International Journal on Islamic Applications in Computer Science And Technology, Vol. 3, Issue 3, September 2015, 24-32

Natural Language Interface for Quran Knowledge Retrieval

Aliyu Rufai Yauri, Rabiah Abdul Kadir, Azreen Azman, Masrah Azrifah Azmi Murad
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia
rufaialeey@yahoo.com, {rabiah, azreen, masrah}@fsktm.upm.edu.my

ABSTRACT

With the introduction of semantic Web technology, recent researches has focuses on Natural Language Interface (NLI) with the aim of using unstructured natural language query to retrieved structured RDF data. The NLI researches focuses on semantically formulating natural language query either manually, semi-automatic or automatically. Most of these systems are search systems, where user's natural language query is semantically formulated to structured query. However, these system rely on the presence of ontology concept in the query, in order to formulate the natural language query and it is consider failure when the concept is not found in the query. This paper proposed a natural language interface that semantically formulate user's natural language query either with the presence of ontology concept in the query or otherwise. A system-based suggestion approach is proposed as solution to semantically formulate natural language query when there is no ontology concept in the query. The system is based on using n-gram maximum likelihood estimation to automatically suggest the possible structured query. The proposed natural language interface has improve the effectiveness retrieved result with 0.03 precision and 0.07 recall improvement.

***Keywords*: Natural Language Interface, ontology, Semantic, Natural Language Query, Information Retrieval.**

1. INTRODUCTION

The shortcomings of the traditional keyword search system has led to the introduction of semantic web was introduced by the W3C consortium. The semantic web, in other words a web of linked data, is an extension of the current version of the web whereby information is given a well-defined meaning to enable human and computers to easily work together. Semantic Web models the meaning of information on the web, as well as applications and services, so as to discover, annotate, process and publish data that is encoded in them (Zou, Finin, & Chen, 2004.). Semantic Web represents data in RDF graphical triple form representation triple.

{Subject, Predicate, Object}

In triple format, a subject represents a concept that appears on the left side of the triple, an object is a concept that appears on the right hand side of the triple, and a predicate stands for explicit relationships that exist between subject and object which may be represented by a word, phrase or sentence.

The challenge is that querying data represented in RDF structure in the knowledgebase requires a structured query such as SPARQL. These structured queries are represented in complex syntax which requires the user to be familiar with how to use the syntax before they can use the query language for retrieving the desired information from the knowledgebase. Studies show that users prefer using natural language Interface where natural language query are used to retrieve these structured RDF data (Tablan, Damljanovic & Bontcheva, n.d., Kaufmann & Bernstein, 2008). Natural language Interface hides the complexity of the structured query thereby enabling users to use natural language in order to retrieve information from structured data.

This paper describes work on natural language interface that semantically formulates Natural language query to structured query in order to retrieve corresponding knowledge from Quran Knowledge base. The proposed approach has features that differentiate it from the existing approaches. The proposed system is able to semantically formulate natural language query to structured query automatically with the present of concept in the query or otherwise which is not the case in the existing system. Existing systems rely on the presence of concepts in the query before natural language query are semantically formulated to structured query.

The proposed natural language interface system was tested using the Quran ontology published by Leeds University of the United Kingdom. Leeds University's Quran ontology is composed of 300 important noun concepts identified from the Holy Quran, and approximately 350 relationships that link the concepts (Dukes & Atwell, 2009). The query set used for the experiment in this research was obtained from the Islamic Research Foundation Website where people send their queries related to Islam and experts answer the questions.

The rest of this paper is organized as follow. Section 2 provides related work. Section 3 over view of the proposed system. Section 4 presents evaluation and section 5 analyses of the proposed system. And section 6 presents collusion and future direction of the research.

2. LITERATURE REVIEW

In recent years, several researches has been presented on the semantic query formulation of retrieving from structured RDF data in the knowledge base. Previously reported systems on semantic query formulation can be categorised into three approaches: mainly manual, semi-automatic and automatic semantic query formulation systems.

Manual semantic query formulation is mostly template –based where user is required to semantically formulate structured query manually. In this approach, user is required to be either be familiar with the syntax of the structured query such as SPARQL syntax or have a knowledge of how the RDF data is represented in the knowledge base in order to retrieve data from the knowledge base. Ontology editors such as Protégée and some query editors like Virtuoso SPARQL, Flint SPARQL Editor, and Drupal SPARQL Query Builder among others

are systems that enable users to manually formulate a formal query language and retrieve knowledge from the knowledgebase. Protégée enables users to manually construct a SPARQL query in order to retrieve knowledge stored in Protégée. In Protégée, a user can use the SPARQL query tab provided to construct a SPARQL query, and execute the query against the knowledge stored in Protégée.

Semi-automatic systems could be template-based or browse like systems where user and machine are involved in the semantic query formulation process. Work in (Popov, Kiryakov, Kirilov, Manov & Dimitar, 2003; Damljanovi, 2011) improve from manually semantic query formulation to semi-automatic semantic query formulation were proposed. In these system computer and human work together in order to semantic formulate structured query. (Popov. et al, 2003) is a template base query formulation approach which presents to user with predefined query templates from where they choose to semantically formulate the structured query SeRQL that is used for retrieval from the knowledgebase. The SeRQL translation is then used to match data in the knowledgebase for retrieval. Another semi-automatic approach for semantic query formulation is (Donderler, Saykol, Arslan, Ulusoy & Gudukbay 2003). Barzdins, Liepins, Veilande, & Zviedris (2008) present an ontology-assisted query formulation based on concepts annotated in the database. The main concept of this approach is to use ontology as the main guide to generating a SPARQL query.

TAP (Guha, McCool & Miller, 2003) proposed some improvements to the approaches mentioned above. TAP goes beyond just providing a template from which the user chooses the variables that are used for query formulation, by providing the users with a search and browse mechanism.

CINDI (Stratica, Kosseim, & Desai, 2005) represent systems that incorporate mechanisms for dealing with ambiguity in user queries where an external dictionary is used to extend formulation of the query to the synonyms of the original query (Cimiano, Haase, Heizmann, Mantel, & Studer, 2008).

In order to fully automate the process of semantic query formulation, various systems attempt to automate the process of semantic query formulation. QuestIO is a natural language interface approach that formulates a user's natural language query to a structured query SeRQL (Tablan, Damljanovic & Bontcheva, 2008). The system is a domain independent base and easy to be use with our need of training. The system is based on small fragment queries, which are able to work with ill-formed or incorrect sentences. It involves binding the ontology structure, the use of fuzzy string matching and ontology similarity metrics to formulate user queries in order to retrieve information from the knowledgebase. The system cannot cope with queries that may be more than one sentence, and it does not have provision for resolving ambiguity in cases where different vocabulary from that in the knowledgebase is used.

PowerAqua (Lopez, Fernández, Motta, & Stieler, 2011) is an ontology-based natural language interface that supports the transformation of a user's natural language query into structured form. It is an extension of the previously discussed Aqualog, mainly designed to

cope with problems in the Aqualog system. Power Aqua has the advantage of being domain independent, where user queries don't have to target specific domains. User queries can be formulated to retrieve information from semantically structured data on the web. However one of the drawbacks of Power Aqua is that the system is limited to single sentence queries.

FREyA (Damljanovic, Agatonovic, & Cunningham, 2010) is a natural language interface for querying ontologies where the system attempts to automatically semantically formulate natural language queries into structured queries. The system provides the user with a clarification dialogue in case the system fails to answer the query. The system uses semantically annotated ontology with syntactic parsing in order to formulate user natural language queries into structured SPARQL query language. Suggestions provided by this system also are based on if concepts are identified in the queries terms. If concepts are not identified in the query terms the system will fail to provide suggestions and thus fail completely. Although the system claims to automatically formulate user queries to SPARQL, therefore, the system's results in term of automatically formulating natural language semantically are not good enough, as they mostly require the user to disambiguate their queries through clarification dialogue, because ambiguity in natural language is almost inevitable. This means the user is greatly engaged with the system for query processing. The system is also based on single sentence queries.

Our focus in this research is to focus on fully automating the natural language interface system in order to automatically formulate user's inputted natural language query. And provides assistance t user in case the system fail to provide answer to the inputted natural language query.

3. SYSTEM OVERVIEW

This section provides a summary of the implementation of the natural language interface for Quran knowledge retrieval proposed by this paper. The detailed description of the step by step implementation of the natural language interface system proposed in this paper can be well described by the flowchart 1.

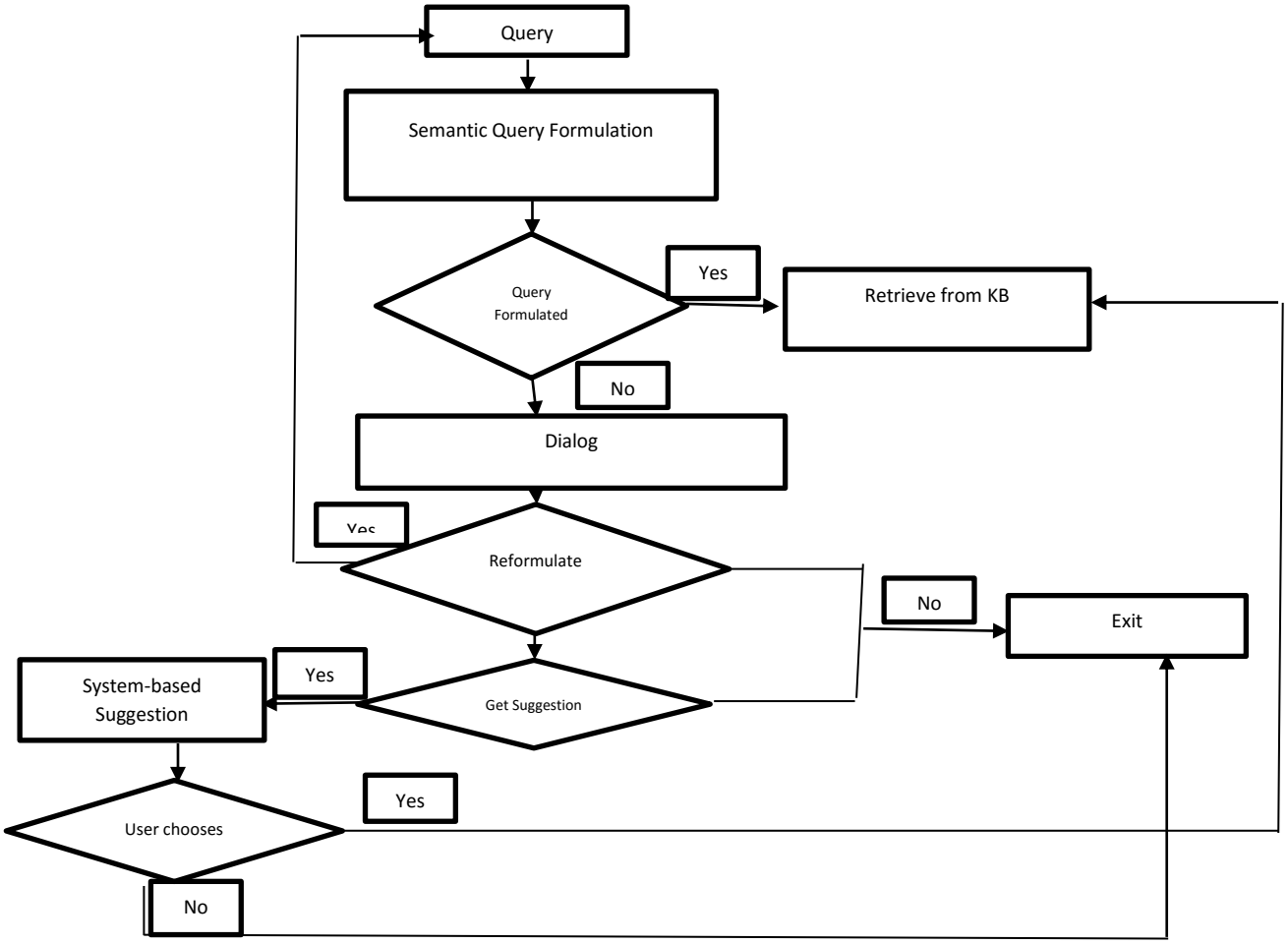

Figure 1: Flowchart of the Natural Language Interface for Quran Knowledge retrieval

Figure 1 provides the implementation of the natural language interface system in this paper which can be broken down into several steps:

3.1 Query pre-processing module

The first step in the automated semantic query formulation process in the proposed approach is query pre-processing module. This module involved tokenization, stop words removal, and lemmatization. The output of the lemmatization used in part-of-speech tagging to assign a part-of-speech for each the query token for the further process of the queries.

3.2 Semantic Query formulation module

The main objective of this module is to take the natural language queries that were processed in the query pre-processing module, concepts and predicates identification, automatically resolve ambiguities, and semantically formulate the query tokens into structured triple format (Subject, Predicate, Object). For semantic query formulation, statistical machine

learning approach was use to automatically identify concepts from natural language query and automatically detect predicate between the identified concepts. This section include

- Concept identification
- Predicate detection

In the concept identification process, the system automatically takes noun query tokens and matched those tokens against the ontology gazetteer. Ontology gazetteers is a container contain the list of all Quran ontology concepts. If any of the user's natural language query tokens match any concept in the gazetteer, such token is automatically identified as concepts. The next step is identifying possible predicates between these identified concepts.

Predicate detection involves using the remaining query tokens after concepts has been identified to detect predicate. Predicate detection involved using statistical machine learning approach based on Ngram maximum likelihood estimation to automatically identify possible predicates between the identified concepts.

When possible predicates are identified, the system automatically formulate triples and are then parse to the retrieval module for retrieval. However, when system fails to automatically formulate the natural language query to structured query. The system performs the next step.

3.3 User Participation Dialog module

When system fails to automatically formulate natural language query to structured triple representation, we proposed a user participation dialog. This dialog involved engaging the user in order to facilitate further processing of the query. To engage the user when the system fails to automatically interpret user queries, the system should provide dialogues where the user has the option of either choosing a system suggestion or reformulating their query in order to retrieve answers from the knowledgebase. This provides a flexible option for users who may prefer suggestions from the system from which they choose so as to formulate their query, or those who prefer not to waste time choosing from suggestions, and would rather just go ahead and reformulate their query.

3.3.1 System based Suggestion:

System suggestion is a mechanism where system uses query tokens suggest to the user variables with which to formulate structured queries. When system fails to automatically formulate natural language to structured query dues to two reasons.

1: Concept is identified, but predicate is not detected.

2: No concept identified from the natural language query

The proposed system based suggestion in the research, proposed solution in case any of the above scenarios occurs:

Concept is identified, but predicate is not detected: When the system is able to automatically identify ontology concepts but fails to automatically detect likely predicates between the

identified concepts, in our approach it automatically pulls out all triples from the triple store that involve the identified concepts and presents them to the user. For example, if the system is able to identify Quran and God which are both ontology concepts, but fails to identify any predicate from the user query token, it will automatically pull out all triples that are either (Quran, any relation, God) or (God, any relation, Quran) or (Quran, literal) or (God, literals) and present them to the user as suggestions. Here the triple chosen by the user is used as the triple for SPARQL query generation

No concept identified from the natural language query: When system is not able to identify concept from the natural language query, the system attempt to predict any possible predicate from the natural language query. The system adopt machine learning approach to automatically learn from triples to predict phrase or sentence as the possible predicate that can be estimated from the natural language query. The system uses the entire triple store as a training set and uses an N-gram maximum likelihood estimate to compute any possible predicate from the query tokens. Triple store contains all the triples in the knowledge base. An N-gram is a probabilistic automata for generating phrase or sentences. N-gram predict a word given the previous word.

When the system is able to detect any possible predicate from the natural language query, all the triples (subject, predicate, objects) that has the predicted predicate as its predicated is presented to the user as suggestion generated by the system from his query.For example "Who is a generous person?" has no ontology concepts in the query token. However, there is a triple in the knowledgebase (Muhammad isAGenerous Messenger), and so the system will be able to compute the predicate "isAGenerous" from the query token. In this case the triple (Muhammad isAgenerous Messenger) is presented to the user as a suggestion. When the user is satisfied with a suggestion and chooses, the system automatically parses to the retrieval module for further processing.

3.4 Retrieval module

This module take the semantically formulated query either semantically formulated automatically by the system or via system-based suggestion and retrieve the corresponding answer. The semantically formulated triple is used to formulate SPARQL query which is used by Jena inference engine to match against the knowledge base for retrieval of corresponding answer

4. EXPERIMENT

In order to experiment using the proposed semantic query formulation based statistical machine learning approach in this paper, the Quran ontology was applied. The Quran ontology was annotated and stored in the Protégée ontology editor, which serves as a knowledgebase that responds to the semantically formulated queries. The statistics shows that a total of 300 nouns, i.e. noun concepts, obtained from Leeds Quran ontology were used. The dataset also contained 1475 verbs, which were adapted from the Leeds University Quran dictionary. The

number of predicate used for the experiment contained 350 relationship obtained from Leeds Quran ontology and additional relationship added during the development of Quran knowledgebase used for this research. A total of 50 queries obtained from the Islamic Research Foundation website were used for the experiment.

5. EVALUATION

A comprehensive evaluation was carried out to compare the proposed approach with the results of existing NLI system Freya. The comparison was performed in terms effectiveness of the based suggestion suggestions. The evaluation of the suggestion approach is based on precision recall of the returned retrieved verses after suggestions were presented to the user and the user made a choice for further processing

Analysis of the effectiveness of the suggestion approach

System Name	Precision	Recall
Proposed NLI	0.61	0.69
FREyA	0.58	0.62

The results show that the suggestions proposed in this paper had a precision of 0.61 and recall of 0.69 in terms of the Quran verses retrieved, while the suggestion approach in FREyA had a precision of 0.58 and recall 0.62 in terms of the retrieved verses after suggestions were provided to the user. The proposed approach of system-based suggestions in in this paper outperformed that of FREyA in terms of precision and recall.

6. CONCLUSION

The main purpose of this paper was to examine the natural language interface which involved the process of the semantic query formulation of natural language queries, including complex and simple queries, in order to retrieve answers from Quran ontology. The proposed Natural language Interface proposes a solution in case the system is not able to semantically formulate natural language query to structured query. The system proposed solution for this by prompting dialog to the user and ask the user if he intends to reformulate the query or needs system-based suggestion. Ig user chooses to reformulate, the system enable query reformulation. And if user chooses to get suggestion from the system, system based suggestion is provided to the user based on concept matching and N-gram maximum likelihood estimation. This solution provided by the system, has assisted in reducing the number of failed queries and thus increase effectiveness of the system.

7. REFERENCES

Zou, Y., Finin, T., & Chen H. (2004). F-OWL: An inference engine for the semantic web, Proceedings of the Third International Workshop (FAABS), April 16–18, 2004 , USA.

Tablan, V., Damljanovic, D., & Bontcheva, K. (2010). A Natural Language Query Interface to Structured Information, In ESWC 2010, volume 6088 of LNCS, pages 106{120. Springer, 2010 .

Dukes, K., Atwell. University of Leeds Quran ontology. http://corpus.quran.com/ontology.jsp. Retrieved 23 September 2010.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D. and Goranov, M. (2003). KIM - Semantic Annotation Platform. In 2nd International Semantic Web Conference, Florida, USA, 2003 834-849.

Guha, R., McCool, R., and Miller, E. (2003). Semantic Search. Proceedings of the WWW2003, Budapest, 2003.

Stratica, N., Kosseim, L., & Desai, B. C. (2005). Using semantic templates for a natural language interface to the CINDI virtual library. Data & Knowledge Engineering, 55(1), 4–19. doi:10.1016/j.datak.2004.12.002

Lopez, V., Fernández, M., Stieler, N., Motta, E., Hall, W., Mkaa, M. K., & Kingdom, U. (2011). PowerAqua : supporting users in querying and exploring the Semantic Web content. Semantic Web Journal.

Damljanovic, D.; Agatonovic, M.; and Cunningham, H. 2012. FREyA: an Interactive Way of Querying Linked Data using Natural Language. The Semantic Web.125–138. Springer.