Design and Development of Question Answering System in Bangla Language from Multiple Documents

[1]Samina Tasnia Islam
Computer Science and Engineering Department
Military Institute of Science and Technology
Dhaka, Bangladeh
saminatasnia113@gmail.com

[2]Mohammad Nurul Huda
Computer Science and Engineering Department
United International University
Dhaka, Bangladeh
mnh@cse.uiu.ac.bd

***Abstract* – The paper has presented about the experiment of design and development of automatic question answering system for Bangla language. The purpose of proposed answering system is to provide answer based on the keyword, lexical and semantic feature of a question. User gives a question for which answer has to be found from multiple documents. For measurement (time and quantity) related question the system gives specific answer otherwise the system retrieves relevant answers.**

Keywords – Question answering; keyword extraction; stemming; information retrieval; sentence ranking.

I. Introduction

In Question Answering (QA) system a user finds relevant answer of a question from unstructured document collection. QA is similar with Information Retrieval (IR) as minimum amount of information is retrieved which is enough to fulfill user demands [1].

Nowadays we need a system which will receive a question in natural language from user and find answers form given documents and provide relevant answers quickly to user. This study designed a QA system in NLP for Bangla language that allows a user to give question and provides relevant answers from a single or multiple documents. The originality of this paper is that the proposed system identifies the question type and for measurement (time and quantity) related question it provides relevant specific answer otherwise it retrieves relevant answers.

In this paper Section-II discusses the problems of QA system in Bangla language, section-III is about the implementation of the proposed system, section-IV analyses the evaluation of the system, section-V is the conclusion and finally references are given in section-VI.

II. Problems Of Question Answering System In Bangla Language

Question answering system in Bangla language faces many problems. Problems of it are following:

1) Automatic QA from unstructured documents is a difficult task as there is a probability that the source text may contain only one answer to any user's question [2].

2) Mapping questions to answers between question strings to answer strings using lexical, syntactic or semantic relationship is another difficult task [2].

3) The greater the answer redundancy in the text document, the more possibility to retrieve an answer in a simple relation to the user's question. Otherwise, we will need to solve the above issues facing NLP systems [2].

4) In Bangla language there is another challenge to identify the keyword or headword from the question as there is no specific rules in which place the *'wh'* word of the question will be appeared.

III. Implementation

The proposed system can be designed by merging the contents of all the documents, removing stop words, stemming of question and text documents, keyword extraction from question, N-grams formation from keywords for approximate matching, retrieving n-best answers and generating specific answers by question type and evaluation of performance and correctness. Details of the steps are following:

A. Removing Stop Words

In the proposed system necessary intelligence about conjunctions, pronouns, verbs and also inexhaustible words have to be provided for removing these stop words from the question as well as from the text documents.

B. Stemming

Intelligence about suffixes is also needed for stemming the words of the given question as well as the text document. Both the training sets of question as well as the documents have to be stemmed to find out the morphological stem of the words for approximate matching and retrieving information from source documents.

C. Keyword Extraction and N-grams formation from Keywords for Approximate Matching:

Keywords or headwords from the question have to be generated. In the proposed system one of the statistical approaches that is the word intermediate distance vector and its mean value was used to extract keyword from sentence [3].

N-grams have to be formed for effective approximate matching. Keywords generated from questions will be used to form n-grams (unigram/bigram/trigram), thus allowing the n-grams to be compared with other sequences for retrieving relevant information from the source text.

978-1-7281-3445-1/19/$31.00 ©2019 IEEE

Table 1: Keyword Extraction Process. Here the keywords/Headwords are the root of the words.

Question	Keywords
বাংলাদেশে প্রথম কম্পিউটার কবে আসে?	Array ([0] => প্থম [1] => কম্পিউটা [2] => আস)
অ্যাবাকাস কবে আবিষ্কৃত হয়?	Array ([0] => অ্যাবাকাস [1] => আবিষ্কৃত)
জন নেপিয়ার (John Napier) এর অস্থি কি?	Array ([0] => নেপিয়া [1] => (John [2] => Napier) [3] => এ [4] => অস্থি [5] =>)
গটফ্রাইড ভন লিবনিজ কিভাবে যান্ত্রিক ক্যালকুলেটর আবিষ্কার করেন?	Array ([0] => গটফ্ইড [1] => যান্ত্িক [2] => ক্যালকুলেট [3] => আবিষ্কা [4] => কেন)
রিকোনিং যন্ত্র কি?	Array ([0] => যন্ত্)
ডিফারেন্স ইন্জিন কি?	Array ([0] => ইন্জিন)

In Table 1, all the questions are set from a document related with computer collected from Wikipedia (Bangla). Here it's shown that keywords are extracted from the question. Here the keywords are stemmed to find out the morphological root of the word.

D. Lexical and Semantic Features:

In Bangla question *"wh-word" is* a vital lexical feature. An important role is played by the end marker. If the end marker is "|" then the given question is definition type [4].

In the proposed system, measurement unit is used as semantic feature. According to interrogative (*wh-type*) type and semantic feature answer will be retrieved.

Here, if the question type is time related কবে (kəbɛ)/ কখন (kəkhən) and quantity related কত (kət) then proposed system will give specific answer.

A document related with "Cox's Bazar" which is one of the tourist spot of Bangladesh and another document related with "Computer" have been collected from wikipedia(Bangla) and several questions given below have been set from these documents.

- **Question 1:** "কক্সবাজার থানা প্রথম কবে প্রতিষ্ঠিত হয়? (When Cox's Bazar Thana has been first established?)"
  Answer 1: "১৮৫৪ সাল" ("1854 year")
  Answer 2: "সাল এবং পৌসভা" ("Year and Municipality")
- **Question 2:** "কম্পিউটারে প্রথম বাংলা লেখা সম্ভব হয় কখন? (When Bangla writing was possible in Computer first?)"
  Answer 1: "১৯৮৭ সাল" ("1987 Year")
- **Question 3:** "বাংলাদেশে প্রথম কম্পিউটার আসে কত সালে ? (When Computer has first come in Bangladesh?)"
  Answer 1: "১৯৬৪ সাল" ("1964 Year")
  Answer 2: "১৯৭১ সাল" ("1971 Year")
  Answer 3: "১৯৮১ সাল" ("1981 Year")

In the first question the first answer is correct; the second answer here matches with the keywords of the question. In question 2, here only one answer is retrieved which is correct and in question 3 first answer is correct and other answers matches with the keywords of the question. In these answers every word is the root of the words itself.

E. Ranking the retrieved sentences

Retrieved sentences will be ranked by Textual Entailment Module (TE) [5] and best ranked retrieved information will be considered as answer.

IV. Evaluation

Evaluation of the proposed system can be carried out by using the following formulas.

$$\text{Precision} = \frac{\text{Relevant Items Retreived}}{\text{Retreived Items}} \dots\dots \text{eq}^{(1)}$$

$$\text{Recall} = \frac{\text{Relevant Items Reteived}}{\text{Relevant Items}} \dots\dots\dots \text{eq}^{(2)}$$

$$\text{F Score} = 2 * \frac{(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \dots\dots\dots \text{eq}^{(3)}$$

For evaluation purpose several documents have been selected from Wikipedia. Approximate 500 questions have been set to test the performance and correctness. At first for every single question relevant items have been retrieved, retrieved items and total no of relevant answer for the question given by user have been identified. Thus for every single question precision and recall have been calculated using $eq^{(1)}$ and $eq^{(2)}$ respectively.

Then average precision and average recall have been calculated. With the value of average precision and average recall F Score/ F Measure has been calculated using $eq^{(3)}$.

After testing approximate 500 question average precision (0.35) average recall (0.65) have been obtained and the F score (0.45) is calculated based on average precision and averave recall. F score/ F measure reveals the system performance. The higher F score indicates that the system is more better.

Table 2: Process for Calculating Precision and Recall

Questions	Relevant Answers Retreived	Retreived Answers	Relevant Answers	Precision	Recall
কম্পিউটার শব্দের উতপত্তি কিভাবে?	1	2	1	0.5	1
যান্ত্রিক ক্যালকুলেটর সর্বপ্রথম কবে আবিষ্কৃত হয়?	1	1	1	1	1
কম্পিউটার শব্দের অর্থ কি?	0	1	1	0	0
বাংলাদেশে প্রথম কম্পিউটার কবে আসে?	1	1	1	1	1
অ্যাবাকাস কবে তৈরি হয়?	0	6	2	0	0
জন নেপিয়ার (John Napier) এর অস্থি কি?	1	2	2	0.5	0.5
গটফ্রাইড ভন লিবনিজ কিভাবে যান্ত্রিক ক্যালকুলেটর আবিষ্কার করেন?	1	2	1	0.5	1
রিকোনিং যন্ত্র কি?	0	9	2	0	0
যান্ত্রিক ক্যালকুলেটর সর্বপ্রথম কবে আবিষ্কৃত হয়?	1	2	1	0.5	1
গণকযন্ত্র কি?	1	1	1	1	1
মাইক্রোপ্রসেসর উদ্ভাবক কোন প্রতিষ্ঠান?	0	5	1	0	0
কম্পিউটারে প্রথম বাংলা লেখা সম্ভব হয় কখন?	1	2	1	0.5	1
বাংলা ওয়ার্ডপ্রসেসিং সফটওয়্যার উদ্ভাবন করে কারা?	1	1	1	1	1
মাইক্রোসফট উইন্ডোজ' এর সঙ্গে ব্যবহারের জন্য ইন্টারফেস 'বিজয়' কবে উদ্ভাবিত হয়?	1	2	1	0.5	1
			Average:	**0.5**	**0.69**

In Table 2 process for calculating precision and recall has been shown. The first question in Table 2, this system has retreived two answers so here retrieved items=2. Among the two answers only one is relavent with the question and in the given document there is only one relevant answer as a result here relevant items retrieved=1 and relevant items=1. So for the first question using eq(1) and eq(2) precision (1/2=0.50) and recall (1/1=1.00) have been calculate. Like this way for every single question precision and recall have been calculated. Using the value of precision and recall average precision and average recall were calculated.

V. Conclusion

This paper has discussed a QA system from multiple documents. This study concludes the following:

- This approach is able to provide relavent answers of a user question for Bangla Language from multiple documents.
- Keywords of a question can be find out in this approach.
- This approach retreives relevant specific answers for time related and quantity related question.
- Precision, recall and F score of the system are 0.35, 0.65 and 0.45 respectively.
- If the document size is large then there is a possibility that the system may retreive some less relevant information as total no of retrieved information increases.

In future the author of this paper would like to do the following tasks:

- There exists some more answer or information retreival system for some other languages like Chinese, English, Japanese [6][7][[8][9]. So this approach can be compared in future with other existing answer retreival approaches.
- Finding other question type by "*wh*" words and retreiving answers accordingly.
- If any relevant answers starts with pronoun then there is a chance that more relavant answers may exist in other sentences of the text. So finding out more relevant answers if any answers starts with pronoun .

VI. Reference

[1] D. Buscaldi, P. Rosso, J. M. Gomez-Soriano, E. Sanchis, "Answering questions with an n-gram based passage retrieval engine," *Journal of Intelligent Information Systems,* vol. 34.2, pp. 113-134, 2010.

[2] E. Brill, S. Dumais, and M. Banko, "An analysis of the AskMSR question-answering system," *Proceedings of the ACL-02 conference on Empirical methods in natural language processing,* vol. 10, Association for Computational Linguistics, 2002.

[3] S. Siddiqi and A. Sharan, "Keyword extraction from single documents using mean word intermediate distance," *International Journal of Advanced Computer Research,* vol. 6.25, pp. 138, 2016.

[4] S. Banerjee and S. Bandyopadhyay, "Bengali question classification: Towards developing qa system," *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, 2012.

[5] P. Pakray, P. Bhaskar, S. Banerjee, B. C. Pal, S. Bandyopadhyay, A. Gelbukh, "A Hybrid Question Answering System based on Information Retrieval," *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[6] L. Zhenqiu, "Design of automatic question answering system base on CBR," *Procedia Engineering,* vol. 29, pp. 981-985, 2012.

[7] E. Sneiders, "Automated question answering using question templates that cover the conceptual model of the database," *International Conference on Application of Natural Language to Information Systems*, Springer Berlin Heidelberg, 2002.

[8] Y. Ke and M. Hagiwara, "An English neural network that learns texts, finds hidden knowledge, and answers questions," *Journal of Artificial Intelligence and Soft Computing Research,* vol 7.4, pp. 229-242, 2017.

[9] T. Sakai, et al, "ASKMi: A Japanese question answering system based on semantic role analysis," *Coupling approaches, coupling media and coupling languages for information retrieval,* pp. 215-231, 2004.

[10] L. Hirschman and R. Gaizauskas, "Natural language question answering: the view from here," *natural language engineering,* vol 7.04, pp. 275-300, 2001.

[11] A. Andrenucci and E. Sneiders, "Automated Question Answering: Review of the Main Approaches," *ICITA (1)*, 2005.

[12] M. Z. Islam, M. N. Uddin and M. Khan, "A light weight stemmer for Bengali and its Use in spelling Checker," *Centre for Research on Bangla Language Processing*, 2007.