مجلة العلوم البحتة والتطبيقية
Journal of Pure & Applied Sciences
www.Suj.sebhau.edu.ly ISSN 2521-9200
Received 05/01/2018 Revised 02/02/2018 Published online 14/02/2018

Automated Answer Extraction for Reading Comprehension System Based on Matching Approach

*Fatimah Amhimmid Mare , Mohamed Abdelgader Matoug , Kamal Mohamed Elsanoussi
Department of Computer Science, Faculty of Education Tragan, Sebha University, Libya
*Corresponding author: fat.mare@sebhau.edu.ly

Abstract One of Reading Comprehension (RC) tasks is inspired by the Information Extraction (IE) application to extract a set of features from a natural language text. Since reading comprehension tests were created to judge the reading ability of humans, there are challenges in using language understanding systems to extract information from comprehension stories. The questions and answer keys already exist in the story. The challenge how use the language understanding system to find automatic an answer for questions. The main target in this study is to review the matching approach of Natural Language Processing techniques to extract information from reading comprehension; the information would be able to answer the WH questions of reading comprehension texts. The matching approach decomposed the story sentences and questions into a container of words that were augmented with additional automated linguistic processing and then the answering engine stage is applied to the matching process after representing the information into a bag of words. Because the answer to a question must come from the given document, the story structure has to be examined in the context of responding to test questions. On WH questions, the experiment tested 15 children's stories that contained 262 sentences (with an average of 18 sentences per story) and 75 WH questions. The result achieved was 67.3% Hum sent accuracy of the correct answer on the questions pertaining to the children's stories.

Keywords: Automated linguistic processing, Information Extraction (IE), Matching approach, Natural Language Processing (NLP), Natural Language Understanding (NLU), Question Answering (QA), Reading comprehension system (RCS).

إستخراج الجواب الآلي لنظام القراءة الاستيعابية وفقاً للنهج القائم على نظام المطابقة

*فاطمة امحمد مرعي و محمد عبدالقادر معتوق و كمال محمد السنوسي
قسم علوم الحاسوب - كلية التربية/تراغن - جامعة سبها، ليبيا
*للمراسلة: fat.mare@sebhau.edu.ly

الملخص تعتبر نظم القراءة الاستيعابية من أحد النظم المستوحاة من تطبيقات استخراج المعلومات لتحديد مدى القدرة على إستيعاب النص وفهمه وهي لاستخراج مجموعة من السمات من نص اللغة الطبيعية. كانت هناك العديد من الصعوبات الناتجه عن استخدام أنظمة فهم اللغة لاستخراج المعلومات من قصص القراءة الاستيعابية منذ إنشاء اختبارات القراءة الاستيعابية للحكم على قدرة البشر على فهم النص المقروء. حيت تكون الاسئلة ومفاتيح الاجابة موجودة بالفعل في القصة لكن التحدي يكمن في كيفية استخدام نظم فهم اللغة للعثور على اجابات الاسئلة بشكل تلقائي و بصورة آلية. النهج المستخدم: الهدف الرئيسي من هذه الدراسة هو استعراض النهج القائم على المطابقة لتقنيات معالجة اللغات الطبيعية لاستخراج المعلومات من قصص القراءة الاستيعابية؛ وهذه المعلومات ستكون قادرة على الاجابة على الاسئلة من النص . يقوم نهج المطابقة بتحليل جمل القصة وكذلك الاسئلة لمجموعة من الكلمات مع معالجتها لغويا ومن تم تطبيق مرحلة البحث عن الاجابة من خلال التطابق بعد استعراض معلومات الاسئلة والاجوبة من خلال مجموعه من الكلمات. نظرا لان الاجابة على السؤال يجب أن تأتي من الوثيقة المعطاة ، فإن تركيب القصة يجب ان يتم فحصه بناء على اجابات الاسئلة. النتائج: تم اختبار التجربه على 15 قصة من قصص القراءة الاستيعابية المخصصة للاطفال وفقا لنظام الاسئلة التي تبدأ ب (من، الى اين، كم العدد، الخ) والتي كانت تحتوي على 262 جملة (بمتوسط 18 جملة لكل قصة) وعدد 75 سؤال ، فكانت النتائج المتحققة بنسبة 67.3% للاجابات الصحيحة على الاسئلة المتعلقة بقصص الاطفال.

الكلمات المفتاحية: المعالجة اللغوية الآلية، استخراج المعلومات ، نهج المطابقة، معالجة اللغات الطبيعية ، فهم اللغة الطبيعية ، اجابة الاسئلة ، نظام القراءة الاستيعابية.

INTRODUCTION

Natural language processing (NLP) is a field related to the area of computer science, artificial intelligence, linguistics and human computer interactions by means of which computational mechanisms are investigated and formulated. These mechanisms allow the development of systems that is capable of understanding the knowledge expressed in texts of a given language. Natural Language Processing is a theoretically range of computational techniques for representing and analyzing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications[1].

The most common applications utilizing NLP include the following: systems of machine translation, information retrieval (IR), information extraction, question answering (QA), recognition of entities, classification and filtrate of documents, generation of summaries, etc. The Natural Language Processing (NLP) community has been granted increased attention to the issues that related to Reading Comprehension systems and utilized as a means to solve, develop and evaluate reading comprehension tasks to attain a better understanding of documents. It endeavors to offer answers to questions expressed in natural language. A reading comprehension system is sufficiently close to information extraction applications [2]. The information extraction and reading comprehension have need of natural language understanding. For both, the difference between them is that reading comprehension strives to understand the entire story [3]. Reading comprehension is the ability to read text, process it, and understand its meaning. Reading comprehension is a dynamic and an interactive process, To understand a text, the reader needs to recognize each word and retrieve its meaning, combine this information with syntactic knowledge to make meaningful sentences and integrate the meanings of each sentence to construct representation of the state of affairs described by the text [4].

However the level of understanding differs from reader to reader. To evaluate their understanding levels, reading comprehension tests are proposed. Such tests ask reader to read a story and to demonstrate his/her understanding of that story by answering questions about it. Reading comprehension task involves reading a short passage of text and answering a series of questions pertaining to that text. The questions from each passage are chosen to measure how well the system has understood the narrative. Therefore, this task was proposed as one of methods for evaluating Natural Language Understanding NLU technologies.

Reading comprehension Task: There has been growing attention to reading comprehension in the branch of Question Answering (QA) systems, where from the early days of artificial intelligence in the 60's, researchers have been fascinated with answering natural language questions [5] and the question answering system has been used in many areas of NLP research as natural language database systems, dialogue systems, reading comprehension systems and open domain question answering. As a matter of fact, reading comprehension tests are a form of single document question answering. A single document task involves questions associated with one particular document. In most cases, the answer appears somewhere in the document.

LITERATURE REVIEW

The idea of a reading comprehension system simulating understanding with respect to reading a story or passage and answering questions pertaining to it, has attracted researchers since the early 1970s [6]. The early work by Lehnert (1978) [7] focused more on knowledge representation and inference issues and the work was planned as devices for modeling human story comprehension. There are many automated reading comprehension systems and their approaches towards solving the problem vary differently from each other. Researchers appeared to pay wide attention to the problem of answering questions based on the comprehension of stories. In the late 1990s, one of the early reading comprehension tasks was proposed by the MITRE Corporation which developed the Deep Read system. Hirschman et al. (1999) [8] in Deep Read describe an automated reading comprehension system that accepts text input (a story) and answers questions about it. This system has been used pattern matching bag of words techniques augmented with other automated linguistic techniques such as stemming, name identification and semantic class identification. The MITRE Corporation group defined the "Remedia Corpus" that consists of 115 short stories to evaluate the RC system. The MITRE group also defined the Hum Sent scoring metric, Hum Sent answers were compiled by a human annotator, who examined the stories and chose the sentence(s) that best answered the questions. Deep Read achieved 36.3% Hum Sent accuracy in the Remedia test set.

Another well.-known Question Answering system is Quarc (Question Answering for Reading Comprehension) by Riloff and Thelen (2000) , [9] Quarc is a rule-based system that uses lexical and semantic heuristics to look for evidence that a sentence contains the answer to question. Each type of WH question looks for different types of answer and the rules are applied to each sentence in the story. Each rule awards a specific number of points to a sentence, depending on how strongly the rule believes that has found the answer. A rule can assign four possible point values: clue (+3), good-clue (+4), confident (+6) and slam-dunk (+20). The main purpose of these values is to assess the relative importance of each clue for different question types, such as the WHERE rule: if it contain(S, LOCATION),

then Score(S)+ = confident

where, questions reward candidate answer sentences with 6 extra points if they contain a named entity LOCATION. Then the sentences with the top score are cut off. Quarc was evaluated on the same data set that was used to evaluate the Deep Read reading comprehension system. Quarc's performance based Hum Sent accuracy on WHAT at 28% , WHEN at 55% and WHY at 28% questions had improved by several percentage points, but performance on WHO 41% and WHERE 47% questions. AQUAREAS (Automated Question Answering upon Reading Stories) by Ng et al. (2000) [10] used a machine-learning approach to determine if a candidate sentence is the answer to a question based on numerous features. It was the early work that reported that the use of a machine learning approach could achieve competitive results on reading comprehension tests. The machine learning approach comprises two steps. First, a set of features was designed to capture the information that helped to distinguish answer sentences from non-answer sentences. Next, a learning algorithm was used to generate a classifier for each question type. This approach achieved 39.3% Hum Sent accuracy on the Remedia data set. Whereas, Qspecific System by Charniak et al. (2000) [11] performs at 41% (HumSentAcc) on the Remedia corpus. It combines the use of manually generated rules with statistical techniques as bag of words and Bag of Verbs (BOV) matching, as well as deeper semantic analysis of nouns. The BOV matching is a disparity of BOW matching in which only verbs are examined instead of all non-stop words. The techniques used different strategies for different questions.

Some previous work has shown that adding some of linguistic technologies to a reading comprehension question answering system will offer some improvement to the performance of RC systems. Xu and Meng (2005b) [12] have developed a reading comprehension system by using bag of words and syntactic features in an attempt to improve the accuracy. Syntactic features are represented by verb dependencies in the system. The context independent meaning of a sentence can be represented by the logical forms, which can be captured using relationships between verbs and noun phrases. Xu and Meng achieved 40% Hum sent accuracy on the Remedia test set. While, Du et al. (2005) [13] , [16] also used bag of words as baseline set and proposes an approach towards RC that attempts to utilize external knowledge to improve the performance. By metadata, they are referring to automatically labeled verbs, named entities as well as base noun phrases in the passage. It is important to achieve match between the question and a candidate answer sentence before the candidate is selected as the final answer. Du et al. (2007) [3] proposed an approach towards RC, other techniques such as linguistic feature matching; the semantic extending and named entity filtering are used incrementally to boost the system performance. Feature matching is an extension of the BOW approach. The system achieved 41.3% Hum Sent accuracy overall.

MATERIALS AND METHODS

The bag of words approach: Bag of words is the traditional method used for extracting information [14] and it is one of the most popular representation methods for reading comprehension which consists of representing each document by the words that belong to it and then process those words in order to generate information to answer reading comprehension questions.

This study aims to generate automatic answers for reading comprehension questions, which concentrate on the factoid questions which involve the five WH questions: WHO, WHAT, WHEN, WHERE and WHY. A very important characteristic of using the representation of information by using BOW is the ability to capture the co-occurring features in both the passage and WH questions. The BOW is proposed approach with other NLP techniques to automatic answered reading comprehension questions, this approach incorporates some of automated linguistic processing including tokenize, remove stop-words and stemming. Every two words are considered a match if they share the same morphological root. Given a question, the BOW matching approach selects the passage sentence with the maximum number of matching words as the answer.

The bag of words treats document as container of words, it is the norm in many applications and has been shown to be surprisingly effective in addressing a broad range of NLP tasks, including extraction information, word sense disambiguation, text categorization, reading comprehension and information filtering system.

Theory and features of bag of words: The text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar. The generation of an electronic document will lead to the following features:

- Text document is represented by the words for every sentence it contains. This representation makes learning far simpler and easier
- Stemming to identify a word by its root, the word with a common stem will usually have similar meanings
- Stop words are also used whereas, the most common words are unlikely to help e.g., “the”, “a”, “an”, “you”. Usually those common stop words are removed to return the most relevant result from a document and to keep content words (verbs, adverbs, nouns, adjectives) in text

Representation information based on bag of words approach: The approach to representing information for passage in a text collection that share content with an input question, is retrieval using a bag of words model. It is based on the hypothesis that text can be signified as a collection of featured elements. It allows a document looks like a “bag”.

A document can be treated as a set of sentences and features extracted from the sentence are considered as the “individual words” and serves as the basic element for further processing.

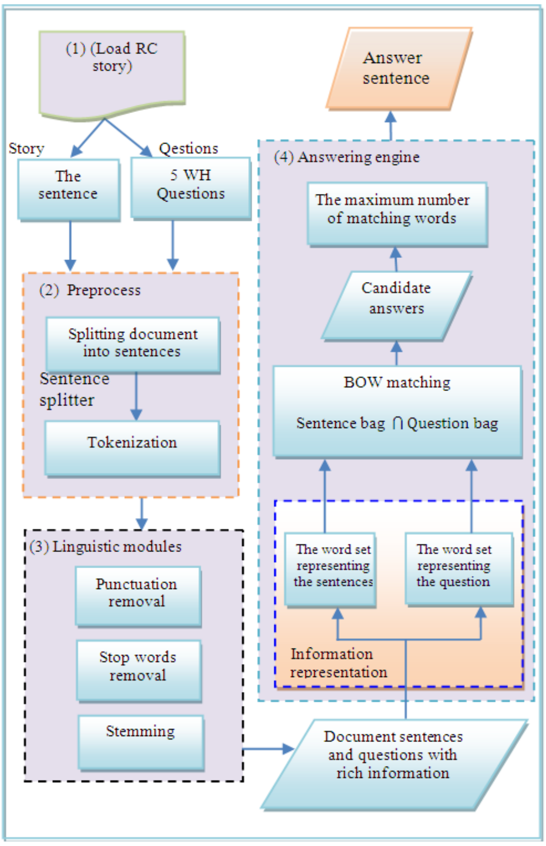

Fig. 1: The flow chart of our reading comprehension system

Story2-9.txt
The Hare and the Tortoise story.

The Hare was once boasting of his speed before the other animals. I have never yet been beaten, said he, when I put forth my full speed. I challenge any one here to race with me. The Tortoise said quietly, I accept your challenge. That is a good joke, said the Hare; I could dance round you all the way. Keep your boasting till you've beaten, answered the Tortoise, Shall we race?

So a course was fixed and in the early morning a start was made. The Hare darted almost out of sight at once, But the Hare soon stopped and, believing that the Tortoise could never catch 34him; The Hare lay down by the wayside to have a nap.

The Tortoise never for a moment stopped, but went on with a slow but steady pace straight to the end of the course. When the Hare awoke from his nap, he saw the Tortoise just near the winning-post. The Hare ran as fast as he could, but it was too late. He saw the Tortoise had reached the goal. Then the Tortoise said to the Hare "slow and steady wins the race".

1. Who said to the Hare "slow and steady wins the race"?
2. What time did the race start?
3. When did the Tortoise stop before the end of race?
4. Where did Hare lie down to have a nap?
5. Why did the Hare stop before the end of race?

Fig. 2: Simple of reading comprehension story

Each sentence is represented by its signature elements of words that will used later in the stage of extract the answer for the task of reading comprehension system. Where the processing of bag of words creates signature elements for each word in the sentence, this approach coverts a feature as tokens by reducing a text to group of words that the system looked-for.

The BOW makes the largest contribution to the RC system results and it can give a significant improvement. Therefore, a number of researchers in the field investigated the use of representation information of a BOW approach for answer retrieval by using matching approach on the information, those information have been extracted when the input question and sentences in passage has been converted to "bag of words" by removing the stop words and the punctuation, after that getting the root of its words.

System architecture: The flow chart of our RC system is shown in Fig. 1. In the current system, an automated linguistic processing is applied to represent and extract information from the story sentences and questions. BOW matching approach used this information to determine the best answer sentences for the questions of comprehension story.

The processes and phases come in four phases that are showing in following points:

- Phase 1: Load RC story
- Phase 2: Preprocessing
- Phase 3: Linguistic processing modules
- Phase 4: Answering engine

Phase 1: Load RC story: Knowing what data is required for this study is an important step, since the scope of this is to automate the answer reading comprehension system; a set of children story is chosen as the data source. Fig. 2 shows simple of reading comprehension story with its WH questions. Due to the fact that data exists today in various sources on different platforms and must be copied from its sources for use.

The system accepts a story as input and stores data in TXT format. The procedures for this phase are summarized as follows:

- Read the specified file that contains the story
- Export the story into a text file and the format of the file should be as follows
- Convert all uppercase characters in a string to lowercase
- As far as the questions follow the story in the file, each question begins with its number (beginning at 1) and a period. Example: 1. Is this a question?
- Put one question per line
- Load each sentence in the story by obtaining the vector of sentence objects that contains each sentence in the story
- Reach the questions when a line begins with the first WH question
- Obtain also the vector of sentence objects that contains each question in the story

Phase 2: Pre-processing: The next stage is the pre-processing phase that is used to initiate linguistic processing. In this phase the program produce the input vector of sentences and question into list of sentence. Therefore, the pre-processing starts with sentence splitter and then split every sentence to its list of word. The steps involved in this phase are as follows:-

Sentence splitter: Sentence splitter is the process of demarcating and classifying text into individual sentences. A simple and limited way of dividing text into sentences would be to use sentence splitter function. It can detect the end of sentence by using some of sentence delimiters, our system treats occurrences of '.', '?' and '!' as sentence delimiters. In this phase the system splits the story into sentence tokens, the resulting sentence tokens will be then passed to the word tokenize phase.

Tokenization: In this phase every sentence is split into tokens. The term "token" adverts to an abstract entity for the smallest item in a text by parsing the sentence into its individual words by using "Tokenizer" that consists of partitioning input text into the constituent lexical units (words, punctuation). The steps involved in the word tokenize are as follows:

- Divided into an individual word in the sentence. (a word is defined to be any string of alphanumeric characters separated by spaces or punctuation)
- Store those individual words into the vector of word to pass on to some other form of linguistic processing modules
- After parses a sentence into its individual words, stores the original form of the sentence without the word tokenize for future use because if this sentence is determined to be the answer sentence, it can easily be output in its original form

Phase 3: Linguistic processing modules: Remove all non-alphabetic characters (e.g. symbols, spaces) from a string and also removing un-necessary words are important step. By Stemming, removes suffixes in order to determine the root or stem of a word.

Removal of punctuation: Punctuation removal aims to keep only the alpha and numeric characters in text ([a-z] [A-Z] [0-9]), so all punctuation is removed from the sentence. The reason for this is that punctuation is tagged differently than words. The absence of the punctuation does not impact the assessment process.

Stop words removal: Stop word removal function is used to remove the common stopwords which occur very often and are not of significant importance. A list of predefined stop words (a and, on, or, to, in, at, of, the and more ...) is used to eliminate words deemed useless, removing those words means reduce noise from a document. The steps involved in the stop words removal function are as follows:

- Create a stop word list (stopwords.txt) – The format for reading and writing is one stop word per line
- Save the stop words to the given file
- Lowercases all the stop-words before the test
- Load the stop words from the given file
- Get the current stop-word list as an array
- Check each word in text input for being a stop word or not by match it with every word in stop-word list
- Return true, if the word is in the current stop-word set
- If true, remove the stop-word; just replace stop-word by null
- After remove the stop words, send the words to make stemmer for every noun and verb in sentence

Stemming: The goal of stemming is to decrease inflectional forms and sometimes derivationally related forms of a word to a common base form. The stemming is implemented by using the natural language processing tools. The most usual algorithm for stemming and one that has frequently been shown to be empirically very effective, is Porter's algorithm [15]. The Porter stemmer is a conflation stemmer developed by Martin Porter at the University of Cambridge in 1980. The stemmer is based on the idea that the suffixes in the English language are predominantly made up of a combination of smaller and simpler suffixes. The next shows some rules of Porter's algorithm:

- With what is left, replace any suffix on the left with suffix on the right

BOW-S.TXT
(Representation of information in RC story)

BOW-s1: [hare][tortoise][story]
BOW-s2: [hare][once][boast][speed][before][animal]
BOW-S3: [never][beaten][when][put][forth][full][speed]
BOW-S4: [challenge][one][here][race][time]
BOW-S5: [tortoise][quietly][accept][challenge]
BOW-S6: [that][good][joke][hare][dance][round][way]
BOW-S7: [keep][boast][till][beaten][answer][tortoise][shall][race]
BOW-S8: [so][course][fix][early][morning][start][made]
BOW-S9: [hare][dart][out][sight][once]
BOW-S10: [soon][stop][believe][tortoise][never][catch]
BOW-S11: [hare][lay][down][wayside][nap]
BOW-S12: [tortoise][never][moment][stop][went][slow][steady][pace][straight][end][course]
BOW-S13: [when][hare][awoke][nap][saw][tortoise][near][win][post]
BOW-S14: [hare][ran][fast] [late]
BOW-S15: [saw][tortoise][reach][goal]
BOW-S16: [tortoise][hare][slow][steady][win][race]

Fig. 3: The representation for whole sentences

BOW-Q.TXT
(Representation of information for every question)

BOW-Q1:[who][hare][slow] [steady] [win][race]
BOW-Q2: [what][time][race][start]
BOW-Q3:[when][tortoise][stop][before][end][race]
BOW-Q4: [where][hare] [lie][down][nap]
BOW-Q5: [why][hare][stop] [before][end][race]

Fig. 4: The representation for the questions

For example:

iveness -> ive	effectiveness --> effective
fulness -> ful	usefulness --> useful
ousness -> ous	nervousness --> nervous

- Remove remaining standard suffixes: al, ance, ence, er, ic, able, ible, ant, ement,ment, ent, sion, tion, ou, ism, ate, iti, ous, ive,ize, ise

For example:
ance ->(null) allowance -> allow
able ->(null) adjustable -> adjust
ment ->(null) adjustment -> adjust
(*s or *t)) ion -> (null) adoption -> adopt

Phase 4: Answering engine: Many reading comprehension systems assume that there are common words shared between questions and answers, they measure the similarity between questions and answers by matching them. The steps involved in the function of answering engine are as follows sub steps.

Representation of information: This process shows how the information that gotten from the above phases is represented. The representation of information is processing as follows:

- The output word that have been get from linguistic processing modules process is obtained to use as input in this process
- Represent information that content of text sentences as the container of words, for every sentence make one bag to gather its feature words, the bag start with the number of sentence as in the following example of the BOW representation

Sentence1: The Hare was once boasting of his speed before the other animals.
(Bag-s1):
{[hare][once][boast][speed]
[before][animal] }

The sentence is represented as collection of words, disregarding grammar. (Bag-s1) used to show the unique information that content in (Sentence1). It is done after removing punctuation from tokens and remove irrelevant words of stop word to keep only the content words (verbs, adverbs, nouns, adjectives) the nouns and verbs in the BOW are replaced by their base forms, which are the outputs of the stemming process. The representation for (Sentence 1) is the set in (Bag-s1). The following Figure illustrates the representation for whole sentences of reading comprehension story.

- Represents the information content of a question as the set of words for WH questions (WHO, WHAT, WHERE, WHEN and WHY) to extract information from every question

The BOW also can represent the information content of a question in the same way that in the representation of sentence. Representation for story that has 5WH questions have been shown in Fig. 4:

- The word sets are considered to have no structure or order, it just focus on the point of contain unique elements

Bag of word matching: In previous step, the story sentences and reading comprehension questions are decomposed and represented as a collection of words. The BOW matching process selects the passage sentence with the maximum number of matching words as the answer. The steps involved in this process are as follows:

- Every two words are considered a match if they share the same morphological root
- The sentence that has matched with given a question will be called “candidate sentence” and will be collected
- Find the best match between the word set representing the question and the sets representing sentences in the document as following formula

Matching words = Sentences bag $\cap$ question bag

- The BOW matching system measures the match by the size of the intersection of the two word sets

BOW matching is conducted between the question word set and the sentence word set. Count the number of matching word between the two word set of sentences and a given question.

- Return the bag of word of candidate sentences that have match with word set in question

BOW matching is conducted between the question word set and the candidate answer word set. The set of candidate sentences will be filtered in next step of the answer extraction task.

Answer extraction: Answer searching focuses on the identification of information encoded in the wording of the question and matching this against information from the story. The search consists of finding the match between the word set representing the question and the sets representing the sentences in the reading comprehension document. The process of the answer extraction is based on the out of the intersection and the best match of the two word sets. This process is depicted as following:

- Return the original sentence for each of candidate sentence- the original sentences of the candidate sentences that have match with a given question will be returned from the original file of RC story
- Among of numerous candidate sentences that are found, the BOW matching approach measures the match by size of the intersection of the word sets, it selects the passage sentence with the maximum number of matching words

For example:
The question Q1: Who said to the Hare “slow and steady wins the race”?

The Hare and the Tortoise story.

<ANSQ5> The Hare was once boasting of his speed before the other animals</ANSQ5>. I have never yet been beaten, said he, when I put forth my full speed. <ANSQ2> I challenge any one here to race with me in any time </ANSQ2>.
The Tortoise said quietly, I accept your challenge. That is a good joke, said the Hare; I could dance round you all the way. <ANSQ3> Keep your boasting till you've beaten, answered the Tortoise, Shall we race? </ANSQ3>.
So a course was fixed and in the early morning a start was made. The Hare darted almost out of sight at once, But he soon stopped and, believing that the Tortoise could never catch him, <ANSQ4> The Hare lay down by the wayside to have a nap</ANSQ4>.
The Tortoise never for a moment stopped, but went on with a slow but steady pace straight to the end of the course. When the Hare awoke from his nap, he saw the Tortoise just near the winning-post. The Hare ran as fast as he could, but it was too late. He saw the Tortoise had reached the goal. <ANSQ1> Then the Tortoise said to the Hare “slow and steady wins the race” </ANSQ1>.

Fig. 5: The answer tags for RC story

The best sentence is chosen by filtering the candidate sentences that have been got as illustrated in Table 1. The candidate sentence that contains the maximum number of matching words will be returned as an answer sentence:

- If multiple candidate sentences contain the maximum number of matching words, the candidate sentence that appeared earlier is returned as an answer

As clear in Table 2, the first and second candidate sentences have an equal and same number of matching words (two words) but the approach will return the sentence that appear first:

- For given question, tag its answer sentence that are generate by the system as the best answer. The answer starts with the number of question

Fig. 5 shows the answer tags for the above example (The Hare and the Tortoise story). For example for question 1 “Q1”, its answer sentence will start with tag <ANSQ1> and at the end of answer it will close with tag </ANSQ1>.

Table 1: The maximum number of matching words

Candidate sentences	Matching words
S1-BOW	
candidate sentence1:The Hare and the Tortoise story.	1
S2-BOW	
candidate sentence2:The Hare was once boasting of his speed before the other animals.	1
S4-BOW	
candidate sentence3:I challenge any one here to race with me in any time.	1
S9-BOW	
candidate sentence4: The Hare darted almost out of sight at once,	1
S11-BOW:	
candidate sentence5:The Hare lay down by the wayside to have a nap.	1
S12-BOW	
candidate sentence6:The Tortoise never for a moment stopped, but went on with a slow but steady pace straight to the end of the course.	2
S14-BOW:	
candidate sentence7:The Hare ran as fast as he could, but it was too late.	1
S16-BOW	
candidate sentence8: The Tortoise said to the Hare, "slow and steady wins the race”.	5

Table2: Example of multiple candidate sentences

Question: When did the Tortoise stop before the end of race?
BOW: [when][tortoise][stop][before][end] [race]

Candidate Sentences	Matching Words
Candidate sentence1: The Tortoise never for a moment stopped, but went with a slow but steady pace straight to the end of the course. BOW: [tortoise][never][moment][stop] [went][slow][steady][pace][straight][end] [course]	2
Candidate sentence2:The Tortoise said to the Hare, “slow and steady wins the race”. BOW:[tortoise][hare][slow][steady][win][race]	2

RESULTS AND DISCUSSION

The experiment was based on the selected answer extraction technique. The experiment was conducted on the set of children story to prove that the selected technique is suitable in giving an automatic answer to the reading comprehension task.

Experiment dataset: The data consists of 15 short stories. The children stories are derived from different children web sites which consist of teaching materials for learning children; they have a different reading grade level (the ages range from 7-11). Each story typically has a particular event or individual as its focus and many stories also contain positive lessons and helpful information to illuminate the importance and relevance of the topic. The stories cover a wide time period and include a wide range of “current event” topics including science, natural disasters, economy, medical and the environment. Each story is approximately 200-350 words in length. Each passage has an average of 18 sentences and every sentence has 8-22 words. Accompanying each story are five short questions: who, what, when, where and why. The questions are simple “factoid” questions that are actually typical of what most education experts consider ideal for stimulating the interest of young human readers.

The RC stories are more suitable for our purposes because those stories are written for young children; the sentences are rather short, ranging from an approximate average of ten words per sentence. In many ways, for today’s NLP systems, this "elementary" reading material can be more challenging than adult material.

Experiential results: The results are evaluated by the Hum Sent as the prime evaluation metric for reading comprehension and the experiments allow us to explore the effect of our technique on system performance. The system applies bag of words matching approach. The Hum Sent answers are sentence that a human judged to be the best answer for each question. It compares the answer phase given by the system for a particular question to the list of sentences marked by human as possible answer for the question. For example, the correct answer to Q1 as comes into view:

Q1: Who said to the Hare "slow and steady wins the race"?

According to the human answer a key which is tagged for story.

Answer keys -story2-9.txt.wdra.xml

<txt>
<ANSQ1> The Tortoise </ANSQ1>
<ANSQ2>In the early morning</ANSQ2>
<ANSQ3>The Tortoise never for a moment stopped </ANSQ3>
<ANSQ4>Down by the wayside </ANSQ4>
<ANSQ5>He was believing that the Tortoise could never catch him, </ANSQ5>
</txt>

By comparing the system output with the answer key that provided by the human, the performance of the RC system can be assessed.

The answer key: <ANSQ1> the Tortoise </ANSQ1>

Table 3: Hum Sent accuracy on every type of questions

Question type	Accuracy hum sent (%)
Who	79.5
What	52.6
When	66.0
Where	78.5
Why	59.8
OVERALL:	**67.3**

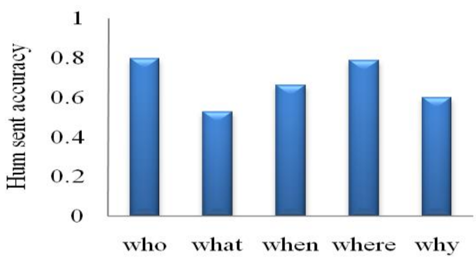

Fig. 6: Comparison of accuracy for the different question types

The system answer: <ANSQ1> The Tortoise said to the Hare, "slow and steady wins the race" </ANSQ1>.

The system answer has the correct answer "the Tortoise "even where the sentence also contained needless information.

Hum sent metric provides a binary score, it assign a suitable score for each answer. Score one point for an acceptable response and zero point otherwise. The score of the set of questions is the average of the score for each question.

Out of 75 questions, system has generated answers for all the given questions. Out of 75 answers, 51 are identified as correct answers, for the remaining 24 questions, system unable to generate the true answer. Table 3 shows the RC results for various question types in overall of data set. The result has achieved 67.3% Hum Sent accuracy of the correct answer on the question of the children stories. Performance varied for the different question types as we can see in Fig. 6, performed the best on who questions nearly 80 percent and achieved 78.5% Hum Sent accuracy for Where questions. The worst on WHAT questions, it achieved 52.6% Hum Sent accuracy.

Comparatively, small gain is illustrated for what questions from the various linguistic techniques, probably because what question has many types, most of which are not answered by a person, place or time. Why questions are by far the hardest because they need understanding of rhetorical structure and because answers tend to be whole clauses rather than phrases embedded in a context that matches the query closely.

The Error Analysis: In order to have a deep analysis of the experiment result, we analyse the cases that may occur when the BOW return incorrect answer, one of the following three cases may happen:

- The incorrect answer has a greater number of matching words than the correct answer
- The incorrect answer and the correct answer have an equal number of matching words but the incorrect answer appears earlier in the document
- So as to answer the question correctly by using bag of word approach, it requires similarity between the question and the candidate sentence. The incorrect answer happens when the true sentence has no matching words against the question

CONCLUSION

A reading comprehension system aims to understand a document in order to be able to automatically answer questions that follow it. This work identified the question information from reading comprehension, this information is to facilitate the process of extracting the answer from text. Through presented bag of words approach, it represented the information content of every sentence as a set of words. The word sets are considered to have no structure, the word order is ignored as well and contain unique elements. Reading comprehension is a useful task for developing and evaluating natural language understanding systems. Crucially, this task is neither too easy nor too hard to extract an answer from a text, as the performance of our system demonstrates. The reading comprehension is sufficiently close to information extraction applications such as ad hoc question answering, fact verification, situation tracking and document summarization, which improvements on the reading comprehension evaluations will in turn improve systems for these applications. The solutions proposed to solve the problem of reading comprehension depend not only on the overlap between questions and correct answers, but also depend on the information that when it is not easy to get from the story. The questions may solicit for information that is not provided in the passage or the information resides in different parts of the passage. It can be used in other sources and world knowledge to confirm the answer or to be provided in the case of non-existence in passage. The world knowledge can be utilized to support the answer extraction procedure.

REFERENCES

[1]- Shelza, Lalit, Priti Aggarwal, and Geetanjali Sharma, 2017. Questioning Answering Mechanism Using Natural Language Processing. International Journal of Engineering and Computer Science (IJECS), https://www.ijecs.in, Volume 6 Issue 7, July 2017, 22020-22026, DOI: 10.18535/ijecs/v6i7.20.

[2]- Du, Y.P. and M. He, 2008. Multi-strategy to improve reading comprehension ‘ cis International Conference on Computational Intelligence and Security, vol. 1, pp.86-89.

[3]- Du, Y.P., He, Ming, Ye and Naiwen, 2007. Mining the Semantic Information to Facilitate Reading Comprehension. In Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues, Ed by Huang, D.S., Heutte, L. and Loog, M. Berlin / Heidelberg: Springer .

[4]- Arivuchelvan, K. M.; and Lakahmi, K.,2017. Reading comprehension system–a review. Indian J. Sci. Res, 2017, 14.1: 83-90.

[5]- Pakray, P., 2007. Multilingual restricted domain qa system with dialogue management department of computer science and engineering . Kolkata: Jadavpur University.

[6]- Rabiah, A.K., 2008. Question Answering for Reading Comprehension Using Logical Inference Model, Malaysia: University Kebangsaan Malaysia.

[7]- Lehnert, W.G., 1978. A conceptual theory of question answering. In Proceedings of the 5th International joint Conference on Artificial Intelligence USA: Morgan Kaufmann Publishers Inc.

[8]- Hirschman, L., M. Light, E. Breck and J.D. Burger, 1999. Deep Read: a reading comprehension system. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics.

[9]- Riloff, E. and M. Thelen, 2000. A rule-based question answering system for reading comprehension tests. Proceeding of the In ANLP/NAACL 2000 Workshop on Reading comprehension tests as evaluation for computer-based language understanding sytems Washington.

[10]- Ng, H.T., L.H. Teo and J.L.P. Kwan, 2000. A Machine Learning Approach to Answering Questions for Reading Comprehension Tests. Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000).

[11]- Charniak, E., et al., 2000. Reading comprehension programs in a statistical-language-processing class. Proceeding of the In ANLP/NAACL 2000 Workshop on Reading comprehension tests. Seattle:Washington: Association for Computational Linguistics.

[12]- Xu, K. and M. Helen, 2005b. Using verb dependency matching in a reading comprehension system. Infor. Retrieval Technol., 3411: 190-201.

[13]- Du, Y. M. Helen, X. Huang and L. Wu, 2005. The use of metadata, web-derived answer patterns and passage context to improve reading comprehension performance. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada.

[14]- Christy, A. and P. Thambidurai, 2008. A Tool for Efficient information extraction with soft matching rules for text mining. J. Comput. Sci. 4: 375-381.

[15]- Porter, M.F., 1980. An algorithm for suffix stripping. 14: 130-137.

[16]- Xu, K. and M. Helen, 2005a. Design and development of a bilingual reading omprehension Corpus. Int. J. Computational Linguistics Chinese Language Processing, 10: 251-276.