Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Exploring Key Concept Paraphrasing based on Pivot Language Translation for Question Retrieval

Wei-Nan Zhang[1], **Zhao-Yan Ming**[2*], **Yu Zhang**[1], **Ting Liu**[1], **Tat-Seng Chua**[3]
1 Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology
2 Department of Computer Science, Digipen Institute of Technology
3 School of Computing, National University of Singapore

Abstract

Question retrieval in current community-based question answering (CQA) services does not, in general, work well for long and complex queries. One of the main difficulties lies in the word mismatch between queries and candidate questions. Existing solutions try to expand the queries at word level, but they usually fail to consider concept level enrichment. In this paper, we explore a pivot language translation based approach to derive the paraphrases of key concepts. We further propose a unified question retrieval model which integrates the key concepts and their paraphrases for the query question. Experimental results demonstrate that the paraphrase enhanced retrieval model significantly outperforms the state-of-the-art models in question retrieval.

Introduction

Question retrieval[1] in community based question answering (CQA) is different from general Web search (Xue, Jeon, and Croft 2008). Unlike the Web search engines that return a long list of ranked documents, question retrieval returns several relevant questions with possible answers directly. While in traditional question answering (QA), the main tasks are answer extraction (Kwok, Etzioni, and Weld 2001; Moldovan et al. 2003), answer matching (Cui, Kan, and Chua 2007) and answer ranking (Ko et al. 2010), with CQA, the main task is to search for relevant questions with good ready answers (Cao et al. 2012).

One of the major challenges for question retrieval is the **word mismatch** between queries and candidate questions. For example, in Table 1, the query and question are relevant to each other, but the same meaning is expressed with different word forms, such as *"get colds"* and *"catch a cold"*, *"lower temperature"* and *"winter months"*. These make it non-trivial for semantic level question matching.

To tackle the word mismatch problem, previous work mainly resorts to query expansion. (Xu and Croft 1996) explored local and global features to expand single terms in queries. (Collins-Thompson and Callan 2005) used synonyms, cue words, co-occurrence and background smoothing to determine query associations. However, the former approach fails to assign explicit weights to the expanded aspects and the later approach overlooks phrase level evidences for query expansion. Meanwhile, pseudo relevance feedback (Baeza-Yates and Ribeiro-Neto 2011) and blending (Belkin et al. 1993) are also two effective approaches to tackle the word mismatch between queries and the candidate documents in the term level. (Zhou et al. 2013) utilized the Wikipedia as an external knowledge base to enhance the performance of question retrieval. Despite their success, literature that considers the concept level expansion by exploiting multiple external knowledge sources is still sparse.

Table 1: An example for illustrating the word mismatch between a query and a relevant question.

Query
Q1: Why do people *get colds* more often in *lower temperature*?
Relevant Question:
Q2:Why are you less likely to *catch a cold* or flu in spring summer and autumn than *winter months*?

In this paper, we take three major actions to solve the word mismatch problem in question retrieval from CQA archives as illustrated in Figure 1. First, we utilize a pivot language translation approach (Callison-Burch 2008) to explore key concept paraphrases in the queries from bilingual parallel corpora[2]. Figure 2 presents an example of pivot language translation for concept paraphrasing. We put the original concept "get colds" on the left column. The arrow directions represent the translation from source to target. English concepts on right column indicate the candidate paraphrases. Pivot languages are on the intermediate columns, German and Chinese for (a) and (b) respectively. Both of the translations in two directions are obtained by using the method of (Koehn, Och, and Marcu 2003).

Second, we estimate the importance of the generated paraphrases for the original query under two considerations. One

*corresponding author

[1]Here we define question retrieval in CQA services as a task that new questions are used as queries to find relevant questions with ready answers. 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.

[2]Bilingual corpora have been verified to be the effective resource in many subjects of information retrieval (Mehdad, Negri, and Federico 2011).

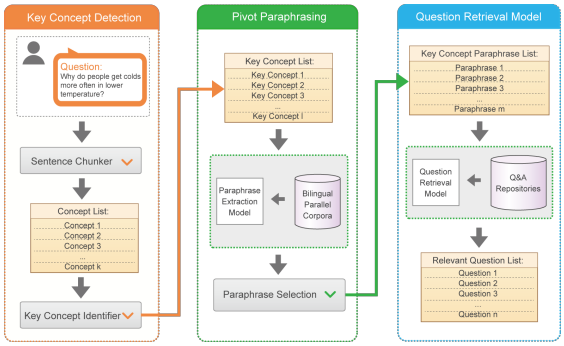

Figure 1: The framework of key concept paraphrase based question retrieval.

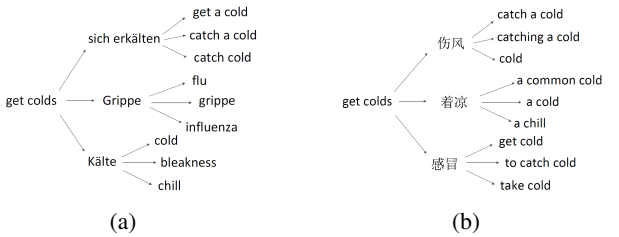

Figure 2: An example of pivot language translation approach to explore concept paraphrases.

is based on the paraphrase generation probabilities obtained from 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.

Finally, we propose a novel probabilistic retrieval model which integrates the state-of-the-art probabilistic retrieval model, key concept model, and key concept paraphrase model. The contributions of this work are twofold:

- To the best of our knowledge, this is the first attempt at using a pivot translation approach with multiple languages to explore concept level paraphrases as a semantic expansion of queries for question retrieval.
- Second, towards question retrieval task, we propose a question retrieval model using key concepts and their paraphrases which can be seamlessly integrated with the state-of-the-art question retrieval frameworks.

Key Concept Paraphrase based Question Retrieval

In this section, we will detail the proposed scheme that uses the key concept paraphrase as the expansions of queries for question retrieval. We will present the framework of the scheme, which consists of three components as shown in Figure 1. It can be decomposed into the following parts.

Key Concept Detection

According to (Bentivogli and Pianta 2003), single words, idioms, restricted collocations or free combination of words can be used to express concepts. Prevalent work (Bendersky and Croft 2008; Bendersky, Metzler, and Croft 2010) used the noun phrases as concepts in verbose queries for web search. Noun phrases have been verified to be reliable in the key concept detection in information retrieval (Bendersky and Croft 2008; Xu and Croft 1996) and natural language processing (Hulth 2003). Moreover, as verb phrases that usually represent events or relations between entities, are important information carriers, we also consider verb phrases as concepts.

In this study, we implement and extend the state-of-the-art key concept detection approach (Bendersky, Metzler, and Croft 2010). Our proposed features include statistical, syntactic and semantic linking information. The summary of these features are presented in Table 2.

Inspired by (Bendersky and Croft 2008), we assume that each concept c_i can be classified into one of the mutually exclusive classes: KC (key concept class) or NKC (non-key concept class). Meanwhile, we directly estimate the $p(c_i|q) = \frac{p_k(c_i)}{\sum_{c_i \in q} p_k(c_i)}$. Here, given the manually ranked concepts as the 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. Meanwhile, we also notice that the named entities are usually stable in form. [3]

Pivot Language Translation Approach to Key Concept Paraphrasing

To overcome the surface word mismatching between semantic similar questions, we propose to use a pivot language to bridge the semantic gap. Basically, the pivot language translation approach translates a concept in the target language into an auxiliary (or pivot) language, such that the concepts in the auxiliary language carry the meaning of the target concept, but strip off the original word form. It then translates the concepts in the auxiliary language back into the target language. In this way, the target concept is expanded into other forms in its own language, with the aid of another language. The pivot language translation approach is formally proposed and extended by (Callison-Burch 2008). In the following, we will describe our approach of using pivot language translation to expand the key concepts for question retrieval.

Candidate Paraphrases Generation Given concept c_i in one language, English for example, we aim to find a plurality of 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

[3]The named entities and the concepts are recognized by using the Stanford core-nlp toolkit (http://nlp.stanford.edu/software/corenlp.shtml) and the openNLP (http://opennlp.apache.org/) toolkit.

Table 2: A summary of features used in the key concept detection task.

Feature Name	Feature Description
$df(c_i)$	Concept document frequency in the corpus.
$ngram\_tf(c_i)$	Concept term frequency counted from Google n-grams data (Brants and Franz 2006).
$dep\_subj(c_i)$	Whether one of the words in the concept has the syntactic role of $nsubj$.
$dep\_obj(c_i)$	Whether one of the words in the concept has the syntactic role of $dobj$.
$ne(c_i)$	Whether part of the concept or the concept itself is a named entity.
$wiki\_link(c_i)$	The proportion of the concept occurring as anchor texts in the Wikipedia articles (Odijk, Meij, and de Rijke).

in the pivot language, and $p(\mathbb{c}_j|f)$ that the pivot language concept f translates as the candidate concept paraphrase $\mathbb{c}_j$.

Meanwhile, we also adopt three strategies which are proposed by (Callison-Burch 2008), to improve the performance of the accuracy of paraphrase generation. They are language model for paraphrase re-ranking, multiple parallel corpora and syntactic constraint. We then obtain the paraphrase probability as follows:

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

where $p(f|c_i, s(c_i)) = \frac{\text{count}(f, c_i, s(c_i))}{\sum_f \text{count}(f, c_i, s(c_i))}$, L and $s(c_i)$ represent the set of multiple languages and the syntactic role of c_i respectively. $\text{count}(f, c_i, s(c_i))$ equals to the co-occurrence times of f and c_i with the same syntactic constraint. $p(\mathbb{c}_j|f, s(c_i))$ can be estimated in a similar way.

Paraphrase Selection To select the generated paraphrases for the question retrieval model, we introduce two schemes for allocating the weights for each candidate paraphrase $\mathbb{c}_j$ of concept c_i, by considering their generating probabilities and the statistical distributions in the whole corpora.

Paraphrase probability based weighting scheme
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:

$$w_{pp}(\mathbb{c}_j) = 1 - \frac{\log p(\mathbb{c}_j|c_i)}{\sum_{\mathbb{c}_j} \log p(\mathbb{c}_j|c_i)} \tag{2}$$

where $p(\mathbb{c}_j|c_i)$ is computed by Equation (1).

In this weighting scheme, we obtain the final weights of the paraphrases as $w_{pp}(\mathbb{c}_j)$. Here, we assume the higher the paraphrase probability, the more important the paraphrase is. As the value of $\log p(\mathbb{c}_j|c_i)$ is negative, the normalization item is in inverse ratio of the paraphrase weight.

Statistical distribution based weighting scheme
The statistical distributions of candidate paraphrase $\mathbb{c}_j$ reflects the importance of candidate paraphrases in the whole Q&A repository. 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 data set. Hence, the weights of $\mathbb{c}_j$ can also be formulated as follows.

$$w_{sd}(\mathbb{c}_j) = \frac{p(\mathbb{c}_j) \log p(\mathbb{c}_j)}{\sum_{\mathbb{c}_j} p(\mathbb{c}_j) \log p(\mathbb{c}_j)} \tag{3}$$

Here, we use the maximum likelihood estimation $p(\mathbb{c}_j) = \frac{df(\mathbb{c}_j)}{\sum_{\mathbb{c}_j} df(\mathbb{c}_j)}$ ($df(\mathbb{c}_j)$ represents the document frequency of $\mathbb{c}_j$.) by counting how often the candidate paraphrase $\mathbb{c}_j$ occurred in the whole data set.

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

$$\hat{p}(\mathbb{c}_j|c_i) = \frac{\delta w_{pp}(\mathbb{c}_j) + (1-\delta) w_{sd}(\mathbb{c}_j)}{\sum_{\mathbb{c}_j} (\delta w_{pp}(\mathbb{c}_j) + (1-\delta) w_{sd}(\mathbb{c}_j))} \tag{4}$$

where δ, which is a free parameter in $[0, 1]$ to balance the two weighting schemes.

Key Concept Paraphrasing based Question Retrieval Model

In this section, we will derive the novel question retrieval model that integrates the key concept and paraphrase model from a general key concept model step by step.

Key Concept based Retrieval Model We start by ranking a question q^* in response to a query q by estimating the ranking score of q^* as in standard language model (Ponte and Croft 1998). Then inspired by (Bendersky and Croft 2008), we obtain the key concept model for question retrieval as:

$$rankScore(q^*) = \sum_i p(q|q^*, c_i) p(c_i|q^*) \tag{5}$$

To estimate the joint conditional probability, we use a linear interpolation of the individual probabilities following (Bendersky and Croft 2008; Wei and Croft 2006).

$$rankScore(q^*) = \lambda' p(q|q^*) + (1-\lambda') \sum_i p(q|c_i) p(c_i|q^*) \tag{6}$$

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

$$rankScore(q^*) \propto \lambda p(q|q^*) + (1-\lambda) \sum_i p(c_i|q) p(c_i|q^*) \tag{7}$$

Concept Paraphrase based 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 (8).

$$\begin{aligned} rankScore(q^*) \propto & \lambda p(q|q^*) \\ & + (1-\lambda) \sum_i \sum_j p(c_i|q) p(c_i|q^*, \mathbb{c}_j) p(\mathbb{c}_j|q^*) \end{aligned} \tag{8}$$

Here, we use an interpolation of $p(c_i|q^*)$ and $p(\mathbb{c}_j|c_i)$ to estimate $\sum_j p(c_i|q^*, \mathbb{c}_j)p(\mathbb{c}_j|q^*)$ as $\theta p(c_i|q^*) + (1-\theta)\sum_j p(c_i|\mathbb{c}_j)p(\mathbb{c}_j|q^*)$.

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. We then obtain the new question retrieval model which integrates the key concept model and paraphrase model as in Equation (9).

$$rankScore(q^*) \propto \alpha p(q|q^*) + \beta \sum_{c_i \in q} p(c_i|q)p(c_i|q^*) + \gamma \sum_{c_i \in q} p(c_i|q) \sum_{\mathbb{c}_j \in q^*} p(\mathbb{c}_j|c_i)p(\mathbb{c}_j|q^*) \quad (9)$$

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 $\alpha + \beta + \gamma = 1$. $p(c_i|q)$ and $p(\mathbb{c}_j|c_i)$ can be estimated by the maximum likelihood estimation and Equation (4) respectively. $p(c_i|q^*)$ and $p(\mathbb{c}_j|q^*)$ can be estimated by the maximum likelihood. We 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'. It is worth noticing that the former model $p(q|q^*)$ can be implemented in any one of the existing probabilistic ranking models. In this paper, we choose the state-of-the-art question retrieval model, namely, translation based language model (TLM) which is proposed by Xue et al., (2008) (Xue, Jeon, and Croft 2008).

Experiment Results

Evaluation on Key Concept Detection

For key concept detection, we randomly selected 1,000 questions from the 1 million plus question data. They had no overlapping concepts with the searching queries. After question chunking, we obtained a total of 3,685 concepts. For a given concept, two annotators manually labeled it as KC or NKC. When conflicts occurred, another annotator was involved to make the final decision.

For comparison, we implemented the state-of-the-art key concept detection approach (Bendersky, Metzler, and Croft 2010) as our baseline. 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 5 concepts. We use 5-fold cross validation on the 3,685 concepts of the 1,000 questions for the key concept detection experiment. Table 3 shows the experiment results of key concept detection. From Table 2, we can see that the baseline can be enhanced by the features proposed in our approach. The reason is that we not only capture the statistical information, such as the document frequency and Google $n\text{-}gram$, but also obtain the advantages of linguistic analysis, such as dependency parsing and named entity recognition, and external knowledge base, such as Wikipedia. We notice that the performance of the baseline in this paper is lower than that in the original paper. This is due to the difference of data set.

Table 3: Experimental results on key concept detection. $\mathscr{F}$ denotes the use of our proposed features. $*$ indicates the statistical significance over the baseline (within 0.95 confidence interval using the t-test)

	Bendersky et al.2010	Bendersky et al.2010($\mathscr{F}$)
MRR	82.14	**84.57***
$p@1$	68.57	**71.42***

Table 4: Experiment results of key concept paraphrase generation on the percentage of correct meaning. $\star$ indicates the statistical significance over the baseline (within 0.95 confidence interval using the t-test)

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

Evaluation on Paraphrase Generation

Paraphrase Generation Results For paraphrase generation, we used the Europarl (Koehn 2005) which contains ten parallel corpora between English and (each of) Danish, Dutch, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish. With approximate 30 million words per language, we obtained a total of 315 million English words. We used Giza++ (Och and Ney 2003) to create automatic word alignments. A trigram language model was trained on the English sentences using the SRI language modeling toolkit (Stolcke 2002).

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 (Quirk, Brockett, and Dolan 2004; Ibrahim, Katz, and Lin 2003; Dolan, Quirk, and Brockett 2004; Marton, Callison-Burch, and Resnik 2009). For comparison, we implemented the method of paraphrase generation from monolingual parallel corpora in (Marton, Callison-Burch, and Resnik 2009), which is the state-of-the-art model, and use it as our baseline. We call it "**MonolingTrans**" for short. For training, we used the similar question pairs in (Bernhard and Gurevych 2009) and Microsoft parallel corpus in (Quirk, Brockett, and Dolan 2004; Dolan, Quirk, and Brockett 2004) as the monolingual parallel corpora.

For evaluation, we invited two native English speakers to provide their judgments on whether the generated concepts have the same meaning as the original concepts. As the experimental results were evaluated by two annotators, 20% of their annotated data are overlapped data for computing the annotation agreements. For the paraphrase generation task, the Cohen's *kappa* (Cohen and others 1960) coefficient equals to 0.617, which is interpreted as "good" agreement.

The experimental results are presented in Table 4 with the evaluation of $average\ accuracy$. From Table 4, 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 similari-

ty 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.

Pivot Languages Analysis To study the performances of different pivot languages on generating paraphrases, we remove one language at a time and use the remaining 9 pivot languages for paraphrase generation. Table 5 shows the experimental results of pivot language analysis. We randomly select 110 concepts paraphrases for analysis.

From Table 5, we observe that German language contributes the most and Danish the least in terms of the accuracy of paraphrase generation. The statistics on our Q&A repository show that NP (Noun Phrase) is the majority type of concept (44.02%). Hence, we further check the part-of-speech (pos) distributions on the generated paraphrases for each language resource. Table 6 shows the pos distributions of the generated paraphrases on percentage.

From Table 6, we found that German and Danish corpora contain the most and least percentage of NPs for generating noun phrase (NP) respectively. It suggests that the pivot languages which are suitable for NP paraphrasing are more likely to perform better on generating accurate paraphrases than other pivot languages. Hence, it may explain the reason of the accuracy changes by removing of the German and Danish corpora respectively.

Second, according to the analysis of the Europarl corpora on machine translation (Koehn 2005), the author had revealed that an apparent reason for the differences of the translations between two languages is the variance of 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, when 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 2005), the Finnish vocabulary is about five times as big as English, due to the morphology. 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.

Question Retrieval Results

Question Retrieval Data Set We collected a total number of $1,123,034$ questions as the retrieval corpus, which covers a range of popular topics, including health, internet, etc. For question retrieval experiment, we randomly selected 140 questions as searching queries and 28 as development set to tune all the involved parameters. Table 7 details the statistics of our data set.

Table 7: Statistics of question retrieval data set.

# of queries	140
# of total questions	1,123,034
# of relevant questions	1028
# of development queries	28

To obtain the relevance ground truth of each question query, we pooled the top 20 results from various methods, namely, the vector space model, okapi BM25 model, language model and our proposed methods. We then asked two annotators, who were not involved in the design of the proposed methods, to independently annotate whether the candidate question is relevant with the query or not. When conflicts occurred, another annotator was involved to make the final decision.

State-of-the-Art Methods To verify the effectiveness of our proposed key concept paraphrase based question retrieval model, we comparatively evaluate the following question retrieval models.

- **TLM**: The translation based language model proposed by (Xue, Jeon, and Croft 2008) is involved as a baseline.
- **STM**: We run the syntactic tree matching model (Wang, Ming, and Chua 2009) as a baseline. It is a structure based approach, which uses the tree kernel function to measure the similarity between query and candidate question.
- **REL**: We choose the pseudo relevance feedback (PRF) on language model (Cao et al. 2008) as a baseline.
- **WKM**: We implement the world knowledge (WK) based question retrieval model Zhou et al. (Zhou et al. 2013) as another state-of-the-art model. The world knowledge can be seen as an external source for query expansion.
- **KCM**: We present the key concept based retrieval model proposed by (Bendersky, Metzler, and Croft 2010) as a baseline.
- ***Mono*KCM**: We employ the *Mono*KCM as a baseline. It utilizes the phrase based statistical machine translation model to obtain the translation probabilities.
- ***Para*KCM**: Our proposed pivot language translation based key concept paraphrase model.

Question Retrieval Results For evaluation, we use precision at position 1 ($p@1$) and 10 ($p@10$) and mean average precision (MAP) (Baeza-Yates and Ribeiro-Neto 2011). The experimental results are presented in Table 8.

Table 8 shows that first, KCM model outperforms TLM model in the question retrieval, which reveals that the key concept based query refinement scheme is more effective in question retrieval task. The reason is that TLM model employs IBM translation model 1 to capture the word translation probabilities. However, questions in CQA repositories are usually verbose and some of the words are noise for question matching. Hence, the quality of word alignment is poorer. Moreover, it will influence the translation accuracy.

Second, STM model captures the structure similarities between queries and questions. It can improve the performance of string matching in question retrieval. However the semantic similarity in STM is measured by WordNet and a rule-

Table 5: Pivot language analysis. %chg of accuracy represents the changes of accuracy on both correct meaning and grammar when a single pivot language is removed. Negative value for a pivot language indicates that the accuracy has decreased after the pivot language is removed.

	Danish	German	Greek	Spanish	Finnish
% accuracy change	-11.57	**-33.02**	-22.88	-18.91	-18.71
# of paraphrases	**1,928**	$2,027$	$3,074$	$4,019$	$5,109$
	French	Italian	Dutch	Portuguese	Swedish
% accuracy change	-17.05	-20.18	-21.11	-20.5	-20.31
# of paraphrases	$5,446$	$6,333$	$6,739$	$7,099$	$7,487$

Table 6: The pos distributions of paraphrases on each pivot language. The values represent the percentages of pos of the generated paraphrases when only used a single pivot language for paraphrasing. Here, "ADJP", "JJ", "NP", "PP" and "VP" represent adjective phrase, adjective word, noun phrase, preposition phrase and verb phrase respectively.

	Danish	German	Greek	Spanish	Finnish	French	Italian	Dutch	Portuguese	Swedish
ADJP	**6.80**	4.94	5.75	3.14	3.07	3.31	3.09	2.99	3.71	3.50
JJ	4.08	**6.17**	3.45	2.35	1.84	1.65	1.55	1.49	1.39	1.31
NP	41.50	**48.15**	44.25	47.06	42.94	42.70	43.56	44.03	42.92	43.11
PP	8.16	8.64	9.20	10.59	9.82	10.19	10.82	10.70	11.37	**11.38**
VP	37.41	32.10	35.63	34.12	**39.26**	37.47	29.90	35.32	33.87	34.14
Others	2.04	0.00	1.72	2.75	3.07	4.68	5.93	5.47	**6.73**	6.56

Table 8: Experimental results among different question retrieval models. The † and ‡ indicate that the results of *Para*KCM and *Mono*KCM are statistical significant over all baselines and the TLM, STM, REL, WKM and KCM models (within 0.95 confidence interval using the t-test) respectively. $\%$ changes denote the improved performance in percentage. The results of our approach are in bold.

Models	p@1	p@10	MAP
TLM	0.1928	0.1759	0.2889
STM	0.2071	0.1864	0.2973
REL	0.2143	0.2015	0.3124
WKM	0.2071	0.1981	0.3203
KCM	0.2143	0.2067	0.3237
*Mono*KCM	$0.2214^{\ddagger}$	$0.2179^{\ddagger}$	$0.3554^{\ddagger}$
***Para*KCM**	$\mathbf{0.2357}^{\dagger}$	$\mathbf{0.2280}^{\dagger}$	$\mathbf{0.3910}^{\dagger}$

based approach, which has the limitation of data sparseness on UGC query expansion.

Third, WKM model 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, the number of synonyms extracted by only using the Wikipedia concepts is quite sparse. Meanwhile, the associative concepts may introduce more relevant terms rather than similar terms.

Fourth, *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 shows that the phrase based translation model can better capture the similarities between query and candidate questions than the word level translation model.

Fifth, the results of *Para*KCM model indicate that question retrieval model can be benefited from concept based query refinement and concept paraphrase based query expansion. Moreover, our proposed *Para*KCM model outperforms *Mono*KCM model, which shows that paraphrases generated from bilingual parallel corpora can enhance the performance of retrieval model more than that from the monolingual parallel corpus. This may be caused by the difference between the above two approaches of the accuracy of paraphrase generation .

Sixth, the proposed *Para*KCM model outperforms the REL model. It illustrates that our proposed model is more effective than the REL model on query expansion for the question retrieval task. This is because the proposed approach capture not only the term importance, but also the concept importance. Hence, it can be seen as the adopting of the term context information, which can overcome the shortage of REL model. Moreover, as the questions are extremely short than the documents, the number of expansion terms obtained by REL model is very limited.

Conclusion

In this paper, we proposed a pivot language translation approach to paraphrase key concept. 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 by considering to jointly estimate their probabilities on the multiple linguistic resources.

Acknowledgments

We would like to acknowledge the reviewers for their insight reviews. We thank Yiming Li for his help on the proofreading of the camera-ready version. This paper is supported by the 973 Program (Grant No. 2014CB340503), National Natural Science Foundation (Grant No. 61133012, 61472105).

References

Baeza-Yates, R. A., and Ribeiro-Neto, B. A. 2011. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England.

Belkin, N. J.; Cool, C.; Croft, W. B.; and Callan, J. P. 1993. The effect multiple query representations on information retrieval system performance. In *ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, 339–346.

Bendersky, M., and Croft, W. B. 2008. Discovering key concepts in verbose queries. In *ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, 491–498.

Bendersky, M.; Metzler, D.; and Croft, W. B. 2010. Learning concept importance using a weighted dependence model. In *WSDM*, 31–40.

Bentivogli, L., and Pianta, E. 2003. Beyond lexical units: enriching wordnets with phrasets. In *European chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, 67–70.

Bernhard, D., and Gurevych, I. 2009. Combining lexical semantic resources with question & answer archives for translation based answer finding. In *ACL-IJCNLP*, ACL '09, 728–736.

Brants, T., and Franz, A. 2006. Web 1t 5-gram version 1.

Callison-Burch, C. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Empirical Methods in Natural Language Processing*, EMNLP '08, 196–205.

Cao, G.; Nie, J.-Y.; Gao, J.; and Robertson, S. 2008. Selecting good expansion terms for pseudo-relevance feedback. In *ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, 243–250. New York, NY, USA: ACM.

Cao, X.; Cong, G.; Cui, B.; Jensen, C. S.; and Yuan, Q. 2012. Approaches to exploring category information for question retrieval in community question-answer archives. *ACM Trans. Inf. Syst.* 30(2):7.

Cohen, J., et al. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1):37–46.

Collins-Thompson, K., and Callan, J. 2005. Query expansion using random walk models. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, 704–711.

Cui, H.; Kan, M.-Y.; and Chua, T.-S. 2007. Soft pattern matching models for definitional question answering. *ACM Trans. Inf. Syst.* 25(2).

Dolan, B.; Quirk, C.; and Brockett, C. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04.

Hulth, A. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Empirical methods in natural language processing*, EMNLP '03, 216–223.

Ibrahim, A.; Katz, B.; and Lin, J. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the second international workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, 57–64.

Ko, J.; Si, L.; Nyberg, E.; and Mitamura, T. 2010. Probabilistic models for answer-ranking in multilingual question-answering. *ACM Trans. Inf. Syst.* 28(3):16:1–16:37.

Koehn, P.; Och, F. J.; and Marcu, D. 2003. Statistical phrase-based translation. In *North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, 48–54.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5.

Kwok, C.; Etzioni, O.; and Weld, D. S. 2001. Scaling question answering to the web. *ACM Trans. Inf. Syst.* 19(3):242–262.

Marton, Y.; Callison-Burch, C.; and Resnik, P. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, 381–390.

Mehdad, Y.; Negri, M.; and Federico, M. 2011. Using bilingual parallel corpora for cross-lingual textual entailment. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 1336–1345. Association for Computational Linguistics.

Moldovan, D.; Paşca, M.; Harabagiu, S.; and Surdeanu, M. 2003. Performance issues and error analysis in an open-domain question answering system. *ACM Trans. Inf. Syst.* 21(2):133–154.

Och, F. J., and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.* 29(1):19–51.

Odijk, D.; Meij, E.; and de Rijke, M. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, 9–16.

Ponte, J. M., and Croft, W. B. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, 275–281.

Quirk, C.; Brockett, C.; and Dolan, W. B. 2004. Monolingual machine translation for paraphrase generation. In *EMNLP*, 142–149.

Stolcke, A. 2002. Srilm - an extensible language modeling toolkit. In *INTERSPEECH*.

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

Wei, X., and Croft, W. B. 2006. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, 178–185. New York, NY, USA: ACM.

Xu, J., and Croft, W. B. 1996. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, 4–11.

Xue, X.; Jeon, J.; and Croft, W. B. 2008. Retrieval models for question and answer archives. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, 475–482.

Zhou, G.; Liu, Y.; Liu, F.; Zeng, D.; and Zhao, J. 2013. Improving question retrieval in community question answering using world knowledge. In *IJCAI*, 2239–2245. AAAI Press.