Capturing the Semantics of Key Phrases Using Multiple Languages for Question Retrieval

Wei-Nan Zhang, Zhao-Yan Ming, Yu Zhang, Ting Liu, and Tat-Seng Chua

Abstract—In the age of Web 2.0, community user contributed questions and answers provide an important alternative for knowledge acquisition through web search. Question retrieval in current community-based question answering (CQA) services do not, in general, work well for long and complex queries, such as the questions. The main reasons are the verboseness in natural language queries and the word mismatch between the queries and the candidate questions in the CQA archive during retrieval. To address these two problems, existing solutions try to refine the search queries by distinguishing the key concepts in the queries and expanding the queries with relevant content. However, using the existing query refinement approaches can only identify the key and non-key concepts, while the differences between the key concepts are overlooked. Moreover, the existing query expansion approaches, not only overlook the weights of key concepts in the queries, but also fail to consider concept level expansion for them. In this paper, we explore a key concept identification approach for query refinement and a pivot language translation based approach to explore key concept paraphrasing. We further propose a new question retrieval model which can seamlessly integrate the key concepts and their paraphrases. The experimental results demonstrate that the integrated retrieval model significantly outperforms the state-of-the-art models in question retrieval.

Index Terms—Key concept paraphrasing, query/question expansion, question retrieval

1 Introduction

Community Question Answering (CQA) services have emerged as popular alternatives for online information acquisition, such as Yahoo! Answers,[1] WikiAnswers[2] and Baiduzhidao,[3] etc. According to Google Trends,[4] all the above three CQA services had more than 10 million searches and visits in 2011. Over times, a huge amount of high quality question and answer (QA) pairs has been accumulated as comprehensive knowledge bases of human intelligence. It helps users to seek precise information by obtaining correct answers directly, rather than browsing through large ranked lists of results. Hence to retrieve relevant questions and their corresponding answers becomes an important task for information acquisition. Here we define question retrieval in CQA services as a task in which new questions are used as queries to find relevant questions for which the answers are already available. For simplicity and consistency, we use the term "query" to denote new questions posed by users and "question" to denote those answered questions available in the CQA archives.

Question retrieval in CQA is different from general Web search [1]. Unlike the Web search engines that return a long list of ranked documents, question retrieval returns several relevant questions with possible answers directly. Meanwhile, question retrieval can also be considered as a traditional Question Answering (QA) problem, but the focus of the QA task is transformed from answer extraction , answer matching and answer ranking to searching for relevant questions with good ready answers [2].

One major challenge is the word verboseness in the queries where important words may be surrounded by other additional words. As Park and Croft [3] described, these additional words are more likely to confuse the current search engines rather than help them. For example, in a query: *"Why are you less likely to catch a cold or flu in spring summer and autumn than winter months ?"*, some of the words are key terms for question retrieval, such as *"catch a cold"* and *"winter months"*, some of them are complementary words which are less important and may cause confusions for retrieval models, such as *"spring summer and autumn"*. The other major challenge is the word mismatch between the queries and the candidate questions for retrieval. For example, *"Why do people get colds more often in lower temperature?"* and *"Why are you less likely to catch a cold or flu in spring summer and autumn than winter months?"* are relevant to each other, but the same meaning is represented with different word forms, such as *"get colds"* and *"catch a cold"*. This makes it difficult for the two questions to be matched in the question retrieval task. In applications based on user generated content (UGC), such as CQA services, where the users tend to use a more diverse and informal vocabulary to express their information needs,

[1] http://answers.yahoo.com/
[2] http://wiki.answers.com/
[3] http://zhidao.baidu.com/
[4] http://www.google.com/trends/

- *W.-N. Zhang, Y. Zhang, and T. Liu are with the Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China 150001.*
- *Z.-Y. Ming is with the Department of Computer Science, Digipen Institute of Technology, Singapore. E-mail: mingzhaoyan@gmail.com.*
- *T.-S. Chua is with the School of Computing, National University of Singapore, Singapore 117417.*

Manuscript received 21 Mar. 2014; revised 3 Nov. 2015; accepted 4 Nov. 2015. Date of publication 23 Nov. 2015; date of current version 3 Mar. 2016.
Recommended for acceptance by N. Ramakrishnan.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2015.2502944

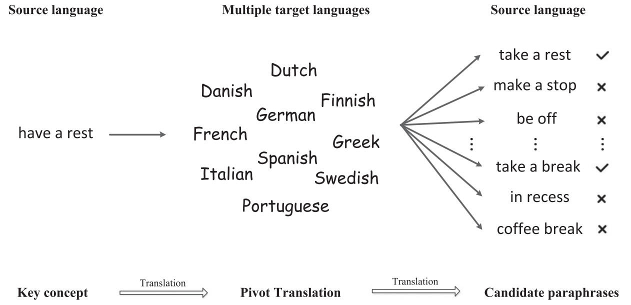

Fig. 1. An example of using multiple languages to obtain the paraphrases of a key concept by using pivot translation.

the word mismatch problem is even more common and severe than in general search.

In order to solve the word verboseness in queries, previous work mainly focused on core term discovery [4], query reformulation [5], key concept identification [6] on verbose queries, etc. Despite the great success achieved, these papers mainly focused on distinguishing the key concepts from the non-key ones and the importance among the key concepts was not taken into consideration. In this paper, we propose a ranking based method for key concept identification, which not only distinguishes the key concepts from the non-key ones, but also captures the differences among key concepts.

To tackle the word mismatch problem, previous work mainly resorts to query expansion [7], [8]. However, the former approach overlooks concept level evidences for query expansion and the latter approach fails to assign explicit weights to the expanded aspects. Jeon et al. [9] compared four different retrieval models and revealed that the translation model (TM) achieved the best performance. Xue et al. [1] combined the language model (LM) and translation model to a translation based language model (TLM) and further improve the performance of question retrieval. However, both of them are based on the term level expansion and the dependence between terms is not considered. Zhou et al. [10] employed the phrasal translation model to capture the contextual information for question retrieval. However, the phrase based translation model makes little or no direct use of syntactic information, which leads to the limitation on the translation performance and further impact the question retrieval results.

Overall, we adopt three approaches to tackle the verboseness and word mismatch problems in question retrieval from CQA archives. First, we utilize a pivot language translation approach to explore key concept paraphrases in the queries from multiple language resource. We try to obtain key concept paraphrases as semantic expansions to bridge the lexical gaps among different concept forms with same meaning. Fig. 1 presents an example of using multiple languages to obtain the paraphrases of a key concept by using pivot translation.[5]

Second, we estimate the weights of key concept paraphrases by considering two issues. One is based on the paraphrase generation probabilities which can be obtained when utilizing a pivot language translation approach. The other is based on the statistical distribution of paraphrases in the Q&A repository, which reflects the importance of the given concept paraphrases over the whole data set. Third, we propose a novel probabilistic retrieval model which can successfully integrate the state-of-the-art question retrieval model, key concept model and key concept paraphrase model to achieve better performance. The contributions of this work are three-fold:

- To the best of our knowledge, this is the first thorough study of using multiple languages to bridge the semantic gaps in question retrieval task.
- Second, we propose a ranking-based approach to capture the importance levels of key concepts in the target questions, which significantly outperforms the state-of-the-art binary classification approach.
- Third, we demonstrate the usability of the paraphrase model to be compatible with existing question retrieval models, and show that it contributes additional semantic connection among the key concepts in the query and the retrieved questions.

2 Related Work

2.1 Key Concept Detection

Turney [12] first proposed a genetic based classification approach to automatically extracting key words or key phrases for academic journal articles. They also compared the performance of C4.5 classifier and the proposed genetic approach and verified that the proposed approach outperformed the C4.5 classifier. Later, Hulth [13] proposed a classifier to further improve the key words extraction performance in the abstracts of the academic articles. They adopted the linguistic information, such as part-of-speech tags, syntactic and NP-chunk etc., as features for the key words extraction. Allan et al. [4] employed several linguistic and statistic approaches to identify core terms in TREC $\langle desc \rangle$ queries. They then verified the improvements on information retrieval task. Callan et al. [5] further converted the $\langle desc \rangle$ query to the structured INQUERY query by using the noun phrases, named entity recognition, exclusionary constraints and proximity operators. They also verified the structured INQUERY query improved the performance of information retrieval. Despite the success of the above work on key words or key phrases extraction, the key concept identification on UGC data has huge difference to that on academic papers and TREC queries. For example, the UGC data has lots of informal expressions, verbose description, non word symbols, etc.

Bendersky and Croft [6] used the AdaBoost M1 meta classifier with the C4.5 decision tree approach to distinguishing key concepts from non-key ones. They then implemented the proposed approach in the Indri query language for information retrieval in. Recently, Bendersky and Croft [14] used the hypergraph model to estimate the concept dependencies in arbitrary queries. The concept dependencies were then applied to impact the term weighting of the arbitrary queries and then were used to improve the

[5] In real deployment, it is easy to obtain the multilingual parallel corpora from the World Wide Web. Inspired by the previous approach [11], we can mine multilingual parallel text from the Web, which includes three steps, namely, locating the web pages that might have parallel translations, generating the candidate pairs that might be translations and structural filtering of the non-translation pairs. Hence, our proposed approach can adapt to the practical applications.

performance of the ranking model on information retrieval. However, the former neglected the correlation of terms in query and the later only explored the relevance of concepts rather than discovering key concepts in query. In this paper, we propose a ranking based method to identify key concept in UGC data so that to explore the correlation of terms in query for question retrieval.

2.2 Query Expansion

An effective method to tackle the word mismatch problem in information retrieval is query expansion. Cui et al. [15] proposed a correlation based query expansion method to extract expansion terms from search log data. The extracted terms were then integrated into the original query in a unified ranking model to improve the performance of Web search. Xu and Croft [8] analyzed the documents which are retrieved by the initial query as the local information. They then explored the word relations in the whole corpus as global information. Finally they combined the local and global information as the expansion of query for information retrieval task. However, both of the two approaches on query expansion are totally based on the statistical information and the semantic information of terms are neglected.

Buscaldi et al. [16] utilized WordNet[6] as a semantic dictionary to capture the similarities between terms in queries and candidate documents. The similarity of terms were computed by the distance in the WordNet tree structure. However, the low coverage, labor-intensive and non-timely nature makes these semantic dictionaries difficult to adapt to information retrieval on UGC, such as question retrieval in CQA services. Riezler et al. [17] adopted the monolingual translation model to capture terms similarities between questions and their corresponding answers. The translated question terms thus can be seen as the expansion terms for query. Recently, Gao and Nie [18] extended the latent concept expansion model for query expansion using search engine query logs. There are many other query expansion methods that have been proposed for IR, a comprehensive review can be found in [19].

Despite the success of previous work, literature regarding the concept level query expansion by automatically exploring the semantic information of concept from UGC data is still sparse. In this paper, we propose a pivot language translation approach, which compensates for the existing paraphrasing research in a suitable granularity, to exploit concept level paraphrases as expansions for question retrieval.

2.3 Pivot Language Approach to Paraphrasing

Bannard and Callison-Burch [20] employed pivot language translation approach to extract paraphrases from bilingual parallel corpora. The so called pivot translation approach is indeed a bi-direction translation process between the source and the target languages. To eliminate the syntactic errors when identifying paraphrases, Callison-Burch [21] further improved the pivot language translation approach by adding syntactic constraint. The constraint can be described as that only the extracted concepts which have the same syntactic roles can be the paraphrase candidates. Zhao et al. [22] extended this approach to generate richer phrasal paraphrase patterns from bilingual parallel corpora. The dependency parsing features were utilized on the English corpus. The extracted paraphrase patterns can be further instantiated as phrase paraphrases. Tomuro [23] introduced a rule based method to derive a set of interrogative paraphrase patterns for question paraphrases recognition. However, sentence-level paraphrasing still faces challenges that are not easy to tackle, such as the deep understanding of complex sentences and sophisticated syntactic, semantic and contextual processing to generate the equivalent candidates.

In this paper, we adapted the state-of-the-art approach on phrasal paraphrase generation to the UGC data and effectively integrated the paraphrase into a unified ranking model for question retrieval.

2.4 Question Retrieval

Berger et al. [24] introduced statistical approaches to bridging the lexical gap in FAQ retrieval. They inspected a collection of answered questions and characterize the relation between question and answer with a statistical model. Riezler et al. [17] utilized a monolingual translation based retrieval model for answer retrieval. They introduced sentence level paraphrasing technique to capture lexical similarities between questions and answers. Duan et al. [25] first detected question topic and focus by using a tree cut method. They then proposed a new language model to capture the relation between question topic and focus for question retrieval. Jeon et al. [9] compared four different retrieval models, i.e., VSM, BM25, LM and translation model for question retrieval in CQA archives. Experimental results reveal that the translation model outperforms the other models. Xue et al. [1] combined the language model and translation model to a translation based language model and obtain better performance in question retrieval. Following that, Wang et al. [26] proposed a syntactic tree matching model to finding similar questions, and demonstrated that the model is robust against grammatical errors. Bernhandrd and Gurevych [27] utilized the monolingual parallel corpora, which are collected from the WikiAnswer website, the definitions and glosses of the same term in different lexical semantic resources, to train the translation model for question retrieval. Cao et al. [2] proposed the category smoothing based and question classification based approaches to enhance the performances of existing question retrieval models. Na and Ng [28] proposed a multilingual translation model to enrich the document representation for information retrieval. They first done the word level translation without considering the word ordering information and then constructed the phrase units using the word translations. However, this process may cause serious ambiguities on phrase translation. Singh [29] focused on identifying the entities in questions and integrated them to improve the performance of question retrieval. Recently, Zhou et al. [30] extended the word embedding [31] with metadata to further improve the question retrieval result. They encoded the question category information into the continuous word embedding and further obtained a better representation of sentence using the Fisher Vector.

[6] http://wordnet.princeton.edu/

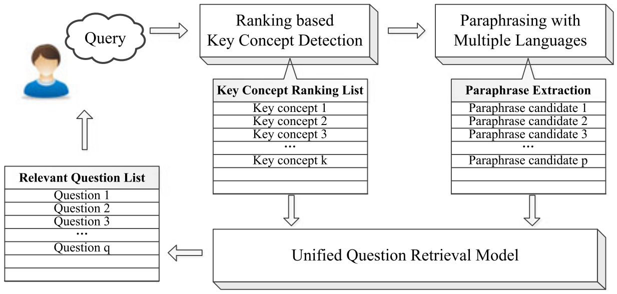

Fig. 2. The framework of key concept paraphrase based question retrieval. It is constructed by three modules, first, the ranking based key concept detection for query refinement, second, the translation based approach to paraphrase mining, by using multiple languages, for query expansion, and third, the unified question retrieval model to integrate key concept and the corresponding paraphrase.

From the above work in question retrieval, using translation model to capture the similarity of terms and using the external knowledge to enrich the context of questions are two state-of-the-art processes for question retrieval. In this paper, we take the advantages of the previous approaches and further integrate the weighted key concepts and their paraphrases into a unified probabilistic ranking model to tackle the word mismatch problem in question retrieval.

3 The Proposed Approach

The framework of our proposed approach is in Fig. 2. Given a question query, the first component *detects the key concepts* in query using a ranking based method. The second component then automatically *explores the key concept paraphrases* using a pivot language translation approach from multiple language resource.

3.1 Key Concept Detection

3.1.1 Concept Definition

According to [32], single words, idioms, restricted collocations or free combination of words can be used to express concepts. Although, noun phrases have been verified to be reliable in key concept detection in information retrieval [6], we also consider verb phrases. We observed that in CQA questions, verb phrases are important information carriers. Questions like *"Why do people get colds more often in lower temperature?"* and *"Why are you less likely to catch a cold or flu in spring summer and autumn than winter months?"* are two similar questions that share less common noun phrases, but their verb phrases are paraphrases. The above examples illustrate that verb phrases are as important as noun phrases in question retrieval. Hence, we use noun phrases and verb phrases extracted from the query questions and the candidate search questions as concepts.

3.1.2 Key Concept Detection

Key concepts can be seen as the refined real intent in user queries. In this section, we introduce a supervised machine learning method for key concept detection with our new features including statistical, syntactic and semantic linking information. First, we introduce two assumptions from [6]. One is that each concept c_i can be classified into one of the mutually exclusive classes: key concept class (KC) or NKC (non-key concept class). The other is that there exists a normalized variant $p_k(c_i)$ which represents the probability that c_i belongs to key concept class. Based on these assumptions, we can directly estimate the $p(c_i|q)$ as follow:

$$\hat{p}(c_i|q) = \frac{p_k(c_i)}{\sum_{c_i \in q} p_k(c_i)}. \tag{1}$$

Here, given the manually ranked concepts as training set, we aim to learn a pair-wise ranking function of the form $p_k : X \rightarrow \mathbb{R}$, such that $p_k(c_i) > p_k(c_j)$ indicates that concept c_i has a higher probability than concept c_j of belonging to class KC. Next, we present the new features, which include a mix of statistical, syntactic and semantic linking features for concept weighting.

$df(c_i)$: *document frequency* of concept in the corpus. Here, a document is a Q&A pair in the collected CQA archive.

$ngram\_tf(c_i)$: As the concept *term frequency* in the experimental corpus may not correctly reflect the importance of concepts due to the size of the corpus, we use *Google n-grams*[7] data set to estimate the concept term frequency.

$dep\_subj(c_i)$ and $dep\_obj(c_i)$: We use linguistic analysis techniques to recognize whether one of the words in the concept has the syntactic role of *nsubj* or *dobj*. Here, *nsubj* and *dobj* respectively represent the subject with the part-of-speech(*pos*) of *NN* and object of verb in current sentence. In linguistic analysis, the words with the syntactic role of *nsubj* or *dobj* are usually important components in the sentence. We use Stanford core-nlp toolkit[8] to get the dependency relations in the questions.

$ne(c_i)$: We also consider whether part of the concept or the concept itself is a named entity, as named entities tend to be key components in the sentence. The named entity recognizer is also from Stanford core-nlp toolkit.

$wiki\_link(c_i)$: Wikipedia[9] is a high quality collaborative encyclopedia constructed manually by users and experts from web pages. It contains a huge number of entities and the entities are usually referred to in each others' description text as anchors. Inspired by [33], [34], we assume anchor text in Wikipedia articles tend to be key concepts.[10]

However, we notice that not all the detected key concepts are suitable to paraphrasing. For example, for human names, product names, location names and organization names etc, we could not obtain diverse forms by paraphrasing. Hence, for the key concept paraphrase generation step, the above name entities are not considered. We recognize them by using named entity recognizer of the Stanford core-nlp toolkit.[11] Meanwhile, the concepts, including noun phrase and verb phrase, were extracted by using the openNLP[12] chunking tool. In this study, we only consider the verb phrases and noun phrases of the chunking results as key concept candidates.

[7] Google n-grams is a counted data set of English words n-grams, which is generated from a large web corpus. One can get the data from LDC website http://www.ldc.upenn.edu/

[8] http://nlp.stanford.edu/software/corenlp.shtml

[9] http://www.wikipedia.org/

[10] In this study, we used the Wikipedia dump resource which is obtained by the end of December 30, 2012.

[11] http://nlp.stanford.edu/software/corenlp.shtml

[12] http://opennlp.apache.org/

3.2 Pivot Approach to Exploring Key Concept Paraphrase

Pivot approach to paraphrase mining can be briefly described as in [20] that using a phrase in one language, which is usually called pivot language, to identify paraphrases in a target language. It first translates the original phrases in the target language into the pivot language phrases. Then it translates these pivot phrases back to the phrases in the target language. In the past few years, researchers started to use word alignment based approaches to generate paraphrase resources. Later, it is extended by Callison-Burch [21] with syntactic constraints when generating phrase paraphrases. In this work, we further extend the approach proposed by Callison-Burch [21] by considering the statistical distribution of paraphrases as selecting evidences to explore key concept paraphrases from bilingual corpora which can overcome the problems of word mismatch in question retrieval.

3.2.1 *Candidate Paraphrases Generation*

Given a concept c_i in one language, English for example, we aim to find all of the other English concepts $\mathbb{c}_j$ with the probability $p(\mathbb{c}_j|c_i) > \tau$, where τ is a threshold for initially filtering out those candidate paraphrases with low quality. The probability that $\mathbb{c}_j$ is the paraphrase of c_i is implemented as a conditional probability $p(\mathbb{c}_j|c_i)$, in terms of the translation probability $p(f|c_i)$ that English concept c_i translates as a particular concept f in the pivot language, and $p(\mathbb{c}_j|f)$ that the pivot language concept f translates as the candidate concept paraphrase $\mathbb{c}_j$. Since f can be multiple concepts in the pivot language, we can evaluate $p(\mathbb{c}_j|c_i)$ as follows:

$$p(\mathbb{c}_j|c_i) = \sum_f p(\mathbb{c}_j|c_i, f) = \sum_f p(f|c_i)p(\mathbb{c}_j|f). \tag{2}$$

We use maximum likelihood estimation [35] to calculate the translation probabilities $p(c|f)$. Here, $count(c, f)$ is equal to counting the co-occurrence of concept c and f aligned in the parallel corpus:

$$p(c|f) = \frac{count(c, f)}{\sum_c count(c, f)} \tag{3}$$

and $p(f|c)$ can be calculated similarly as $p(c|f)$. Meanwhile, we also adopt several strategies similar to [21], to improve the performance of the accuracy of paraphrase generation as follows.

Language model for paraphrase re-ranking [20]: simply substitute the concepts by their paraphrases in a question and use the language model score to re-rank these questions. Hence, the generated paraphrases can also be reranked by introducing context information.

Multiple parallel corpora [20]: automatic word alignment in single bilingual language pair is not always reliable for paraphrase generation. Hence, we use multiple parallel corpora L to reduce the systematic errors by voting the correct word alignment results. $|L|$ is the number of pivot languages,

$$p(\mathbb{c}_j|c_i) = \frac{1}{|L|}\sum_{l \in L}\sum_f p(f|c_i)p(\mathbb{c}_j|f). \tag{4}$$

Syntactic constraint [21]: use syntactic type s of c_i $(s(c_i))$ to refine the paraphrase probability, i.e., only the paraphrases $\mathbb{c}_j$ in the same syntactic role with c_i $(s(\mathbb{c}_j) = s(c_i))$ and in the different word forms with c_i $(\mathbb{c}_j \neq c_i)$ are taken into consideration,

$$p(\mathbb{c}_j|c_i) = p(\mathbb{c}_j|c_i, s(c_i)) \tag{5}$$

$$p(\mathbb{c}_j|c_i, s(c_i)) \approx \sum_{l \in L} \frac{\sum_f p(f|c_i, s(c_i))p(\mathbb{c}_j|f, s(c_i))}{|L|}, \tag{6}$$

where $p(f|c_i, s(c_i))$ and $p(\mathbb{c}_j|f, s(c_i))$ are calculated by $count(f, c_i, s(c_i))$ and $count(f, \mathbb{c}_j, s(c_i))$ which are computed by counting the co-occurrence of concept f and c_i, $\mathbb{c}_j$ which have the same syntactic role of $s(c_i)$, respectively, as $p(f|c_i, s(c_i)) = \frac{count(f,c_i,s(c_i))}{\sum_f count(f,c_i,s(c_i))}$ and $p(\mathbb{c}_j|f, s(c_i)) = \frac{count(f,\mathbb{c}_j,s(c_i))}{\sum_{\mathbb{c}_j} count(f,\mathbb{c}_j,s(c_i))}$.

3.2.2 *Paraphrase Selection*

As stated in the above discussion, we can estimate the concept paraphrase probabilities for each concept c_i in query through Equation (5). However, our final goal is to integrate the generated paraphrases into the question retrieval model. Hence, we need to allocate the weights for the integrated paraphrases in the question retrieval task. It is better to consider not only the paraphrase generation probabilities, but also the statistical distribution of paraphrases in the whole question dataset. Next, we will introduce the two schemes for allocating the weights for each candidate paraphrase $\mathbb{c}_j$ of concept c_i.

Weighting scheme based on paraphrase probability. As not all the generated paraphrases are considered to be integrated into the retrieval model, we need to normalize the paraphrase generation probabilities to help distinguish the important paraphrases by using the following equation:

$$w_{pp}(\mathbb{c}_j) = \frac{p(\mathbb{c}_j|c_i)}{\sum_{\mathbb{c}_j} p(\mathbb{c}_j|c_i)}, \tag{7}$$

where $p(\mathbb{c}_j|c_i)$ is computed by Equation (5).

Statistical distribution based weighting scheme. Meanwhile, we also consider the statistical distributions of candidate paraphrase $\mathbb{c}_j$, since it can reflect the importance of candidate paraphrases in the whole Q&A repository for the question retrieval task. Here, we introduce the *entropy* of the candidate paraphrase $\mathbb{c}_j$ to represent its weight, as *entropy* is defined to describe the importance of particular sample in the whole dataset. Hence, the weights of $\mathbb{c}_j$ can also be formulated as follows:

$$w_{sd}(\mathbb{c}_j) = \frac{p(\mathbb{c}_j)\log p(\mathbb{c}_j)}{\sum_{\mathbb{c}_j} p(\mathbb{c}_j)\log p(\mathbb{c}_j)}. \tag{8}$$

Here, we use the maximum likelihood estimation by counting the frequency of candidate paraphrase $\mathbb{c}_j$ occurred in the whole dataset. It is defined as:

$$p(\mathbb{c}_j) = \frac{df(\mathbb{c}_j)}{\sum_{\mathbb{c}_j} df(\mathbb{c}_j)}. \tag{9}$$

Here, $df(\mathbb{c}_j)$ represents the document frequency of $\mathbb{c}_j$.

Uniform paraphrase weighting scheme. Finally, we use linear integration to combine the proposed weighting scheme w_{pp} and w_{sd} as follow:

$$\hat{p}(\mathbb{c}_j|c_i) = \frac{\delta w_{pp}(\mathbb{c}_j) + (1-\delta) w_{sd}(\mathbb{c}_j)}{\sum_{\mathbb{c}_j} \left(\delta w_{pp}(\mathbb{c}_j) + (1-\delta) w_{sd}(\mathbb{c}_j)\right)}, \tag{10}$$

where δ is a free parameter in $[0, 1]$ to balance the two weighting schemes. Based on the above weighting schemes, we can choose the candidate paraphrase with the highest weight as the final concept paraphrase:

$$\hat{\mathbb{c}}_j = \arg \max_{\mathbb{c}_j : \mathbb{c}_j \neq c_i} \hat{p}(\mathbb{c}_j|c_i). \tag{11}$$

4 Integrating with the Existing Question Retrieval Models

In Section 3.2, we obtain the key concept paraphrases for query expansion and allocate weights for them. Next, we will derive the novel question retrieval model from general key concept model step by step. Finally, to obtain better performance in question retrieval, we integrate the key concepts and their paraphrases into the existing question retrieval model.

4.1 Key Concept Based Retrieval Model

We start by ranking a candidate question q^* in response to a question query q by estimating the ranking score of question q^* as in standard language model [36]. Then inspired by [6], we obtain the key concept model for question retrieval as:

$$rankScore(q^*) = \sum_i p(q|q^*, c_i) p(c_i|q^*). \tag{12}$$

Similar to [6], we use an interpolation to estimate $p(q|q^*, c_i)$ as:

$$\begin{aligned} rankScore(q^*) &= \lambda' p(q|q^*) + (1-\lambda') \sum_i p(q|c_i) p(c_i|q^*) \\ &= \lambda' p(q|q^*) + (1-\lambda') \sum_i p(c_i|q) \frac{p(q)}{p(c_i)} p(c_i|q^*). \end{aligned} \tag{13}$$

We assume a uniform distribution for $p(q)$ and $p(c_i)$, then $\frac{p(q)}{p(c_i)}$ equals to a constant C. Hence, we use a normalized parameter $\lambda = \frac{\lambda'}{\lambda' + (1-\lambda')C}$ ($\lambda \in [0, 1]$), and we obtain the ranking function as:

$$rankScore(q^*) \propto \lambda p(q|q^*) + (1-\lambda) \sum_i p(c_i|q) p(c_i|q^*). \tag{14}$$

Note that it is better that we can obtain the true probabilities of $p(c_i)$ and $p(q)$. However, if we obtain the statistics from the searching corpus, the two probabilities become corpus dependent, while the retrieval model is better to be corpus independent and applicable to both large and small corpora. One of the solutions here is to use a big standalone general corpus to obtain unbiased statistics, which requires additional storage and computation resources, and "unbiased" still remains a problem. Our current choice is simple and easy, without sacrificing the overall strength of the proposed approach.

4.2 Concept Paraphrase Enhanced Retrieval Model

For the concept c_i in query q, we use $\mathbb{c}_j$ to represent the corresponding paraphrase of c_i in the candidate question q^*. First, we want to explore the paraphrases potentially generated the actual concepts in query q. And then we get Equation (15):

$$\begin{aligned} rankScore(q^*) \propto & \lambda p(q|q^*) \\ & + (1-\lambda) \sum_i \sum_j p(c_i|q) p(c_i|q^*, \mathbb{c}_j) p(\mathbb{c}_j|q^*). \end{aligned} \tag{15}$$

And a common way to estimate a joint conditional probability is using a linear interpolation of the individual probabilities [6], [37]. Similar to the derivation from Equation (12) to (13), we use an interpolation to estimate $p(c_i|q^*, \mathbb{c}_j)$ as:

$$\begin{aligned} rankScore(q^*) \propto \lambda p(q|q^*) + (1-\lambda) \sum_i p(c_i|q) \\ \times \left(\theta p(c_i|q^*) + (1-\theta) \sum_j p(c_i|\mathbb{c}_j) p(\mathbb{c}_j|q^*) \right). \end{aligned} \tag{16}$$

Here, we again assume a uniform distribution for $p(\mathbb{c}_j)$ and $p(c_i)$, and thus $\frac{p(c_i)}{p(\mathbb{c}_j)}$ equals to a constant C'. Hence the above ranking function is equivalent to:

$$\begin{aligned} rankScore(q^*) \propto \lambda p(q|q^*) + (1-\lambda) \sum_i p(c_i|q) \\ \times \left(\theta p(c_i|q^*) + (1-\theta) \sum_j p(\mathbb{c}_j|c_i) C' p(\mathbb{c}_j|q^*) \right). \end{aligned} \tag{17}$$

For implementation, we may only consider the explicit concepts and their corresponding paraphrases, i.e., the concepts and the paraphrases that appear in the actual query q and candidate question q^* respectively. By normalizing the parameters of each part of the model as α, β and γ, we finally obtain the new question retrieval model which integrates the key concept model and paraphrase model as in Equation (18). Here, given the query question q, q^* represents the candidate question for ranking,

$$\begin{aligned} rankScore(q^*) \propto \alpha p(q|q^*) + \beta \sum_{c_i \in q} p(c_i|q) p(c_i|q^*) \\ + \gamma \sum_{c_i \in q} p(c_i|q) \sum_{\mathbb{c}_j \in q^*} p(\mathbb{c}_j|c_i) p(\mathbb{c}_j|q^*), \end{aligned} \tag{18}$$

where $\alpha = \frac{\lambda}{Z}$, $\beta = \frac{(1-\lambda)\theta}{Z}$, $\gamma = \frac{(1-\lambda)(1-\theta)C'}{Z}$. $Z = \lambda + (1-\lambda)\theta + (1-\lambda)(1-\theta)C'$, α, β and γ are three free parameters in $[0, 1]$ to balance the three parts of the model and $\alpha + \beta + \gamma = 1$. c_i and $\mathbb{c}_j$ represent the concepts and their corresponding paraphrases respectively. For all possible c_i in q, the weight of concept c_i in q equals to $p(c_i|q)$ as it is estimated in Section 3.1. $p(c_i|q^*)$ which represents the weight of concept c_i in q^* is estimated through maximum likelihood.

To be specific, the values of $p(c_i|q)$ and $p(\mathbb{c}_j|c_i)$ in Equation (18) are calculated by Equation (1) and (10) respectively. Specially, $p(\mathbb{c}_j|c_i)$ can be interpreted from the following three angles.

First, it indicates the probability that $\mathbb{c}_j$ can be the paraphrase of concept c_i. Second, for paraphrase selection, it represents the importance of concept paraphrase $\mathbb{c}_j$ among all the other candidate paraphrases. Third, it is the weight of concept paraphrase $\mathbb{c}_j$, which is integrated into the key concept paraphrase based question retrieval model.

4.3 Integrating with the Existing Question Retrieval Models

It is worth noticing that the former model $p(q|q^*)$ can be implemented in any one of the existing ranking models. Here, we re-implement there classic information retrieval models and the state-of-the-art question retrieval model and integrate them into the proposed question retrieval framework (QRF) respectively.

For the classic information retrieval model, we re-implement the vector space model (VSM) [38], okapi BM25 model (BM25) [39] and Language model [36]. Given the query q and the candidate question q^c, we use RS to represent the ranking score of the existing IR models. The forms of the three IR models are as follows:

$$RS_{VSM} = \frac{\sum_{t \in q \bigcap q^c} w_{t,q} w_{t,q^c}}{\sqrt{\sum_t w_{t,q}^2} \sqrt{\sum_t w_{t,q^c}^2}}. \tag{19}$$

Here $w_{t,q} = \ln(1 + \frac{N}{f_t})$, $w_{t,q^c} = 1 + \ln(tf_{t,q^c})$. N is the number of questions in the collection, f_t is the number of questions that contain term t, and tf_{t,q^c} is the frequency of term t in q^c,

$$RS_{BM25} = \sum_{t \in q \bigcap q^c} w_{t,q} w_{t,q^c}. \tag{20}$$

Here $w_{t,q} = \ln(\frac{N+f_t+0.5}{f_t+0.5})$, $w_{t,q^c} = \frac{(k+1)tf_{t,q^c}}{k(1-b)+b\frac{W_{q^c}}{W_A}+tf_{t,q^c}}$. k and b are two empirical parameters. W_{q^c} is the question length of q^c and W_A is the average question length in the whole question set,

$$\begin{aligned} RS_{LM} &= \prod_{t \in q} P(t|q^c) \\ &= \sum_{t \in q} P(t|M_q) \times \log P(t|M_{q^c}). \end{aligned} \tag{21}$$

Here $P(t|M_q) = tf_{t,q^c}$, $P(t|M_{q^c}) = \frac{|q^c|}{|q^c|+\delta} \times \frac{tf_{t,q^c}}{|q^c|} + \frac{\delta}{|q^c|+\delta} \times \frac{tf_{t,C}}{|C|}$. C is the collection which contains about 20 millions question and answer pairs. $tf_{t,C}$ is the frequency of term t in C and δ is a smoothing parameter. Dirichlet smoothing is used in language model.

For the state-of-the-art question retrieval model, we re-implement the translation based language model, which is proposed by Xue et al. [1]. The ranking score of TLM is computed as follows:

$$RS_{TLM} = \prod_{w \in q} \beta p_{ml}(w|q^c) + (1-\beta) \sum_{t \in q^c} p(w|t) p(t|q^c). \tag{22}$$

Here, $p(w|q^c)$ and $p(w|t)$ denote the language model and translation model respectively. β is the parameter to balance the two models.

We then integrate the classic information retrieval models and the state-of-the-art question retrieval model into the proposed key concept paraphrasing based question retrieval framework by replacing the $p(q|q^c)$ in Equation (18) to the RS_{VSM}, RS_{BM25}, RS_{LM} and RS_{TLM} respectively.

TABLE 1
Statistics of the Experimental Data Set

# of total questions	1,123,034
# of queries in T	251
# of relevant questions in T	1,624
# of questions in D	83
# of relevant questions in D	644

# represents the number of corresponding questions.

5 Experiment Results

For question retrieval, we collected a large question data set from Yahoo! Answers, which contains $1,123,034$ questions as the retrieval corpus. It covers a range of popular topics, including health, internet, etc. For question retrieval experiment, we utilize the experiment data set (**T**) which is used in [40]. It contains 251 queries[13] and 1,624 manually labeled relevant questions. We also randomly select 83 additional queries with 644 manually labeled relevant questions as our development set (**D**) to tune all the involved parameters. For the ground truth of **D**, two annotators who were not involved in the design of the proposed methods, are employed to independently annotate whether the candidate question is relevant with the query question or not. When conflicts occurred, a third annotator was involved to make the final decision. The development set has no overlap with the 251 search queries. Table 1 details the statistics of the experimental data set.

For key concept detection, we randomly selected 1,000 questions which had no overlapping concepts with the searching queries. After question chunking, we obtained a total of 3,685 concepts. We used four annotators to give their judgements of concept importance at three levels: *definitely important*, *partially important*, or *not important*. They labeled each concept in one of the three levels. In our experiments, there are two kinds of chunking errors. One is the chunks which have wrong boundaries or have no meaning, such as "save that he", "often you have", "of those wet" etc. We have annotated this kind of chunking error as "*not important*". The other is the chunks which have explicit meaning but not be a noun or verb phrase. We have annotated this kind of chunking error as "*partially important*". The concepts with the correct chunking results of noun or verb phrases and having the correct meanings are labeled as "*definitely important*". The final label for each concept was decided via label voting. When for one concept the above three annotators gave three different labels, a fourth annotator will decide the final label of the concept.

For paraphrase generation, we used the Europarl [21] which contains ten parallel corpora between English and (each of) Danish, Dutch, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish. With approximately 30 million words per language, we obtained a total of 315 million English words. We used Giza++ [41] to create automatic word alignments. A trigram language model was trained on the English sentences using the SRI language modeling toolkit [42].

[13] We remove 1 query from the original query set as the query has only 1 relevant question which is the same to it.

TABLE 2
Word Distribution and Coverage between Bilingual Corpora (English Part) and Current Dataset Used in the Question Retrieval Task

	# of unique words	# of shared words
Q&A repository	$324,380$	$54,773(16.89\%)$

% in the bracket indicates the coverage of bilingual corpora on the Q&A repository in percentage.

To check the vocabulary coverage of the bilingual corpora we compare the English word distributions between bilingual corpora and the question retrieval dataset. Table 2 summarizes the English word distribution and the coverage of the bilingual corpora and our current Q&A repository.

We can see that the coverage of bilingual corpora on Q&A repository is not high. Hence, to better analyze the uncovered words in our dataset, we randomly select 10,000 uncovered words from the vocabulary of Q&A repository. We manually cluster these uncovered words into five categories: "multi-words/sub-words/typo", "special sequences", "non-English words", "proper noun" and "others". We present the ratio of each category in the selected uncovered vocabulary and some example cases in Table 3. We draw the following observations:

The cases in "multi-words/sub-words/typo" category are usually caused by the typo or ellipsis when user input questions. For example, "abovetopsecret" should be "above top secret", "aboutu" should be "about you" and "systeam" should be "system".

The "special sequences" category contains some sequences that have no actual meanings or represent some special meanings that are hard to understand. For example, "zzzzzz" may represent sleeping or other state. Hence the cases in these two categories are ill formatted words and hard to adapt to downstream processes, such as paraphrase generation and question retrieval.

In "Non-English words" category indicated, there are some words in other languages in question and answer data. This is difficult to understand in pure English context and certainly hard to be covered by other English dataset.

In the "Proper noun" category, some name entities, such as human and brands names, are usually OOV (Out Of Vocabulary) words. They are difficult to be matched by other dataset. Fortunately, these proper nouns do not need to be paraphrased for question retrieval.

The "others" category takes 4.14 percent of the whole uncovered vocabulary. It mainly consists of words in different grammatical tenses and possessive cases.

From above, we can conclude that most of the uncovered words in our current question retrieval dataset are ill formed and do not influence the effects of paraphrase generation and question retrieval. At last, for our testing set, we obtain a total of $1,100$ concepts and correspondingly obtain $7,752$ paraphrases by using the proposed pivot language translation approach. Hence, for each concept in testing set, we obtain 7.05 paraphrases on average.

For training TLM, we used the similar question pairs in [27] and Microsoft parallel corpus in [43], [44] as the monolingual parallel corpora. The similar question pairs were selected by users in WikiAnswer CQA service. We obtained a total of 16,448,892 monolingual parallel sentence pairs. Giza++ [41] was used to create automatic word alignments.

5.1 Key Concept Detection Results

To assess the effectiveness of our approach on key concept detection, we utilize the SVM^{rank}[14] tool for concept ranking. The SVM^{rank} model is selected for two reasons. First, key concept detection is essentially a ranking task. As we have presented in Section 3.1, once we obtain the concept ranking list, we can obtain the key concepts. Second, ranking methods are more suitable than classification methods in practice as it not only compares the differences between concepts in KC and NKC, but also compares the differences among the concepts in KC.

For the experiment comparison, we choose two baselines. The first is [6] which used the AdaBoostM1 model with lexical, term frequency, Google n-gram and query log features to discover key concept in verbose query. The second is [45] which used the Markov Random Field to model the term dependencies for key concept identification in verbose query. Precision at position one $(p@1)$ and mean reciprocal rank (MRR) are adopted as our evaluation metrics. And the MRR calculated on the returned top five concepts. We use 5-fold cross validation on the 3,685 concepts of the 1,000 questions for the key concept detection experiment. Table 4 presents the experimental results on baseline1 [6] and baseline2 [45].

From Table 4, we can see that: First, the baseline1 can be enhanced by the features proposed in our approach. The reason may be that we not only capture the statistical information, such as the document frequency and Google $n\text{-}gram$, but we also obtain the advantages of linguistic analysis, such as dependency parsing and named entity recognition, and external knowledge base, such as Wikipedia.

Second, our proposed ranking-based model to key concept detection (RbKCD) outperforms the classification based models. The reason may be that the RbKCD not only can capture the differences between positive instance (key concept) and negative instance (non-key concept), but can also capture the differences among positive instances. As verified by Bendersky and Croft [6], not all of the key concepts are useful for information retrieval. Meanwhile, it is clear that there is no need to add more concepts into IR model. This is consistent with the result in [6]. In our experiments, the best performance is achieved when only one key concept was added into the question retrieval model.

Third, the proposed approach outperforms the baseline2 [45] at both p@1 and MRR. This is because that the baseline2 approach only model the unigram, bigram and unordered window terms. However, the unigram and bigram are usually ambiguous in sense. Our proposed approach captures the weights of concepts in query by using the statistic and linguistic information. Moreover, the phrase structure can better represent the independent semantic unit.[15]

[14] www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

[15] We should note that we do not employ the query log information as features for key concept detection. This is because we cannot obtain the query log resources. According to the experimental results in [6], [45], we deduce that the current performance of our proposed approach may further be improved by considering the features exploited from query log.

TABLE 3
The Distribution of Uncovered Words through Random Selection in the Four Categories and Example Cases

	Multi-words/Sub-words/Typo	Special sequences	Non-English words	Proper noun	Others
Example Cases	abovetopsecret bestgamblingsites westsi	zzzzzz abab xoxoxoxoxox	zhejiang mawaived hadoiiii	xbox skii suzuki	earrings renewed girlygirly
%	40.65	29.96	11.50	13.75	4.14

% indicate the percentage of each categories in the selected uncovered vocabulary.

We also analyze the utility of various features used in our key concept detection task as described in Section 3.1.2. In each iteration, we remove one single feature from feature set and leave the other features for training and prediction. We assume that the features are independent with each other, and the decreasing accuracy thus indicates the contribution of the removed feature to the overall accuracy. Table 5 presents the experimental results of feature analysis.

From Table 5, we note that all of the above features contribute more or less to key concept detection task. The features $df(c_i)$ and $dep\_subj(c_i)$ contribute the most on MRR and $p@1$ respectively. This is because for the ranking task, document frequency of concept usually reflects its statistical distribution on the whole dataset, and hence the lower the document frequency of concept, the more important it is. Meanwhile, we can conclude that the top rank concepts are more likely to be the subjects in the given queries. In future work, we plan to consider these differences in features for further improving the performance of key concept detection.

5.2 Concept Paraphrase Generation Results

5.2.1 Evaluation on Paraphrase Generation

As the bilingual parallel corpora are used for paraphrase generation in our proposed approach, we call it "**BilingPivot**" for short. Meanwhile, paraphrase generation can also be done from monolingual parallel corpora by using monolingual translation model [43], [44], [46], [47]. For comparison, we implement the the state-of-the-art method of paraphrase generation from monolingual parallel corpora in [43] as our baseline, which is treated as a statistical machine translation problem that utilized a monotone phrasal decoder to generate paraphrases in same meaning. We call it "**MonolingTrans**" for short. For training, we use two data set as the monolingual parallel corpora. First is the similar question pairs in [27] which are collected by the users' clicking of the similar questions of the search queries in WikiAnswer service. Here, the similar question pairs which are chosen by users are paraphrases. Second is the Microsoft parallel corpus in [43], [44] which is constructed by automatically aligning the similar news articles of the same topic and then extract the sentence level paraphrases.

TABLE 4
Experimental Results on Key Concept Detection (KCD)

KCD Models	MRR	$p@1$
Bendersky&Croft2008	79.14	64.29
Bendersky et al.2010	81.45	65.71
Bendersky&Croft2008($\mathcal{F}$)	82.14	68.57
Bendersky et al.2010($\mathcal{F}$)	84.57	71.42
Our Approach	**85.89**	**73.85**

$\mathcal{F}$ indicates using our proposed feature set. The results in bold are obtained by our approach.

For evaluation, we invited two native English speakers to provide their judgments on whether the generated paraphrases can replace the original concepts. As the experimental results were evaluated by two annotators, we set 20 percent of overlap data to compute their agreements. For the paraphrase generation task, the *kappa* value κ equals to 0.617, which is interpreted as "good" agreement. Meanwhile, the number of key concept for paraphrasing is equal to 1,000. According to Equation (11), we totally get 1,000 paraphrases for evaluation in Table 6 and Fig. 3.

The experimental results are presented in Table 6 with the evaluation of $average\ accuracy$.

From Table 6, we can see that BilingPivot outperforms MonolingTrans on the correct meaning. It is because monolingual method uses the translation model to capture the similarity between each term pair in monolingual parallel sentences. In this case, the similarity is calculated by the statistical co-occurrence between two terms in the same language. Hence, it may cause error in paraphrase generation as the most co-occurrent phrases are not always paraphrases.

5.2.2 Pivot Languages Analysis

As described at the beginning of Section 5, we use 10 bilingual parallel corpora to generate the concept paraphrases.

TABLE 5
Feature Analysis

	$p@1$ (% chg)	MRR (% chg)
$df(c_i)$	-4.60	$\mathbf{-7.10}$
$ngram\_tf(c_i)$	-1.40	-3.33
$dep\_subj(c_i)$	$\mathbf{-7.50}$	-5.40
$dep\_obj(c_i)$	-3.40	-6.90
$ne(c_i)$	-2.30	-3.50
$wiki\_link(c_i)$	-1.70	-3.30

% of change (% chg) in accuracy when a single feature is removed. Negative value for a feature indicates that accuracy has decreased after feature removal and vice versa.

TABLE 6
Experiment Results of key Concept Paraphrase Generation on Percentage of Correct Meaning

	MonolingTrans	BilingPivot
Average Accuracy	55.47%	**59.29**%

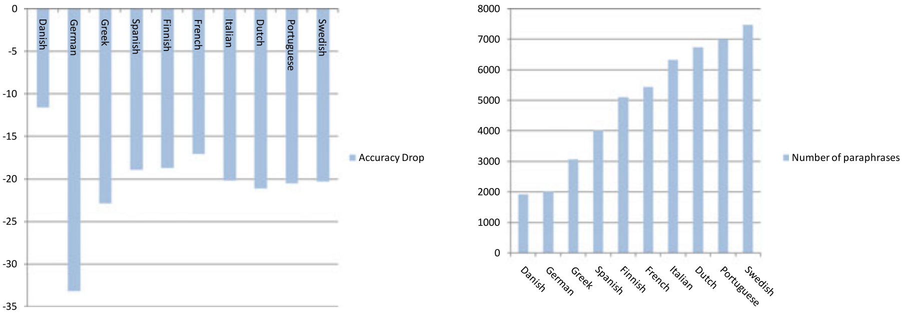

Fig. 3. Accuracy drop of paraphrasing and the number of paraphrases when a single pivot language is removed.

We actually obtain 10 pivot languages. However, different pivot languages may not have the same performance. To verify this, we design to remove one language at a time and use the rest of nine pivot languages for paraphrase generation. We can then distinguish the different abilities for paraphrase generation among these pivot languages. Fig. 3 reports the experimental results of pivot language analysis. We randomly select 110 concepts as input to obtain the paraphrases for manual evaluation.

From Fig. 3, we observe that German language contributes the most in terms of the accuracy of paraphrase generation and Danish the least.

As the statistics on our Q&A repository show that Noun Phrase is the majority type of concept (44.02 percent). We further check the part-of-speech (*pos*) distributions on the generated paraphrases for each language resource. Fig. 4 shows the result.

Here, “ADJP”, “JJ”, “NP”, “PP” and “VP” represent adjective phrase, adjective word, noun phrase, preposition phrase and verb phrase respectively.

From Fig. 4, we found that the most percentage of paraphrases in all the 10 pivot languages are NP (noun phrase), followed by the VP (verb phrase) paraphrases. It shows that most of the paraphrases are NP and VP. It reveals the language habit on paraphrasing and may be indicative to the paraphrase generation task.

According to the analysis of the Europarl corpora on machine translation [48], one reason for the differences of the translations between two languages is morphological richness. Noun phrases in German are marked with cases, which manifests themselves as different word endings at nouns, determiners etc. Hence, The richness of German may explain the highest contributions of it on the paraphrasing performance by using it as the pivot language. Moreover, with Danish language is removed, we obtain the smallest number of generated paraphrases. Although each of the language resource is about the same scale in terms of sentence number, the sparsity of the vocabularies on each pivot approach are different, which may lead to the different performance on paraphrasing. According to the statistics by Koehn [48], the Finnish vocabulary is about five times as big as English, due to the morphology. By checking the number of unique words on each language resource, we find that the Danish and Swedish corpora have the largest and smallest numbers of unique words respectively. Hence, we can deduce that the differences on the quantities of generating paraphrases may be cause by the different scales of vocabularies of each corpus.

Overall, we can also see that when any of the 10 pivot languages is removed, the corresponding performance decreases. It suggests that all of the 10 pivot languages are contributing to paraphrase generation.

5.3 Question Retrieval Results

5.3.1 Parameter Tuning

In the experiments, we use grid search to obtain the optimal values of parameters on development set which contains 83 questions.

For the δ in Equation (10), the best performance is achieved when $\delta = 0.5$. It reveals that the paraphrase weighting schemes based on paraphrase probability and statistical distribution may be equally important for the paraphrase selection experiment. For the parameters of α, β and γ in Equation (18), the best performance is achieved when $\alpha = 0.4$, $\beta = 0.2$ and $\gamma = 0.4$. Moreover, we also check the impact of the number of paraphrases, which are integrated into the retrieval model, on the performance of the question retrieval result. The best performance is obtained when adding one paraphrase to the retrieval model. It indicates that not all of the generated key concept paraphrases are useful for the question retrieval task.

5.3.2 Comparison Systems

To evaluate the proposed key concept paraphrase based question retrieval model, we compare with the following question retrieval models.

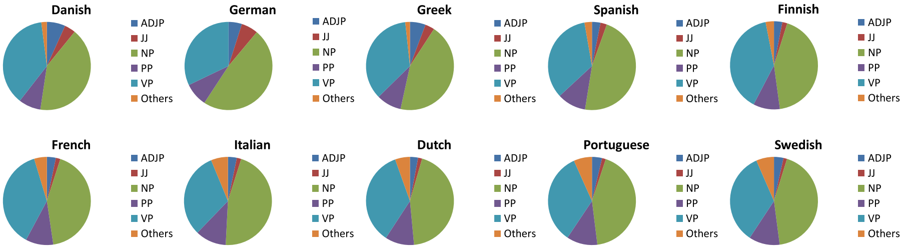

Fig. 4. The percentages of pos of the generated concept paraphrases when only a single pivot language is used to generate the paraphrases.

TABLE 7
Experimental Results among Different Question Retrieval Models

	TLM	STM	REL	PBTM	ETLM	WKM	M-NET	KCM	*Mono*KCM	***Para*KCM**
MAP	0.3957	0.3971	0.3967	0.4095	0.4073	0.4116	0.4507	0.4118	0.4509†	**0.4578***
p@5	0.3238	0.3259	0.3232	0.3318	0.3314	0.3413	0.3686	0.3414	0.3688†	**0.3722***
p@10	0.2548	0.2564	0.2548	0.2603	0.2603	0.2715	0.2848	0.2722	0.2848†	**0.2889***

The $*$ and † indicate that the results of ParaKCM *and* MonoKCM *are statistical significant over the TLM, STM, REL, PBTM, ETLM, and WKM models (within* 0.95 *confidence interval using the t-test), respectively.*

TLM. The translation based language model proposed by Xue et al. [1], which is the state-of-the-art question retrieval model which combines the translation model and the language model to estimate the parameters in ranking function. (baseline 1).

STM. The syntactic tree matching model [26], which is mainly based on a syntactic tree kernel function to compute the structure similarity of the query and candidate questions. (baseline 2).

REL. The improved pseudo relevance feedback (PRF) model [49] with new optimized term selection scheme (baseline 3).

KCM. The key concept based retrieval model proposed [45], which is the state-of-the-art model for key concept detection in verbose queries (baseline 4). It uses the AdaBoostM1 model to classify the key concept from nonkey ones with multiple features.

***Mono*KCM**. The key concept paraphrase based question retrieval model, where the paraphrases are obtained by using the monolingual based paraphrase generation approach [47].

PBTM. The phrase based translation model for question retrieval in CQA archives [10], which is the first work to use machine translation probabilities to estimation the term similarity for question retrieval.

ETLM. The entity based translation language model for CQA question retrieval [29], which is an extension of TLM by replacing the word translation to entity translation for ranking.

WKM. The world knowledge (WK) based question retrieval model [50], which used the Wikipedia as an external resource to add the estimation of the term weights from Wikipedia space into the ranking function.

M-NET. The M-NET [30] which is a state-of-the-art approach to CQA question retrieval using continuous word embedding, which added the meta-data (category information) of the questions to obtain the updated word embedding and Fish Vector is utilized to regularize the question length.

***Para*KCM**. The proposed key concept paraphrase based question retrieval model in CQA archives.

5.3.3 Question Retrieval Results

For evaluation, we use precision at position n ($p@n; n = 5, 10$) and mean average precision (MAP). The experimental results are shown in Table 7.

We can conclude from Table 7: KCM model outperforms TLM model. It indicates that the key concept based query refinement scheme is effective in question retrieval task. The reason is that TLM model employs IBM translation model 1 to capture the word translation probabilities. However, as we described in Section 1, questions in CQA repositories are usually verbose and some of the words are noise for question matching. Zhou et al. [10] also revealed that the translation between the query and candidate question requires a *distillation*. Hence, the quality of word alignment is lower in TLM and it may have negative impact on the translation accuracy.

STM model captures the structure similarities between queries and questions. It can fairly improve the performance of string matching in question retrieval. However, first, most of the similar questions in UGC data share less common structures in syntactic tree. Second, the semantic similarity in STM is measured by WordNet and partial matching of production rules, which may face to the data sparseness problem on UGC query expansion.

PBTM model outperforms the TLM model. It validates the effectiveness of the content information of terms, which is modeled by phrase or consecutive sequence of words, for question retrieval.

ETLM and WKM models are based on the external resources, e.g., Wikipedia. ETLM constrain the translations between query and question based on the entities in them. The translation probabilities are then estimated through the QA pair alignment and the Wikipedia co-occurrence. While, WKM utilizes a more widely information from Wikipedia. It generalizes the concepts in queries by exploiting their synonyms, hypernyms, associative concepts etc., through Wikipedia thesaurus. These synonyms and associative concepts can be seen as an expansion for query and perform better than traditional bag-of-word (BoW) models. However, both of their performance are limited by the low coverage of the concepts (or entities) of Wikipedia on the UGC expressions.

M-NET model outperforms all the baselines and the KCM model. The reasons are two-fold. First, the M-NET employ the continuous word representation which can better capture the semantic similarity of words. Second, the metadata is utilized as a regularization item for learning better word embedding.

*Mono*KCM model outperforms the KCM model. It shows that the concept paraphrase resources can further improve the performance of concept based question retrieval model. It verifies that both query refinement and expansion are important to question retrieval. Meanwhile, we can see that *Mono*KCM model outperforms the TLM model by a large margin. It again verifies that the phrase based translation model can better capture the similarities between query and candidate questions than the word level translation model.

The proposed *Para*KCM model outperforms the *Mono*KCM model. It may be because that the different corpus sizes for obtaining the paraphrases by using the "*BilingPivot*" and the "*MonolingTrans*" approaches respectively, as the

TABLE 8
Experimental Results over Different IR Models that Integrated into the Proposed QRF

	MAP	p@5	p@10
VSM	0.3703	0.2847	0.2536
VSM+QRF	0.4024	0.3095	0.2757
BM25	0.3749	0.3017	0.285
BM25+QRF	0.4074	0.313	0.279
LM	0.4177	0.3167	0.285
LM+QRF	0.4578	0.3233	0.2697
TLM	0.3957	0.3238	0.2548
TLM+QRF	0.4578	0.3722	0.2889

corpus size used in the former is much larger than that used in the latter. We will collect more, if any, monolingual parallel corpus to verify the performance variation in the future work.

5.3.4 Performance Variation by Integrating Different IR Models

As described in Section 4.3, we also check the variation of the performance of question retrieval over different IR models that are integrated into the proposed question retrieval framework. Table 8 shows the experimental results of these models in question retrieval.

From Table 8, we can see that the performance of all the four models are boosted by being integrated into the proposed question retrieval framework. It again reveals that the paraphrase model is compatible with the existing IR models and contributes effective semantic connection among the key concepts in the query and the retrieved questions.

6 Conclusion

In this paper, we proposed a key concept paraphrasing based approach to effectively tackle the major problems of word verboseness and word mismatch in question retrieval by exploring the translations of pivot languages. Further, we expanded queries with the generated paraphrases for question retrieval. The experimental results showed that the key concept paraphrase based question retrieval model outperformed the state-of-the-art models in the question retrieval task.

In the future, we plan to generate the concept paraphrases to jointly estimating their probabilities on the multiple linguistic resources. Meanwhile, we will consider to adopt the word or phrase embedding approach to explore the phrasal paraphrases due to its power on measuring words or phrases similarities using the context of monolingual resource. In addition, we plan to distinguish the differences of the *pos* on the concept paraphrases generation by using the diverse combinations of pivot languages and reallocate their weights for different pivot languages.

Acknowledgments

This paper is supported by the 973 Program (Grant No. 2014CB340503), National Natural Science Foundation (Grant No. 61133012, 61502120). Z.-Y. Ming is the corresponding author.

References

[1] X. Xue, J. Jeon, and W. B. Croft, "Retrieval models for question and answer archives," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008, pp. 475–482.

[2] X. Cao, G. Cong, B. Cui, C. S. Jensen, and Q. Yuan, "Approaches to exploring category information for question retrieval in community question-answer archives," *ACM Trans. Inf. Syst.*, vol. 30, no. 2, p. 7, 2012.

[3] J. H. Park and W. B. Croft, "Query term ranking based on dependency parsing of verbose queries," in *Proc. ACL*, 2010, pp. 829–830.

[4] J. Allan, J. P. Callan, W. B. Croft, L. Ballesteros, J. Broglio, J. Xu, and H. Shu, "INQUERY at TREC-5," in *Proc. TREC*, 1996, pp. 119–132.

[5] J. P. Callan, W. B. Croft, and J. Broglio, "TREC and tipster experiments with INQUERY," in *Proc. Inf. Process. Manage.*, 1995, pp. 327–343.

[6] M. Bendersky and W. B. Croft, "Discovering key concepts in verbose queries," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008, pp. 491–498.

[7] K. Collins-Thompson and J. Callan, "Query expansion using random walk models," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manage.*, 2005, pp. 704–711.

[8] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," in *Proc. 19th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1996, pp. 4–11.

[9] J. Jeon, W. B. Croft, and J. H. Lee, "Finding similar questions in large question and answer archives," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manage.*, 2005, pp. 84–90.

[10] G. Zhou, L. Cai, J. Zhao, and K. Liu, "Phrase-based translation model for question retrieval in community question answer archives," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Lang. Technol. - Vol. 1*, 2011, pp. 653–662.

[11] P. Resnik and N. A. Smith, "The web as a parallel corpus," *Comput. Linguist.*, vol. 29, no. 3, pp. 349–380, Sep. 2003.

[12] P. D. Turney, "Learning algorithms for keyphrase extraction," *Inf. Retr.*, vol. 2, pp. 303–336, 2000.

[13] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2003, pp. 216–223.

[14] M. Bendersky and W. B. Croft, "Modeling higher-order term dependencies in information retrieval using query hypergraphs," in *Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2012, pp. 941–950.

[15] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma, "Query expansion by mining user logs," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 829–839, Jul. 2003.

[16] D. Buscaldi, P. Rosso, and E. S. Arnal, "A WordNet-based query expansion method for geographical information retrieval," *Work. Notes Clef Workshop*, 2005.

[17] S. Riezler, A. Vasserman, I. Tsochantaridis, V. O. Mittal, and Y. Liu, "Statistical machine translation for query expansion in answer retrieval," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguistics*, 2007, pp. 464–471.

[18] J. Gao and J.-Y. Nie, "Towards concept-based translation models using search logs for query expansion," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012.

[19] R. A. Baeza-Yates and B. A. Ribeiro-Neto, *Modern Information Retrieval - The Concepts and Technology Behind Search*, 2nd ed. Harlow, England, U.K: Pearson Education Ltd., 2011.

[20] C. Bannard and C. Callison-Burch, "Paraphrasing with bilingual parallel corpora," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2005, pp. 597–604.

[21] C. Callison-Burch, "Syntactic constraints on paraphrases extracted from parallel corpora," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2008, pp. 196–205.

[22] S. Zhao, H. Wang, T. Liu, and S. Li, "Pivot approach for extracting paraphrase patterns from bilingual corpora," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2008, pp. 780–788.

[23] N. Tomuro, "Interrogative reformulation patterns and acquisition of question paraphrases," in *Proc. 2nd Int. Workshop Paraphrasing*, 2003, pp. 33–40.

[24] A. L. Berger, R. Caruana, D. Cohn, D. Freitag, and V. O. Mittal, "Bridging the lexical chasm: statistical approaches to answer-finding," in *Proc. 23rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2000, pp. 192–199.

[25] H. Duan, Y. Cao, C.-Y. Lin, and Y. Yu, "Searching questions by identifying question topic and question focus," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2008, pp. 156–164.

[26] K. Wang, Z. Ming, and T.-S. Chua, "A syntactic tree matching approach to finding similar questions in community-based qa services," in *Proc. 32nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2009, pp. 187–194.

[27] D. Bernhard and I. Gurevych, "Combining lexical semantic resources with question & answer archives for translation-based answer finding," in *Proc. ACL*, 2009, pp. 728–736.

[28] S. H. Na and H. T. Ng, "Enriching document representation via translation for improved monolingual information retrieval," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2011, pp. 853–862.

[29] A. Singh, "Entity based q&a retrieval," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2012, pp. 1266–1277.

[30] G. Zhou, T. He, J. Zhao, and P. Hu, "Learning continuous word embedding with metadata for question retrieval in community question answering," in *Proc. ACL*, 2015, pp. 250–259.

[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. (2013). Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst. 26*, pp. 3111–3119. [Online]. Available: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[32] L. Bentivogli and E. Pianta, "Beyond lexical units: Enriching wordnets with phrasets," in *Proc. 10th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2003, pp. 67–70.

[33] D. Odijk, E. Meij, and M. de Rijke, "Feeding the second screen: Semantic linking based on subtitles," in *Proc. 10th Conf. Open Res. Areas Inf. Retrieval*, 2013, pp. 9–16.

[34] E. Meij, M. Bron, L. Hollink, B. Huurnink, and M. Rijke, "Learning semantic query suggestions," in *Proc. 8th Int. Semantic Web Conf.*, Berlin, Heidelberg, 2009, pp. 424–440.

[35] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. NAACL*, 2003, pp. 48–54.

[36] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1998, pp. 275–281.

[37] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2006, pp. 178–185.

[38] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 10, p. 16, 1974.

[39] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proc. TREC*, 1994, pp. 109–125.

[40] X. Cao, G. Cong, B. Cui, C. S. Jensen, and C. Zhang, "The use of categorization information in language models for question retrieval," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2009, pp. 265–274.

[41] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Comput. Linguist.*, vol. 29, no. 1, pp. 19–51, Mar. 2003.

[42] A. Stolcke, "SRILM—An extensible language modeling toolkit," in *Proc. 7th Int. Conf. Spoken Language Process.*, 2002, pp. 901–904.

[43] C. Quirk, C. Brockett, and W. B. Dolan, "Monolingual machine translation for paraphrase generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2004, pp. 142–149.

[44] B. Dolan, C. Quirk, and C. Brockett, "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources," in *Proc. 20th Int. Conf. Comput. Linguistics*, 2004, pp. 350–356.

[45] M. Bendersky, D. Metzler, and W. B. Croft, "Learning concept importance using a weighted dependence model," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining*, 2010, pp. 31–40.

[46] A. Ibrahim, B. Katz, and J. Lin, "Extracting structural paraphrases from aligned monolingual corpora," in *Proc. 2nd Int. Workshop Paraphrasing*, 2003, pp. 57–64.

[47] Y. Marton, C. Callison-Burch, and P. Resnik, "Improved statistical machine translation using monolingually-derived paraphrases," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2009, pp. 381–390.

[48] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proc. MT Summit*, vol. 5, 2005, pp. 3–4.

[49] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson, "Selecting good expansion terms for pseudo-relevance feedback," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008, pp. 243–250.

[50] G. Zhou, Y. Liu, F. Liu, D. Zeng, and J. Zhao, "Improving question retrieval in community question answering using world knowledge," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2239–2245.

Wei-Nan Zhang is a lecturer in the Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. His research interest includes human-computer dialogue, natural language processing, and information retrieval.

Zhao-Yan Ming is an assistant professor in the Department of Computer Science, Digipen Institute of Technology. Her research interest includes community question answering, information organization, information retrieval, and social media mining.

Yu Zhang is a professor in the Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. His primary research interest is question answering and personalized information retrieval.

Ting Liu is a professor in the Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology. His primary research interest is natural language processing, information retrieval, and social computing.

Tat-Seng Chua is a KITHCT chair professor in the School of Computing, National University of Singapore. He was the acting and founding dean of the School during 1998 to 2000. His main research interest is in multimedia information retrieval and social media analysis. In particular, his research focuses on the extraction, retrieval and question-answering (QA) of text, video and live media arising from the Web and social networks.