International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
International Conference on Industrial Automation and Computing (ICIAC) (12th&13th April 2014)

RESEARCH ARTICLE OPEN ACCESS

Semantic Approach For Processing Natural Language Query

Ms. Ankita A. Kendhe*, Prof. S. J. Karale**

* (Dept. of Computer Technology, Yeshwantrao Chavan College of Engineering,Nagpur (Maharashtra), India
Email: ankitakendhe@gmail.com)

** (Dept. of Computer Technology, Yeshwantrao Chavan College of Engineering,Nagpur (Maharashtra), India
Email: sjkarale@in.com)

ABSTRACT

Automatic question answering system plays important role in current search engines. General search engines are based on keyword based searching mechanism. It retrieves enormous amount of data from which sometime it become difficult for user to recognize relevant information. Another problem is that, users get confused if there are same words with different meaning. To solve these problems semantic based searching mechanism are used. It searches information by understanding the intent of user and meaning of words in searched sentence. We show our approach by describing an implementation and a step-wise answering scenario with a sample query.

Keywords - Question answering system, Ontology, Semantic web, OWL, Triples.

I. INTRODUCTION

Automatic question answering system plays important role in current search engines. General search engines are based on keyword based searching mechanism. It retrieves enormous amount of data from which sometime it become difficult for user to recognize relevant information. Another problem is user get confuse if there are same words with different meaning. To solve these problems semantic based searching mechanism are used. It searches information by understanding the intent of user and meaning of words in searched sentence. It uses semantics to produce highly relevant results. This technique can be used to retrieve information for knowledge bases like ontology. Ontology (Fernandez et al., 2009) is "a technology used to enable the domain knowledge at a high level and improve the query time used in Question Answering system" [1].

II. SEMANTIC APPROACH

Semantic web is an extension of World Wide Web. It is used to disambiguate words so that human and computer works in cooperation. It allow user to find, share and combine information more easily and efficiently. Ontologies an important component of semantic web is used to enhance understanding the intent of user and meaning of words.

To solve questions related to semantics, ontologies are one of the main approaches used for knowledge management. Ontologies are defined as conceptualization which contains set of concepts, their interrelation and rules that governs these concepts to be interpreted by machines. Most ontologies illustrate individuals, classes, attributes, and relations.

For creating ontologies, Web Ontology Language (OWL) is used. OWL is based on W3C standards and help in defining ontologies which contain information representation features. OWL builds on XML and permit users to give machine readable semantic annotations for particular communities of interest. OWL is used to describe classes, properties and individuals. These descriptions can be in single ontology or in the combination of multiple joined ontologies.

III. TRIPLES BASED MODEL

To translate NL query to intermediate triple-based representation linguistic components are used. Linguistic components consist of English tokenizer, sentence splitter, POS tagger and VP chunker. The annotations returned after the sequential execution of these resources include information about sentences, tokens, nouns and verbs. These annotations are used to query ontology. It is preprocessing step which help in accurate classification of query. It is needed to understand particular NL query and also guide NL query in creating equivalent triple based representation [2].

Tokenizer is used to separate a stream of line into words, phrases, symbols or other meaningful elements called tokens. These tokens become input to POS tagging, parsing etc. Automatic allocating descriptors to given tokens is called Tagging. Tag may specify one of the parts of speech, semantic

information and so on. The process of allocating one of the parts of speech to the given word is called Parts Of Speech tagging. Parts of speech contain nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories. Then parser is used which compares grammar against input sentence to produce parsed structure called parse tree [3].

IV. PROPOSED WORK

When user input question in natural language, that question is first processed to get query triples. Triples are in the form of {Subject, Object & Predicates). Linguistic components are used to classify query in triples. The result we get is "Query Triples". There are two main reasons for adopting a triple-based data model. First of all, although not all possible queries can be represented in the binary relational model, in practice these exceptions occur very infrequently. Secondly, RDF-based knowledge representation (KR) formalisms for the semantic web, such as OWL also subscribe to this binary relational model and express statements as <subject, predicate, object> [1]. Hence, it makes sense for a query system targeted at the semantic web to adopt a triple-based model that shares the same format as many millions of other triples on the Semantic Web.

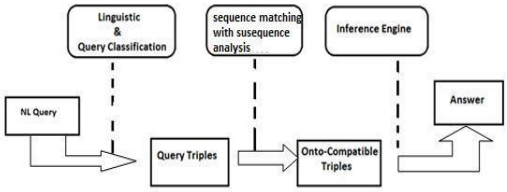

Fig 1. natural language query processing

After getting Query Triples next step is to map these triples to onto compatible triples using sequence matching with subsequence analysis. Sequence matching with sequence analysis task is to match these query triples in opposition to an existing knowledge base, which contain semantically described words or phrases.

While trying to classify parts of the user input a comparison between those two must eventually be made. Because of user input being an infinite set, either a reduction of input words or expansion of the knowledge base must be made. Accepted approaches that solve this problem include stemming, lemmatization and various distance functions. Stemming is based upon a set of rules, which determine word morphing, and is therefore limited to weakly inflected languages, where such rule collections exist. Lemmatization is used in conjunction with large language specific dictionaries, which are used to expand the knowledge base dictionary. This information is then used to derive morphed words into their lemma. For measuring the results we had to create a test set, which would allow us to compare sequences against each other and would at the same time contain the information about the closest match [4].

Fig. 1, shows user's NL query gets translated into "Query Triples" using linguistic and query classification. Then these triples are mapped to "Onto-compatible triples using sequence matching with subsequence analysis" to get desired efficient and relevant answer to user's query.

V. IMPLEMENTATION

Implementation of this project includes various steps that have been shown in Fig. 2. These are: Normalization, interpretation, Ontotriples and finally the answer finding mechanism.

Normalization process gives the "would be" condition where the answer can be very specific to the question. For example as shown in the example, the question is "who is the hod of computer department?" and the normalization would be like "who be the hod of computer department". Here it is normalization that took us to next step in order to find the answer. Interpretation is process of finding objectives of the question, for example Subject, Predicate and Object i.e triplets.

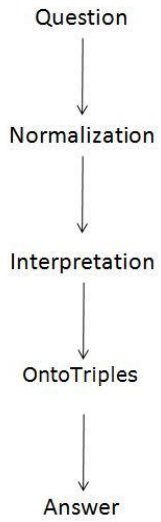

Fig 2. implementation flow

RDF-based knowledge representation (KR) formalisms for the semantic web, such as RDF itself [14] or OWL [5] also subscribe to this binary relational model and express statements as <subject, predicate, object>. It gives where should be the focus in order to find the answer. For example, in context of computer department in YCCE college, the

question is "who is the hod of computer department?" and the interpretation will be "Subject: who is, Predicate: the hod and Object: computer department". Everything is resource <subject> that is connected with other resource <object> via predicate. Predicate are also defined as resources, but they are used in order to define relations between resources.

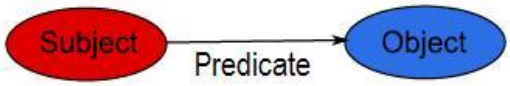

Fig 3. graph model for triples

Fig. 3, shows a Directed Labeled Graph (DLG) with the subjects and the objects as nodes, and the predicate as edges connecting subjects to objects.

OWL was introduced as a proposition for an accurate and flexible representation language for ontologies; it is based on the RDF language – the class and property-structure of RDFS. So we created ontology in OWL language using Standard Protégé tool which is using for creating and editing OWL documents.

Example: Fig. 4, shows sample hierarchy formed for OWL document using Protégé tool. It shows the class instance tree for Mr. A. R. Patil Bhagat. Here we see that Mr. A. R. Patil Bhagat is HOD and Associate professor for computer technology department of YCCE. Blue line indicates link between class and subclasses. Red line indicates link between class and its instances.

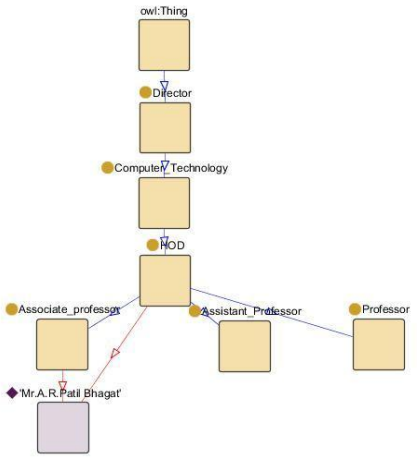

Fig 4. class instance tree for Mr. A. R. Patil Bhagat

After creating OWL file, next task is to map triples i.e. (subject, predicate, and object) to ontotriples using relation matching algorithms such as Jaccard Distance string matching algorithm [6][7], WordNet.

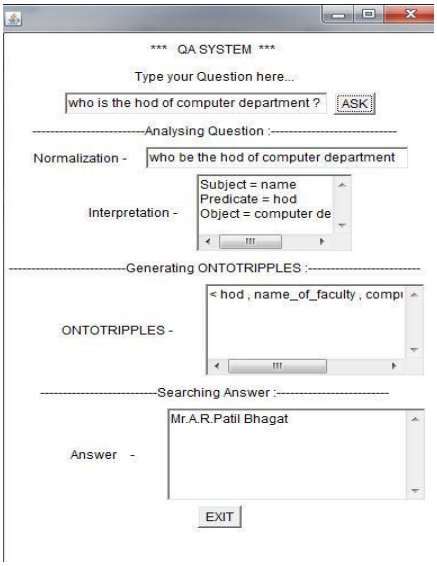

Fig 5. current screenshot for question answering system giving relevant answer

Fig. 5, shows current screenshot of the system. Our system is able to translate question "Who is the hod of computer department?" into query triples as<name, hod, computer department> which is mapped to onto-compatible triples as <hod, name_of_faculty, computer department>.Currently it works for small ontology only. But final system will work for large ontology.

VI. CONCLUSION

Currently our question answering system is under construction. It works for small domain specific ontology only. We are trying to create large ontology which will be capable of answering all possible questions for particular domain specific ontology.

REFERENCES

[1] S. Kalaivani and K. Duraiswamy, "Comparison of Question Answering Systems Based on Ontology and Semantic Web in Different Environment", Department of Computer Science and Engineering, Institute of Road and Transport Technology, Erode-638 316, Tamilnadu, India K.S. Rangasamy College of Technology, Tiruchengode-637 209, Tamilnadu,India, Journal of Computer Science 8 (9), 1407-1413, 2012.

[2] Vanessa Lopez, Victoria Uren, Enrico Motta, Michele Pasin, "AquaLog: An ontology-driven question answering system for organizational semantic intranets", Knowledge Media Institute & Centre for Research in Computing, The Open University, Walton Hall, Milton Keynes MK7 6AA, United Kingdom, March 2007.

[3] nlp.stanford.edu
[4] Marko Ferme, Milan Ojsteršek ,"Sequence matching with subsequence analysis", ISBN: 978-960-474-250-9.
[5] Mc Guinness, D., van Harmelen, F.: OWL Web Ontology Language Overview, W3C Recommendation 10 (2004) http://www.w3.org/TR/owl-features/.
[6] WilliamW. Cohen, Pradeep Ravikumar, Stephen E. Fienberg," A Comparison of String Distance Metrics for Name-Matching Tasks", American Association for Artificial Intelligence (www.aaai.org), 2003.
[7] Ankush Maind, Prof. Anil Deorankar and Dr. Prashant Chatur, "Measurement Of Semantic Similarity Between Words: A Survey", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2(6), December 2012.