REVIEW OF ACHIEVING MULTIPLE LANGUAGE TRANSLATION FOR ENHANCED QUESTION-ANSWER PAIR RETRIEVAL IN CQA USING NMF

Priyanka sanvatsarkar[1], Dr.Sulochana Sonkamble[2]

[1]Miss.Priyanka yadav Sanvatsarkar, Computer,
JSPM NTC, PUNE, INDIA-411041

[2]Dr,Sulochana Sonkamble, Computer,
JSPM NTC, PUNE, INDIA-411041

---***---

Abstract – *CQA helpful in answering real world question's provided answer to human. Address recovery in CQA can naturally and the most important and late inquiries that have been tackled by different clients. In existing system elective approach to address the word ambiguity and word miss match issues by exploiting conceivably rich semantic data drawn from different language. 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 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 re trivial with query likelihood model for answer. If these translate the question word by word, it discards the contextual information. We would expect that such a translation would not be able to solve word ambiguity problem.*

Key Words: Natural Language Processing, Information Retrieval, Community Question Answering, Question Retrieval, Text Mining.

1. INTRODUCTION

In resent year community question answering like Yahoo! Answer is most popular service to use in business industry. The motivation behind this archive is to convict, break down and characterize abnormal state needs and elements of the taking in the multilingual interpretation representations for question recovery in group address noting by means of non-negative matrix factorization. It concentrates on the abilities required by the partners, and the objective clients. to give the fundamental and suitable data to a Mean Average Precision (MAP) user/examiner as content. The points of Interest of how the taking in the multilingual interpretation representations for question recovery in group address noting through non-negative matrix factorization satisfies these necessities are definite in the utilization case and supplementary determinations. The plan of the objective framework is given. The different parts of programming like information, program, and interfaces are planned. The venture estimating and booking, work breakdown structure is finished. The test arrange taking in the multilingual interpretation representations for question recovery in group address noting by means of non-negative matrix factorization is likewise given through a similar report.

To make community address noting entrances more helpful, it is essential for the framework to have the capacity to bring the inquiries asked in different dialects too. This will give the client an extensive variety of pre addressed inquiries to search for arrangement of his/her issue. Current frameworks neglect to do as such. Additionally these frameworks bring related inquiries in light of the watchwords in it. Along these lines, if there is a question which is identified with the theme yet having different catchphrases, then that question is not recovered; this is a noteworthy disadvantage of a framework as there can be numerous conditions where a semantically related question however not having comparable watchwords is not recovered. The proposed framework demonstrates an approach to recover questions which are identified with the made inquiry however asked in other dialect and the inquiries that are identified with the point yet not having comparative catchphrases. The proposed framework demonstrates this can be accomplished when these inquiries are recovered semantically as opposed to utilizing catchphrases. comprehensive and taxonomic tutorial information. The paper must emphasize concepts and the underlying principles and should provide authentic contribution to knowledge. If your paper does not represent original work, it should have educational value by presenting a fresh perspective or a synthesis of existing knowledge. The purpose of this document is to provide you with some guidelines. You are, however, encouraged to consult additional resources that assist you in writing a professional technical paper.

It is found that, much of the time, robotized approach can't get comes about that are in the same class as those produced by human insight. Alongside the expansion and change of basic correspondence advances, community Question Answering (CQA) has risen as a to a great degree famous other option to get in-arrangement internet, owning to the accompanying truths. Data seekers can post their particular inquiries on any point and acquire answers gave by different members. By utilizing community endeavors, they can show signs of improvement answers.

2. RELATED WORK

In this research paper the real test for question recovery in CQA is the word equivocalness and word jumble issues. Analysts have utilized question-answer sets to learn different translation models.

A] **SMT FOR QUERY EXPANSION IN AR**

To determine Experimental results show that SMT based expansion improves retrieval performance over local expansion and over retrieval without expansion. Only one technique are used in this paper and on that hard to detect the answered.

B] FINDING SIMILAR QUESTIONS IN LARGE QA

To fetch Question retrieval that is 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 conceivable to discover semantically comparative inquiries with moderately little word cover. It’s mainly work on similar word of question only but is cannot work on the other question answered and that why its performance is bad.

C] **LEXICAL SEMANTIC RESOURCES**

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.

D] QUESTION RETRIEVAL IN CQA USING WORLD KNOWLEDGE

To find out best method experiments conducted on a real QA data set show that with the help of Wikipedia thesaurus, the performance of question retrieval is improved as compared to the traditional methods. In this paper other kind of data set such as categorized of question not available on the forum.

E] **SEMANTIC RELEVANCE MODELING FOR CHINESE QA PAIRS**

(1) 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

F] **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.

G] WORD-BASED TRANSLATION MODEL

In this system word-based translation model which misuses the semantic similitude between answers of existing inquiries to learn translation probabilities, which permits them to coordinate semantically comparative inquiries regardless of lexical gap. However, these word-based translation models are considered to be context independent in that they don't take into record any contextual data in modeling word translation probabilities. In order to further improve the word-based translation model with some contextual data.

3. PROBLEM STATEMENT

There are many pre-existing community question answering (CQA) services like yahoo answers, quora etc. in which a user can ask/post a question and users from all over the world can answer that question. As their answers are based on real life experience, they can be pretty useful for the questioner. The CQA also retrieves and shows the user pre-answered questions which might be related to user's questions. This is done mostly using keywords. But this keyword strategy can be little less useful when it comes to searching questions which may not contain the same words or keywords but are still related to the user's questions. Also there can be word ambiguities when a word/question is translated from other language.

4. MOTIVATION

In existing framework when the client can seek the topic around then word ambiguity and mismatch problem happen. To conquer this problem, we attempting to actualizes existing framework. To provide the necessary and appropriate information to a user/questioner in the Form of text.

5. EXISTING SYSTEM ARCHITECTURE

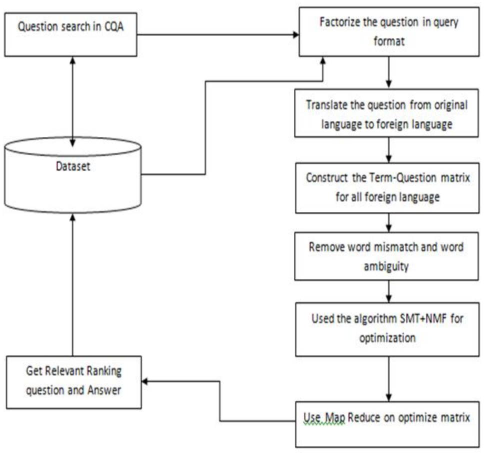
Question search in CQA
Factorize the question in query format
Translate the question from original language to foreign language
Dataset
Construct the Term-Question matrix for all foreign language
Remove word mismatch and word ambiguity
Used the algorithm SMT+NMF for optimization
Get Relevant Ranking question and Answer
Use Map Reduce on optimize matrix

Fig. Question Retrival Framework

1. Enter question in CQA.

2. To check the question in dataset.

3. Factorize the question in query format

4. To search the question in historical dataset.

5. We translate the English questions into other four languages using Google Translate, which takes into account contextual information during translation.

6. Remove word mismatch and word ambiguity.

7. Use the algorithm SMT+NMF for optimization.

8. Use Map Reduce on optimize matrix and using ranking get Expected result with best answer.

6. APPROACH PROBLEM STATEMENT

Let $L = \{l1, l2, \ldots, lP\}$ denotes language set, where P is the number of language, $l1$ denotes the natural language (for e.g., English) while $l2$ to lP are the foreign languages. Let $D1 = \{d(1)\ 1, d(1)\ 2, \ldots, d(1)\ N\}$ be the set of historical question set in original language, where N is number of historical questions in $D1$ with vocabulary size $M1$. Now translate each original historical question from language $l1$ into other languages lp ($p \in [2, P]$) by Google Translate. Thus, they can obtain $D2, \ldots, DP$ in different languages, and MP is the vocabulary size of Dp. A question $d(p)\ i$ in Dp is represented as a MP dimensional vector $\mathbf{d}(p)\ i$, in which each entry is calculated by tf-idf. The N historical questions in Dip are represented in a $Mp \times N$ term-question matrix $\mathbf{D}p = \{\mathbf{d}(p)\ 1, \mathbf{d}(p)\ 2, \ldots, \mathbf{d}(p)\ N\}$, in which each row corresponds to a term and each column corresponds to a question. they can enrich the original question representation by add the translated words from language $l2$ to $l\Sigma P$, the original vocabulary size is increase from $M1$ to $P\ p{=}1Mp$. the term-question matrix becomes $\mathbf{D} = \{\mathbf{D}1, \mathbf{D}2, \ldots, \mathbf{D}P\}$ and $\mathbf{D} \in R(\ \Sigma Pp{=}1Mp) \times N$.

Be that as it may, there are two issues with this improvement: (1) advancing the first inquiries with the interpreted words from different dialects makes the question representation considerably sparser; (2) measurable machine translation may present noise (Statistical machine translation quality is a long way from agreeable in real applications.). To tackle these two issues, we propose to leverage measurable machine translation to enhance question retrieval by means of network factorization, where qi represents a queried question, and $\mathbf{q}i$ is a vector representation of qi.

ALGORITHM OPTIMIZATION FRAMEWORK

Input: $\mathrm{D}p \in \mathbb{R}^{\mathrm{Mp\ x\ N}}, p \in [1, P]$

```
for $p = 1 : P$ do
    $\mathrm{V}(0)\ p \in \mathbb{R}^{K \times N}$ ← random matrix
    for $t = 1 : T$ do _ $T$ is iteration times
        $\mathrm{U}(t)\ p \leftarrow \mathrm{Update}(\mathrm{D}p, \mathrm{V}(t-1)^{p})$
        $\mathrm{V}(t)\ p \leftarrow \mathrm{Update}(\mathrm{D}p, \mathrm{U}(t)^{p})$
    end for
    return $\mathrm{U}(T)^{p}, \mathrm{V}(T)^{p}$
end for
```

7.PROPOSED SYSTEM:

A Web forum is a website or section of a website that allows visitors to communicate with each other by posting messages. Most forums allow anonymous visitors to view forum postings, but require you to create an account in order to post messages in the forum. When posting in a forum, you can create new topics or post replies within existing message. In proposed system introduce the forum site discussion about the related question and answering related selected category of datasets. Also support multilingual support for our proposed forum site.

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. Future work motivates further investigate the use of this method for other kinds of data sets, such as categorized questions from forum sites.

REFERENCES

[1] Guangyou Zhou, Zhiwen Xie, Tingting He, Jun Zhao and Xiaohua Tony Hu, "Learning the Multilingual Translation Representations for Question Retrieval in Community Question Answering via Non-negative Matrix Factorization". 2016, pp.2329-9290

[2]Stefan Riezler, IoannisTsochantaridis, Vibhu Mittal and Yi Liu, "Statistical Machine Translation for Query Expansion in Answer Retrieval". International Joint Conference on Artificial Intelligence.

[3] D. D. Lee and H. S. Sung, \Algorithms for non-negative matrix factorization", in NIPS, 2000, pp. 556-562.

[4] J. Jeon, W. B. Croft, and J. H. Lee, \Finding similar questions in large question and answer archives", in CIKM, 2005, pp. 84-90.

[5] D. Bernhard and I. Gurevych, \Combining lexical semantic resource swath question and answer archives for translation-based answer finding", in ACL, 2009, pp. 728-736

[6] Guangyou Zhou, Yang Liu, Fang Liu, Daojian Zeng, and Jun Zhao," Improving Question Retrieval in CQA UsingWorld Knowledge". International Joint Conference on Artificial Intelligence, 2010.

[7] baoxun wang, bingquan liu, xiaolong wang, chengjie sun, and deyuan zhang" Deep Learning Approaches to Semantic Relevance Modeling for Chinese Question-Answer Pairs"2011

[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.

[9] A. Singh, \Entity based q and a retrieval", in EMNLP, 2014, pp. 1266-1277.