2014 International Conference on Recent Trends in Information Technology

Automatic Question Generation System

Dr.P Pabitha
Dept of Computer Technology
MIT, Anna University
Chennai, India
pabithap@gmail.com

M.Mohana
Dept of Computer Technology
MIT, Anna University
Chennai, India
mona28892@gmail.com

S.Suganthi
Dept of Computer Technology
MIT, Anna University
Chennai, India
sugikushal@gmail.com

B.Sivanandhini
Dept of Computer Technology
MIT, Anna University
Chennai, India
firstproject245@gmail.com

***Abstract*- The process of automating the question generation consists of many tasks. Selecting the target content (what to ask), question type (who, why, how) and actual question generation are the major issue of Automatic Question Generation. Certain definitions retrieved is available in Wikipedia either directly or is the outcome of executing set of sub queries for each key phrase categories The problem in the existing system is that some of the definition sentences which are taken out from Wikipedia were implicit. The proposed system overcomes the problems by using Supervised Learning Approach, Naïve Bayes method. It also extends its work to use Summarization, Noun Filtering and Question Generation in the aim of generating semantically correct questions.**

Keywords— Key phrases; Supervised Machine Learning; Naïve Bayes; Stemming; Automatic Question Generation; Summarization; Noun Filtering.

I. INTRODUCTION

Researchers have proved that the humans are not skilled in generating the questions. The Automatic Question Generation is necessary one to generate good questions. Questions are indeed to have in-depth knowledge in any domain. The existing question generation system in [1] paves a way to generate questions based on extracting key phrases by using unsupervised method and then mapping it with Wikipedia article in order to get additional information about the key phrases. The obtained information may tend to be implicit.

A numerous number of question generation techniques were also proposed [2][3][4][5]. All the methods have stuck to either reading comprehension or assessing the vocabulary of the person using the tool.

II. RELATED WORKS

Stemming is the process of mitigating the derived words to its parental form. The automatically removing suffixes of word in English in specific interest of information retrieval. The algorithm [6] which is described is used for removing the suffix stripping. A theoretical and practical attributes of stemming algorithm and a new version of context- sensitive and longest matching stemming algorithm [7] for English is proposed, though developed for use in library information of transfer systems.

Academic journals and articles has a list of key words also called as phrases as. Supervised learning [8] is used to generate key phrases. The key phrases are documented by classifying the whole document into a set of positive set and Algorithm. A fuzzy set theoretic approach [9], fuzzy n-gram indexing, is used to extract n-gram keywords.

Text summarization is a process of providing a shortened summary of the existing text containing only the actual idea. Text summarization based on sentence clustering [10] consists of three steps: clustering, similarity calculation and topic recognition. Some statistical methods [11] are used to improve the content selection while doing summarization. Fuzzy logic [12] and word-net are used to get the most inter-related sentences from the document. The most significant sentence is found based on fuzzy measures and inference.

The difficulties in understanding meaning of noun compounds such as day light, green card, bee honey, human biology is studied and explores the usage of predicates [13] shows the relatedness that holds between the compounds and its nouns. For example, honey is taken by bee from blossoms. Here verb describes the relationship between honey and bee which is the semantic interpretation of noun compounds.

Automatic Question Generation (AQG) tool [14] that serves as a guide for students by providing questions which are meaningful and grammatically correct. A system to assess the vocabulary of the human knowledge is given in [15]. Questions are to be generated automatically to test the vocabulary.

III. PROPOSED SYSTEM

The Proposed system architecture is shown in Fig 1. It consists of following components: Stemmer, Key phrases Extractor, Phrase Mapper, Summarization, Noun Filtering and Question Generator. Stemmer does the job of stemming the words in the given document. Supervised Learning is a machine learning technique in which the systems are trained and produces an inferred function, which can be used for mapping new examples. The proposed system initially extracts key phrases from the documents using the Key phrase Extractor. The Extractor does the Extraction of Key phrase based on the model. Each key phrase is matched with the database. The file name which is mapped with Key phrases is extracted. If the extracted file is too large, we do summarization of the file. Nouns are filtered out from the summarized document. The Generation of question will take place using the Nouns.

978-1-4799-4989-2/14/$31.00 © 2014 IEEE

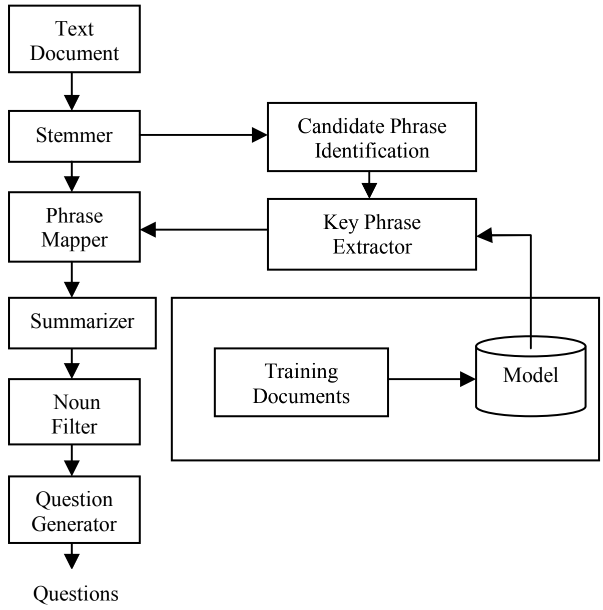

Fig. 1. Architecture of Automatic Question Generation System

A. Stemming

The process of reducing the derived words to root words is called Stemming. Stemmer has great influence in the process of key phrase extraction. A stemming algorithm reduces the words "addition", "adding", and "adder" to the root word, "add". This avoids confusion in extracting key phrases because it extracts same key phrase for many times if stemming is not applied. Therefore Stemming is applied before the key phrase extraction. The given text document is initially stemmed by the Stemmer and the stemmed file is given to the key phrase extractor for key phrase extraction.

B. Key phrase Extractor

Key phrase Extractor extract the key phrase from the stemmed file. Supervised Naive Bayes approach is applied to draw key phrases. Key phrase extraction chooses key phrases from the text itself. In this approach, the training data is used to train the key phrase extractor.

Supervised Naive Bayes' Method:

Naive Bayes works on existing available data sets and classifies them according to Bayes theorem and probability. This classifier is chosen because it is an independent feature model. Bayes method relies on the fact that one feature is independent on the other. Only limited training set is enough to calculate the parameter necessary for classification.

Bayes theorem,

$$\mathrm{P}(\mathrm{C}|\mathrm{F1},\ldots.,\mathrm{Fn}) = \frac{\mathrm{p(C)p(F1,\ldots,Fn|C)}}{\mathrm{p(F1,\ldots,Fn)}} \qquad \text{(Eq. 1)}$$

Where,

C- Dependent class variable

F1... Fn - Features

Applying this theorem to key phrase extraction,

$$1).\mathrm{p(kp1)} = \frac{\mathrm{P(kp1)}}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{P(kpi)}}$$

$$2).\mathrm{p(kp2)} = \frac{\mathrm{P(kp2)}}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{P(kpi)}}$$

Where,

P(kp1)=P(kp1)p(lang|kp1)p(sw|kp1)p(maxpl|kp1) p(minpl|kp1)p(minno|kp1)

P(kp2)=P(kp2)p(lang|kp2)p(sw|kp2)p(maxpl|kp2) p(minpl|kp2)p(minno|kp2)

kp1- key phrase 1

kp2- key phrase 2

lang- language

sw- stop word

maxpl- maximum phrase length

minpl- minimum phrase length

minno- minimum no of occurrences

C. Phrase Mapper

The Key phrases which are extracted using key phrase Extractor are used in the Phrase Mapper. Each key phrase is mapped to a file in the database. The file name which is matched with the name of the key phrase is extracted from the database.

D. Summarizer

Summarizer does the process of compressing the contents of a document in order to get concise and compact information from the document. Summarization should involve extracting only the required sentences from the text and omitting the unwanted contents. The important points of the document should not be missed. Summaries can be done for both single document as well as multi document.

E. Noun Filter

The summarized document is used to filter the Nouns from the document. Nouns are Person, Location, Possession, Thing, Animal, and Idea. Noun is the key content in a sentence and forms the basis for generating questions.

F. Question Generation

Questions are generated from the summarized document and noun filtered from it based on Questions generation rules.

IV. IMPLEMENTATION

The central aim of designing the system is to generate questions from the given document. Implementation is done by using Eclipse Kepler. The input given to the system is text document.

A. Stemming

Stemming is done to avoid confusion during extraction of key phrase. This is implemented based on the Porter stemmer rules. There will be no duplication in the extracted key phrases.

B. Key phrase Extraction

Key phrases are to be extracted from the text document. To extract key phrases, supervised Naive Bayes method is used. In supervised learning, training should be given to the system. The training dataset used here contains 25 text files and 25 key phrase files. It uses the five parameters to calculate the probability for each key phrase.

The probabilities calculated for the key phrase with the parameters is done using hash map. The <key value, feature> is the representation of hash map. In the system the hash mapping is done by <key phrase, parameter>. Finally, the key phrases with highest probabilities are chosen to generate. The key phrases generated for a text file called “a.txt” will be stored in “a.key”.

C. Phrase Mapper

The generated key phrases are used to retrieve a file named same as key phrases from the database. If the file retrieved is large, then Summarization is taken place.

D. Summarization

Summarization is a process of compressing the contents of a document in order to get concise and compact information from the document. The process used to summarize is shown in Fig.2. Summarization should involve extracting only the required sentences from the text and omitting the unwanted contents. The important points of the document should not be missed. Summaries can be done for both single document as well as multi document. The proposed summarizer first eliminates the stop words like is, of, the, etc. Then frequencies of words are calculated and arranged in descending order.

The sentences that contain words with highest frequency are extracted from the document. When the contents of the document are found to be irrelevant, the system eliminates that document and searches for a document titled with the semantic similar word of the key word. Thus the system aims at using multi document summarization during exceptional cases.

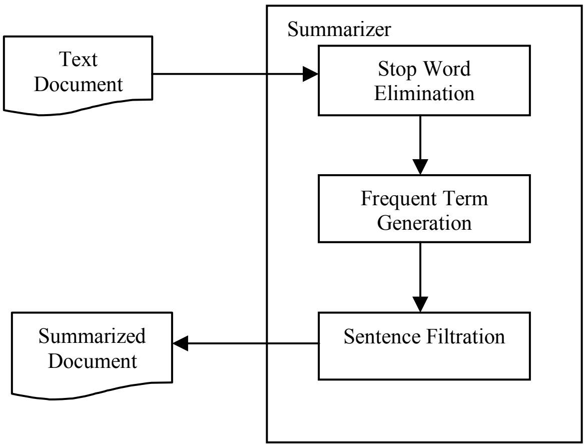

Fig. 2. Summarization Architecture

```
Algorithm: Summarization
Input: Text Document $T_D$
Output: Summarized document $S_D$
  Begin
        For each stopwords $S_W$ in $T_D$
        Begin
                Remove  $S_W$
        End
        Wordfrequency $W_F$ ($T_D$)
        Word  W= Arrangewordbyfrequency($W_F$)
        Splitsentences= getsentences($T_D$)
        $S_D$ ={}
        For each sentence S in $T_D$
        Begin
                Matchingsentence $M_S$=search(S, W)
                $S_D$.add($M_S$)
        End
          For each S in $S_D$
        Begin
                If sentence in $S_D$
                        Summary=Summary+ “  ” + S
        End
End
```

E. Noun Filtering

Once the summarization is done, the next thing is to taken out the Noun from the summarized document. Maxent tagger is used in the process of noun filtering. Nouns are filtered out based on the regular expressions which are defined to extract only the nouns.

```
Algorithm: Noun Filtering
Input: Summarized Document $S_D$ (.txt)
Output: Nouns N (.txt)
Begin
Create an object for Maxent Tagger Mt for tagging the English text
 Get $S_D$
for each line in the $S_D$
Begin
        Tags are assigned accordingly based on Mt and stored it in another file $A_f$
End
 Define Regular Expression for selecting N from the Af
Define also Special Character and specify Tag as NNP and NN
for each line in $A_f$
Begin
        N = Select words which is tagged as NNP or NN
   End
End
```

F. *Question Generation*

Questions are generated from the summarized document and noun filtered from it based on Questions generation rules. A standard template is used to generate questions by using the nouns extracted from the summarized document which is given in Fig. 3.

- Concept(Limits, Application, Benefits) – Who, What ,When, Where
- Person: Who is X?
- Definition: What is meant by X?
- Example: Give an example of X?
- Calculation: Calculate or compute X?
- Procedure: How do you perform X?
- Result: What results to X?
- Consequence: What are the consequences of X ?
- Objective: What is the objective of X?
- Reason: Why X?
- Perspective: What is your opinion about X?
- Abbreviation: What is X?

Fig. 3. Template for question generation

V. RESULTS AND DISCUSSION

Two properties are measured after summarization. They are Compression Ratio (how much shorter is than the original) and Omission Ratio (how much of the central information is retained).

$$Compression\ Ratio = \frac{Total\ No\ of\ words\ in\ summarized\ doc}{No\ \ of\ Frequent\ terms}$$

(Eq. 2)

$$Omission\ Ratio = \frac{Total\ \ No\ of\ words\ in\ original\ doc - Total\ No\ of\ words\ in\ summarized\ doc}{No\ \ of\ Frequent\ terms}$$

(Eq. 3)

The histogram which is shown in Fig 4 compares the compression ratio for the different summarized documents. Each summarized document along the X-axis has varying number of words in ascending order. Compression Ratio is represented along the Y-axis. The green, red, blue bars represents the compression ratio for 5, 10, 15 Frequent Terms respectively. From this histogram, the inference is that if the number of words relative to frequent terms in the summarized document increases, Compression Ratio also increases.

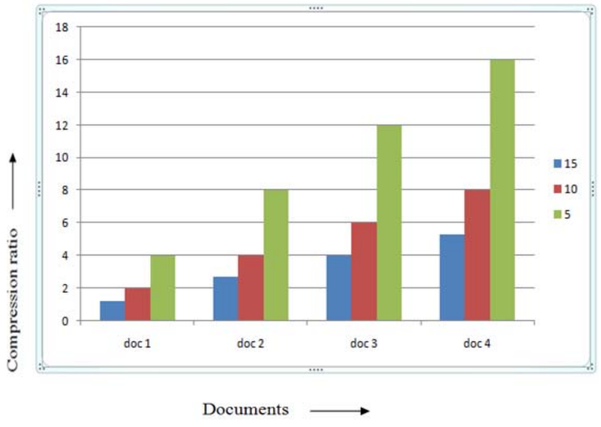

Fig 4 Compression ratio for different documents

All nouns are extracted from the summarized document. This is not useful in all cases to generate questions. It may also lead to irrelavent questions and more questions will be generated.

VI. CONCLUSION AND FUTURE WORK

Thus, the proposed system has so far implemented the stemming part using the porter algorithm. Porter stemmer utilizes suffix stripping. The System then using the stemmed file starts extracting the key phrases based on features. The system will access the database to know more information about the key phrase. Summarization is done for the larger file which is taken. The sentences are searched for the Noun and it is filtered out. Finally, questions are generated based on the Nouns.

Using phrase mapper needs the user to give the document contents which causes overhead to the system. So, instead of using phrase mapper, accessing the Google contents will reduce the burden. Extracting all nouns also hinders the system. Work should be done in extracting only the required nouns. Generating more number of questions leads to unnecessary questions. Thus for small documents, many questions will be generated.

VII. REFERENCES

[1] Ming Liu,Rafael A. Calvo, Anindito Aditomo, and Luiz Augusto Pizzato,"Using Wikipedia and Conceptual Graph Structures to Generate Questions for Academic Writing Support", IEEE Transactions on Learning Technologies, Vol 5, NO. 3, pp.251-263, july-september - 2012.

[2] R. Mitkov and L.A. Ha, "Computer-Aided Generation of Multiple-Choice Tests," Proc. HLT-NAACL Workshop Building Educational Applications Using Natural Language Processing, pp. 17- 22, 2003.

[3] H.Kunichika, et al,"Automaitc Question Generation Methods for Intelligent English Learning Systems and its Evaluation,"Proc.Int'lConf.Computersin Education, pp.1117-1124, 2002.

[4] M. Heilman and N.A. Smith, "Good Question! Statistical Ranking for Question Generation," Proc. Ann. Conf. North Am. Chapter of the Assoc. for Computational Linguistics - Human Language Technologies, pp 609-617, 2010.

[5] J. Mostow and W. Chen, "Generating Instruction Automatically for the Reading Strategy of Self-Questioning," Proc. Int'l Conf. Artificial Intelligence in Education, pp. 465-472, 2009.

[6] M.F. Porter, "An Algorithm for Suffix Stripping", Program, 14(3), P: 130-137, 1980.

[7] S. Santhana Megala, Dr.A.Kavitha, Dr.A.Marimuthu "Improvised Stemming Algorithm – TWIG" volume 3,issue 7, july- 2013.

[8] P.D. Turney, "Learning Algorithms for Key phrase Extraction," Information Retrieval, vol. 2, pp. 303-336, October 4, 1999..

[9] Bidyut Das, Subhajit Pal, Suman Kr. Mondal, Dipankar Dalui, Saikat Kumar Shome, "Automatic Keyword Extraction From any Text Document Using N-gram Rigid Collocation" ,International Journal of Soft Computing and Engineering, ISSN :2231-2307, Vol 3, issue 2, pp.1-5. Xia, Zhihua, et al. "An Efficient and Privacy-Preserving Semantic Multi-Keyword Ranked Search over Encrypted Cloud Data." (2013).

[10] Zhang Pei-ying , Li cun He, "Automatic text summarization based on sentences clustering and extraction", IEEE 2009.

[11] SemanticWeb , http://en.wikipedia.org/wiki/semantic_web

[12] F. Kyoomarsi, H. Khosravi, E. Eslami And M. Davoudi "Extraction-Based Text Summarization Using Fuzzy Analysis" ,Iranian Journal of Fuzzy Systems Vol. 7, No. 3, pp. 15-32 , March 2010.

[13] Preslav I. Nakov, Marti A. Hearst, " Semantic Interpretation of Noun Compounds Using Verbal and Other Paraphrases" , ACM Transactions on Speech and Language Processing, Vol.10,No.3,Article 13, July 2013.

[14] Ming Liu, Rafael A. Calvo, "An Automatic Question Generation Tool for Supporting Sourcing and Integration in Students' Essays" , Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009.

[15] Jonathan C. Brown Gwen A. Frishkoff Maxine Eskenazi, "Automatic Question Generation for Vocabulary Assessment" Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, October 2005.