

Survey on Community Question Answering System to Solve the Lexical Gap

B. Deeppikaa¹, Geetha Vinodh², Sri Heera³

Student, Associate Professor

Department of Computer Science and Engineering

Easwari Engineering College, Chennai, India

Corresponding Authors' email id: deeppikaa19@gmail.com

Abstract

Web search engines give a ranked list of related documents based on users keywords which depends on various aspects like popularity measures, keyword match, and frequency of accessing documents in which users have to check every specific document for getting the desired information and it causes information retrieval a prolonged process. Community Question Answering (CQA) system focus to deliver users short and precise answers instead of irrelevant documents. CQA is a specialized application which deals with information retrieval which has an ability to retrieve the right answers to questions posed in natural language. Natural Language Processing (NLP) techniques used to process a question, then searches for the required information regarding user questions to determine the answer accurately. This survey mainly focus on different approaches to solve the problems arising due to the lexical gap and also rank the accurate answers in question and answering blogs such as community question answering websites.

Keywords: *Community Question Answering system, Information Retrieval, Lexical Gap, Natural Language Processing.*

INTRODUCTION

In recent years, large amount of memory is employed by the historical web pages to retrieve the vital information, from those

pages mainly in the blogs such as traditional Frequently Asked Questions (FAQ) archives and the emerging Community Question Answering (CQA)

services, such as Yahoo! Answers, Live QnA, HealthTap and Zhihu.com. The web content of these web sites is usually organized as questions and the answers associated with metadata from which the requesting users categorize the questions and the respondents reply the best answers. This results in CQA archives to have valuable resources for various tasks like question-answering and knowledge mining, etc. One fundamental task for reusing the contents in CQA is to find similar questions for newly queried questions, as questions are the keys to accessing the knowledge in CQA. Then, the best answers to those similar questions used to answer the queried questions, results in Question retrieval.

QUESTION ANSWERING TYPES

Question classification:

Open-domain QAS is deals with questions of nearly everything and Closed-domain QAS deals with questions in a specific domain.

Data Source classification:

Structured data deals with relational database and unstructured data deals with documents / web pages in Internet

Answer classification:

Extracted answer lies directly in the database and Generated answer needs to be generated or formulated from the retrieved data.

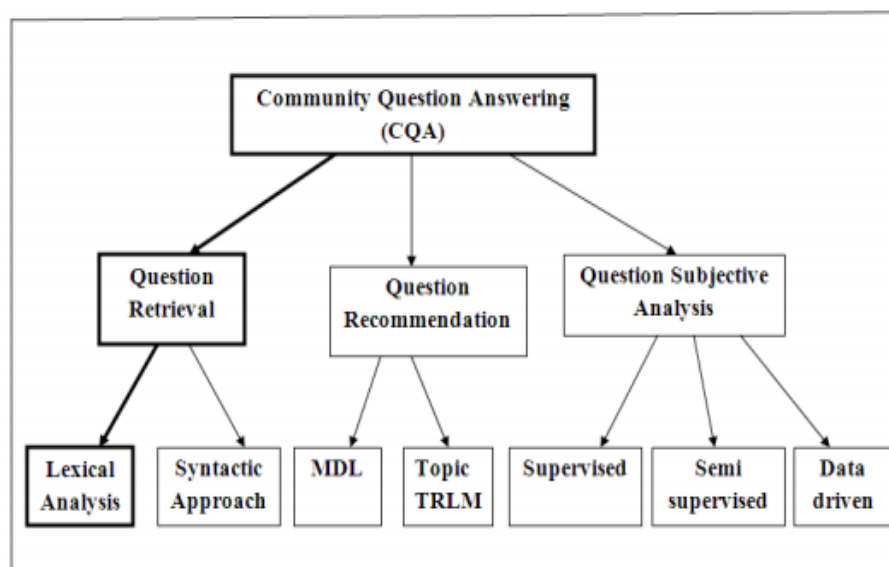


Fig 1.Question answering system

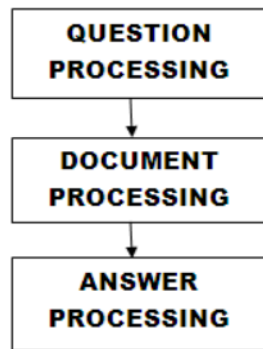


Fig 2. Architecture of question answering system

Question Processing used to analyze sentence tagging, sentiment analysis and also classify question to determine the intention of the question. It also does question reformulation so that the document processing can handle it. Document Processing in Information Retrieval (IR) module used to retrieve the most Relevant data from the system. Answer processing used to identify useful information from the document and also rank the answers if there are multiple answers and returns the most relevant answer [4].

Question Classification and Answer

In Question processing the system first should analyze the type of question. Questions can be classified into two categories: Question with 'WH' question words such as what, where, who, whom, which, how, why and etc.

Questions with 'modal' or 'auxiliary' verbs that their answers are Yes/No[4]. Table 1 shows question words, type of questions and answers.

Table 1 Question classification and answer

Question classification	Entity name	Example
When	Date/time representation	When did the first world war start?
What	Date/Number	What is temperature of Spain?
Where	Location	Where is president located?
Who	Person	Who is the first president of India?
Which	Location/person	Which is the oldest language?
Why	Reason	Why is the sky blue?

Approaches in Question Answering Systems Linguistic Approach

Linguistic approach understands natural language text, linguistic & common knowledge. Linguistic techniques such as tokenization, POS tagging and parsing. These were implemented to user's question for formulating it into a precise query that extracts the respective response from the structured database.

Statistical Approach

Availability of huge amount of data on internet increased the importance of statistical approaches. A statistical learning

method gives the better results than other approaches. Online text repositories and statistical approaches are independent of structured query languages and can formulate queries in natural language form. Mostly all Statistical Based QA system applied a statistical technique in QA system such as Support vector machine classifier, Bayesian Classifiers, maximum entropy models.

Pattern Matching Approach

Pattern matching approach deals with expressive power of text pattern, it replace the sophisticated processing involved in other computing approaches. Most of the pattern matching QA systems uses the surface text pattern, while some of them also rely on templates for response generator.

Deep Learning approach

Most of the Deep Learning methods is used to implement one or more component of QAS such as question classification, sentence selection, etc. It converts Natural Language into a computable form e.g. using word embedding or using Neural Language Model e.g. using RNN / LSTM or C.

RESEARCH BACKGROUND

Community Question Answering

Community Question answering (CQA) is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that automatically answer questions posed by humans in a natural language. A CQA implementation, usually a computer program, may construct its answers by querying a structured database of knowledge or information, usually a knowledge base. More commonly, CQA systems can pull answers from an unstructured collection of natural language documents. Table 2 shows the comparison between the CQA and QA In many cases, the community generated content, however, may not be directly usable due to the vocabulary gap. Users with diverse backgrounds do not necessarily to share the same vocabulary. Stack Overflow, one of the technical question answering sites for users can ask technical related questions. The entire technical questions are solved by technical experts, where different users ask different questions for the similar answer and there is a gap exists between what is asked and what is answered either syntactically or semantically and such gaps ends in lexical gap.

Table 2 Comparison of CQA and QA

	Community QA	QA
Question Type	Factoid Single Sentence Questions	Multi Sentence Questions
Source of Answers	Extracted from Documents in a corpus	User generated
Quality of Answers	Generally very high	Varies a lot, depending on the community
Availability of Meta Data	None	Best answer selection and upvotes/downvotes
Time Lag	Automatic and Immediate	Generally has to wait For an answer

NATURAL LANGUAGE PROCESSING (NLP) Natural language processing can be defined as the ability of a machine to analyze, understand, and generate human speech. The goal of NLP is to make interactions between computers and humans feel exactly like interactions between humans and humans [1]. QA system take a natural language question as input, convert the question into a query and forwards it to the next module. When a set of required documents is retrieved, the CQA system extracts an accurate answer for user posted question [5].

Deep Analytics:

Deep analytics involves the application of advanced data processing techniques in order to extract specific information from large or multi-source data sets. Deep analytics is often used in the financial sector, the scientific community, the pharmaceutical sector, and biomedical industries. Increasingly, however, deep

analysis is also being used by organizations and companies interested in mining data of business value from expansive sets of consumer data. Machine Translation: Natural language processing is increasingly being used for machine translation programs, in which one human language is automatically translated into another human language.

Information Retrieval

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines.

In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy. User queries are matched against the database information. However, as opposed to classical SQL

queries of a database, in information retrieval the results returned may or may not match the query, so results are typically ranked. This ranking of results is a key difference of information retrieval searching compared to database searching.

Lexical Gap

Lexical gaps are instances of lack of lexicalization detected in a language while comparing two languages or in a target language during translation. Although the problem seems to be minor and clear, one gets rather the opposite impression after an excursion through the linguistic literature on lexical gaps. Different living environment and experience causes some lexical meanings in one language does not exist in another language. Lexical gaps are filled by means of Hypernyms or by Antonymous Expressions

Rank Based Answer Generator

Rank based answer generation is to determine that the user post their exact questions and obtain answers from other users. This helps to get better answers than compared with search engines. The generated answers will be better than the answers searched by the browsers and accurate result is obtained. A enormous number of QA pairs have been gathered in their repositories, and it also preserve the answers of the search results. The Support

Vector Machine algorithm is a retrieval function that uses pairwise ranking methods to sort the accurate results based on how relevant they are to a particular user query.

RELATED WORK ON COMMUNITY QUESTION ANSWERING

Al-Harbi O., Jusoh S., and Norwawi N. (2011) [1]: Resolving the lexical ambiguity problem by combining two pieces of knowledge; context knowledge and ontology of concepts knowledge of interesting domain, into shallow natural language processing (SNLP). The combination of these knowledge is used to decide the most possible meaning of the word. It lacks in resolving the syntactic ambiguity in natural language questions.

Bernhard D., and Gurevych I. (2009) [2]: Monolingual translation probabilities have recently been introduced in retrieval models to solve the lexical gap problem. It is evaluated with three datasets for training statistical word translation models for use in answer finding question-answer pairs, manually-tagged question reformulations and glosses for the same term extracted from several lexical semantic resources. The existing system lacks in question analysis by automatically identifying question topic and question focus.

Guangyou Zhou, Zhiwen Xie, Tingting He. (2016) [3]: The State-of-the-art approaches address these issues by implicitly expanding the queried questions with additional words or phrases using monolingual translation models. The task of question retrieval in CQA and represent a question as Bag-of-Embedded-Words (BoEW) in a continuous space. The existing system lacks in pairs to learn various translation models to bridge the lexical gap problem.

Qiu X., Huang X. (2015)[6]: The convolutional neural tensor network architecture to encode the sentences in semantic space and model their interactions with a tensor layer and also help to learn better word embeddings. The existing system lacks in to efficiently detect local reuses at the semantic level for large scale problems.

Shen Y., Rong W., Jiang N., Peng B., Tang J., and Xiong Z. (2017) [7]: A Word Embedding based Correlation (WEC) model is proposed by integrating advantages of both the translation model and word embedding and also leverages the continuity and smoothness of continuous space word representation to deal with new pairs of words that are rare in the training parallel text. It is necessary to focus on questionquestion matching

tasks, or Multilanguage question retrieval task. It does not solve parallel detection problems.

Wei-Nan Zhang, Zhao Yan Ming, Yu Zhang. (2016) [8]: It Explore the key concept identification approach for query refinement and a pivot language translation based approach to explore key concept paraphrasing. These word embedding models contribute the most the performance. The existing system generates noise samples for each input word to estimate the target word causes inefficiency.

Zhang K., Wu W., Wu H., Li Z., Zhou M. (2014) [9]: They are heterogeneous for both the literal level and user behaviours. Conduct a series of experiments to evaluate our proposed approaches automatically on large-scale data sets. The existing system cannot be directly used for large scale problems Zhou G., Chen Y.,

Zeng D., and Zhao J. (2013) [11]: A novel Question-Answer Topic Model (QATM) to learn the latent topics aligned across the question-answer pairs to alleviate the lexical gap problem. A faster and better retrieval model for question search by leveraging user chosen category. The existing system lacks in the localness

and hierarchy intrinsic to the natural language problems.

Zhou G., He T., Zhao J., Hu P. (2015) [12]: The framework of fisher kernel to aggregated them into the fixed length vectors. That metadata of category information benefits the word embedding learning for question representation. The existing system have problem from different aspects such as extraction methods with or without linguistic knowledge.

Zhou G., Cai L., Zhao J., and Liu K. (2011) [10]: A machine learning algorithm that aims to predict a ranking among all the possible labels, to perform question classification. Training process does not need many training data, which are always expensive to obtain in the CQA services. Combining the bilingual translation or category information can be done for better question retrieval. Sometimes when we search a query in the services, the system always tells us it cannot find any results.

PROPOSED WORK

We present a novel approach to determine accurate answer in CQA based on user questions instead of retrieving the entire document. It categorized into offline learning and online search component. In

the offline learning component, instead of prolonged process we construct the positive, neutral, and negative training samples in the forms of preference pairs. We then propose a robust pairwise learning to rank model to incorporate these three types of training samples. In the online search component, for a given question, we first collect a pool of answer candidates via finding its similar questions. We then employ the offline learned model to rank the answer candidates via pair wise comparison. IV.

CONCLUSION AND FUTURE WORK

The main goal of a community question answering system is to retrieving answers to questions rather than full documents as most information retrieval systems. Therefore we analyzed the different approaches to solve the problem arising due to lexical gap and also determine to rank the accurate answers for the user question instead of retrieving the entire document. Lexical Gap is complex phenomenon, in which it requires many factors to obtain exact answers for the causes of classification and translation strategies. The future work towards multilingual, multi-knowledge source CQA systems that are capable of understanding noisy, human natural language input and also to overcome the problems in pairwise ranking such as

noise-sensitive, largescale preference pairs, and loss of information about the finer granularity in the field of CQA.

Ontology’ in International Journal of Computer Applications (0975 – 8887) Volume 97 – No.20.

REFERENCES

1. Al-Harbi O., Jusoh S., and Norwawi N. (2011), ‘Lexical Disambiguation in Natural Language Questions (NLQs)’ in IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 2, July 2011
2. Bernhard D., and Gurevych I. (2009), ‘Combining lexical semantic resources with question & answer archives for translation-based answer finding’, in Proceedings of the ACL, pp. 728–736.
3. Guangyou Zhou, Zhiwen Xie, Tingting He. (2016), ‘Question-answer topic model for question retrieval in community question answering’, in IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 24, Issue:7).
4. Pawan Kumar., Raj Kumar Goel., and Prem Sagar Sharma. (2014), ‘A New Architecture of Automatic Question Answering System using
5. Poonam Gupta., and Vishal Gupta. (2012), ‘A Survey of Text Question Answering Techniques’ in International Journal of Computer Applications (0975 – 8887) Volume 53– No.4.
6. Qiu X., Huang X .(2015), ‘Convolutional neural tensor network architecture for community-based question answering’, in IJCAI proceedings of the twenty fourth international joint conference on artificial intelligence.
7. Shen Y., Rong W., Jiang N., Peng B., Tang J., and Xiong Z. (2017), ‘Word Embedding Based Correlation Model for Question/Answer Matching’, in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence.
8. Wei-Nan Zhang, Zhao Yan Ming, Yu Zhang .(2016), ‘Capturing the Semantics Of Key Phrases Using Multiple Languages For Question Retrieval’, in IEEE Transactions on

-
- Knowledge and Data Engineering (Volume: 28, Issue: 4).
9. Zhang K., Wu W., Wu H., Li Z., Zhou M. (2014), 'Question Retrieval With High Quality Answers In Community Question Answering', Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management pp. 371-380.
10. Zhou G., Cai L., Zhao J., and Liu K. (2011), 'Phrasebased translation model for question retrieval in community question answer archives', in Proceedings of the ACL, pp. 653–662.
11. Zhou G., Chen Y., Zeng D., and Zhao J. (2013), 'Towards Faster And Better Retrieval Models For Question Search', CIKM Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pp. 2139- 2148.
12. Zhou G., He T., Zhao J., Hu P. G.(2015), 'Learning Continuous Word Embedding With Metadata For Question Retrieval In Community Question Answering', Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 250–259.
-