BFQA: A Bengali Factoid Question Answering System

Somnath Banerjee, Sudip Kumar Naskar, and Sivaji Bandyopadhyay

Department of Computer Science and Engineering,
Jadavpur University, India
s.banerjee1980@gmail.com,
{sudip.naskar,sbandyopadhyay}@cse.jdvu.ac.in

Abstract. Question Answering (QA) research for factoid questions has recently achieved great success. Presently, QA systems developed for European, Middle Eastern and Asian languages are capable of providing answers with reasonable accuracy. However, Bengali being among the most spoken languages in the world, no factoid question answering system is available for Bengali till date. This paper describes the first attempt on building a factoid question answering system for Bengali language. The challenges in developing a question answering system for Bengali have been discussed. Extraction and ranking of relevant sentences have also been proposed. Also extraction strategy of the ranked answers from the relevant sentences are suggested for Bengali question answering system.

Keywords: BFQA, Question Answering (QA), Bengali Factoid QA.

1 Introduction

A QA system is an automatic system capable of answering natural language questions in a human-like manner: with a concise, precise answer. Generally questions can be classified into five broad categories ([12]): factoid questions, list questions, definition questions, complex questions and speculative questions.

As there exists no Bengali QA system till date and the development of Bengali QA system is at its nascent stage, our aim for the work reported here was to address the factoid questions for the following factors:- (i) a considerable percentage of the questions actually submitted to a search engine belongs to factoid questions. (ii) The percentages of factoid questions are increased each year in TREC due to frequent occurrences in daily usage. (iii) Sophisticated state-of-the-art approaches to open-domain QA use named entity recognition as a core process for detecting candidate answers.

2 Related Work

Designing a QA system for European languages particularly for English is not new in natural language processing. A number of QA systems have been developed since the 1960s. Two such early QA systems were BASEBALL[3] and LUNAR[4]. The most notable QA system available to date is IBM Watson [2] which was developed under...
IBM’s DeepQA project. Research in QA received significant boost when a shared task on factoid QA was included in the 8th Text REtrieval Conference (TREC).

A number of QA systems were developed for European languages particularly for English ([5][6][7]), Middle Eastern languages ([8][9][10]) and Asian languages, e.g., Japanese ([11][12]) Chinese ([13][14]), etc. The aforesaid systems are capable of providing answers with reasonable accuracy. However, for Bengali, which is a widely spoken language in India and among the most spoken languages in the world, very little work ([12][13][14]) have been reported so far in QA research like other Indian languages.

3 Challenges

To the best of our knowledge, there exists no QA system till date for Bengali. Developing a QA system for low resource language is very much challenging. Several issues were confronted for developing the system which includes-

– *Presence of many interrogatives*: Unlike English there are many interrogatives present in the Bengali. A study [15] identified a total of twenty six interrogatives and classified them into three categories – Unit Interrogative (UI), Dual Interrogative (DI) and Compound/Composite Interrogative (CI).

– *Interrogative position*: A Bengali interrogative can appear in all potential positions, i.e., three positions (first, in between, last) of a question text [12]. This makes it difficult to propose rule-based question analysis.

– *Resource scarcity*: The language processing tools for Bengali are either under development phase or not developed yet. Even a fully-fledged parser has not been developed yet and no NER system is publicly available for Bengali. Besides, gold-standard corpora for QA research are not developed yet.

4 BFQA Architecture

Our proposed factoid QA system for Bengali language, named BFQA, has a pipeline architecture having three components, namely question analysis, sentence extraction and answer extraction. The question analysis module accepts natural language question in Bengali as input posed by the user. The question analysis step processes in five stages, namely question type (QType) identification, expected answer type (EAT) identification, named entity identification, question topical target (QTT) identification and keyword identification. The valid keywords are ‘AND’ed together to form the query. Sentences are extracted from the passages based on the query and are ranked based on the answer score value. Finally, extracted answers are validated using the EAT module. The architecture of the proposed model is depicted in Figure 1

5 Question Analysis

Question analysis plays a crucial role for an automatic QA system. Acquiring the information embedded in a question is a primary task that allows the QA system