) \qquad (2)$$

2.2 Statistical Translation Language Model

Another interesting way of estimating $p(w|d)$ introduced by Berger and Lafferty [2] is based on statistical machine translation [3]. In order to assess the relevance of a document to a user's query, they have estimated the probability that the query would have been generated as a translation of the document. In other words, they allow the query likelihood to be computed based on a *translation model* of form $p(w|u)$, which is the probability that word u is semantically translated to word w.

To put it more formally, in their model, the query likelihood can be calculated by using the following "translation document model":

$$p_t(w|d) = \sum_{u \in d} p_t(w|u)p(u|d)$$

where $p_t(w|u)$ is the probability of "translating" word u into word w and it allows us to score a document by counting the matches between a query word and semantically related words in the document. If $p_t(w|u)$ only allows a word to be translated into itself, the simple exact matching query likelihood would be achieved. However, $p_t(w|u)$ would in general allow us to translate u into other semantically related words with non-zero probabilities, thus achieving "semantic smoothing" of the document language model.

2.3 Smoothing for Translation Language Model

In this section, we consider statistical machine translation when combined with two basic smoothing methods described in section 2.1.

The basic component in the translation language model is $p_t(w|d) = \sum_{u \in d} p_t(w|u)p(u|d)$ which can be used to replace $p_{ml}(w|d)$ in all basic language model approaches. This will give us 1) translation language model with Dirichlet prior smoothing and 2) translation language model with Jelinek-Mercer smoothing. When we replace $p_{ml}(w|d)$ with $p_t(w|d) = \sum_{u \in d} p(u|d)p_t(w|u)$ in equation 2, we have the following:

$$p_t(w|d) = \frac{|d|}{|d|+\mu}[\sum_{u \in d} p(u|d) \cdot p_t(w|u)] + \frac{\mu}{|d|+\mu} p(w|C) \quad (3)$$

And when p_t is replaced with p_{ml} in equation 1, we have the following:

$$p_t(w|d) = (1-\lambda)[\sum_{u \in d} p(u|d) \cdot p_t(w|u)] + \lambda p(w|C) \qquad (4)$$

Equations 3 and 4 give us Dirichlet prior smoothing and Jelinek-Mercer (JM) smoothing with translation language model, respectively.

Authors in [2] only considered translation language model with Jelinek-Mercer smoothing.

2.4 Estimation of Translation Model

The key part for translation language model is to learn the word-to-word translation probability, $p_t(w|u)$. It is clear that the performance of the proposed smoothed translation model depends on the quality of the word-to-word translation probabilities. In the scenario of statistical machine translation [3], a parallel corpus of two languages is often assumed to be available, and the EM algorithm [5] can be used to estimate a translation model.

In order to gain word-to-word probabilities in monolingual scenario, ideally, we should have a sample of queries and relevant documents, but since we do not often have, Berger and Lafferty [2] use the idea of *synthetic queries* as their training data. The idea is to take a document and synthesize a query to which the document would be relevant. They proposed a sampling technique which distinguishes a document from other documents.

In order to select words which are representative of a document, for each document $\mathbf{d} \in D$, they compute the mutual information statistics [7] for each of its words according to: $I(w, \mathbf{d}) = p(w, d) \log \frac{p(w|d)}{p(w|D)}$, where $p(w|d)$ is the probability of word w in document d, and $p(w|D)$ is the probability of word w in the collection. Their proposed algorithm for generating synthetic queries is shown in figure 1, where synthetic queries are sampled based on normalized mutual information $\tilde{I}$, and the Poisson parameter λ is set to 15. The resulting $(\mathbf{d}, \mathbf{q})$ of documents and synthetic queries are used to estimate the probabilities with the EM algorithm. More details can be found in [2].

```
Begin
    Do for each document $\mathbf{d} \in D$
        Do for $x = 1$ to 5
        Begin
            Select a length $m$ for this query according to
            Poisson distribution
            Do for $i = 1$ to $m$
                Select the next query word by sampling the
                scaled distribution: $q_i \sim \tilde{I}$
            Record ($\mathbf{d}$, $\mathbf{q}$)
        End
End
```

Figure 1: A sampling for synthetic queries

Although generating synthetic queries is a reasonable way to estimate the translation probabilities, this method has two deficiencies: (1) it is inefficient; (2) there is no guarantee that a query word is covered. In the next section, we propose a mutual information-based estimation which is more efficient than this method and has a better word coverage.

3. ESTIMATION OF TRANSLATION MODEL BASED ON MUTUAL INFORMATION

In this section, we propose a more efficient way to estimate translation probabilities which can have a better coverage of query words than the existing method discussed in the previous section. We will also present a way to combine translation language model with pseudo-relevance feedback.

3.1 Mutual Information-Based Approach

Mutual information [26] is a good measure to assess how two words are related. In our method, for each word in the collection, we compute all words which have high mutual information scores with it and normalize the computed mutual information scores as follows:

First, we compute the mutual information scores for each pair of two words w and u in the collection. Informally, mutual information compares the probability of observing w and u *together* (the joint probability) with the probabilities of observing w and u *independently.* The mutual information between words w and u are calculated as follows:

$$I(w;u) = \sum_{X_w=0,1} \sum_{X_u=0,1} p(X_w, X_u) \log \frac{p(X_w, X_u)}{p(X_w)p(X_u)} \qquad (5)$$

where X_u and X_w are binary variables indicating whether u or w is present or absent.

The probabilities are estimated as follows:

$$\begin{aligned}
p(X_w = 1) &= \frac{c(X_w = 1)}{N} \\
p(X_w = 0) &= 1 - p(X_w = 1) \\
p(X_u = 1) &= \frac{c(X_u = 1)}{N} \\
p(X_u = 0) &= 1 - p(X_u = 1) \\
p(X_w = 1, X_u = 1) &= \frac{c(X_w = 1, X_u = 1)}{N} \\
p(X_w = 1, X_u = 0) &= \frac{(c(X_w = 1) - c(X_w = 1, X_u = 1))}{N} \\
p(X_w = 0, X_u = 1) &= \frac{(c(X_u = 1) - c(X_w = 1, X_u = 1))}{N} \\
p(X_w = 0, X_u = 0) &= 1 - p(X_w = 0, X_u = 1) \\
&\quad - p(X_w = 1, X_u = 0) - p(X_w = 1, X_u = 1)
\end{aligned}$$

where $c(X_w = 1)$ and $c(X_u = 1)$ are the numbers of documents containing word w and u, respectively, $c(X_w = 1, X_u = 1)$ is the number of documents that contain both w and u, and N in the total number of documents in the collection.

We then normalize the mutual information score to obtain a translation probability:

$$p_{mi}(w|u) = \frac{I(w;u)}{\sum_{w'} I(w';u)} \qquad (6)$$

$p_{mi}(w|u)$ gives us the probability of translating word u to another word w; intuitively, the probability would be higher if the two words tend to co-occur with each other.

3.2 Optimizing Self-Translation Probability

The approaches described in sections 3.1 and 2.4 might under-estimate the self-translation probabilities, i.e., it is possible that $p(w|u) > p(w|w)$. This may lead to non-optimal retrieval performance because it is possible that a document that matches a query word exactly $(p(w|w))$ gets

Table 1: Sample word translation probabilities using synthetic queries (left) and mutual information (right). Note that words are stemmed.

w=everest	
q	$p(q\|w)$
everest	0.079
climber	0.042
climb	0.0365
mountain	0.0359
mount	0.033
reach	0.0312
expedit	0.0314
summit	0.0253
whittak	0.016
peak	0.0149

w=everest	
q	$p(q\|w)$
everest	0.1051
climber	0.0423
mount	0.0339
028	0.0308
expedit	0.0303
peak	0.0155
himalaya	0.01532
nepal	0.015
sherpa	0.01431
hillari	0.01431

less score contribution from matching the query word exactly than a document that "matches" a query word through translation ($p(w|u)$). To overcome this bias, we introduce a parameter α to control the effect of self-translation. This is a general method that can be applied to adjust the estimated probabilities from any given estimation method.

$$p_t(w|u) = \begin{cases} \alpha + (1-\alpha)p(u|u) & w = u \\ (1-\alpha)p(w|u) & w \neq u \end{cases}$$

and $p(w|u)$ is estimated either with mutual information or synthetic queries. α is a parameter that controls the effect of self-translation probability and when we set $\alpha = 1$, we recover the basic query likelihood method.

The "regularized" translation model $p_t(w|u)$ can then be used in Equations 3 and 4 to rank documents.

3.3 Translation Language Model with Feedback

Feedback techniques have been shown to improve retrieval accuracy substantially[13, 27, 41]. A natural question with translation model is whether translation model can benefit from feedback techniques. In this section, we use pseudo-relevance feedback to expand our query model [41] and then score the expanded query model with translation language model based on the negative cross entropy of the expanded query language model and the translation document model (also equivalent to scoring based on negative KL-divergence):

$$\sum_{p(w|\theta_q)>0} p(w|\theta_q). \log p_t(w|d)$$

where $p(w|\theta_q)$ is the query model generated by pseudo-relevance feedback and $p_t(w|d)$ is a smoothed translation model and can be computed using either of equations 3 or 4.

4. EXPERIMENTS

4.1 Data Set

The experiments in this section use four main document collections: (1) news articles (AP90) with TREC topics 51-100 and 78,321 articles. (2) San Jose Mercury News (SJMN) articles with TREC topics 51-100 and 90,250 articles (3) ad hoc data in TREC7 with topics 351-400 and 528,155 articles and (4) TREC8 with topics 401-450 and 528,155 articles.

In the experiments, we only use title of the queries. As for preprocessing, we do stemming using Porter stemmer [23] and stop word removal. All experiments are done using the Lemur toolkit [1]. The performance is measured using two standard measures: MAP(mean average precision) and precision @10 (precision at 10).

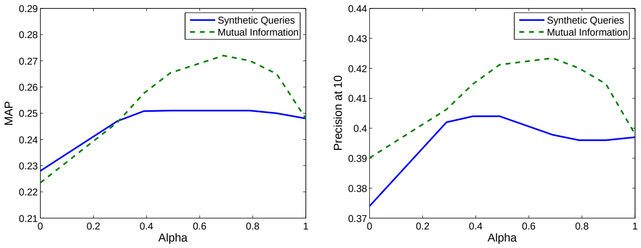

Figure 2: Comparison of mutual information and synthetic queries according to MAP (Left) and Precision at 10 (right). (Both are according to Dirichlet prior smoothing).

The optimal value for Dirichlet prior smoothing for baseline is 1000 for all data sets and optimal value for JM smoothing for baseline method is gained when coefficient is set to 0.5 for AP90 data set and 0.3 for the rest of data sets.

The methods used for experiments in the following sections are: BL (baseline), i.e., either Dirichlet prior smoothing or JM smoothing [42], TM-MI (translation language model with mutual information[2] for word-to-word translation probabilities), TM-SYN (translation language model with synthetic queries), fb (pseudo-relevance feedback on baseline) and fb+TM(pseudo-relevance feedback combined with translation language model using mutual information).

4.2 Comparing Synthetic Queries with Mutual Information

We first look into the question whether mutual information (MI) can be an alternative way of estimating translation model. Table 2 shows the results for both TM-SYN and TM-MI methods with both Dirichlet prior smoothing and JM smoothing, respectively. The results indicate that TM-MI method is able to better capture word relatedness. Indeed, statistical significance tests indicate that the difference between TM-MI and TM-SYN is statistically significant. In addition, estimating translation probabilities by mutual information for all data sets is more efficient than learning translation probabilities by synthetic queries. Table 1 shows a document word together with ten most probable query words that it will translate to by both synthetic queries and mutual information estimation methods. The table shows that the related words for word "everest" in case of mutual information are more specific than for words learned via synthetic queries.

Figure 2 shows the sensitivity of mutual information and synthetic queries to α parameter according to MAP measure (left) and Precision@ 10 (right). The difference indeed makes clearer that mutual information works better than synthetic queries. (Our results for synthetic queries are comparable to those reported in [2].)

According to these results, we can conclude that mutual

[1] http://www.lemurproject.org/

[2] We use mutual information throughout the paper for simplicity but we mean the normalized mutual information described in section 3.1.

Table 2: Performance of Translation Language model with synthetic queries and mutual information estimation according to Dirichlet prior smoothing (left) and JM smoothing (right), * means improvements over TM-SYN are statistically significant with Wilcoxon signed-rank test. We only show the significance tests for MAP measure.

Data	MAP		Precision @10	
	TM-MI	TM-SYN	TM-MI	TM-SYN
AP-90	0.272*	0.251	0.423	0.404
SJMN	0.2*	0.195	0.28	0.266

Data	MAP		Precision @10	
	TM-MI	TM-SYN	TM-MI	TM-SYN
AP-90	0.264*	0.25	0.381	0.357
SJMN	0.197*	0.189	0.252	0.267

Table 3: Performance of Translation Language model on different datasets with Dirichlet Prior smoothing (left) and JM smoothing (right), * means improvements over baseline are statistically significant with Wilcoxon signed-rank test. We only show the significance tests for MAP measure.

Data	MAP		Precision @10	
	BL	TM-MI	BL	TM-MI
AP-90	0.248	0.272*	0.398	0.423
SJMN	0.195	0.2*	0.266	0.28
TREC7	0.183	0.187*	0.412	0.404
TREC8	0.248	0.249	0.452	0.456

Data	MAP		Precision @10	
	BL	TM-MI	BL	TM-MI
AP-90	0.246	0.264*	0.357	0.381
SJMN	0.188	0.197*	0.252	0.267
TREC7	0.165	0.172	0.354	0.362
TREC8	0.236	0.244*	0.428	0.436

information works better than synthetic queries and it is also more efficient.

Because of the high computational complexity of synthetic queries, we cannot compare mutual information with it on larger collections, but later we will further experiment with mutual information on larger collections.

4.3 Comparing Translation Language Model with Standard Query Likelihood

We now look into how well a translation model with our mutual information-based estimation method performs as compared with the standard query likelihood method. Table 3 shows the results for BL and TM-MI methods according to two measures MAP and Precision @10.

Comparing the columns TM-MI with BL in both tables indeed indicates that the TM-MI outperforms method BL. Significant tests using Wilcoxon signed-rank test [37] show the difference between these two methods for cases marked in the tables are statistically significant. Comparing TM-MI with Dirichlet prior smoothing and TM-MI with JM smoothing shows that TM-MI with Dirichlet prior smoothing has higher MAP than TM-MI with JM smoothing.

Stress Tests: In order to have a better understanding of the translation language model, we applied some stress tests on AP90 data set[3]. This experiment is to help us understand when exactly the translation language model would be most beneficial. For the stress test, we gradually and randomly remove query words from relevant documents and compare the performance of BL method with TM-MI method. The results of MAP and Precision @10 are shown in Figure 3.

The results indeed indicate that the baseline method (BL) is purely based on exact matching and the performance will drop significantly if the exact matching does not happen. On the other hand, translation language model (TM-MI) is still able to find relevant documents by translating query words to semantically related words in the documents. This indicates that the translation language model works significantly better than the baseline when there is a vocabulary gap between queries and documents.

4.4 Effect of Smoothing on Translation Language Model

Understanding the influence of smoothing on translation language model is important and no previous work has looked into this. We have a good understanding of smoothing methods for basic language models [42], but it is not clear how smoothing affects the performance of statistical translation language models. In this section, we look into how statistical translation model behaves with the smoothing parameters.

We vary the smoothing parameters (both JM and Dirichlet prior smoothing) for both BL and TM-MI methods. Figure 4 (left and middle) shows the variation of the JM smoothing parameter and Dirichlet prior smoothing parameter on AP90, respectively (we do not show the results on other data sets since they are similar). The result of TM-MI with JM smoothing indicates that the translation model does need a very little smoothing. As shown, the optimal values for translation language model with Dirichlet prior smoothing is 1000 and with JM smoothing is 0.1. As a result, translation language model is less sensitive to the choice of smoothing parameter than the baseline method. And this is intuitively expected, as smoothing is implicitly gained by translating a document word to other *semantically related words*.

Please note that in the translation language model, we have one other parameter to tune, i.e., the number of words used for translation. Figure 4 (right) shows the sensitivity of the number of the words according to MAP measure. As shown in the figure, the translation language model is not so sensitive to the number of words used for translation.

4.5 Results with Pseudo-Relevance Feedback

Both statistical translation model and pseudo-relevance feedback are to capture word associations, so it would be interesting to see whether they are essentially taking advantage of the same associations or they can be combined to achieve even more improvement.

Table 4 shows the pseudo-relevance feedback results for baseline (fb) and when pseudo-relevance feedback is combined with translation language model (fb+TM). For fb+TM method, we first apply pseudo-relevance feedback on initial results (i.e., KL-divergence retrieval model [11]), and then this new query model from pseudo-relevance feedback is used with translation language model to score documents. The feedback parameters are fixed to extract 20 expanded

[3] We got the same trends on other data sets, but we only show the results for AP90 data set.

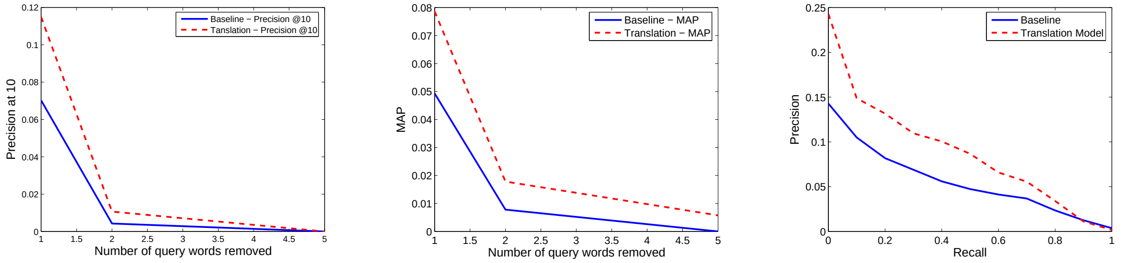

Figure 3: Stress Tests on AP90 Collection, Precision @10 (left) and MAP (middle). Precision-Recall curve when only ”one query word” is removed from relevant documents (right)

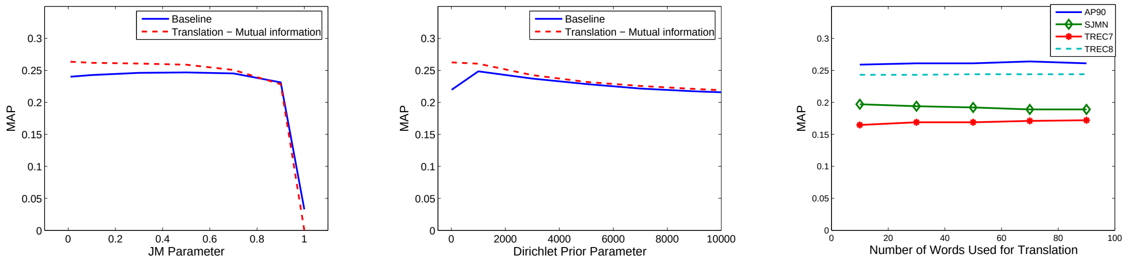

Figure 4: JM parameter variation on AP90 (left), Dirichlet prior parameter variation on AP90 (middle) and Sensitivity of number of words used for translation to MAP (right).

words from the top 10 retrieved documents in the initial run. As shown in table 4, fb-TM method indeed outperforms fb method when used with JM smoothing. Statistical significant tests reveal that the difference is indeed statistically significant. However, fb+TM method does not significantly outperform fb method when used with Dirichlet prior smoothing. An interesting observation is that although the performance of pseudo-feedback (fb) method with JM smoothing is lower than pseudo-feedback with Dirichlet prior smoothing, when pseudo-feedback (fb) is combined with translation language model, i.e., fb+TM method, the better performance is gained with JM smoothing. In fact, the performance of fb+TM with JM smoothing is consistently better than the fb+TM with Dirichlet prior smoothing.

Figure 5 shows the P-R curves for BL, fb and fb+TM methods with JM Smoothing on AP90[4]. This figure indeed indicates that the precision of fb+TM method at different recall points is higher than BL and fb methods. This is an interesting conclusion that translation language model brings in co-occurrence word knowledge that once combined with pseudo-relevance feedback, significant improvement is gained.

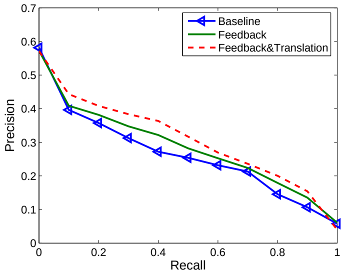

Figure 5: Comparison of Baseline with Translation Language model combined with pseudo-feedback and pseudo-feedback alone on AP90 data set with JM smoothing

4.6 The Need for Self-Translation Regularization

A potential problem of the estimated translation probabilities is that it is possible that $p(w|u) > p(w|w)$. This may lead to non-optimal retrieval performance because it is possible that a document that matches a query word exactly $(p(w|w))$ gets less score contribution from matching the query word exactly than a document that “matches” a query word through translation $(p(w|u))$. The interpolation formula (with α) can help alleviate this problem; indeed, if $\alpha \geq 0.5$, we can always ensure that this constraint be satisfied. So, it would be interesting to see how α affects the performance. Figure 6 shows the sensitivity of α parameter according to MAP measure. We indeed observe that when α is very small (close to no interpolation) the performance is poor, suggesting that it is important to regulate the self-translation probabilities to ensure that it is sufficiently large. In Figure 6, we can see that when $0.5 \leq \alpha \leq 0.8$ for most data sets, we can gain the optimal value. Note that when $\alpha = 1$, we reach the baseline.

4.7 Findings

1. Translation language model is statistically significant bet-

[4] We do not show other curves due to their similarity.

Table 4: Performance of Translation Language model combined with pseudo-feedback with Dirichlet Prior smoothing (left) and JM smoothing (right), * and + mean improvements over baseline and fb, respectively, are statistically significant with Wilcoxon signed-rank test. We only show the significance tests for MAP measure.

Data	MAP			Precision @10		
	BL	fb	fb+TM	BL	fb	fb+TM
AP-90	0.248	0.285	0.285*	0.3978	0.404	0.406
SJMN	0.195	0.231	0.232*	0.266	0.295	0.3
TREC7	0.183	0.226	0.226*	0.412	0.38	0.38
TREC8	0.248	0.270	0.278*	0.452	0.456	0.438

Data	MAP			Precision @10		
	BL	fb	fb+TM	BL	fb	fb+TM
AP-90	0.246	0.271	**0.298***+	0.357	0.383	0.411
SJMN	0.188	0.229	**0.234***+	0.252	0.316	0.313
TREC7	0.165	0.209	**0.222***+	0.354	0.38	0.384
TREC8	0.236	0.240	**0.281***+	0.428	0.4	0.452

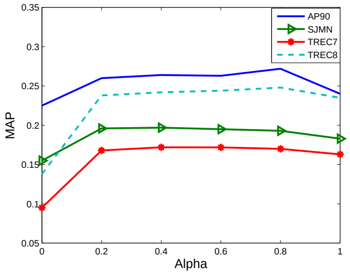

Figure 6: Sensitivity of α parameter to MAP measure

ter than the baseline query likelihood especially when there is a vocabulary gap.
2. Normalized mutual information can be used for word-to-word translation effectively and the results in the previous sections indicate that it is more accurate than synthetic queries. Synthetic queries are inefficient for a large collection such as TREC7 or TREC8.
3. The performance of translation language model combined with pseudo-relevance feedback outperforms pseudo-relevance feedback alone; this indicates that translation language model brings in co-occurrence knowledge in addition.
4. Translation language model is less sensitive to the choice of smoothing parameter than the baseline.
5. Translation language model is robust as it improves over all individual queries.

5. RELATED WORK

Language modeling approaches received considerable attentions recently [22]. One of the most important challenges in language model-based information retrieval is to estimate a better document model. Smoothing is an important approach for document model estimation and has been shown to be critical for information retrieval [42]. To further improve the estimation of document models, different heuristics have been proposed in the past. For example, cluster or topic-model based approaches have been studied in [16, 36]. Tao et al. [33] proposed a document expansion approach to enrich document representation before estimating document models.

Statistical translation models were originally studied in machine translation with the goal of automatically translating sentences between different languages (e.g., French and English) [3] where authors proposed five different translation models. The simplest model (i.e., IBM 1) [3] ignores position information when learning word-to-word translation probabilities. This model has been adopted in information retrieval by Berger and Lafferty [2]. To train translation models, they synthetically generated (query, document) pairs. An alternative way of estimating the translation model is based on document titles [8]. In this work, the authors proposed to use (title, document) pairs as training data. These estimation methods are inefficient and the coverage of query words is low. Our proposed mutual information-based estimation is more efficient and has a better query words coverage.

Translation models have been naturally used in cross-lingual information retrieval domain [20, 39]. For example, Nie et al. [20] used parallel corpus as training data to learn translation models. The work by Lavrenko et al. [12] has adapted the relevance model in two different ways based on KL-divergence retrieval models to perform cross-lingual information retrieval. The cluster-based query likelihood proposed in [10] can be regarded as a form of a translation model where the whole document is translated into the query. Recently, translation models have been applied in many applications including question answering, sentence retrieval and tracking information flow [18, 19, 40]. For example, Xue et al [40] has applied translation model on question-answer archives where question and answer pairs are used to train the translation model. In Contrary to all these works, we studied statistical translation model in *ad hoc retrieval* context.

Vocabulary gap has also been studied in the past. Many studies have tried to bridge the vocabulary gap between documents and queries both based on co-occurrence thesaurus [1, 9, 14, 21, 24, 31, 32, 38] and hand-crafted thesaurus [15, 35]. Some other works have considered to combine both approaches [4, 17]. In this paper, we considered word co-occurrence relationship based on mutual information and incorporated it into translation language model in a more principled way.

6. CONCLUSIONS AND FUTURE WORK

As a principled approach to capturing semantic relation of words in information retrieval, statistical translation models have been shown to outperform simple language models which rely on exact matching of words in the query and documents. In this paper, we propose a new simple way to estimate translation probabilities based on mutual information. Our experiment results indicate that the proposed mutual information estimation method is both more efficient and more effective than the existing synthetic query estimation method. We also proposed to regularize translation probability to ensure sufficient self-translation probability mass, which has been shown to be effective for both estimation methods we experimented with. Our results also show that the translation language model is not so sensitive to the effect of smoothing, and it can be combined with pseudo-relevance feedback to further improve the performance.

For future, it would be interesting to propose some other efficient estimation methods. It would also be interesting to explore other ways of incorporating the translation probabilities into the retrieval formula. Another interesting direction is to study how to transfer the knowledge learned from one collection to another collection.

7. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their useful comments. This material is based upon work supported by a Sohaib and Sara Abbasi Fellowship, Yahoo! Key Scientific Challenge Award, the National Science Foundation under Grant Numbers IIS-0347933, IIS-0713581, IIS-0713571, and CNS-0834709, and by NIH/NLM grant 1 R01 LM009153-01. Any opinions, findings, conclusions, or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

8. REFERENCES

[1] J. Bai, D. Song, P. Bruza, J. Y. Nie, and G. Cao. Query expansion using term relationships in language models for information retrieval. *ACM CIKM*, pages 688–695, 2005.

[2] A. Berger and J. Lafferty. Information retrieval as statistical translation. *ACM SIGIR*, pages 222–229, 1999.

[3] P. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[4] G. Cao, J. Y. Nie, and J. Bai. Integrating word relationships into language models. *ACM SIGIR*, pages 298–305, 2005.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *ACM SIGKDD*, 39(B):1–38, 1997.

[6] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.

[7] F. Jelinek. *Statistical Methods for speech recognition.* MIT Press., 1997.

[8] R. Jin, A. G. Hauptmann, and C. X. Zhai. Title language model for information retrieval. In *ACM SIGIR*, pages 42–48, 2002.

[9] Y. Jing and B. Croft. An association thesaurus for information retrieval. *RIAO*, pages 141–160, 1994.

[10] O. Kurland and L. Lee. Corpus structure, language models, and ad hoc information retrieval. *ACM SIGIR*, pages 194–201, 2004.

[11] J. Lafferty and C. Zhai. Document language models, query models and risk minimization for information retrieval. *ACM SIGIR*, pages 111–119, 2001.

[12] V. Lavrenko, M. Choquette, and B. Croft. Cross-lingual relevance models. *ACM SIGIR*, pages 175–182, 2002.

[13] V. Lavrenko and B. Croft. Relevance-based language models. *ACM SIGIR*, pages 120–127, 2001.

[14] M. Lesk and B. Croft. Word-word associations in document retrieval systems. *American Documentation*, 20:20–27, 1969.

[15] S. Liu, F. Lin, C. Yu, and W. Meng. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. *ACM SIGIR*, pages 266–272, 2004.

[16] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *ACM SIGIR*, pages 186–193, 2004.

[17] R. Mandala, T. tokunaga, H. Tanaka, and K. Satoh. Ad hoc retrieval experiments using wordnet and automatically constructed thesauri. *TREC-7*, pages 475–481, 1998.

[18] D. Metzler, Y. Bernstein, B. Croft, A. Moffat, and J. Zobel. Similarity measures for tracking information flow. *ACM CIKM*, pages 517–524, 2005.

[19] V. Murdock and B. Croft. Simple translation models for sentence retrieval in factoid question answering. *ACM SIGIR*, pages 31–35, 2004.

[20] J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *ACM SIGIR*, pages 74–81, 1999.

[21] H. J. Peat and P. Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. *J. of Information science*, 42(5):378–383, 1991.

[22] J. Ponte and W. B. Croft. A language modeling approach to information retrieval. *ACM SIGIR*, pages 275–281, 1998.

[23] M. Porter. An algorithm for suffix stripping. *Program*, 14(3), 1980.

[24] Y. Qiu and H. Frei. Concept based query expansion. *ACM SIGIR*, pages 160–169, 1993.

[25] C. J. V. Rijbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, pages 106–119, 1977.

[26] C. J. V. Rijsbergen. Information retrieval. *Butterworths*, 1979.

[27] S. Robertson and K. Sparck. Relevance weighting of search terms. *Journal of American Society for Information Science*, 27:129–146, 1976.

[28] G. Salton. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer.* Addison-Wesley, 1989.

[29] G. Salton and M. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill., 1983.

[30] G. Salton, C. S. Yang, and C. T. Yu. A theory of term importance in automatic text analysis. *Journal of American Society for Information Science*, 26(1):33–44, 1975.

[31] H. Schutze and J. O. Pedersen. A co-occurrence based thesaurus and two applications to information retrieval. *Information and processing management*, 33(3):307–318, 1997.

[32] A. F. Smeaton and C. J. V. Rijsbergen. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26(3):239–246, 1983.

[33] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In *HLT-NAACL*, pages 407– 414, 2006.

[34] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, 1991.

[35] E. M. Voorhess. Query expansion using lexical-semantic relations. *ACM SIGIR*, pages 61–69, 1994.

[36] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *ACM SIGIR*, pages 178–185, 2006.

[37] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

[38] J. Xu and B. Croft. Query expansion using local and global document analysis. *ACM SIGIR*, pages 4–11, 1996.

[39] J. Xu, R. Weischedel, and C. Nguyen. Evaluating a probabilistic model for cross-lingual information retrieval. *ACM SIGIR*, pages 105–110, 2001.

[40] X. Xue, J. Jeon, and W. B. Croft. Retrieval models for question and answer archives. In *ACM SIGIR*, pages 475–482, 2008.

[41] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. *ACM CIKM*, pages 403–410, 2001.

[42] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. *ACM SIGIR*, pages 334–342, 2001.