XSearch: A Domain-Specific Cross-Language Relevant Question Retrieval Tool

Bowen Xu
Zhejiang University
China
max_xbw@zju.edu.cn

Zhenchang Xing
Australian National University
Australia
Zhenchang.Xing@anu.edu.au

Xin Xia
University of British Columbia
Canada
xxia02@cs.ubc.ca

David Lo
Singapore Management University
Singapore
davidlo@smu.edu.sg

Xuan-Bach D. Le
Singapore Management University
Singapore
dxb.le.2013@smu.edu.sg

ABSTRACT

During software development process, Chinese developers often seek solutions to the technical problems they encounter by searching relevant questions on Q&A sites. When developers fail to find solutions on Q&A sites in Chinese, they could translate their query and search on the English Q&A sites. However, Chinese developers who are non-native English speakers often are not comfortable to ask or search questions in English, as they do not know the proper translation of the Chinese technical words into the English technical words. Furthermore, the process of manually formulating cross-language queries and determining the importance of query words is a tedious and time-consuming process. For the purpose of helping Chinese developers take advantages of the rich knowledge base of the English version of Stack Overflow and simplify the retrieval process, we propose an automated cross-language relevant question retrieval tool (*XSearch*) to retrieve relevant English questions on Stack Overflow for a given Chinese question. This tool can address the increasing need for developer to solve technical problems by retrieving cross-language relevant Q&A resources.
Demo Tool Website: http://172.93.36.10:8080/XSearch.
Demo Video: https://goo.gl/h57sed.

CCS CONCEPTS

• **Software and its engineering → Software libraries and repositories**; • **Information systems → Social networking sites**;

KEYWORDS

Domain-Specific Translation, Cross-Language Question Retrieval

ACM Reference format:
Bowen Xu, Zhenchang Xing, Xin Xia, David Lo, and Xuan-Bach D. Le. 2017. XSearch: A Domain-Specific Cross-Language Relevant Question Retrieval Tool. In *Proceedings of 2017 11th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering, Paderborn, Germany, September 4–8, 2017 (ESEC/FSE'17),* 5 pages.
https://doi.org/10.1145/3106237.3122820

ESEC/FSE'17, September 4–8, 2017, Paderborn, Germany

ACM ISBN 978-1-4503-5105-8/17/09…$15.00
https://doi.org/10.1145/3106237.3122820

1 INTRODUCTION

Ten percent of the world's programmers are in China[1]. Even without a localized version of Stack Overflow in Chinese, it would still be very desirable to support developers in China to easily access the knowledge repository of the English version of Stack Overflow. Developers in China usually graduate with a Bachelor degree. To fulfill the degree requirements, they need to pass the national college English test (Level 4). As such, developers in China are often equipped with basic English reading comprehension skills, and they could be fluent enough to read posts in English. However, most of them often are not comfortable asking questions in English [2, 5]. Furthermore, they often cannot accurately translate Chinese technical words into proper English technical words, as the general translation tools usually do not understand the domain-specific meaning of those technical words. This makes it difficult for them to formulate English queries to search the Internet.

This reality of English reading and writing skills of developers in China indicates a potential to make the content of the English version of Stack Overflow more easily accessible to developers in China. In our previous work [12], we propose a domain-specific cross-language relevant question retrieval approach that takes as input a question written in Chinese and returns relevant questions written in English from Stack Overflow. These relevant English questions are the keys to accessing the knowledge in the English version of Stack Overflow. In this work, we strengthen the approach by proposing a new word embedding based retrieval algorithm and implement it as a publicly accessible web site. We also conduct an user study to evaluate our tool and the result shows that compared with our previous work, the performance is significantly improved.

Using our tool, developers can write query in Chinese (may be mixed with English words such as programming languages, tools, parameters). Given a Chinese query, our tool *XSearch* retrieves relevant English questions from Stack Overflow. Our tool improves the efficiency of Chinese developers to find solutions in a repository of English questions by solving the problem of query understanding, domain-specific translation of technical words and retrieval

[1] https://goo.gl/U8uFTO

algorithm. A key benefit of our tool is that it allows Chinese developers to more easily take advantage of high-quality English Q&A resources on Stack Overflow.

2 DESIGN CHALLENGES

Cross-language relevant question retrieval is a very complex process. To achieve the above objective of the cross-language question retrieval, we must address the following three challenges.

Challenges in keyword extraction: A question may contain many words, and we would like to extract the essential information for query formulation. To that end, we should use keyword extraction algorithms to summarize the essential information in the question. Many keyword extraction algorithms have been proposed in the natural language processing field. Different algorithms are based on different heuristics to evaluate the importance of a word. As they are heuristic-based, some keyword extraction algorithms may perform better than others on some cases, but worse on other cases. One way to address the weaknesses of these keyword extraction algorithms is to combine them together in order to make a comprehensive judgment. In this paper, we use two different keyword extraction algorithms (FudanNLP [2] and IctclasNLP [3]) to extract Chinese keywords in the title and description of the Chinese question, and take the union of the two sets of keywords as the final Chinese keywords.

Challenges in domain-specific translation: Cross-language question retrieval has to translate the words in the source language into some appropriate words in the target language. The accuracy of this translation will directly affect the relevance of the questions retrieved in the target language. Kluck and Gey [6] point out that in many cases there exists a clear difference between the domain-specific meaning and the common meaning of a word. This means that it can be difficult to use the common translation result of a word for domain-specific information retrieval. For example, for the Chinese word "代码", the general domain translation tool returns several English translations, such as "code" and "word". In the context of software engineering, the translation "code" is more appropriate than the other translations. Several studies propose domain-specific translation techniques which are based on building domain-specific dictionary [4, 9, 11]. A few research studies have been carried out on domain-specific translation, which are based on domain-specific translation lexicon [4, 9, 11]. However, developing a domain-specific dictionary requires a significant effort. In this paper, we propose a new approach to support domain-specific translation.

Challenges in question retrieval algorithm: When searching on Stack Overflow, we find that the question retrieval algorithm is not very robust. For example, when we search the question *"Joda Time sometimes returns wrong time"*, we can retrieve a question on Stack Overflow successfully. However, if we change the word "returns" to "return", the search for *"Joda Time sometimes return wrong time"* returns no matches. It seems that Stack Overflow question retrieval algorithm does not take word stemming into consideration. Furthermore, we observe that keywords extracted from different parts of the question (such as title versus description) often have different levels of importance for question retrieval. However, existing question retrieval algorithms do not take this into account. In this paper, we design a question retrieval algorithm to address these limitations by considering word stemming and assigning different weights to the words from title and description. Furthermore, we propose three different similarity metrics to determine the relevance between query and questions.

[2]FudanNLP, available at http://nlp.fudan.edu.cn
[3]IctclasNLP, available at http://ictclas.nlpir.org/docs

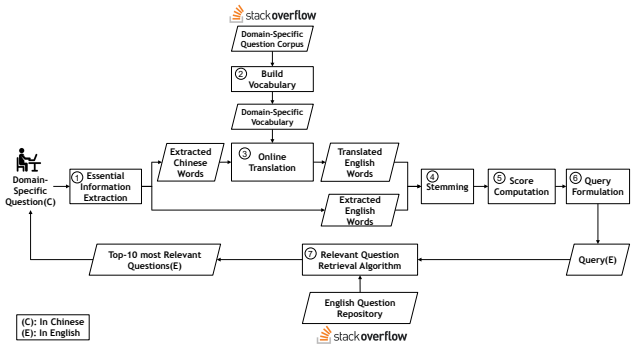

Figure 1: Overall Framework

3 TOOL DESIGN

3.1 Overall Framework

Figure 1 presents the overall framework of our domain-specific cross-language relevant question retrieval. Given a software engineering related question in Chinese, essential information extraction (step 1) extracts Chinese words and English words from the given Chinese question. Given the extracted Chinese words, domain-specific cross-language translation (step 3) translates the Chinese words into domain-specific English words, based on a domain-specific vocabulary derived from a corpus of Stack Overflow questions (step 2). Given a list of candidate English words (extracted or translated) from the input Chinese question, the system formulates an English query as follows: first, it stems the English words (step 4), then it assigns different weights to the different types of English words (step 5) depending on whether the words are from the question title or description, and finally it takes a subset of English words with the highest scores to formulate the English query (step 6). The system uses the English query to search a repository of Stack Overflow questions (step 7), and the retrieval algorithms using different similarity metrics return the top-10 most relevant English questions to the user.

3.2 Essential Information Extraction

Given a question in Chinese, we extract the essential information based on the following two observations:

(1) Question title sums up the core issue of the question better than the question description.
(2) Most technical words are written in English. Developers always ask some technical questions with some domain-specific words in English. Those English words are important for retrieving relevant English questions.

Thus, we divide the question essential information into four kinds: (i) Chinese words in Title (*CT*), (ii) English words in Title(*ET*), (iii) Chinese Keywords in Description(*CKD*), (iv) English words in Description(*ED*), and they are given different weights.

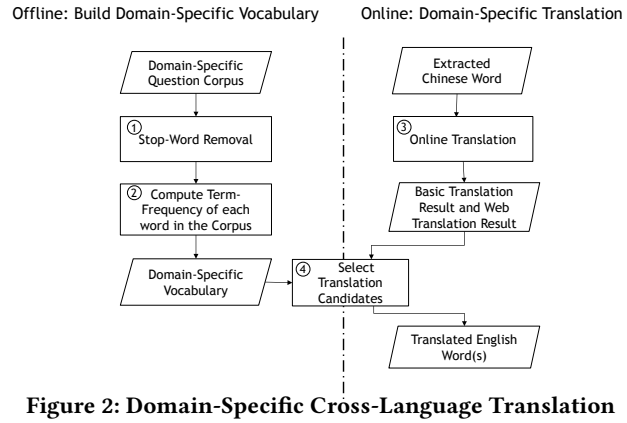

Figure 2: Domain-Specific Cross-Language Translation

The input is the query, the algorithm uses two different popular Chinese keywords extraction algorithms (FudanNLP and IctclasNLP) to extract Chinese keywords. The algorithm of keywords extraction of FudanNLP is based on TextRank algorithm. The algorithm of keywords extraction of IctclasNLP is based on entropy. The two algorithms produce two different but complementary sets of Chinese keywords. To reduce the bias caused by a single method, the algorithm takes the union of the two keyword sets as the final set of Chinese keywords to summarize the Chinese question query.

3.3 Domain-specific Cross-language Translation

Translation is a critical step in cross-language information retrieval [1, 7]. Recent results show that the challenge lies in how to differentiate a word's domain-specific meaning from its common meaning. General translation tools like Youdai Translation, Google Translate, may not perform well for domain-specific translation, because it does not consider any domain knowledge. We proposed a method to address the limitation. Figure 2 presents the details of our method. We divide the domain-specific cross-language translation into an offline vocabulary building step and an online translation step. To build a domain-specific vocabulary, we make use of crowdsourced knowledge in Stack Overflow discussions. We collect a corpus of randomly-selected 30,000 Stack Overflow questions tagged with *java*. The corpus contains a total of 111,174 English words. We first remove stop words (step 1), such as "hello", "the" and "you". The stop-word list we use is available at Snowball[4]. Then, we compute term frequency for each word in the corpus and build a vocabulary of each word and their term frequency in the corpus (step 2).

For online translation, given a Chinese word, the system first uses a general Chinese-English translation tool to obtain a list of candidate English words which includes both basic translation result and several web translation results (step 3). Then, the system checks the term frequency of these English word candidates in the domain-specific vocabulary (step 4). Finally, the system selects the English words with the term frequency above the mean term frequency of all the candidate English words as the translation of the given Chinese word. If none of the candidate English words exist in the domain-specific vocabulary, the system returns the basic translation (i.e., the most common translation) as the translation result.

[4]Stop-word list, available at http://snowball.tartarus.org/algorithms/english/stop.txt

3.4 Relevant Question Retrieval Algorithm

Our cross-language question retrieval contains three steps: stemming, word score computation and query formulation, and question retrieval.

3.4.1 Stemming. In the stemming step, we reduce each word to its root form, for example, words "write" and "written" are both reduced to "writ". We use a popular stemming algorithm (the Porter stemmer [10]) in this work.

3.4.2 Word Score Computation. We set the weights of different kinds of words based on the observations mentioned in Section 3.2. We assign CT:ET:CKD:ED a weight of 2:2:1:1, respectively. For each kind of words, we update its *Score* by $(termfrequency \times Weight)/WordsetSize$. If a word belongs to different kinds of word set at the same time, the score of the word will be accumulated. After computing the score of all the words, we use all words with the scores to generate an English query.

3.4.3 Relevant Question Retrieval. Question retrieval algorithms calculate the relevance between the query and each English question in the repository, and recommend the top-10 most relevant English questions to users that may help them solve the problem. Different with our previous work [12], we propose a new retrieval algorithm based on word embeddings. Word Embeddings exploit distributed word representation. Distributed word representations assume that words appear in similar context tend to have similar meanings [3]. Therefore, individual words are no longer treated as unique symbols, but mapped to a dense real-valued low-dimensional vector space. Each dimension represents a latent semantic or syntactic feature of the word. Semantically similar words are close in the embedding space. Thus, given a word, it will be converted into a high dimensional vector with real value by looking up the dictionary of word embeddings. Semantically close words, such as *JPanel*, *JButton*, *JFrame* and *JLabel* which are GUI components are close in the vector space.

Given a query word vector v_{Qw} and a word vector v_{qw} in English question, we define their semantic similarity as the cosine similarity between their learned word embeddings, i.e.,$Rel(v_{Qw}, v_{qw}) = CosineSim(v_{Qw}, v_{qw})$. It is simply the inner product of the two vectors, normalized by their Euclidean norm. To compute the relevance between the query and the question text, we use the text-to-text similarity measure introduced by Mihalcea et al. [8]. According to [8], the relevance between a query word Qw and a English question Q is computed as the maximum similarity between Qw and any word $w^{'}$ in Q, i.e., $Rel(Qw, Q) = \max_{w^{'} \epsilon Q} Rel(Qw, w^{'}) \times Qw.score$.

Then, we define the relevance between a given query *Query* and an English question Q in repository as

$$Rel(Query, Q) = \sum_{Qw \epsilon Query} Rel(Qw, Q)$$

4 TOOL IMPLEMENTATION

We have implemented our *XSearch* tool based on Browser/Server structure. *XSearch* is organized into two separate components: Data Collection and Data Searching.

Data Collection:

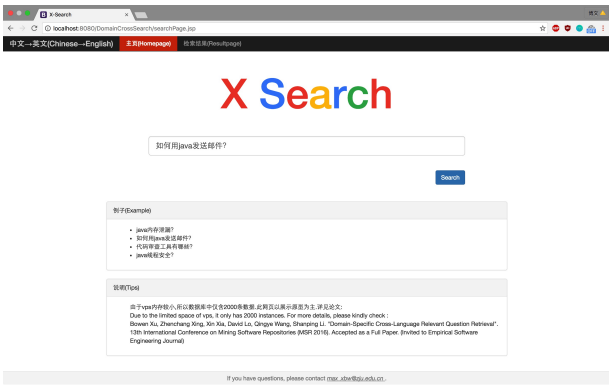

Figure 3: Homepage of XSearch

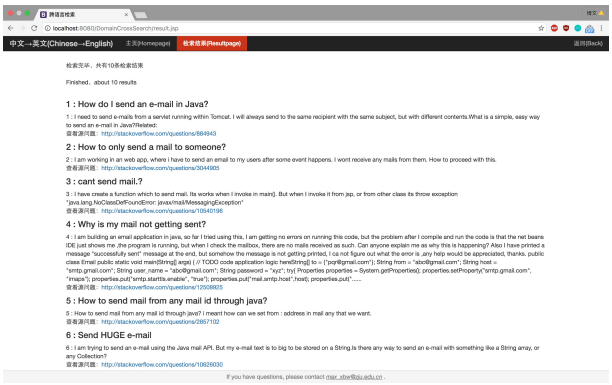

Figure 4: Result of XSearch

Step 1: English Question Databases. We extract 714,599 Java questions from Stack Exchange Data Dump[5] released by Stack Exchange, Inc. We use 30,000 Java questions to build domain-specific vocabulary and another 684,599 as the repository of English questions for question retrieval.

Step 2: Word Embedding Corpora. For building the word embedding corpora, we randomly select 300,000 English questions tagged with "java" as the word embedding corpora. It contains totally 329,845 different english words.

Data Searching:

Figure 3 shows the homepage of *XSearch*. When a developer sends a query to the server for relevant question search, the query is processed. Essential information is extracted and translated. Next, every extracted word is converted into a vector by word embeddings. Then, retrieval algorithm returns top-10 most relevant questions in repository and send them back to the developer as shown in Figure 4.

5 USER STUDY

We crawl 200 Java questions from SegmentFault and V2EX (two Chinese Q&A websites for computer programming) as query Chinese question set and randomly choose 80 questions for the evaluation.

User Study. We conduct an user study to evaluate the top-10 most relevant questions generated by our approach. The evaluator group included 5 master students, all of whom have industrial experience in Java programming (ranging from 3-6 years) and pass the national college English test (Level 4). We provide these five users 80 Chinese Java questions from SegmentFault and V2EX. For each Chinese question, we provide a questionnaire of the top-10 most relevant English questions generated by our approach. The user study evaluation has two steps. First, we ask the participants to read the same question at the same time and independently evaluate whether each retrieved English question is relevant or not to the given Chinese question. If more than half of the participants refer an English question is relevant with the given Chinese question, then it will be record as an actually relevant question.

Results. The evaluation result shows that *XSearch* achieves precision@1, precision@5, precision@10, top-1 accuracy, top-5 accuracy, top-10 accuracy, MRR and MAP of 0.83, 0.79, 0.75, 0.83, 0.91, 0.93, 0.87 and 0.42 which outperforms our previous approach [12] by 48%, 88%, 134%, 48%, 32%, 31%, 40%, 61%, respectively. We apply Wilcoxon signed-rank test to compare our approach with our previous approach [12], the result shows that the improvement is significant at the confidence level of 95% score.

[5]Stack Exchange Data Dump, available at https://archive.org/download/stackexchange

6 CONCLUSION AND FUTURE WORK

We propose a novel tool to retrieve relevant English questions for a given Chinese query. We mine domain-specific knowledge from Stack Overflow to improve the accuracy of translation of domain-specific Chinese words. Considering the difference between different types of words, we assign query words with different weights in query formulation and query retrieval. To overcome the lexical gap issue, we propose to adopt neural language model (word embeddings). All the steps of our approach are automated, and thus it can help developers save time in terms of query translation, query formulation, and question retrieval. As a result, it can potentially help Chinese developers improve their efficiency to solve the technical problems.

In the further work, we will crawl more questions on Stack Overflow or some other Q&A sites. This will help improve the relevance and usefulness of retrieved questions. Furthermore, we observe that most of the questions in Stack Overflow contain code. In the current approach, we treat code as natural language text. In the future, we can extract code segments of questions and process them in a different way from regular English texts. Another improvement is to expand our domain-specific stop words list and vocabulary to improve the accuracy of domain-specific translation. Moreover, we plan to prove the generality of our proposed framework which indicates that our tool can be used for other natural languages, not limited to Chinese and English. In that way, our approach can be exploited to link the Q&A resources in different localized versions of Stack Overflow or Q&A sites for computer programming in different languages.

ACKNOWLEDGMENTS

Xin Xia is the corresponding author. This work was partially supported by NSFC Program (No. 61602403 and 61572426), and National Key Technology R&D Program of the Ministry of Science and Technology of China (No. 2015BAH17F01).

REFERENCES

[1] Rita Marina Aceves-Pérez, Manuel Montes-y Gómez, and Luis Villaseñor-Pineda. 2007. Enhancing cross-language question answering by combining multiple question translations. In *Computational Linguistics and Intelligent Text Processing*. Springer, 485–493.

[2] Harkness. 2017. Why are some Chinese students who have learnt English for years still poor in English? https://goo.gl/7ltMLy. (2017).

[3] Zellig S Harris. 1954. Distributional structure. *Word* 10, 2-3 (1954), 146–162.

[4] Djoerd Hiemstra, Franciska De Jong, and Wessel Kraaij. 1997. A domain specific lexicon acquisition tool for cross-language information retrieval. In *Computer-Assisted Information Searching on Internet*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 255–268.

[5] Shang-Ling Jui. 2010. *Innovation in China: the Chinese software industry*. Routledge. 142–144 pages.

[6] Michael Kluck and Fredric C Gey. 2001. The domain-specific task of CLEF-specific evaluation strategies in cross-language information retrieval. In *Cross-Language Information Retrieval and Evaluation*. Springer, 48–56.

[7] Wessel Kraaij, Jian-Yun Nie, and Michel Simard. 2003. Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics* 29, 3 (2003), 381–419.

[8] Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, Usa*. 775–780.

[9] Anselmo Peñas, Bernardo Magnini, Pamela Forner, Richard Sutcliffe, Álvaro Rodrigo, and Danilo Giampiccolo. 2012. Question answering at the cross-language evaluation forum 2003–2010. *Language resources and evaluation* 46, 2 (2012), 177–217.

[10] Martin F Porter. 1980. An algorithm for suffix stripping. *Program* 14, 3 (1980), 130–137.

[11] Philip Resnik and I Dan Melamed. 1997. Semi-automatic acquisition of domain-specific translation lexicons. In *Proceedings of the fifth conference on Applied natural language processing*. Association for Computational Linguistics, 340–347.

[12] Bowen Xu, Zhenchang Xing, Xin Xia, David Lo, Qingye Wang, and Shanping Li. 2016. Domain-specific cross-language relevant question retrieval. In *Proceedings of the 13th International Workshop on Mining Software Repositories*. ACM, 413–424.