Answer Sentence Ranking of Factoid Question

Nitish Kumar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY

MIZORAM-796012, INDIA

NOVEMBER - 2017

Answer Sentence Ranking of Factoid Question

Report submitted to

National Institute of Technology, Mizoram

Of

Bachelor of Technology

By

Nitish Kumar (Enrolment No: BT14CS007)

Supervisor/ Supervisors

Mr. Goutam Majumder

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY

MIZORAM-796012, INDIA

NOVEMBER - 2017

APPROVAL SHEET

This report entitled **Answer Sentence Ranking of Factoid Question** by **Nitish Kumar** is approved for the degree of **Bachelor of Technology** in **Computer Science & Engineering**.

Supervisor (s)

____

____

____

Head of Department

____

____

____

Date: _____________

Place: _____________

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

____

(Signature)

____

(Name of the student)

____

(Roll No.)

Date: __________

CERTIFICATE

It is certified that the work contained in the report titled "**Answer Sentence Ranking of Factoid Question**," by "**Nitish Kumar**," has been carried out under my/our supervision for 7th Semester **Project Phase-I**.

Signature of Head of Dept.
Name
Department
NIT Mizoram
Month, Year

Signature of Supervisor(s)
Name(s)
Department(s)
NIT Mizoram
Month, Year

ACKNOWLEDGEMENT

I am profoundly grateful to **Mr. Goutam Majumde**r for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

I would like to express deepest appreciation towards **Prof. Rajat Gupta** Director NIT Mizoram **Dr. Partha Pakray** Head of Department of Computer Science Engineering, Whose invaluable guidance supported us in completing this project.

At last I must express our sincere heartfelt gratitude to all the staff members of Computer Engineering Department who helped me directly or indirectly during this Course of work.

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

IR:	Information Retrieval
IE:	Information Extraction
NLP:	Natural Language Processing
FAQ:	Frequently Asked Question
QA:	Question and Answer
WWW:	World Wide Web
SQuAD:	Stanford Question Answering Dataset
MC:	Machine Comprehension
RNN:	Recurrent Neural Network

ABSTRACT

In this Project, I've purpose R-NET method. It is end-to-end neural networks model reading comprehension style question answering, which aims to answer questions from a given passage. I first match the question and passage with gated attention-based recurrent networks to obtain the question-aware passage representation. Then I propose a self-matching attention mechanism to refine the representation by matching the passage against itself, which effectively encodes information from the whole passage. Then finally employ the pointer networks to locate the positions of answers from the passages.

CONTENTS

Chapter 1
Introduction

Question answering is a specialized area in the field of Information Retrieval (IR), Information Extraction (IE) and Natural Language Processing (NLP). In the current scenario, information retrieval systems or search engine are able to display relevant information and list of web pages containing that document according to their rank as per the search engine. [1], [2]. Hence, the main objective of all question-answer system is to retrieve the answer to a question rather than full documents.

However, in recent, there are mainly two types of a question, Factoid, and Non-factoid. The main type of question is submitted by the user in natural language, which is a factoid question ("A question for which a complete answer can be given in 50 bytes or less, which are roughly a few words" According to Radu Soricut, Eric Brill, and his team member). Another definition of factoid question, Question that can ask frequently (FAQ). These questions are sometimes called factoid questions. An example has been listed below:

Q: Who is the president of India?

Q: What is India's men national costume?

A Non-factoid question is beyond factoid question. It can be anything. Answering of non-factoid question is even more complex than factoid question. You can be asked to provide an answer to a math problem, to explain how to fix a specific model of a car, and so on, so forth. An important sub-problem of non-factoid QA consists in finding already existing answers posted on community QA sites such as Quora! This is very much an IR task. An example has been listed below:

Q. How does a film 3-idiots qualify for a National Award?

(Above question requires an answer along the following lines: A feature film must screen in an Indian country theater in 35 or 70mm or in 24-frame progressive scan digital format suitable for exhibiting in existing commercial digital camera site for paid admission for seven consecutive days. The seven-day run must begin before midnight, November 15, of the qualifying year [2017]).

But the recent research trend is shifting toward the various types of questions.

(1) List Type Questions: An example has been listed below.

Q. List the country that who won the world cup cricket?

(2) Why Type Questions: An example has been listed below.

Q. Why Mahatma Gandhi was killed

Above, we see there are various types of question, but mainly I am concentrating on factoid Type of questions. The answering of a factoid question is not an easy task. The challenges posed by answering factoid question have been addressed using a large variety of methods, such as [3] question parsing, [4] question-type determination, [3] WordNet exploitation, [3]Web exploitation, [4] noisy-channel transformations, [3]semantic analysis and [3] inferencing.

The question answering (QA) task has received a great deal of attention from the computational linguistic research community in the last few years, for example, Text Retrieval Conference (TREC) [5]. It is a conference series co-sponsored by NIST, initiated the Question-Answering Track in 1999 which tested system ability to retrieve short text snippets in response to factoid question [5]. Following the success of (TREC), in 2002 the workshops of the Cross-Language Evaluation Forum (CLEF) and NII Test Collection for IR System (NTCIR) started multilingual and cross-lingual QA tracks, focusing on European and Asian languages respectively [6].

Question answer systems are classified into two main categories. (1). namely open-domain QA system and (2). Closed-domain QA systems. Open-domain question answering deals with a question about nearly everything and can only rely on universal ontology and information such as World Wide Web (WWW). On the other hand, closed-domain question answering deals with question under the specific domain (music, weather etc.) The domain-specific QA system involves heavy use of natural language processing systems formalized by building a domain-specific ontology [7].

In this Project, I've purpose R-NET method. It is end-to-end neural networks model reading comprehension style question answering, which aims to answer questions from a given passage. I am working on SQuAD dataset for developing system. More detail of the method see in the next chapter.

1.1 MOTIVATION

Data is swelling up with every passing second. Thus, the demand to arrange it. Ordering of the Data will help readability and accessibility, thus will save our valuable time. This idea motivated me to explore this field. I especially aimed to research in Fact-based Dataset since I found it most relevant to our daily life. My project is a little step towards this. My project consists of Fact-based questions which can be answered in just a few words. This will help many to save their time and will provide ease of search. For better understanding, I have gone through Squad Dataset thoroughly which helped me to understand and implement my Project. I learnt the basics of arranging a Dataset so as to find out valuable information out of that. But, I think there is still scope of improvement in this research field, which I will try to discover in coming time. For instance, Reading multiple reviews of a product at any online store might turn out to be hectic task as it will take lot of time, Moreover drawing conclusions from those reviews can turn out to be a tough call. My project can turn out to be a lot more worthy in this case, which can generate an overall conclusion within few seconds and that too accurately!

Chapter 2
Literature Review

Question answering is a part of a computer discipline in which we perform IR and IE. Generally we have many types of questions. But I'm concentrating only on factoid type of questions. Lots of work on question answering system has been already done. such as R-net (Machine Reading Comprehension with self-matching networks), Reinforced Mnemonic Reader for Machine Comprehension, MEMEN (Multi-layer Embedding with Memory Networks For Machine Comprehension), Reasonet (Learning to stop Reading in Machine Comprehension), Structural Embedding of Synthetic Trees for Machine Comprehension. Their working is discussed in the below paragraph.

Firstly, I extracted the data for my project from Microsoft research Asian group. They introduce R-NET model. It is end-to-end neural networks model for reading comprehension question answering. They match the question and passage based on the recurrent network model. They finally employ the pointer networks to locate the position of answer from the passage. They perform an experiment on the SQuAD [8] and Microsoft Machine Reading Comprehension (MS-MACRO) dataset and archive the best result on both datasets among all public result.

R-NET model consist of four parts:

(1) The Recurrent Network Encoder to Build Representation for Question and Passages Separately.

(2) The gated matching layer to match the question and passage.

(3) The self-matching layer to aggregate information from the whole passage.

(4) The pointer-network based answer boundary prediction layer.

By using R-NET model, they achieved 72.3% exact match accuracy on the hidden SQuAD test-set [8]. While the ensemble model further boosts the result to 78.92%, which currently holds the first place on the SQuAD

leader board [8]. Besides, their model also archives the best-published result on the MS-MACRO dataset.
After I completed the observance of R-NET model, I started gathering information from another method proposed (Reinforced Mnemonic Reader for Machine Comprehension). The reinforced Mnemonic reader for machine comprehension (MC) task, which aims to answer a query about a given context document. They overviewed several novel mechanisms that address a critical problem in MC that is not adequately solved by previous work, such that enhancing the capacity of an encoder, modeling long-term dependencies of contexts, refining the predicted answer span, and directly optimizing the evaluation matric. Extensive experiments on TriviaQA and Stanford question answering (SQuAD) [8] show that their model achieves state-of-the-art results.

They tell about that all recent works are the use of an "encoder-interaction-pointer" framework. In such a framework, word sequence of both query and context are projected into distributed represented and encoded by recurrent neural networks. The attention mechanism is then used to model the complex interaction between the query and the context. Finally, a pointer network is used to predict the boundary of the answer. They use SQuAD [8] dataset model and achieve 77.61% exact match accuracy.

Now another method was developed by Boyuan Pan and his team, name of this method is novel neural network architecture called Multi-layer embedding with memory networks (MEMEN) for machine comprehension. In the encoding layer, we employ the classical skip-gram model to the syntactic and semantic information of the word to train a new kind of embedding layer. They also proposed a memory network of full-orientation matching of the query and passage to catch more pivotal information. The experiment shows that their model competitive results both from the perspectives of precision and efficiency in Stanford question answering dataset (SQuAD) [8] among all published results and achieves the-state-of-the-art results on TrivaQA dataset.

They introduced the (MEMEN), an end-to-end neural network for machine comprehension task. Their model consists of three parts:

(1) The encoding of context and query, in which they add useful syntactic-semantic information in the embedding of every word,

(2) The high-efficiency multilayer memory network of full orientation matching to match the question and context,

(3) The pointer network based answer boundary prediction layer to get the location of the answer in the passage.

Their machine reading model consists of three parts. First, they concatenate several layers of embedding of question and contexts and pass them into a bidirectional RNN (Mikolov et al, 2010). Then they obtain the relationship between query and context through a novel full-orientation matching and apply memory networks in order to deeply understand. In the end, output layer help locate the answer in the passage. They launched their method yields competitive results on the large machine comprehension benchmarks SQuAD [8] and the state-of-the-art result on TriviaQA dataset. On SQuAD, [8] their model archives 75.37% exact match and 82.66 F1 scores.

Heading ahead for my research I get to know about another method proposed by Yelong Shen and his team. They told about that teaching a computer to read and answer general question per-training to a document is the challenging yet unsolved problem. They describe a novel neural network architecture called the reasoning network (ReasoNet) for machine comprehension tasks. ReasoNet makes use of multiple turns to effectively exploit and then reason over the relation among queries, documents, and answers. Different from previous approaches using a fixed number of turns during interface, ReasoNets introduced a termination state to relax this constraint on the reasoning depth. ReasoNets achieve superior performance in machine comprehension dataset, including unstructured CNN and Daily mail datasets, the Stanford SQuAD [8] dataset, and a structured graph Reachability dataset.

They examine the performance of reasons on CNN, Daily Mail Dataset and SQuAD [8]. They calculate the different type of constraints. The constraints are: 1) vocal size 2) Embedding layer 3) Bi-GRU Encoder 4) Memory and Attention. After that, they perform an operation on different types of datasets and achieved results based on the datasets. For instance they apply ReasoNets method on CNN and Daily Mail dataset, and achieves 72.9% (valid), 74.7% (test) and 77.6% (valid), 76.6% (test) respectively. And then apply ReasoNets method on SQuAD [8] and achieve 69.1% (exact match), 78.9% (F1 score) for single model and 73.4% (exact match), 81.8% (F1 score) for ensemble model.

The last but not least, I went through was “Name of the method is the structural embedding of syntactic trees (SEST)”, an algorithm framework to utilize structured information and encode them into vector representations that can boost the performance of algorithms for the machine comprehension. They evaluate our approach using a state-of-the-art neural attention model on the Squad [8] dataset.
They apply deep learning based methods demonstrated great potential for question answering, none they take syntactic information of the sentences such as constituency tree and dependency tree into consideration. They told about that such type of techniques have been proven too useful in many natural language understanding tasks in the past and illustrated noticeable improvements such as work by [8].

The constituency tree of a sentence defines the internal node and terminal nodes to represent phrase structure grammars and the actual words. By using the constituency tree, we can reduce the size of the candidate space and the help the algorithm to identify the correct answer. On the other hand, a dependency tree is constructed based on the dependency structure of a sentence.

They detail the procedure of two alternative improvements of their methods the structural Embedding of constituency tree model (SECT) and the structural embedding of dependency model (SEDT). They assume that the syntactic information has already been generated in the pre-processing step using tools such as the Stanford CoreNLP.

Rank	Model	Exact match	F1 (score)
1	R-net	78.926	85.722
2	Reinforced Mnemonic reader	77.678	84.888
3	MEMEN	75.370	82.658
4	ReasoNet	75.034	82.552
5	SEDT+BiDAF	73.723	81.530

Figure 2.1: comparison between several methods proposed

Chapter 3
Dataset Description

I used SQuAD [8] dataset for development of this system. It is a set of Wikipedia articles, where the answer to every question is a segment of text and spans, from the corresponding passage. SQuAD [8] contains 107,785 question-answer pairs of 536 articles.
In contrast to prior datasets, SQuAD [8] does not provide a list of answer choices for each question. Rather, systems must select the answer from all possible spans in the passage, thus needing to cope with a fairly large number of candidates. While questions with span-based answers are more constrained than the more interpretative questions found in more advanced standardized tests, they still find a rich diversity of questions and answer types in SQuAD [8]. They develop automatic techniques based on distances in dependency trees to quantify this diversity and stratify the questions by difficulty. The span constraint also comes with the important benefit that span-based answers are easier to evaluate than free-form answers.

3.1 Dataset collection:

I collect their dataset in three stages: curating passages, crowdsourcing question- answers on those passages, and obtaining additional answers.

3.1.1 Passage curation: To retrieve high-quality articles, they used Project Nayuki's Wikipedia's internal PageRank to obtain the top 10000 articles of English Wikipedia, from which we sampled 536 articles uniformly at random. From each of these articles, we extracted individual paragraphs, stripping away images, figures, tables, and discarding paragraphs shorter than 500 characters. The result was 23,215 paragraphs for the 536 articles covering a wide range of topics, from musical celebrities to abstract concepts. I partitioned the articles randomly into a training set (80%), a development set (10%) and a test set (10%).

3.1.2 Question-answer collection. Next, they employed crowd workers to create questions. They used the Daemon platform with Amazon Mechanical Turk as its backend. Crowd workers were required to have a 97% HIT acceptance rate, a minimum of 1000 HITs, and be located in the United States or Canada. Workers were asked to spend 4 minutes on every paragraph and paid $9 per hour for the number of hours required to complete the article. The task was reviewed favorably by crowd workers, receiving positive comments on Turkopticon. On each paragraph, crowd workers were tasked with asking and answering up to 5 questions on the content of that paragraph. The questions had to be entered in a text field, and the answers had to be highlighted in the paragraph. To guide the workers, tasks contained a sample paragraph and examples of good and bad questions and answers on that paragraph along with the reasons they were categorized as such. Additionally, crowd workers were encouraged to ask questions in their own words, without copying word phrases from the paragraph. On the interface, this was reinforced by a reminder prompt at the beginning of every paragraph, and by disabling copy-paste functionality on the paragraph text.

3.1.3 Additional answers collection. To get an indication of human performance on SQuAD [8] and to make our evaluation more robust, we obtained at least 2 additional answers for each question in the development and test sets. In the secondary answer generation task, each crowd worker was shown only the questions along with the paragraphs of an article and asked to select the shortest span in the paragraph that answered the question. If a question was not answerable by a span in the paragraph, workers were asked to submit the question without marking an answer. Workers were recommended a speed of 5 questions for 2 minutes and paid at the same rate of $9 per hour for the number of hours required for the entire article. Over the development and test sets, 2.6% of questions were marked unanswerable by at least one of the additional crowd workers. The dataset Description of SQuAD is presented in Figure 3.1.

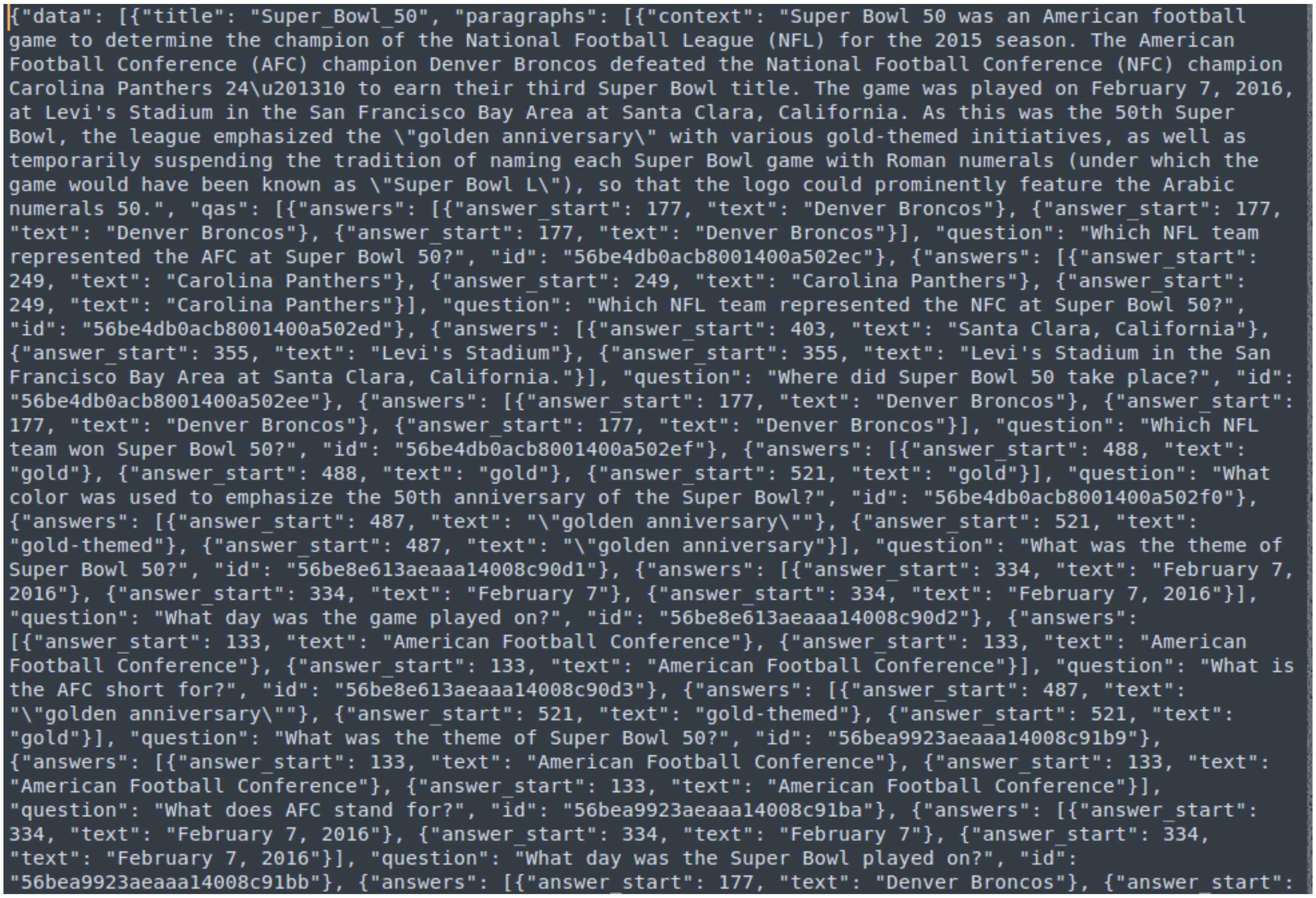
{"data": [{"title": "Super_Bowl_50", "paragraphs": [{"context": "Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24\u201310 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the \"golden anniversary\" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as \"Super Bowl L\"), so that the logo could prominently feature the Arabic numerals 50.", "qas": [{"answers": [{"answer_start": 177, "text": "Denver Broncos"}, {"answer_start": 177, "text": "Denver Broncos"}, {"answer_start": 177, "text": "Denver Broncos"}], "question": "Which NFL team represented the AFC at Super Bowl 50?", "id": "56be4db0acb8001400a502ec"}, {"answers": [{"answer_start": 249, "text": "Carolina Panthers"}, {"answer_start": 249, "text": "Carolina Panthers"}, {"answer_start": 249, "text": "Carolina Panthers"}], "question": "Which NFL team represented the NFC at Super Bowl 50?", "id": "56be4db0acb8001400a502ed"}, {"answers": [{"answer_start": 403, "text": "Santa Clara, California"}, {"answer_start": 355, "text": "Levi's Stadium"}, {"answer_start": 355, "text": "Levi's Stadium in the San Francisco Bay Area at Santa Clara, California."}], "question": "Where did Super Bowl 50 take place?", "id": "56be4db0acb8001400a502ee"}, {"answers": [{"answer_start": 177, "text": "Denver Broncos"}, {"answer_start": 177, "text": "Denver Broncos"}, {"answer_start": 177, "text": "Denver Broncos"}], "question": "Which NFL team won Super Bowl 50?", "id": "56be4db0acb8001400a502ef"}, {"answers": [{"answer_start": 488, "text": "gold"}, {"answer_start": 488, "text": "gold"}, {"answer_start": 521, "text": "gold"}], "question": "What color was used to emphasize the 50th anniversary of the Super Bowl?", "id": "56be4db0acb8001400a502f0"}, {"answers": [{"answer_start": 487, "text": "\"golden anniversary\""}, {"answer_start": 521, "text": "gold-themed"}, {"answer_start": 487, "text": "\"golden anniversary"}], "question": "What was the theme of Super Bowl 50?", "id": "56be8e613aeaaa14008c90d1"}, {"answers": [{"answer_start": 334, "text": "February 7, 2016"}, {"answer_start": 334, "text": "February 7"}, {"answer_start": 334, "text": "February 7, 2016"}], "question": "What day was the game played on?", "id": "56be8e613aeaaa14008c90d2"}, {"answers": [{"answer_start": 133, "text": "American Football Conference"}, {"answer_start": 133, "text": "American Football Conference"}, {"answer_start": 133, "text": "American Football Conference"}], "question": "What is the AFC short for?", "id": "56be8e613aeaaa14008c90d3"}, {"answers": [{"answer_start": 487, "text": "\"golden anniversary\""}, {"answer_start": 521, "text": "gold-themed"}, {"answer_start": 521, "text": "gold"}], "question": "What was the theme of Super Bowl 50?", "id": "56bea9923aeaaa14008c91b9"}, {"answers": [{"answer_start": 133, "text": "American Football Conference"}, {"answer_start": 133, "text": "American Football Conference"}, {"answer_start": 133, "text": "American Football Conference"}], "question": "What does AFC stand for?", "id": "56bea9923aeaaa14008c91ba"}, {"answers": [{"answer_start": 334, "text": "February 7, 2016"}, {"answer_start": 334, "text": "February 7"}, {"answer_start": 334, "text": "February 7, 2016"}], "question": "What day was the Super Bowl played on?", "id": "56bea9923aeaaa14008c91bb"}, {"answers": [{"answer_start": 177, "text": "Denver Broncos"}, {"answer_start":

Figure 3.1 Data-set Descriptions

3.2 Dataset Prepossessing

Dataset consisted of Question and their multiple Answers which were jumbled up. A systematic system should contain Questions and Answers in an orderly fashion. So, From the Dataset passages, Question with their corresponding Answer were filtered out. Therefore, Dataset is prepossessed to obtain a paragraph which consisted of Questions and Answers.

3.2.1 Passage Prepossessing

The Dataset were a combination of multiple Passages. A Passage should be convenient to read and understandable by any user. Therefore these Passages were extracted one by one. This will provide flexibility for user to ask any Question whose corresponding Answer(s) will be fetched from the Passage. A sample Passage which was extracted from the Dataset is shown in the figure below:

```
1) Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The
American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers
24\u201310 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at
Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the \"golden anniversary\" with various gold-themed
initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would
have been known as \"Super Bowl L\"), so that the logo could prominently feature the Arabic numerals 50.
2) The Panthers finished the regular season with a 15\u20131 record, and quarterback Cam Newton was named the NFL Most Valuable Player (
MVP). They defeated the Arizona Cardinals 49\u201315 in the NFC Championship Game and advanced to their second Super Bowl appearance since
the franchise was founded in 1995. The Broncos finished the regular season with a 12\u20134 record, and denied the New England Patriots a
chance to defend their title from Super Bowl XLIX by defeating them 20\u201318 in the AFC Championship Game. They joined the Patriots,
Dallas Cowboys, and Pittsburgh Steelers as one of four teams that have made eight appearances in the Super Bowl.
3) The Broncos took an early lead in Super Bowl 50 and never trailed. Newton was limited by Denver's defense, which sacked him seven times
and forced him into three turnovers, including a fumble which they recovered for a touchdown. Denver linebacker Von Miller was named Super
Bowl MVP, recording five solo tackles, 2\u00bd sacks, and two forced fumbles.
4) CBS broadcast Super Bowl 50 in the U.S., and charged an average of $5 million for a 30-second commercial during the game. The Super Bowl
50 halftime show was headlined by the British rock group Coldplay with special guest performers Beyonc\u00e9 and Bruno Mars, who headlined
the Super Bowl XLVII and Super Bowl XLVIII halftime shows, respectively. It was the third-most watched U.S. broadcast ever.
5) In early 2012, NFL Commissioner Roger Goodell stated that the league planned to make the 50th Super Bowl \"spectacular\" and that it
would be \"an important game for us as a league\".
6) The league eventually narrowed the bids to three sites: New Orleans' Mercedes-Benz Superdome, Miami's Sun Life Stadium, and the San
Francisco Bay Area's Levi's Stadium.
7) The league announced on October 16, 2012, that the two finalists were Sun Life Stadium and Levi's Stadium. The South Florida/Miami area
has previously hosted the event 10 times (tied for most with New Orleans), with the most recent one being Super Bowl XLIV in 2010. The San
Francisco Bay Area last hosted in 1985 (Super Bowl XIX), held at Stanford Stadium in Stanford, California, won by the home team 49ers. The
Miami bid depended on whether the stadium underwent renovations. However, on May 3, 2013, the Florida legislature refused to approve the
funding plan to pay for the renovations, dealing a significant blow to Miami's chances.
8) On May 21, 2013, NFL owners at their spring meetings in Boston voted and awarded the game to Levi's Stadium. The $1.2 billion stadium
opened in 2014. It is the first Super Bowl held in the San Francisco Bay Area since Super Bowl XIX in 1985, and the first in California
since Super Bowl XXXVII took place in San Diego in 2003.
9) For the third straight season, the number one seeds from both conferences met in the Super Bowl. The Carolina Panthers became one of only
ten teams to have completed a regular season with only one loss, and one of only six teams to have acquired a 15\u20131 record, while the
Denver Broncos became one of four teams to have made eight appearances in the Super Bowl. The Broncos made their second Super Bowl
appearance in three years, having reached Super Bowl XLVIII, while the Panthers made their second Super Bowl appearance in franchise
history, their other appearance being Super Bowl XXXVIII. Coincidentally, both teams were coached by John Fox in their last Super Bowl
appearance prior to Super Bowl 50.
10) Despite waiving longtime running back DeAngelo Williams and losing top wide receiver Kelvin Benjamin to a torn ACL in the preseason, the
Carolina Panthers had their best regular season in franchise history, becoming the seventh team to win at least 15 regular season games
since the league expanded to a 16-game schedule in 1978. Carolina started the season 14\u20130, not only setting franchise records for the
best start and the longest single-season winning streak, but also posting the best start to a season by an NFC team in NFL history, breaking
the 13\u20130 record previously shared with the 2009 New Orleans Saints and the 2011 Green Bay Packers. With their NFC-best 15\u20131
regular season record, the Panthers clinched home-field advantage throughout the NFC playoffs for the first time in franchise history. Ten
players were selected to the Pro Bowl (the most in franchise history) along with eight All-Pro selections.
```

Figure 3.2 Passage Description

3.2.2 Question and Answer Prepossessing

In general, A Passage will contain many Question and Answer. Now a Passage is subdivided into another Passage which will contain all the Question set. Further, each Question will have multiple Answer(s), so a Question and its corresponding Answer(s) will be grouped together. A sample output for the above mentioned is shown below:

```
Paragraph: 1
1) Which NFL team represented the AFC at Super Bowl 50
1.1) Denver Broncos
1.2) Denver Broncos
1.3) Denver Broncos
2) Which NFL team represented the NFC at Super Bowl 50
2.1) Carolina Panthers
2.2) Carolina Panthers
2.3) Carolina Panthers
3) Where did Super Bowl 50 take place
3.1) Santa Clara, California
3.2) Levi's Stadium
3.3) Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.
4) Which NFL team won Super Bowl 50
4.1) Denver Broncos
4.2) Denver Broncos
4.3) Denver Broncos
5) What color was used to emphasize the 50th anniversary of the Super Bowl
5.1) gold
5.2) gold
5.3) gold
6) What was the theme of Super Bowl 50
6.1) \"golden anniversary\"
6.2) gold-themed
6.3) \"golden anniversary
7) What day was the game played on
7.1) February 7, 2016
7.2) February 7
7.3) February 7, 2016
8) What is the AFC short for
8.1) American Football Conference
8.2) American Football Conference
8.3) American Football Conference
9) What was the theme of Super Bowl 50
9.1) \"golden anniversary\"
9.2) gold-themed
9.3) gold
10) What does AFC stand for
10.1) American Football Conference
10.2) American Football Conference
10.3) American Football Conference
11) What day was the Super Bowl played on
11.1) February 7, 2016
11.2) February 7
11.3) February 7, 2016
```

Figure 3.2 Question and Answer Description

Chapter 4

Question Answer System Components

The architecture of our QA system is presented in Figure 4.1. There are 3 separate modules that handle various stages in the system's pipeline: The first model is called Question Processing module. It is identify the focus of the question, classifies the question type derives the expected answer type, and reformulates the question into semantically equivalent multiple questions.
The second model is a Document Processing. In which mainly dealing information retrieval (IR), which takes a query as input and returns a list of documents deemed to be relevant to the query in a sorted manner. IR system recall is very important for question answering, because if no correct answers are present in a document, no further processing could be carried out to find an answer [9].
A third model, called answer processing. In which mainly dealing Answer Extraction, analyses the content presented and chooses the text fragment deemed to be the best answer to the posed question[7].

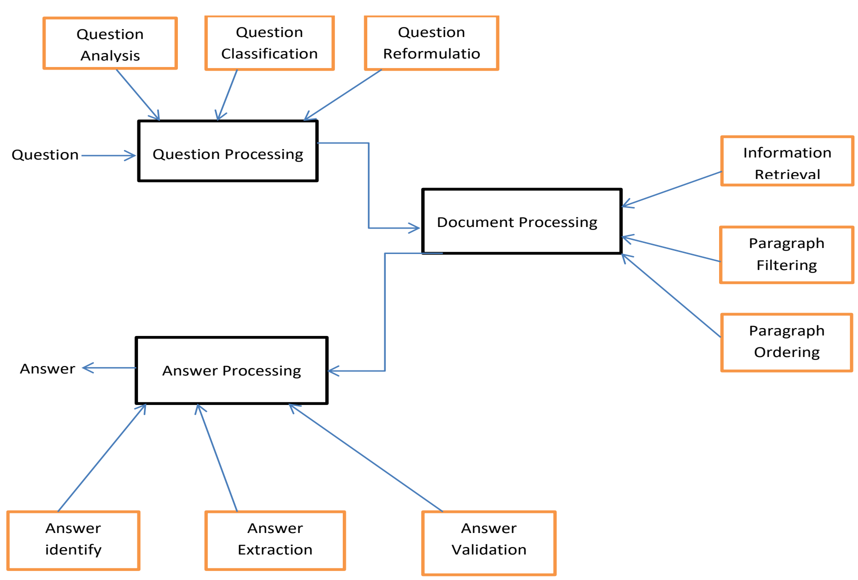

Figure 4.1 Question Answer System Architecture

Typically, the following scenario occurs in the QA system:
1. First, the user posts a question to the QA system.
2. Next the question analyser determines the focus of the question in order to enhance the accuracy of the QA system.
3. Question classification plays a vital role in the QA system by identifying the question type and consequently the type of the expected answer.
4. In question reformulation, the question is rephrased by expanding the query and passing it the information retrieval system.
5. The information retrieval component is used to retrieve the relevant documents based upon important keywords appearing in the question.
6. The retrieved relevant documents are filtered and shortened into paragraphs that are expected to contain the answer.
7. Then, these filtered paragraphs are ordered and passed to the answer processing module.
8. Based on the answer type and other recognition techniques, the candidate answers are identified.
9. A set of heuristics is defined in order to extract only the relevant word or phrase that answers the question.
10. The extracted answer is finally validated for its correctness and presented to the user.

4.1 Question Processing Module

Given a natural language question as input, the overall function of the question processing module is to analyze and process the question by creating some representation of the information requested. Therefore, the question processing module is required to:

- **Analyze:** The question, in order to represent the main information that is required to answer the user's question.
- **Classify:** The question type, usually based on taxonomy of possible questions already coded into the system, which in turn leads to the expected answer type, through some shallow semantic processing of the question.

- **Reformulate:** The question, in order to enhance the question phrasing and to transform the question into queries for the information retrieval (search engine).

These steps allow the question processing module to finally pass a set of query terms to the document processing module, which uses them to perform the information retrieval

4.1.1 Question Analysis

Question analysis is also referred to as "Question Focus". Unfortunately, classifying the question and knowing its type is not enough for finding answers to all questions. The "what" questions in particular can be quite ambiguous in terms of the information asked by the question [10]. In order to address this ambiguity, an additional component which analyses the question and identifies its focus is necessary.

The focus of a question has been defined by Moldovan et al. [3] to be a word or sequence of words which indicate what information is being asked for in the question. For instance, the question "What is the longest river in India?" has the focus "longest river". If both the question type (from the question classification component) and the focus are known, the system is able to more easily determine the type of answer required.

Identifying the focus can be done using pattern matching rules, based on the question type classification.

4.1.2 Question Type Classification:

In order to correctly answer a question, it is required to understand what type of information the question asks for, because knowing the type of a question can provide constraints on what constitutes relevant data (the answer), which helps other modules to correctly locate and verify an answer.

The question type classification component is therefore a useful, if not essential, component in a QA system as it provides significant guidance about the nature of the required answer. Therefore, the question is first classified by its type: what, why, who, how, when, where questions, etc.

4.1.3 Answer Type Classification

Answer type classification is a subsequent and related component to question classification. It is based on a mapping of the question classification. Once a question has been classified, a simple rule based mapping would be used to determine the potential answer types. Again, because question classification can be ambiguous, the system should allow for multiple answer types.

4.1.4 Question Reformulation

Once the "focus" and "question type" are identified, the module forms a list of keywords to be passed to the information retrieval component in the document processing module. The process of extracting keywords could be performed with the aid of standard techniques such as named-entity recognition, stop-word lists, and part-of-speech taggers, etc.
Other methods of expanding the set of question keywords could include using an online lexical resource such as the WordNet ontology. The synsets (synonym sets) in WordNet could be used to expand the set of question keywords with semantically related words that might also occur in documents containing the correct question answer [11].

4.2 Document Processing Module

The document processing module in QA systems is also commonly referred to as paragraph indexing module, where the reformulated question is submitted to the information retrieval system, which in turn retrieves a ranked list of relevant documents. The document processing module usually relies on one or more information retrieval systems to gather information from a collection of document corpora which almost always involves the www as at least one of these corpora [10]. The documents returned by the information retrieval system is then filtered and ordered.

Therefore, the main goal of the document processing module is to create a set of candidate ordered paragraphs that contain the answer(s), and in order to achieve this goal, the document processing module is required to:

- **Retrieve:** A set of ranked documents that are relevant to the submitted question.
- **Filter:** The documents returned by the retrieval system, in order to reduce the number of candidate documents, as well as the amount of candidate text in each document.
- **Order:** The candidate paragraphs to get a set of ranked paragraphs according to a plausibility degree of containing the correct answer.

The motivation for shortening documents into paragraphs is making a faster system. The response time of a QA system is very important due to the interactive nature of question answering. This ensures that a reasonable number of paragraphs are passed on to the answer processing module.

4.2.1 Information Retrieval

Information domains, such as the web, have enormous information content. Therefore, the goal of the information retrieval system is to retrieve accurate results in response to a query submitted by the user, and to rank these results according to their relevancy.

One thing to be considered is that it is not desirable in QA systems to rely on IR systems which use the cosine vector space model for measuring similarity between documents and queries. This is mainly because a QA system usually wants documents to be retrieved only when all keywords are present in the document. This is because the keywords have been carefully selected and reformulated by the Question Processing module. IR systems based on cosine similarity often return documents even if not all keywords are present.

Information retrieval systems are usually evaluated based on two metrics precision and recall. Precision refers to the ratio of relevant documents returned to the total number of documents returned. Recall refers to the number of relevant documents returned out of the total number of relevant documents available in the document collection being searched. In general, the aim for information retrieval systems is to optimize both precision and recall. For question answering, however, the focus is subtly different. Because a QA system performs post processing on the documents returned, the recall of the IR system is significantly more important than its precision [10].

4.2.2 Paragraph Filtering

As mentioned before, the number of documents returned by the information retrieval system may be very large. Paragraph filtering can be used to reduce the number of candidate documents, and to reduce the amount of candidate text from each document. The concept of paragraph filtering is based on the principle that the most relevant documents should contain the question keywords in a few neighbouring paragraphs, rather than dispersed over the entire document. Therefore, if the keywords are all found in some set of N consecutive paragraphs, then that set of paragraphs will be returned, otherwise, the document is discarded from further processing

4.2.3 Paragraph Ordering

The aim of paragraph ordering is to rank the paragraphs according to a plausibility degree of containing the correct answer. Paragraph ordering is performed using standard radix sort algorithm. The radix sort involves three different scores to order paragraphs:
(i)**. same word sequence score:** The number of words from the question that are recognized in the same sequence within the current paragraph window.
(ii). Distance score: The number of words that separate the most distant keywords in the current paragraph window;
(iii). missing keyword score: The number of unmatched keywords in the current paragraph window.
A paragraph window is defined as the minimal span of text required to capture each maximally inclusive set of question keywords within each

paragraph. Radix sorting is performed for each paragraph window across all paragraphs.

4.3 Answer Processing Module

As the final phase in the QA architecture, the answer processing module is responsible for identifying, extracting and validating answers from the set of ordered paragraphs passed to it from the document processing module. Hence, the answer processing module is required to:

- **Identify:** The answer candidates within the filtered ordered paragraphs through parsing.
- **Extract:** The answer by choosing only the word or phrase that answers the submitted question through a set of heuristics.
- **Validate:** The answer by providing confidence in the correctness of the answer.

4.3.1 Answer Identification

The answer type which was determined during question processing is crucial to the identification of the answer. Since usually the answer type is not explicit in the question or the answer, it is necessary to rely on a parser to recognize named entities (e.g. names of persons and organizations, monetary units, dates, etc.). Also, using a part-of-speech tagger (e.g., Brill tagger) can help to enable recognition of answer candidates within identified paragraphs. The recognition of the answer type returned by the parser creates a candidate answer. The extraction of the answer and its validation are based on a set of heuristics [3].

4.3.2 Answer Extraction

The parser enables the recognition of the answer candidates in the paragraphs. So, once an answer candidate has been identified, a set of heuristics is applied in order to extract only the relevant word or phrase that answers the question.

Researchers have presented miscellaneous heuristic measures to extract the correct answer from the answer candidates. Extraction can be based on measures of distance between keywords, numbers of keywords matched and other similar heuristic metrics. Commonly, if no match is found, QA systems would fall back to delivering the best ranked paragraph. Unfortunately, given the tightening requirements of the TREC QA track, such behaviour is no longer useful. As in the original TREC QA tracks, systems could present a list of several answers, and were ranked based on where the correct answer appeared in the list. From 1999-2001, the length of this list was 5. Since 2002, systems have been required to present only a single answer [12].

4.3.3 Answer Validation

Confidence in the correctness of an answer can be increased in a number of ways. One way is to use a lexical resource like WordNet to validate that a candidate response was of the correct answer type. Also, specific knowledge sources can also be used as a second opinion to check answers to questions within specific domains. This allows candidate answers to be sanity checked before being presented to a user. If a specific knowledge source has been used to actually retrieve the answer, then general web search can also be used to sanity check answers. The principle relied on here is that the number of documents that can be retrieved from the web in which the question and the answer co-occur can be considered a significant clue of the validity of the answer. Several people have investigated using the redundancy of the web to validate answers based on frequency counts of question answer collocation, and found it to be surprisingly effective. Given its simplicity, this makes it an attractive technique.

Chapter 5

Method's Purpose

5.1 R-NET METHOD

Figure 5.1 gives an overview of R-NET. First, the question and passage are processed by a bidirectional recurrent network [13] separately. then match the question and passage with gated attention-based recurrent networks, obtaining question-aware representation for the passage. On top of that, I apply self-matching attention to aggregate evidence from the whole passage and refine the passage representation, which is then fed into the output layer to predict the boundary of the answer span.

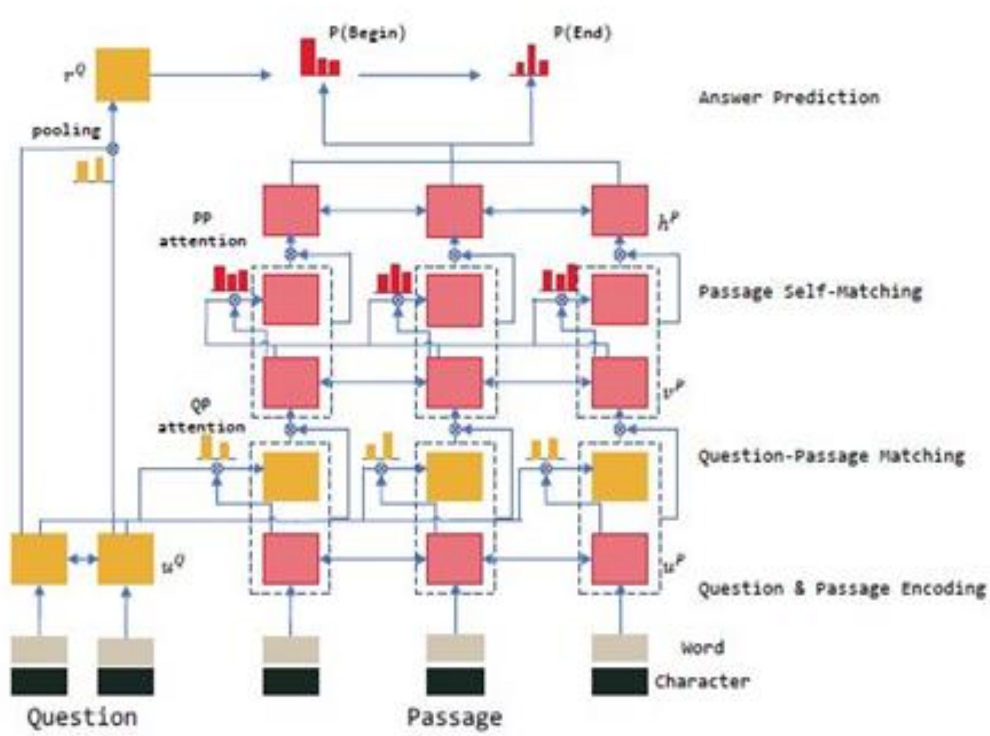

Figure 5.1 R-NET structure overview

5.1.1 QUESTION AND PASSAGE ENCODER

Consider a question Q and a passage p. I first convert the words to their respective word-level embedding and character level embedding's. The character level embedding are generated by taking the final hidden state of a bi-directional recurrent neural network (RNN) applied to embedding's of characters in the token. Such character-level embedding's have been shown to be helpful to deal with out-of-vocab (OOV) tokens. Then I use a bi-directional RNN to produce new representation
I choose to use Gated Recurrent Unit (GRU) **[14]** in our experiment since it performs similarly to LSTM [15] but is computationally cheaper.

5.1.2 GATED ATTENTION-BASED RECURRENT NETWORKS

I propose a gated attention-based recurrent network to incorporate question information into passage representation. It is a variant of attention-based recurrent networks, with an additional gate to determine the importance of information in the passage regarding a question. Given question and passage representation propose generating sentence-pair representation via soft-alignment of words in the question and passage.

5.1.3 SELF-MATCHING ATTENTION

Through gated attention-based recurrent networks, question-aware passage representation is generated to pinpoint important parts in the passage. One problem with such representation is that it has very limited knowledge of context. One answer candidate is often oblivious to important cues in the passage outside its surrounding window. Moreover, there exists some sort of lexical or syntactic divergence between the question and passage in the majority of SQuAD dataset [8]. Passage context is necessary to infer the answer. To address this problem, we propose directly matching the question-aware passage representation against itself. It dynamically collects evidence from the whole passage for words in passage and encodes the evidence relevant to the current passage word and its matching question information into the passage representation

Self-matching extracts evidence from the whole passage according to the current passage word and question information.

5.1.4 OUTPUT LAYER

I follow Wang & Jiang [16] and use pointer networks [17] to predict the start and end position of the answer. In addition, we use an attention-pooling over the question representation to generate the initial hidden vector for the pointer network. Given the passage representation the attention mechanism is utilized as a pointer to select the start position (p1) and end position (p2) from the passage.

References

[1] L. Hirschman and R. Gaizauskas, "Natural language question answering: the view from here," Natural Language Engineering, vol. 7, no. 4, pp. 275-300, 2001.

[2] D. Zhang and W. Lee, "A Web-based Question Answering System," Massachusetts Institute of Technology (DSpace@MIT), 2003.

[3] D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Girju and V. Rus, "Lasso: A Tool for Surfing the Answer Net," in Proceedings of the Eighth Text Retrieval Conference (TREC-8), 1999.

[4] Abdessamad Echihabi and Daniel Marcu. 2003. A Noisy-Channel Approach to Question Answering. Proceedings of the ACL 2003. Sapporo, Japan.

[5] P. Banerjee and H. Han, "Drexel at TREC 2007: Question Answering," in Proceedings of the Sixteenth Text Retrieval Conference (TREC 2007), 2007.

[6] M. M. Sakre, M. M. Kouta and A. M. N. Allam, "Automated Construction of Arabic-English Parallel Corpus," Journal of the Advances in Computer Science, vol. 3, 2009.

[7] M. Ramprasath and S. Hariharan, "A Survey on Question Answering System," International Journal of Research and Reviews in Information Sciences (IJRRIS), pp. 171-179, 2012.

[8] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016.

[9] S. Stoyanchev, Y. Song and W. Lahti, "Exact phrases in information retrieval for question answering," in Proceedings of the 2nd workshop on Information Retrieval for Question Answering, 2008.

[10] A. Lambert, "A Quick Introduction to Question Answering," CSIRO ICT Centre, 2004.

[11] M. M. Sakre, M. M. Kouta and A. M. N. Allam, "Weighting Query Terms Using Wordnet Ontology," International Journal of Computer Science and Network Security, vol. 9, no. 4, pp. 349-358, 2009.

[12] E. Voorhees, "Overview of the TREC 2002 Question Answering Track," in Proceedings of the Text Retrieval Conference (TREC 2002), 2002.

[13] Ankur P. Parikh, Oscar T¨ackstr¨om, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. "*In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*," EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, 2016.

[14] Yichen Gong and Samuel R Bowman. Ruminating reader: Reasoning with gated multi-hop attention. ArXiv preprintarXiv: 1704.07415, 2017.

[15] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In ACL (System Demonstrations), pp. 55–60, 2014.

[16] Wenhui Wang, Nan Yang, FuruWei, Baobao Chang, and Ming Zhou. Gated self-matching networks for reading comprehension and question answering. In Association for Computational Linguistics (ACL), 2017.

[17] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pp. 2692–2700, 2015.