International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 4, Issue 4, July - August 2015 **ISSN 2278-6856**

Automatic / Smart Question Generation System for Academic Purpose

[1]Deepshree S. Vibhandik , [2]Rucha C. Samant

[1]P.G.Student Dept. of Computer Engg. G.E.S's. R.H.Sapat College of Engg. Nashik,
Affiliated to Savitribai Phule Pune University

[2]Asst. Prof. Dept. of Computer Engg. G.E.S's. R.H.Sapat College of Engg. Nashik,
Affiliated to Savitribai Phule Pune University

Abstract

In this paper, we present a novel approach designed for automatic / smart question generation system for academic purpose. Preparation of highly standard questions which encourage student's thinking ability is very challengeable task that need to be performed by the academicians. So, there arises a need for automatic question generation. Basically, Question Generation from text is a Natural Language Processing task combining Natural Language Understanding and Natural Language Generation. In this paper, we consider Automatic Question Generation system that generates specific trigger questions and multiple choice questions from students literature review paper. To facilitate generation of specific trigger question, the system extracts key concepts from student's paper using Lingo algorithm. Also, to bring out generation of multiple choice questions the system pulling out abbreviations from student's review paper using regular expression pattern matching technique.

Index Terms: Automatic Question Generation, Natural Language Processing, Lingo Algorithm.

1. INTRODUCTION

In the last years, we have renewed interest in the development of automated systems. Natural Language Processing is a technique that explores how computers can be used to understand and manipulate natural language text. Natural Language Processing is a field of interest where many computer based technology tend to move us from manual system to automated system. The development of several kinds of automated system affected in various areas and Education field is one of them. In the education field, there are many examples of automated systems viz. Automatic Question Generator, Domain Module Generator, Intelligent Tutor, Learning Management System etc. In this paper, we are going to proposed Automatic Question Generation system.

Automatic Question Generation (AQG) is a challenging task which involves natural language understanding and generation. Three major aspects of AQG have been addressed in the study: selection of the target content (what to ask about), selection of the question types (e.g., Who, Why, Yes/No) and construction of the actual questions [1]. Most of the research, in this field focused on different application domain like reading comprehension, vocabulary assessment etc. Various Question Taxonomies have been proposed according to different application domains. A well-known taxonomy was proposed by Graesser and Pearson [2].

Rest of the paper is organized as follows. Section II presents related research of AQG systems. Section III highlights system architecture. Section IV gives results and discussion and finally section V presents conclusion as well as future scope of the system.

2. RELATED RESEARCH

2.1 Automatic Question Generation Systems

Many Automatic Question Generation systems have been evolved for different domain like reading comprehension, vocabulary assessment etc. There are some approaches that have been developed for generating factual questions from reading material. In [3], Hidenobu Kunichika and Tomoki Katyama employs highly advanced NLP technology based system. They have developed AQG system to generate factual questions from the contents of story. The system employed the five methods for question generation: to ask about the content of one sentence, to use synonyms or antonyms, to use modifiers appeared in plural sentences, to ask the contents represented by plural sentences by using a relative pronoun, and to ask relationship of time and space. Husam Ali and Yllias Chali has proposed AQG from sentences in [4]. This system is able to generate only factual questions which are content based, so it is useful for building an automated trainer for learners to ask better questions, also for reading comprehension and vocabulary assessment.

Also, some research has been done for preparing multiple choice questions. In [5] Ayako Hoshino suggest the MCFBQ system to generate MCQ to test the students knowledge about the vocabulary and grammar. The most recent research in the field of Automatic Question Generation is mainly worked for preparing conceptual or logical questions. Ming Liu and R.A Calvo described and evaluated a G-Ask AQG system that focuses on citation sentences in a literature review [6]. The system proposed here will generate conceptual trigger questions and multiple choice questions from students' review paper intended to prompt student to focus on key concepts of their area of study.

2.2 Key Phrase Extraction Techniques

Key phrase is a meaningful and significant expression consisting of one or more words in a document. Key phrase provide important information about content of document. Key phrase extraction is an important research

topic in the NLP and IR field. Key phrase extraction is the process obtaining the key phrases which are available in the body of text document.

Two popular single document key phrase extraction algorithms are: GenEx by Turney[7] and KEA by Frank[8]. GenEx uses machine learning to extract key phrases from individual documents, which employs genetic algorithm to tune its parameter. The KEA (Key Extraction Algorithm) automatically extracts key phrases from the full text of documents. The set of all candidate phrases in a document are identified using rudimentary lexical processing. Features are computed for each candidate, and machine learning is used to generate a classifier that determines which candidate should be assigned as key phrases. Mani and Bloedorn suggested a method for multi-document summarization based on graph representation based on concepts in the text. CorePhrase [9] algorithm works by extracting a list of candidate key phrases by intersecting documents using graph based model of phrases in the document.

The Lingo [10] algorithm is unsupervised approach used for key phrase extraction. It is based on Singular value Decomposition. The proposed system is mainly based on Lingo algorithm to perform the task of key phrase extraction.

2.3 Knowledge Base Wikipedia

Wikipedia can be considered as resource that includes knowledge about named entity and domain specific terms. It is very large knowledge base as it covers multiple domains with more than three million pages. Wikipedia articles are easily available as it is free encyclopedia. Researcher can access Wikipedia articles in XML dumps using JWPL (Java Wikipedia Library) which is an open source Java based APIs. JWPL can parse Wikipedia articles with Wikipedia Markup Language and convert them into relational database. Midiawiki is a free and open-source wiki application. MediaWiki has an extensible web API (application programming interface) that provides direct, high-level access to the data contained in the MediaWiki databases. In this study, wikipedia article gives adequate support for key phrase classification and also for conceptual graph construction.

2.4 Conceptual Graph

Conceptual graph is a way of knowledge representation means a graph in which nodes are concepts and edges indicate the relationship between them [11]. Here, a conceptual graph contains the triple, which is the combination of white node, black node and the relationship between white node and black node. The basic idea of graph construction is to create triple by assigning key phrase as black node and target sentence as white node. Conceptual graphs are not intended as a means of storing data but as means of describing data and interrelationships.

3. SYSTEM ARCHITECTURE

The Fig.1 shows the system architecture of proposed system. This architecture comprises seven different stages as Key Phrase Extraction, Key Phrase Classification, Conceptual Graph Construction and Conceptual Question Generation, Abbreviation Extraction, Possible Option Extraction and MCQ generation. Initially the system has a corpus of different literature review papers. By using the Lingo algorithm [8] key phrases extracted from this corpus. Then extracted key phrases matched with Wikipedia article using JWPL and then key phrase are classified into five concept categories Research Field, Technology, System, Term and other. Using the contents of Wikipedia articles Conceptual Graph [9] is constructed for each extracted key phrase using the Trgex Rules . Finally generated triples from conceptual graph matched with different questions templates and then conceptual questions will be generated. In the same way, abbreviations from the literature review paper will be extracted. According to the extracted abbreviations, possible options are need to generate and finally MCQ will be generated.

3.1 Key Phrase Extraction

Key phrase can be defined as a phrase of one to three words to capture the main topic from the document. Key phrase implies two features phraseness and informativeness [1]. Phraseness is a abstract notion which describe the degree to which a given word sequence is considered to be a phrase. Informativeness refers to how well a phrase captures or illustrate the key ideas in a document. Key phrase list is short list of phrases that captures the main topics discussed in a given document. The

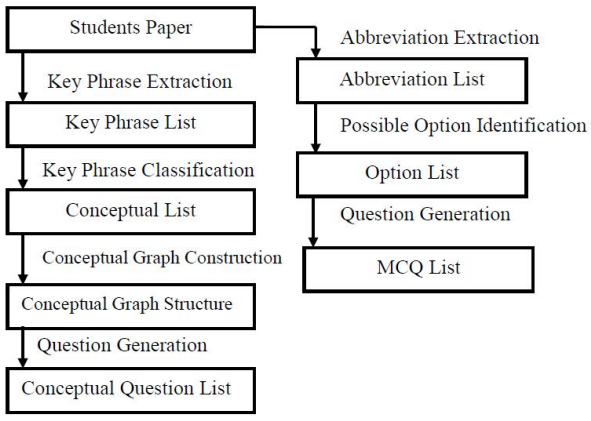

Fig.1: System Architecture

key phrase extraction task is performed here by using Lingo algorithm [10]. The general idea behind unsupervised Lingo algorithm is first to find meaningful descriptions of clusters and then based on descriptions determine their content. Along with key phrase extraction using Lingo algorithm, the system also extracts the important phrase by using regular expression matching technique.

3.2 Key Phrase Classification

The key phrases that extracted using Lingo algorithm may be relevant or not relevant. In order to find out exact phrases the extracted key phrases are linked to wikipedia articles using MediaWiki. Mediawiki provide programmatic access to wikipedia database. If a key

phrase matches with a title of Wikipedia article, then content of that article will be retrieved. Key phrases that cannot be matched to any Wikipedia article are discarded. From each retrieved Wikipedia page, the system will identify the definition sentence as one type of the conceptual taxonomy by using pattern matching rules. Key phrases that are extracted in previous phase are now classified into following conceptual categories: Research Field, Technology, System, Algorithm, Term, and Other.

3.3 Conceptual Graph Construction

After the key phrases are classified, conceptual graph structures are created based on content of Wikipedia article. The basic idea of graph construction is to create triple by assigning key phrase as black node and target sentence as white node. The relation include is a (definition of the concept), has-limitation (drawback of the concept), has-strength (advantages of the concept) and apply-to (application of the concept) etc.

The system uses the various pattern matching rules to extract the informative sentences from wikipedia article for each key phrase. From the Wikipedia article, each section is checked against the cue phrase as given in Table 1 [1]. A target section in a page is identified by using the cue phrases matched with the section title. Each cue phrase belongs to a relation type. Table 1 gives the idea about how to establish edge relation between key phrase and sentence.

Table 1: Cue phrases in Edge Relation Category

Edge Relation	**Cue Phrase**
Has-Limitation	limitation, issue, challenge, problem, drawback, disadvantage
Has-Strength	benefit, advantage, overcome
Apply-to	application, use, apply
Include-Technology	technologies, approaches, technique, algorithm, method

3.4 Question Generation

The trigger questions are generated based on the triples in the conceptual graph. So that the system need to the generated triples with predefined question templates. Each pattern matching rule contains triple and question template. By using various kind regular expression the system generate different patterns to form question template. The questions are generated based on the triples in conceptual graph. Question templates are shown in Table 2.

Table 2: Question Generation Rule

Triple	**Question Template**
Is-A (Research Field, Sentence)	Sentence + What impact would the proposed project have on the field of + ConceptName + ?
Has-Strength	Sentence + How do you see + ConceptName + being applied in your project?
Has-Limitation (Concept, Sentence)	Sentence + How do you address these issues in your project?
Has-Strength (Concept, Sentence)	Sentence + Have you considered this strength in relation to your project?
Apply-to (Concept, Sentence)	Do you know that + Sentence + How is this application relevant to your project?

3.5 Abbreviation Extraction

The task of implementation of abbreviation extraction basically involves sentence selection and abbreviation selection. The sentence selection step is the process of determination of such sentences that are comprising of abbreviation. The process of selecting abbreviation requires to find out successive word sequences in the sentence having first letter of each word is capitalized followed by parenthesized sequence of capitalized letters of that words. To perform this task, The system again uses various pattern matching rules.

3.6 Possible Option Identification

For each extracted abbreviation from document and to frame a MCQ based on abbreviation, it needs to discover possible options from the given document. Here, possible option means to find out all nearest word sequence from document in which each word start with corresponding letter of the extracted abbreviation. It is the process of finding out possible expansion or full form for the acronym.

4. IMPLEMENTATION SETUP

We have implemented our system in client-server architecture. This system is built in java. For server side setup Apache tomcat is used. The system is built in eclipse IDE with jdk-7 and tomcat-7. For database storage system make use of mysql 5.6.

The system has 3 users. Admin manages the complete system. The data corpus is uploaded by the administrator and system perform document clustering and extracts key phrases using Lingo algorithm. User uploads the input document for which he needs to generate question. System determines appropriate phrases from input document using Lingo algorithm and regular expression. From extracted phrases system generates question and return to the user. The extracted key phrases and generated questions are stored in database. Examiner is one more user of the system that examines the system generated questions according to different quality measure that are stored in database and give ranking to the questions. The quality measures that are available in database are correctness, clarity, usefulness etc. Using question assessment the administrator evaluates the system performance.

5. RESULTS AND EVALUATION

Initially, from the Administrator side the data corpus of 24 papers from different domain viz. data mining, network security, natural language processing etc. feed to the system. For the given corpus of 24 papers the system generates 9 clusters and extracts 248 key phrases.

From the User side 17 papers were uploaded to the system in one by one fashion. For the uploaded 17 papers the system extracts total 363 key phrases

from which 212 phrases ate relevant to the context. Table 3 shows classification result on 212 key phrases.

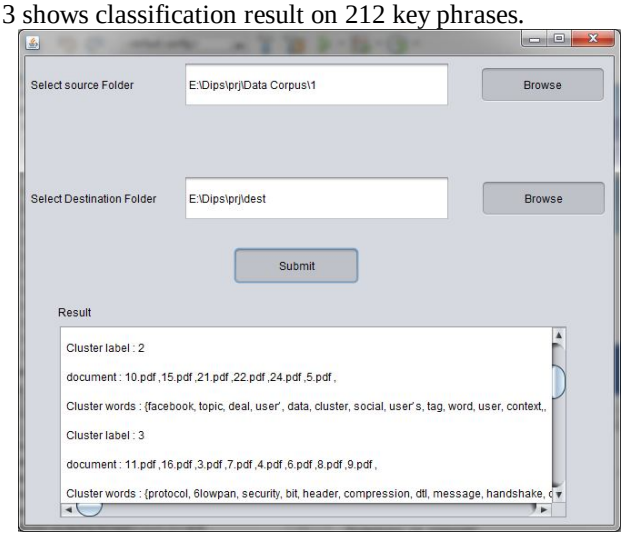

Fig.2: Clustering of Documents

By matching the predefined question template with 212 phrases the system generates 212 trigger conceptual questions. Also the system extracts 224 abbreviated forms and generates MCQs by finding out possible options according to the abbreviation. Out of 224 generated MCQs the expert identifies 79 MCQs that are corect.

Table 3: Key Phrase Classification Matrix

Predicted KP / True KP	Algo	Sys.	Tech.	Proto.	RF
Algorithm	72				
System		54			
Technology			76		
Protocol				0	
Reserach Field					16
Total	102	87	141	2	32

The computer generated question are evaluated manually by using the following quality measures.

QM1: Correctness

QM2: Clarity

QM3: Useful for learning important concepts

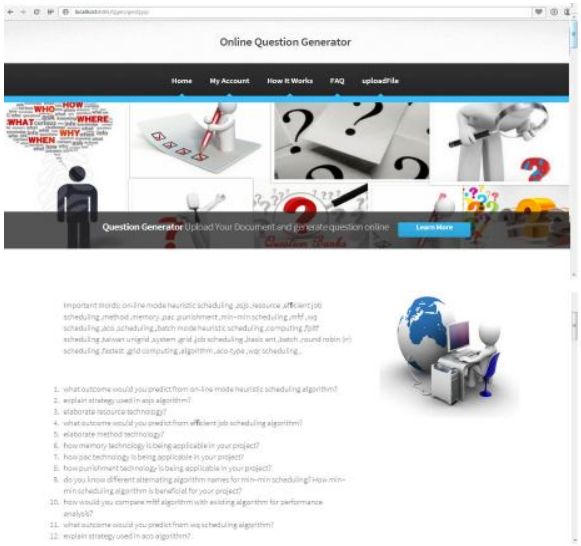

Fig.3: List of Generated Questions

6. CONCLUSION

In this paper, we presented an smart question generation tool. The tool is based on conceptual graph structures constructed from Wikipedia articles. The questions are intended to prompt students to reflect on key concepts in their area of study. The headings of Wikipedia sections were found to be useful for identifying the target sentence or phrase list in the content. Using information from section headings in Wikipedia reduces the computational cost needed to scan and classify each sentence in a Wikipedia article. The drawback of using section headings is that some target sentences or phrase lists cannot be extracted because they appeared in sections whose heading does not contain the cue phrases. One of limitation of our current AQG system is that it is domain dependent, because we only defined a limited number of concepts for generating questions (e.g., Research Field, Technology, Algorithm, System, Protocol and Other). Although these concepts are common in the science disciplines, they may not be directly suitable for other humanity or social science disciplines (e.g., English literature). The question generated by the system for technical study are useful for students while preparing for their examination. The system is also helpful for teachers to ask questions to students to judge students knowledge and understanding of their subject study.

References

[1] Ming Liu, Rafael A. Calvo, "Using Wikipedia and Conceptual Graph Structures to Generate Questions for Academic Writing Support", IEEE transaction on learning technology, vol. 5, pp. 251-263, July-Sept. 2012.

[2] Ming Liu and Rafael A. Calvo, "Question Taxonomy and Implications for Automatic Question Generation", Proc. Intl Conf. Artificial Intelligence in Education, pp. 504-506, 2011.

[3] Hidenou Kunichika, Tomoki Katayama, and Tsukusa Hirashima and Akira Takeuchi, "Automated Question Generation Methods for Intelligent English Learning

Systems and its Evaluation", Proc. Intl Conf. Computers in Education, pp. 1117-1124, 2002.

[4] Husam Ali, Yllias Chali, Sadid A. Hasan, "Automatic Question Generation from Senetences", TALN 2010, Montreal, 19-23, July 2010.

[5] Ayako Hoshino, Lunan Huan, and Hiroshi Nakagawa, "A framework for automatic generation of grammar and vocabulary questions", Proc. of WorldCALL 2008, Fukuoka, Japan, August 2008.

[6] M. Liu, R.A. Calvo, and V. Rus, "Automatic Question Generation for Literature Review Writing Support", Proc. 10th Intl Conf. Intelligent Tutorial Systems, pp. 45-54, 2010.

[7] Turney, P.D.: "Learning algorithms for Keyphrase Extraction", Information Retrieval 2 (2000) 303–336.

[8] Frank, E., Paynter, G., Witten, I., Gutwin, C., Nevill-Manning, C.: "Domain-specific Keyphrase Extraction", In: 16th International Joint Conference on Artificial Intelligence (IJCAI-99), Stockholm, Sweden (1999) 668–673.

[9] Khaled M., Diego N. Matute, "CorePhrase Extraction for Document Clustering".

[10] S. Osinski, J. Stefanowski, and D. Weiss, "Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition", Proc. Int'l Conf. Intelligent Information Systems, 2004.

[11] Hadi Amiri, Abolfazl AleAhmad, Masoud Rahgozar, "Keyword Suggestion Using Conceptual Graph Construction from Wikipedia Rich Documents", Int'l Conf. on Information and Knowledge Engineering (IKE'08), Las Vegas, USA, July 14-17, 2008.