Reading Notes

Open-Domain Textual Question Answering Techniques – UT Dallas (2003)

*“In 1998 TREC initiated an experimental task, the Question Answering Track, with the aim to foster research in domain-independent textual Q/A. The technology that emerged from this TREC-initiated task approximates an information seeking (IS) application in which a user poses a question in natural language and receives the answer as a text snippet as short as a word or as long as a sentence”*

Information Seeking Q/A was good for trivial questions but lacked context. Another issue is that IS Q/A can’t handle systematically ambiguous questions such as “Where is the Taj Mahal?” The system doesn’t know if the user is referring to the one in India, or the Casio in Atlantic City. IS Q/A does not recognise open ended nature of questions such as “What are the causes of violence in the Middle East?”

At the time, 1998 – 2003, it was not efficient to locate information in a large document collection. The TREC Q/A track was an important step to the final goal of creating a system which will locate and return the information that the user wants. The system answers the user with a set of at most 5 answer-strings, of “shorts (50 bytes) or long (250 bytes). 736,794 documents were used to derive answers from. Using multiple document sources makes finding an answer more complex, as it means processing thousands of documents to answer the question. Multiple sources may contain the same answer, or the answer may not be complete.

To give the user an answer the IS Q/A systems use IR techniques for processing documents. The collection of documents are indexed and a retrieval mechanic must be made. The paper mentions three different types of indexing which could be used; word based, conceptual, and paragraph indexing. The indexes are then used to implement different types of retrieval; one which ranks documents containing terms found in the question, another which retrieves the documents and then ranks their passages, or finally simply retrieval and rankings of individual document passages.

At the time, recent results from TREC evaluations showed that IR techniques we not sufficient for finding answers to some questions, at least not alone. Two criteria should be met when processing a natural language question. The system needs to know the type of answer the user expects, to use a previous example, do we want to know about the Taj Mahal in India, or the Casino in Atlantic City. Secondly the system needs to know where to find the answer, using specific keywords in the question.

**2. System Architectures for QA**

Usually there will be three modules for the answer part of a QA system, the question processing module, the document processing module, and the answer processing module.

**Question Processing Module**

The purpose of this module is to capture the semantics of the question so the system can figure out which answer type to return. It is also used to highlight keywords which will be used to retrieve relevant text passages from documents.

**Document Processing Module**

This module uses the semantics captured in the Question Processing module to essentially cross check the keywords and expected answer types against the documents.

**The Answer Processing Module**

Before an answer is given, this module will check the semantics of the answer against that of the question. If they are a close enough match the module will return an answer to the user.

IR – Information Retrieval

IS – Information Seeking

TREC - The Text Retrieval Conference (TREC) is a series of workshops organized by the National Institute of Standards and Technology (NIST) designed to advance the state-of the-art in IR

Open Domain Question Answering System using Cognitive Computing

Ruby Bhati Department of CSE, JSS Academy of Technical Education, Noida rubybhati04@gmail.com Prof. S.S Prasad JSS Academy of Technical Education, Noida [profssprasad@jssaten.ac.in](mailto:profssprasad@jssaten.ac.in)

“Over many years, various question answering systems have been developed. Some of them were database oriented and provide access to specific domains, while few were open domain oriented aiming to provide answers to general trivia-like questions. There can be two basic types of question answering systems, one which tries to provide answer by accessing only structured information stored in the form of database and the other one tries to answer by analysing unstructured information such as text files. The main challenge of the systems that use database to find answers is to transform a natural language query into database query. Database construction is a tedious process whether it is manual or automatic and their domain in most cases is restricted.”