Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Improving Question Retrieval in Community Question Answering Using World Knowledge

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

Abstract

Community question answering (cQA), which provides a platform for people with diverse background to share information and knowledge, has become an increasingly popular research topic. In this paper, we focus on the task of question retrieval. The key problem of question retrieval is to measure the similarity between the queried questions and the historical questions which have been solved by other users. The traditional methods measure the similarity based on the bag-of-words (BOWs) representation. This representation neither captures dependencies between related words, nor handles synonyms or polysemous words. In this work, we first propose a way to build a *concept thesaurus* based on the semantic relations extracted from the world knowledge of Wikipedia. Then, we develop a unified framework to leverage these semantic relations in order to enhance the question similarity in the concept space. Experiments conducted on a real cQA data set show that with the help of Wikipedia thesaurus, the performance of question retrieval is improved as compared to the traditional methods.

1 Introduction

Over the past years, question answering (QA) has attracted much attention in natural language processing (NLP) and information retrieval (IR) fields. However, most of the QA researches mainly focus on locating the concise answer for a given factoid question in the related documents [Maybury, 2004; Wang *et al.*, 2010; Gupta and Gupta, 2012]. In real world, more complex questions are usually asked, and users are more willing to expect a longer and more comprehensive answer. In this situation, traditional QA systems fail to give the satisfactory answers.

With the development of Web 2.0, 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 (cQA) 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].

In cQA, the systems can directly return answers to the queried questions instead of a list of relevant documents, thus providing an effective alternative to the traditional adhoc information retrieval. To make full use of the large scale archives of question-answer pairs, it is critical to have functionality helping users to retrieve historical answers [Duan *et al.*, 2008]. Therefore, it is a meaningful task to retrieve the questions that are semantically equivalent or relevant to the queried questions.

The most widespread representations for question retrieval today are based on BOWs, These include various term-weighting retrieval schemes, such as tf-idf and BM25 [Robertson *et al.*, 1994]. The pertinent feature of these BOWs representation is that they represent individual words. However, we observe that the user generated questions via social media are always short texts. The limitations of lengths leads to the sparsity of the word features. In addition, the BOWs representation models do not provide sufficient word co-occurrence or context shared information for effective similarity measure. Because of this situation, the traditional similarity models (e.g., tf-idf, BM25, etc.) based on the BOWs representation are not effective for question retrieval.

To solve the above limitations, researchers have proposed the use of translation models [Berger *et al.*, 2000; Jeon *et al.*, 2005; Riezler *et al.*, 2007; Xue *et al.*, 2008; Lee *et al.*, 2008; Bernhard and Gurevych, 2009; Zhou *et al.*, 2011] to capture the semantic word or phrase relations. While useful, the effectiveness of these models in the literature 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 paper, we propose to use Wikipedia as the world knowledge for question re-

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

trieval, which surpasses other knowledge bases by the coverage of concepts, rich semantic information and up-to-date content. In particular, we first build an easy-to-use thesaurus from Wikipedia, which explicitly derive the concept relationships based on the structural knowledge in Wikipedia, including synonymy, polysemy, hypernymy and associative relation. The thesaurus facilitates the integration of the rich world knowledge of Wikipedia into questions, because it resolves synonyms and introduces more general and associative concepts which may help identify related topics between the queried questions and the historical questions. Besides, it provides a rich context for polysemy concept sense disambiguation. Then we treat the different relations in the thesaurus according to their different importance, in order to improve the traditional similarity measure for question retrieval. Experiments conducted on a real cQA data set show that with the help of Wikipedia thesaurus, the performance of question retrieval is improved as compared to the traditional methods.

The rest of this paper is organized as follows. Section 2 describes a way to build a concept thesaurus based on the semantic relations extracted from Wikipedia. Section 3 describes the method to leverage the world knowledge of Wikipedia for question retrieval. Section 4 presents the experimental results. In section 5, we conclude with ideas for future research.

2 Wikipedia Thesaurus

In this section, we propose a way to mine synonym, hypernym and associative relations explicitly for each concept through analyzing the rich links in Wikipedia, and build it as an easy-to-use thesaurus.

Wikipedia is today the largest encyclopedia in the world and surpasses other knowledge bases in its coverage of concepts, rich semantic knowledge and up-to-date content. Recently, Wikipedia has gained a wide interest in IR community and has been used for many problems ranging from document classification [Gebrilovich and Markovitch, 2006; Wang and Domeniconl, 2008; Wang *et al.*, 2007] to text clustering [Hu *et al.*, 2008; 2009a; 2009b]. Each article in Wikipedia describes a single topic: its title is a succinct, well-formed phrase that resembles a term in a conventional thesaurus [Milne *et al.*, 2006]. Each article belongs to at least one category, and hyperlinks between articles capture their semantic relations. These semantic relations include: equivalence (synonym), hierarchical relations (hypernym) and associative relation. However, Wikipedia is an open data resource built for human use, so it inevitable includes much noise and the semantic knowledge within it is not suitable for direct use in question retrieval in cQA. To make it clean and easy-to-use as a thesaurus, we first preprocess the Wikipedia data to collect Wikipedia concepts, and then explicitly derive relationships between Wikipedia based on the structural knowledge of Wikipedia.

2.1 Wikipedia Concepts

Each article of Wikipedia describes a single topic and its title can be used to represent the concept, e.g., "United States". However, some articles are meaningless – they are only used for Wikipedia management and administration, such as "1980s", "List of newspapers", etc. Following the literature [Hu *et al.*, 2008], we filter Wikipedia titles according to the rules described (titles satisfy one of below will be filtered):

- The article belongs to categories related to chronology, e.g., "Years", "Decades" and "Centuries".
- The first letter is not a capital one.
- The title is a single stopword.

2.2 Semantic Relations in Wikipedia

Wikipedia contains rich relation structures, such as synonym (redirect link pages), polysemy (disambiguation page), hypernym (hierarchical relation) and associative relation (internal page link). All these semantic relations are expressed in the form of hyperlinks between Wikipedia articles [Milne *et al.*, 2006].

Synonym

Wikipedia contains only one article for any given concept by using redirect hyperlinks to group equivalent concepts to the preferred one. These redirect links cope with capitalization and spelling variations, abbreviations, synonyms, and colloquialisms. Synonym in Wikipedia mainly comes from these redirect links. For example, "IBM" is an entry with a large number of redirect pages: synonyms (I.B.M, Big blue, IBM Corporation) [Cai *et al.*, 2011]. In addition, Wikipedia articles often mention other concepts, which already have corresponding articles in Wikipedia. The anchor text on each hyperlink may be different with the title of the linked article. Thus, anchor texts can be used as another source of synonym [Hu *et al.*, 2008].

Polysemy

In Wikipedia, disambiguation pages are provided for a polysemous concept. A disambiguation page lists all possible meanings associated with the corresponding concept, where each meaning is discussed in an article. For example, the disambiguation page of the term "IBM" lists 3 associated concepts, including "Inclusion body myositis", "Injection blow molding", and "International Business Machine" [Cai *et al.*, 2011].

Hypernym

In Wikipedia, both articles (concepts) and categories belong to at least one category, and categories are organized in a hierarchical structure. The resulting hierarchy is a directed acyclic graph, in which multiple categorization schemes coexist simultaneously [Milne *et al.*, 2006]. To extract the real hierarchical relations from Wikipedia categories, we utilize the methods proposed in [Ponzetto and Strube, 2007] to derive generic hierarchical relation from category links. Thus, we can get hypernym for each Wikipedia concept.

Associative Relation

Each Wikipedia article contains a lot of hyperlinks, which express relatedness between them. As Milne et al. [2006] mentioned that, links between articles are only tenuously related. For example, comparing the following two links: one from

the article "IBM" to the article "Apple Inc.", the other from the article "IBM" to the article "Software engineer" [Cai *et al.*, 2011]. It is clear that the former two articles are more related than the later pair. So how to measure the relatedness of hyperlinks within articles in Wikipedia is an important issue. In this paper, we adopt three measurements: *Content-based*, *Out-link category-based* and *Distance-based*, which has been described in [Wang *et al.*, 2007].

Content-based measure is based on the BOWs representation of Wikipedia articles. Clearly, this measure (denoted as S_{BOWs}) has the same limitations of the BOWs approach since it only considers the words appeared in text documents.

Out-link category-based measure (denoted as S_{OLC}) compares the out-link categories of two associative articles. The out-link category of a given article are the categories to which out-link articles from the original one belong.

Distance-based measure captures the length of the shortest path connecting the two categories they belong to, in the acyclic graph of the category taxonomy. This measure is normalized by taking into account the depth of the taxonomy and denoted as D_{cat}.

Following [Wang *et al.*, 2007; Wang and Domeniconl, 2008; Cai *et al.*, 2011], the overall strength of an associative relation between concepts can be written as:

$$S_{overall} = \lambda_1 S_{BOWs} + \lambda_2 S_{OLC} + (1 - \lambda_1 - \lambda_2)(1 - D_{cat}) \quad (1)$$

where λ_1 and λ_2 reflect the relative importance of the individual measure. Using equation (1), we rank all the out-linked concepts for each given concept. Then we denote the out-link concepts with relatedness above certain threshold as associative ones for each given concept.

3 Improving Question Retrieval with Wikipedia Concepts

3.1 Traditional BOWs Question Similarity

Traditional methods represent each question as BOWs. After removing the stopwords[3] and stemmed by Porter stemmer[4], the stemmed terms construct a tf-idf vector representation $\mathbf{q}$ for each queried question q. Similarly, for each historical question d, the stemmed terms also construct a tf-idf vector representation $\mathbf{d}$. Finally, the similarity between the queried question and the historical question in the vector space is, then, calculated as the cosine similarity between $\mathbf{q}$ and $\mathbf{d}$:

$$S_{term}(q, d) = \frac{< \mathbf{q}, \mathbf{d} >}{\|\mathbf{q}\|_2 \cdot \|\mathbf{d}\|_2} \quad (2)$$

3.2 Mapping Questions into Wikipedia Concepts

To use the Wikipedia thesaurus to enrich questions, one of the key issues is how to map words in questions to Wikipedia concepts. Following the literature [Hu *et al.*, 2008], we build a phrase index which includes the phrases of Wikipedia concepts, their synonym, and polysemy in Wikipedia thesaurus. Based on the generated Wikipedia phrases index, all candidate phrases can be recognized in the web page. We use the Forward Maximum Matching algorithm [Wong and Chan, 1996] to search candidate phrases, which is a dictionary-based word segmentation approach. By performing this process, it is necessary to do word sense disambiguation to find its most proper meaning mentioned in questions if a candidate concept is a polysemous one. Wang et al. [2007] proposed a disambiguation method by considering the document similarity and contextual information, and the experiments showed a high disambiguation accuracy. We adopt Wang et al. [2007]'s method to do word sense disambiguation for the polysemous concepts in questions.

Figure 1 shows an example of the identified Wikipedia concepts for question Q_1 using the above method [Cai *et al.*, 2011]. The phrase "software engineer" in Q_1 is mapped into Wikipedia concept "Software engineer", "Big Blue" in Q_1 is mapped into Wikipedia concept "IBM".

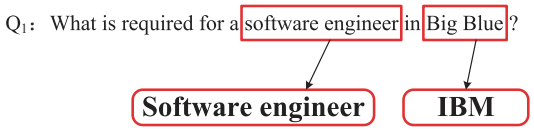

Figure 1: An example of the identified Wikipedia concepts for question Q_1.

3.3 Measuring Question Similarity with Hypernyms

In Wikipedia, each concept belongs to one or more categories. Moreover, these categories further belong to more higher level categories, forming an acyclic category graph [Hu *et al.*, 2008]. The set of categories contained in the category graph of a given concept c is represented as $Cate(c) = \{cate_{c_1}, \ldots, cate_{c_m}\}$. In the category graph, a category may have several paths linked to a concept. We calculate the distance $dis(c, cate_i)$ $(i \in [c_1, \ldots, c_m])$ by the length of the shortest path from concept c to category $cate_i$.

As noted by Hu et al. [2008], the higher level categories have less influence than those lower level categories since the lower level categories are more specific and therefore can depict the articles more accurate. In this paper, we present the influence of γth layer categories on concept c as $Inf_\gamma(c)$ and define $lnf_1(c) = 1$. For higher levels of categories, similar to [Hu *et al.*, 2008; Cai *et al.*, 2011], we introduce a decay factor $\mu \in [0, 1]$. Thus, we have $lnf_\gamma(c) = \mu lnf_{\gamma-1}(c) = \mu^{\gamma-1} lnf_1(c)$. As each Wikipedia concept has more than one categories, and each category has more than one parent categories, a large value of γ will introduce too many categories. Therefore, we set $\gamma \leq 3$ in our experiments. Thus, for each concept c we can build a category vector $\mathbf{cate}_c = \{lnf(c, cate_{c_1}), \ldots, lnf(c, cate_{c_m})\}$, where $lnf(c, cate_{c_1}) = lnf_{dis}(c, cate_{c_i}(c))$, which indicates the influence of category $cate_{c_i}$ on concept c. Let $\mathbf{C}_d$ denote the concept representation of historical question d, the corresponding category vector can be represented as $\mathbf{Cate}_d = \bigcup_{c \in \mathbf{C}_d} \mathbf{cate}_c$. Similarity, for queried question q, we also represent it in the category vector space as $\mathbf{Cate}_q$. The similarity between the queried question and the historical question in the category space is, then, calculated as the

[3]http://truereader.com/manuals/onix/stopwords1.html

[4]http://tartarus.org/martin/PorterStemmer/.

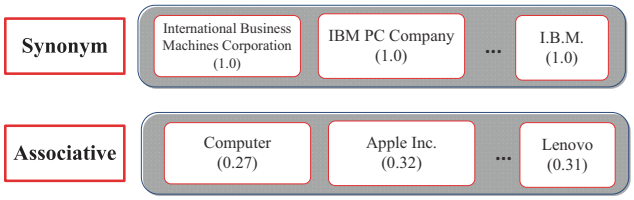

Figure 2: An example of the synonyms and associative concepts for Wikipedia concept "IBM".

cosine similarity between $\mathbf{Cate}_q$ and $\mathbf{Cate}_d$:

$$S_{cate}(q,d) = \frac{< \mathbf{Cate}_q, \mathbf{Cate}_d >}{\|\mathbf{Cate}_q\|_2 \cdot \|\mathbf{Cate}_d\|_2} \quad (3)$$

3.4 Measuring Question Similarity with Synonymies and Associative Concepts

To further improve the performance, synonyms and associative concepts in Wikipedia can be used to include more related concepts to overcome the data sparseness. For each concept c in Wikipedia, a set of related concepts $\mathbf{rela}_c = \{(c_1, w(c_1,c)), (c_2, w(c_2,c)), \ldots, (c_k, w(c_k,c))\}$ are selected from its synonyms and associative concepts, in which c_k is the kth related concepts of c and $w(c_k,c)$ is the relatedness between c_k and c. The relatedness is defined as follows:

$$w(c_k,c) = \begin{cases} 1 & \text{if } c_k \text{ and } c \text{ are synonyms;} \\ S_{overall} & \text{if } c_k \text{ and } c \text{ are associative relations} \end{cases}$$

where $S_{overall}$ is defined by equation (1). Let $\mathbf{C}_d$ denote the concept representation of historical question d, the corresponding synonym and associative vector can be represented as $\mathbf{SA}_d = \bigcup_{c \in \mathbf{C}_d} \mathbf{rela}_c$.

Figure 2 gives an example of the synonyms and associative concepts for Wikipedia concept "IBM". For concept "IBM", a set of related concepts $\mathbf{rela}_{\text{IBM}} = \{$("International Business Machines Corporation", 1.0), ("IBM PC Company", 1.0), ..., ("I.B.M.", 1.0), ("Computer", 0.27), ("Apple Inc.", 0.32), ..., ("Lenovo", 0.31) $\}$. For question Q_1 in Figure 1, the concept vector is $\mathbf{C}_{Q_1} = \{$"Software engineer", "IBM"$\}$, the corresponding synonym and associative vector can be represented as $\mathbf{SA}_{Q_1} = \mathbf{rela}_{\text{Software engineer}} \bigcup \mathbf{rela}_{\text{IBM}}$.

We expand the synonym and associative concepts to $\mathbf{C}_d$ and get the final concept representation $\mathbf{SAC}_d = \mathbf{SA}_d \bigcup \mathbf{C}_d$. Similarity, for queried question q, we also represent it in the final synonym and associative concept space as $\mathbf{SAC}_q = \mathbf{SA}_q \bigcup \mathbf{C}_q$. The similarity between the queried question and the historical question in the final synonym and associative concept space is, then, calculated as the cosine similarity between $\mathbf{SAC}_q$ and $\mathbf{SAC}_d$:

$$S_{sac}(q,d) = \frac{< \mathbf{SAC}_q, \mathbf{SAC}_d >}{\|\mathbf{SAC}_q\|_2 \cdot \|\mathbf{SAC}_d\|_2} \quad (4)$$

3.5 The Combination of Question Similarity

In the previous sections, we describe the methods to exploit Wikipedia category, synonym and associative relations for question similarity computation. In this section, we combine the above question similarity scores into term matching score $S_{term}(q,d)$ using a linear combination:

$$Score(q,d) = \alpha S_{cate}(q,d) + \beta S_{sac}(q,d) + \gamma S_{term}(q,d) \quad (5)$$

where $\alpha + \beta + \gamma = 1$, the relative importance of hierarchical relation (category) score, synonym and associative relation score, and the term matching score is adjusted through α, β and γ, respectively. That is to say, for a queried question q (or a historical question d), we represent the question through three different vectors: "Words", "Hypernyms" $\mathbf{Cate}_q$, "Synonyms" and "Associative Concepts" $\mathbf{SAC}_q$. An example of the feature vectors for question Q_1 in Figure 1 are shown in Table 1 [Cai *et al.*, 2011].

Words	require(0.019), software(0.031), engineer(0.027), big(0.018), blue(0.022)
Hypernyms	Software engineering(1.0), Software engineers(1.0), Computer hardware companies(1.0), Cloud computing(0.5) International business(0.5), computer companies(0.5) Multinational companies(1.0), Technology companies(0.25) ...
Synonyms	International Business Machines Corporation(1.0), International Business Machines(1.0), IBM computer(1.0), IBM Corporation(1.0), I.B.M.(1.0), ...
Associative Concepts	Apple Inc.(0.32), IBM Personal Computer(0.60), Corporation(0.36), Computer science(0.47), software architecture(0.72), ...

Table 1: An example of the different vector representation for question Q_1 in Figure 1.

4 Experiments

4.1 Wikipedia Data for Building the Thesaurus

Wikipedia data can be obtained easily from http://download.wikipedia.org for free research use. It is available in the form of database dumps that are released periodically. The version we used in our experiments was released on Sep. 9, 2007. We identified over 4 million distinct entities that constitute the vocabulary of thesaurus. The Wikipedia dump we use contains about 126,465 categories and 1,590,321 concepts after pre-precessing and filtering.

4.2 cQA Data for Question Retrieval

We collect the data set from Yahoo! Answers and use the *getByCategory* function provided in Yahoo! Answers API[5] 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 2 shows the distribution across first-level categories of the questions in the archives.

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

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; 2010]. This set contains 252 queried questions and can be freely downloaded for research communities.[6] For each method, the top 20 retrieval results are kept. Given a returned result for each queried question, an 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.

4.3 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.

4.4 Experimental Results

In this subsection, we present the experimental results. To demonstrate the effectiveness of the proposed method, we introduce the following methods for comparison:

- *BOWs*: This method measures the question similarity based on BOWs representation by using cosine similarity.
- *Category*: This method measures the question similarity with Hypernyms derived from Wikipedia by using equation (3).
- *SAC*: This method measures the question similarity with Synonymies and Associative Concepts derived from Wikipedia by using equation (4).
- *BOWs_Category*: This method measures the question similarity by using equation (5) except that we set $\beta = 0$.
- *BOWs_SAC*: This method measures the question similarity by using equation (5) except that we set $\alpha = 0$.

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

#	Methods	MAP	P@10
1	*BOWs*	0.242	0.226
2	*Category*	0.364	0.235
3	*SAC*	0.397	0.244
4	*BOWs_Category*	0.425	0.252
5	*BOWs_SAC*	0.448	0.266
6	***BOWs_Category_SAC***	**0.463**	**0.272**

Table 3: Experimental results for question retrieval.

- *BOWs_Category_SAC*: This method measures the question similarity by using equation (5).

The parameters λ_1, λ_2, α, β and γ used in the paper are set in the following way:

- For *SAC*, the parameters λ_1 and λ_2 in equation (1) are tuned according to the methodology suggested in [Wang and Domeniconl, 2008; Cai *et al.*, 2011]. As a result, the values $\lambda_1 = 0.4$ and $\lambda_2 = 0.5$ are used in our experiments.
- For *BOWs_Category*, the parameter β is set to 0. We tune the parameters α and γ 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. We tune the value of α and γ from 0.1, 0.2 up to 1.0, and thus we can find the proper values of α and γ on the development set. Finally, α is set to 0.3 and γ is set to 0.7.
- For *BOWs_SAC*, the parameter α is set to 0. We tune the value of β and γ from 0.1, 0.2 up to 1.0, and thus we can find the proper values of β and γ on the development set. Finally, β is set to 0.4 and γ is set to 0.6.
- For *BOWs_Category_SAC*, we tune the value of α, β and γ from 0.1, 0.2 up to 1.0, and thus we can find the proper values of α, β and γ on the development set. Finally, α is set to 0.2, β is set to 0.3 and γ is set to 0.5.

Table 3 presents the experimental results for question retrieval. From Table 3, we have several observations:

- Comparing our proposed *Category* and *SAC* experimental results, we find that hierarchical relation, synonym and associative relation can significantly improve the performance of question retrieval (row 1 vs. row 2 and row 3, the comparisons are statistically significant at $p < 0.05$).
- synonym and associative relation plays a more importance role than hierarchical relations (*Category* only gets 12.2% improvement of MAP over *BOWs*, while *SAC* gets 15.5% improvement of MAP over *BOWs*).
- Combining the different methods can further improve the performance of question retrieval (row 2 vs. row 4 and row 6; row 3 vs. row 5 and row 6).

4.5 Comparison with the State-of-the-art

This paper aims to tackle the limitation of BOWs for question retrieval. Many researchers have proposed the use of translation models [Berger *et al.*, 2000; Jeon *et al.*, 2005;

Riezler *et al.*, 2007; Xue *et al.*, 2008; Lee *et al.*, 2008; Bernhard and Gurevych, 2009; Zhou *et al.*, 2011] to capture the semantic word relations and solve the limitation of BOWs by using the translated words. Jeon et al. [2005] (*Jeon2005*) 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. Xue et al. [2008] (*Xue2008*) 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] (*Zhou2011*) proposed a monolingual phrase-based translation model for question retrieval. To implement the word-based translation models, we use the GIZA++ alignment toolkit[7] trained on one million question-answer pairs from another data set[8] to learn the word-to-word translation probabilities. For phrase-based translation model described in [Zhou *et al.*, 2011], we employ Moses toolkit[9] to extract the phrase translation and set the maximum length of phrases to 5. Recently, Singh et al. [2012] (*Singh2012*) 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. Zhou et al. [2012] (*Zhou2012*) employed the bilingual translation and expanded queried questions with translated words for question retrieval. Furthermore, Zhou et al. [2013] (*Zhou2013*) borrowed the statistical machine translation to expand the question representation via a matrix factorization framework. Besides, Cao et al. [2009] (*Cao2009*) and Cao et al. [2010] (*Cao2010*) also proposed to utilize category information 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 also 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 4 shows the comparison with the state-of-the-art for question retrieval. From this table, we can see that our proposed method is better than previous work. The results indicate that using the world knowledge of Wikipedia for question retrieval is more helpful than the translation models. In addition, we also find that some other methods (Zhou2012, Zhou2013, Cao2010) obtain better performance than ours (***BOWs_Category_SAC***) by using the external information beyond the texts. However, these external information (e.g., bilingual translation or category information) is orthogonal to ours, and we suspect that combining the bilingual translation or category information into our proposed method might get even better performance. We leave it for future research.

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

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

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

#	Methods	MAP	P@10
1	*BOWs*	0.242	0.226
2	*Jeon2005*	0.405	0.247
3	*Xue2008*	0.436	0.261
4	*Zhou2011*	0.452	0.268
5	*Singh2012*	0.450	0.267
6	*Zhou2012*	0.483	0.275
7	*Zhou2013*	0.564	0.291
8	*Cao2009*	0.408	0.247
9	*Cao2010*	0.456	0.269
10	***BOWs_Category_SAC***	**0.463**	**0.272**

Table 4: Comparison with the state-of-the-art for question retrieval.

5 Conclusions and Future Work

In this paper, we first propose a way to build a *concept thesaurus* based on the semantic relations extracted from world knowledge of Wikipedia. Then, we develop a unified framework to leverage these semantic relations in order to enhance the similarity measure for question retrieval in the concept space. Experiments conducted on a real cQA data set show that with the help of Wikipedia thesaurus, the performance of question retrieval is improved as compared to the traditional methods.

In the future, we would like to combine the bilingual translation (e.g., [Zhou *et al.*, 2012; 2013]) or category information (e.g., [Cao *et al.*, 2010]) into our proposed method for question retrieval. Besides, 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

[Berger *et al.*, 2000] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. Bridging the lexical chasm: statistical approach to answer-finding. In *Annual International ACM SIGIR Conference (SIGIR)*, pages 192–199, 2000.

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

[Cai *et al.*, 2011] L. Cai, G. Zhou, K. Liu, and J. Zhao. Large-scale question classification in cqa by leveraging

wikipedia semantic knowledge. In *Conference on Information and Knowledge Management (CIKM)*, pages 1321–1330, 2011.

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

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

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

[Gebrilovich and Markovitch, 2006] E. Gebrilovich and S. Markovitch. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorication with encyclopedia knowledge. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1301–1306, 2006.

[Gupta and Gupta, 2012] P. Gupta and V. Gupta. A survey of text question answering techniques. *International Journal of Computer Applications*, 53(4):1–8, 2012.

[Hu *et al.*, 2008] J. Hu, L. Fang, Y. Cao, H. Zeng, H. Li, Q. Yang, and Z. Chen. Enhancing text clustering by leveraging wikipedia semantics. In *Annual International ACM SIGIR Conference (SIGIR)*, 2008.

[Hu *et al.*, 2009a] X. Hu, N. Sun, C. Zhang, and T.-S. Chua. Exploting internal and external semantics for the clustering of short texts using world knowledge. In *Conference on Information and Knowledge Management (CIKM)*, 2009.

[Hu *et al.*, 2009b] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou. Exploiting wikipedia as external knowledge for document clustering. In *KDD*, 2009.

[Jeon *et al.*, 2005] J. Jeon, W. Croft, and J. Lee. Finding similar questions in large question and answer archives. In *Conference on Information and Knowledge Management (CIKM)*, pages 84–90, 2005.

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

[Maybury, 2004] M. Maybury. New directions in question answering. *AAAI/MIT Press*, 2004.

[Milne *et al.*, 2006] D. Milne, Q. Medelyan, and I. H. Witten. Mining domain-specific thesauri from wikipedia: a case study. In *WI*, 2006.

[Ponzetto and Strube, 2007] S. P. Ponzetto and M. Strube. Deriving a large scale taxonomy from wikipedia. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2007.

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

[Robertson *et al.*, 1994] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *TREC*, pages 109–126, 1994.

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

[Wang and Domeniconl, 2008] P. Wang and C. Domeniconl. Building semantic kernels for text classification using wikipedia. In *KDD*, 2008.

[Wang *et al.*, 2007] P. Wang, J. Hu, H.-J. Zeng, L. Chen, and Z. Chen. Improving text classification by using encyclopedia knowledge. In *International Conference on Data Mining (ICDM)*, 2007.

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

[Wang *et al.*, 2010] B. Wang, X. Wang, C. Sun, B. Liu, and L. Sun. Modeling semantic relevance for question-answer pairs in web social communities. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1230–1238, 2010.

[Wong and Chan, 1996] P. Wong and C. Chan. Chinese word segmentation based on maximum matching and word binding force. In *International Conference on Computational Linguistics (COLING)*, 1996.

[Xue *et al.*, 2008] X. Xue, J. Jeon, and W. B. Croft. Retrieval models for question and answer archives. In *31st Annual International ACM SIGIR Conference (SIGIR)*, pages 475–482, 2008.

[Zhou *et al.*, 2011] G. Zhou, L. Cai, J. Zhao, and K. Liu. Phrase-based translation model for question retrieval in community question answer archives. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 653–662, 2011.

[Zhou *et al.*, 2012] G. Zhou, K. Liu, and J. Zhao. Exploiting bilingual translation for question retrieval in community-based question answering. In *24th International Conference on Computational Linguistics (COLING)*, pages 3153–3170, 2012.

[Zhou *et al.*, 2013] G. Zhou, F. Liu, Y. Liu, S. He, and J. Zhao. Statistical machine translation improves question retrieval in community question answering via matrix factorization. In *51th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2013.