Survey on Question Retrieval in Community Question Answering via NON-Negative Matrix Factorization

Yogita Puri
M.S.Bidve Engineering College,Latur

Pankaj Kokane
M.S.Bidve Engineering College,Latur

ABSTRACT: CQA is useful in answering real-world questions. CQA provide a solution to human. Question retrieval in CQA can automatically find the foremost relevant and up to date questions that are solved by other users. We propose an alternative thanks to addressing the word ambiguity and word mismatch problems by taking advantage of probably rich semantic information drawn from other languages. The translated words from other languages via non-negative matrix factorization. Contextual information is exploited during the interpretation from one language to a different language by using Google Translate. Thus, word ambiguity is often solved supported the contextual information when questions are translated. Multiple words that have similar meanings in one language are also translated into a singular word or some words in a very foreign language. it's a word-based translation language model for retrieval with a question likelihood model for an answer. We use a translated representation by alternative enriching the first question with the words from other languages in CQA. We translate English questions into other four languages using Google translate which takes into account contextual information during translation. If we translate the question word by word, it discards the contextual information. We'd expect that such a translation would not be able to solve the word ambiguity problem.

Keywords

Community Question Answering, Statically Machine Translation, Non Matrix Factorization, Google Translator, Recursive Neural Network.

1. INTRODUCTION

To make community question answering portals more useful, it is necessary for the system to be able to fetch the questions asked in other languages moreover. this may give the user a wide range of pre answered inquiries to rummage around for solutions to his/her problem. Current systems fail to try to so. Also, these systems fetch related questions supported the keywords in it. Thus, if there's an issue which is said to the subject but having other keywords, then that question isn't retrieved, this is a serious drawback of a system as there is many circumstances where a semantically related question but not having similar keywords isn't retrieved. The proposed the system shows the way to retrieve questions which are associated with the asked question but asked in other languages moreover because the questions that are associated with the subject but not having similar keywords. The proposed system shows that this will be achieved when these questions are retrieved semantically instead of using keywords. It is found that, in most cases, an automatic approach cannot obtain results that are nearly as good as those generated by human intelligence. together with the proliferation and improvement of underlying communication technologies, community Question Answering (CQA) has emerged as an extremely popular alternative to accumulate information online, owning to the subsequent facts. a. Information seekers are able to post their specific questions on any topic and acquire answers provided by other participants. By leveraging community efforts, they're able to bounce back answers than simply using search engines. as compared with automated CQA systems, CQA usually receives answers with better quality as they're generated supported human intelligence. c. Over times, an amazing number of QA pairs are accumulated in their repositories, and it facilitates the preservation and search of answered questions.

Related Work

Learning the Multilingual Translation Representations for Question Retrieval in Community Question Answering via Non-negative Matrix Factorization In this paper they propose to employ statistical machine translation to improve question retrieval and enrich the question representation with the translated words from other languages via matrix factorization. They also Proposes a way of fetching previously asked questions which are asked in different different languages but are related to the asked question after the development of web 2.0, www became very interactive and lot of new kinds of applications emerged based on web 2.0.

1.1 Word Based Translation Language Model

In this paper they proposed a word-based translation language model for question retrieval with a query likelihood model for the answer. Experiments consistently reported that the word-based translation model could yield better performance than the traditional methods (e.g., VSM, BM25 and LM). However, these word-based translation models are considered to be context independent in that they don't take into account any contextual information in modeling word translation probabilities. In order to further improve the word-based translation model with some contextual information.

1.2 Finding Similar Questions in Large Question and Answer Archives

In this paper they show that a question and answer archive from a community-based Q&A service can serve as a valuable resource to train retrieval models that can recognize semantically similar questions. Specifically, they showed that a retrieval model based on translation probabilities learned from the archive significantly outperforms other approaches in terms of finding semantically similar questions despite a considerable amount of lexical mismatch. Because of the computational cost, we initially used a relatively small subset of the available archive. As they increase the number of the training samples, they expect to get more accurate word translation probabilities and better retrieval performance.

1.3 Question-answer topic model for question retrieval in community question answering

In this paper they proposed a question-answer topic model to learn the latent topics aligned across the question answer pairs to alleviate the lexical gap problem, with the assumption that a question and its paired answer share the same topic distribution.

1.4 Tapping on the potential of q&a community by recommending answer providers

In this paper they proposed a supervised question-answer topic modeling approach, which assumes that questions and answers share some common latent topics and are generated in a question language and answer language. Besides, other researchers also applied the topic models for the related tasks in cQA.

1.5 Lexical Semantic Resources

In this paper they proposed to fetch the data from system to use as a parallel training dataset the definitions and glosses provided for the same term by different lexical semantic resource. Not work such as question paraphrase retrieval, and larger datasets. Not improve question analysis by automatically identifying question topic and question focus

1.6 Semantic Relevance Modeling Chinese QA Paris

In this paper they proposed two deep belief networks with different architectures have been presented based on the QA joint distribution and the answer-to-question reconstruction principles respectively. Both the models show good performance on modeling the semantic relevance for the QA pairs, using only word occurrence features. Taking the data driven strategy, our DBN models learn semantic knowledge from large amount of QA pairs to quantify the semantic relevance between questions and their answers. (2) We have investigated the textual similarity between the CQA and the forum datasets for QA pair extraction, which provides the basis to our approaches to avoid hand-annotating work and show good performance on both the CQA and the forum corpora

1.7 Entity Based Q&A Retrieval

To highly dependent the availability of quality corpus in the absence of which they are troubled by noise. Semantic concepts for addressing the lexical gap issue in retrieval models for large online Q&A collections.

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

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

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

2. OUR APPROCH

2.1 Problem Statement

This paper aims to leverage statistical machine translation to complement the question representation. In order to handle the word ambiguity and word mismatch problems, we expand a matter by adding its translation counterparts. Statistical AI (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 = {l1, l2, . . . , lP } denote the language set,

where P is the number of languages considered in the paper, l1 denotes the original language (e.g., English) while l2 to lP are the foreign languages.

Let D1 = {d (1) 1 , d(1) 2 , . . . , d(1) N } be the set of historical question collection in original language, where N is the number of historical questions in D1 with vocabulary size M1.

Now we first translate each original historical question from language l1 into other languages lp (p ∈ [2, P]) by Google Translate. Thus, we can obtain D2, . . . , DP in different languages, and Mp is the vocabulary size of Dp.

A question d (p) i in Dp is simply represented as a Mp dimensional vector d (p) i , in which each entry is calculated by tf-idf. The N historical questions in Dp are then represented in a Mp × N term-question matrix

Dp = {d (p) 1 , d (p) 2 , . . . , d (p) N }, 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 l2 to lP , the original vocabulary size is increased from M1 to ∑P p=1 Mp. Thus, the term-question matrix becomes

D = {D1, D2, . . . , DP } and D ∈ R (∑P p=1 Mp)×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 qi represents a queried question, and qi is a vector representation of qi.

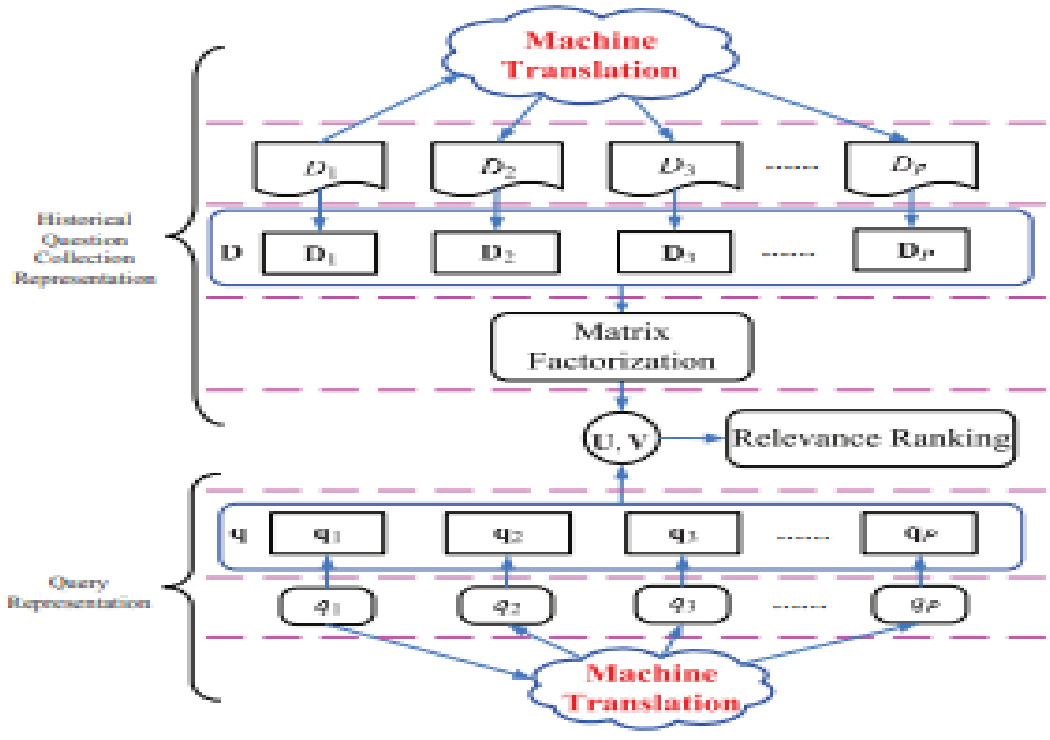

Fig 1: Framework of our proposed approach for question retrieval.

3. Our Proposed Algorithm

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** ▷ 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

To tackle the info scantiness of question illustration with the translated words, we have a tendency to hope to search out 2 or additional lower dimensional matrices whose product provides a decent approximate to the first one via matrix resolving. Previous studies have shown that there's psychological and physiological evidence for parts-based illustration within the human brain. The non-negative matrix resolving (NMF) is proposed to find out the elements of objects like text documents. NMF aims to search out 2 nonnegative matrices whose product provides a decent approximation to the first matrix and has been shown to be superior to SVD in document clump.

4. Conclusion

As we all know the CQA system is getting incredible popularity over the years. But since the existence of the CQA system it is just giving the information to a question, posed by user, in the form of textual contents. We are work on system with use of translated representation we will propose in this paper. In this, we work on the original questions are enhanced with semantically similar word from other languages. This can help in retrieving questions which are related to the questions which are from other languages.

5. REFERENCES

[1] Learning the Multilingual Translation Representations for Question Retrieval in Community Question Answering via Non-negative Matrix Factorization Guangyou Zhou, Zhiwen Xie, Tingting He, Jun Zhao and Xiaohua Tony Hu.2017.

[2] Adamic, J. Zhang, E. Bakshy, and M. Ackerman. 2008. Knowledge sharing and yahoo answers: everyone knows and something.

[3] X. Xue, J. Jeon, and W. B. Croft, “Retrieval models for question and answer archives,” in SIGIR, 2008, pp. 475–482.

[4] Z. Ji, F. Xu, B. Wang, and B. He, “Question-answer topic model for question retrieval in community question answering,” in CIKM, 2012, pp. 2471–2474.

[5] J. Guo, S. Xu, S. Bao, and Y. Yu, “Tapping on the potential of q&a community by recommending answer providers,” in CIKM, 2008, pp. 921–930.

[6] G. Zhou, J. Zhao, T. He, and W. Wu, “An empirical study of topicsensitive probabilistic model for expert finding in question answer communities,” Knowledge-Based Systems, pp. 136–145, 2014.

[7] J. Jeon, W. B. Croft, and J. H. Lee, “Finding similar questions in large question and answer archives”, in CIKM, 2005, pp. 8490.

[8] 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 ACL, 2013, pp. 852-861. 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 ACL, 2013, pp. 852861. Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[9] X. Cao, G. Cong, B. Cui, and C. S. Jensen, “A generalized framework of exploring category information for question retrieval in community question answer archives,” in CIKM, 2010, pp. 201–210..

[10] Statistical Machine Translation for Query Expansion in Answer Retrieval, ser. ACL, 2007.

[11] G. Zhou, L. Cai, J. Zhao, and K. Liu, “Phrase-based translation model for question retrieval in community question answer archives,” in ACL, 2011, pp. 653–662.