Vision Document

1.Introduction

Traditionally, Question Answering (QA) systems query search engines or have access to a crawled collection of the documents on the web to use as data sources in their search and answering process. When the questions that the QA system will be working on is restricted to some domain or topics, using the whole Web as the source corpus becomes unnecessary and might even affect the performance of the QA system. So, having a reduced and enhanced corpus specific to the domain can improve the QA system's performance on future unseen questions on the same domain. This project aims at building a system that can generate this domain specific corpus based on input training questions from the domain that the QA system will be working on.

2. Problem Description

Given a set of specific domain questions, search the data sources and build a small and enhanced corpus that is fed to the QA system.

3. Solution Overview

1. Given the initial set of questions, extract phrases from the questions as seeds to give to search engines to get related documents to those seeds.
2. Process the retrieved documents and add to the corpus.
3. Extract the required phrases from the retrieved document as seeds to give to the search engine in the next round.
4. Evaluate the corpus output by the system in terms of recall with respect to the development questions.
5. Save the corpus to be given to the QA system as input if the the recall of the system has crossed a predefined threshold.

4.Scope & Limitations(including data, tools, etc.)

1. Input: A set of questions and corresponding answers
2. Output: A domain specific corpus based on the input questions and answers
3. Measurement: Recall of the results of Watson QA system with the corpus generated by MR VS Recall of using just the bing/wikipedia top results as corpus
4. Size: Explored data size should be bounded by some function of number of questions
5. Time: No bound on computation time
6. Data sources : Bing, DBpedia