International Journal of Latest Trends in Engineering and Technology
Vol.(13)Issue(2), pp.009-014
DOI: http://dx.doi.org/10.21172/1.132.02
e-ISSN:2278-621X

QUESGEN USING NLP

Pawan NGP[1], Pooja Bahuguni[2], Pooja Dattatri[3], Shilpi Kumari[4], Vikranth B.M[5]

Abstract— when people read for long hours, they seldom are able to grasp concepts and it gives them false sense of understanding it. The aim of this project is to tackle this problem by processing given text and generating applicable questions and answer. The steps followed are: 1. Candidate key sentences are selected (using Text Rank). 2. Candidate key words are selected from candidate key sentences (RAKE). 3. These selected key sentences and words are stored in the database (MongoDB) and presented to the user through chatbot interface.
Keywords— NLP, NLP toolkit, Sentence extraction, Keyword extraction, ChatBot, RAKE, TextRank

1. INTRODUCTION

Humans are the most curious by nature. Asking Questions to meet their never-ending quest for information and knowledge. For Example,teachers ask students, questions to evaluate performance of the students, pupils learn by asking questions to teachers,and even our normal life conversation consists of asking questions. Questions are the major part of countless learning interactions. However, with the advent of technology, attention spans of individuals have significantly gone down and they are not able to ask good questions. It has been noticed that when people try to read for long hours, they seldom are able to grasp concepts. But having spent some time reading gives people a false sense of understanding it. Learning and understanding consists of asking and answering questions . Apart from that, by answering questions,one can quickly check if the person who is listening/speaking has fully understood the concepts are not[2]. Nowadays e-learning has been prominently used by a lot of people in understanding concepts as it is easy to use. Amount of e-learning students are increasing day by day [3]. the aim is to generate questions automatically in e-learning systems [4] with the hope that the possible benefits from this could assist us in meeting our useful exploring needs. The goal of question generation is to generate questions according to the information in e-learning materials.

Our work is divided into two modules. The first one is used by the readers to monitor the achieved knowledge level. The second can be used by the teachers to generate questions on different bloom's levels from an e-book. The first module is implemented as an interactive ChatBot [9] [10]. The ChatBot uses quizzing as the measure to measure the learners' learning effect. One common quizzing form is the multiple-choice question [5]. A second form would be fill-in-the-blank type of question. The second module is for large scale learning materials such as e-books. Questions are generated, topic-wise and segregated by bloom's level of difficulty [11].

The paper describes principles and methods to analyze the given text content using Natural language Processing (NLP) and generate appropriate questions.

2. USAGE OF NATURAL LANGUAGE PROCESSING

2.1 What is NLP?

The application of computational techniques for analyzing and synthesizing of natural language.

2.2 NLP tools used:

- Keyword extraction
- Tokenization
- Stemmer
- Parts of speech tagging
- Lexical semantics
- Co-Occurrence matrices
- Corpus/Corpora

2.3 Stemmer:

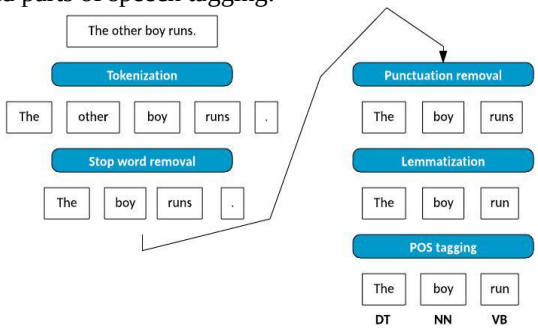
Stemming is carried out to eliminate confusion while extracting the key phrase. Stemming is enforced based on the Porter-stemmer rules. There will be no presence of duplicate entries in the extracted key phrases.

[1,2,3,4,5] Department of Computer Science and Engineering, BMS College of Engineering, Bangalore, Karnataka, India.

2.4 Tokenization:
Tokenization is a process of chopping up a defined document and a character sequence into pieces called tokens.along with that throwing away certain characters like punctuation.

2.5 Parts of speech tagging
the process of marking up a word in a text (corpus) as corresponding to a particular part of speech is known to be part-of-speech tagging (POS tagging or PoS tagging or POST) in the area of corpus linguistics, it is also known as grammatical tagging or word-category disambiguation, which is based on both its definition and its context—i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph. A common form of this is taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs [7], etc.

Here is an example of tokenization and parts of speech tagging:



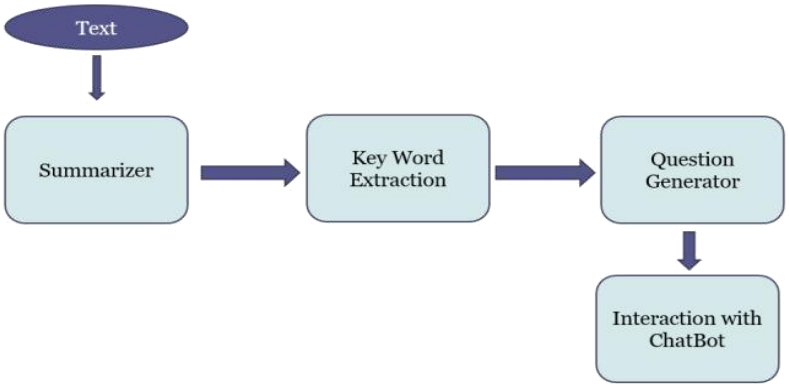
2.6 Co-occurrence matrices
Generally speaking, a co-occurrence matrix will have specific entities in rows (ER) and columns (EC) [8].This matrix presents the number of times each ER appears in the same context as each EC. As a consequence, in order to use a co-occurrence matrix, you have to define your entities and the context in which they co-occur.
In NLP, the most classic approach is to define each entity (i.e., lines and columns) as a word present in text, and the context as a sentence.Consider the following text: “Roses are red. Sky is blue.”

With the classic approach described before, we'll have the following matrix:

	Roses	are	red	Sky	is	blue
Roses	1	1	1	0	0	0
are	1	1	1	0	0	0
red	1	1	1	0	0	0
Sky	0	0	0	1	1	1
is	0	0	0	1	1	1
Blue	0	0	0	1	1	1

2.7 Corpus
Corpus means a collection of text. It could be data sets of poems by a certain poet, bodies of work by a certain author, etc. In this case, we are going to use a data set of predetermined stop words.

3. STRUCTURE OF THE PROPOSED SYSTEM



3.1 Summarizer:
Summarizer is the process to extract the most important and central ideas while ignoring irrelevant information from the text. Summarization just contains the required sentences from the text block. Summarization tries to create abstract or representative summary of the entire text.Text Summarization can be extracted for both single document and multi document [1]

3.2 KeyWord extraction [12]:
Key Word describes the main topics expressed in a document. Keyword extraction is extracting the most significant words and phrases that appear in the summarized text. Key Words are Location, Person, Thing, Possession, Animal, and Idea. Key Word is the potential key in a sentence and forms the basis for generating questions [6].

3.3 Question Generator:
Questions are generated from the summarized sentences and the key words extracted [1]. Each sentence from the summary paraphrases into a question for which the key words extracted from that sentence is the answer.

3.4 Interaction With Chatbot:
Questions that are sent to the application from the server end and are then presented to the user via a chatbot interface

4. SUMMARIZER WITH TEXTRANK
Summarization is creating short and accurate summary of a longer text document. The task contains picking a subset of a text so that the information of the subset is as close to the original text as possible.
The algorithm used for Summarization is TextRank. To compute the summary of the text without any training we are using an Unsupervised graph based algorithm called TextRank. The importance of each vertex in a graph is decided based on Graph based training algorithms.
The key idea of text rank is to provide a score for each sentence in a text ,then can take top n sentences and sort them as they appear in the text document to build an automatic summary.
What textrank does? : It finds out how similar each sentence is from all other sentences. The important sentence is the one that is most similar to all the others.
Algorithm involves following steps :
1) Concatenate all the text given in the article.
2)Extract all the sentences from the text. This can be done just by splitting the text by "." or newline character.
Find the vector representation of each sentence .
Similarity between sentences called as similarity score for each sentence is calculated and stored in a matrix.
The similarity matrix is converted into a graph with each sentences as node and similarity scores as weighted edges for sentence rank calculation.
Finally certain number of top ranked sentences form the summary.

5. KEYWORD EXTRACTION WITH RAKE
RAKE is short for Rapid Automatic Keyword Extraction algorithm, and is a domain independent keyword extraction algorithm which tries to determine key phrases in a body of text by analysing the frequency of word occurrence and its co-occurrence with other words in the text.
The advantages of the RAKE algorithm are
As Rake Algorithm can operate on documents without depending on corpus, they are domain independent; and
It's precision despite its speed, simplicity and computational efficiency.
RAKE algorithm follows the following steps to extract keywords from the given text documents:
The document will be broken into array of words (like spaces and punctuation).
Array of words will be split into sequence of contiguous words by breaking the sequence at stopwords.Now these sequences form "candidate keywords"
Now we have list of candidate key words. Next step is to calculate "score" of these . This can be calculated using the following formula.
word_score=degree (word)/ frequency (word)
Next the word scores of constituent words in each candidate keyword will be added to find the score of that "candidate keyword".
From the list of candidate keywords we take the highest one-third keywords and that forms the final list of extracted keywords from the document.
In the above formula the frequency(word) is the number of times that word occurs in the candidate keywords list.Suppose this problem is represented as a graph then degree of a word is nothing but the degree of each node in the graph. In this undirected graph , two nodes are connected if they appear in the same candidate keyword. If the degree of the is high then it

means that the word occurs more often.Hence, the degree of a word will represent how often it occurs in the candidate keywords
According to the formula shown above the "word score" is proportional to the degree of the word and it is inversely proportional to the frequency of the word. Therefore, the RAKE does not support the words that occur more often in in long candidates, and favours words that occur more often in longer candidate keywords.

6. QUESTION GENERATION

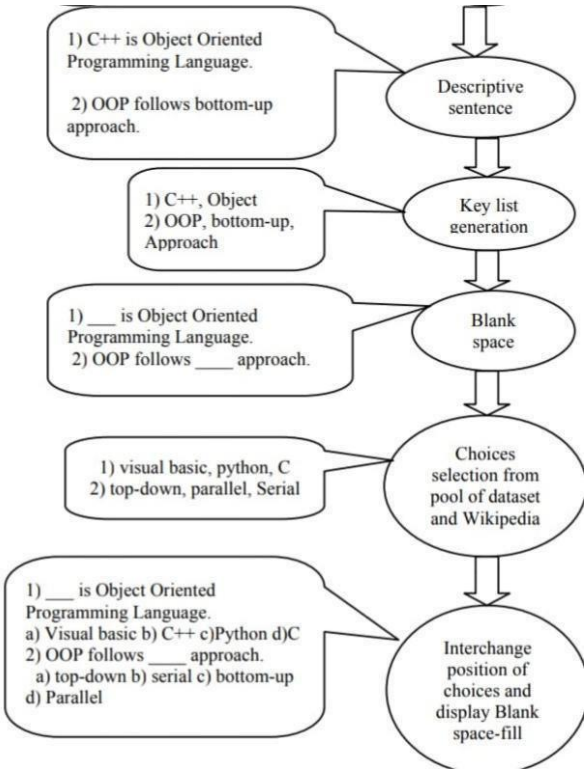
The extracted sentences and the corresponding keywords for those sentences are stored into a database in the key-value form. The database used is MongoDB which stores JSON objects in a document.
For each user a random and unique string is created when he/she opens our website and it is sent to through the url to the next page where user can input the text for which the questions should be generated.
This token is used as unique id for each text. Using this id each questions are queried from the database.
There are two collection:
Contents: token,text.
Questions : token,question,answer and status can be stored
Status: can be correct,wrong or not attempted.By default it will be "not attempted". This status will be helpful when user revists the same page.User can continue from he/she had stopped.
For example:
The sentences in a summary:
C++ is Object Oriented Programming language.
OOP follows Bottom-Up approach.
Keywords for each sentence:
C++, Object
OOP, bottom-up, Approach.
Displaying blank space for the keywords: this can be done by replacing key words by few underscores.
_____ is Object Oriented Programming language.
OOP follows _____ approach.
In MongoDB this is stored as:
Contents collection:
{_id : Mongo generated id,
token : xyz234we,
text : "C++ is a object oriented program.OOP follows, bottom-up approach"
}
Question collection:

{_id : Mongo generated id, serial_no: 1, token : xyz234we,

question : ____ is a object oriented
program,
answer : c++,
status: "unanswered",
user_answer: ""
}

{_id : Mongo generated id, serial_no:2, token : xyz234wi,

question : OOP follows ____ approach,
answer : bottom-up ,status: "unanswered",
user_answer: ""
}



7. CHATBOT INTERFACE

The figure shows the overview of the backend of the ChatBot interface. We interact with the bot, the bot reads questions from database and asks us, we answer the questions. Then the bot compares the results with the answer in the database.
The system has 3 main parts.

7.1 Node.js server

Responsible for handling the quiz processes and all things that are associated with the ChatBot interface.It Greets users, reads questions from MongoDB and asks the users, result of answered questions are compared with answers in the database, and sends data to Google Analytics. User can pause/resume anytime during the quiz.

7.2 MongoDB

The user data is stored in the MongoDB database along with the above defined structure. It is used to keep track of all the questions a user answered, in order to generate results.

7.3 Google Analytics

It generates an analytics report of how much the user answered correct questions.

8. REFERENCES

[1] Dr.P Pabitha, M.Mohana, S.Suganthi, B.Sivanandhini, “Automatic Question Generation System”, Dept of Computer Technology, MIT, Anna University, Chennai, India, 2014.

[2] Che-Hao Lee1, Tzu-Yu Chen1, Liang-Pu Chen2, Ping-Che Yang2, Richard Tzong-Han Tsai1, “Automatic Question Generation from Children’s Stories for Companion Chatbot”, 1Department of Computer Science and Information Engineering, National Central University, 2Digital Service Innovation Institute, Institute for Information Industry.

[3] A.S. Omarbekova, A.A. Sharipbay, G.T. Bekmanova, G. Sh. Nurgazinova, A.Barlybayev, “Automatic formation of questions and answers on the basis of the knowledge base”, Faculty Of Information Technologies, L.N Gumilyov Eurasian National University, Kazakhstan.

[4] Laszlo Bednarik, Laszlo Kovacs, “Automated EA-type Question Generation from Annotated Texts”, Department of Information Technology, University of Miskolc, Hungary.

[5] Yi-Chien Lin, Li-Chun Sung, Meng Chang Chen, “An Automatic Multiple-Choice Question Generation Scheme for English Adjective Understanding”, Institute of Information Science, Academia Sinica, Taiwan.

[6] Brendan WYSE, Paul PIWEK, “Generating Questions from OpenLearn study units”, Computing Department, Open University, UK.

[7] Gauri Nalawade, Department of Computer Technology, Shah and Anchor Kutchhi Engineering College, Chembur, Rekha Ramesh, Department of educational Technology, Indian Institute of Technology Bombay, “Automatic Generation of Question Paper from User Entered Specifications using a Semantically Tagged Question Repository”, Mumbai, India.

[8] Kyo-Joong Oh1, Ho-Jin Choi1, Gahgene Gweon2, Jeong Heo3, Pum-Mo Ryu3, “Paraphrase Generation Based on Lexical Knowledge and Features for a Natural Language Question Answering System”, Dept. of Computer Science, Korea Advanced Institute of Science and Technology (KAIST) 1, Dept. of Knowledge Service Engineering, Korea Advanced Institute of Science and Technology (KAIST) 2, Knowledge Mining Team, Electronics and Telecommunications Research Institute (ETRI) 3, Daejeon, Republic of Korea.

[9] Dr.P Pabitha, M.Mohana, S.Suganthi, B.Sivanandhini, “Automatic Question Generation System”, Dept of Computer Technology, MIT, Anna University, Chennai, India, 2014.
[10] Che-Hao Lee1, Tzu-Yu Chen1, Liang-Pu Chen2, Ping-Che Yang2, Richard Tzong-Han Tsai1, “Automatic Question Generation from Children’s Stories for Companion Chatbot”, 1Department of Computer Science and Information Engineering, National Central University, 2Digital Service Innovation Institute, Institute for Information Industry.
[11] A.S. Omarbekova, A.A. Sharipbay, G.T. Bekmanova, G. Sh. Nurgazinova, A.Barlybayev, “Automatic formation of questions and answers on the basis of the knowledge base”, Faculty Of Information Technologies, L.N Gumilyov Eurasian National University, Kazakhstan.
[12] Laszlo Bednarik, Laszlo Kovacs, “Automated EA-type Question Generation from Annotated Texts”, Department of Information Technology, University of Miskolcl, Hungary.
[13] Yi-Chien Lin, Li-Chun Sung, Meng Chang Chen, “An Automatic Multiple-Choice Question Generation Scheme for English Adjective Understanding”, Institute of Information Science, Academia Sinica, Taiwan.
[14] Brendan WYSE, Paul PIWEK, “Generating Questions from OpenLearn study units”, Computing Department, Open University, UK.
[15] Gauri Nalawade, Department of Computer Technology, Shah and Anchor Kutchhi Engineering College, Chembur, Rekha Ramesh, Department of educational Technology, Indian Institute of Technology Bombay, “Automatic Generation of Question Paper from User Entered Specifications using a Semantically Tagged Question Repository”, Mumbai, India.
[16] Kyo-Joong Oh1, Ho-Jin Choi1, Gahgene Gweon2, Jeong Heo3, Pum-Mo Ryu3, “Paraphrase Generation Based on Lexical Knowledge and Features for a Natural Language Question Answering System”, Dept. of Computer Science, Korea Advanced Institute of Science and Technology (KAIST) 1, Dept. of Knowledge Service Engineering, Korea Advanced Institute of Science and Technology (KAIST) 2, Knowledge Mining Team, Electronics and Telecommunications Research Institute (ETRI) 3, Daejeon, Republic of Korea.
[17] AM Rahman1, Abdullah Al Mamun1, Alma Islam2, “Programming challenges of Chatbot: Current and Future Prospective”, IEEE1, International Islamic University Chittagong2.
[18] Ashay Argal1, Siddharth Gupta1, Ajay Modi1, Pratik Pandey1, Simon Shim1, Chang Choo2, “Intelligent Travel Chatbot for Predictive
[19] Recommendation in Echo Platform”, Computer Engineering Department1 and Electrical Engineering Department2, San Jose State University, San Jose, USA.
[20] Shubham Dhainje,Renuka Chatur, Komal Borse ,Vishal Bhamare, Student, “An Automatic Question Paper Generation: Using Bloom's Taxonomy”,
[21] Dept. of Computer Engineering, R. H. Sapat College of Engineering Management Studies and Research, Nashik, Maharashtra, India.
[22] Stuart Rose, Dave Engel, Nick Cramer, Wendy Cowley, “Automatic keyword extraction from individual documents”.