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

Deshmukh Ashvini B[1] Shelke Pooja P[2] Kokare Sayali A[3] Taware Saksha S[4] Prof. Chatse R.V[5]

[1,2,3,4]B.E Student [5]Guide

[1,2,3,4,5]Department of Information Technology

[1,2,3,4,5]SVPM's C.O.E. Malegaon (Bk), 413115, Savitribai Phule, Pune University, Maharashtra, India

Abstract— CQA helpful in answering real world question. CQA provide answer to human. Question retrieval in CQA can automatically find the most relevant and recent questions that have been solved by other users. 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. The translated words from other languages via non-negative matrix factorization. Contextual information is exploited during the translation from one language to another language by using Google Translate. Thus, word ambiguity can be solved based on the contextual information when questions are translated. Multiple words that have similar meanings in one language may be translated into a unique word or a few words in a foreign language. It is a word-based translation language model for retrieval with query likelihood model for answer. We use a translated representation by alternative enriching the original question with the words from other language in CQA. We translate the English question into other four language using Google translate which take into account contextual information during translation. If we translate the question word by word, it discard the contextual information. We would expect that such a translation would not be able to solve word ambiguity problem.

Key words: Community Question Answering, Statically Machine Translation, Non Matrix Factorization, Google Translator, Recursive Neural Network

I. 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 as well. This will give the user a wide range of pre answered questions to look for solution of his/her problem. Current systems fail to do so. Also these systems fetch related questions based on the keywords in it. Thus, if there is a question which is related to the topic but having other keywords, then that question is not retrieved, this is a major drawback of a system as there can be many circumstances where a semantically related question but not having similar keywords is not retrieved. The proposed system shows a way to retrieve questions which are related to the asked question but asked in other language as well as the questions that are related to the topic but not having similar keywords. The proposed system shows that this can be achieved when these questions are retrieved semantically instead of using keywords.

It is found that, in most cases, automated approach cannot obtain results that are as good as those generated by human intelligence. Along with the proliferation and improvement of underlying communication technologies, community Question Answering (CQA) has emerged as an extremely popular alternative to acquire information online, owning to the following facts. a. Information seekers are able to post their specific questions on any topic and obtain answers provided by other participants. By leveraging community efforts, they are able to get better answers than simply using search engines. In comparison with automated CQA systems, CQA usually receives answers with better quality as they are generated based on human intelligence. c. Over times, a tremendous number of QA pairs have been accumulated in their repositories, and it facilitates the preservation and search of answered questions.

II. System Description

A. *Functionality summary*

User Enter question in CQA.CQA check the question in dataset. CQA Factorize the question in query format. We translate the English questions into other four languages using Google Translate, which takes into account contextual information during translation. Remove word mismatch and word ambiguity. Use the algorithm SMT+NMF for optimization. Use Map Reduce on optimize matrix. Using ranking get expected result with best answer.

B. *User*

User first do the registration to the system. He login to the system by entering user id and password. Then, he ask the question in the system.

C. *CQA*

The CQA gives the users answer in textual format. After that it selects the answer medium for particular question.

D. *Google Translator*

Semantic information drawn from other languages. The word ambiguity and word mismatch problems have been solved.

III. Literature Survey

A. *Learning the Multilingual Translation Representations for Question Retrieval in Community Question Answering via Non-negative Matrix Factorization*

This paper proposes a way of fetching previously asked questions which are asked in different language 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.

B. *Finding Similar Questions in Large Question and Answer Archives*

In this paper, we discuss methods for question retrieval that are based on using the similarity between answers in the archive to estimate probabilities for a translation-based retrieval model. We show that with this model it is possible to find semantically similar questions with relatively little word overlap. Question retrieval that are based on using the similarity between answers in the archive to estimate probabilities for a translation-based retrieval model.

C. Entity based Q and A retrieval

Bridging the lexical gap between the users question and the question answer pairs in the Q and A archives has been a major challenge for Q and A retrieval. While useful, the eff ectiveness of these models is highly depend ant on the availability of quality corpus in the absence of which they are troubled by noise issues .Moreover these models perform word based expansion in a context agnostic manner resulting in translation that might be mixed and fairly general. This results in degrade retrieval performance. We explore strategies to learn the translation probabilities between words and the concepts using the Q and A archives and a popular entity catalog. Experiments conducted on a large scale real data show that the proposed techniques are promising. Semantic concepts for addressing the lexical gap issue in retrieval models for large online Q and A collections.

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

Question retrieval in CQA can automatically find the most relevant and recent questions that have been solved by other users. The word ambiguity and word mismatch problems bring about new challenges for question retrieval in CQA. 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. They can helpful in answering real world question. Challenge is word mismatch between the queried question and historical question.

IV. MODEL FORMULATION/ ALGORITHM

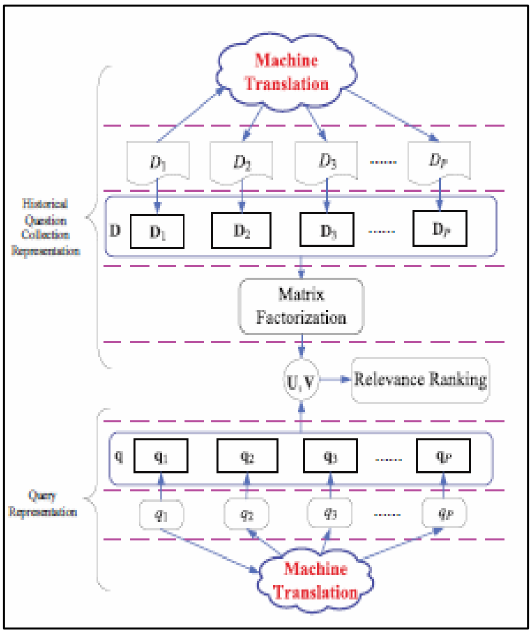
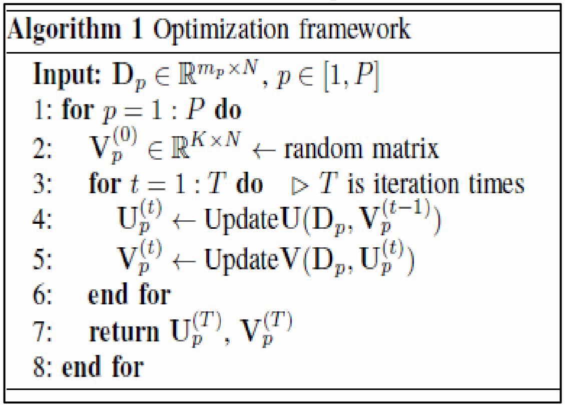


Fig. 1:

To tackle the info scantness 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 non-negative matrices whose product provides a decent approximation to the first matrix and has been shown to be superior to SVD in document clump.

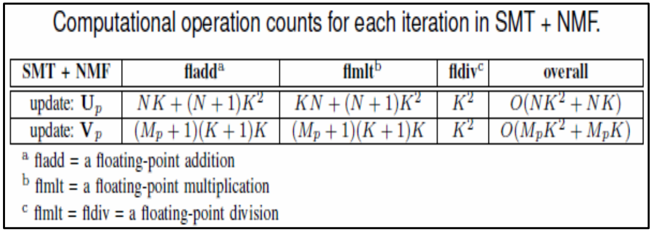

Algorithm 1 Optimization framework

```
Input: $D_p \in \mathbb{R}^{m_p \times N}$, $p \in [1, P]$
for $p = 1 : P$ do
    $V_p^{(0)} \in \mathbb{R}^{K \times N} \leftarrow$ random matrix
    for $t = 1 : T$ do    $\triangleright$ $T$ is iteration times
        $U_p^{(t)} \leftarrow \mathrm{UpdateU}(D_p, V_p^{(t-1)})$
        $V_p^{(t)} \leftarrow \mathrm{UpdateV}(D_p, U_p^{(t)})$
    end for
    return $U_p^{(T)}$, $V_p^{(T)}$
end for
```

V. ANALYSIS

In this segment, we have a tendency to discuss the time quality of our proposed methodology SMT + NMF. Besides expressing the quality of the algorithmic rule victimisation massive O notation, we have a tendency to conjointly count the number of arithmetic operations to produce a lot of details about time period. we have a tendency to show the ends up in Table two. Based on the change rules summarized in algorithmic rule one, it is not laborious to count the arithmetic operations of inner loop in SMT + NMF for every single language question illustration. Suppose in every inner iteration, the computation of (VpVTp + μpI)−1 in equation and (UTpUp + λpI)−1 in equation only need one time, the optimization Up and Vp take O(NK2+MpNK) and O(MpK2+MpNK) operations, respectively.

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 ΣPp=1 T × O(NK2 + MpK2 + 2MpNK). For each language Dp, the size of vocabulary Mp is almost constant as the number of questions increases. Besides, K ≪ min(Mp,N), so the overtime complexity depends on ΣP p=1 T × O(MpNK).


Computational operation counts for each iteration in SMT + NMF.

SMT + NMF	fladd[a]	flmlt[b]	fldiv[c]	overall
update: U_p	$NK+(N+1)K^2$	$KN+(N+1)K^2$	K^2	$O(NK^2+NK)$
update: V_p	$(M_p+1)(K+1)K$	$(M_p+1)(K+1)K$	K^2	$O(M_pK^2+M_pK)$

[a] fladd = a floating-point addition
[b] flmlt = a floating-point multiplication
[c] flmlt = fldiv = a floating-point division

Theoretically, the computational time is almost linear with the number of questions N and the number of languages P considered in the paper. Therefore, the proposed method can be easily adapted to the large-scale information retrieval task.

VI. Conclusion

As we all know the CQA system is getting tremendous 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. A system with use of translated representation is proposed in this paper. In this, 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.

References

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

[2] Singh, “Entity based q and a retrieval”, in EMNLP, 2012, pp. 12661277.

[3] 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.

[4] D. Bernhard and I. Gurevych, “Combining lexical semantic resource swith question and answer archives for translation-based answer finding”, in ACL, 2009, pp. 728736.

[5] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization”, in NIPS, 2000, pp. 556562.

[6] W. Xu, X. Liu, and Y. Gong, “Document clustering based on nonnegative matrix factorization” , in SIGIR, 2003, pp. 267273.