Abstract

The goal of this tool is to address a scenario in which multiple inter-related WH-questions are asked in a particular topic reflecting basically content of a web page and they give answer for Frequently ask Question in an intelligent manner by retrieving information from web page. We focused on the issues involved in selecting information on different levels targeting questions of several types.

This project is completed in 4 phases. First, conversion of the web pages into a corpus by removing HTML tags and separating tables and pure text. Second, understanding semantics of questions that are likely to be asked and generating regular expression for them. Third, processing the asked question and answer in best possible manner using annotation corpus. Last we ensure that updation of web pages is taken care and answer to query comes from updated content.

In 7th semester we have achieved to complete first two phase. We see major difficulty in processing of question types and finding unambiguous semantic for the same. We have taken web page from Tezpur University website.

In 8^{th} semester more profound work is be done in achieving efficient answer by focusing on searching techniques like Co-reference Resolution, efficient Chunking, exploring Tag Pattern and Text Corpus, Synonym handling, Morphological Analysis. Neglecting words or frequently used words and focusing on words having low weightage of occurrence can effectively reduce time taken to extract information.

In future, we would like to expand it to give result for different other types of question that will make it more user friendly. The question could be ask from various fields.

Also, increasing its efficiency with the growth in our knowledge of subject, we would be incorporating new ideas to make it more relevant and closer to idealistic. However, we can't achieve ideal situation at this stage due to lack of much needed expertise, as this subject is new to us as a student. Still our efforts are in process of generating a new tool to serve better