Improving Search Relevance for Short Queries in Community Question Answering

Haocheng Wu*
University of Science and Technology of China
Hefei, China
ustcwhc@outlook.com

Wei Wu
Microsoft Research
No. 5 Danling Street, Haidian District
Beijing, China
wuwei@microsoft.com

Ming Zhou
Microsoft Research
No. 5 Danling Street, Haidian District
Beijing, China
mingzhou@microsoft.com

Enhong Chen
University of Science and Technology of China
Hefei, China
cheneh@ustc.edu.cn

Lei Duan
Microsoft
1020 Enterprise Way
SunnyVale, CA, US
Lei.Duan@microsoft.com

Heung-Yeung Shum
Microsoft Research
One Microsoft Way
Redmond, WA, US
hshum@microsoft.com

ABSTRACT

Relevant question retrieval and ranking is a typical task in community question answering (CQA). Existing methods mainly focus on long and syntactically structured queries. However, when an input query is short, the task becomes challenging, due to a lack information regarding user intent. In this paper, we mine different types of user intent from various sources for short queries. With these intent signals, we propose a new intent-based language model. The model takes advantage of both state-of-the-art relevance models and the extra intent information mined from multiple sources. We further employ a state-of-the-art learning-to-rank approach to estimate parameters in the model from training data. Experiments show that by leveraging user intent prediction, our model significantly outperforms the state-of-the-art relevance models in question search.

Categories and Subject Descriptors

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

Keywords

community question answering, question search, short query, user intent

1. INTRODUCTION

In the last decade, many community question answering (CQA) archives such as Yahoo! Answers, Quora, and Baidu Knows have emerged and accumulated a large number of questions, answers, and users. Many question-answer pairs on these sites are of high quality, attracting more and more attention on how to leverage historical content to answer new queries.

One way to leverage content in CQA archives is for the system to search the existing questions when users submit their queries and reuse the answers for relevant questions. Many studies have been conducted along this line [28, 26, 29, 15]. Existing methods usually focus on long and syntactically structured queries. However, we have found that when searching CQA archives, users influenced by web search are used to issuing short queries. For example, we collected a one-day search log of Yahoo! Answers and found that 24% of queries were shorter than 4 words and incomplete sentences. Basically speaking, for many users, organizing their information needs via long queries is more difficult than simply typing several keywords. Therefore, besides long queries, considering short queries in CQA question search is also important. On the other hand, on many CQA sites, the search results are not satisfactory when an input query is short. For example, we found that in Yahoo! Answers, the first result for the query "surface battery" was irrelevant to the new "Microsoft Surface Tablet", although most likely the user making the query expected to see questions about the battery performance of the new product[1]. The bad performance of search engines on popular CQA sites further demonstrates the importance and challenge of improving search relevance for short queries.

In this paper, we examine how to improve search relevance for short queries in CQA question search. One big challenge for short query search in CQA is how to understand the underlying search intent in short queries. On the one hand, understanding user search intent is the key to improving search relevance. For example, in the log of Yahoo! Answers, we found that someone issued a query "bad esn", browsed several search results, and finally clicked on the question, "How to fix a bad ESN number?" For this query, people usually want to see something about "fixing the bad

*The work is done when the author is an intern at Microsoft Research Asia

WSDM'14, February 24–28, 2014, New York, New York, USA.

http://dx.doi.org/10.1145/2556195.2556239.

[1]The first returned question was "How long should my laptop have lasted? Is this worth complaining to customer care over?" before we submit the draft.

esn issue on their cell phones", and thus results like "how to fix a bad esn number" should be highly ranked. However, without an understanding of the intent, it is hard to achieve this goal. On the other hand, understanding the underlying search intent in short queries is quite difficult, because short queries usually convey less information. More seriously, signals in CQA data that can facilitate search intent understanding are very sparse, which makes the task even more challenging. For example, unlike web pages, there is little anchor information in CQA data.

In this paper, we propose a model for improving the search relevance of short queries in CQA. To the best of our knowledge, we are the first to study this problem. Specifically, we mine search intent from question descriptions in CQA archives, web query logs, and web search results. Different types of sources provide us different types of signals and reveal user intent from different perspectives. From question descriptions, we obtain some specific and question-oriented intent; from web search logs, we gain an understanding of the pervasive preference of common web users; and from the top web search results, we further distill some of the popular topics related to the short queries. With the intent prediction, we propose a new intent-based language model. The model measures the relevance between a short query and a candidate question based on not only the query itself but also the underlying intent hidden behind the query. The final relevance is calculated as a linear combination of the similarity from the original query and the similarity from the associated intent. To further improve the model, we employ LambdaMART [6], which is a state-of-the-art learning to rank approach, to learn the parameters in the model from training data. With all the extra sources and supervision, our model can select highly relevant questions that match the information needs of short queries.

We conducted our experiments with data sets from Yahoo! Answers and Quora for short queries with an average length of 1.94. Yahoo! Answers is the largest and the most popular CQA site in the English world, while Quora is a rapidly growing CQA site and famous for its high quality content [25]. We compared our model with a number of state-of-the-art relevance models in CQA question search. The experimental results showed that our model significantly improves the performance of question ranking for short queries over the baseline methods.

Our contributions in this paper are 3-fold: (1) a proposal for improving search relevance for short queries in CQA; (2) a proposal for an intent-based language model by leveraging search intent that is mined from CQA, web search logs, and web search results; (3) the empirical verification of the efficacy of the proposed method on data from two popular CQA sites, Yahoo! Answers and Quora.

The rest of the paper is organized as follows. Section 2 introduces related work. Section 3 describes the methods of search intent mining. Section 4 provides the specifics of our model. Section 5 presents the experiments and finally Section 6 concludes the paper.

2. RELATED WORK

2.1 Question Search

Burke, et al. [7] retrieved semantically similar questions from frequently asked questions (FAQ) using a WordNet dictionary and a maker-passing algorithm [22]. In recent years, most retrieval models are based on language models [23]. Translation-based language models were proposed by interpolating translation probabilities into language models. The translation probabilities, including word-to-word translation probability [15] and phrase-to-phrase translation probability [29], are learnt from question-question pairs [15], question-description pairs [18], and question-answer pairs [28]. In addition to translation-based models, Cao, et al. [8] proposed a language model with leaf category smoothing in which they considered leveraging category information specified by users to improve search relevance. Duan, et al. [12] proposed a mixture model to extract topic and focus from questions and incorporated the information into language models. Wang, et al. [26] developed a tree kernel-based syntactic tree matching method to find similar questions. These work provides different ways to retrieve relevant questions from CQA archives. However, none of them considers the question retrieval problem for short queries.

Recently, Liu, et al. [19] proposed predicting web searcher satisfaction with question-answers from CQA. They extracted features that measure query clarity, query-question matching, and answer quality, and learned a regression model to predict how a question-answer pair in CQA could satisfy a web searcher's information needs. Compared with their work, we focus on understanding intent behind short queries and leveraging query expansion techniques to match query and questions in CQA, which goes beyond keyword matching as in Liu, et al.'s work.

2.2 User Intent Mining

User intent can be mined from a variety of sources. In CQA archives, in order to make their information needs clear, askers sometimes supply descriptions for their questions. These descriptions specify the askers' requests and thus can be used to mine the askers' information needs. Li, et al. [18] treated questions and the corresponding question descriptions as source-target pairs and employed a translation model to predict the askers' intent for ambiguous questions. In web search, user behavior when searching and clicking reflects their preference for particular search results. Therefore, user behavior implicitly indicates their search intent. Hu, et al. [13] leveraged web query logs and represented search intent in queries by subtopics mined from co-clicked queries and documents, where each subtopic was represented by a cluster of query expansions and clicked URLs. In addition to question descriptions in CQA and web search logs, search results can also be utilized to mine user intent for queries. For example, Wang, et al. [27] extracted text fragments containing query terms from the search results and clustered them into subtopics. In this paper, we have leveraged user intent mined from various sources to improve the relevance of short queries in CQA question search. Recently, Chen, et al. [10] proposed improving question search relevance with user intent. There is a stark difference between our work and theirs. We focus on short queries, while they deal with long questions.

2.3 Relevance Ranking

Relevance ranking is one of the key components of web search. Given a query, documents[2] are retrieved and ordered according to their relevance to the query. The relevance between a query and a document can be measured by relevance

[2]In this work, the documents are questions.

models. Traditional relevance models, such as BM25 [24], Language Model for Information Retrieval [16, 23], and dependency model [2, 3] based on Markov Random Field [21], are hand-crafted with a few parameters left for manually tuning. Recently, a supervised learning approach, namely learning-to-rank, has proven to be very affective at the automatic construction of relevance models for web search [17, 20]. In this paper, we also employ a learning-to-rank approach to learn the parameters of a model particularly designed for short query search in CQA.

3. USER INTENT MINING

Understanding user intent is the key to improving the relevance of a short query in a CQA question search due to the heterogeneity of query and question. It is common for one short query to relate to multiple candidate questions. On the other hand, searchers are usually only interested in questions whose content matches their intent. In this paper, we aim to select questions that match a user's main search intent hidden behind short queries. Search intent can be represented via subtopics of queries, and the subtopics can be discovered by tracking user behavior from various sources. Basically speaking, after issuing a short query, most question searchers expect to see results that relate to two aspects of the query: (1) the most interesting or the most important aspect of the query. For example, "pyramids", "history", and "ancient" for query "egypt". (2) The most popular subtopic of the query. For example, "brotherhood" and "revolution" for "egypt". The two types of intent represent different user preferences related to their queries, and it is hard to mine them from a single source.

We mine user intent from three different sources: question descriptions in CQA archives, web search logs, and the top search results from a commercial search engine. Basically speaking, data from the three different sources contains different aspects of a query: the intent signals mined from descriptions reveal an asker's specific needs for a question; the web search log conveys common preferences about the query; and the top search results displayed by popular search engines such as Google, Bing, and Yahoo contain popular subtopics related to the query. For each type of intent, we extract several keywords and assign each keyword a weight.

3.1 Intent Mining from CQA Archives

In CQA archives, askers are allowed to supply supplementary descriptions to their questions. The descriptions further clarify the askers' information needs in addition to the questions. For example, for the question "Why do you love Baltimore?", the words "Maryland" and "Charm City" in the description are quite useful to distinguish which Baltimore city is asking about, since there are more than 10 cities named "Baltimore"[3]. Therefore, descriptions are particularly useful for discovering an asker's specific intent.

To extract intent from the descriptions, we treat the questions as sources and the corresponding descriptions as targets, and train a term-to-term translation model. We employ the method proposed by [18] to predict user intent for short queries. Formally, given a short query Q, we rank terms by $P_{cqa}(t|q)$, where $P_{cqa}(t|q)$ is defined as

$$P_{cqa}(t|q) = \varepsilon \sum_{w \in q} P_{tp}(t|w) P_{ml}(w|q) + (1-\varepsilon) P_{ml}(t|C). \quad (1)$$

[3] http://en.wikipedia.org/wiki/Baltimore_(disambiguation)

In Equation (1), $P_{tp}(t|w)$ represents the translation probability from term w in query q to term t. $P_{ml}(w|q)$ is the maximum likelihood for term w estimated from query q, and $P_{ml}(t|C)$ is the smoothing item estimated from a large corpus C. $0 \leq \varepsilon \leq 1$ is the parameter for a trade-off between the translation based likelihood and the smoothing item. In practice, we have found $\varepsilon = 0.3$ gives the optimal results.

With Equation (1), we rank the terms and select those with high probabilities as defined by $P_{cqa}(t|q)$ to represent the hidden intent. By this means we get the intent word set $W = \{(t, \varphi)\}$ from CQA archives, where t is the intent term and φ is the weight as defined by $P_{cqa}(t|q)$.

3.2 Intent Mining from Query Log

Most question searchers are also web users. Many searchers are accustomed to submitting their queries to search engines and browsing web pages before they search for answers in a CQA archive [11]. Therefore, user behavior tracked by a search engine or browser is another valuable source for understanding search intent in short queries.

In this paper, in addition to descriptions of questions, we also leverage large-scale query logs to mine user intent for short queries. Compared with those mined from question descriptions, intent mined from query logs represents more general information needs of common searchers. For example, for the query "beijing", the top intent mined from the query log is "travel". Therefore, we can guess that most searchers of "beijing" are interested in travel guides. On the other hand, the top intent mined from question descriptions is "olympics", which means that most askers would like to know something about the 2008 Summer Olympics in Beijing.

We mine user intent from a query log of a commercial search engine. Specifically, we employ the method proposed by Hu, et al. [13] and represent search intent behind short queries with subtopics. Given a query, the method collects queries that share the same suffix or prefix and aggregates the co-clicked URLs of these queries. After that, queries and URLs are clustered based on word overlap and the similarity of co-click patterns. Each cluster represents one subtopic of the original query and thus can be used to mine user intent. We extract intent from both the queries and URLs in the top subtopic cluster, since the top cluster represents a user's primary intent. For queries, we weight each suffix or prefix with click numbers and select the terms with high click numbers; For URLs, we first remove stop words and conduct stemming and then we calculate term frequencies and select the terms with high frequencies. After individual normalization of the numbers, we merge the terms from queries and URLs to get the intent word set $W = \{(t, \varphi)\}$, where t is the intent term and φ is the weight from queries or URLs. In this way, we leverage both the related queries and clicked documents to predict the pervasive intent of searchers.

3.3 Intent Mining from Web Search Results

With question descriptions in CQA archives and query logs, it is easy for us to mine some interesting subtopics related to short queries. However, the intent of some queries is time-sensitive. For these queries, understanding the current context of the intent is more important but hard to achieve with the sources mentioned above. For example, for the query "xbox", since Microsoft just announced the new console "Xbox ONE", questions about the new product may

Algorithm 1: Intent Mining from Web Search Results

Input: Query q, top M search results R, window size l, weight parameters η, σ, τ
Output: Intent term set $W = \{(t, \varphi)\}$

$H \leftarrow$ titles of documents in R;
$S \leftarrow$ snippets of documents in R;
$U \leftarrow$ URLs of documents in R;
$A \leftarrow$ concatenate T,S and U as a single string;
$L \leftarrow$ length of A;
$F \leftarrow \{(t, f)\} = \emptyset$;
for i: 1 to L **do**
  if $A[i] \in q$ **then**
    for j: $-l$ to l **do**
      if F contains key $A[i+j]$ **then**
        $F[A[i+j]] \leftarrow F[A[i+j]] + 1$;
      else
        $F \leftarrow F \cup (A[i+j], 1)$;
$W \leftarrow \{(t, \varphi)\} = \emptyset$;
for each $(t, \varphi) \in F$ **do**
  $\varphi \leftarrow 0$;
  for each $h \in H$ **do**
    $\varphi += \eta \cdot BM25(t, h, H)$;
  for each $s \in S$ **do**
    $\varphi += \sigma \cdot BM25(t, s, S)$;
  for each $u \in U$ **do**
    $\varphi += \tau \cdot BM25(t, u, U)$;
  $\varphi \leftarrow \varphi \cdot f$;
  $W \leftarrow W \bigcup \{(t, \varphi)\}$;
rank W in descending order of φ;
return W;

be more attractive than those about "Xbox 360". The intent "Xbox ONE" can be mined from the top search results (particularly, news about the Xbox), but it is hard to discern from query logs and question descriptions, since the Xbox 360 still dominates these sources.

Mining user intent from web search results has been studied by several researchers [1, 27]. The existing work adopts offline processing on web pages and thus is not able to capture the timely evolution of search intent on web pages. In this paper, we propose an online method to extract popular intent for short queries from the top search results.

Given a query, we crawl the newest search results from a commercial search engine and parse URLs, titles, and snippets of the top M documents. For each query term, we calculate the frequency of co-occurrence terms within a fixed window size in URLs, titles, and snippets, and then collect these terms and their frequencies to form an intent candidate set. For each term in the candidate set, we calculate its BM25 scores with the URLs, the titles, and the snippets. We heuristically set a weight for each field (URL/title/snippet) and define a final score as a weighted linear combination of BM25 scores in the three fields. Each term in the candidate set is ranked based on the final score, and we pick terms with high scores as user intent mined from web search results.

Algorithm 1 shows the pseudo code for mining intent that is currently relevant for short queries from top search results. We heuristically set $\{\eta, \sigma, \tau, M, l\} = \{0.4, 0.2, 0.1, 10, 3\}$ in experiments.

Table 1 gives examples of user intent mined from question descriptions, query logs, and top search results. From the examples, we can see that intent mined from the three sources are quite different. Take "usain bolt" for instance. The intent mined from descriptions is more specific and it shows that askers on CQA website care more about his characteristics, such as being the fastest man in the world, a world record owner, and an Olympic star. Intent mined from search logs is more general and shows that search engine users care more about personal information, such as his biography, twitter account, and his girlfriend. The intent mined from the search engine shows the recent achievements of Usain Bolt, such as wining three gold medals at the 2013 Moscow World Championship.

4. MODELS

We incorporate the intent mined from the three sources into language models and propose a new model to improve the search relevance for short queries in CQA.

4.1 Language Model for Information Retrieval

In recent years, the language model for information retrieval (LMIR) [23] has proven to be effective at question retrieval. Formally, given a query q and a candidate question Q, LMIR measures the relevance between q and Q through

$$P(q|Q) = \prod_{w \in q} [(1-\lambda) P_{ml}(w|Q) + \lambda P_{ml}(w|C)], \qquad (2)$$

where $P_{ml}(w|Q)$ represents the maximum likelihood of term w estimated from Q, and $P_{ml}(w|C)$ is a smoothing item that is calculated as the maximum likelihood in a large corpus C. The smoothing item avoids zero probability, which stems from the terms appearing in the candidate question but not in the query. $\lambda \in (0, 1)$ is a parameter that acts as a trade-off between the likelihood and the smoothing item.

LMIR performs well when there is a great deal of overlap between a query and a candidate question, but when the two present similar meanings with different words, LMIR fails to capture their similarity.

4.2 Translation-based Language Model

LMIR cannot measure the similarity between a query and a candidate question when there are lexical gaps. For example, when a query is "How do I get knots out of my cat's fur?" and a candidate question is "How can I remove a tangle in my cat's fur?", LMIR gives a very low similarity score. To address this issue, a translation-based language model [4] is proposed. Translation-based language models improve traditional LMIR by learning term-term or phrase-phrase translation probability from question-description or question-answer pairs and incorporating the information into maximum likelihood. Formally, given a query q and a candidate question Q, translation-based language is defined as

$$P_{trb}(q|Q) = \prod_{w \in q} \left[(1-\lambda) P_{mx}(w|Q) + \lambda P_{ml}(w|C)\right], \qquad (3)$$

where

$$P_{mx}(w|Q) = \alpha P_{ml}(w|Q) + \beta P_{tr}(w|Q)$$
$$P_{tr}(w|Q) = \sum_{v \in Q} P_{tp}(w|v) P_{ml}(v|Q).$$

Table 1: Examples of query intent words mined from three sources

Query	Intent words from CQA	Intent words from query log	Intent words from search results
usain bolt	fastest, world, record, olympics	biography, twitter, girlfriend	2013, gold, moscow, championship
superbowl	patriots, steelers, giants, nfl	story, history, 2012, ticket, xlvi	ticket, 2013, nfl, history, commercial
egypt	cairo, country, arabic, pyramids	morsi, election, ancient, history	revolution, brotherhood, police

Here λ, α, and β are parameters, satisfying $\alpha + \beta = 1$. $P_{tp}(w|v)$ represents the translation probability from term v in Question Q to term w.

Xue, et al. [28] further considered the likelihood from answers and proved some improvement when interpolating a translation-based language model with an extra answer likelihood term $P_{ml}(w|a)$ from the answer a of question Q. The model is defined as

$$P_{trba}(q|Q) = \prod_{w \in q} \left[(1-\lambda) P_{mx}(w|Q,a) + \lambda P_{ml}(w|C) \right], \quad (4)$$

where

$$P_{mx}(w|Q,a) = \alpha P_{ml}(w|Q) + \beta P_{tr}(w|Q) + \gamma P_{ml}(w|a).$$

γ is an extra parameter satisfying $\alpha + \beta + \gamma = 1$.

The existing models work well for long question queries. However, when an input query is short, these models cannot accurately measure the relevance between the query and the candidates due to the lack of intent information.

4.3 Intent-based Language Model

We are aiming for improved relevance of short queries in CQA question search. To this end, given a query q, we mine different types of search intent from various sources, as presented in the previous section. Each type of intent is represented as a series of terms associated with weights. We rank the terms in a descending order based on their weights and pick the top N terms for creating an intent-based language model. Formally, suppose that intent from source i is $W_i = \{(t_{ij}, \varphi_{ij})\}$, $1 \leqslant i \leqslant 3$, $1 \leqslant j \leqslant N$, where t_{ij} is the term, and φ_{ij} is the weight. Our intent-based language model is defined as

$$P_{ib}(q|Q) = \pi_0 P_{trba}(q|Q) + \sum_{i=1}^{3} \pi_i \sum_{j=1}^{N} \varphi_{ij} P_{trba}(t_{ij}|Q), \quad (5)$$

where P_{trba} represents the translation-based language model plus answer language model. $\{\pi_i\}_{0 \leq i \leq 3}$ are parameters that can be learned from training data. φ_{ij} is the weight of intent word t_{ij} in intent word set W_i from the source i. Equation (5) means that the relevance between a query q and a candidate Q is determined by not only the query itself, but also the underlying user intent hidden behind the query. The final model is a linear combination of the relevance from the original query and the relevance from the intent of the query.

Note that there is stark difference between our model and the one proposed by [10]. They focus on long question queries with clear information needs and their intent means different types of questions (objective/subjective/opinion), while we are addressing the relevance issue for short queries and therefore have to first predict what the possible intent is for the queries and then take the predictions into account when calculating relevance.

5. EXPERIMENT

We conducted experiments to test the mined intent and the intent-based language model.

5.1 Experiment Setup

5.1.1 CQA Data Preparation

We crawled a large number of questions associated with descriptions and answers from Yahoo! Answers using a public API[4] and from Quora.com using the approach described in [25].

Table 2 presents overviews of the two CQA data sets. 81% of the questions in Yahoo! Answers and 58% of the questions in Quora have descriptions and the average number of answers in the two data sets are 7.0 and 2.7, respectively. Besides the two large-scale CQA data sets, we also collected a one-year query log from a commercial search engine and randomly sampled 1,782 queries. 94.2% of the sampled queries were shorter than 4 words and the average length of the queries was 1.94. In other words, we considered the retrieval task of short queries in our experiments. Table 3 gives the length distribution of the sampled queries.

We separately indexed the crawled questions from Yahoo! Answers and Quora using an open source Lucene.Net System[5]. For each sampled query, we retrieved several candidate questions from the indexed data based on the inline-ranking algorithm in Lucene.Net and created two subsets. One subset consisted of query-question pairs with questions coming from Yahoo! Answers, and the other contained candidate questions from Quora. Each question was associated with its description and answers as metadata. Table 4 gives the statistics of the two subsets.

We recruited human judges to label the relevance of the candidate questions regarding the queries and compared different methods for the labeled data[6]. Specifically, we hired two judges. Each judge followed our guidelines and labeled a question with one of four levels: "Excellent", "Good", "Fair", and "Bad". If the judges disagreed on a question, we invited a third expert to make the final decision. Table 5 presents the specifics of our labeling guideline and label distributions. We randomly split each of the two labeled data sets into a training set, a validation set, and a test set with a ratio 2:1:1. We learned the parameters in the intent-based language model on the training set, tuned the parameters of the learning to rank algorithm on the validation set and tested the performance of our model and other baseline models on the test set. The training set and the validation set are also used for tuning parameters in baseline models.

To evaluate the performance of different models, we calculated NDCG [14] at positions 1, 3, and 5.

[4] http://developer.yahoo.com/answers/

[5] http://lucenenet.apache.org/

[6] The labeled data from Yahoo! Answers is available at http://home.ustc.edu.cn/~ustcwhc/

Table 2: Overview of two CQA data sets

Statistics	Yahoo	Quora
Question #	127,787,139	649,843
Description #	103,605,696	375,829
Answer #	894,855,746	1,743,259

Table 3: Length distribution of the labeled queries

Total	Query Length				Avg.
	1	2	3	≥4	Length
1782	658	732	289	103	1.94

5.1.2 Data Sources for Intent Mining

There are three different sources for intent mining: question descriptions in CQA, query logs, and the top search results of a web search engine. We first employed the method in [18] and trained two translation models using the crawled 103.6M and 375.8K question-description pairs of Yahoo! Answers and Quora, respectively. Secondly, we collected one-year (May 1st, 2012 - April 30th, 2013) query logs from a commercial search engine and employed the method in [13] to mine user intent. Finally, for the online intent prediction from web search results, we used Bing API[7], crawled top 10 documents for each query, and distilled user intent following Algorithm 1. In the experiments on both Yahoo data and Quora data, we used the same intent signals from search logs and search results, and individual intent signals from each question-description translation model. For each source, we selected the top intent words, which were three times as the length of the input query.

5.1.3 Baselines

We first employed four variants of language models as baseline methods: the traditional language model for information retrieval (LMIR) given by Equation (2) [23], translation model (TR) [4] given by Equation (3), translation-based language model (TBL) [15], and translation-based language model plus answer language model (TAL) given by Equation (4) [28]. For the translation-based models (TR, TBL, TAL), we individually trained question-answer translation models on Yahoo! Answers and Quora with the pooled approach proposed by Xue, et al. [28], which creates a parallel corpus from both question-to-answer pairs and answer-to-question pairs. To enhance efficiency, we sampled 10% of the pairs from Yahoo data in the translation model learning. The total number of pairs were 179M for Yahoo! Answers and 3.4M for Quora.

In addition to these models, we also considered the method proposed by Li, et al. [18] as a baseline, as it is among the state-of-the-art methods leveraging user intent in question retrieval. This method measures question-question similarity with an LDA-based model [9]. In this method, the description of a query question and a user's information needs are concatenated along text denoted as D_i, while the description of a candidate question is denoted as D_j. Question similarity is defined as the similarity of D_i and D_j calculated using a KL divergence on topic distributions. In our experiments, we concatenated each question and its description as a document. We trained LDA models [5] on 6.4M sampled documents from Yahoo! Answers and the whole 650K documents from Quora using GibbsLDA++[8], and heuristically set the topic number in the training to be 26 and other parameters to be default for both data sets.

For a fairer comparison, we linearly combined the top-performing language model TAL and the LDA-based model and denoted the model as TAL+LDA. The combination weights were learnt from training data using LambdaMART [6].

[7] http://datamarket.azure.com/dataset/bing/search

Table 4: Overview of two labeled data sets

Statistics	Yahoo	Quora
Queries #	1,782	1,782
Questions #	12,947	13,739
Quesitons #/query	7.27	7.71

Table 5: Labeling guideline and label distribution

Label	Yahoo	Quora
Excellent	15.1%	1.4%
Meaning: the question is quite related to the query, and would have pleased the searcher or made him or her want to follow up or explore further.		
Good	29.2%	33.7%
Meaning: the question is related to the query, and the searcher is likely to have found it interesting.		
Fair	42.2%	51.8%
Meaning: the question is related to the query, but is not likely to be interesting to the searcher.		
Bad	13.3%	13.1%
Meaning: the question is not related to the query.		

5.1.4 Parameter Tuning

Language models have smoothing parameters (e.g., {λ, α, β, γ} in TAL). We selected the best parameters in {0, 0.1, 0.2,. . . , 1} from training and validation sets. We found that the optimal parameter combinations of TAL are {0.7, 0.1, 0.1, 0.8} on Yahoo data and {0.3, 0.1, 0.5, 0.4} on Quora data.

We used RankLib[9] to train a LambdaMART [6] for learning parameters in both TAL+LDA and our intent-based language model. There are four parameters in LambdaMART: {nt, nl, lr, mil}, which stand for the number of trees, the number of leaves, the learning rate, and the minimum instances per leaf, respectively. We chose nt from {10, 50, 100}, nl from {2, 4, 16}, lr from {0.1, 0.2, 0.3, 0.4}, and mil from {10, 50, 100} for each model on the validation data.

5.2 Results

We denote our intent-based language model given by Equation (5) as IBLM. Table 6 reports the experimental results on the two data sets. From Table 6 we can see that IBLM outperformed all the baseline methods on both Yahoo data and Quora data for all metrics, and the improvements were statistically significant(sign-test, p value < 0.01). The results demonstrate both the usefulness of the intent mined from various sources and the efficacy of the proposed model on leveraging the mined intent for improving search relevance of short queries.

[8] http://sourceforge.net/projects/gibbslda/

[9] http://people.cs.umass.edu/~vdang/ranklib.html

Table 6: Evaluation results on Yahoo data and Quora data

	Yahoo data			Quora data		
	NDCG@1	NDCG@3	NDCG@5	NDCG@1	NDCG@3	NDCG@5
LDA[18]	63.35	69.91	72.50	55.72	61.03	65.74
LMIR[23]	67.02	73.83	76.07	60.72	68.06	72.39
TR[4]	68.27	73.84	75.98	60.52	67.87	71.97
TBL[15]	68.13	73.89	76.37	60.97	68.28	72.46
TAL[28]	69.44	74.95	76.53	61.96	67.11	72.48
TAL+LDA	70.03	74.02	75.92	62.71	68.54	72.93
IBLM	**71.33**	**77.12**	**77.70**	**64.04**	**69.98**	**74.14**

Table 8: Evaluation results of different intent models

(a) Yahoo data

	NDCG@1	NDCG@3	NDCG@5
TAL [28]	69.44	74.95	76.53
TAL+TAL(W_{cqa})	70.33	75.55	76.85
TAL+TAL(W_{log})	70.25	75.60	**77.23**
TAL+TAL(W_{web})	**71.14**	**75.94**	77.17

(b) Quora data

	NDCG@1	NDCG@3	NDCG@5
TAL [28]	61.96	67.11	72.48
TAL+TAL(W_{web})	63.31	69.21	73.51
TAL+TAL(W_{log})	63.70	69.00	**73.73**
TAL+TAL(W_{cqa})	**64.02**	**69.24**	73.65

5.3 Discussions

We investigated the reasons why IBLM can outperform baseline methods. Table 7 gives an example. The query "priceline" represents a commercial web site that provides users discount rates for travel-related purchases such as airline tickets and hotel stays[10]. For this query, questions about hotels or tickets are more relevant than others. Therefore, given two questions "Is Priceline the cheapest way to buy hotel rooms online?", and "Why did Priceline acquire Kayak for $1.8 billion?", it is better to rank the former in a higher position than the latter. The best performing baseline TAL is not aware of the underlying intent in the query and thus cannot do a good job of ranking. In contrast, our model can leverage the intent mined from different sources and successfully identity the searcher's preference.

5.3.1 Comparison between Intent from Different Sources

We also studied the different effects of intent mined from different sources. In this paper, we mined user intent from question descriptions in CQA archives, query logs, and web search results. To compare different intents, we calculated TAL models using the intent words mined from an individual source and linearly combined them with the original TAL model. The combination weights were learned using LambdaMART on training data, and the parameters of LambdaMART were tuned in the same way with IBLM. We denoted the three individual intent models as TAL+TAL(W_{cqa}), TAL+TAL(W_{log}) , and TAL+TAL(W_{web}) , which means using intent mined from CQA archives, query logs, and web search results, respectively. Table 8 reports the results of individual intent models. We can see that all individual models improved the original TAL model and the improvement was statistically significant (sign test, $p < 0.01$).

TAL+TAL(W_{cqa}) leverages intent mined from question descriptions. The intent reflects askers' specific needs when asking questions about a short query. Query 1 in Table 9 gives an example that demonstrates the effect of considering intent mined from question descriptions. The query "Baltimore" refers to more than 10 cities and therefore was ambiguous[11]. On the other hand, many askers were interested in Baltimore, Maryland, which is the most famous Baltimore and expressed their specific needs through descriptions after their questions. We mined the intent and took it into account in our model. Therefore, our model TAL+TAL(W_{cqa}) can successfully rank the question about Baltimore in Maryland in a high position and fulfill the user's information needs.

TAL+TAL(W_{log}) makes use of intent mined from query logs. User intent is represented by query subtopics and co-clicked URLs that reflect searchers' pervasive preference towards a short query. Query 2 in Table 9 gives an example. "Gangnam style" was a popular song in 2012, and many users searched for "gangnam style video" and "psy gangnam style" on search engines. Therefore, questions about videos for Gangnam Style or the singer of Gangnam Style may attract more attention from searchers. We mined this common intent from query logs and with the information our model successfully ranked the question "What meaning is intended to be conveyed by the different scenes in the Gangnam Style **video**?" with the label "Excellent" in a high position.

The TAL+TAL(W_{web}) model leverages the top search results from a search engine to predict user intent. Modern search engines usually mix recent material such as news and general content such as Wikipedia pages in the top retrieval results. The diverse search results allowed us to mine both recent aspects and general aspects related to a short query. This makes intent mined from web search results particularly useful for improving the search relevance of time-sensitive queries such as sports events and new products. Query 3 in Table 9 gives an example. Every year, the Superbowl is the most popular sporting event for football fans in America, and every year there are new things happening in at that year's Superbowl. When a user issues this query, it is very likely that he or she would like to see questions about the most recent or upcoming superbowl rather than those about historical superbowls. Our model captured this information need (i.e., superbowl 2013) from top search results and pro-

[10] http://www.priceline.com/

[11] http://en.wikipedia.org/wiki/Baltimore_(disambiguation)

Table 7: Example for comparison between IBLM and the best performing baseline TAL

Query 1	priceline			
Intent Words	From three sources: ***hotel***, ***cheap***, ***bid***, flight, ticket			
Label	TAL	IBLM	Question	Answer
Excellent	Rank 3	Rank 1	Is Priceline the ***cheapest*** way to buy ***hotel*** rooms online?	It's a very good way to get a good rate if you have the time/patience to do re-***bids*** and have the flexibility in location of your stay.....don't know the exact ***hotel*** you're getting...suppose y-ou want a 5* ***hotel*** in Beverly Hills...that LAX doesn't have a 5* ***hotel***
Fair	Rank 1	Rank 3	Why did Priceline acquire Kayak for $1.8 billion?	...joint industry relations for better air and ***hotel*** deals...

vided question searchers with the most recent and relevant results[12].

It is interesting to note that TAL+TAL(W_{web}) performs best on Yahoo! Answers data (See Table 8(a)), while on Quora data the best performing model is TAL+TAL(W_{cqa}) (See Table 8(b)). On the one hand, Yahoo! Answers is the largest CQA website, and there are a large number of active users on it. It is easier for recent content to emerge on Yahoo! Answers. Therefore using a mixture of recent intent and general intent mined from web search results can lead to the best performance on Yahoo data. On the other hand, data in Quora has higher quality than data in Yahoo! Answers. Users in Quora are better educated and more capable of presenting their questions, descriptions, and answers in clear statements [25]. Therefore, there is less noise in descriptions in the Quora data, which leads to better intent understanding through question-description translation. Thus, it is reasonable that the TAL+TAL(W_{cqa}) model would achieve the best performance on Quora data.

5.3.2 *Unsupervised Intent-based Model*

In our intent-based language model (see Equation (5)), each intent word is used to calculate a new TAL model and the final model is a linear combination of all the TAL models. An alternative for leveraging the intent could be concatenating the intent words with the original query and performing retrieval using the new, longer query in TAL. This renders an unsupervised intent-based language model. An interesting question is whether the unsupervised model performs as well as or even better than IBLM. In this section, we compared the unsupervised alternative with the supervised model and the best performing baseline, TAL model.

Figure 1 presents a comparison of the supervised model, the unsupervised model, and the baseline model across intent mined from all three sources. We denote the unsupervised alternative as TAL(Q+W_{cqa}), TAL(Q+W_{log}), and TAL(Q+W_{web}). From the experimental results, we can see that the unsupervised intent-based models performed the worst. This may stem from the fact that in the unsupervised models all intent words and the original query were equally weighted and the concatenation drifts the topic of a user's query and makes its intent more ambiguous. On the other hand, the supervised models can leverage the labeled data and calibrate the effect of different intent words in measuring the relevance between the original query and candidate questions.

[12]We conducted the experiments when Superbowl 2013 was still ongoing.

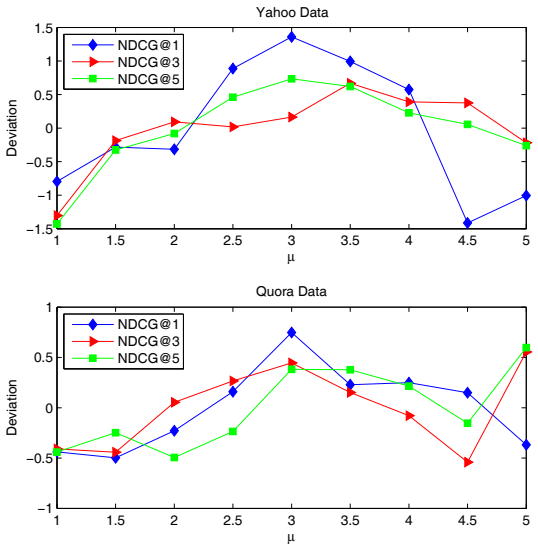

Figure 2: The influence of μ on the performance of IBLM

5.3.3 *Influence of Intent Word Number*

In our experiments, we fixed the number of intent words at three times the length of the input query. In practice, it is worth understanding the influence of the intent word number to the final relevance. Intuitively, too few intent words may lack the power of disclosing underlying search intent while too many intent words may bring noise. To verify this intuition, we defined a parameter μ as the ratio of intent word number and query length and selected it from $\{1, 1.5, 2, 2.5, ..., 5\}$. Figure 2 reports the performance of IBLM with different numbers of intent words on the two data sets. Note that in Figure 2, the vertical axis represents the deviation of the NDCG@i, which is defined as NDCG@i - Avg(NDCG@i). We employed this measure to highlight the gradual change of NDCG with respect to μ. The experimental results verified our intuition and indicated that $\mu = 3$ is the best choice for the number of intent words.

Table 9: Examples for comparison between three variants of IBLM and TAL model

Query 1	baltimore			
Intent Words	From question-description tranlsation model: ***maryland***, ***city***, ***area***, raven, harbor			
Label	TAL	TAL + TAL(W_{cqa})	Question	Answer
Excellent	Rank 3	Rank 1	What do people like about Baltimore?	...most affordable truly urban ***areas*** in the United States...the cheapest "big" ***city*** in the north east corridor...in many ways by the ***Maryland*** Institute College of Art and its alumni...you would pay in another major ***city***...
Fair	Rank 1	Rank 4	Is Baltimore considered in the North or the South?	Being right near the border, Baltimore gets some of both: northern charm and southern efficiency.
Query 2	gangnam style			
Intent Words	From query search log: youtube, ***video***, ***psy***, oppa, lyrics, dance			
Label	TAL	TAL + TAL(W_{log})	Question	Answer
Excellent	Rank 4	Rank 1	What meaning is intended to be conveyed by the different scenes in the Gangam Style ***video***?	... The ***video*** is a satire about Gangnam...I think a lot of what ***Psy*** is pointing out is...The Subversive Message Within South Korea's Music ***Video*** Sensation...
Fair	Rank 1	Rank 7	What are some similarities and differences between Gangam Style and Kolaveri Di?	Blind following!!
Query 3	superbowl			
Intent Words	From web search engine: ticket, ***2013***, ***nfl***, ***history***, commercial			
Label	TAL	TAL + TAL(W_{web})	Question	Answer
Good	Rank 3	Rank 1	As of January 14, ***2013***, which teams do you predict for the Superbowl ***2013*** and who will win it?	...one of the ***NFL***'s deadliest weapons...The most complete team in the ***NFL*** is going to destroy Atlanta if they...three most explosive single-season offenses in ***NFL History***.
Fair	Rank 1	Rank 9	Which member of the QB class of 2012 will win a Super-Bowl first?	Put me down for Russell Wilson. I think Andrew Luck is still likely to have the best career of the three,...

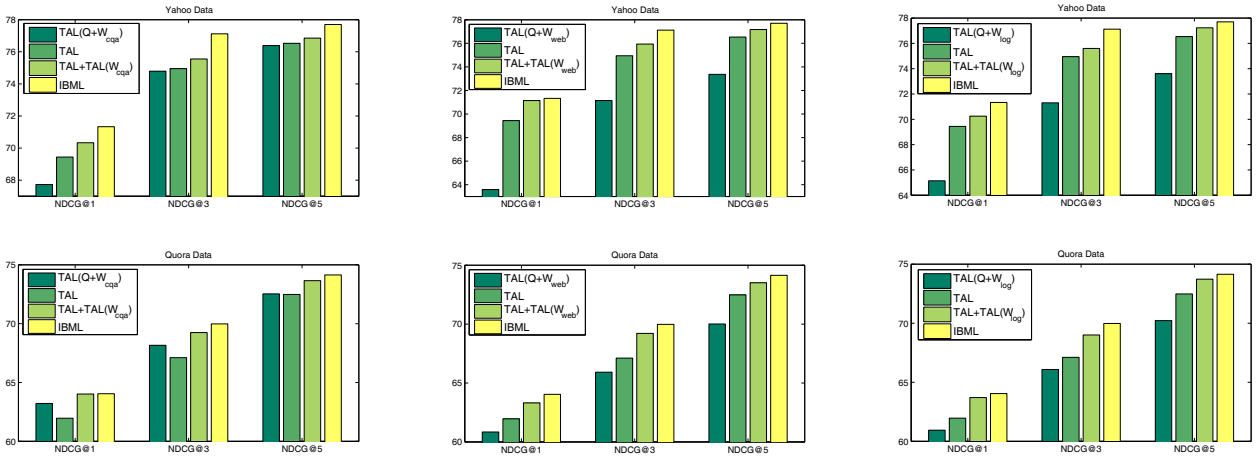

Figure 1: Comparisons between unsupervised IBLM, TAL and supervised IBLM

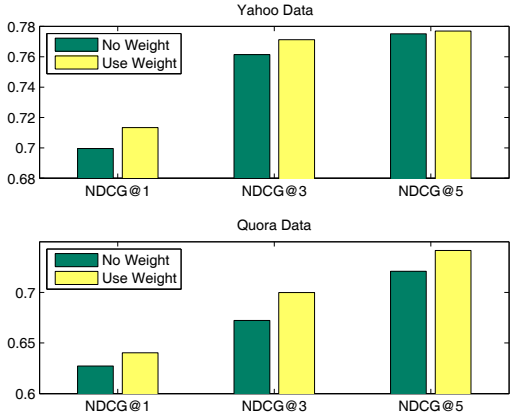

Figure 3: The influence of intent word weight on the performance of IBLM

5.3.4 *Influence of Intent Word Weight*

In Equation (5), we not only leverage the content of intent but also consider the weight of different intent words. An interesting question is whether the intent word weight really contributes to the final relevance. In this section, we set each weight φ_{ij} as 1 and compared it with IBLM.

Figure 3 shows the results when using and not using intent word weight. We can see that weighted IBLM performs better over all the metrics. This is because the weight encodes user behavior such as click numbers and thus can provide extra information other than intent content.

6. CONCLUSION

In this paper, we have investigated the problem of question retrieval for short queries in community question answering. We propose an intent-based language model that takes advantage of both the state-of-the-art question retrieval models and the extra intent information mined from three data sources. The parameters of the model are automatically learned with a state-of-the-art learning-to-rank approach. We empirically study the efficacy of the proposed model on data sets from Yahoo! Answers and Quora. The evaluation results show that with user intent prediction, our model can significantly improve state-of-the-art relevance models on question retrieval for short queries.

7. ACKNOWLEDGMENT

We would like to thank Fengze Jiang for his help in drawing the figures.

8. REFERENCES

[1] Bandyopadhyay, A., Ghosh, K., Majumder, P., and Mitra, M. Query expansion for microblog retrieval. *IJWS 1*, 4 (2012), 368–380.

[2] Bendersky, M., Metzler, D., and Croft, W. B. Learning concept importance using a weighted dependence model. In *WSDM* (2010), pp. 31–40.

[3] Bendersky, M., Metzler, D., and Croft, W. B. Parameterized concept weighting in verbose queries. In *SIGIR* (2011), pp. 605–614.

[4] Berger, A. L., Caruana, R., Cohn, D., Freitag, D., and Mittal, V. O. Bridging the lexical chasm: statistical approaches to answer-finding. In *SIGIR* (2000), pp. 192–199.

[5] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. In *NIPS* (2001), pp. 601–608.

[6] Burges, C. From ranknet to lambdarank to lambdamart: An overview. *Microsoft Research Technical Report* (2010).

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

[8] Cao, X., Cong, G., Cui, B., Jensen, C. S., and Zhang, C. The use of categorization information in language models for question retrieval. CIKM '09, ACM, pp. 265–274.

[9] Celikyilmaz, A., Hakkani-Tur, D., and Tur, G. Lda based similarity modeling for question answering. In *NAACL HLT Workshop on Semantic Search* (2010), pp. 1–9.

[10] Chen, L., Zhang, D., and Levene, M. Question retrieval with user intent. In *SIGIR* (2013), pp. 973–976.

[11] Dror, G., Maarek, Y., Mejer, A., and Szpektor, I. From query to question in one click: suggesting synthetic questions to searchers. In *WWW* (2013), pp. 391–402.

[12] Duan, H., Cao, Y., Lin, C.-Y., and Yu, Y. Searching questions by identifying question topic and question focus. In *ACL* (2008), pp. 156–164.

[13] Hu, Y., nan Qian, Y., Li, H., Jiang, D., Pei, J., and Zheng, Q. Mining query subtopics from search log data. In *SIGIR* (2012), pp. 305–314.

[14] Järvelin, K., and Kekäläinen, J. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR* (2000), pp. 41–48.

[15] Jeon, J., Croft, W. B., and Lee, J. H. Finding similar questions in large question and answer archives. In *CIKM* (2005), pp. 84–90.

[16] Lafferty, J. D., and Zhai, C. Document language models, query models, and risk minimization for information retrieval. In *SIGIR* (2001), pp. 111–119.

[17] Li, H. A short introduction to learning to rank. *IEICE Transactions 94-D*, 10 (2011), 1854–1862.

[18] Li, S., and Manandhar, S. Improving question recommendation by exploiting information need. In *ACL* (2011), pp. 1425–1434.

[19] Liu, Q., Agichtein, E., Dror, G., Gabrilovich, E., Maarek, Y., Pelleg, D., and Szpektor, I. Predicting web searcher satisfaction with existing community-based answers. In *SIGIR* (2011), pp. 415–424.

[20] Liu, T.-Y. *Learning to Rank for Information Retrieval*. Springer, 2011.

[21] Metzler, D., and Croft, W. B. A markov random field model for term dependencies. In *SIGIR* (2005), pp. 472–479.

[22] Minsky, M. L. *Semantic Information Processing*. The MIT Press, 1969.

[23] Ponte, J. M., and Croft, W. B. A language modeling approach to information retrieval. In *SIGIR* (1998), pp. 275–281.

[24] Robertson, S. E., and Walker, S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR* (1994), pp. 232–241.

[25] Wang, G., Gill, K., Mohanlal, M., Zheng, H., and Zhao, B. Y. Wisdom in the social crowd: an analysis of quora. In *WWW* (2013), pp. 1341–1352.

[26] Wang, K., Ming, Z., and Chua, T.-S. A syntactic tree matching approach to finding similar questions in community-based qa services. In *SIGIR* (2009), pp. 187–194.

[27] Wang, Q., nan Qian, Y., Song, R., Dou, Z., Zhang, F., Sakai, T., and Zheng, Q. Mining subtopics from text fragments for a web query. *Inf. Retr. 16*, 4 (2013), 484–503.

[28] Xue, X., Jeon, J., and Croft, W. B. Retrieval models for question and answer archives. In *SIGIR* (2008), pp. 475–482.

[29] Zhou, G., Cai, L., Zhao, J., and Liu, K. Phrase-based translation model for question retrieval in community question answer archives. In *ACL* (2011), pp. 653–662.