Retrieval Models for Question and Answer Archives

Xiaobing Xue
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts, Amherst, MA, 01003, USA
xuexb@cs.umass.edu

Jiwoon Jeon*
Google, Inc.
Mountain View, CA 94043, USA
jjeon@google.com

W. Bruce Croft
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts, Amherst, MA, 01003, USA
croft@cs.umass.edu

ABSTRACT

Retrieval in a question and answer archive involves finding good answers for a user's question. In contrast to typical document retrieval, a retrieval model for this task can exploit question similarity as well as ranking the associated answers. In this paper, we propose a retrieval model that combines a translation-based language model for the question part with a query likelihood approach for the answer part. The proposed model incorporates word-to-word translation probabilities learned through exploiting different sources of information. Experiments show that the proposed translation based language model for the question part outperforms baseline methods significantly. By combining with the query likelihood language model for the answer part, substantial additional effectiveness improvements are obtained.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Performance

Keywords

Question and Answer Retrieval, Translation Model, Language Model, Information Retrieval

1. INTRODUCTION

Large scale question and answer (Q&A) archives have become an important information resource on the Web. These include the FAQ archives constructed by companies for their products and the archives generated from Web services such as Yahoo Answers! and Live QnA, where people answer questions posed by other people. The retrieval task in a Q&A archive is to find relevant question-answer pairs for new questions posed by the user [6]. Q&A retrieval has several advantages over Web search. First, the user can use natural language instead of only keywords as a query, and thus can potentially express his/her information need more clearly. Second, the system returns several possible answers directly instead of a long list of ranked documents, and can therefore increase the efficiency of finding the required answers. Q&A retrieval can also be considered as an alternative solution to the general Question Answering (QA) problem. Since the answers for each question in the Q&A archive are generated by humans, the difficult QA task of extracting a correct answer is transformed to the Q&A retrieval task.

The major challenge for Q&A retrieval, as for most information retrieval tasks, is the word mismatch between the user's question and the question-answer pairs in the archive. For example, "what is francis scott key best known for?" and "who wrote the star spangle banner?" are two very similar questions, but they have no words in common. This problem is more serious for Q&A retrieval, since the question-answer pairs are usually short and there is little chance of finding the same content expressed using different wording.

To solve the word mismatch problem, many different approaches have been proposed. In this paper, we focus on translation-based approaches since the relationships between words can be explicitly modeled through word-to-word translation probabilities.

Berger and Lafferty [2] proposed using the classic IBM translation model 1 for information retrieval tasks[1]. However, because of various fundamental differences between machine translation and information retrieval, the pure IBM model performs worse than other state of the art retrieval algorithms. We explain the reasons for the poor performance of the pure IBM model in the comparison with the query likelihood language model. This comparison also gives us insights that enable us to address problems with the IBM model. We propose a mixed model that leverages the benefit of both approaches.

Besides designing the translation based retrieval model, another important problem is how to learn good word-to-word translation probabilities. In a Q&A archive, since the asker and the answerer may express similar meanings with different words, it is natural to use the question-answer pairs as the "parallel corpus" that is used for estimation in machine

*The contributions of this author were done during graduate studies at UMass Amherst.

[1] IBM model 1 will be described in the following section.

SIGIR'08, July 20–24, 2008, Singapore.

translation. Since the question part and answer part are written in the same language, the word-to-word translation probabilities can be learned with either part as the source language and the other part as the target. Intuitively, the same word will be related to different sets of words whichever part it appears in. Thus, combining the word-to-word translation probabilities learned with different source and target configurations is beneficial. We propose two combination techniques to improve the translation probability estimates.

Question-answer pairs can also be viewed as documents with different fields, and that probabilities associated with these fields may be estimated in different ways. If we assume a language model approach, we can consider how to estimate probabilities of generating queries. Given the word mismatch problem between the user question and questions in the archive is particularly acute, for the question part, the query is generated by our proposed translation-based language model. For the answer part, the query is simply generated by the query likelihood language model. Our final model for Q&A retrieval is a combination of the above models. Experiments show that our proposed translation-based language model for the question part outperforms three types of representative baseline methods significantly. After combining with query likelihood language model for the answer part, further improvement is observed.

Most previous studies on translation-based information retrieval did not recognize the weakness of the original translation model and adopted the IBM model "as is". They also suffered from low-quality word to word translation probability estimates. In this paper, we overcome these drawbacks and successfully propose a translation-based language model to solve the word mismatch problem in Q&A retrieval.

2. THE RETRIEVAL MODEL

A typical Q&A archive consists of a huge number of question-answer pairs. Here, C denotes the whole archive, $C = \{(q,a)_1, (q,a)_2, ..., (q,a)_L\}$. Q denotes the set of all questions in C, $Q = \{q_1, q_2, ..., q_M\}$ and A denotes the set of all answers in C, $A = \{a_1, a_2, ..., a_N\}$. For each $(q,a)_i \in C$, $q \in Q$ and $a \in A$. Note that $M \leq L$ and $N \leq L$, since the same question can be provided with different answers and the same answer can correspond to different questions. Given the user question $\mathbf{q}$[2], the task of Q&A retrieval is to rank $(q,a)_i$ according to $score(\mathbf{q}, (q,a)_i)$. Under the language modeling framework, this score can be modeled by the probability that $\mathbf{q}$ is generated by $(q,a)_i$. Thus, the following parts of this section focus on how to calculate $P(\mathbf{q}|(q,a)_i)$.

2.1 A Translation-Based Language Model for the Question Part

Both the IBM model and the query likelihood language model are generative models: the former for translation, the latter for information retrieval. Although they were proposed for different purposes, they share many common aspects and assumptions. In this subsection, we compare these two approaches and propose a new model that combines advantages of both approaches[3].

The language modeling approach to information retrieval [12] has been successfully applied to many different applications because of its flexibility and theoretically solid background. The probabilities of sampling (or generating) the query from document language models are used to rank documents. Typically, unigram language models with the maximum likelihood estimator are used to estimate document language models that are smoothed by background collections with the Dirichlet smoothing technique [16].

IBM model 1 [3] does not require any linguistic knowledge of the source or the target language and treats every possible word alignment equally. Because of its simplicity and proven performance, this model has been widely used for many translation tasks. Berger and Lafferty [2] proposed to directly use this model for retrieval tasks.

The ranking function for the query likelihood language model with Dirichlet smoothing and IBM model 1 are compared in Table 1.

Here, $\mathbf{q}$ is the query, D is the document, C is the background collection, λ is the smoothing parameter, $|D|$ and $|C|$ are the lengths of D and C, respectively. $\#(t, D)$ denotes the frequency of term t in D. $P(w|null)$ is the probability that the term w is translated (generated) from the $null$ term. $P(w|t)$ is the the translation probability from word t to word w.

From Table 1, it is easy to recognize that the equations used to describe the query likelihood language model and the IBM model look similar to each other. There are three different parts in the equations, which can be compared as follows:

- $P_{ml}(w|C)$ vs. $P(w|null)$
  $P(w|null)$ is introduced in the IBM model to generate spurious terms in the target sentence. $P_{ml}(w|C)$ has a very similar role in the language models. This background distribution generates common terms that connect content words. Therefore, they basically play the same role in both approaches. However, the concept of the spurious term is a little awkward and the estimated values are less stable than the background distribution used in the language modeling framework. So we choose $P_{ml}(w|C)$ instead of $P(w|null)$ for our model.

- λ vs. 1
  The translation model assumes only one null word and it is not easy to control the effect of background smoothing. On the other hand, the language modeling approach explicitly uses the parameter λ to adjust the amount of smoothing. Considering the smoothing parameters have been shown to have a big impact in retrieval performance, the lack of a mechanism to control background smoothing in the IBM model leads to the relatively poor performance. Therefore, we decided to use λ in our model.

- $P_{ml}(w|D)$ vs. $P_{tr}(w|D)$
  The third difference comes from different sampling strategies. The query likelihood model uses the maximum likelihood estimator. This method gives zero probabilities for unseen words in the document. The word mismatch problem occurs because of this naive sampling method. The IBM model use a more sophisticated sampling method. Every word in the document has some probability of being translated into a target

[2] In this paper, the user question has the same meaning as the user query.

[3] This part was first published in Jeon's Ph.D. thesis [4].

Table 1: Comparisons of language model and IBM model 1.

Language Model	IBM model 1
$P(\mathbf{q}\|D) = \prod_{w\in\mathbf{q}} P(w\|D)$	$P(\mathbf{q}\|D) = \prod_{w\in\mathbf{q}} P(w\|D)$
$P(w\|D) = \frac{\|D\|}{\|D\|+\lambda} P_{ml}(w\|D) + \frac{\lambda}{\|D\|+\lambda} P_{ml}(w\|C)$	$P(w\|D) = \frac{\|D\|}{\|D\|+1} P_{tr}(w\|D) + \frac{1}{\|D\|+1} P(w\|null)$
$P_{ml}(w\|D) = \frac{\#(w,D)}{\|D\|}$, $P_{ml}(w\|C) = \frac{\#(w,C)}{\|C\|}$	$P_{tr}(w\|D) = \sum_{t\in D} P(w\|t) P_{ml}(t\|D)$

word and these probabilities are added up to calculate the sampling probability. Therefore, if a document has many semantically related terms to a target term, then the term gets high probability from the document. This sampling approach that considers word to word relationships helps to overcome the word mismatch problem.

However, we cannot simply choose the sampling method used in the IBM model because of the self translation problem. Since the target and the source languages are the same, every word has some probability to translate into itself. In some cases, low self-translation probabilities reduce retrieval performance by giving very low weights to the matching terms. In the opposite case, very high self-translation probabilities do not exploit the merits of the translation approach.

To overcome this problem, a few different approaches have been proposed. Murdock and Croft [10] use the translation based sampling method only when the target term does not exist in the document and use the maximum likelihood estimator if the target term is present in the document. This method does not fully exploit the power of the translation method. Jeon et al. [6] set $P(w|w) = 1$ for all w while maintaining other word translation probabilities unchanged. This approach produces inconsistent probability estimates and makes the model unstable. Jin et al. [8] force other terms to have lower translation probabilities than self translations: $P(w|w) \geq P(w' \neq w|w)$. This constraint can reduce the problem but very low or very high self translations are still possible. All these improvised modifications gave significant improvements over the original translation model.

Instead of using makeshift solutions, here, we propose to linearly mix two different estimations: maximum likelihood estimation and translation based estimation. Our final ranking function is given as,

$$P(\mathbf{q}|D) = \prod_{w\in\mathbf{q}} P(w|D) \tag{1}$$

$$P(w|D) = \frac{|D|}{|D|+\lambda} P_{mx}(w|D) + \frac{\lambda}{|D|+\lambda} P_{ml}(w|C) \tag{2}$$

$$P_{mx}(w|D) = (1-\beta) P_{ml}(w|D) + \beta \sum_{t\in D} P(w|t) P_{ml}(t|D) \tag{3}$$

In our translation based language model (TransLM), we can control the impact of the translation component by β. If we set a small value for β, the model behaves like the query likelihood model and the importance of matching terms is emphasized. This is similar to increasing the self translation probability. Therefore, we can control the amounts of self translation using β. The background smoothing is adjusted using λ.

In the situation of Q&A retrieval, after applying our translation based language model to the question part, the above equations are changed as follows.

$$\begin{aligned} P(\mathbf{q}|(q,a)) &= \prod_{w\in\mathbf{q}} P(w|(q,a)) && (4) \\ P(w|(q,a)) &= \frac{|(q,a)|}{|(q,a)|+\lambda} P_{mx}(w|(q,a)) + \\ &\quad \frac{\lambda}{|(q,a)|+\lambda} P_{ml}(w|C) && (5) \\ P_{mx}(w|(q,a)) &= (1-\beta) P_{ml}(w|q) + \\ &\quad \beta \sum_{t\in q} P(w|t) P_{ml}(t|q) && (6) \end{aligned}$$

2.2 Incorporating the Answer Part

Note that in Eq. 6, $P_{mx}(w|(q,a))$ is only calculated based on the question part and the answer part of a question-answer pair is not considered. Although it has been shown that doing Q&A retrieval based *solely* on the answer part does not perform well [6], the answer part should provide additional evidence about relevance and, therefore, it should be combined with the estimation based on the question part. We use the query likelihood language model for the answer part, which is combined with the translation-based language model for the question part to form the final retrieval model. In this combined model, $P(\mathbf{q}|(q,a))$ and $P(w|(q,a))$ are estimated with Eq. 4 and Eq. 5. The estimation for $P_{mx}(w|(q,a))$, however, is changed to the following:

$$\begin{aligned} P_{mx}(w|(q,a)) &= \alpha P_{ml}(w|q) + \beta \sum_{t\in q} P(w|t) P_{ml}(t|q) + \\ &\quad \gamma P_{ml}(w|a) && (7) \end{aligned}$$

where $\alpha + \beta + \gamma = 1$.

In Eq. 7, the generation probability of the question part is modeled by $\alpha P(w|q) + \beta \sum_{t\in q} P(w|t) P(t|q)$ and the generation probability of the answer part is modeled by $\gamma P(w|a)$. The relative importance of these components is adjusted through α, β and γ. When $\gamma = 0$, the retrieval model is based only on the translation-based language model for the question part. When $\gamma = 1$, the retrieval model is based on the query likelihood language model for the answer part. In addition, when $\beta = 0$, the retrieval model becomes a combination model which combines the language model estimated from different fields [11].

3. LEARNING WORD-TO-WORD TRANSLATION PROBABILITIES

The performance of the proposed retrieval model heavily depends on the quality of the learned word-to-word translation probabilities. In this section, techniques for estimating word-to-word translation probabilities are discussed in detail.

3.1 The Basic Algorithm

IBM translation model 1 incorporated an EM-based algorithm to learn the word-to-word translation probabilities. Suppose there is a parallel corpus consisting of English-French sentence pairs, $S = \{(\mathbf{e}_1, \mathbf{f}_1), (\mathbf{e}_2, \mathbf{f}_2), ..., (\mathbf{e}_N, \mathbf{f}_N)\}$. The translation probability from an English word e to an French word f is calculated as:

$$P(f|e) = \lambda_e^{-1} \sum_{i=1}^{N} c(f|e; \mathbf{f}_i, \mathbf{e}_i) \quad (8)$$

$$c(f|e; \mathbf{f}_i, \mathbf{e}_i) = \frac{P(f|e)}{P(f|e_1) + ... + P(f|e_l)} \#(f, \mathbf{f}_i) \#(e, \mathbf{e}_i) \quad (9)$$

Here, $\lambda_e = \sum_f \sum_{i=1}^{N} c(f|e; \mathbf{f}_i, \mathbf{e}_i)$ is a normalization factor to make the sum of translation probabilities for the word e equal to 1. $\{e_1, ..., e_l\}$ are English words that appear in $\mathbf{e}_i$. $\#(f, \mathbf{f}_i)$ and $\#(e, \mathbf{e}_i)$ are the number of times the French word f appears in $\mathbf{f}_i$ and the number of times the English word e appears in $\mathbf{e}_i$.

Given the initial value of $P(f|e)$, Eq. 8 and Eq. 9 are used to calculate the updated $P(f|e)$ repeatedly until the probability converges. Brown et. al. [3] showed that this process converges to the same final probability no matter what initial values are set.

3.2 Translation probabilities in Q&A archives

In a Q&A archive, question-answer pairs can be considered as a type of parallel corpus, which is used for estimating word-to-word translation probabilities. In IBM translation model 1, English is the source language and French is the target language. Since the questions and answers in a Q&A archive are written in the same language, the word-to-word translation probability can be calculated through setting either as the source and the other as the target. $P(A|Q)$ is used to denote the word-to-word translation probability with the question as the source and the answer as the target. $P(Q|A)$ is used to denote the opposite configuration.

For a given word, the related words differ when it appears in the question or in the answer. For example, when the word "cheat" appears in the question part, words such as "trust", "forgive", "dump" and "leave" usually appear in the corresponding answer part. These words represent the answerer's suggestion when the asker poses some question about how to react to cheating behaviors. On the other hand, when the word "cheat" appears in the answer, words such as "husband" and "boyfriend" will be observed in the question, which implies most cheating related questions are about the asker's husband and boyfriend. Clearly, all these words are useful to attack the word mismatch problem, thus it is reasonable to combine $P(Q|A)$ and $P(A|Q)$ instead of choosing just one of them.

In addition, the correspondence of words in the question-answer pair is not as strong as in the English-French sentence pair, thus noise will be inevitably introduced for both $P(Q|A)$ and $P(A|Q)$. Suppose a word w_2 appears in the corresponding answer or question part whenever the word w_1 appears in the question or answer part. Another word w_3 only appears in the corresponding answer part when w_1 appears in the question part. Intuitively, w_2 should be more similar to w_1 than w_3. This intuition will be considered implicitly by combining $P(Q|A)$ and $P(A|Q)$, since $P(w_2|w_1)$ will get contributions from both $P(Q|A)$ and $P(A|Q)$, but $P(w_3|w_1)$ only gets the contribution from $P(A|Q)$.

Two methods are used to combine $P(Q|A)$ and $P(A|Q)$. These two methods differ in the stage that the combination occurs. The first method linearly combines the trained word-to-word translation probabilities, which is shown as follows:

$$P_{lin}(w_i|w_j) = (1-\delta) P(w_i, Q|w_j, A) + \delta P(w_i, A|w_j, Q) \quad (10)$$

The second method first pools the question-answer pairs used for learning $P(A|Q)$ and the answer-question pairs used for learning $P(Q|A)$ together, and then uses Eq. 8 and Eq. 9 to learn the combined word-to-word translation probabilities. Suppose we use the collection $\{(q,a)_1, ..., (q,a)_n\}$ to learn $P(A|Q)$ and use the collection $\{(a,q)_1, ..., (a,q)_n\}$ to learn $P(Q|A)$, then $\{(q,a)_1, ..., (q,a)_n, (a,q)_1, ..., (a,q)_n\}$ is used here to learn the combination translation probability $P_{pool}(w_i|w_j)$.

3.3 Examples

Table 2 shows some example word-to-word translations learned from the Wondir dataset[4], using three different ways of estimating the translation probabilities. It can be seen that most top target words are semantically related to the source word, regardless of the estimation method.

To clarify the differences between using questions and answers as sources and targets, consider the word "everest". When this word appears in the question part, the words "29,035", "8,850", "feet" and "height" often appears in the corresponding answers, as shown by the $P(A|Q)$ column, since the user often asks about the height of the mountain everest. On the other hand, if this word appears in the answer part, the corresponding question part often contains words such as "tallest", "highest", and "mountain" as shown by the $P(Q|A)$ column, because "everest" is used as the answer to the questions such as "what is the highest mountain?". Furthermore, P_{pool} shows that after combining $P(A|Q)$ and $P(Q|A)$, we can obtain both these important words.

It is also interesting to note that for the source term "xp", the rank of "drive" is higher than "window" according to $P(A|Q)$. However, P_{pool} assigns the opposite order for these two words, since "window" is also among the top words according to $P(Q|A)$ but "drive" is not. Intuitively, "window" should be more similar to "xp" than "drive". This intuition supports our assumption that two words are more similar when they are related with different source-target configurations. Also, our proposed combination method indeed boosts such words implicitly.

4. EXPERIMENTS

In this section, experiments are conducted on a real Q&A archive to demonstrate the effect of our proposed retrieval model.

[4] The details will be introduced in the following part.

Table 2: Word-to-word translation probability examples. Each column shows the top 10 target terms for a given source term. TTable denotes the type of the word-to-word translation probability table used.

Source	**everest**			**xp**			**search**		
TTable	$P(A\mid Q)$	$P(Q\mid A)$	P_{pool}	$P(A\mid Q)$	$P(Q\mid A)$	P_{pool}	$P(A\mid Q)$	$P(Q\mid A)$	P_{pool}
1	everest	mountain	everest	xp	xp	xp	search	search	search
2	29,035	tallest	mountain	drive	window	window	google	information	google
3	ft	everest	tallest	install	computer	install	page	website	information
4	mount	highest	29,035	click	system	drive	list	free	internet
5	8,850	mt	highest	system	pc	computer	engine	info	website
6	feet	discover	mt	window	version	system	internet	internet	web
7	measure	hillary	ft	computer	edition	click	click	web	list
8	expedition	edmund	measure	pc	install	pc	web	address	free
9	height	mountin	feet	program	software	program	information	picture	info
10	nepal	biggest	mount	microsoft	98	microsoft	result	online	page

4.1 Experimental Configuration

The Wondir collection consists of roughly 1 million question and answer pairs collected from a community based question answering service run by Wondir[5]. The collection has been used in previous research on question and answer services (e.g., [9]). Topics for questions are very diverse, ranging from restaurant recommendations to rocket science. The average length for the question part and the answer part is 27 words and 28 words respectively. Spelling errors are very common in this collection, which makes the word mismatch problem very serious.

For a practical Q&A retrieval system, the search results should be presented to the user in a hierarchical way, where a list of questions is first presented, and after the user selects a specific question the corresponding answers are presented. This hierarchical strategy is more effective for the user than directly presenting a list of question-answer pairs, since this result list could easily be overwhelmed a single question with many answers. Since the relevance of the answer to its corresponding question is usually guaranteed (with the exception of "spam" answers [5]), the retrieval performance of a system can be measured by the rank of relevant questions it returns. Thus, in our experiments, relevance judgments are based on questions. It is easy to transform a question-answer pair rank into a question rank by taking the highest rank among a group of question-answer pairs for the same question. In all the following experiments, ranking algorithms first output question-answer pair ranks that are then transformed into question ranks.

50 questions from the TREC-9 QA track[6] are used for testing. These questions are selected from search engine logs collected from Excite and Encarta. A pooling technique was used to find candidate questions for each query. After being manually judged, 220 semantically similar questions are found in total for 50 queries.

Mean Average Precision (MAP) and Precision at 10 (P@10) are used as the performance measures. A two-sided paired t-test is used for significance testing.

Three types of baselines are used to compare with our proposed retrieval model, which are summarized as follows:

- **Type I:** Query Likelihood Language Model (LM), Okapi BM25 (Okapi) and Relevance Model (RM). This type of baseline represents state-of-the-art retrieval models.
- **Type II:** The combination model which combines the language model estimated from the question part and the answer part at the word level (LM-Comb) [11]. This model is equivalent to setting β as zero in Eq. 7. This type of baseline represents the usual technique used to rank documents with several fields.
- **Type III:** Other translation-based models proposed by Murdock and Croft (Murdock) [10] and Jeon at. al. (Jeon) [6]. This type of baseline represents previous work on translation-based language models. As our proposed translation-based model, this type of baseline uses only the question part.

Table 3: Preliminary Results.

Retrieval	Wondir	
Unit	MAP	P@10
Question	0.2936	0.2105
Answer	0.1625	0.1053
Q&A pair	0.3217	0.2211

4.2 Results

A preliminary experiment was conducted to show the importance of the question part and the answer part for Q&A retrieval. The query likelihood retrieval model was used with the question parts, the answer parts, and the question-answer pairs, respectively. Table 3 shows the retrieval performance with these different fields.

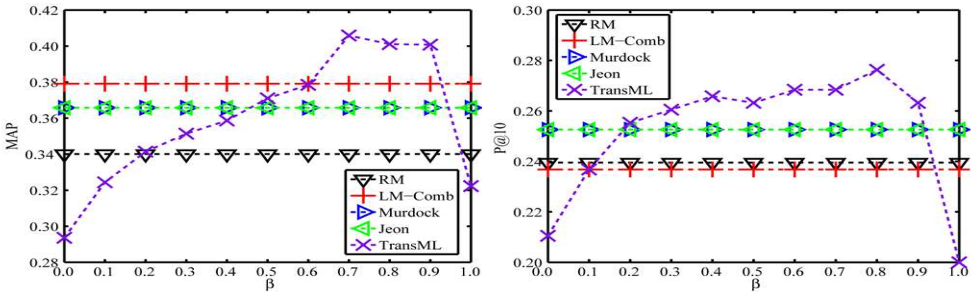
Table 3 shows that on Wondir dataset the question part is more important than the answer part for Q&A retrieval, which supports the observation of previous research [6, 7]. As expected, the performance of using the question-answer pair is better than using each field alone. Thus, it is reasonable for Type I baselines to work on the question-answer pair instead of any field.

Our proposed translation-based retrieval model based on the question part is compared with the state-of-the-art retrieval systems (Type I), the combination technique for documents with different fields (Type II), and other translation-based language models based on the question part (Type III). For each method, the best performance after parameter tuning is reported. For translation-based methods, $P(Q|A)$ and $P(A|Q)$ are both used. The results are summarized in Table 4. Note that, because of the many baselines and estimation methods, we present a comparison of all retrieval runs using significance tests after reporting all results.

Table 4 shows that our proposed TransLM model per-

[5] `http://www.wondir.com`

[6] `http://trec.nist.gov/data/qa/t9_qadata.html`



(a) MAP *(b) P@10*

Figure 1: The influence of β on the performance of TransLM

Table 4: Comparisons with three types of baselines.

			Wondir	
Type	Model	Trans Prob	MAP	P@10
Type I	LM		0.3217	0.2211
	Okapi		0.3207	0.2158
	RM		0.3401	0.2395
Type II	LM-Comb		0.3791	0.2368
Type III	Murdock	$P(Q\|A)$	0.3566	0.25
	Murdock	$P(A\|Q)$	0.3658	0.2526
	Jeon	$P(Q\|A)$	0.3546	0.25
	Jeon	$P(A\|Q)$	0.3658	0.2526
	TransLM	$P(Q\|A)$	0.379	0.2658
	TransLM	$P(A\|Q)$	0.4059	0.2684

Table 5: Results for the combined word-to-word translation probability.

	Wondir	
Trans Prob	MAP	P@10
$P(A\|Q)$	0.4059	0.2684
$P(Q\|A)$	0.379	0.2658
P_{lin}	0.4149	0.2842
P_{pool}	0.4238	0.2868

Table 6: Results for the combined retrieval model.

	Wondir	
Model	MAP	P@10
LM-Comb	0.3791	0.2368
TransLM	0.4238	0.2868
TransLM+QL	0.4885	0.3053

forms better than both the state-of-the-art retrieval systems and the combination technique for documents with different fields. Compared with other translation-based approaches, TransLM shows improvement no matter what kind of word-to-word translation probability is used. In addition, it seems that $P(A|Q)$ is more effective than $P(Q|A)$, which can be explained as the question source being more important than the answer source for generating the user question. Overall, with $P(A|Q)$ our proposed TransLM model outperforms any baseline method noticeably.

Fig. 1 shows the influence of β for the performance of TransLM with $P(A|Q)$. The RM, LM-Comb, Murdock and Jeon methods are used for reference. Fig. 1 shows that, for MAP, TransLM performs better than the baseline methods when β is between 0.6 and 0.9, whereas for P@10, TransLM performs better than the baselines methods when β is between 0.2 and 0.9. In both cases, a relatively broad set of good parameter values is observed.

Although $P(A|Q)$ shows good performance when used for TransLM, we carried out experiments to see whether the performance of TransLM can be further improved with the word-to-word translation probability combination techniques proposed in Section 3.

Table 5 compares the effect of P_{lin} and P_{pool} with $P(A|Q)$ and $P(Q|A)$ when used with TransLM. Here, the best performance of P_{lin} is reported after tuning its linear interpolation parameter δ. Table 5 shows that both combined translation probability estimates are better than $P(Q|A)$ and $P(A|Q)$. The method P_{pool} performs slightly better than P_{lin}. Fig. 2 shows the influence of the parameter δ on the performance of P_{lin}.

After exploring translation based language models and methods for learning word-to-word translation probabilities, we then tested the performance of our retrieval model for question-answer pairs that incorporates the answer part (Eq. 7). This model is called TransLM+QL. Table 6 compares TransLM+QL with TransLM and LM-Combine. P_{pool} is used as the method for estimating translation probabilities.

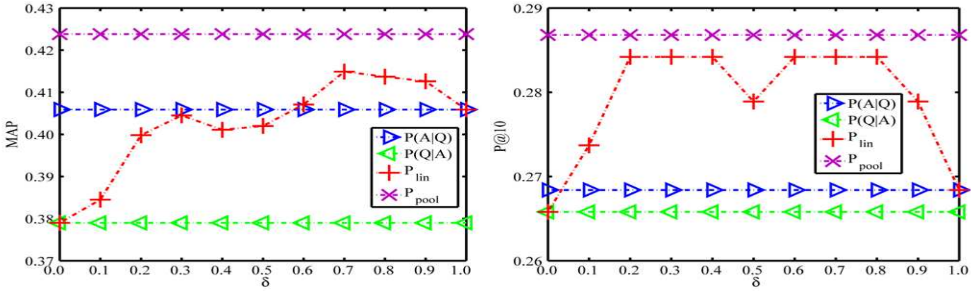
Table 6 shows that by incorporating the query likelihood language model from the answer part, TransLM+QL further improves TransLM significantly, even though the latter showed significant improvement over other baseline methods. This observation can be explained as follows. Sometimes, different wording to the concept of the question can be directly observed in the corresponding answer, thus the query likelihood language model for the answer can be considered as a good complement to the translation-based language model for the question.

Fig. 3 shows the influence of β and γ on the performance of TransLM+QL. Different colors are used to denote the parameter configurations where the performance is inferior to LM-Comb, the performance is between LM-Comb and TransLM and the performance is better than TransLM. No color areas denote the invalid parameter configurations. Clearly, better performance can be observed when γ is small and β is relatively big.

A comparison using the paired t-test is conducted for TransLM($P(Q|A)$), TransLM(($P(A|Q)$), TransLM(P_{lin}), TransLM(P_{pool}) and TransLM(P_{pool})+QL over baseline methods. The results are summarized in Table 7, which clearly shows that our proposed retrieval model TransLM(P_{pool})+QL outperforms all baseline methods significantly.

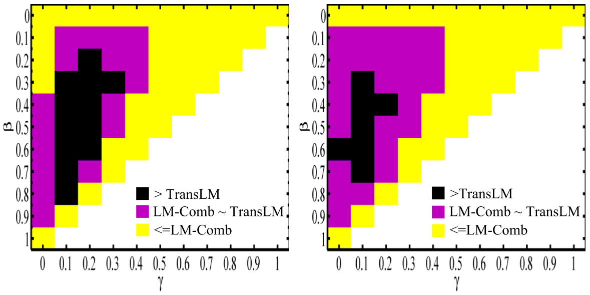
Finally, two retrieval examples are shown in Table 8. These examples indicate that the questions retrieved by our models are more reasonable than the language model results[7].

[7] Since the LM is based on question-answer pairs, it is possi-



(a) MAP *(b) P@10*

Figure 2: The influence of δ on the performance of P_{lin}



(a) MAP *(b) P@10*

Figure 3: The influence of β and γ on the performance of TransLM+QL

5. RELATED WORK

There has been research related to our approach in question and answer retrieval, FAQ retrieval, and translation-based retrieval. Apart from Jeon's previous work ([6, 5, 4], there has been other work on retrieval from FAQ data, which is very similar to Q&A data. Berger et al. [1] reported some of the earliest work using statistical retrieval models, including translation-based approaches, with FAQ data. This work used the "pure" IBM model 1 to find relevant answers among multiple candidate answers for call center users. Their experiments were done with small data sets that consisted of only a few thousand Q&A pairs. Riezler et al. [13] also demonstrated the potential advantages of a translation-based approach to retrieval with FAQ data using a more sophisticated translation model trained on a large amount of data extracted from FAQ pages on the Web. Soricut and Brill [15] used one million FAQs collected from the web to train their answer passage retrieval system. They also used the original IBM model 1 without any modification. Other work on FAQ retrieval has used simpler retrieval models. For example, FAQ Finder [14] used the conventional vector space model to calculate the statistical similarities between questions and WordNet to help calculate the semantical similarities. These two types of similarities were combined heuristically to rank FAQs.

Recently, Jijkoun and Rijke [7] automatically collected approximately three million FAQs from the web and implemented a retrieval system for the collected FAQ collection. Their retrieval system was constructed based on the vector space model. Several combinations of scores for different fields were attempted in their experiments and the results showed the importance of the question part. These results were interesting due to the scale of the collection they used.

There are two main issues with applying translation methods. The first is to modify the model to make it suitable for monolingual transformation rather than translation, and the second is the estimation of translation probabilities. In section 2, we reviewed some previous attempts [6, 10, 8] to address these problems.

6. CONCLUSION AND FUTURE WORK

Q&A retrieval has become an important issue due to the popularity of Q&A archives on the web. In this paper, we propose a novel translation-based language model to solve this problem. Our approach combines the translation-based language model estimated using the question part and the query likelihood language model estimated using the answer part. A new technique was described for using different configurations of question-answer pairs to improve the quality of the translation probability estimates. The retrieval experiments demonstrated the effectiveness of both the retrieval model and the estimation technique.

Our experiments were conducted on one type of Q&A archive, which was collected from a web service where people answer questions posed by other people. Further work will focus on testing the effect of the proposed retrieval model on FAQ archives. We also plan to work on data from Yahoo! Answers, which is potentially a much larger collection than Wondir. The new experiments will use questions derived from this data. Other techniques for combining the models estimated from the question and answer parts will be investigated. In addition, phrase-based machine translation models have shown superior performance compared to word-based translation models in translation applications. We plan to study the effectiveness of these models in the Q&A setting.

7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-0711348. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. Briding the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on*

ble that some questions without query words are retrieved.

Table 7: T-Test result summary. m and p denote p-value < 0.05 based on MAP and P@10 respectively.

T-Test	LM	OKapi	RM	LM-Comb	TransLM($P(Q\|A)$)	TransLM($P(A\|Q)$)
TransLM($P(Q\|A)$)					n/a	
TransLM($P(A\|Q)$)	m,p	m,p				n/a
TransLM(P_{lin})	m,p	m,p		p	m	p
TransLM(P_{pool})	m,p	m,p		p	m	p
TransLM(P_{pool})+QL	m,p	m,p	m,p	m,p	m,p	m,p

Table 8: Retrieval Examples. Top 5 retrieved questions are listed for each query.

	LM	TransLM(P_{pool})+QL
Query:	who is the leader of india	
1	who is agashthy	who is the prime minister of india
2	who is the air chief marshall of airforce of india	who is current vice prime minister of india
3	who is veerappan	who is the army chief of india
4	how is father of india	who is the finance minister of india
5	who is the general seceratary of india	who is the first prime minister of india
Query:	who made the first airplane that could fly	
1	what is the oldest aiirline that still fly airplane	what is the oldest aiirline that still fly airplane
2	who is bin ladin	who was the first one who fly with plane
3	which airplne of the world has been fly the longest	who was the first person to fly a plane
4	what has 4 wheel and fly	who the first one fly to the spase
5	how do airplane fly	who the first one who fly to sky

Research and Development in Information Retrieval, pages 192–199, 2000.

[2] A. Berger and J. Laffery. Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 222–229, 1999.

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

[4] J. Jeon. *Searching Question and Answer Archives.* IR, University of Massachusetts, August 2007.

[5] J. Jeon, W. B. Croft, J. Lee, and S. Part. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 228–235, 2006.

[6] J. Jeon, W. B. Croft, and J. H. Lee. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM Conference on Information and Knowledge Management*, pages 84–90, 2005.

[7] V. Jijkoun and M. de Rijke. Retrieving answers from frequently asked questions pages on the web. In *Proceedings of the 14th ACM Conference on Information and Knowledge Management*, pages 76–83, 2005.

[8] R. Jin, A. G. Hauptmann, and C. Zhai. Title language model for information retrieval. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–48, 2002.

[9] X. Liu, W. B. Croft, and M. Koll. Finding experts in community-based question-answering services. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 315–316, New York, NY, USA, 2005. ACM.

[10] V. Murdock and W. B. Croft. A statistical model for sentence retrieval. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 684–691, 2005.

[11] O. Paul and J. Callan. Language models and structured document retrieval. In *Proceedings of 1st INEX workshop*, 2003.

[12] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, 1998.

[13] S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. Liu. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 464–471, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[14] R. D. Rurke, K. J. Hammond, V. A. Kulyukin, S. L. Lytinen, N. Tomuro, and S. Schoenberg. Question answering from frequently asked question files: experiences with the faq finder system. *AI Magazine*, 18(2):57–66, 1997.

[15] R. Soricut and E. Brill. Automatic question answering: beyond the factoid. In *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics meeting*, pages 57–64, 2004.

[16] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342, 2001.