MULTILINGUAL TRANSLATION VIA NON NEGATIVE MATRIX FACTORIZATION USING HADOOP.

Suyog S. Majgaonkar ,Atharva A. Karmarkar , Tushar Kisan Chau, Atul Ashoke Sonawane ,Dr.(Mrs.)A.P.Adsul

Abstract:

CQA helpful in answering real world questions provide answer to human. In this work, System proposes elective approach to address the word ambiguity and word mismatch issues by exploiting conceivably rich semantic data drawn from different language. The translated words from other languages via non-negative matrix factorization. Contextual 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 system translate the question word by word, it discards the contextual information. Then such translations would not be able to solve word ambiguity problem.

Keywords: NaturalLanguage Processing, Information Retrieval,CommunityQuestionAnswering,Question Retrieval,Text Mining.

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.

System Architecture

In community question answering system we are provide textual answer. 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. It will be helpful for ending out best answer and it will be easy to understand the user. There are various methodologies are use in Community question answer (CQA) system:

1. TF-IDF Calculation

TF: Term Frequency number of time that term occurs in document if we denote that row frequency.

IDF: The inverse document frequency is a measure of how much information the word provides that is whether the term is common or across the all document.

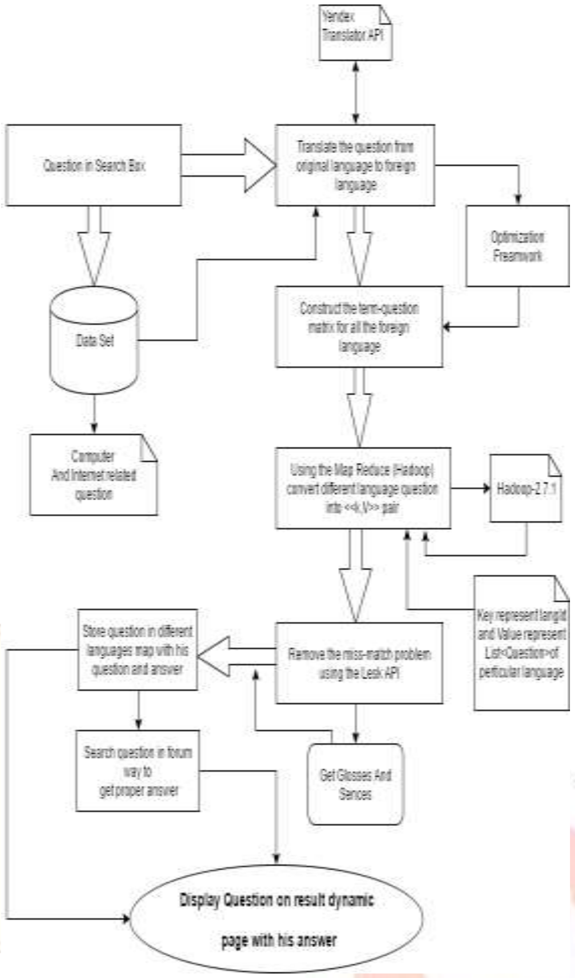
Yendex Translator API
Question in Search Box
Translate the question from original language to foreign language
Optimization Freamwork
Data Set
Construct the term-question matrix for all the foreign language
Computer And Internet related question
Using the Map Reduce (Hadoop) convert different language question into <<k,V>> pair
Hadoop-2.7.1
Key represent langId and Value represent List<Question>of perticular language
Store question in different languages map with his question and answer
Remove the miss-match problem using the Lesk API
Search question in forum way to get proper answer
Get Glosses And Sences
Display Question on result dynamic page with his answer

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

2. Optimization Framework

The optimization is use SMT + NMF. The data are stored in the format of the matrix. That matrix having whole information about the terms that are present in our dataset (Computer and internet). Terms as in the vocabulary of question calculated.

(a) Matrix Factorization: Matrix factorization is used to solve the data sparseness and noise. Proposed system Signiant improves performance by using matrix factorization.

(b) NMF:

Non-Negative matrix factorization is used to create term and question matrix.

3. Lesk API.

Lesk API is nothing but an algorithm which is use to solve word ambiguity and word miss match problem. Also use to finding the index of best answer senses glosses. Glosses is nothing but number of meanings of that question.

4. Map Reduce.

Map Reduce (using the hadoop) is a computing model that supports distributed computing on large datasets. Map Reduce expresses a computing task as a series of Map and Reduce operations and performs the task by executing the operations in a distributed computing environment. In this section, we describe the implementation of SMT + NMF +Lesk on Map Reduce, referred to as distributed SMT + NMF +lesk . At each iteration, the algorithm updates Up and Vp in matrix format using the mapper and reducer functionality.

5. CQA Forum.

CQA forum site which is like chat application. CQA forum is used when user is not satisfied any answer then he will asked that answer to another user which is online on chat application. This chat application is very useful to user.

Related Work:

In proposed system we demonstrate different techniques by the references in that, John Lafferty examine problem of language model smoothing and its influence on retrieval performance so because of smoothing adjusts the maximum likelihood estimator so as to correct the inaccuracy due to data sparseness [1]. 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 possible to find semantically similar questions with relatively little word overlap[2].Experimental results show that SMT based expansion improves retrieval performance over local expansion and over retrieval without expansion.[3]. Parallel training dataset the definitions and glosses provided for the same term by different lexical semantic resources.[4].The Xiao long Wang introduced two concept as follows a)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 modelling the semantic relevance for the QA pairs, using only word occurrence features. And b) 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 [5]. Sometime inputs are typically a single sentence and outputs are either continuous or a limited discrete set. Neural networks have not yet shown to be useful for tasks that require mapping paragraph-length inputs to rich output spaces so we get the limited and meaningful output [6]. Amit Singh introduce highly dependent the availability of quality corpus in the absence of which they are troubled by noise in the question answering pair for that they get on semantic concepts for addressing the lexical gap issue in retrieval models for large online Q&A collections[7].On next Guan you Zhou1 introduced the cross lingual language supported in question answer pair but it was takes only two language support in this the author using pretext classification methods for tasks like factoid question answering typically use manually defined string matching rules or bag of words representations technique[8].To learn continuous word embeddings with metadata of category information within CQA pages for question retrieval. To deal with the variable size of word embedding vectors, we employ the framework of fisher kernel to aggregate them into the fixed length vectors. Experimental results on large-scale real world CQA data set show that our approach can significantly outperform state-of-the-art translation models and topic-based models for question retrieval in CQA[9].That main aim is work on described an answer ranking engine for non-factoid questions built using a large community-generated question-answer collection[10].

Methodology and Algorithms

Algorithm 1 Optimization framework

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

for p = 1 : P do

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

Create the matrix Dp with TF-TDF format with each keyword format

Factorize the main matrix Dp in two parts for optimization like Up and Vp

matrix Dp size define as Mp*N

where Mp is a vocabulary wise

N is historical question

We want to factorize this matrix in two part for optimization we are given some 'k' value so our matrix Up is like that mp*K and matrix Vp k*N

Now we done following step

1)To remove the noise in the word those are repeated

whaen user enter the question in search bar

Update Up

2)optimize the row by row to check the tf-idf of each and every term/vocabulary

3)and reduce the stop word and give the main word related to questions

Update Vp

4) optimize the column by column to check the tf-idf of each and every term/vocabulary

5) and reduce the stop word and give the main word related to question

6) decompose the matrix upto 't' iteration time

7) Do like that with our original language and foreign language data set

8)The least squares problem with opiginal language norm regularization find on next step

9)The least squares problem with foreign language norm regularization find on next step

where regularization is: to find the sparceness of matrix Dp and matrix Vp

Algorithm 2:

Lesk Algorithm:

```
1)function SIMPLIFIED LESK(word, sentence) returns best sense of word
2)best-sense <- most frequent sense for word
3)max-overlap <- 0
4)context <- set of words in sentence
5)for each sense in senses of word do
6)signature <- set of words in the gloss and examples of sense
7)overlap <- COMPUTEOVERLAP (signature, context)
8)if overlap > max-overlap then
9)max-overlap <- overlap
10)best-sense <- sense
11)end return (best-sense)
```

Algorithm 3:

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 we are introduce the forum site discussion about the related question and answering related selected category of datasets. we also support multilingual support for our proposed forum site.

Forum Site:

```
1)User U={u1,u2,u3......un); complete the registration(User)
2)getUniqueId() <-0;
  message<-null
  subject<-null
  language<-L=(l1,l2,l3....ln);
foreach(i in U) do
if(User register)then
//generate_unique_userId
getUniqueId<-i;
```

```
else
//not_register
 registration(User)
end if
3)includeCurrentDiscussionTopic()
//privious user are talking on perticular subject
getSubject()
4)getFamiliarLanguage(language l)
5)discussForum()
//send message in the group
message<-sent message;
6)return message;
```

Results and Discussion:

In this system, dataset of 'Yahoo.com/questions' data is used for experiment purpose, in that 'computer and internet' category have more than 5000 question present. That question translated into different languages using yandex translator. But in all question having more similar word for that first calculate the tf-idf ratio of all languages. This all question divided into set of different languages using the Map Reduce technology (map<langid,List<question>>,Reducer<AllLang,Map<langid,List<questions>>)

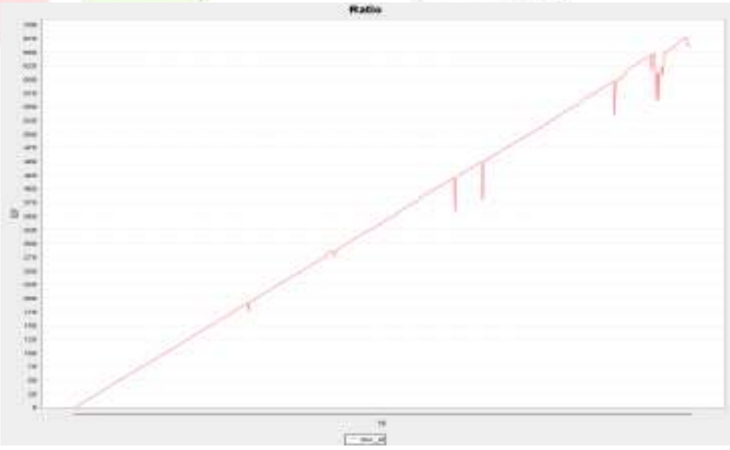

Fig 2.TF-IDF Ratio

After finding the IF-IDF the optimization with lesk finds the most relevant question in to datasets and it gives the all languages related question.

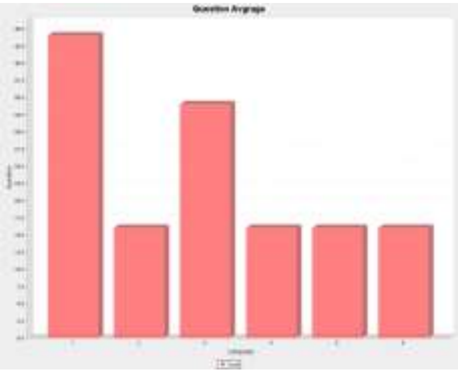

Fig 3. Question count after Map Reduce

The performance of the system calculate as, when out Map Reduce work is going to start and how much time required for fetching the relevant question using optimizing framework.

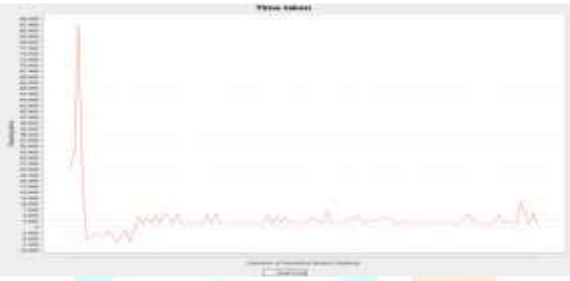

Fig 4. Performance of CQA

\

This graph shows comparison of simple and hadoop search question, simple search it takes too much time for searching question rather than Hadoop

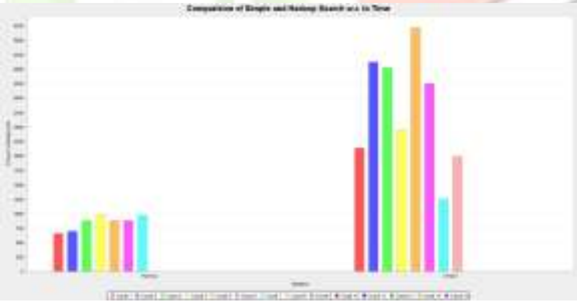

Fig 5.Simple search VS Hadoop search

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

Reference:

[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] A. 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 swath 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 non negative Matrix factorization" , in SIGIR, 2003, pp. 267273.