EUREQA: OVERCOMING THE DIGITAL DIVIDE THROUGH A MULTIDOCUMENT QA SYSTEM FOR E-LEARNING

Saket Gupta, Sparsh Mittal a and Ankush Mittal

Department of Electronics and Computer Engineering, Indian Institute of Technology Roorkee
Roorkee, 247667, India
email: {freshuce, sparsuch, ankumfee} @iitr.ernet.in,
web: www.iitr.ernet.in

ABSTRACT
In this paper, we present a Natural Language Processing (NLP) based Multi-document Question Answering system (EUREQA) which intelligently constructs the answers from single and multiple documents. Answers to the questions may not be present in a single document but may require multiple documents. To be used in any real scenario, QA systems must go beyond local search in a document and deal with possibility of finding answer from different locations from one or multiple documents. As a tool for E-learning, our system provides quick and effective reach to the students over the vast information present on web. By this way, information-network service system will be expansively reached throughout the country empowering the general masses, especially the rural people. The comparison of EUREQA with other search engines like Google shows the need of a QA system over and above a search engine such as Google in e-learning.

1. INTRODUCTION
E-Learning has always played a vital role in providing cheap and resourceful access to the ocean of knowledge currently available on the internet and in the form of e-books. Modern Search engines such as Google have huge storehouse of information but the limitation faced by users is of manually searching the documents obtained for the queries thus posed, which not only are vast in number but also respond improperly when a query is posed[1]. These searches are based on keyword search. The meaning of query is very relevant and not only keywords. (e.g. “How is snake poison employed in counteracting poison?” , “When is snake poison employed in counteracting poison?” and “Why is snake poison employed in counteracting poison?” all have different meanings).

In the field of e-learning where e-books and digital texts are made available to the user, the user finds the need of searching for a particular concept, or if a novice requests to know something about a new field to which he doesn’t belong (for e.g. an engineering student may want to know the medicine for sinus [2]), it becomes quite cumbersome to search in the wide field of documents, or huge indexes of books and contents table.

The e - learning system can help to solve the problems of teacher shortage, differing teacher quality, differing learning places and materials, which especially occur in the rural areas .In fact, rural people, will enjoy an equal opportunity to access information service, as city people. It will effectively enable them to access the vast storehouse of knowledge present over digital media.

Moreover, e-learning through question answering provides much more flexibility for everyone, for example, senior citizens and rural residents, workers unable to quit their day jobs to attend college. It is available anytime according to learners’ convenience and fosters life-long learning for each member of the rural area.

Another challenge that the users face today is that many questions are framed to contrast among different competing entities. Such questions have not been considered into the field by any of the existing QA Systems. HITIQA [3] implemented analytical framework for the development of logical answers, but being limited by the scope of their analysis which confined to the basis of frame construction by keyword search and categorization.

In this paper, we present a system which answers the most relevant passage to the user instead of a document or link. The system recognizes the entities of a domain and constructs its answer by exploiting the syntactic structure of the question, and phrase matching to increase the accuracy of the result. We also expand the working of the system by incorporating the answering of questions whose answer lie in multiple documents. Generally such questions are “comparison” type questions. The algorithms are elaborated in subsequent sections.

The main contributions of our work are as follows:

1. In the field of multi-document search, our QA system is a step in the direction of next-generation system which has the capability of extracting answers from diverse documents.
2. It enables the users to access precise and reliable information from a global library (i.e., the web), hence requiring minimum efforts and time providing generic algorithm that intelligently provides answers to users based on their queries and the system’s feedback loop.
3. The teaching system in educational institutions can act as a medium to develop an efficient system for distribution of lecture notes and other educational content like e-books to the students via intranet and internet.
4. The main motivation behind developing e-learning system such as EUREQA is that it will benefit rural people and those living in the remote parts of country much more than it will benefit the people in the mainstream of the society. By this way, information network service system will be expansively reached throughout the country. In a country like India where most of the users are
situated in the remote part of the country like villages with little access to the schools and colleges, this system can enable them to access the organized set up of information, thus taking the benefit of ICT to even the general public.

The paper is organized as follows: Section 2 describes the previous work in Data mining realm with regard to QA systems. Section 3 describes the corpus that was available to us, and our methods of mining answers for contrast seeking questions and extracting answers from multiple documents. Section 4 presents and discusses our experimental results. Section 5 gives conclusions and briefly discusses improvements and future research.

2. LITERATURE REVIEW

In the field of Natural Language Understanding, the different applications that researchers work on are classified as two major classes:

i. Text-based applications: Text-based natural language research is ongoing in applications such as summarizing texts ([4]), exploring pertinent documents from the database of texts, extracting information from messages or articles on certain topics, etc.

ii. Dialogue-based applications. Some applications are Question-Answering systems, where natural language is used to query a database (LUNAR [5]), automated customer service over the telephone, tutoring systems. Some of the problems faced by dialogue systems are language used is very different, and the system needs to participate actively in order to maintain a natural, smooth-flowing dialogue. Also, dialogue requires the use of acknowledgements to verify that things are understood.

E-learning realm

E-learning is defined as formal and informal education and information – sharing that uses digital technology. Using an innovative notion of semantic exploration of a collection of RFC documents [6], Mendes and Slacks have employed fuzzy clustering algorithms to ascertain knowledge domains, TopicMap, although the algorithm is still restrained by drawback accredited to the Table-of-Content page. COVA (Content-based retrieval), enables remote users to access specific parts of interest from a large lecture database [7]. But manually configuring and operating large databases is always an arduous undertaking, which limits the system.

Question Answering

Most common search engines like Google’s search is based on keyword matching. AskJeeves, a commercial search engine, responds to natural-language questions but its recall is very limited due to it’s partially hand constructed database. Cody, et al developed QA system that answers natural language questions by conferring with a repository of documents [8]. Their system MULDER is asserted as the first general-purpose, fully-automated question-answering system on the web. MULDER's recall is at par with that of Google, though, imperfection to give accurate answer to domain specific questions, due to NLP difficulties confines their capability. Semantic construal, elusiveness decisions, general intellect, etc, all are the difficulties yet to be solved. Another elucidation was the use of a locality based similarity heuristic for decision making [11] with the initial assembly of corpus entities using parsers and stemmers. The results were achieved with reasonable accuracy, yet the system confined itself to a particular classification of questions and did not deal with questions pertaining to knowledge in diverse documents, and those which are of the nature of contrasting between elements, i.e. questions of the nature of “Contrast between TTL and CMOS.” are not resolved by any domain dependent/independent knowledge based system. Users generally desire the distinguishing features, rather than plain text of either or both.

Multi-document summarization:

Research on multi-document summarization of news has seen a surge of activity in the past five years, with the development of many multi-document news summarization systems ( [4], [9], [10], [11]) and several that run online on a daily basis [10, 11] generating hundreds of summaries per day. Since 2001, DUC (Document Understanding Conference), a NIST-run annual evaluation conference, has organized quantitative evaluations of multi-document summarization systems. Multi-document Question Answering poses many new challenges over multi-document summarization. Answering a comparison based question is a more challenging task as it requires not mere description of topics being asked, but also includes sentences describing the “distinguishing” features of the topics. For example a summary in response to the query on “features of POP3” would include sentences which match the words present in query such as POP3, feature etc. But to answer a question such as “What is difference between POP3 and SMTP” matching of words present in question would lead one astray and a mere listing of their features will not be sufficient.

3. IMPLEMENTATION

A Architecture Overview

The system is based on searching in context and entities of a domain for effective extraction of answers. The system recognizes the entities by searching from the course material. It is fully automatic as it does not require any manual intervention for configuring it to any particular domain. For context based retrieval a retrieval engine working on locality-based similarity heuristics is used to retrieve relevant passages from the collection. The system utilizes natural-language parsers and heuristics in order to return high-quality answers. The different phases of the system are now discussed.
**B Answer Mining**

This module tries to recognize the entities in a particular course (domain specific entity) to which the user wants to pose questions. This configures the system automatically to any type of course domain. The Question classifier uses pattern matching based on wh-words (such as when, why, what, where, how etc) and simple part-of-speech information to determine question types. Questions similar to: “Differentiate between circuit switching and packet switching” i.e. questions containing the keywords like ‘difference’, ‘compare’, ‘contrast’ may require answer to be extracted from more than one passages. These are separately classified and described in Section C.

In this QA system, Link Grammar Parser decides the question’s syntactic structure to extract Part of speech information. The Question focus is identified by finding the object of the verb. This information is used to select plausible answers from the E-learning materials. The Query Formulation module converts question to query words for providing input to the retrieval engine. The system constructs a hash table of the entities identified from the question on the basis of entity file and the default file provided by the user. Individual words in the question are compared from this table to identify the entities. These keywords are considered most important and are given the maximum weight of 4.

Information Retrieval engine is needed for retrieving the documents and passages. The answers are most appropriate when there is a local similarity in the text with the query. For example for the question “How is the s-plane mapped into the z-plane?” the query terms ‘s-plane’, ‘mapped’, ‘z-plane’ have local similarity which is identified in the text, by locality based similarity algorithm.

The top ranked passages which are now returned (after weighting and ranking on basis of locality and context) are answer candidate. Our system avoids the higher ranking of passages just due to the frequent occurrence of noun words (as in search engines) thus phrase matching based re-ranking is performed. After phrase matching, the system processes the passages according to the classification done in question classification. For example if the question required any date or numerical expression then the system searches for these terms in the passages to match the answer type

**C Multi-document Retrieval**

C1. The Segregate Algorithm

Answers to domain questions involving a “comparison” or “differentiation” or “contrast” between two different entities of the corpus, generally lie in different documents. For example the questions like:-

- Compare between POP3 and SMTP
- Contrast among Binary Trees and B-Trees.

We developed the “Segregate Algorithm” that maps the two separate ingredients (components) of the question (for example ‘Binary Trees’ and ‘B-Trees’) in their respective knowledge domain documents. The actual answer may need to be extracted from different documents, for the information may be present in different documents. The documents are scanned for the central entities and the top 10 documents thus obtained are re-ranked on the basis of entity matching between the documents thus obtained for both components. The re-ranking is now described.

C2. Entity Cluster Matching Based Passage Sieving

The obtained passages will be most accurately depicting a contrast when their parameters or entity clusters are very
Similar. For example if two types of motors are to be compared then the entity cluster contains parameters such as speed, rating, efficiency, mounting, etc, on the basis of which comparison is performed. Thus, re-ranking is performed and the passages are then selectively chosen to obtain top 5 passages according to the new ranks thus scored. The link parser in the system recognizes the entities of the passages obtained for the first component, and thus builds entity cluster, and compares it with entity cluster of passages corresponding to the second component. A correlative rank is awarded to the passages which achieve highest correlation among the entity cluster parameters. The highest ranked passages are then displayed along with their percentage accuracies.

**Comparison of our system with GOOGLE**

We compared our system with the most sophisticated search engine of the age, Google. The questions were posed to Google and 5 documents were checked for presence of answer in them.

**Evaluation metrics**

For general questions we used the popular metric mean reciprocal ranking suggested in TREC [13] for the assessment of question answering systems. The mean reciprocal ranking is defined as follows:

\[
RR = \frac{1}{\text{rank}[i]} \quad \text{MRAR} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\text{rank}[i]}
\]

Where \( n \) is the number of questions.

For evaluation of comparison based questions no metric has been suggested in the literature. To evaluate EUREQA’s performance for such questions we defined a new metric which is called “Mean Correlational Reciprocal Rank” which is defined as follows:

Let \( \text{rank}_1 \) and \( \text{rank}_2 \) be the ranks of the correct answers given by system for both the components respectively. Then

\[
MCRR = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(\text{rank}_1[i] \times \text{rank}_2[i])}
\]

Where \( n \) is the number of questions.

Also if answer to a question is not found in passages presented to user than it is assumed that rank of that question is \( \zeta \) where value of \( \zeta \) is large compared to number of passages. For calculation of MRAR this \( \zeta \) is taken as infinity. To calculate MCRR \( \zeta \) is taken as a much smaller value as it avoids punishing the case where the system provided answer to only one of the components.

In our experiments we took \( \zeta \) to be 10.

While a formal derivation of formula for MCRR is not necessary but it is intuitively obvious being defined very similar to MRAR. Still, we can justify it use with the following arguments:

a) It is symmetric w.r.t. objects being compared so it takes “difference between A and B” and “difference between B and A” to be the same.

b) The answer to a comparison question is complete when both the components (e.g. SRAM and DRAM) are described, not just one. So, it punishes the answers where only one component has been answered.

**Results:** We calculated both MRAR and MCRR for our system and Google search engine. The results were plotted and are shown here.

---

**Figure 1 Output of the system for the contrast type question “What is the difference between program counter and stack pointer?”**

---

**4. EXPERIMENTATION AND RESULTS**

A Sample Test Resource Digital e-documents from the well recognized “Complete Digital Design- A Comprehensive Guide to Digital Electronics and Computer System Architecture” by Marc Balch published by McGraw Hill Publications was adopted to experiment on the system. Unlike the open domain evaluations, where test questions can be mined from question logs (Encarta, Excite, AskJeeves), no question sets are at the disposal of restricted domain evaluators [12]. To build a set of questions we took a pair of 60 questions from conducting a survey among the students of computer science. The group comprised of beginners, sophomores and research scholars as well. This was to simulate use of an e-learning system in a real-life scenario by a general user as the students’ (the potential users) question better reflected questions likely to be posed in a QA system used for a practical purpose.

The students were requested to make questions of all type including comparison type questions. The questions thus received were of widely varying difficulty level covering various topics of the subject.

For each question the system presents five top answers to the user. A question is answered if the answer to the question is available in the text only which is presented to the user (and not in the document from which the text is retrieved).
The following table summarizes the result of our experiments.

<table>
<thead>
<tr>
<th></th>
<th>MRAR</th>
<th>MCRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUREQA</td>
<td>0.617</td>
<td>0.510</td>
</tr>
<tr>
<td>Google</td>
<td>0.481</td>
<td>0.343</td>
</tr>
</tbody>
</table>

**Evaluation of the results:**

Even though the difference in the numerical values of the metrics MRAR and MCRR is large enough we claim that the evaluation was inherently lenient towards the Google. As opposed to EUREQA where the answers must be present in only the passages provided to the user to be taken as correct, for Google the authors had to manually search in the whole document returned by it to check whether it somewhere contained the answer. This makes the user effort exorbitantly large for Google. Moreover, this strategy completely fails for comparison based questions if it does not happen to find a direct answer in the same words as presented in the question.

Figure 5 presents a snapshot of the answer provided by Google for the question “What is the difference between volatile memory and non-volatile memory.”

**5. CONCLUSION AND FUTURE WORK**

In this paper we presented the synthesis of an automatic domain independent QA system which uses the technique of entity recognition and matching. The corpus of the system can be multiple resources, which is generally so in the E-learning field, and the answer extracted from multiple documents. The system is based on searching in context and utilizes syntactic information. We discussed the implementation of the system. The experimentation done on the test questions submitted by various students were then shown, and the results compared with Google. The work on comparison type question and the methodology of re-ranking adopted has yielded promising results in the understanding NLP and data mining concepts of such questions.

Online educational materials are more easily updated and can be presented in an attractive manner to benefit even an uninterested learner. This is a major advantage of our system which will solve the problem of indifference and lack of awareness, so widely present in rural people.

Our future work will focus on developing a systematic framework for image (jpeg, bmp, etc) extraction and method for its contextual presentation. Along with images, focus will be on incorporating audio lectures along with text lectures available in the e-learning facilities and adapting the system to the needs of a lay man. This will greatly enhance the efficacy of the system.
6. REFERENCES