International Journal of Recent Engineering Research and Development (IJRERD)
Volume No. 01 – Issue No. 05, ISSN: 2455-8761
www.ijrerd.com, PP. 09-14

Conversion of Natural Language Queries into SPARQL Queries Using Triple Based Data Model

Mr. Sanket V. Jambhulkar
Dept. of Computer Technology
YeshwantraoChavan College of Engineering
Nagpur, India

Prof. Shivkumar J.Karale
Dept.of Computer Technology
YeshwantraoChavan College of Engineering
Nagpur,India

Abstract: In Natural Language Interface which can take the Natural Language Queries and mapped them into the SPARQL queries, the traditional keyword-base system the process on information retrieval is done only by keywords, and not by meaning of that words. The machine cannot understand the meaning and relationship between the entities which are presenting in the system. In the semantic web, the meaning and relationship are described in structured knowledge representation or ontology. In Semantic search, it provides the facility to machine to understand meaning of terms and concepts. Normal user overwhelmed by the formal logic of Semantic web. The fundamental problem is to help the users for querying ontology whose logic they do not understand. For this, the solution is provide the Natural Language Interface which can accept the Natural Language Queries and mapped them into the SPARQL queries. But such tools faces the problem of the entering the ungrammatical queries. Furthermore, these kinds of system are hard to adopt the new domains. In this proposed system, we have used the simple approach to solving these problems. Although the method is simple, it achieve the right information retrieval performance. It elaborate the things on Natural Language queries to SPARQL for ease of use and portability.

Index Terms: Ontology, Natural Language Processing (NLP), Domain-based ontology, OWL, RDF, and SPARQL

INTRODUCTION

A. SEMANTIC WEB: As the field of information technology grows, the system of the information retrieving is usually getting more and more complicated, contains the multiple sub-systems and integration of the system is remarkably important.[1] In the traditionally based system, the retrieving process is done by keyword. The majority of the data contents which are present on the web are suitable for human use. The rearrangement of the data is not done in the form of machine understandable manner. For this reason, computer applications have the problem to understand this kind of data. Fortunately, the solution of this sort of problem can be solved with the help of Semantic Web. In the words of the founder of Semantic Web, Tim Berners-Lee, "Semantic web is the extension of the current World Wide Web." [1] The purpose of Semantic web is to define the machine understandable metadata; thus, this gives freedom to computer and people to work in cooperation. The idea behind the use of Semantic web in current Web is, it allows to use the full potential of current Web and provides great flexibility to users to share, find and combine information more conveniently. The primary objective is, the application in context will try to determine the meaning of the text and creating the correlations between the terms that will represent for a user. One of the aspects to show the relationship in Semantic Web is Ontologies that enhance the understanding and description of information. [7][8]

B. ONTOLOGY: Ontology is the formal terms that describe a list of words which represent the important concepts, like classes of objects and the corresponding correlation between them. The creation of ontology is divided into three parts: ontology capture, ontology coding and possible integration with existing ontologies.[9][10][12]

C. Natural Language Interface: The relationships (properties) and the concepts (classes) of the domain are identified by with the help of Ontology. RDF triples i.e. Subject-Object-Predicate that stores the data. SPARQL is the Query Language that used to query the RDF data store which is useful to search and locate the needed RDF data. For the purpose of retrieving the data from RDF data store; the user must have mastered in SPARQL language, which is the new technology. To doing ease of this, if we provide natural language interface for the user, it will convenient for the user to refer the system easily. [13] The flow goes like this; the first user enters his query in natural language with the help of Natural Language Interface; the application converts the NL Query into SPARQL query, this query will be fired on the RDF datastore, and the required triples will be retrieved. The system is adaptable to the new domain that it doesn't need to configure the system for the new domain. [11]

D. RDF: Resource Description Framework (RDF) is a subdivision of World Wide Web Consortium (W3C) specifications originally outlined being a metadata knowledge representation. It is a regulated, labeled graph data format and its general-purpose is to represent the information on the web. [2]In many applications which are relevant to Natural Language Processing (NLP) or the Semantic Web, RDF is broadly utilized to organize data. It plays a significant part in knowledge representation and ontology. An RDF representation consists of a collection of triples. A triple combines three parts: subject, predicate, and object. Its formulation is <subject, predicate, object>. For instance, we can represent "Isaac Newton's given name is Newton" as a triple <Isaac_Newton, hasGivenName, Newton>. We can read, write and operate RDF quickly by the open source Java Project "Jena" .

E. SPARQL:SPARQL [14] is the query language for the RDF data, it is comparable to SQL and broadly utilized in the query processing and inference engine [6] like "ARQ", "Pellet", "Jena" etc. We can query a triple by any segment of the triple. SPARQL holds constraining queries, arbitrary pattern matching, optional graph pattern along with the process of conjunctions and disjunctions. We can also perform regular expression confinement by the keyword "FILTER". Either the results of SPARQL queries are results set or RDF graphs.

I. RELATED WORK

1.PANTO-Portable Natural Language Interface System: Separated from NLI system, NLIKB, which is based on the remote analytical parser, Stanford Parser couples tools such as WordNet. Numerous metric procedures are mapped into Natural Language (NL) question terminologies to *Query Triples*, which is described as the intermediate. The brief explanation of this semantics is also mapped onto the *Onto Triples* and that are linked to elements from the fundamental part of the ontology. PANTO comes with a set of eleven fact-finding mapping rules. Remain onto Triples which are described as the SPARQL queries. In general, PANTO is based on the experimental observation in which two actual phrases from the lexical tree are mapped to triple existing in the ontology. [23]

2.QACID: Ontology-Based NLI System: This system incorporates the various entertaining domain, i.e. Cinema/movie domain. This system used Spanish as the targeted language which consists of two main areas. It is *User query establishment database* and *textual-consequent engine*. But, the earlier unit of functionality is done mainly for the enlargement and system instruction plan, the other is intended to treat the unknown query. The primary objective of the QACID system is the establishment of the query in the database. The knowledge system incorporates 54 clusters, and each one of them offers one type of question, and it has a model query format that is extracted from the group of instruction data. Every cluster is associated with a single SPARQL query. A hindrance of this QACID is it is not able to counter to the unknown ontological theory, and therefore this system fails when the user posts a query which is not present in the lexicon.

3. NLP-Reduce System: This system does not correlate with any advance grammatical or semantic tools and does not store on obtaining the identical query commands to the particular memory instances. The primary phase of this system is to *generate the query* that is pledged in creating an SPARQL query for the given words. As there is no dependency on any complex NLP query management, it becomes useful portability which is defined as the primary strength of this system.[14]

4. ORAKEL- an Ontology-Based NLI System: Another NLI system that supports English factoid questions. These issues are translated into initial level logical forms. This transformation utilizes the schema of parsing and the methodology in a limited style. ORAKEL is requiring domain expert to port to another domain, and therefore it becomes the domain dependent lexicon. [13]. Ontology for appropriate knowledge base (KB) is utilized to manage the figuring process of the lexicon. A division of the terminology is naturally generated through the considered ontology. In ORAKEL, ontology is presented as the core of the entire lexicon process which will adjust randomly to the defined knowledge Base and domain. ORAKEL is one of the well-defined methods for resulting the user defined queries. The lexicon is used for accurate measuring for the development of natural language for ontology entities. ORAKEL mainly concentrate on conquering the exercise of adapting

International Journal of Recent Engineering Research and Development (IJRERD)
Volume No. 01 – Issue No. 05, ISSN: 2455-8761
www.ijrerd.com, PP. 09-14

the system in the given field. [22]. Although ORAKEL holds with a significant disadvantage which is not capable of handling the ungrammatical question and also unknown words. [22]

PROPOSED SYSTEM

In traditional parsing system, if the user enters wrong formulated or ungrammatical query the parser does not parse the sentence. In this system, we have to use the simple approach in which it only uses the stemming and synonym expansion. It only tries to map and link the words which are present in the query and their synonyms expansion that are found in the knowledge base.

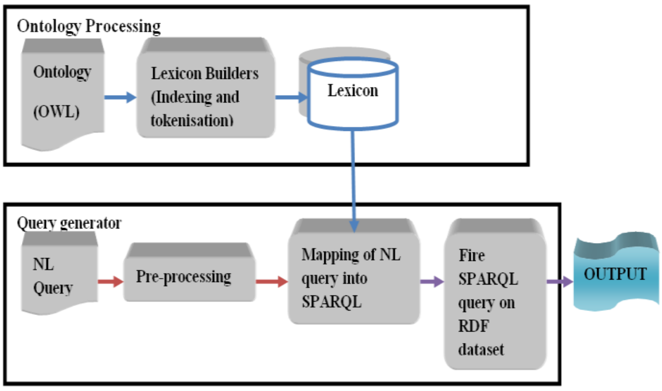


Figure 1: System Architecture

A. FLOW OF THE PROCESS

The ontology provides the vast vocabulary about the content domain. The input data in various formats have to be converted to owl/RDF data model, based on the vocabulary of Ontologies.

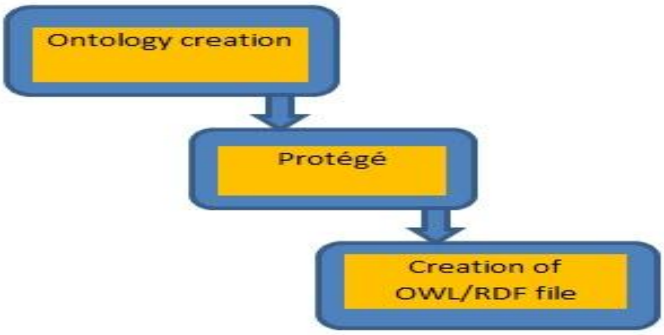


Figure 2: Creation of Ontology

B. STEPS IN PROCESSING

The different processes have explained in detail as follows:
The flow of the system starting from the entering input in Natural Languages up to the display of result to user is givenhere .

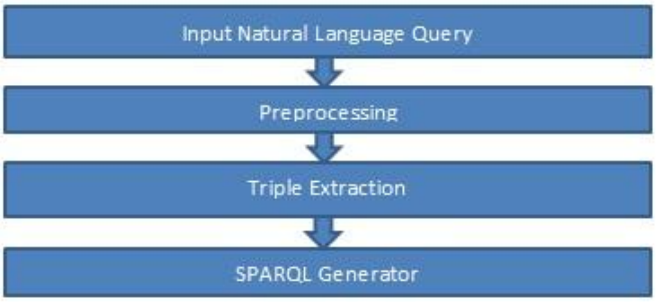


Figure 3: Different Steps in Processing

IMPLEMENTATION DETAILS AND EXPERIMENTS

We are constructing a semantic data for the educational domain that involves the construction of ontologies for various colleges such as YCCE, NYSS, SVSS but the scope not only restricted to the educational domain related data. As well as multiple ontologies are mapped by using equivalent classes and properties.Natural language interface is constructed to map the Natural Language Query to SPARQL. It is working for simple queries containing classes, properties andliterals,This system is constructed and tested using more than 500 queries related to the educational domain. The interface of Ontology is done with the help of JENA API, and Core Java.

Approach

The ontological concepts are queried by mapping the NLQ into the SPARQL query and distinguish the relations among the NL Query and the concepts of Ontology. The Natural Language Query is converted to SPARQL-based on whatever the query type is. The SPARQL query is fired with the help of Jena API. The results are retrieved and converted to the appropriate format and displayed on the screen.

Tools and Technologies

Protégé ontology editor is used to construct the ontology [3]. The above modules developed using JEE, Core Java, Jena API. The GUI is developed using JSP and Jquery. We have not used any of NLP parser to parse the sentence like Stanford Core NLP API.[4][5] The Pellet reasoneris used for Reasoning. Jena API is providing the interface between the Ontology and Java applications.

Domain Ontology Creation

This module involves the construction of educational domain ontologies of different colleges

- YCCE
- NYSS
- SVSS

Classes, object properties (attributes having value as other attributes i.e. relationships between classes) and data properties (attributes having value as a literal) are identified.

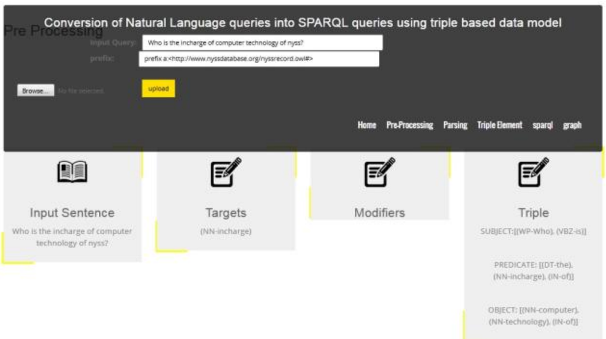
Classes	Individuals	Example of Instance
Person	HoD	*Name of HoD*
Person	Lecturer	*Name of Lecturer*
Department Name	Staff	*Total Staff*
College	Principal	*Name of Principal*

User Interface

The User Interface allows the Natural Language query and delivers the results to the user on the console.

Figure5:User Interface

Mapping of Natural Language Query into SPARQL Query and Output .This module performs the conversion of NL Query into SPARQL. The mapping of concepts in the NL Query performed with the concepts in the ontologies. The Semantic searchis used for searching the OWL data. The TRIPLE element gives us subject,object and predicate of inputted query.



.

Figure 6: Triple elemrnt

PRELIMINARY EVALUATION AND RESULT

To evaluate the performance, we implement this prototype in JAVA. We have collected 500 natural language queries of the educational domain of 3 different colleges and translated into OWL and ran the provided 300 queries on the YCCE, 100 queries on NYSS and 100 queries on SVSS Knowledge Base. As we are not using any sophisticated linguistic analysis, therefore, some queries that are provided in the educational domain could not be answered. The system successfully answered 272 queries of the YCCE queries, thereby achieving 88.23% average recall and 90.66% average precision. The system could also provide an answer for 87 queries of the NYSS queries with an average recall of 80.45% and an average precision of 87%. Note that we calculated recall and precision in a very strict manner, i.e., we assigned 0% recall as well as 0% precision to queries such as "how many lecturers are there in computer technology?" if system found 20 lecturers instead of the correct number 23 to the query.

We believe that the approach is promising. Our system processes queries as the generating triple element for finding the target and modifiers to mapping the query triple with onto-triple as they converts NLQueries to SPARQL.

Parameters for Checking:

$$\text{Precision} = \frac{\text{Retrieved relevant}}{\text{Total retrieved}}$$

$$\text{Recall} = \frac{\text{Retrieved relevant}}{\text{Total relevant}}$$

Database	#question	total result	valid result	Precision	Recall
YCCE	300	155	140	0.51	0.9032
NYSS	100	87	70	0.87	0.8045
SVSS	100	85	69	0.85	0.8117
Total	500		Average	0.7456	0.8468

Table1: Precision and Recall of the system

CONCLUSION

We think that natural language interfaces show a potential for end-user access to the Semantic Web but suffer from their inapplicability to new domains i.e. they are hard to adapt to any new system. To overcome this problem, we have introduced the system, which is completely portable and robust.

REFERENCE

[1]. Chong Wang, Miao Xiong, PANTO: A Portable Natuaral language interface to Ontologies

[2]. D. Brickley and R. V. Guha (Eds), RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation, 10 February 2004. Available at http://www.w3.org/TR/rdf-schema/

[3]. Protege Overview, http://protege.stanford.edu/overview/

[4]. The Stanford Parser: A statistical parser, http://nlp.stanford.edu/software/lex-parser.shtml

[5]. Marie-Catherine de Marneffe, Christopher D. Manning, "The Stanford typed dependencies manual" in Revised for Stanford Parser v1.6.2, February, 2010.

[6]. D.Anglunin, "Inference of Reversible Languages", *J. ACM29(1982) 741-765.*

[7]. T.Berners-Lee, J.Hendler and O. Lassila "The Semantic

[8]. Web," *Scientific American*. 284(5):35-43, 2001.A.J Gerber, A.Barnard, A.J Van der Merwe"Towards aSemantic Web Layered Architecture"

[9]. Lars Marius Garshol (2004) *Metadata?Thesauri?Taxonomies? Topic Maps! Making sense of it all* onwww.ontopia.net. 13 October 2008.

[10]. A. Maedche, S. Stabb (2001) "Ontology Learning for theSemantic Web" *IEEE intelligent Systems, Special Issue on thesemantic Web,* 16(2).

[11]. Gruber Tom (1993): "A translation approach to portableontology specifications". In: *Knowledge Acquisition*. 5: 199.

[12]. A. Maedche,V.Pekar&S.Staab, "Ontology Learning PartOne-On Discovering Taxonomic Relations from the Web"*IEEE intelligent Systems, Special Issue on the semantic Web.*

[13]. Khadija Elbedweihy, Stuart N. Wrigley, and Fabio Ciravegna. 2012. Evaluating semantic search query approaches with expert and casual users. In Proceedings of the 11th international conference on The Semantic Web - Volume Part II (ISWC'12), Springer-Verlag, Berlin, Heidelberg, 274-286

[14]. Kaufmann, E., Bernstein, A., Fischer, L.: NLP-Reduce: A naive but domain independent natural language interface for querying ontologies. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, Springer, Heidelberg (2007) Conference on , vol., no., pp.397,402, 1-2 March 2013

[15]. Fernandez, M., Cantador, I., Lopez, V., Vallet, D., Castells, P., Motta, E..Semantically enhanced Information Retrieval: an ontology-based approach. Web Semantics: Science, Services and Agents on the World Wide Web, North America, 9, jan. 2012.

[16]. The Stanford Parser: A statistical parser, http://nlp.stanford.edu/software/lex-parser.shtml Issue 2, Volume 5, 2011, pg 141-148

[17]. Hai Wang, ShepingZhai "Query for SemanticWeb Services Using SPARQL-DL" 2009 Second International Symposium on Knowledge Acquisition and Modelingpg 367-370

[18]. Damljanovic´, D., &Bontcheva, K. (2009). Towards enhanced usability of natural language interfaces to knowledge bases. In: V. Devedzˇic´, & D. Gaševic´(Eds.),Web 2.0 & semantic web (pp. 105–133). US: Springer.

[19]. Lopez, V., Fernández, M., Motta, E., &Stieler, N. (2011). PowerAqua: Supporting users in querying and exploring the semantic web.Semantic Web, 1–17.

[20]. Cimiano, P., Haase, P., Heizmann, J., Mantel, M., &Studer, R. (2008). Towards portable natural language Interfaces to knowledge bases – The case of the ORAKEL system. Data and Knowledge Engineering, 65(2), 325–354.

[21]. Philipp Cimiano, Peter Haase, J¨orgHeizmann, Matthias Mantel March 1, 2007. ORAKEL: A Portable Natural Language Interface to Knowledge Bases. (http://www.aifb.kit.edu/images/e/e7/2007_1439_Cimiano_ORAKEL_A_Porta_1.pdf)

[22]. Wang, C., Xiong, M., Zhou, Q., & Yu, Y. (2007). Panto: A portable natural language interface to ontologies. In Proceedings of the 4th European conference on the semantic web: research and applications, ESWC '07 (pp. 473–487). Berlin, Heidelberg: Springer-Verlag.

[23]. Ferrández, Óscar, Izquierdo, R., Ferrández, S., &Vicedo, J. L. (2009). Addressing ontology-based question answering with collections of user queries. Information Processing & Management, 45(2), 175–188.

[24]. BorutGorenjak, Marko Ferme " A Question Answering System on DomainSpecific Knowledge with Semantic WebSupport" INTERNATIONAL JOURNAL OF COMPUTERS