Exploiting Bilingual Translation for Question Retrieval in Community-Based Question Answering

Guangyou Zhou, Kang Liu and Jun Zhao
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Beijing 100190, China
{gyzhou, kliu, jzhao}@nlpr.ia.ac.cn

ABSTRACT
Community-based question answering (CQA) has become an important issue due to the popularity of CQA archives on the web. This paper is concerned with the problem of question retrieval. Question retrieval in CQA archives aims to find historical questions that are semantically equivalent or relevant to the queried questions. However, question retrieval is challenging partly due to the word ambiguity and lexical gap between the queried questions and the historical questions in the archives. To deal with these problems, we propose the use of translated words to enrich the question representation, going beyond the words in the original language to represent a question. In this paper, each original language question (e.g., English) is automatically translated into an foreign language (e.g., Chinese) by machine translation services, and the resulting translated questions serves as a semantically enhanced representation for supplementing the original bag of words. Experiments conducted on real CQA data set demonstrate that our proposed approach significantly outperforms several baseline methods and achieves the state-of-the-art performance.

TITLE AND ABSTRACT IN ANOTHER LANGUAGE, L_2 (OPTIONAL, AND ON SAME PAGE)

利用双语翻译对社区问答进行问题检索

由于互联网上社区问答数据集的流行，使得社区问答的研究变得越来越流行。本文关注的是问题检索。 问题检索的目的是从历史问题数据集中查找与查询问题语义等价或相关的历史问题。然而，问题检索的挑战主要是词汇歧义和查询问题与历史问题 之间的词汇鸿沟。为了解决这些问题，我们提出利用翻译词来丰富问题的表示，而不单纯利用原始语言的词来表示问题。 在本文中，通过机器翻译，每个原始语言（例如：英语）的问题都被自动翻译成另一种外国语言（例如：汉语）， 经过翻译后的问题可以作为一种增强的语义表示来辅助原始的基于词袋的表示方法。在真实社区问答数据集上的实验表明，我们的方法可以极大提升基线系统的方法 并取得了最好的性能。

KEYWORDS: Community Question Answering, Question Retrieval, Bilingual Translation.

KEYWORDS IN L_2: 社区问答，问题检索，双语翻译

Proceedings of COLING 2012: Technical Papers, pages 3153–3170,
COLING 2012, Mumbai, December 2012.

1 引言

在过去的若干年中，大规模的问答数据集成了互联网上的重要信息资源。 这些资源包括传统的由专家或公司为他们的产品提供的常见问题解答集以及新出现的基于社区的在线服务，例如Yahoo! Answers和Live QnA，在这些在线社区上，人们可以回答他人提出的问题。这种在线社区称为基于社区的问答服务。 在这些社区中，任何人都可以提问和回答关于任何主题的问题，寻找信息的人与那些知道答案的人就联系起来了。 由于社区问答上的答案通常以显式的形式由人们提供，它们对回答真实问题起到了很好的作用 (Wang et al., 2009)。

为了更好地利用大规模的问答对，具备帮助用户检索先前答案的功能非常必要 (Duan et al., 2008)。因此， 检索与查询问题语义等价或相关的问题是一件非常有意义的任务。然而，问题检索的挑战主要是词汇歧义和查询问题与历史问题 之间的词汇鸿沟。词汇歧义通常会引发问题检索模型检索出许多与用户查询意图不匹配的历史问题。这也是由问题和用户的高度多样化造成的。例如，依据不同的用户，词"interest"既可以指"curiosity"也可以指"a charge for borrowing money"。另外一个挑战是查询问题与历史问题的词汇鸿沟。查询问题中的词不同于历史问题中的词但是它们之间是相关的词。 词汇鸿沟问题对社区问答的问题检索而言更加严重，主要是问答对通常很短，查找相同的内容表达往往使用不同的词(Xue et al., 2008)。

为了解决词汇鸿沟问题，大多数学者将问题检索任务看作是一个统计机器翻译的问题，并利用IBM模型1(Brown et al., 1993)来学习词与词之间的翻译概率(Berger et al., 2000; Jeon et al., 2005; Xue et al., 2008; Lee et al., 2008; Bernhard and Gurevych, 2009)。 实验结果一致表明基于词的翻译模型取得了比传统检索方法更好的性能。 最近，Riezler et al. (2007)和Zhou et al. (2011)提出了基于统计短语翻译的问题和答案检索方法。基于短语的翻译模型可以刻画上下文信息，在翻译的过程中对整个短语建模， 从而在某种程度上降低了词汇歧义的问题。然而，目前公开发表的工作都是基于单语的方法，仅仅利用了原始语言的信息， 而没有利用来自其它语言潜在的丰富的语义信息。通过其它语言，可以利用各种方法增加原始问题的语义信息，从而提高仅仅利用原始语言方法的性能。

通过利用外国语言，我们提出利用翻译表示通过外国语言词汇来替换原始语言中的词，其中外国语言是指不同于原始语言的。 利用双语信息进行问题检索的基本思想如下：（1）从一种语言翻译成另一种语言的过程中可以利用上下文信息，如表1所示，英文单词"interest"和 "bank"在不同的上下文中有多种意思，在利用**Google Translate** (Google-Trans)翻译的过程中正确的意思可以得到纠正。因此，问题中词的歧义在翻译的过程中可以根据上下文信息得到解决。（2）多个语言相关的词在某种语言中可以被翻成另外一种语言的唯一表示。如表1所示，英文单词例如"company"和"firm"可以被翻译成中文单词"公司 (gōngsī)"，"rheum"和"catarrh"可以被翻译成中文单词"感冒(gǎnmào)"。

在本文中，通过机器翻译，每个原始语言（例如：英语）的问题都被自动翻译成另一种外国语言（例如：汉语）， 经过翻译后的问题可以作为一种增强的语义表示来辅助原始的基于词袋的表示方法。具体来说， 原始语言与外国语言的词汇之间通过翻译联系起来，对解决上述两个问题的解决起到重要的作用。 首先，每个原始语言句子中的词可以被翻译成另一种语言中的多个词，因此在给定原始语言中词的上下文的情况下， 词汇歧义在翻译的过程中可以得到解决。同时，语义相关的多个词可以被翻译成另一外国语言中的一个词。 因此，原始语言中的词汇鸿沟在某种程度上可以通过另一种外国语言中的翻译词来解决。

我们利用来自Yahoo! Answers的大规模数据集做实验。采用两种商业翻译服务（例如，Google Translate和Yahoo Babel Fish和一种基于词典的基线翻译将大规模的英文问题翻译成中文问题。 实验表明，我们的方法可以极大提升基线系统的方法并取得了最好的

	英语	汉语
词汇歧义	How do I get a loan from a **bank**?	我(wǒ) 如何(rúhé) 从(cóng) **银行(yí nháng)** 贷款(dàikuǎn) ？
	How to reach the **bank** of the river?	如何(rúhé) 前往(qiánwǎng) **河岸(héàn)** ？
词汇鸿沟	company firm	公司(gōngsī) 公司(gōngsī)
	rheum catarrh	感冒(gǎnmào) 感冒(gǎnmào)

Table 1: 谷歌翻译（Google translate）：一些例子。

性能。

论文的组织结构如下。第三部分介绍了我们方法的框架。第四部分详细介绍了我们的方法。第五部分给出了实验结果。 在第六部分，我们总结了全文并对未来工作做了展望。

2 Introduction

Over the past few years, large-scale question and answer archives have become an important information resource on the Web. These include the traditional FAQ archives constructed by the experts or companies for their products and the emerging community-based online services, such as Yahoo! Answers[1] and Live QnA[2], where people answer questions posed by other people. This is referred as the community-based question answering services. In these communities, anyone can ask and answer questions on any topic, and people seeking information are connected to those who know the answers. As answers are usually explicitly provided by human, they can be helpful in answering real world questions (Wang et al., 2009).

To make use of the large-scale archives of question-answer pairs, it is critical to have functionality of helping users to retrieve previous answers (Duan et al., 2008). Therefore, it is a meaningful task to retrieve the semantically equivalent or relevant questions to the queried questions. However, question retrieval is challenging partly due to the **word ambiguity** and **lexical gap** between the queried questions and the historical questions in the archives. **Word ambiguity** often causes a question retrieval model to retrieve many historical questions that do not match the user's intent. This problem is also amplified by the high diversity of questions and users. For example, depending on different users, the word "interest" may refer to "curiosity", or "a charge for borrowing money". Another challenge is **lexical gap** between the queried questions and the historical questions. The queried questions may contain words that are different from, but related to, the words in the relevant historical questions. The lexical gap is substantially bigger for question retrieval in CQA largely due to the fact that the question-answer pairs are usually short and there is little chance of finding the same content expressed using different wording (Xue et al., 2008).

To solve the lexical gap problem, most researchers regarded the question retrieval task as a statistical machine translation problem by using IBM model 1 (Brown et al., 1993) to learn the word-to-word translation probabilities (Berger et al., 2000; Jeon et al., 2005; Xue et al., 2008; Lee et al., 2008; Bernhard and Gurevych, 2009). Experiments consistently reported that the word-based translation models could yield better performance than the traditional methods. Recently, Riezler et al. (2007) and Zhou et al. (2011) proposed a phrase-based translation model for question and answer retrieval. The phrase-based translation model can capture some contextual information in modeling the translation of phrases as a whole, thus the word ambiguity problem is somewhat alleviated. However, most existing studies in the literature are basically *monolingual approaches* which are restricted to the use of original language of questions, without taking advantage of potentially rich semantic information drawn from other languages. Through other languages, various ways of adding semantic information to a question could be available, thereby leading to potentially more improvements than using the original language only.

Taking a step toward using other languages, we propose the use of *translated representation* by alternatively presenting the original questions with the words of a foreign language, one that is different from the original language of questions. The idea of exploiting bilingual information for question retrieval is based on the following observations: (1) Contextual information is exploited during the translation from one language to another. As shown in Table 2, English words "interest" and "bank" that have multiple meanings under different contexts are correctly

[1] http://answers.yahoo.com
[2] http://qna.live.com

	English	Chinese
Word ambiguity	How do I get a loan from a **bank**?	我(wǒ) 如何(rúhé) 从(cóng) **银行(yínháng)** 贷款(dàikuǎn) ?
	How to reach the **bank** of the river?	如何(rúhé) 前往(qiánwǎng) **河岸(héàn)** ?
Lexical gap	company firm	公司(gōngsī) 公司(gōngsī)
	rheum catarrh	感冒(gǎnmào) 感冒(gǎnmào)

Table 2: Google translate: some illustrative examples.

addressed by **Google Translate**[3] (GoogleTrans). Thus, word ambiguity based on contextual information is naturally involved when questions are translated. (2) Multiple words that are semantically similar in one language may be translated into unique words or a few words in a foreign language. For example in Table 2, English words such as "company" and "firm" are translated into "公司 (gōngsī)", "rheum" and "catarrh" are translated into "感冒(gǎnmào)" in Chinese.

In this paper, each original question is automatically translated into a foreign language by machine translation services, and the resulting translated questions serve as a semantically enhanced representation for supplementing the original bag of words. Specially, the vocabularies of the original and foreign languages are connected via translation, which could bring about important benefits in dealing with the two addressed problems. First, an original language word can be translated into multiple candidate words in a foreign language. Therefore, the word ambiguity problem can be resolved during the translation in a given context of an original language word. Conversely, various different original language words that refer to similar meanings are translated into a single word or a few words in a foreign language. Thus, the lexical gap problem in the original language is to some extent ameliorated by using the translated words in a foreign language.

We conduct experiments on a large-scale data set from Yahoo! Answers. Two commercial machine translation services (e.g., Google Translate and Yahoo Babel Fish[4]) and a baseline dictionary-based system are used for translating English questions into Chinese questions. Experimental results show that our proposed method significantly outperforms several baseline methods and achieves state-of-the-art performance.

The remainder of this paper is organized as follows. Section 3 introduces the framework of the proposed method. Section 4 describes our proposed method in detail. Section 5 presents the experimental results. In Section 6, we conclude with ideas for future research.

3 Framework of the Proposed Approach

The framework of the proposed approach for question retrieval is summarized in Figure 1. Each historical question in original language (e.g., English) is translated into the corresponding foreign language (e.g., Chinese) via machine translation services. Note that in the framework, different machine translation services can be used to obtain different translation. When a queried question is given, the queried question is translated using the same machine translator. Next, question retrieval on both representations (e.g., English representation and Chinese

[3]http://translate.google.com/translate_t
[4]http://babelfish.yahoo.com/translte_txt

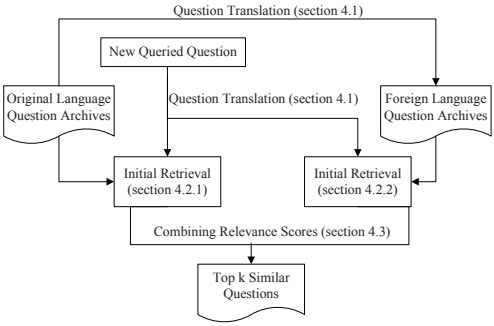

Figure 1: Framework of our approach by using question translated representation.

representation) is performed, and the two resulting relevance scores are combined to produce a final ranked list of semantically similar questions.

4 Our Approach

4.1 Question Translation

Translating historical questions in original language (e.g., English) into the corresponding foreign language is the first step of the proposed approach. Manual translation is time-consuming and labor-intensive, and it is not feasible to manually translate a large amount of questions in original language in real applications. Fortunately, machine translation techniques have been well developed in the NLP filed, though the translation quality is far from satisfactory. A few commercial machine translation services can be publicly accessed. In this paper, the following two commercial machine translation services and one baseline system are used to translate English questions into Chinese questions.

Google Translate (GoogleTrans): Google Translate is one of the state-of-the-art commercial machine translation systems used today. Google Translate employs statistical machine learning methods to build a translation model based on large-scale bilingual parallel corpus. Contextual information is utilized during the translation from one language text to the aligned text in anther language.

Yahoo Babel Fish (YahooTrans): Different from Google Translate, Yahoo Babel Fish uses SYSTRAN's rule-based translation engine. SYSTRAN is one of the earliest developers of machine translation software. SYSTRAN employs complex sets of specific rules defined by linguists to analyze and then transfer the grammatical structure of the source language into the target language. During the translation, word ambiguity can be resolved based on the contextual information.

Baseline Translate (DicTran): We simply develop a translation method based only on one-to-one word translation using an English to Chinese lexicon in StarDict[5].

[5]StarDict is an open source dictionary software, available at http://stardict.sourceforge.net/.

4.2 Bilingual Retrieval Method

4.2.1 Retrieval model

Language models have been performed quite well empirically in many information retrieval tasks (Zhai and Lafferty, 2001), and also have performed very well in question retrieval (Jeon et al., 2005; Cao et al., 2009, 2010). In this paper, we use the language modeling approach for question retrieval. In the language modeling approach to question retrieval, language models are constructed for each queried question $\mathbf{q}$ and each historical question $\mathbf{d}$ in CQA archives C. The historical questions in C are ranked by the distance to a given queried question $\mathbf{d}$ according to the language models. The most commonly used language model in question retrieval is the unigram model, in which words are assumed to be independent of each other.

One of the commonly used measures of the similarity between query model and historical question model is negative Kullback-Leibler (KL) divergence (Zhai and Lafferty, 2001). With unigram model, the negative KL-divergence between model $\theta_{\mathbf{q}}$ of query $\mathbf{q}$ and model $\theta_{\mathbf{d}}$ of historical question $\mathbf{d}$ is computed as follows:

$$\begin{aligned} Score(\mathbf{q},\mathbf{d}) &= -\sum_{w\in V} p(w|\theta_{\mathbf{q}}) \log \frac{p(w|\theta_{\mathbf{q}})}{p(w|\theta_{\mathbf{d}})} \\ &= \sum_{w\in V} p(w|\theta_{\mathbf{q}}) \log p(w|\theta_{\mathbf{d}}) - \sum_{w\in V} p(w|\theta_{\mathbf{q}}) \log p(w|\theta_{\mathbf{q}}) \\ &= \sum_{w\in V} p(w|\theta_{\mathbf{q}}) \log p(w|\theta_{\mathbf{d}}) + E(\theta_{\mathbf{q}}) \end{aligned} \tag{1}$$

where $p(w|\theta_{\mathbf{q}})$ and $p(w|\theta_{\mathbf{d}})$ are the generative probabilities of a word w from the models $\theta_{\mathbf{q}}$ and $\theta_{\mathbf{d}}$, V is the vocabulary of C, and $E(\theta_{\mathbf{q}})$ is the entropy of $\mathbf{q}$.

Let $tf(w,\mathbf{q})$ and $tf(w,\mathbf{d})$ as the frequencies of w in $\mathbf{q}$ and $\mathbf{d}$, respectively. Generally, $p(w|\theta_{\mathbf{q}})$ is calculated with maximum likelihood estimation (MLE):

$$p(w|\theta_{\mathbf{q}}) = \frac{tf(w,\mathbf{q})}{\sum_{w'\in\mathbf{q}} tf(w',\mathbf{q})} \tag{2}$$

To calculate $p(w|\theta_{\mathbf{d}})$, several smoothing methods have been proposed to overcome the data sparseness problem of a language model constructed from one historical question (Zhai and Lafferty, 2001). Therefore, $p(w|\theta_{\mathbf{d}})$ with the Dirichlet prior smoothing can be calculated as follows:

$$p(w|\theta_{\mathbf{d}}) = \frac{tf(w,\mathbf{d}) + \lambda p(w|\theta_C)}{\sum_{w'\in V} tf(w',\mathbf{d}) + \lambda} \tag{3}$$

where λ is the prior parameter in the Dirichlet prior smoothing method, and $p(w|\theta_{\mathbf{C}})$ is the probability of w in C, which is often computed with MLE:

$$p(w|\theta_C) = \frac{\sum_{\mathbf{d}\in C} tf(w,\mathbf{d})}{\sum_{\mathbf{d}\in C}\sum_{w'\in V} tf(w',\mathbf{d})} \tag{4}$$

4.2.2 Retrieval model for translated representation

We now extend the retrieval model described in subsection 4.2.1 in order to support translated representation. Let $\pi(\mathbf{d})$ be the translated representation result by using the machine translation service π (e.g., Google Translate) for a given historical question $\mathbf{d}$, and $\pi(\mathbf{q})$ be the translated representation result by using the machine translation service π for a queried question $\mathbf{q}$. Therefore, the query language model $p(w|\theta_{\pi(\mathbf{q})})$ based on the translated representation can be calculated as follows:

$$p(w|\theta_{\pi(\mathbf{q})}) = \frac{tf(w, \pi(\mathbf{q}))}{\sum_{w' \in \pi(\mathbf{q})} tf(w', \pi(\mathbf{q}))} \tag{5}$$

Similarly, by replacing $tf(w, \mathbf{d})$ in equation (3) with $tf(w, \pi(\mathbf{d}))$, we obtain the following smoothed model $p(w|\theta_{\pi(\mathbf{d})})$:

$$p(w|\theta_{\pi(\mathbf{d})}) = \frac{tf(w, \pi(\mathbf{d})) + \lambda p(w|\theta_{\pi(C)})}{\sum_{w' \in V_F} tf(w', \pi(\mathbf{d})) + \lambda} \tag{6}$$

where V_F is vocabulary of the translated foreign language, and $p(w|\theta_{\pi(\mathbf{C})})$ is defined by

$$p(w|\theta_{\pi(C)}) = \frac{\sum_{\mathbf{d} \in C} tf(w, \pi(\mathbf{d}))}{\sum_{\mathbf{d} \in C} \sum_{w' \in V_F} tf(w', \pi(\mathbf{d}))} \tag{7}$$

Finally, we calculate the relevance score of the historical question $\mathbf{d}$ with respect to the queried question $\mathbf{q}$ using $Score(\pi(\mathbf{q}), \pi(\mathbf{d}))$ based on their translated representation.

4.3 Combining Relevance Score for Bilingual Representation

After obtaining the two relevance scores from the original and translated representation perspective, we can rank the final similar historical questions based on the linear combination and refined ranking approach, respectively.

4.3.1 Linear combination

To produce a single ranked list from the two relevance scores using equation (1) on the original and translated representation, we use the following linear combination:

$$Score_{E+F}(\mathbf{q}, \mathbf{d}) = \alpha Score(\mathbf{q}, \mathbf{d}) + (1 - \alpha) Score(\pi(\mathbf{q}), \pi(\mathbf{d})) \tag{8}$$

In equation (8), the importance of relevance scores on the original and translated representation is adjusted through α. When $\alpha = 1$, the final retrieval model is based on the original representation. When $\alpha = 0$, the final retrieval model is based on the translated representation.

4.3.2 Refined ranking approach

Based on the original and translated representation, we can obtain two kinds of ranked lists $\vec{R}_E(\mathbf{q})$ and $\vec{R}_F(\pi(\mathbf{q}))$, which reflect the similarity between a queried question and a historical

questions from two different perspectives. If the retrieval model based on the original representation cannot capture the similarity due to the word ambiguity and lexical gap, then the retrieval model based on the translated representation should be good for dealing with the word ambiguity and lexical gap problems. Therefore, we consider a refined ranking approach to boost the question retrieval performance.

In order to measure the similarity between the two ranked results, we utilize a measurement, similar to Jaccard coefficient, which is defined as the size of the intersection divided by the size of the union of these two top k ranked results,

$$J = \frac{|\vec{R}_E(\mathbf{q}) \bigcap \vec{R}_F(\pi(\mathbf{q}))|}{|\vec{R}_E(\mathbf{q}) \bigcup \vec{R}_F(\pi(\mathbf{q}))|} \tag{9}$$

This measurement implies the following meaning: a large value is reached if the retrieval model based on the translated representation could retrieve many common relevant historical questions within the top-k results. Based on this scheme, we adopt a measurement for an adaptive ranking refinement. Let $R_E(\mathbf{q}, \mathbf{d})$ be the rank of historical question $\mathbf{d}$ for a given queried question $\mathbf{q}$, and let $R_F(\pi(\mathbf{q}), \pi(\mathbf{d}))$ be the rank of translated representation $\pi(\mathbf{d})$ for a given translated queried question $\pi(\mathbf{q})$. Therefore, we define a refined score $Score(\mathbf{q}, \mathbf{d})$ based on the following function:

$$Score(\mathbf{q}, \mathbf{d}) = \frac{1}{R_E(\mathbf{q}, \mathbf{d})} + \varphi(\mathbf{d}) \cdot J \cdot \frac{1}{\hat{R}_F(\pi(\mathbf{q}), \pi(\mathbf{d}))} \tag{10}$$

where $\varphi(\mathbf{d}) = 1$ if $\mathbf{d} \in \vec{R}_E(\mathbf{q})$ and $\pi(\mathbf{d}) \in \vec{R}_F(\pi(\mathbf{q}))$, otherwise $\varphi(\mathbf{d}) = 0$. By applying the refined ranking strategy, we obtain the refined ranking model shown in Algorithm 1.

Algorithm 1 Refined Model for Question Retrieval

Input: Given a queried question $\mathbf{q}$;

Step1: Retrieve the top-k most relevant historical questions based on the original representation using equations (1) to (4), and then obtain the ranked results $\vec{R}_E(\mathbf{q})$;

Step2: Retrieve the top-k most relevant historical questions based on the translated representation using equations (5) to (7), and then obtain the ranked results $\vec{R}_F(\pi(\mathbf{q}))$;

Step3: Refine with equation (10) and get the final ranked results.

Output: Return the ranked historical questions $\{\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_k\}$.

4.4 Category-Sensitive Language Model for Bilingual Representation

In CQA, when a user asks a question, the user typically needs to choose a category for the question from a predefined hierarchy of categories. Hence, each question in CQA archive has a category label and questions in CQA services are organized into hierarchies of categories (Cao et al., 2009, 2010). Based on these observations, it is naturally to employ the category information for bilingual representation. Let $c(\mathbf{d})$ be the leaf category of historical question $\mathbf{d}$, then *category-term frequency* of $\mathbf{d}$ for word w is defined as follows:

$$tf(w, \mathbf{d} \cup c(\mathbf{d})) = tf(w, \mathbf{d}) + \mu \cdot tf(w, c(\mathbf{d})) \tag{11}$$

where μ is the weight of category frequency, and $tf(w, c(\mathbf{d}))$ is frequency of word w in $c(\mathbf{d})$. Finally, model $\theta_{\mathbf{d}}$ defined in equation (3) is written as:

$$\begin{aligned} p(w|\theta_{\mathbf{d},c(\mathbf{d})}) &= \frac{tf(w, \mathbf{d} \cup c(\mathbf{d})) + \lambda p(w|\theta_C)}{\sum_{w' \in V} tf(w', \mathbf{d} \cup c(\mathbf{d})) + \lambda} \\ &= \frac{tf(w, \mathbf{d}) + \mu \cdot tf(w, c(\mathbf{d})) + \lambda p(w|\theta_C)}{\sum_{w' \in V} tf(w', \mathbf{d}) + \sum_{w' \in V} tf(w', c(\mathbf{d})) + \lambda} \end{aligned} \tag{12}$$

Similarly, we could define the translated representation for model $p(w|\theta_{\pi(\mathbf{d})})$ as follows:

$$p(w|\theta_{\pi(\mathbf{d}),\pi(c(\mathbf{d}))}) = \frac{tf(w, \pi(\mathbf{d})) + \mu \cdot tf(w, \pi(c(\mathbf{d}))) + \lambda p(w|\theta_{\pi(C)})}{\sum_{w' \in V} tf(w', \pi(\mathbf{d})) + \sum_{w' \in V} tf(w', \pi(c(\mathbf{d}))) + \lambda} \tag{13}$$

Given the bilingual representation, we again combine the two category-sensitive relevance scores with the above linear combination and refined ranking approach, respectively.

5 Experiments

5.1 Data Set and Evaluation Metrics

We collect the data set from Yahoo! Answers and use the *getByCategory* function provided in Yahoo! Answers API[6] to obtain CQA threads from the Yahoo! site. More specifically, we utilize the *resolved* questions and the resulting question repository that we use for question retrieval contains 2,288,607 questions. Each resolved question consists of four parts: "question title", "question description", "question answers" and "question category". For question retrieval, we only use the "question title" part. It is assumed that the titles of the questions already provide enough semantic information for understanding the users' information needs (Duan et al., 2008). There are 26 categories at the first level and 1,262 categories at the leaf level. Each question belongs to a unique leaf category. Table 3 shows the distribution across first-level categories of the questions in the archives.

Category	#Size	Category	# Size
Arts & Humanities	86,744	Home & Garden	35,029
Business & Finance	105,453	Beauty & Style	37,350
Cars & Transportation	145,515	Pet	54,158
Education & Reference	80,782	Travel	305,283
Entertainment & Music	152,769	Health	132,716
Family & Relationships	34,743	Sports	214,317
Politics & Government	59,787	Social Science	46,415
Pregnancy & Parenting	43,103	Ding out	46,933
Science & Mathematics	89,856	Food & Drink	45,055
Computers & Internet	90,546	News & Events	20,300
Games & Recreation	53,458	Environment	21,276
Consumer Electronics	90,553	Local Businesses	51,551
Society & Culture	94,470	Yahoo! Products	150,445

Table 3: Number of questions in each first-level category

We use the same test set in previous work (Cao et al., 2009, 2010). This set contains 252 queried questions and can be freely downloaded for research communities.[7] For each method, the top 20 retrieval results are kept. Given a returned result for each queried question, an

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

[7] The data set is available at http://homepages.inf.ed.ac.uk/gcong/qa/

annotator is asked to label it with "relevant" or "irrelevant". If a returned result is considered semantically equivalent to the queried question, the annotator will label it as "relevant"; otherwise, the annotator will label it as "irrelevant". Two annotators are involved in the annotation process. If a conflict happens, a third person will make judgement for the final result. In the process of manually judging questions, the annotators are presented only the questions.

Evaluation Metrics: We evaluate the performance of question retrieval using the following metrics: **Mean Average Precision** (MAP) and **Precision@N** (P@N). MAP rewards methods that return relevant questions early and also rewards correct ranking of the results. P@N reports the fraction of the top-N questions retrieved that are relevant. We perform a significant test, i.e., a t-test with a default significant level of 0.05.

Parameter Selection: We tune the parameters on a small development set of 50 questions. This development set is also extracted from Yahoo! Answers, and it is not included in the test set. For the smoothing parameter λ, we set $\lambda = 2000$ empirically in the language modeling approach for both English representation and Chinese translated representation. For parameter μ used in equation (11), we set $\mu = 0.8$ empirically. For parameter α, we do an experiment on the development set to determine the optimal values among 0.1, 0.2, $\cdots$, 0.9 in terms of MAP. As a result, we set $\alpha = 0.6$ in the experiments empirically as this setting yields the best performance. For parameter k described in Algorithm 1, we try several different values on the development set. Finally, we set $k = 30$ empirically as this setting gives the better performance.

5.2 Question Retrieval Results using Language Model

Table 4 shows a comparison of the results obtained using monolingual and bilingual representation using language model (LM) defined in subsection 4.2.1 and subsection 4.2.2 for question retrieval. In Table 4, **E** denotes the baseline LM using English representation (queried questions and historical questions). **C** denotes the run of LM using Chinese representation via English-Chinese translation (queried questions and historical questions). **E + C** denotes the run of LM with the combination of English and Chinese representation, where **Linear E + C** denotes the linear combination using equation 8, and **Refined E + C** denotes the refined ranking approach using equations 9 and 10. There are some clear trends in the results of Table 4:

Translation tools	#	Methods	MAP	P@10
-	1	E	0.385	0.242
GoogleTrans	2	C	0.350	0.234
	3	Linear E + C	0.468†	0.269†
	4	Refined E + C	**0.483**†	**0.275**†
YahooTrans	5	C	0.327	0.214
	6	Linear E + C	0.441†	0.258†
	7	Refined E + C	0.465†	0.267†
DicTran	8	C	0.246	0.178
	9	Linear E + C	0.398	0.246
	10	Refined E + C	0.414†	0.249

Table 4: Comparison of bilingual and monolingual representation using language model (LM) for question retrieval. The mark † indicates statistical significance over E.

(1) Using the bilingual translated representation, question retrieval performance can be significantly improved (row 1 vs. row 3 and row 4; row 1 vs. row 6 and row 7; row 1 vs. row 9 and

row 10). The reason is that various different words in English that refer to similar meanings can be translated into only a few words or a single word in Chinese. Thus, the lexical gap problem in English is to some extent ameliorated by using translated words in Chinese.

(2) We can see that question retrieval performance relies positively on the translated bilingual representation, and **GoogleTrans** performs the best while **DicTran** performs the worst (row 3 vs. row 9; row 4 vs. row 10), which is consistent with the fact **GoogleTrans** is deemed the best of the three machine translation systems, while **DicTran** is the weakest one. Moreover, **DicTran** performs translation without taking into account the surrounding words as contextual information, while **GoogleTrans** and YahooTrans are context-dependent and thus produce different translated Chinese words depending on the context of an English word. Therefore, the word ambiguity problem can be resolved during the English-Chinese translation in a given context of an English word.

(3) Comparing the two combination strategies **Linear** and **Refined**, it is seen that the refined ranking strategy (Refined E + C) gives the better results than linear combination regarding the different translation tools (row 3 vs. row 4; row 6 vs. row 7; row 9 vs. row 10).

5.3 Question Retrieval Results using Category-Sensitive Language Model

Table 5 shows the comparison results of monolingual and bilingual representation using category-sensitive language model (CSLM) for question retrieval. In Table 5, **E** denotes the baseline CSLM using the English representation only, and **E + C** denotes the run of CSLM based on the English and Chinese representation.

Translation tools	#	Methods	MAP	P@10
-	1	E	0.441	0.258
GoogleTrans	2	C	0.396	0.247
	3	Linear E + C	0.493†	0.282†
	4	Refined E + C	**0.525**†	**0.290**†
YahooTrans	5	C	0.358	0.237
	6	Linear E + C	0.476†	0.272†
	7	Refined E + C	0.492†	0.281†
DicTran	8	C	0.283	0.191
	9	Linear E + C	0.455	0.263
	10	Refined E + C	0.470†	0.270†

Table 5: Comparison of bilingual and monolingual representation using category-sensitive language model (CSLM) for question retrieval. The mark † indicates statistical significance over E.

Category-sensitive language model (CSLM) without considering the translated representation (e.g., row 1 in Table 5) is highly effective for question retrieval, achieving about 5.6% MAP increase over the baseline LM (e.g., row 1 in Table 4), with statistical significance. Similar findings have also been found by Cao et al. (2009) and Cao et al. (2010). Additionally, using the bilingual translated representation (E + C) achieves further improvements over CSLM (e.g., row 1 vs. row 3 and row 4). Specially, our refined ranking approach (Refined E + C) using GoogleTrans achieves about 8.4% further increase of MAP over the baseline CSLM (E) for question retrieval, finally leading to a noticeable increase of 14% MAP over the baseline LM.

5.4 Comparison with Different Methods

The motivation of this paper is to solve the lexical gap and word ambiguity problems for question retrieval. Jeon et al. (2005) proposed a word-based translation model for automatically fixing the lexical gap problem. Experimental results demonstrated that the word-based translation model significantly outperformed the traditional methods (i.e., VSM, BM25, LM). Xue et al. (2008) proposed a word-based translation language model for question retrieval. The results indicated that word-based translation language model further improved the retrieval results and obtained the state-of-the-art performance. Zhou et al. (2011) proposed a monolingual phrase-based translation model for question retrieval. This method can capture some contextual information in modeling the translation of phrases as a whole. To implement the word-based translation models, we use the GIZA++ alignment toolkit[8] trained on one million question-answer pairs from another data set[9] to learn the word-to-word translation probabilities. For phrase-based translation model described in (Zhou et al., 2011), we employ Moses toolkit[10] to extract the phrase translation and set the maximum length of phrases to 5. Recently, Singh (2012) extended the word-based translation model and explored strategies to learn the translation probabilities between words and the concepts using the CQA archives and a popular entity catalog. However, these existing studies in the literature are basically monolingual translation, which are restricted to the use of the original language of the CQA archives, without taking advantage of potentially rich semantic information drawn from other languages. In this paper, we propose the use of translated words to enrich question representation, going beyond the words in original language to represent the questions.

#	Methods	MAP	P@10
1	Jeon et al. (2005)	0.405	0.247
2	Xue et al. (2008)	0.436	0.261
3	Zhou et al. (2011)	0.452	0.268
4	Singh (2012)	0.450	0.267
5	**Refined E + C (LM, GoogleTrans)**	**0.483**†	**0.275**†

Table 6: Comparison with different methods for question retrieval without considering the category information. The mark † indicates statistical significance over previous work.

The comparisons with different methods for question retrieval are shown in Table 6. The results in Table 6 show that we propose the use of translated words to enrich question representation is much better than traditional monolingual approaches (row 1, row 2, row 3 and row 4 vs. row 5). Significant tests using t-test show the difference between our proposed approach and traditional monolingual approaches for cases marked in the table are statistically significant.

To further analyze why traditional monolingual approaches fail to give the satisfactory results for solving the word ambiguity and lexical gap problems, we identify two key challenges in adapting traditional monolingual translation approaches for question retrieval(Jeon et al., 2005; Xue et al., 2008; Zhou et al., 2011). First, unlike bilingual text, question-answer pairs are not semantically equivalent, leading to a wider range of possible phrases for a given phrase.

[8] http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

[9] The Yahoo! Webscope dataset Yahoo answers comprehensive questions and answers version 1.0, available at http://reseach.yahoo.com/Academic_Relations.

[10] http://www.statmt.org/moses/

Furthermore, both sides of question-answer parallel text are written in the same language (e.g., English). Thus, the most strongly associated word or phrase pairs found by the off-the-shelf word alignment and phrase extraction tools are identical pairs. Second, in question-answer pairs, there are far more unaligned words than in bilingual pairs. Also, there are more large phrase pairs that cannot be easily decomposed. These difficult cases confuse the IBM word alignment models. To the best of our knowledge, it is the first work to give a thorough analysis the key challenges in adapting traditional monolingual translation approaches for question retrieval.

Besides, we are aware of only two published studies (Cao et al., 2009) and (Cao et al., 2010) on utilizing category information for question retrieval. Now we compare our proposed category-sensitive language model (CSLM) for bilingual representation with these two studies. Cao et al. (2009) employed classifiers to compute the probability of a queried question belonging to different categories, and then incorporated the classified categories into language model for question retrieval. Cao et al. (2010) introduced the different combinations to compute the global relevance and local relevance, the combination VSM + TRLM showed the superior performance than others. In this paper, we compare the proposed method with the combination VSM + TRLM. To implement these two methods, we employ the same parameter settings with Cao et al. (2009) and Cao et al. (2010). Table 7 shows the comparison. From this table, we can see that our proposed category-sensitive language model (CSLM) for bilingual representation can significantly improve the performance. The results also validate the effectiveness of the proposed method.

#	Methods	MAP	P@10
1	Cao et al. (2009)	0.408	0.247
2	Cao et al. (2010)	0.456	0.269
3	**Refined E + C (CSLM, GoogleTrans)**	**0.525**†	**0.290**†

Table 7: Comparison with previous work for question retrieval by considering the category information. The mark † indicates statistical significance over previous work.

5.5 Parameter Sensitivity of Combination

To combine the relevance scores for question retrieval, we propose to use the linear combination and the refined ranking approach to rank the final similar questions. In linear combination, we use parameter α to control the relative importance of original question representation and translated representation. In refined ranking approach, we retrieve the top-k similar questions from two perspectives for each queried question. To investigate the effect of these two parameters, we design the following experiments.

To examine the effect of α, we choose the best translation service GoogleTrans and evaluate α with different values among 0, 0.1, $\cdots$, 0.9 in terms of MAP on a small development set of 50 questions. This development set is also extracted from Yahoo! Answers data, and it is not included in the test set. The experimental results for different α are illustrated in Figure 2. Monolingual baselines E and C are used for reference. Figure 2(Left) shows that, for MAP, **E + C** performs better than baselines E and C when $\alpha \in (0.2, 0.9)$. Therefore, a relative broad set of good parameter value is observed. When $\alpha = 0.6$, **E + C** gives the best performance.

To investigate the effect of parameter k, we also choose the best translation service GoogleTrans with several different values from 10 to 50 in terms of MAP on this development set.

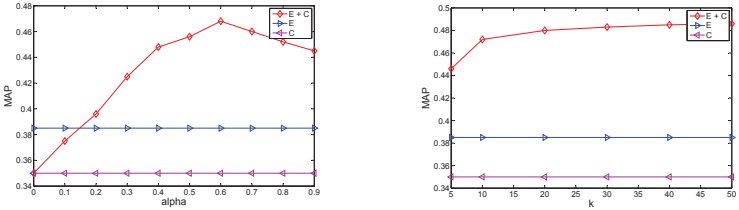

Figure 2: Left: The effect of parameter α for the linear combination using MAP metric; Right: The effect of parameter k for the refined ranking using MAP metric.

The experimental results for different k are illustrated in Figure 2(Right). We can see the performance becomes better for greater k used in the refined ranking approach. We believe the reason is that more historical questions may contain more similar questions. However, a larger k may result in longer processing time. Therefore, a good tradeoff is to set $k = 30$.

6 Related Work

6.1 Question Retrieval in CQA

The research of question retrieval has been further extended to the CQA data. The major challenge for question retrieval in CQA is the word ambiguity and lexical gap problems. Jeon et al. (2005) proposed a word-based translation model for automatically fixing the lexical gap problem. Xue et al. (2008) proposed a word-based translation language model for question retrieval. The results indicated that word-based translation language model further improved the retrieval results and obtained the state-of-the-art performance. Subsequent work on word-based translation models focused on providing suitable parallel data to learn the translation probabilities. Lee et al. (2008) tried to further improve the translation probabilities based on question-answer pairs by selecting the most important terms to build compact translation models. Bernhard and Gurevych (2009) proposed to use as a parallel training data set the definitions and glosses provided for the same term by different lexical semantic resources. Cao et al. (2010) explored the category information into the word-based translation model for question retrieval.

Recently, Riezler et al. (2007) and Zhou et al. (2011) proposed a phrase-based translation model for question and answer retrieval. The phrase-based translation model can capture some contextual information in modeling the translation of phrases as a whole, thus the word ambiguity and lexical gap problems are somewhat alleviated. Singh (2012) addressed the lexical gap issues by extending the lexical word-based translation model to incorporate semantic information (entities).

However, most existing works in the literature are basically monolingual approaches which are restricted to the use of the original language of the CQA archives, without taking advantage of potentially rich semantic information drawn from other languages. In this paper, we intend to address two fundamental issues in question retrieval: word ambiguity and lexical gap. To solve these problems, we enrich the question representation via bilingual translation. Compared to the traditional monolingual approaches, our proposed bilingual translation is much more effective due to the recent advance in statistical machine translation. To the best of our knowledge, it is the first work to improve question retrieval in CQA via bilingual translation.

6.2 WSD and Query Expansion for Monolingual Information Retrieval

Besides in CQA, word ambiguity and lexical gap have been investigated in information retrieval (IR). Zhong and Ng (2012) proposed a novel approach to incorporate word senses into the language modeling approach to IR. Experimental results showed that word sense disambiguation (WSD) can significantly improve a state-of-the-art IR system. Query expansion has been one of the most effective approaches to resolve the lexical gap problem, which enrich the original query by adding some additional words (Lv and Zhai, 2010; Xu et al., 2009). Recently, Trieschnigg et al. (2010) enriched the original word-based representation with a concept-based representation, thereby proposing the translation of the original word language to a concept language. However, their translation models are based solely on the use of translation at the lexical level (e.g., word-to-concept), and thus their method is very different from our context-dependent style of translation. Na and Ng (2011) also applied automatic translation for monolingual retrieval. However, they used the expected frequency of a word computed from all possible translated representations, while we use the state-of-the-art commercial machine translation service (e.g., Google Tranlate), which is much simpler than their translation strategies.

6.3 Machine Translation for Cross-Lingual Information Retrieval

Cross-lingual retrieval information retrieval (CLIR) addresses the problem of retrieving documents written in a language different from the query language. The common approach in CLIR is to perform query translation or document translation using a machine translation system(Chen and Gey, 2004; Kraaij et al., 2003). However, the major difference is that our goal is to improve monolingual question retrieval and not CLIR. Moreover, these studies performed translation without taking into account the context information of an original word(Chen and Gey, 2004; Kraaij et al., 2003). On the contrary, our approach is context-dependent and thus produces different translated words depending on the context of a word in original language.

Conclusion and Future Work

In this paper, we intend to address two fundamental issues in question retrieval: word ambiguity and lexical gap. To solve these problems, we propose the use of bilingual question representation, encouraged by the fact that a translated word in a foreign language can be used to enrich the original question representation. We employ the statistical machine translation services to automatically translate all questions, producing bilingual representations. Then, the relevance score between a queried question and a historical question is computed by combining two evidences derived from the bilingual perspectives. Experimental results conducted on large-scale CQA data set from Yahoo! Answers show that by using English-Chinese translation, our approach achieves improvements over monolingual approaches, and the improvements are in many cases statistically significant.

There are some ways in which this research could be continued. First, we would like to extend the current experiments by considering other languages (e.g., English-French, Chinese-English, etc.). We want to see how strongly the linguistic diversity between original and foreign languages affects question retrieval performance. Second, we will try to investigate the use of the proposed approach for other kinds of data set, such as categorized questions from forum sites and FAQ sites.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61070106, No. 61272332 and No. 61202329), the National High Technology Development 863 Program of China (No. 2012AA011102), the National Basic Research Program of China (No. 2012CB316300), Tsinghua National Laboratory for Information Science and Technology (TNList). We thank the anonymous reviewers for their insightful comments. We also thank Dr. Gao Cong for providing the data set.

References

Berger, A., Caruana, R., Cohn, D., Freitag, D., and Mittal, V. (2000). Bridging the lexical chasm: statistical approach to answer-finding. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, pages 192–199.

Bernhard, D. and Gurevych, I. (2009). Combining lexical semantic resources with question & answer archives for translation-based answer finding. In *Annual Meeting of the Association for Computational Linguistics (ACL 2009)*, pages 728–736.

Brown, P., Pietra, V., Pietra, S., and Mercer, R. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Cao, X., Cong, G., Cui, B., and Jensen, C. (2010). A generalized framework of exploring category information for question retrieval in community question answer archives. In *International Conference on World Wide Web (WWW 2010)*, pages 201–210.

Cao, X., Cong, G., Cui, B., Jensen, C., and Zhang, C. (2009). The use of categorization information in language models for question retrieval. In *ACM Conference on Information and Managemenet (CIKM 2009)*, pages 265–274.

Chen, A. and Gey, F. (2004). Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval*, 7(1).

Duan, H., Cao, Y., Lin, C., and Yu, Y. (2008). Searching questions by identifying questions topics and question focus. In *Annual Meeting of the Association for Computational Linguistics (ACL 2008)*, pages 156–164.

Jeon, J., Croft, W. B., and Lee, J. H. (2005). Finding similar questions in large question and answer archives. In *ACM Conference on Information and Managemenet (CIKM 2005)*, pages 84–90.

Kraaij, W., Nie, J., and Simard, M. (2003). Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics*, 29.

Lee, J. T., Kim, S. B., Song, Y. I., and Rim, H. C. (2008). Bridging lexical gaps between queries and questions on large online q&a collections with compact translation models. In *Empirical Methods on Natural Language Processing (EMNLP 2008)*, pages 410–418.

Lv, Y. and Zhai, C. (2010). Positional relevance model for pseudo-relevance feedback. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*.

Na, S. and Ng, H. (2011). Enriching document representation via translation for improved monolingual information retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*.

Riezler, S., Vasserman, A., Tsochantaridis, I., Mittal, V., and Liu, Y. (2007). Statistical machine translation for query expansion in answer retrieval. In *Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 464–471.

Singh, A. (2012). Entity based q&a retrieval. In *Empirical Methods on Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pages 1266–1277.

Trieschnigg, D., Hiemstra, D., Jong, F., and Kraajj, W. (2010). A cross-lingual framework for monolingual biomeidcal information retrieval. In *ACM Conference on Information and Managemenet (CIKM 2010)*.

Wang, K., Ming, Z., and Chua, T.-S. (2009). A syntactic tree matching approach to finding similar questions in community-based qa services. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, pages 187–194.

Xu, Y., Jones, G., and Wang, B. (2009). Query dependent pseudo-relevance feedback based on wikipedia. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*.

Xue, X., Jeon, J., and Croft, W. B. (2008). Retrieval models for question and answer archives. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pages 475–482.

Zhai, C. and Lafferty, J. (2001). A study of smooth methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pages 334–342.

Zhong, Z. and Ng, H. (2012). Word sense disambiguation improves information retrieval. In *Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 273–282.

Zhou, G., Cai, L., Zhao, J., and Liu, K. (2011). Phrase-based translation model for question retrieval in community question answer archives. In *Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 653–662.