Journal of Theoretical and Applied Information Technology
15th April 2019. Vol.97. No 7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DEEP BELIEF NETWORK BASED QUESTION ANSWERING SYSTEM USING ALTERNATE SKIP-N GRAM MODEL AND NEGATIVE SAMPLING APPROACHES

[1]K YOGESWARA RAO, [2]GORTI SATYANARYANA MURTY, [3]T.PANDURANGA VITAL

[1]Associate professor, Aditya Institute of Technology and management, Department of Computer Science and engineering, Tekkali, Andhra Pradesh, India.
[2]Professor, Aditya Institute of Technology and management, Department of Computer Science and engineering, Tekkali, Andhra Pradesh, India.
[3]Associate professor, Aditya Institute of Technology and management, Department of Computer Science and engineering, Tekkali, Andhra Pradesh, India.
E-mail: [1]yogiindusiu@gmail.com, [2]gsn_73@yahoo.co.in,3vital2927@gmail.com

ABSTRACT

The Question Answering (QA) system becomes essential owing to the increasing amount of web content and high demand for the right and short information. Intending to enhance QA results towards the Natural Language Processing (NLP) community, most of the Question answering system exploits machine learning algorithms to generate an appropriate answer to the user query. Even though, it lacks to predict accurately over the large-scale data by itself, needs an external force to make adjustments in the answer prediction. With the recent evolution in deep learning, the neural network architecture reflects its potential for QA. The deep learning model can determine the issues in answer prediction on their own and resolves it. The class of deep neural network such as Deep Belief Network (DBN) is widely applied in question answering especially for text processing. Moreover, most of the works on the text processing exploits the skip-gram model for representing the relevant words in the vector over a massive volume of unstructured text data. However, it results in inefficient outcomes, especially when processing the combination of frequent data and stop words. To resolve these issues, this paper introduces the Deep Neural network for Answering user queries (DNA). The proposed DNA approach performs the QA system over DBN by applying alternate skip-N gram and negative sampling. The conditional probability measurement develops an alternate skip-N gram model and alternatively applying the normal N-gram and skip-N gram model. It improves the efficiency of relevant word-pair detection without increasing the computational complexity. By only using samples, the negative sampling reduces the impact of noise on the accuracy of alternate skip-N gram model and improves the efficiency of the QA system. Finally, the DNA is evaluated using Java and compared with the existing Unified model for Document-Based Question Answering (UDBQA). The results show the efficiency of DNA, for instance, the UDBQA approach reduces the F-measure by 16.3%, compared to the DNA approach with 2000 number of queries.

Keywords: *Deep Belief Network, Question Answering, Skip-N Gram Model, Negative Sampling, And Hidden Layers.*

1. INTRODUCTION

Recently, the research works on Big data processing have emerged owing to the dramatic growth of data sources on the web. On the other hand, the growth of online data sources also enables people all over the world to extract the required information from the web. The retrieval system such as search engine has done a great job of extracting the information from the web for the user in which the extracted information, probably in the form of structured, unstructured or links on web pages. While the user lacks to get the desired answer during questing. Among the set of retrieved documents, the users need to determine the relevant answer from the ranked list of documents which is a time-consuming process and there is a chance of selecting the extraneous information over the listed documents and links. To cope up with this constraint, the Question Answering (QA) system into play. The QA system is designed to answer the online users' query posted in natural language automatically. The QA system incorporates Information Retrieval (IR) and Natural Language

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Processing (NLP) techniques for answering the user posted query. Ordinarily, online users submit thousands of questions to retrieve the short answer from the Web source [1]. In order to answer a submitted query, a QA system exploits the number of documents or pre-structured database in the natural language format. The QA system seeks to deal with the various types of the query of users that including factoid based, definition based, reason-based, definition based, and explanation based types. The processing of big data improves efficiency in the retrieval of an appropriate answer to a user query. However, handling the variety of high volume data is the arduous task, which intends to degrade the performance of the question answering system [2]. The primary concern of the question answering system is to provide the most relevant result based on the natural language questions submitted by the user. In the question answering system, two approaches are present such as the manual and the automatic process. The manual question classification methods utilize the advantage of handmade rules for capturing the relevant answer over the document. Even though it helps to find the relevant answer in the document, it is more time consuming, arduous, and non-extensible in the environment. Hence, the machine learning technique has been widely employed as a powerful tool in extracting the hidden patterns from the heterogeneous data in an automatic manner [3].

In order to retrieve the best answer in the question answering system, most of the conventional work makes use of a machine learning algorithm for answer selection. Moreover, the conventional machine learning method significantly provides the most relevant answers to the users over the small-scale data. Despite, exploring the hidden features from the complex data representation is the quite challenging task [4]. In addition, in order to effectively provide the relevant answer, analyzing the hidden features is essential. Recently, the researchers have focused on deep learning techniques to resolve the issues of conventional machine learning technique in terms of handling the complex representation of large-scale data. The deep learning technique exploits the layered hierarchical procedure based on the rich features to retrieve the best solution for the user query [5]. Accordingly, deep learning methods are widely used in text processing, computer vision, speech recognition, and image recognition [6]. However, there is a challenge for deep learning based QA system for answering the factoid questions represented in the natural language.

The wealth of data available over the internet and the enormous expressions of natural language queries made it tedious for the deep learning to situate the relevant answers. In the natural language query processing, framing the appropriate question from the free form users' text for utilizing by an information retrieval engine is essential, but anyone mistake happens at the moment are susceptible to render the inefficient results. However, in such circumstances, where the reformulation of simple questions or keyword method will not satisfy, more advanced semantic, syntactic and contextual processing needs to be implemented to capture the answer. Thus, the proposed scheme enhances the deep learning approaches, such as Skip-N gram model and contrastive noise estimation over Deep Belief Network (DBN) whereas it discovers the desired answer concerning a user query with the help of Wordnet and web search engine. The proposed DNA has taken into the account of lexical, syntactic, and semantic structure for precise answer prediction. Further, the proposed scheme utilizes the advantages of DBN. The DBN is configured with additional features, such as 1) unlabeled data handling, 2) Bayesian probabilistic generative model, 3) efficient detection of hidden variables in the deepest layer, 4) Solving over the fitting problem.

1.1 Contributions

The contributions of the proposed Deep Neural network for Answering user queries (DNA) are as follows.

- The proposed DNA approach deeply analyzes the Big data source for retrieving an appropriate answer to the user query using multiple hidden layers over DBN.
- The information retrieval layer processes the user query, by extracting the query arguments and returns multiple relevant documents from a web search engine.
- The filtering layer processes the retrieved documents to filter the irrelevant word-pairs using alternate skip-N gram model and negative sampling.
- To improve the skip-N gram model without increasing the complexity and processing the irrelevant word-pairs, the conditional probability is utilized over DBN.
- Instead of requiring the data and the numerical probabilities of the noise distribution, the DNA implements the negative sampling with only samples and improves the efficiency of the question answering system.
- To show the superior performance of the DNA approach the experimental evaluation compares the proposed approach with the existing Unified Document-Based Question Answering (UDBQA).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2. RELATED WORK

With the demand for accurate results retrieval, the research works on a question answering system is emerging. The question answering system enhances the information retrieval process in the search engine and meets the user satisfaction level. The QA system often faces the difficulties during processing the queries, analyzing the context of the information source, extracting the relevant answers and finally formulating the answers. The advent of deep learning techniques in the question answering system helps to cope up with the constraints and to ensure precise results. The deep learning approaches face massive data in the detection of hidden patterns, even the handling of overwhelming data intends to a memory constraint. The work [7] employs the Recursive Neural Network (RNN) to reduce the issues of question answering based on task compositionality. It effectively retrieves the appropriate answer when the query comprises of the few numbers of individual words. However, the answer selection disregard of knowledge source such as Wikipedia leads to irrelevant information retrieval while answering the user query. By employing the deep similarity technique, the model automatically computes the best answer for the questions from the public question answering database. It computes the best solution based on the similarity score of the corresponding question and answer. Additionally, it applies the confidence measurement to improve the quality of answer selection [8]. The neural network [9] gives attention to the question in various perspectives to the precise question answering which exploits the global collection of knowledge base information to avert the out of vocabulary (OOV) issue. It seeks to represent the queries dynamically based on the various focuses of different candidate answer perspective. The research work [10] has developed the attentional, compositional model that comprises both the images and structured information sources which converts the different questions into the dynamic neural network to retrieve the answer. It has the ability to answer the questions from the structured and unstructured documents. The Conditional FOcused neural network-based question answering approach (CFO) combines the methods of deep recurrent neural networks and neural embedding to answer the factoid questions over the large-scale knowledge base. In addition, it employs the pruning method to mitigate the search space which helps to reduce the time-complexity of the system [11]. To deal with the natural language and the memory issues in question answering, the Dynamic Memory Network (DMN) approach enhances with the input fusion model [12] that involves two processes such as encoding of sentences and the identification of relationships between the sentences. The positional encoder enables the sentence encoding, and the bidirectional gated recurrent units (GRU) allow the identification of the interaction of sentences with the help of attention gates. In addition, it has the ability to retrieve the relevant answer even the queries comprise the text and the images.

The attentive pooling method [13] provides the most relevant answers based on the convolutional neural network (CNN) and the max-pooling method which helps to achieve the higher coverage of entities while answering the questions. Moreover, it introduces the entity linker for discovering the potential top-N entity candidates for the query. The convolution layer with the max-pooling helps to model the relation within the question pattern and the predicate more efficiently. Dual Attentive Neural Network Framework (DANN) [14] enhances the answer selection process with the help of the deep semantic matching and user community metadata representation. It employs the CNN for semantic matching of question entities with the attention. The cross attention model [15] precisely answer the question by focusing on the different aspects of answers based on the neural network. It exploits global KB information to avert the OOV issue. To enhance the quality of answer retrieval in the complex sentence, the system [16] employs the framework of the neural network and the deep case to identify the content which eases the faster the extraction of answer for the complicated questions. Also, it attempts to imitate the human brain information recalling system during answer selection, whereas the correlated words in the documents are coupled through the deep cases. In order to overcome the issue of information loss while selecting the answer based on the single view aspect, the Multi-View Fusion Neural Network (MVFNN) [17] has analyzed the inherent relation between the question and the answer in various aspects which facilitates the efficient answer selection based on methods of fusion RNN and the attention mechanism. However, it lacks to make use of the heterogeneous resources for the answer selection.

The method [18] integrates the local view of the answer with the global view of complete question and answers to accomplish the superior performance based on the newly developed attention mechanism using a neural network for answer selection. The framework of the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

convolutional neural network and the recurrent neural network [19] attempts to combine the semantic relation between the specific question and answer and the semantic correlation of answer sequence which assists in extracting the potential context among the answer sequences using the discovered matching patterns of question and answers. In order to resolve the issue of question answering based on the document, the model [20] exploits the information belongs to the title through the various human reading strategies to gain the valuable content of the document. The title information provides the generalized understanding of the document in which it makes use of LDA and LSA model for acquiring the topic of the document. However, It fails to reduce the unnecessary stop words which induce the complexity in the system.

2.1 Problem statement

Most research works have focused on the deep learning method on the question answering system to provide better results while dealing with the massive data. Despite, the processing of natural language queries and the continuous Bag Of Words (BOW) is the prominent issue in deep learning method while answering the questions. Most of the conventional work on deep learning learns the entire keywords from the retrieved documents. However, an increasing number of keywords does not contribute to improving the accuracy of question answering. Identifying the most relevant words to the query arguments is a challenging task. In addition, vector representation of most relevant keywords is an essential task in NLP. However, the vector representation schemes fail in differentiating the words with similar context, but different term vocabulary. Thus, it necessitates the distributed vector representations in a continuous space. The distributed representation of words over large corpora can capture the word meaning and context in the document. It is essential to represent the vectors in a distributed manner with precise syntactic and semantic word relationships. Another main problem in vector representation is noise contrastive estimation. Mostly, the noise contrastive estimation techniques use samples and the numerical probabilities of the noise distribution but meet memory challenges in question answering systems. Designing an efficient noise contrastive estimation using samples has to be focused in the future.

3. AN OVERVIEW OF QUERY ANSWERING SYSTEM

The proposed DNA divides the question-answering process into three layers, such as input, filtering, and output layer, which executes the query processing, information retrieval and processing, and answer processing respectively as shown in Figure.1.

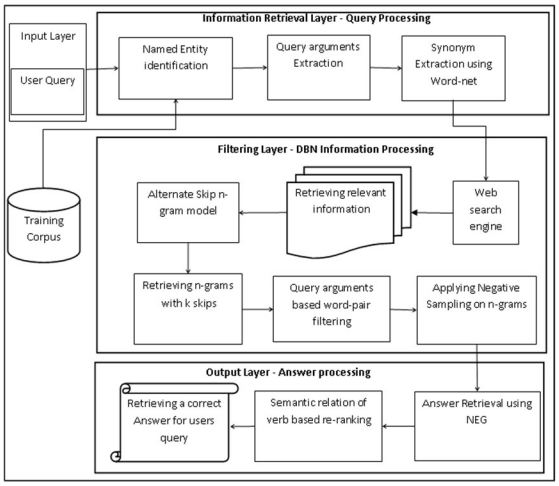

Figure 1: DNA Approach

Input and Information Retrieval Layer –Query processing: The first layer of DBN is the input, which inserts a user's query into DBN. The primary process in query processing is retrieving the relevant information against user query using the web search engine. To search in web engine, the DNA has to extract the query arguments and relevant words using Wordnet. The web search engine gives a vast number of documents, and thus, it is essential to filter out the documents that do not contain the answer. To do it, the DNA needs named entity recognition to map the entity types of the user query. With the aim of identifying the named entity type, the training corpus is employed in DNA. The DNA trains the corpus with different question-answer pairs for factoid questions and its rules.

Filtering Layer- DBN Information Processing: The filtering layer of the DNA approach exploits the web search engine for retrieving the relevant documents using query relevant words. In order to retrieve the n-grams for each sentence from the documents, the DBN exploits the alternate skip-gram model and negative sampling. It results in n-gram vector representation with k-skips and makes the deep learning process extremely efficient.

Output Layer- Answer Processing: The output layer consists of sentence filtering and sentence re-ranking process. The DNA approach filters the sentences using altered Skip-N gram and NEG sampling. In sentence filtering, the semantic relation between the query and retrieved sentences are used. The DNA approach generates the syntactic patterns for both filtered sentences and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

questions using POS tagger. Moreover, the DNA approach finds the named entity of a given query to map with the named entity of sentences. Finally, it predicts the correct answer from the filtered sentences.

3.1 Query Processing by using NLP

Mostly, the user queries are in the form of natural language, and the QA systems retrieve the relevant answer for a user query from an unstructured collection of natural language documents. Thus. The Natural Language Processing (NLP) is an essential tool for processing user's queries. Moreover, the Named Entity Recognition (NER) is the central component in the QA system. Once the user query has been processed, the DNA exploits the web search engine to find a set of documents containing the query arguments. After filtering the irrelevant sentences from the retrieved documents, the NER is used to find relevant "Person," "Location," and "Time" from the retrieved sentences using the named entity recognizer, such as "Who," "Where," "When," and so on. To achieve this objective, the training corpus is built on the DNA approach, where the question type is beneficial to detect the specific named entity of the answer. The DNA approach utilizes the NLP tools, such as a stop word remover and porter stemming to extract the essential arguments from the query, POS tagger to detect the named entity recognizer, and web search engine to retrieve the sentences containing the query arguments.

3.2 Deep Belief Network for Information Processing

The Deep Belief Network (DBN) is a multilayer network, which consists of many layers, such as input, filtering, and output layers. The DBN consists of the number of restricted boltzmann machines. Each boltzmann machine is a two-layer undirected graphical model, which consists of the visible and hidden layers. The visible layers are input and result. The hidden layer includes information retrieval and filtering layers. Initially, the boltzmann machine receives input from a user and executes the NLP and web search engine based information retrieval. In filtering layer, the boltzmann machine divides the document into several sentences D = {st1 , st2 , . . . stn}. The filtering layer utilizes the alternate Skip-N gram and negative sampling for two levels of sentence filtering in DNA. The alternate skip-N gram model results in log probability measurement for each sentence, and the negative sampling scheme improves the efficiency of log probability and reduces the noise from the filtered sentences. Thirdly, the output layer determines the correct answer to a user query using the named entity recognizer.

3.2.1 Alternate Skip-N gram Model for Vector Representation

The string similarity measurement is widely applied for question answering system. The N-gram similarity is the most commonly used string similarity measure. It slices the longer string of N characters and estimates the similarity value between the two words. A similarity value between a pair of words denotes the degree of relationship between them. However, the n-gram is assigned a randomly initialized vector, which affects the result accuracy. The n-gram model represents a consecutive sub-sequence of length n of words. A k-skip-n-gram represent a length-n subsequence with k distance from each other. For example, the input text is "deep learning is suitable for question answering". The set of 1-skip-2-grams provides all the bi-grams with the sub-sequences.

"Deep is, learning suitable, is for, suitable question, for answering".

The objective of the Skip-N gram model training is to detect the word representations and identify the surrounding words in a sentence. The Skip-gram model aims at improving the average Log Probability (LP). The DNA removes the word pair, when all the n-grams are stop words, excluding the named entities. It reduces the time complexity of Skip-N gram model.

$$LP = \sum_{t=1}^{T} \sum_{-n \le k \ge n, k \ne 0} \log p(w(n+k)|wn) \quad (1)$$

$$p\,(w(n+k) + |wn) = \exp\left[(w(n+k) \Big/ \sum_{t=1}^{T} \exp wn \right] \quad (2)$$

While applying equation (2) in (1), the LP is estimated. Where T is the total number of words in a sentence, n represents the size of a gram, and k represents the size of skipping. Larger n generates large training examples and thus can lead to higher accuracy. However, the skip-N gram model only considers the probability of the word in the position of n and 'n+k' individually, but not takes the conditional probability. It affects the accuracy in some cases. For example, in the sentence "The Times is a world-famous newspaper," the words of 'the' and 'Times' give meaning when they are represented with 0 skips. However, as per the skip-N gram model, they are split with non-zero skip value. Thus, the proposed DNA alters the operation of skip-N gram model, such as alternate Skip-N gram model. While applying the 'two gram with one skip,' the DNA results in 'The Times, The is, Times is, Times a, is a, is the world, a world, world newspaper, and famous newspaper.' Even though, it increases the number of word pairs; the DNA

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

plans to compensate for the complexity of more number of word pairs with conditional probability measurement. The DNA replaces the exponential measurement with conditional probability in log probability estimation. It reduces the number of word pairs along with the stop word removal. If both the words are equivalent in frequency, both are dependent with each other. In such a case, considering a single word frequency reduces the importance of frequency of word pair. The conditional probability measurement is applied in

DNA to solve this problem. Finally, equation (2) is updated using equation (3).

$$p\,(w(n+k)+|wn) = {}^{1}/_{T}\sum_{t=1}^{T}\sum_{-n\le k\ge n,k\neq 0}\log p\left({(w(n+k)|wn)}\Big/{[(p\,(w(n+k))(p(w))]}\right)$$

,,,,,(3)

3.2.2 Negative Sampling Based Sentence Relevant Score Measurement

The sentence filtering in DNA is to perform using the Skip-N gram, which is improved with the negative sampling. The DNA applies the negative sampling to approximately maximize the results of the log probability of the Skip-N gram model. The conditional probability term in the alternative Skip-N gram is improved in negative sampling. Thus, the task of negative sampling is to distinguish the target word w_t from the noise distribution $P_n(w_t)$ using logistic regression. Most of the conventional schemes exploit the noise contrastive estimation scheme. However, it requires both the data and the numerical probabilities of the noise distribution. To solve this issue, the DNA exploits negative sampling with only samples. The following equation results in Accurate LP (ALP).

$$ALP = {}^{1}/_{T}\sum_{t=1}^{T}\sum_{-n\le k\ge n,k\neq 0}\log p\left({(w(n+k)|wn)}\Big/{\sum w'\epsilon T\log(w'(n+k)|w'n)}\right)$$

----(4)

3.3 Answer Processing by using Named Entity Recognition

The final phase of the DNA system is to correctly extract the answer from a set of sentences using answer validation. In answer validation process, DNA approach exploits the named entity type in both the user's query and answer sentences.

3.3.1 Answer validation

The DNA approach retains the relevant answer sentences with high-rank value. The retrieved answer sentences are listed in descending order of LP value to sort the relevant answer sentences. Moreover, the list carries the most suitable answer in the top of the ranked list. To accurately answer a user's query, understanding the context of the user's query is essential. In answer validation, the named entity type of the selected answer sentences is identified. The accuracy of the answer validation process is an essential factor in DNA to avoid plausible answers in the question answering system. The DNA system enables the answer validation process to derive the named entity type from the sentences. If the retrieved sentences are not named entity with a type compatible with the user's query, the retrieved answer sentence is discarded. Using such a way the DNA system provides the topmost answer against a query to the user.

//DNA Question Answering System//

Input: Retrieved query arguments

Output: Filtered answer sentences

// Information Collection Processing

for all query arguments Є user's query do

{

Retrieve most relevant documents, RD from web search engine

}

for each relevant document rd Є RD do

{

Extracts Answer sentences (ASen) from each document;

}

// Filtering Layer

For each sentence Є ASen do {

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Applies alternate skip-n gram model;

Extracts Word-Pair from ASen;

for each WP do {

If WP has stop words

Remove the WP form Asen List;

end if

Measures accurate ALP for each WP Є ASen using negative sampling;

}

//Answer Processing using Named Entity Recognition

WP List = {ALP};

for each sentence ASen do {

for (i=1; i< |WP|ASen; i++) {

for (j=1; j< |WP|ASen; j++) {

If (number [j]>number [i]) {

int temp = number [j] ;

number [j] = number [i];

number [i] = temp;

}

}

} }

Retrieve answer from top most ASen using named entity;

End

Figure 2: Algorithm for Question Answering System

4. EXPERIMENTAL EVALUATION

The experimental evaluation analyzes the accuracy of the DNA system and compares it with the existing UDBQA approach [20].

4.1 Experimental setup

The DNA approach is implemented using the Java platform. The user queries are in the form of NLP. Hence, the proposed scheme utilizes the WordNet ontology to provide a semantic relation of each word in a query. The implementation of the proposed approach depends on the Java API for WordNet Searching (JAWS) that provides the interface for retrieving the information from the WordNet database. Moreover, the Scribd web search existing UDBQA approach engine is employed as the information retrieval engine. The Scribd web search engine provides the question of relevant information that includes documents and blogs.

4.1.1 Dataset

The DNA approach collects the questions from WikiQA corpus that comprises 3,047 questions. The WikiQA provides a label to the 1,473 sentences as answer sentences to the corresponding queries ion the dataset. Among the 3,047 queries, the DNA approach has taken into account of 2000 queries as a training set for generating training corpus. The WikiQA utilizes the Bing query logs as the query source to mimic the true user queries. It connects each query with a specified sentence in Wikipedia page, which ha relevant answer to the user's query. With the help of crowdsourcing, the WikiQA consists of 29,258 sentences in the dataset.

4.1.2 Evaluation metrics

Recall-Oriented Understudy for Gisting Evaluation (ROGUE): It refers to the ratio of overlapped N-grams between the candidate answer and reference answer.

$$RLCS = {LCS\ (c, r)}/{|r|}$$

$$PLCS = {LCS\ (c, r)}/{|c|}$$

The ROUGE-L is the Longest Common Subsequence (LCS) based statistics. RLS is the ratio of LCS length to reference answer length, namely, recall. Moreover, the PLCS is the ratio of LCS length to candidate answer length, namely precision. Where c is the candidate answers, and r is the reference answer. Substituting the RPLCS and PLCS in the following equation, the ROUGUE is measured.

$$ROUGUE = {[(1 + Y2)\ *\ RLCS\ *\ PLCS]}/{RLCS\ +\ Y2PLCS}$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Bi-Lingual Evaluation Understudy (BLEU): It estimates the quality of the text, which is translated between the machine and the natural language. Where n represents the number of queries.

$$\mathrm{BLEU} = e^{\min(1-r/c,0)}\left(\prod_{i=1}^{n}\left(\mathrm{PLCS}^{1/n}\right)\right)$$

F-measure: F-measure is the discrepancy and balance between precision and recall.

$$\mathrm{F-Measure} = \frac{\mathrm{PLCS} * \mathrm{RLCS}}{\mathrm{PLCS} + \mathrm{RLCS}}$$

4.2 Experimental results

4.2.1 ROGUE

Figure 2 shows the comparative result of DNA and UDBQA approaches by varying user queries from 1000 to 3000. From the figure, it is observed that the ROGUE value of both the DNA and the UDBQA declines when the number of queries increases because the softmax layer in UDBQA applies the exponential calculation, but not considering the importance of both the words in a particular sentence. It leads the UDBQA to leave some essential word pairs in a document, resulting in reduced ROGUE value. The proposed schemes always attain better results, compared to the UDBQA. The precision and recall value of DNA approach is higher than the UDBQA approach, due to the consideration of lexical, syntactic, and semantic structure based accurate candidate answer prediction. However, the reduction of UDBQA recall value shows a decrement in the ROGUE value, compared to the proposed DNA approach. The ROGUE value of DNA approach decreases by 5.5%, but the UDBQA approach decreases by 31.3% when the number of queries increases from 1000 to 3000. At the point of 2000 queries, the DNA approach improves the ROGUE value by 25.4% more than that of the UDBQA.

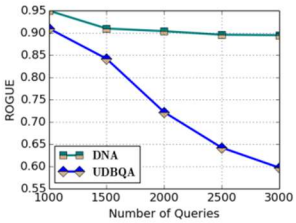

Figure 2: Number of Queries vs. ROGUE

4.2.2 BLEU

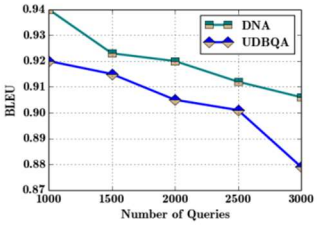

Figure 3: Number of Queries Vs BLEU

The BLEU of DNA approach is shown in Figure 3. It shows the impact of variation of the number of queries from 1000 to 3000 on BLEU of both the DNA and UDBQA. The BLEU value of both the question answering systems linearly declines when increasing the number of queries. However, BLEU value of the DNA approach shows moderate variation when comparing with the UDBQA variation in BLEU. The DNA processes the user's query on the basis of alternate skip-N gram model, which is improved by the conditional probability measurement and alternatively applying the normal N-gram and skip-N gram model. Initially, the query terms are processed and enlarged using the Word-net, which facilitates the DNA to search relevant sentences in the retrieved documents. The existing approach does not consider the presence of essential keywords along with the stop word. As a result, the performance of BLEU in UDBQA approach slightly decreases more than that of the DNA approach. With 2000 number of queries, the UDBQA approach lags by 1.1%, compared to the DNA approach, due to the absence of stop word removal and semantic relationship between the query arguments and answer documents.

4.2.3 F-Measure

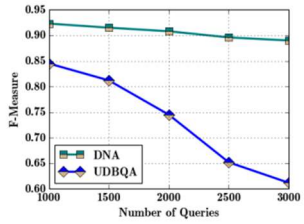

Figure 4: Number of Queries vs. F-Measure

Figure 4 shows the F-measure of DNA and UDBQA approaches, by varying the number of queries from 1000 to 3000. The F-measure value declines while increasing the number of questions in the QA system. The DNA approach improves the performance even while processing more number of user queries, due to the alternate skip-N gram and negative sampling. The negative sampling

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

scheme only requires the samples. However, the noise contrastive estimation scheme requires both the data and the mathematical probabilities of the noise distribution. Thus, the negative sampling and alternate Skip-N gram model reduces the computation complexity, while maintaining the QA accuracy. The irrelevant word pair removal and provision of higher preference to the word-pair, when both the words are essential to improve the efficiency of DNA for both the PLCS and RLCS. It is the primary reason behind the improvement in F-measure of DNA, compared to the UDBQA. The F-Measure value of the DNA approach decreases by 5.5%, but the UDBQA approach decreases BLEU by 3.3% when the number of queries increases from 1000 to 3000. At the point of 2000 queries, the DNA approach improves the ROGUE value by 16.3% more than that of the UDBQA.

5. PERFORMANCECOMPARISON BETWEEN THE PROPOSED METHOD AND THE EXISTING METHOD

Scenario under various queries	DNA method	UDBQA method	Justification
ROGUE	0.95-0.895	0.91-0.597	DNA considers lexical, syntactic, and semantic structure for answer prediction.
BLEU	0.94-0.906	0.92-0.879	Due to the alternate skip-N gram model, which is improved by conditional probability measure
F-measure	0.923-0.89	0.845-0.612	Due to the alternate skip-N gram and negative sampling.

Table:1 The performance evaluation of proposed DNA and existing UDBQA method

6. CONCLUSION

With the exponential raise of availability of information in the internet, the retrieval of desired result with respect to the users' query is crucial. In order to capture the intended information regarding the users' query in the best way, this paper discussed a DNA question answering approach that is based on the schemes of Alternate Skip N-gram model and negative sampling. The DBN based DNA approach identifies a precise answer to a corresponding user query with the support of WorldNet and web search engine. Moreover, by exploiting negative sampling method, the DNA generates the semantically equivalent answers to avoid the plausible answers. Finally, the DNA validates the answer sentences using named entity type and provides a most appropriate answer to a user query. The experimental evaluation shows that the DNA outperforms the UDBQA approach. For instance, the UDBQA approach lags in BLEU by 1.1%, compared to the DNA approach for 2000 queries.

6.1 Open Research Issues

- The most prominent challenge is the lexical gap and the ambiguity which occurs due to the deviation between the natural language questions and the semantically structured data on the knowledge base
- The important challenge for question answering systems using knowledge base was the question entity recognition, notably in questions concerning numerous entities.
- In addition, users have various native languages. Thus, the QA system is intended to identify a specific language and bring the results. The multilingual QA systems still entail most effective algorithm for dealing with the heterogeneous multilingual collection of data.
- The QA systems seek to read and comprehend an entire knowledge base or Wikipedia article that perhaps comprises the key solution to the question which results in computational complexity.

REFERENCES:

[1] Bouziane, Abdelghani, Djelloul Bouchiha, Noureddine Doumi, and Mimoun Malki, "Question answering systems: survey and trends", Procedia Computer Science, Vol.73, pp.366-375, 2015.

[2] Chandio, Aftab Ahmed, Nikos Tziritas, and Cheng-Zhong Xu, "Big-data processing techniques and their challenges in transport domain", ZTE Communications, Vol.1, No.010, 2015.

[3] Gharehchopogh, Farhad Soleimanian, and Yaghoub Lotfi, "Machine learning based question classification methods in the question answering systems", International Journal of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Innovation and Applied Studies, Vol.4, No.2, pp.264-273, 2013.

[4] Dwivedi, Sanjay K., and Vaishali Singh, "Research and reviews in question answering system", Procedia Technology, Vol.10, pp.417-424, 2013.

[5] Sharma, Yashvardhan, and Sahil Gupta, "Deep Learning Approaches for Question Answering System." Procedia Computer Science 132 (2018): 785-794.

[6] Deng, Li, and Dong Yu. "Deep learning: methods and applications." Foundations and Trends® in Signal Processing7, no. 3–4 (2014): 197-387.

[7] Iyyer, Mohit, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III, "A neural network for factoid question answering over paragraphs", In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 633-644, 2014.

[8] Minaee, Shervin, and Zhu Liu, "Automatic question-answering using a deep similarity neural network", arXiv preprint arXiv:1708.01713, 2017.

[9] Zhang, Yuanzhe, Kang Liu, Shizhu He, Guoliang Ji, Zhanyi Liu, Hua Wu, and Jun Zhao, "Question answering over knowledge base with neural attention combining global knowledge information", arXiv preprint arXiv:1606.00979, 2016.

[10] Andreas, Jacob, Marcus Rohrbach, Trevor Darrell, and Dan Klein, "Learning to compose neural networks for question answering", arXiv preprint arXiv:1601.01705, 2016.

[11] Dai, Zihang, Lei Li, and Wei Xu, "Cfo: Conditional focused neural question answering with large-scale knowledge bases", arXiv preprint arXiv:1606.01994, 2016.

[12] Xiong, Caiming, Stephen Merity, and Richard Socher, "Dynamic memory networks for visual and textual question answering", In International conference on machine learning, pp.2397-2406, 2016.

[13] Yin, Wenpeng, Mo Yu, Bing Xiang, Bowen Zhou, and Hinrich Schütze, "Simple question answering by attentive convolutional neural network", arXiv preprint arXiv:1606.03391, 2016.

[14] Liu, Zhiqiang, Mengzhang Li, Tianyu Bai, Rui Yan, and Yan Zhang, "A Dual Attentive Neural Network Framework with Community Metadata for Answer Selection", In proceedings of Springer in National CCF Conference on Natural Language Processing and Chinese Computing, pp.88-100, 2017.

[15] Hao, Yanchao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao, "An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge", In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol.1, pp. 221-231, 2017.

[16] Ansari, Ahlam, Moonish Maknojia, and Altamash Shaikh, "Intelligent question answering system based on artificial neural network', In IEEE International Conference on Engineering and Technology (ICETECH), pp.758-763, 2016.

[17] Sha, Lei, Xiaodong Zhang, Feng Qian, Baobao Chang, and Zhifang Sui, "A Multi-View Fusion Neural Network for Answer Selection", In AAAI, 2018.

[18] Bachrach, Yoram, Andrej Zukov-Gregoric, Sam Coope, Ed Tovell, Bogdan Maksak, Jose Rodriguez, Conan McMurtie, and Mahyar Bordbar, "An Attention Mechanism for Neural Answer Selection Using a Combined Global and Local View", In proceedings on IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), pp.425-432, 2017.

[19] Zhou, Xiaoqiang, Baotian Hu, Qingcai Chen, and Xiaolong Wang, "Recurrent convolutional neural network for answer selection in community question answering", Neurocomputing, Vol.274, pp. 8-18, 2018.

[20] Li, Weikang, Wei Li, and Yunfang Wu, "A Unified Model for Document-Based Question Answering Based on Human-Like Reading Strategy", The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), pp.604-611, 2018.