NATURAL LANGUAGE BASED QUESTION ANSWERING USING DISTRIBUTIONAL SEMANTIC MODEL

Toral Desai

ME, computer Science and Engineering
HGCE, Ahmadabad, India

ABSTRACT: QA is a application or specific type of information Retrieval. Today’s world of the web , information is store with large repositories of information. The Big Problem we face is that large amount of information available that allow us to find what is relevant and cannot be managed without automatic Search. We solve this problem by using Question answering System(QAS). The main goal of QA system is able to answer users Question in Natural Language. The QA system can Fulfil the needs of user as they provide with faster and appropriate answers to user question.

Keywords: Question answering System, Natural Language Processing, Information retrieval, closed domain, open Domain

1. INTRODUCTION

The first question answering system introduce in the 1960s, Baseball was able to answer domain-specific natural language questions which was about the baseball games played in American league over one season. There are various types of QA systems are developed so far such as, closed –domain QAS, open domain QAS , web based QAS, information retrieval(IR) QAS , Information Extraction (IE) based QAS, rule based QAS.

1.1. Architecture of Question Answering Systems

In Figure 1., a Basic Architecture of QA system consists of three Common modules, each of which has a core component beside other supplementary components: “Question Processing Module” whose heart is the question classification, the “Document Processing Module” whose heart is the information retrieval, and the “Answer Processing Module” whose heart is the answer extraction[6].

The Question processing module include three parts: Query Interface, question analyzer and Question classification . The information retrieval component is used to retrieve the relevant documents based upon important keywords appearing in the question. The answer processing module is responsible for identifying, extracting and validating answers from the set of ordered paragraphs passed to it from the information retrieval module.

The detailed description on the working of QA system is given in section 2.1.with the study of various approaches used to develop different modules of QA system.
Fig 1. Basic Architecture Question Answering System.

1.2 Random Indexing
Random Indexing (RI) introduced an effective and scalable method for constructing DSMs from large volumes of text. The method descends from work by Kanerva on Sparse Distributed Representations. The method is scalable because it performs a type of implicit dimensional reduction and it performs this in an incremental fashion. The distributed representation for a context is an index vector:

$$A_{ij} = \begin{cases} +1 & \text{with probability } \frac{\epsilon}{2^k} \\ 0 & \text{with probability } \frac{k - \epsilon}{2^k} \\ -1 & \text{with probability } \frac{\epsilon}{2^k} \end{cases}$$

where $\epsilon$ is the number of non-zero elements and $k$ is the dimensionality.

Index vectors are binary vectors with a small number of non-zero elements, which are either +1 or -1, with equal amounts of both. For example, if the index vectors have twenty non-zero elements in a 1024-dimensional vector space, they have ten +1s and ten -1s. It is easy to see that the expected value of the scalar product of two such randomly generated vectors is zero. When we multiply each coordinate in one vector by the corresponding coordinate in the other vector, in most cases at least one of the multiplicands is zero. In the rare cases that both multiplicands are non-zero, half the time we expect the signs of the multiplicands to be the same, contributing +1 to the scalar product, and the other half of the time we expect the signs of the multiplicands to differ, contributing -1. So on average, these contributions will cancel out. To judge orthogonality we actually calculate the angle between the vectors, so we normalize by the lengths of the vectors, leaving any non-zero results even closer to zero.

1.3 Circular Convolution
Circular Convolution does not increase the dimensionality of vectors. The Circular Convolution is binding operator. The Circular Convolution of Two vectors A and B is:

$$c = a \otimes b$$

Where each element $C_j$ of C is

$$C_j = \sum_{k} a_k b_j - k \mod n$$

(2)
Figure 2.2 Convolution of two vectors $a$ and $b$ for $n=3$.

A Convolution of Two vectors $a$ and $b$ for $n=3$ has the following form (see fig.2.2):

$$
C_0 = a_0b_0 + a_1b_2 + a_2b_1 \\
C_1 = a_0b_1 + a_1b_0 + a_2b_2 \\
C_2 = a_0b_2 + a_1b_1 + a_2b_0
$$

For Example, let $A = [1,2,3]$ and $B=[4,5,6]$. Then, $c = a \otimes b$ can be computed:

$$
C_0 = a_0b_0 + a_1b_2 + a_2b_1 = 1*4 + 2*6 + 3*5 = 31 \\
C_1 = a_0b_1 + a_1b_0 + a_2b_2 = 1*5 + 2*4 + 3*6 = 31 \\
C_2 = a_0b_2 + a_1b_1 + a_2b_0 = 1*6 + 2*5 + 3*4 = 28
$$

2. LITERATURE REVIEW

Piero Molino, Pierpaolo Basile, Annalina Caputo, Pasquale Lops, Giovanni Semeraro [1], investigates the role of distributional Semantic Models in Question Answering System (QAS). The Method to integrate DSM into QAS, Called QuestionCube. The QuestionCube is a framework for QA that merge several techniques to retrieve passages containing the exact answers for Natural Language Question. The authors propose several kinds of DSMs based on classical Term-Term co-occurrence Matrix (TTM), latent semantic analysis (LSA), Random Indexing (RI) and combination of last two Techniques. Authors idea is that DSMs approaches can help to Semantic relatedness between users’ questions and candidate answers by exploiting paradigmatic relations between words.

Walke P.P, Karale, S.[2], shows the all implementation approaches for different categories of QAS. First approaches is Closed-domain QAS. It is introduce in 1961 (e.g., BASEBALL) and 1973 (e.g., LUNAR). In Open Domain based QAS, The most important challenge of an open domain system is its database The efficiency of any system depends on how well the database is arranged and maintained. A vector space model is a kind of model which can be used for classifying the candidate answers. In WEB BASED QAS, the most important property is “snippet – tolerant” property which allows it to provide correct responses to the QAS while searching answer through search engines like Google, yahoo etc.

In Information Retrieval or Information Extraction (IR/IE) based QAS, IR system works on the interaction between human and computer when used to search the answer for posed
question. IE systems are used for extracting the correct answer from the retrieved documents. Developing Rule based QAS is bit challenging task as the developer needs to consider virtually all the possible topics on which the system may get tested.

Jaspreet Kaur, Vishal Gupta [3], discussion regarding different Question Answering types. In addition they describe Mean Reciprocal Rank (MRR) used to evaluate the performance of different question answering systems. They also discuss the recent question answering systems developed and their corresponding techniques.

Semantic relatedness measures quantify the degree in which some words or concepts are related, considering not only similarity but any possible semantic relationship among them. Relatedness computation is of great interest in different areas, such as Natural Language Processing, Information Retrieval, or the Semantic Web. Emadzadeh, E., Nikfarjam, A., Muthaiyah, S [4], we explore the use of a semantic relatedness measure between words, that uses the Web as knowledge source. This measure exploits the information about frequencies of use provided by existing search engines. Semantic measures can also be defined between lexically expressed word senses, or between whole texts. Three main kind of measures are defined in this paper: semantic similarity, semantic relatedness and semantic distance. QA system consists of three Common modules, each of which has a core component beside other supplementary components: “Question Processing Module” whose heart is the question classification, the “Document Processing Module” whose heart is the information retrieval, and the “Answer Processing Module” whose heart is the answer extraction.

Gregory E. Cox & George Kachergis & Gabriel Recchia & Michael N. Jones[5] In this article, we have demonstrated how various proposed word-form encodings can be implemented as holographic reduced representations and how the resulting representations may be used to make predictions about performance in masking priming tasks and in unprimed LD and word naming tasks. We have also introduced a novel holographic representation for word-forms that is relatively simple to compute and satisfies a variety of empirical constraints on word similarity and have shown how this orthographic representation (and, in principle, others) can be integrated with semantic and syntactic information in a unified model of the lexicon.

Ali Mohamed Nabil Allam, Mohamed Hassan Haggag [6] , all question answering system approaches is defined by author and all described the minor lamination of all QA researchers. In Question Processing Module, different approach such as Machine learning based approach, rule based approach, hierarchical taxonomy, flat taxonomy are mainly used in different systems. In Document Processing Module, used the web corpus and knowledge-based corpus approach. In Answer Processing Module, used the text pattern and named entity approach.

It summarized and organized recent research results in a novel way that integrated and added understanding to work in the question-answering system field. It is impossible for a survey to include all or even most of previous research, this survey included only the work of the top-publishing and top-cited authors in the QAS field. This survey [6] also included research containing minor limitations to show how these limitations were discovered, faced and treated by other researchers.

Ehsan Emadzadeh, Azadeh Nikfarjam, Saravanan Muthaiyah [7], research different semantic relatedness functions called “Measure of Semantic Relatedness (MSR)” are discussed and
compared. They found that the quality and accuracy of MSRs are different when applied in various contexts. In this paper they compared several MSR algorithms using different corpuses and have analyzed the results.

Poonam Gupta, Vishal Gupta [8], they discussed some of the approaches like Web Based QAs, IR / IE Based QAS, Restricted Domain QAS, Rule Based QAs used in the existing QA system and proposed a new architecture for QA system retrieve the exact answer. In [8] discussed all basis component of Question Answering system. Answering system has become an important component of the online education platform.

R.Mervin [9], presents a survey of various types of QA systems. These QA systems are classified as Text based QA systems, Factoid QA systems, Web based QA systems, Information Retrieval or Information Extraction based QA systems, Restricted Domain QA systems and Rule based QA systems. The paper further investigates a comparative study of these models for different type of questioners which led to a breakthrough for new directions of research in this area.

Marta Tatu and Dan Moldovan [10], They present a logic-based semantic approach for the recognizing textual entailment task. The system participating in the RTE competition used a set of world-knowledge, NLP, and lexical chain-based axioms and an in-house logic prover which received as input the logic forms of the two texts enhanced with semantic relation instances. Because the state-of-the-art semantic parsers cannot extract the complete semantic information encoded in text, the need for semantic calculus in NLP became evident. They introduce semantic axioms that either combine two semantic instances or label relations between the frame elements of a given frame.

Reinland Kim T. Amplayo, Sean Carlo C. Bermejo and Michael John N. Pedros [11], Holographic reduced representation offers new unconventional solution to one of the basic problems of artificial intelligence and cognitive science, which is to find a suitable distributive coding of structured information. The used distributed representation is based on two binary operations: unary operation “involution” and binary operation “convolution” over a domain of n-dimensional randomly generated conceptual vectors, which elements content normal distribution N(0,1/n).

Gregory E. Cox & George Kachergis & Gabriel Recchia & Michael N. Jones [12], we have demonstrated how different proposed the word-form encodings can be implemented as HRR and how the resulting representations may be used to make forecast about performance in unprimed LD and word naming tasks. We have also represent the novel holographic presentation for word-forms that is relatively simple to compute and satisfies a variation of empirical constraints on” word similarity and have shown how this orthographic presentation can be integrated with semantic in a unified model of the lexicon.

3. PROPOSED METHODOLOGY

Step 1. Generate index vectors for each terms using equation 1.
Step 2. Calculate the context vector of the target words.
Step 3. Create full sentence vector by composing context vectors of the target terms using circular Convolution.
Step 4. Find similarity score of document vector of all sentences and user question using cosine similarity.
Step 5. Re-rank the candidate answers according to their similarity scores. In algorithm, Generate index vectors for each terms using equation 1 show in the section 1.2 in step 1. For each target word generate the context vector in step 2. After generating context vector for each term generate the full sentence vector by composing context vector of the target term using circular Convolution. Find similarity score of document vector of all sentences and user question using cosine similarity. All sentences score is generate using the cosine similarity then after Re-rank the candidate answers according to their similarity scores.

4. EXPERIMENT RESULTS

Figure 4.1. shows GUI for the Display the result of the DSM algorithm. In GUI Enter the Question and click on the search Button. And three algorithm Term Term matrix, Random Indexing using addition and Random Indexing using Circular Convolution is show. you can select check Box and the result is show in GUI.

Figure 4.1 GUI for Show the result Figure 4.2 (A)
In figure 4.2 (A) we can select all check box and enter the Question "who is prime minister of India ?" and click on search button. In Figure 4.2 (B) the 15 related Answer is shows in GUI. In console more then 400 Answer is generate but in this thesis we extract only top 15 Answer. The Random Indexing using circular Convolution is better because give the accurate Answer compare to other two algorithm.

Figure 4.2(B)

Figure 4.2(A),4.2(B) are Sample run of the question “Who is prime minister of India?

5.Result Analysis

After completion of implementation work, comparison between Term Term algorithm, Random Indexing and Random Indexing using circular Convolution. In Figure 5.3 compared three algorithm and which the proposed algorithm is accurate compared to other two.

Figure 5.3 Accuracy@n comparison of DSMs Model
In Figure 5.4 shows the comparison of Random indexing with addition and random indexing with circular convolution operators. In Figure 5.5 shows the comparison of Term term and random indexing with circular convolution operators.

6. CONCLUSION AND FUTURE WORK

Question answering system mainly consists three steps such as Question Processing, Document Processing and Answers Processing. Due to increased need of Question answering system in real world applications, it is necessary to build the QA system which not only give the answer as fast as possible but also gives the correct answer. In this thesis we studied different distributional semantic models like term-term matrix, Random Indexing and Random Indexing using circular Convolution. The Random Indexing using circular Convolution is better than other two algorithm.

My search module is implemented with Google Custom Search API which takes much time in retrieving the search results and if no internet connection then it is not work. so prepare the offline- local corpus, then this QA system can retrieve the answers fast enough to be used in real time applications.

In future, we can perform permutation operation on Random Indexing using circular convolution.

REFERENCES


