Implementation of Automatic Question Paper Generator System

Aleena Susan Mathew[1], Vidya. N[2]

[1]PG Scholar, Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India
[2]Assistant Professor, Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India

---

Abstract – *Student assessment is a crucial part of teaching and is done through the process of examinations and preparation of exam question papers has consistently been a matter of interest. Here proposing an Automatic Question Paper Generating System which is fast, streamlined, randomized and secure. The proposed automatic question paper generating system having the input as text or document or pdf file. In proposed system, stop words are removed and NLP is used. Key phrase extraction is done by using TF – IDF algorithm and checking the wiki presence of terms. Generate triplets for question paper generation and also conducting input clarity checking using some linguistic rules by using WordNet tool.*

***Key Words*: Question Paper Generation, stop words, NLP, Key phrase Extraction, wiki presence, Triplets Generation, Input Clarity Checking.**

1. INTRODUCTION

Today, education is the most important way of achieving success. Proper examinations help students to improve their quality. So, having a proper examination paper is very necessary. The traditional way of generating question paper has been manual. Here proposing an Automatic Question Paper Generator System. Every task performed by this system is automated so that storage space, bias and security is not a concern anymore. The proposed system can be helpful to many educational institutes.

Automatic Question Generation (AQG) involves natural language understanding and generation. Three major aspects of AQG have been addressed: 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. In this system, proposing a novel approach to address three key challenges of automatic trigger question generation for supporting writing. The first challenge concerns the identification of key/central concepts from the potentially many concepts that are contained in an academic paper. The second is related to the system's lack of knowledge about the domain discussed in an academic paper. And the third is how to evaluate whether the questions generated by the system are considered useful by authors/students.

To address the first challenge of identifying key concepts, the system uses an unsupervised extraction algorithm to extract key phrases from an academic paper. The system then classifies each key phrase based on a Wikipedia article matched with the key phrase by using a rule-based approach.

The key phrases can belong to one of the following five concepts adapted from a conceptual taxonomy proposed by Lehnert et al. [1] [2].

A. Research Field:

The key phrase is about a research field. For example, "Social sciences are the fields of academic scholarship that study society."

B. Technology:

The key phrase is related to technology/method/model/algorithm/protocol, e.g., "SOAP, originally defined as Simple Object Access Protocol, is a protocol specification for exchanging structured information in the implementation of web services in computer networks."

C. System:

The key phrase refers to a software system or hardware device. For example, "An image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images."

D. Term:

The key phrase describes a technical term. For example, "The term cognitive load is used in cognitive psychology to illustrate the load related to the executive control of WM."

To address the second challenge, Wikipedia was used as a domain knowledge base. Knowledge from a single article is used to build conceptual graphs used to generate questions.

The third challenge of evaluating the input quality of generated questions is addressed by conducting a Linguistic rule by WordNet tool.

2. RELATED WORKS

2.1 Writing Support Systems

Haswell [3] reviewed systems for automated feedback tracing back to the 1950s. These systems have focused on assessment of end products, and less on providing formative feedback [4], [5]. The Writer Workshop [6] and Editor [7] both focus on grammar and style. Sourcer's Apprentice Intelligent Feedback system (SAIF) [8] is a computer assisted essay writing tool used to detect plagiarism, uncited quotations, lack of citations, and limited content integration problems. It uses a rule-based approach and Latent Semantic Analysis (LSA), a technique used to measure the semantic similarity of texts [9]. SaK, a writing tutoring system [10] developed at the University of Memphis, assesses student compositions. It uses multiple animated characters to provide different aspects of feedback, such as coherence, purpose, topic, and overall quality. Like SaK, a number of automated essay assessment tools or scoring systems [11] has been built based LSA. For example, Apex [12] uses LSA to assess student essays on topic coverage, discourse structure, and coherence.

2.2 Automatic Question Paper Generation Systems

One of the first automatic QG systems proposed for supporting novices to learn English was AUTOQUEST [13]. Kunichika et al. [12] who proposed a question generation method based on both syntactic and semantic information (Space, Time, and Agent) so that it can generate more question types (Where, When, and Who). More recently, Mostow and Chen [13] proposed an approach to generate deep questions based on a situation model. It can generate what, how and why questions. Several approaches have been proposed for automatic multiple-choice QG [14] from reading materials. Coniam [15] removed every nth-word in the text to be a test item, and distractors were identified by choosing the same part of speech (e.g., noun, verb, or adjective) and similar word frequency to a tagged corpus. Mitkov and Ha [16] removed Key Terms, which are noun phrases with a frequency over a certain threshold.

2.3 Key phrase Extraction Techniques

Key phrases provide important information about the content of a document. Two approaches for the automatic extraction of key phrases have been studied. Supervised techniques require labeled data to train the system and tend to be more accurate but also more restricted. Unsupervised techniques do not require training sets and tend to be applicable to wider knowledge domains, but they are also less accurate.

Turney [17] introduced a system for key phrase extraction called GenEx, which is based on a set of parameterized heuristic rules tuned by a genetic algorithm. Frank et al. [18] applied a Naive Bayes classifier for key phrase extraction on the same data used by Turney, which improved the results. Both GenEx and the Naive Bayes classifier are examples of supervised approaches to key phrase extraction. In general, supervised approaches require an annotated training set, which is often not practical.

To eliminate the need for training data, several authors have developed unsupervised approaches to key phrase extraction. Barker and Cornacchia [19] ranked noun phrases extracted from a document by using simple heuristics based on the length and the frequency of their head noun. Bracewell et al. [20] clustered terms which share the same noun term from a list of extracted noun phrases. Another widely adopted unsupervised approach for key phrase extraction is to use graph-based ranking methods. Mihalcea and Tarau [21] represented a document as a term graph based on term relatedness; a graph based ranking algorithm is then used to assign importance scores to each term.

The Lingo algorithm [22], another unsupervised approach, is generally used for clustering web search results. It is based on singular value decomposition (SVD). The cluster-label induction phrase in Lingo involves following steps. First, a term-document matrix A is built from the input documents. Second, the term-document matrix is broken into three matrix (U, S, and V) by performing SVD, such that A ¼ USVT. Third, k column vectors of U are extracted. Each column vector refers to a cluster or latent concept. Fourth, the semantic similarities between latent concepts and single words\phrases are calculated by using classic cosine distances, M ¼ UT k P, where each column vector of matrix P represents a single word or phrase. Last, we choose the most similar single word or phrase as the concept label by finding the largest value in each row of matrix M. Rows of the matrix M represent latent concepts, its columns represent phrases or single words, and individual values are the cosine similarities.

3. Proposed System

Implementation of Automatic Question Paper Generator System consisting different steps:
A. Converting PDF/DOC file into Text file
B. Preprocessing
C. Natural Language Processing
D. Key Phrase Extraction
E. Checking the Wiki-Presence
F. Constructing Concept List
G. Triplets Generation
H. Question Paper Generation
I. Input Clarity Checking

A. Converting PDF/DOC file into Text file

The input which we are given to the system is in the form of pdf file or doc file must convert into text file. Because the system can only read the text file.

B. Preprocessing

In the preprocessing stage, all input documents are split into sentences and performs stop word removal. Stop words are natural language words which have very little meaning, such as "and", "the", "a", "an", and similar words.

C. Natural Language Processing

In this proposed system, NLP is used to breakup text into tokens using tokenizer and also use to generate keyword topic tags from a document using **LDA** (Latent Dirichlet Allocation) based on mallet tool, which determines the most relevant words from a document.

D. Key Phrase Extraction

Key phrases, key terms, key segments or just keywords are the terminology which is used for defining the terms that represent the most relevant information contained in the document. Many applications including text summarization, indexing, and characterization use key phrase extraction. In the proposed system, TF-IDF Algorithm is used to find the scores of a term in a document.

TF*IDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF). The TF*IDF weight of the term is the product of the TF and IDF scores of a term. It checks how relevant the keyword is throughout the web, which is referred to as corpus.

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)…………………1

IDF(t) = log_e (Total number of documents / Number of documents with term t in it)………………………………………..2

Value = TF(t)* IDF(t)……………………………………………………3

E. Checking the Wiki-presence

Extracted key phrases are linked to Wikipedia articles using JWPL (Java Wikipedia Library). If a key phrase matches the title of a Wikipedia article, that article will be retrieved. Key phrases that cannot be matched to any Wikipedia article are discarded. The definition sentence refers to the first sentence that describes the key phrase in the first section of a Wikipedia page. From each retrieved Wikipedia page, identifying the definition sentence as one type of the conceptual taxonomy by using Regex expression rules. Using this definition to classify the associated key phrase.

F. Constructing Concept List

After the key phrases are classified, conceptual graph structures are then created based on information such as the section headings and the section content in a Wikipedia article.

G. Triplets Generation

The conceptual graph structure can be considered as a list of triples. A triple contains a black node, a white node, and their relation. The basic idea of this graph construction algorithm is to create each triple by setting the key phrase as the black node and a target sentence (or phrase list extracted from the target section) as the white node. 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. A target sentence containing the page title or its abbreviation is retrieved and classified as one of these relations (Definition, Apply-to, Has-Limitation, Has-Strength, IS-A, Include-Technology) by using Regex expression rules.

H. Question Paper Generation

The questions are generated based on the triples in the conceptual graph. Here using some question generation rules. Each rule contains a triple and a question template.

I. Input Clarity Checking

Here, conducting input clarity checking using some linguistic rules using tool like WordNet. Some linguistic rules are Grammatical rule, Morphological rule, Semantic rule, Syntactic rule, etc.

4. CONCLUSION

Proposed an Automatic Question Paper Generating System which is fast, streamlined, randomized and secure. Every task performed by this system is automated so that storage space, bias, and security is not concern anymore. The proposed system is very helpful for many educational intitutes. In the future, more advanced concepts will be explored for improving the efficiency of the project. Also consider improving the current question paper generation method through more advanced applications such as image identification and so on.

REFERENCES

[1] Ming Liu, Rafael A. Calvo, Senior Member, IEEE, Anindito Aditomo, and Luiz Augusto Pizzato, Using Wikipedia and conceptual graph sructures to generate questions for academic writing support", IEEE Transactions On Learning Technologies, Vol. 5, No. 3, July – September 2012.

[2] W. Lehnert, C. Cardie, and E. Riloff, " Analyzing Research Papers Using Citation Sentences", Proc. 12th Ann. Conf. Cognitive Science Soc., pp. 511 – 518, 1990.

[3] R. Haswell, "The complexities of Responding to Student Writing; or, Looking for Shortcuts via the Road of Excess," Across the Disciplines, vol. 3, 2006.

[4] M.D. Shermis and J.C. Burstein, Automated Essay Scoring: A CrossDisciplinary Perspective, vol. 16. MIT, 2003.

[5] J. Anderson, Mechanically Inclined: Building Grammar, Usage, and Style into Writer's Workshop. Stenhouse, 2005.

[6] E.C. Thiesmeyer and J.E. Theismeyer, Editor: A System for Checking Usage, Mechanics, Vocabulary, and Structure. Modern Language Assoc., 1990.

[7] M.A. Britt, P. Wiemer-Hastings, A.A. Larson, and C.A. Perfetti, "Using Intelligent Feedback to Improve Sourcing and Integration in Students' Essays," Int'l J. Artificial Intelligence, vol. 14, pp. 359374, 2004.

[8] T.K. Landauer, D.S. McNamara, S. Dennis, and W. Kintsch, Handbook of Latent Semantic Analysis. Lawrence Erlbaum, 2007.

[9] P. Wiemer-Hastings and A.C. Graesser, "Select-a-Kibitzer: A Computer Tool that Gives Meaningful Feedback on Student Compositions," Interactive Learning Environments, vol. 8, pp. 149169, 2000.

[10] D. Wade-Stein and E. Kintsch, "Summary Street: Interactive Computer Support for Writing," Cognition and Instruction, vol. 22, pp. 333-362, 2004.

[11] T.K. Landauer, D. Laham, and P.W. Foltz, "The Intelligent Essay Assessor," IEEE Intelligent Systems, vol. 15, no. 5, pp. 27-31, Sept./ Oct. 2000.

[12] B. Lemaire and P. Dessus, "A System to Assess the Semantic Content of Student Essays," J. Educational Computing Research, vol. 24, pp. 305-320, 2001.

[13] J. Villalon, P. Kearney, R.A. Calvo, and P. Reimann, "Glosser: Enhanced Feedback for Student Writing Tasks," Proc. IEEE Eighth Int'l Conf. Advanced Learning Technologies (ICALT '08), pp. 454-458, 2008.

[14] R.A. Calvo and R.A. Ellis, "Students' Conceptions of Tutor and Automated Feedback in Professional Writing," J. Eng. Education, pp. 427-438, 2010.

[15] J.H. Wolfe, "Automatic Question Generation from Text - An Aid to Independent Study," SIGCUE Outlook, vol. 10, pp. 104-112, 1976.

[16] D. Coniam, "A Preliminary Inquiry into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests," CALICO J., vol. 14, no. 2, pp. 15-33, 1997.

[17] P.D. Turney, "Learning Algorithms for Keyphrase Extraction," Information Retrieval, vol. 2, pp. 303-336, 2000.

[18] E. Frank, G.W. Paynter, I.H. Witten, C. Gutwin, and C.G. NevillManning, "Domain-Specific Keyphrase Extraction," Proc. 16th Int'l Joint Conf. Artificial Intelligence, 1999.

[19] K. Barker and N. Cornacchia, "Using Noun Phrase Heads to Extract Document Keyphrases," Proc. 13th Biennial Conf. Canadian Soc. Computational Studies of Intelligence, 2000.

[20] D.B. Bracewell, F. Ren, and S. Kuriowa, "Multilingual Single Document Keyword Extraction for Information Retrieval," Proc. Int'l Conf. Natural Language Processing and Knowledge Eng., pp. 517522, 2005.

[21] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts," Proc. Conf. Empirical Methods in Natural Language Processing, 2004.

[22] 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.

BIOGRAPHIES

Aleena Susan Mathew received her Bachelor's Degree in Computer Science and Engineering from Sree Buddha College Of Engineering, Ayathil, Elavumthitta, P.O. Pathanamthitta, kerala, India in May 2017. She is currently pursing Master's Degree in Computer Science and Engineering in Sree Buddha College of Engineering, Kerala, India.

Vidya .N received her Bachelor's Degree in Computer Science and Engineering from Sasurie College Of Engineering, Anna University, India in April 2017 and Master's Degree in Computer Science and Engineering from Coimbatore College Of Engineering and Technology, Anna University, India in May 2010. She is a lecturer in the Department of Computer Science and Engineering, Sree Buddha College of Engineering, Ayathil, Elavumthitta P.O. Pathanamthitta, Kerala, India.