Statistical Machine Translation Improves Question Retrieval in Community Question Answering via Matrix Factorization

Guangyou Zhou, Fang Liu, Yang Liu, Shizhu He, and Jun Zhao
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Beijing 100190, China
{gyzhou,fliu,liuyang09,shizhu.he,jzhao}@nlpr.ia.ac.cn

Abstract

Community question answering (CQA) has become an increasingly popular research topic. In this paper, we focus on the problem of question retrieval. Question retrieval in CQA can automatically find the most relevant and recent questions that have been solved by other users. However, the word ambiguity and word mismatch problems bring about new challenges for question retrieval in CQA. State-of-the-art approaches address these issues by implicitly expanding the queried questions with additional words or phrases using monolingual translation models. While useful, the effectiveness of these models is highly dependent on the availability of quality parallel monolingual corpora (e.g., question-answer pairs) in the absence of which they are troubled by noise issue. In this work, we propose an alternative way to address the word ambiguity and word mismatch problems by taking advantage of potentially rich semantic information drawn from other languages. Our proposed method employs statistical machine translation to improve question retrieval and enriches the question representation with the translated words from other languages via matrix factorization. Experiments conducted on a real CQA data show that our proposed approach is promising.

1 Introduction

With the development of Web 2.0, community question answering (CQA) services like Yahoo! Answers,[1] Baidu Zhidao[2] and WkiAnswers[3] have attracted great attention from both academia and industry (Jeon et al., 2005; Xue et al., 2008; Adamic et al., 2008; Wang et al., 2009; Cao et al., 2010). In CQA, 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.

In this paper, we focus on the task of question retrieval. Question retrieval in CQA can automatically find the most relevant and recent questions (historical questions) that have been solved by other users, and then the best answers of these historical questions will be used to answer the users' queried questions. However, question retrieval is challenging partly due to the **word ambiguity** and **word mismatch** between the queried questions and the historical questions in the archives. **Word ambiguity** often causes the retrieval models to retrieve many historical questions that do not match the users' 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 **word mismatch** 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. For example, if a queried question contains the word "company" but a relevant historical question instead contains the word "firm", then there is a mismatch and the historical

[1] http://answers.yahoo.com/
[2] http://zhidao.baidu.com/
[3] http://wiki.answers.com/

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 852–861,
Sofia, Bulgaria, August 4-9 2013.

	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)**？
word mismatch	company firm	公司(gōngsī) 公司(gōngsī)
	rheum catarrh	感冒(gǎnmào) 感冒(gǎnmào)

Table 1: Google translate: some illustrative examples.

question may not be easily distinguished from an irrelevant one.

Researchers have proposed the use of word-based translation models (Berger et al., 2000; Jeon et al., 2005; Xue et al., 2008; Lee et al., 2008; Bernhard and Gurevych, 2009) to solve the word mismatch problem. As a principle approach to capture semantic word relations, word-based translation models are built by using the IBM model 1 (Brown et al., 1993) and have been shown to outperform traditional models (e.g., VSM, BM25, LM) for question retrieval. Besides, Riezler et al. (2007) and Zhou et al. (2011) proposed the phrase-based translation models for question and answer retrieval. The basic idea is to capture the contextual information in modeling the translation of phrases as a whole, thus the word ambiguity problem is somewhat alleviated. However, all these existing studies in the literature are basically *monolingual approaches* which are restricted to the use of original language of questions. While useful, the effectiveness of these models is highly dependent on the availability of quality parallel monolingual corpora (e.g., question-answer pairs) in the absence of which they are troubled by noise issue. In this work, we propose an alternative way to address the word ambiguity and word mismatch problems by 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 enriching the original questions with the words from other languages. The idea of improving question retrieval with statistical machine translation is based on the following two observations: (1) Contextual information is exploited during the translation from one language to another. For example in Table 1, English words "interest" and "bank" that have multiple meanings under different contexts are correctly addressed by using the state-of-the-art translation tool −−**Google Translate**.[4] Thus, word ambiguity based on contextual information is naturally involved when questions are translated. (2) Multiple words that have similar meanings in one language may be translated into an unique word or a few words in a foreign language. For example in Table 1, English words such as "company" and "firm" are translated into "公司 (gōngsī)", "rheum" and "catarrh" are translated into "感冒(gǎnmào)" in Chinese. Thus, word mismatch problem can be somewhat alleviated by using other languages.

Although Zhou et al. (2012) exploited bilingual translation for question retrieval and obtained the better performance than traditional monolingual translation models. However, there are two problems with this enrichment: (1) enriching the original questions with the translated words from other languages increases the dimensionality and makes the question representation even more sparse; (2) statistical machine translation may introduce noise, which can harm the performance of question retrieval. To solve these two problems, we propose to leverage statistical machine translation to improve question retrieval via matrix factorization.

The remainder of this paper is organized as follows. Section 2 describes the proposed method by leveraging statistical machine translation to improve question retrieval via matrix factorization. Section 3 presents the experimental results. In section 4, we conclude with ideas for future research.

[4]http://translate.google.com/translate_t

2 Our Approach

2.1 Problem Statement

This paper aims to leverage statistical machine translation to enrich the question representation. In order to address the word ambiguity and word mismatch problems, we expand a question by adding its translation counterparts. Statistical machine translation (e.g., Google Translate) can utilize contextual information during the question translation, so it can solve the word ambiguity and word mismatch problems to some extent.

Let $L = \{l_1, l_2, \ldots, l_P\}$ denote the language set, where P is the number of languages considered in the paper, l_1 denotes the original language (e.g., English) while l_2 to l_P are the foreign languages. Let $D_1 = \{d_1^{(1)}, d_2^{(1)}, \ldots, d_N^{(1)}\}$ be the set of historical question collection in original language, where N is the number of historical questions in D_1 with vocabulary size M_1. Now we first translate each original historical question from language l_1 into other languages l_p ($p \in [2, P]$) by Google Translate. Thus, we can obtain $D_2, \ldots, D_P$ in different languages, and M_p is the vocabulary size of D_p. A question $d_i^{(p)}$ in D_p is simply represented as a M_p dimensional vector $\mathbf{d}_i^{(p)}$, in which each entry is calculated by tf-idf. The N historical questions in D_p are then represented in a $M_p \times N$ term-question matrix $\mathbf{D}_p = \{\mathbf{d}_1^{(p)}, \mathbf{d}_2^{(p)}, \ldots, \mathbf{d}_N^{(p)}\}$, in which each row corresponds to a term and each column corresponds to a question.

Intuitively, we can enrich the original question representation by adding the translated words from language l_2 to l_P, the original vocabulary size is increased from M_1 to $\sum_{p=1}^{P} M_p$. Thus, the term-question matrix becomes $\mathbf{D} = \{\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_P\}$ and $\mathbf{D} \in \mathbb{R}^{(\sum_{p=1}^{P} M_p) \times N}$. However, there are two problems with this enrichment: (1) enriching the original questions with the translated words from other languages makes the question representation even more sparse; (2) statistical machine translation may introduce noise.[5] To solve these two problems, we propose to leverage statistical machine translation to improve question retrieval via matrix factorization. Figure 1 presents the framework of our proposed method, where q_i represents a queried question, and $\mathbf{q}_i$ is a vector representation of q_i.

[5]Statistical machine translation quality is far from satisfactory in real applications.

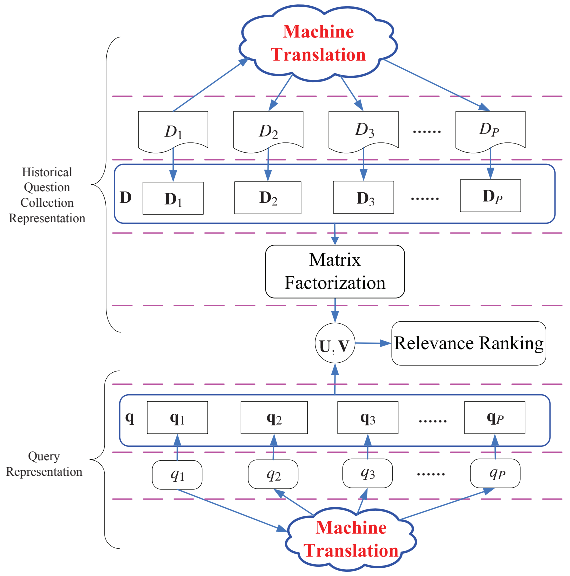

Figure 1: Framework of our proposed approach for question retrieval.

2.2 Model Formulation

To tackle the data sparseness of question representation with the translated words, we hope to find two or more lower dimensional matrices whose product provides a good approximate to the original one via matrix factorization. Previous studies have shown that there is psychological and physiological evidence for parts-based representation in the human brain (Wachsmuth et al., 1994). The non-negative matrix factorization (NMF) is proposed to learn the parts of objects like text documents (Lee and Seung, 2001). NMF aims to find two non-negative matrices whose product provides a good approximation to the original matrix and has been shown to be superior to SVD in document clustering (Xu et al., 2003; Tang et al., 2012).

In this paper, NMF is used to induce the reduced representation $\mathbf{V}_p$ of $\mathbf{D}_p$, $\mathbf{D}_p$ is independent on $\{\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_{p-1}, \mathbf{D}_{p+1}, \ldots, \mathbf{D}_P\}$. When ignoring the coupling between $\mathbf{V}_p$, it can be solved by minimizing the objective function as follows:

$$\mathcal{O}_1(\mathbf{U}_p, \mathbf{V}_p) = \min_{\mathbf{U}_p \geq 0, \mathbf{V}_p \geq 0} \|\mathbf{D}_p - \mathbf{U}_p \mathbf{V}_p\|_F^2 \quad (1)$$

where $\|\cdot\|_F$ denotes Frobenius norm of a matrix. Matrices $\mathbf{U}_p \in \mathbb{R}^{M_p \times K}$ and $\mathbf{V}_p \in \mathbb{R}^{K \times N}$ are the reduced representation for terms and questions in the K dimensional space, respectively.

To reduce the noise introduced by statistical machine translation, we assume that $\mathbf{V}_p$ from language $\mathbf{D}_p$ ($p \in [2, P]$) should be close to $\mathbf{V}_1$

from the original language $\mathbf{D}_1$. Based on this assumption, we minimize the distance between $\mathbf{V}_p$ ($p \in [2, P]$) and $\mathbf{V}_1$ as follows:

$$\mathcal{O}_2(\mathbf{V}_p) = \min_{\mathbf{V}_p \geq 0} \sum_{p=2}^{P} \|\mathbf{V}_p - \mathbf{V}_1\|_F^2 \tag{2}$$

Combining equations (1) and (2), we get the following objective function:

$$\begin{aligned}&\mathcal{O}(\mathbf{U}_1, \ldots, \mathbf{U}_P; \mathbf{V}_1, \ldots, \mathbf{V}_P) \qquad (3)\\ &= \sum_{p=1}^{P} \|\mathbf{D}_p - \mathbf{U}_p\mathbf{V}_p\|_F^2 + \sum_{p=2}^{P} \lambda_p \|\mathbf{V}_p - \mathbf{V}_1\|_F^2\end{aligned}$$

where parameter λ_p ($p \in [2, P]$) is used to adjust the relative importance of these two components. If we set a small value for λ_p, the objective function behaves like the traditional NMF and the importance of data sparseness is emphasized; while a big value of λ_p indicates $\mathbf{V}_p$ should be very closed to $\mathbf{V}_1$, and equation (3) aims to remove the noise introduced by statistical machine translation.

By solving the optimization problem in equation (4), we can get the reduced representation of terms and questions.

$$\begin{aligned}&\min \mathcal{O}(\mathbf{U}_1, \ldots, \mathbf{U}_P; \mathbf{V}_1, \ldots, \mathbf{V}_P) \qquad (4)\\ &\text{subject to} : \mathbf{U}_p \geq 0, \mathbf{V}_p \geq 0, p \in [1, P]\end{aligned}$$

2.3 Optimization

The objective function $\mathcal{O}$ defined in equation (4) performs data sparseness and noise removing simultaneously. There are $2P$ coupling components in $\mathcal{O}$, and $\mathcal{O}$ is not convex in both $\mathbf{U}$ and $\mathbf{V}$ together. Therefore it is unrealistic to expect an algorithm to find the global minima. In the following, we introduce an iterative algorithm which can achieve local minima. In our optimization framework, we optimize the objective function in equation (4) by alternatively minimizing each component when the remaining $2P - 1$ components are fixed. This procedure is summarized in Algorithm 1.

2.3.1 Update of Matrix $\mathbf{U}_p$

Holding $\mathbf{V}_1, \ldots, \mathbf{V}_P$ and $\mathbf{U}_1, \ldots, \mathbf{U}_{p-1}, \mathbf{U}_{p+1}, \ldots, \mathbf{U}_P$ fixed, the update of $\mathbf{U}_p$ amounts to the following optimization problem:

$$\min_{\mathbf{U}_p \geq 0} \|\mathbf{D}_p - \mathbf{U}_p\mathbf{V}_p\|_F^2 \tag{5}$$

Algorithm 1 Optimization framework

Input: $\mathbf{D}_p \in \mathbb{R}^{m_p \times N}$, $p \in [1, P]$
for $p = 1 : P$ **do**
  $\mathbf{V}_p^{(0)} \in \mathbb{R}^{K \times N} \leftarrow$ random matrix
  for $t = 1 : T$ **do** $\triangleright$ T is iteration times
    $\mathbf{U}_p^{(t)} \leftarrow \text{Update}\mathbf{U}(\mathbf{D}_p, \mathbf{V}_p^{(t-1)})$
    $\mathbf{V}_p^{(t)} \leftarrow \text{Update}\mathbf{V}(\mathbf{D}_p, \mathbf{U}_p^{(t)})$
  end for
  return $\mathbf{U}_p^{(T)}$, $\mathbf{V}_p^{(T)}$
end for

Algorithm 2 Update $\mathbf{U}_p$

Input: $\mathbf{D}_p \in \mathbb{R}^{M_p \times N}$, $\mathbf{V}_p \in \mathbb{R}^{K \times N}$
for $i = 1 : M_p$ **do**
  $\bar{\mathbf{u}}_i^{(p)*} = (\mathbf{V}_p\mathbf{V}_p^T)^{-1}\mathbf{V}_p\bar{\mathbf{d}}_i^{(p)}$
end for
return $\mathbf{U}_p$

Let $\bar{\mathbf{d}}_i^{(p)} = (d_{i1}^{(p)}, \ldots, d_{iK}^{(p)})^T$ and $\bar{\mathbf{u}}_i^{(p)} = (u_{i1}^{(p)}, \ldots, u_{iK}^{(p)})^T$ be the column vectors whose entries are those of the i^{th} row of $\mathbf{D}_p$ and $\mathbf{U}_p$ respectively. Thus, the optimization of equation (5) can be decomposed into M_p optimization problems that can be solved independently, with each corresponding to one row of $\mathbf{U}_p$:

$$\min_{\bar{\mathbf{u}}_i^{(p)} \geq 0} \|\bar{\mathbf{d}}_i^{(p)} - \mathbf{V}_p^T\bar{\mathbf{u}}_i^{(p)}\|_2^2 \tag{6}$$

for $i = 1, \ldots, M_p$.

Equation (6) is a standard least squares problems in statistics and the solution is:

$$\bar{\mathbf{u}}_i^{(p)*} = (\mathbf{V}_p\mathbf{V}_p^T)^{-1}\mathbf{V}_p\bar{\mathbf{d}}_i^{(p)} \tag{7}$$

Algorithm 2 shows the procedure.

2.3.2 Update of Matrix $\mathbf{V}_p$

Holding $\mathbf{U}_1, \ldots, \mathbf{U}_P$ and $\mathbf{V}_1, \ldots, \mathbf{V}_{p-1}, \mathbf{V}_{p+1}, \ldots, \mathbf{V}_P$ fixed, the update of $\mathbf{V}_p$ amounts to the optimization problem divided into two categories.

if $p \in [2, P]$, the objective function can be written as:

$$\min_{\mathbf{V}_p \geq 0} \|\mathbf{D}_p - \mathbf{U}_p\mathbf{V}_p\|_F^2 + \lambda_p\|\mathbf{V}_p - \mathbf{V}_1\|_F^2 \tag{8}$$

if $p = 1$, the objective function can be written as:

$$\min_{\mathbf{V}_p \geq 0} \|\mathbf{D}_p - \mathbf{U}_p\mathbf{V}_p\|_F^2 + \lambda_p\|\mathbf{V}_p\|_F^2 \tag{9}$$

Let $\mathbf{d}_j^{(p)}$ be the j^{th} column vector of $\mathbf{D}_p$, and $\mathbf{v}_j^{(p)}$ be the j^{th} column vector of $\mathbf{V}_p$, respectively. Thus, equation (8) can be rewritten as:

$$\min_{\{\mathbf{v}_j^{(p)} \geq 0\}} \sum_{j=1}^{N} \|\mathbf{d}_j^{(p)} - \mathbf{U}_p \mathbf{v}_j^{(p)}\|_2^2 + \sum_{j=1}^{N} \lambda_p \|\mathbf{v}_j^{(p)} - \mathbf{v}_j^{(1)}\|_2^2 \tag{10}$$

which can be decomposed into N optimization problems that can be solved independently, with each corresponding to one column of $\mathbf{V}_p$:

$$\min_{\mathbf{v}_j^{(p)} \geq 0} \|\mathbf{d}_j^{(p)} - \mathbf{U}_p \mathbf{v}_j^{(p)}\|_2^2 + \lambda_p \|\mathbf{v}_j^{(p)} - \mathbf{v}_j^{(1)}\|_2^2 \tag{11}$$

for $j = 1, \ldots, N$.

Equation (12) is a least square problem with L_2 norm regularization. Now we rewrite the objective function in equation (12) as

$$\mathcal{L}(\mathbf{v}_j^{(p)}) = \|\mathbf{d}_j^{(p)} - \mathbf{U}_p \mathbf{v}_j^{(p)}\|_2^2 + \lambda_p \|\mathbf{v}_j^{p} - \mathbf{v}_j^{(1)}\|_2^2 \tag{12}$$

where $\mathcal{L}(\mathbf{v}_j^{(1)})$ is convex, and hence has a unique solution. Taking derivatives, we obtain:

$$\frac{\partial \mathcal{L}(\mathbf{v}_j^{(p)})}{\partial \mathbf{v}_j^{(p)}} = -2\mathbf{U}_p^T(\mathbf{d}_j^{(p)} - \mathbf{U}_p \mathbf{v}_j^{(p)}) + 2\lambda_p(\mathbf{v}_j^{(p)} - \mathbf{v}_j^{(1)}) \tag{13}$$

Forcing the partial derivative to be zero leads to

$$\mathbf{v}_j^{(p)*} = (\mathbf{U}_p^T \mathbf{U}_p + \lambda_p \mathbf{I})^{-1} (\mathbf{U}_p^T \mathbf{d}_j^{(p)} + \lambda_p \mathbf{v}_j^{(1)}) \tag{14}$$

where $p \in [2, P]$ denotes the foreign language representation.

Similarly, the solution of equation (9) is:

$$\mathbf{v}_j^{(p)*} = (\mathbf{U}_p^T \mathbf{U}_p + \lambda_p \mathbf{I})^{-1} \mathbf{U}_p^T \mathbf{d}_j^{(p)} \tag{15}$$

where $p = 1$ denotes the original language representation.

Algorithm 3 shows the procedure.

2.4 Time Complexity Analysis

In this subsection, we discuss the time complexity of our proposed method. The optimization $\bar{\mathbf{u}}_i^{(p)}$ using Algorithm 2 should calculate $\mathbf{V}_p \mathbf{V}_p^T$ and $\mathbf{V}_p \bar{\mathbf{d}}_i^{(p)}$, which takes $O(NK^2 + NK)$ operations. Therefore, the optimization $\mathbf{U}_p$ takes $O(NK^2 + M_p NK)$ operations. Similarly, the time complexity of optimization $\mathbf{V}_i$ using Algorithm 3 is $O(M_p K^2 + M_p NK)$.

Another time complexity is the iteration times T used in Algorithm 1 and the total number of languages P, the overall time complexity of our proposed method is:

$$\sum_{p=1}^{P} T \times O(NK^2 + M_p K^2 + 2M_p NK) \tag{16}$$

For each language $\mathbf{D}_p$, the size of vocabulary M_p is almost constant as the number of questions increases. Besides, $K \ll \min(M_p, N)$, theoretically, the computational time is almost linear with the number of questions N and the number of languages P considered in the paper. Thus, the proposed method can be easily adapted to the large-scale information retrieval task.

Algorithm 3 Update $\mathbf{V}_p$

Input: $\mathbf{D}_p \in \mathbb{R}^{M_p \times N}$, $\mathbf{U}_p \in \mathbb{R}^{M_p \times K}$
$\mathbf{\Sigma} \leftarrow (\mathbf{U}_p^T \mathbf{U}_p + \lambda_p \mathbf{I})^{-1}$
$\mathbf{\Phi} \leftarrow \mathbf{U}_p^T \mathbf{D}_p$
if $p = 1$ **then**
  for $j = 1 : N$ **do**
    $\mathbf{v}_j^{(p)} \leftarrow \mathbf{\Sigma}\boldsymbol{\phi}_j$, $\boldsymbol{\phi}_j$ is the j^{th} column of $\mathbf{\Phi}$
  end for
end if
return $\mathbf{V}_1$
if $p \in [2, P]$ **then**
  for $j = 1 : N$ **do**
    $\mathbf{v}_j^{(p)} \leftarrow \mathbf{\Sigma}(\boldsymbol{\phi}_j + \lambda_p \mathbf{v}_j^{(1)})$
  end for
end if
return $\mathbf{V}_p$

2.5 Relevance Ranking

The advantage of incorporating statistical machine translation in relevance ranking is to reduce "word ambiguity" and "word mismatch" problems. To do so, given a queried question q and a historical question d from Yahoo! Answers, we first translate q and d into other foreign languages (e.g., Chinese, French etc.) and get the corresponding translated representation q_i and d_i ($i \in [2, P]$), where P is the number of languages considered in the paper. For queried question $q = q_1$, we represent it in the reduced space:

$$\mathbf{v}_{q_1} = \arg\min_{\mathbf{v} \geq 0} \|\mathbf{q}_1 - \mathbf{U}_1 \mathbf{v}\|_2^2 + \lambda_1 \|\mathbf{v}\|_2^2 \tag{17}$$

where vector $\mathbf{q}_1$ is the tf-idf representation of queried question q_1 in the term space. Similarly, for historical question $d = d_1$ (and its tf-idf representation $\mathbf{d}_1$ in the term space) we represent it in the reduced space as $\mathbf{v}_{d_1}$.

The relevance score between the queried question q_1 and the historical question d_1 in the reduced space is, then, calculated as the cosine similarity between $\mathbf{v}_{q_1}$ and $\mathbf{v}_{d_1}$:

$$s(q_1, d_1) = \frac{< \mathbf{v}_{q_1}, \mathbf{v}_{d_1} >}{\|\mathbf{v}_{q_1}\|_2 \cdot \|\mathbf{v}_{d_1}\|_2} \qquad (18)$$

For translated representation q_i ($i \in [2, P]$), we also represent it in the reduced space:

$$\mathbf{v}_{q_i} = \arg\min_{\mathbf{v}\geq 0} \|\mathbf{q}_i - \mathbf{U}_i\mathbf{v}\|_2^2 + \lambda_i\|\mathbf{v} - \mathbf{v}_{q_1}\|_2^2 \qquad (19)$$

where vector $\mathbf{q}_i$ is the tf-idf representation of q_i in the term space. Similarly, for translated representation d_i (and its tf-idf representation $\mathbf{d}_i$ in the term space) we also represent it in the reduced space as $\mathbf{v}_{d_i}$. The relevance score $s(q_i, d_i)$ between q_i and d_i in the reduced space can be calculated as the cosine similarity between $\mathbf{v}_{q_i}$ and $\mathbf{v}_{d_i}$.

Finally, we consider learning a relevance function of the following general, linear form:

$$Score(q, d) = \boldsymbol{\theta}^T \cdot \boldsymbol{\Phi}(q, d) \qquad (20)$$

where feature vector $\boldsymbol{\Phi}(q, d) = (s_{VSM}(q, d), s(q_1, d_1), s(q_2, d_2), \ldots, s(q_P, d_P))$, and $\boldsymbol{\theta}$ is the corresponding weight vector, we optimize this parameter for our evaluation metrics directly using the Powell Search algorithm (Paul et al., 1992) via cross-validation. $s_{VSM}(q, d)$ is the relevance score in the term space and can be calculated using Vector Space Model (VSM).

3 Experiments

3.1 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 question title already provides 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 2 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 2: Number of questions in each first-level category.

We use the same test set in previous work (Cao et al., 2009; Cao et al., 2010). This set contains 252 queried questions and can be freely downloaded for research communities.[7]

The original language of the above data set is English (l_1) and then they are translated into four other languages (Chinese (l_2), French (l_3), German (l_4), Italian (l_5)), thus the number of language considered is $P = 5$) by using the state-of-the-art translation tool $--$Google Translate.

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.

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 parameter K, we do an experiment on the development set to determine the optimal values among 50, 100, 150, $\cdots$, 300 in terms of MAP. Finally, we set $K = 100$ in the experiments empirically as this setting yields the best performance. For parameter λ_1, we set $\lambda_1 = 1$ empirically, while for parameter λ_i ($i \in [2, P]$), we set $\lambda_i = 0.25$ empirically and ensure that $\sum_i \lambda_i = 1$.

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

[7]http://homepages.inf.ed.ac.uk/gcong/qa/

#	Methods	MAP	P@10
1	VSM	0.242	0.226
2	LM	0.385	0.242
3	Jeon et al. (2005)	0.405	0.247
4	Xue et al. (2008)	0.436	0.261
5	Zhou et al. (2011)	0.452	0.268
6	Singh (2012)	0.450	0.267
7	Zhou et al. (2012)	0.483	0.275
8	SMT + MF ($P = 2, l_1, l_2$)	0.527	0.284
9	**SMT + MF ($P = 5$)**	**0.564**	**0.291**

Table 3: Comparison with different methods for question retrieval.

3.2 Question Retrieval Results

Table 3 presents the main retrieval performance. Row 1 and row 2 are two baseline systems, which model the relevance score using VSM (Cao et al., 2010) and language model (LM) (Zhai and Lafferty, 2001; Cao et al., 2010) in the term space. Row 3 and row 6 are monolingual translation models to address the word mismatch problem and obtain the state-of-the-art performance in previous work. Row 3 is the word-based translation model (Jeon et al., 2005), and row 4 is the word-based translation language model, which linearly combines the word-based translation model and language model into a unified framework (Xue et al., 2008). Row 5 is the phrase-based translation model, which translates a sequence of words as whole (Zhou et al., 2011). Row 6 is the entity-based translation model, which extends the word-based translation model and explores strategies to learn the translation probabilities between words and the concepts using the CQA archives and a popular entity catalog (Singh, 2012). Row 7 is the bilingual translation model, which translates the English questions from Yahoo! Answers into Chinese questions using Google Translate and expands the English words with the translated Chinese words (Zhou et al., 2012). For these previous work, we use the same parameter settings in the original papers. Row 8 and row 9 are our proposed method, which leverages statistical machine translation to improve question retrieval via matrix factorization. In row 8, we only consider two languages (English and Chinese) and translate English questions into Chinese using Google Translate in order to compare with Zhou et al. (2012). In row 9, we translate English questions into other four languages. There are some clear trends in the result of Table 3:

(1) Monolingual translation models significantly outperform the VSM and LM (row 1 and row 2 vs. row 3, row 4, row 5 and row 6).

(2) Taking advantage of potentially rich semantic information drawn from other languages via statistical machine translation, question retrieval performance can be significantly improved (row 3, row 4, row 5 and row 6 vs. row 7, row 8 and row 9, all these comparisons are statistically significant at $p < 0.05$).

(3) Our proposed method (leveraging statistical machine translation via matrix factorization, SMT + MF) significantly outperforms the bilingual translation model of Zhou et al. (2012) (row 7 vs. row 8, the comparison is statistically significant at $p < 0.05$). The reason is that matrix factorization used in the paper can effectively solve the data sparseness and noise introduced by the machine translator simultaneously.

(4) When considering more languages, question retrieval performance can be further improved (row 8 vs. row 9).

Note that Wang et al. (2009) also addressed the word mismatch problem for question retrieval by using syntactic tree matching. We do not compare with Wang et al. (2009) in Table 3 because previous work (Ming et al., 2010) demonstrated that word-based translation language model (Xue et al., 2008) obtained the superior performance than the syntactic tree matching (Wang et al., 2009). Besides, some other studies attempt to improve question retrieval with category information (Cao et al., 2009; Cao et al., 2010), label ranking (Li et al., 2011) or world knowledge (Zhou et al., 2012). However, their methods are orthogonal to ours, and we suspect that combining the category information or label ranking into our proposed method might get even better performance. We leave it for future research.

3.3 Impact of the Matrix Factorization

Our proposed method (SMT + MF) can effectively solve the data sparseness and noise via matrix factorization. To further investigate the impact of the matrix factorization, one intuitive way is to expand the original questions with the translated words from other four languages, without considering the data sparseness and noise introduced by machine translator. We compare our SMT + MF with this intuitive enriching method (SMT + IEM). Besides, we also employ our proposed matrix factorization to the original question representation (VSM + MF). Table 4 shows the comparison.

#	Methods	MAP	P@10
1	VSM	0.242	0.226
2	VSM + MF	0.411	0.253
3	SMT + IEM ($P = 5$)	0.495	0.280
4	SMT + MF ($P = 5$)	0.564	0.291

Table 4: The impact of matrix factorization.

(1) Our proposed matrix factorization can significantly improve the performance of question retrieval (row 1 vs. row2; row3 vs. row4, the improvements are statistically significant at $p < 0.05$). The results indicate that our proposed matrix factorization can effectively address the issues of data spareness and noise introduced by statistical machine translation.

(2) Compared to the relative improvements of row 3 and row 4, the relative improvements of row 1 and row 2 is much larger. The reason may be that although matrix factorization can be used to reduce dimension, it may impair the meaningful terms.

(3) Compared to VSM, the performance of SMT + IEM is significantly improved (row 1 vs. row 3), which supports the motivation that the word ambiguity and word mismatch problems could be partially addressed by Google Translate.

3.4 Impact of the Translation Language

One of the success of this paper is to take advantage of potentially rich semantic information drawn from other languages to solve the word ambiguity and word mismatch problems. So we construct a dummy translator (DT) that translates an English word to itself. Thus, through this translation, we do not add any semantic information into the original questions. The comparison is presented in Table 5. Row 1 (DT + MF) represents integrating two copies of English questions with our proposed matrix factorization. From Table 5, we have several different findings:

(1) Taking advantage of potentially rich semantic information drawn from other languages can significantly improve the performance of question retrieval (row 1 vs. row 2, row 3, row 4 and row 5, the improvements relative to DT + MF are statistically significant at $p < 0.05$).

(2) Different languages contribute unevenly for question retrieval (e.g., row 2 vs. row 3). The reason may be that the improvements of leveraging different other languages depend on the quality of machine translation. For example, row 3 is better than row 2 because the translation quality of English-French is much better than English-Chinese.

#	Methods	MAP
1	DT + MF (l_1, l_1)	0.352
2	SMT + MF ($P = 2, l_1, l_2$)	0.527
3	SMT + MF ($P = 2, l_1, l_3$)	0.553
4	SMT + MF ($P = 2, l_1, l_4$)	0.536
5	SMT + MF ($P = 2, l_1, l_5$)	0.545
6	SMT + MF ($P = 3, l_1, l_2, l_3$)	0.559
7	SMT + MF ($P = 4, l_1, l_2, l_3, l_4$)	0.563
8	SMT + MF ($P = 5, l_1, l_2, l_3, l_4, l_5$)	0.564

Table 5: The impact of translation language.

Method	Translation	MAP
SMT + MF ($P = 2, l_1, l_2$)	Dict	0.468
	GTrans	0.527

Table 6: Impact of the contextual information.

(3) Using much more languages does not seem to produce significantly better performance (row 6 and row 7 vs. row 8). The reason may be that inconsistency between different languages may exist due to statistical machine translation.

3.5 Impact of the Contextual Information

In this paper, we translate the English questions into other four languages using Google Translate (GTrans), which takes into account contextual information during translation. If we translate a question word by word, it discards the contextual information. We would expect that such a translation would not be able to solve the word ambiguity problem.

To investigate the impact of contextual information for question retrieval, we only consider two languages and translate English questions into Chinese using an English to Chinese lexicon (Dict) in StarDict[8]. Table 6 shows the experimental results, we can see that the performance is degraded when the contextual information is not considered for the translation of questions. The reason is that GTrans is context-dependent and thus produces different translated Chinese words depending on the context of an English word. Therefore, the word ambiguity problem can be solved during the English-Chinese translation.

4 Conclusions and Future Work

In this paper, we propose to employ statistical machine translation to improve question retrieval and

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

enrich the question representation with the translated words from other languages via matrix factorization. Experiments conducted on a real CQA data show some promising findings: (1) the proposed method significantly outperforms the previous work for question retrieval; (2) the proposed matrix factorization can significantly improve the performance of question retrieval, no matter whether considering the translation languages or not; (3) considering more languages can further improve the performance but it does not seem to produce significantly better performance; (4) different languages contribute unevenly for question retrieval; (5) our proposed method can be easily adapted to the large-scale information retrieval task.

As future work, we plan to incorporate the question structure (e.g., question topic and question focus (Duan et al., 2008)) into the question representation for question retrieval. We also want to further investigate the use of the proposed method 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), We thank the anonymous reviewers for their insightful comments. We also thank Dr. Gao Cong for providing the data set and Dr. Li Cai for some discussion.

References

L. Adamic, J. Zhang, E. Bakshy, and M. Ackerman. 2008. Knowledge sharing and yahoo answers: everyone knows and something. In *Proceedings of WWW*.

A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. 2000. Bridging the lexical chasm: statistical approach to answer-finding. In *Proceedings of SIGIR*, pages 192-199.

D. Bernhard and I. Gurevych. 2009. Combining lexical semantic resources with question & answer archives for translation-based answer finding. In *Proceedings of ACL*, pages 728-736.

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

X. Cao, G. Cong, B. Cui, C. Jensen, and C. Zhang. 2009. The use of categorization information in language models for question retrieval. In *Proceedings of CIKM*, pages 265-274.

X. Cao, G. Cong, B. Cui, and C. Jensen. 2010. A generalized framework of exploring category information for question retrieval in community question answer archives. In *Proceedings of WWW*, pages 201-210.

H. Duan, Y. Cao, C. Y. Lin, and Y. Yu. 2008. Searching questions by identifying questions topics and question focus. In *Proceedings of ACL*, pages 156-164.

C. L. Lawson and R. J. Hanson. 1974. Solving least squares problems. *Prentice-Hall*.

J. -T. Lee, S. -B. Kim, Y. -I. Song, and H. -C. Rim. 2008. Bridging lexical gaps between queries and questions on large online Q&A collections with compact translation models. In *Proceedings of EMNLP*, pages 410-418.

W. Wang, B. Li, and I. King. 2011. Improving question retrieval in community question answering with label ranking. In *Proceedings of IJCNN*, pages 349-356.

D. D. Lee and H. S. Seung. 2001. Algorithms for non-negative matrix factorization. In *Proceedings of NIPS*.

Z. Ming, K. Wang, and T. -S. Chua. 2010. Prototype hierarchy based clustering for the categorization and navigation of web collections. In *Proceedings of SIGIR*, pages 2-9.

J. Jeon, W. Croft, and J. Lee. 2005. Finding similar questions in large question and answer archives. In *Proceedings of CIKM*, pages 84-90.

C. Paige and M. Saunders. 1982. LSQR: an algorithm for sparse linear equations and sparse least squares. *ACM Transaction on Mathematical Software*, 8(1):43-71.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1992. *Numerical Recipes In C*. Cambridge Univ. Press.

S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of ACL*, pages 464-471.

A. Singh. 2012. Entity based q&a retrieval. In *Proceedings of EMNLP-CoNLL*, pages 1266-1277.

J. Tang, X. Wang, H. Gao, X. Hu, and H. Liu. 2012. Enriching short text representation in microblog for clustering. *Front. Comput.*, 6(1):88-101.

E. Wachsmuth, M. W. Oram, and D. I. Perrett. 1994. Recognition of objects and their component parts: responses of single units in the temporal cortex of teh macaque. *Cerebral Cortex*, 4:509-522.

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

B. Wang, X. Wang, C. Sun, B. Liu, and L. Sun. 2010. Modeling semantic relevance for question-answer pairs in web social communities. In *Proceedings of ACL*, pages 1230-1238.

W. Xu, X. Liu, and Y. Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of SIGIR*, pages 267-273.

X. Xue, J. Jeon, and W. B. Croft. 2008. Retrieval models for question and answer archives. In *Proceedings of SIGIR*, pages 475-482.

C. Zhai and J. Lafferty. 2001. A study of smooth methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR*, pages 334-342.

G. Zhou, L. Cai, J. Zhao, and K. Liu. 2011. Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of ACL*, pages 653-662.

G. Zhou, K. Liu, and J. Zhao. 2012. Exploiting bilingual translation for question retrieval in community-based question answering. In *Proceedings of COLING*, pages 3153-3170.

G. Zhou, Y. Liu, F. Liu, D. Zeng, and J. Zhao. 2013. Improving Question Retrieval in Community Question Answering Using World Knowledge. In *Proceedings of IJCAI*.