RESEARCH ARTICLE OPEN ACCESS

QUrdPro: Query processing system for Urdu Language

Rukhsana Thaker, Dr.Ajay Goel
CSE Department, BGSB University, Rajouri ,J&K
Head,CSE Department, Baddi University, Baddi,H.P

Abstract:
The tremendous increase in the multilingual data on the internet has increased the demand for efficient retrieval of information. Urdu is one of the widely spoken and written languages of south Asia. Due to unstructured format of Urdu language information retrieval of information is a big challenge. Question Answering systems aims to retrieve point-to-point answers rather than flooding with documents. It is needed when the user gets an in depth knowledge in a particular domain. When user needs some information, it must give the relevant answer. The question-answer retrieval of ontology knowledge base provides a convenient way to obtain knowledge for use, but the natural language need to be mapped to the query statement of ontology. This paper describes a query processing system QUrdPro based on ontology. This system is a combination of NLP and Ontology. It makes use of ontology in several phases for efficient query processing. Our focus is on the knowledge derived from the concepts used in the ontology and the relationship between these concepts. In this paper we describe the architecture of QUrdPro ,query processing system for Urdu and process model for the system is also discussed in detail.
Keywords: ontology, query processing, closed domain

I. INTRODUCTION

The growing interest in providing different information on the web has increased the need for a complicated search tool. Most existing information retrieval systems only provides documents, and this often makes users read a relatively large amount of full text [1].The study of question answering systems (QAS), which enable people to locate the information they need directly from large free-text databases by using their queries, has become one of the important aspects of information retrieval research[2].

Question-answering (QA) study emerged as an effort to deal this in sequence-excess problem. QA systems are classified in two main parts: namely open domain QA system and closed domain QA system. Question which deals with nearly everything and can only relies on worldwide information, such type systems are called as open domain question answering system. On the other hand, closed-domain question answering deals with questions under a particular domain (music, weather forecasting etc.) The domain specific QA system involve deep use of natural language processing systems formalized by building a domain specific ontology[3]

In the recent years, it has been seen that the information is also available over the web in various languages such as Chinese, French, English, Hindi, Urdu, Arabic, etc. With the advent of the Unicode scheme, users can contribute their knowledge over the web in their own language that can be used by others who speak the same language. As a result, the volume of multilingual information is continuously increasing. Moreover, the availability

of information written in native languages provides us valuable insight into the cultural, political, and social condition of that country. Among variety of languages, the Urdu language is among the largest spoken languages of the world and state language of J&K state. Due to the availability of large amount of Urdu language documents over the web, it is

required to develop a sophisticated information retrieval system to utilize the information efficiently. We propose a Query Processing System Architecture that uses words of the sentence (question) typed as a source to search answer. Principle of our Question Processing system is that the system gives user a set of sentences in Urdu language containing the words of sentence typed. Here we described a model of a system, which accepts user queries in natural language (Urdu) after analyzing those queries on the bases of ontology match them with the information stored and result is displayed to users, thus helping users to get the most wanted information without penetrating the huge amount of information existing on the web.

The outline of the paper is as follows. The paper is organized as follows In Section II we discussed related work of this object; Section III discussed the proposed architecture of the query processing system. Section IV discusses the process model in detail. Section V presents the conclusion.

II. LITERATURE REVIEW

Machine translation (MT) was the first computer-based application related to natural language which came into existence in 1940's but first project was used in 1946 in World War II to break enemy code [4]. In 1969 Natural Language Question Answering Systems was made which focused on syntactic, semantic, and logical analysis of English strings [5].

In 2004 a web based Chinese Automatic Question Answering System was made by Cai Donfeng and Cui Huan [6] which uses Google Web services API. In 2006 the geographic core model was made which supported every Geographic Feature (GF) detected in collection of text or image documents [7]. Later on in 2008 A Question Answering Systems was made which automatically generates questions related to the document . QA System based on the analysis of interrogative sentence and the answer type was made for Chinese language in 2009 by CunLi Mao and others which use to carry the text retrieval with the help of the domain knowledge and obtain the relevant paragraphs of the question [8].

Kanaan et el [9] described Arabic QA system which makes use of data redundancy rather than complicated linguistic analyses of either questions or candidate answers, to achieve its task. Akour et el [16] introduced a QA system (QArabPro) for Arabic Language. The system handles all types of questions including (How and Why). But generally similar to those used in a rulebased QA for English text , the authors uses a set of rules for each type of WH question.

Recently, researchers have been attracted to the task of developing Ontology based QA systems such as Querix [10], and AquaLog [11, 12]. Querix introduced by Kaufmann et al. is another ontology-based question answering system that translates generic natural language queries into SPARQL. In case of ambiguities, Querix relies on clarification dialogues with users. In this process users need to disambiguate the sense from the system-provided suggestions. AquaLog Proposed by Lopez et al. [11, 12] is a portable semi-automatic QA system that combines natural language processing, ontologies, logic, and information retrieval technologies. We say that AquaLog is portable, because the configuration time required to customize the system for a particular ontology is negligible. In AquaLog the ontology is used intensively as it is utilized in the refinement of the initial query, the reasoning process, and in the similarity algorithm AquaLog provides the best balance between domain customization effort and performance.

Vanessa Lopez et al "Question Answering on the Real Semantic Web"[13] to conclude with, Power Aqua balances the heterogeneous and large scale semantic data with giving results in real time across ontologies, to translate user terminology into distributed semantically sound terminology, so that the concepts which are shared by assertions taken from different ontology's have the same sense. The goal is to handle queries which require to be answered not only by consulting a single knowledge source but combining multiple sources, and even domains

III. ARCHITECTURE OF QUrdPro SYSTEM

This section presents the architecture and functionality of QUrdPro system. The working of the proposed model is as follows. Here user is provided with the facility to type his question in Urdu . This question is used to extract all the possible answers for it. The architecture for QUrdPro System is as revealed in Figure 1.

1. Query interface is used to retrieve the question posted by the user in Urdu language.
2. NLP parser is used to parse the user query.
3. Interpreter is used to remove the stop words and for stemming.
4. Query formulation is used to formulate the query based on the domain ontology
5. Ontology is used to define classes and concepts of the dataset for a particular domain in Urdu language.

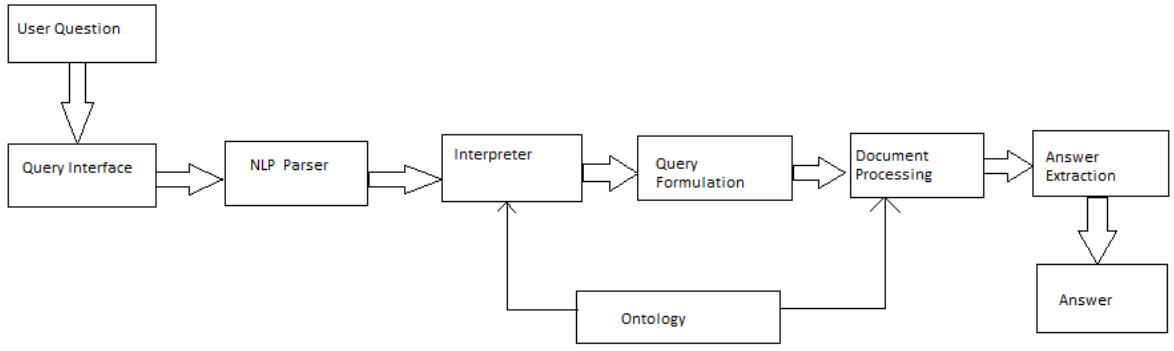

Fig 1: The QUrdPro Architecture

IV. QUrdPro PROCESS MODEL

The QUrdPro model provides a framework which integrates NLP and Ontologies for Urdu language. In our work we have focused on creating a model for our QUrdPro system. In this model there are four phases.

Phase 1. User Interaction phase
Phase 2: Query Processing / Analysis
Phase 3: Document Processing using Ontology
Phase 4: Answer Extraction and retrieval

These phases are discussed in detail.

1. User Interaction Phase : In this phase the user inputs the query in urdu language in the interface (a simple text box or dialog box) provided by the system.
2. Query Processing / Analysis Phase : Query processing is an important phase of the system. It is performed to understand the question asked by the user. Following steps are performed during query processing:
   NLP Parsing: The question asked by the user is taken as a string. This string is tokenizes into subject, verb, preposition etc in this step.
   Interpreter: It interprets the words generated by the NLP parser. It removes the stops words, and performs POS tagging. This can be done using urdu stemmer and POS tagger.
   Query formulation: This classifies the question to the question types(like what, when , why etc).
3. Document Processing: In this phase document is processed to find the answer for the query. Document is searched on the basis of ontology. Words are searched on the basis of concept and there relation that are defined in the ontology.
4. Answer Extraction and retrieval: In this phase answers are selected from the document. These answers are displayed on the interface provided by the system.

V. CONCLUSION

In this paper we have described QUrdPro system which is based on ontology. Basically this is a system for query processing of Urdu language which is an unstructured language. Information retrieval for unstructured data is a very tedious job. So in this paper we have proposed a model for processing such data. We have used domain knowledge for constructing ontology for this system.

REFERENCES

[1] R. Baeza-yates & B. Ribeiro-Neto, (2011) "Modern Information Retrieval", second edition. [2] Breck, E., Burger, J., House, D., Light, M. & Mani, I. (1999) ``Question answering from large document collections'', Question Answering Systems: Papers from the 1999 AAAI Fall Symposium, 5-7 November, North Falmouth, MA, AAAI Press, Menlo Park, CA, pp. 26-31.

[3] Muthu krishnan Ramprasath1 and Shanmuga sundaram Hariharan2,"A Survey on Question Answering System", International Journal of Research and Reviews in Information Sciences (IJRRIS), pp171-179,2012

[4] Liddy, NY. Marcel Decker "Natural Language Processing" , in Encyclopedia of Library and Information Science in 2001

[5] ROBERT F. SIMMONS, "Natural Language Question Answering Systems: 1969"

[6] Cai Dongfeng, Cui Huan, Miao Xuelei, Zhao, Ren Xiangshi "A Web-based Chinese Automatic Question Answering System",pp1141-1146, IEEE 2004

[7] Sallaberry Christian, Etcheverry Patrick, Marquesuzaà Christophe, "Information Retrieval And Visualization Based On Documents" Geospatial Semantics", International Conference on Information Technology: Research and Education, pp277-281, 2006

[8] Cun-Li Mao, Li-Na Li, Zheng-Tao Yu, Lu Han, Jian-Yi Guo , Xiong-Li Lei "Research on Answer Extraction Method for Domain Question Answering System (QA)", 2009 International Conference on Computational Intelligence and Security, pp79-83, IEEE 2009

[9] Kanaan, G., Hammouri, A., Al-Shalabi, R. and Swalha, M.(2009). "A New Question Answering System for the Arabic Language". American Journal of Applied Sciences 6 (4): 797-805

[10] E. Kaufmann, A. Bernstein, and R. Zumstein., "Querix: A natural language interface to query ontologies based on clarification dialogs," In proceeding 5th International Semantic Web Conference (ISWC 2006), pp 980–981, 2006.

[11] V. Lopez, M. Pasin, and Enrico Motta, "AquaLog: An Ontology-Portable Question Answering System for the Semantic Web," Lecture Notes in Computer Science, Vol. 3532, Springer, Berlin, pp. 546-562, 2005.

[12] V. Lopez, and E. Motta, "Ontology-Driven Question Answering in AquaLog," Lecture Notes in Computer Science, Vol. 3136. Springer-Verlag, Berlin, pp. 89–102, 2004.

[13] Vincent Barbier, Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba and Anne Vilnat "Semantic Knowledge in Question Answering Systems"