CONTEMPORARY SEMANTIC SEARCH USING
NATURAL LANGUAGE PROCESSING

Manjusha VRK1*, Lokesh Saravaran1 and Sneha Reddy1

*Corresponding Author: Manjusha VRK

In advanced search engines queries are being expressed in natural language as it seems to be the next step for developing future systems. Hence, users can express what they are searching for in a sentence form. The Semantic Web framework is our chosen field. This approach is to build a complete and meaningful Web Search using a Natural Language system. The process is complete when it covers pre-processing, Semantic analysis, semantic interpretation, and the way of executing a SPARQL query in order to obtain the results. The domain deals with the existing portal of accommodation. A corpus of different queries are being used from the facebook campaign. Furthermore, the Natural Language Understanding (NLU) module is evaluated with respect to another domain (public transportation) and a language (English). The performance of Semantic web enhances a decently sized knowledge base and also othe.

INTRODUCTION

As the count of data users are increasing day by day, it is not accomplished as a part of the web. The ultimate target of the semantic web is to heighten the ethics of the same from documents in the form of data. These are accessed through the architecture of Semantic Web and it literally means the formation of a common platform that makes the data to be shared and reused over application, outfit and community boundaries that are executed automatically. It also provides the possibility to fetch them mutually.

Semantic web doesn’t come with any explicit definition, but there are many gates. Wherever the difficulty and applications referring to the semantic web is becoming widespread now-a-days i.e. developers, implements would originate different form the Semantic web technology. The various range of applications have data integration, knowledge representation and analysis, denomination services, improving searching algorithm and as well as designs.

COMPONENTS OF SEMATIC WEB

In order to reach the goal, the utmost importance is to explain the relationship of the data over the web. This is in par to the utility of present web that connect to the current page with another. This way of hookup elaborates the relationship between the current page and the target page. The main distinction is, on semantic web, such relationship is created between any two resources, there is no sign as the “current page”.

1 Department of Computer Science and Engineering, SRM University - City Campus, Vadaupalani, Chennai

This article can be downloaded from http://www.ijerst.com/International-Conference-on-ICARSM-2015.php#1
There is also one other important difference with relationship, i.e. The hookup itself is named whereas the hookup used by the human is not inferring by the reader. Those relations provide a better and automatic data interchange.

XML plays an important role in semantic web technology. It is sketched out in such a way that information can be transported and restored, and also designed to carry the data instead of displaying the data in the form of static content. Since XML self-descriptive it is easy to develop and establish. The XML document is a tree structured form which start with the root element and it should contain the child and sub-child elements. XML is validated using DTD in order to become well-formed XML document which will define the legal building block of an XML document. The most important aspect while creating XML document is namespaces. In order to avoid the name conflicts, XML provides namespaces which will make the XML document unique from other XML documents and it is created by the developer of their own.

The basic building block for data interchange is Resource Description Framework (RDF). RDF is a World Wide Web Consortium standard for describing the resources available over the web. It identifies the URI and uses simple comments called triples to describe. To read and write the RDF in XML syntax, one should be convenient with graph-based information model which is regarded as the heart of the RDF. RDF graphs are plotted in XML and this will make use of some characteristic features. It is necessary to be strong in the basic structure of XML and the way the XML reveals as nested hierarchies of opening and closing tags and also comfortable with DTD and XML schema concepts. The schema of RDF specifies the blueprint for encoding the point-edge-point or terminal structure that RDF bothers about in terms of the entity hierarchy that XML cares about. RDF 1.0 version embraces a pattern that we termed ‘striped’. In the following example “there is a person named ‘John’ and he ‘dwell With’ the one who has a father named ‘Fred’”. That’s all our example pieces of RDF/XML as briefed above. The RDF terminal-and-terminal view of this is shown below graphically. To understand the stripe, we need a comparison of the abstract graph structure of RDF with the details of XML nesting structure.

Ontology has many translations and for the purpose of this paper, Ontology is a directed description concept in a domain of thesis, properties of each concept describing various features, aspects of the concept and limitations on slot. Ontology gives a general lexicon for researchers to share the data in a region. The reasoning for enhancing Ontology is to distinguish domain knowledge from operational knowledge. This will provide the user an idea in order to get the accurate result. Ontology jointly with the set of individual occurrences of classes covers a knowledge base. In order to develop the ontology, there are few terms such as,

- Arrangement of classes in an ecological hierarchy.
- Defining the slots and specifying the values for such souls.
- Filling the values of slots for occurrences.

Tools that describe the query information through relationship is done by SPARQL. It helps to extract the values from the structured and semi-structured data represented in RDF. It helps in performing complex joins of dissimilar RDF pool in a single query. SPARQL is a RDF query language which will get and process the data that
stored the RDF pool. It is made as a standard format by the W3C. SPARQL provides access to the user to write the queries adjacent to the data that are likely to be called as "key-value" in which data will follow the specification by RDF and the entire database will become a package of "grammatical" triples. SPARQL/RDF becomes robust and easy for the lists that contain mixed values and also the lists themselves contain the joining variable.

RELATED WORKS
Prior to the related work, there has to be a brief discussion about the jargon of present research in the Natural Language Interface. Naturally, while accessing the system through natural language the Natural Language Interface (NLI) are used. Such kind of systems is operating on layout information and for the given the Natural Language questions it finds out the exact and absolute answer. These systems are classified into division of classes. Natural Language Interface to Database (NLIDB) which contains the data in a relational database. The fundamental of NLIDB is suitable for the semantic web which results in a Natural Language Interface to Knowledge Base (NLIKB). The information stored in the form of an ontology plays a vital part in the semantic web. In this system, NLIKB is developed in a special case that search engines can fetch the set of accurate result according to the user’s Natural Language Queries.

PowerAqua-An ontology based NLI System
This paper deals with the poweraqua which is an ontology based NLI system which exceed the existing system by holding multiple ontology source and high scalability. This system is also defined as the multi—ontology based question answering (QA) system which fetch the input (given by the user as a query) and it is capable of retrieving the result through ranking and aggregate the result of the compatible distributed resource present in the semantic web. PowerAqua doesn’t have constrains with single ontology where present existing Natural Language (NL) system falls under the constraints of using only one ontology. Therefore PowerAqua provides the first complete set of supporting the open domain question answering under the semantic web. Hence poweraqua becomes the famous Question and Answering system over the wide scale and multi semantic web. [4]

ORAKEL- An ontology based NLI System
Another NLI system that supports English factoid questions. These questions are initially translated into primary level logical forms. This transformation uses the schema of parsing and the methodology in a confined style. ORAKEL requiring domain expert in order to port to another domain and hence it becomes the domain dependent lexicon [13]. An ontology for particular knowledge base (KB) is used to lead the figuring process of the lexicon. A sector of the terminology is naturally generated thought the considered ontology. In ORAKEL, ontology is presented as the core of the entire lexion process which will adapt randomly to the defined domain and knowledge Base. ORAKEL is one of the well defined process for resulting the user defined queries. The lexicon is used for accurate measuring for the construction of natural language for ontology entities. ORAKEL mainly focus on reducing the exercise of adapting the system in the given field. [6]. Although ORAKEL holds with an important drawback which is not capable of handling the ungrammatical question and also unknown words. [13]
PANTO-Portable NLIKB system

Apart from NLI system, NLIKB which is based on remote analytical parser ,Standford Parser combines tools such as WordNet. Several metric procedures are mapped into Natural Language (NL) question terminologies to Query Triples which is represented as the intermediate. The brief description of this semantics is also mapped onto the Onto Triples and that are linked to elements from the basic part of ontology. PANTO comes with a set of eleven fact-finding mapping rules. Outlast onto Triples which are represented as the SPARQL queries. In general, PANTO is based on the experimental observation in which two factual phrases from the lexical tree are mapped to triple present in the ontology. [7]. In our proposed system, the set of eleven mapping rules are refined in order to make the system more effective.

QACID-ontology based NLI system

This system covers the most entertaining domain, i.e. Cinema/movie domain. This system took Spanish as the targeted language which consists of two main sectors. It is User query establishment database and textual-consequent engine. But, the previous unit of functionality is presented especially for the expansion and system instruction plan, the other is designed to process the unknown query. The main objective of the QACID system is the establishment of query in the database. The information system incorporates 54 clusters and each one of them produces one type of question and it has a model query format that is extracted from the group of instruction data. Every cluster is correlated along a single SPARQL query. A hindrance of this QACID is it is not able to respond to the unknown ontological theory and hence this system fails when the user post a query which is not present in the lexicon. [8]

NLP-Reduce System

This system does not associate with any leading grammatical or semantic tools and does not store on finding the identical query commands to the particular memory instances. The main phase of this system is to generate query that is pledged in creating a SPARQL query for the given words. As there is no dependency on any complex NLP query management, it becomes good portability which is defined as the major strength of this system. [3]

SYSTEM ARCHITECTURE

This system is a C# & .Net based operation which consists of various elements. Fig. 1 shows the entire architecture of the Semantic Web Search Using Natural Language system along with the data-flow with a sample query.

The user makes the request in the Natural Language and that is given as an input to the system. In the proposed system, there is no constraint on query length and hence users are free to use any form of keywords, sentence(s) or as a paragraph. We figure out this system in the English language. The details about the NL query are explained in Section 5.

User queries are analyzed under the Natural Language Query component. The independent Natural Language Query-KB is produced through this component. The Natural Language Understanding (NLU) element is entangled with three blocks like Pre-processing, Named Entity Recognition (NER) and Statistical Semantic Analyser. The details of NLU model are described in Section 6. Here, the proposed system is field independent with a set of limitations for the manifestation of experimental and philosophical values of the Natural Language System. The proposed system is suitable for any domain and
it presents both capacity and limitation of the system. We determined not to create any collection of data from the existing work and three matters are compressed here. Firstly, the existing English corpus has the constrains for the user queries which has to be in a single sentence and Secondly, from our point of view, most familiar Mooney Geoquery [5] is a most common substance for question and responding task where the proposed system substitutes the existing systems. In all existing system’s user are restricted to give the input query in a single sentence form and our main aim is to break the restriction and allow the users to give the queries in one or more sentence. On the other hand, the KB in our system would not contain answers for all the NL queries so in order to overcome this, we request each user to enter three queries and which are stored in Natural Language Corpus.

Table 1 shows the statistical report of our Natural Language Corpus even though there was no restriction of queries in the form of length.

<table>
<thead>
<tr>
<th>Table 1: Various NL Query Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus statistics</td>
</tr>
<tr>
<td>Queries consisting of 1 sentence</td>
</tr>
<tr>
<td>Queries consisting of 2 sentences</td>
</tr>
<tr>
<td>Queries consisting of 3 sentences</td>
</tr>
<tr>
<td>Average query length</td>
</tr>
</tbody>
</table>

In the semantic web, ontology plays a vital role for domain modelling. Hence we are using knowledge Base (KB) along with the perspective ofontology. For extracting data from the website, Source Website Component is used where Novotel, milestoneinternet and hotelmarivaux sites are used which provides large database for the different accommodation domain. By using
the Web crawler hyperlink of the websites are identified and add them to the list of URL's for future visit. Data's of those websites are extracted and exchanging through Information Exchange. Both Web crawler and Information Exchange are present in the Knowledge Base Creator. In this system Knowledge Base (KB) is completely developed to transact with the existing data as the work required to develop ontology from the scratch is difficult. Result Searcher contains two kinds of searching methods. They are 1.Structured Search and 2.Fulltext Search. On one hand, for each and every NL query an equivalent SPARQL query is developed manually. We define some constrains for NL query as well as SPARQL query to ensure the accommodation domain satisfies that the constraints are fetched from the knowledge base using ontology (KB). This may harm the output i.e. Partially satisfied results are discarded. On the other hand, every accommodation text has its own data which is presented in structured properties. These fields are indexed and searched through the fulltext Search. The entire or total process of drafting the proper and accurate result is performed as follows.

- If the Natural query is fully conveyed in SPARQL, then create such kind of query and summate to the appropriate result.
- If the query consists of other terms which are not expressed in SPARQL, the structured search output with fulltext search has to be refined and validated.

As the output, lists of matching results are displayed to the user with their properties.

**OVERVIEW OF SEMANTIC WEB TECHNOLOGY**

The semantic Web is an advance technology that continues to signify the research which introduces the standardized Knowledge Base representation language for ontology, the root part of semantic web. This technology includes RDF Schema, Web Ontology Language (OWL), rules and SPARQL which is represented as query language. Here we are discussing about the some topics under Semantic Web Technologies which are listed below.

### i. Ontologies

They define the theories and alliances that describe and act as an area of knowledge and it is also used to categorise the terms in certain applications, characterize possible relationships and define possible limitations on using those relationships. Partially, Ontologies can be very complex or very simple depends on the place where we implement it. The part of ontology and laws of the semantic web is unique, i.e. To help data consistency for example, uncertainty may exist in the terms used in various information sets, or when a portion of added knowledge can lead to the discovery of new relationship.

Transitional which must finally convert into machine understandable format i.e. Ontology. The Process of gaining, maintaining, enhancing the domain ontologies is defined as “Ontology Engineering”. The complexity of the Ontology depends on the scale of the website. Fetching ontological information from the web is a very difficult charge to handle and so [14] the term ontology learning was introduced. Ontology learning is a way of extracting meaningful or atomic information from the web.

### ii. Ontology Language

In the recent course of development, many ontology languages have grown with the industry and very soon they will become the languages to build Ontology under the circumstance of
This article can be downloaded from http://www.ijerst.com/International-Conference-on-ICARSM-2015.php#1

Resource Description Framework (RDF) and RDF Schema

It is the standard form for the data interchange mechanism on the web. RDF has the feature which facilitates data merging even if there is a difference in the hidden schema, and it also supports the transformation of schema over time without any requirement of all the information to be changed. RDF intensifies the linking structure of the web to use the URIs for naming the interconnection between things as well as the nodes of the link (which is called “triples”). With this sample model, it allows structured and semi-structured data to be joint, exposed and transferred across different applications. The hookup structure designs a guided, labelled representation, and the nodes show the named hookup or link between two resources, presented by the graph nodes. The graph view is the easiest and best possible model for RDF and it is simple to understand imaged justifications.

RDF statements can be encoded in a number of ways, if it’s XML based or not. No matter which of the above formats are used to express the information represented as RDF triples or statements and the particular format is meant to be a “Syntactic Sugar”. Most of the RDF tools can parse several types of serialized formats. The Resource Description Framework data model is not providing the mechanisms to describe the link between properties and resources; hence RDF Schema allows to define the relationship between them, which is the frame based approach and this RDF (S) are widely used in many tools as well as projects. [11]

Even though RDF (S) provides various features on developing the ontology, there are also some problems.

- RDF (S) is too weak to describe the resource with all necessary details.
- There are No localized level and field limitations.
- There are no presence or entity constraints.
- There are no transitive, inverse and symmetrical properties.

iii. SWRL & SQWRL

Every language has its own rules and regulations, similar way, Semantic Web-Ontology languages has its own criteria with regard to language like Semantic Web Rule Language (SWRL) and it is determined to be the scale language for Semantic Web. All the criteria are given in Ontology Web.
Language concepts (classes, properties, individuals, literals…) It also expresses the logics by combining with subdivision of ‘Rule Mark-up Language’.

SQWRL is nothing but, Semantic Query-Enhanced Web Rule Language which is pronounced as a squirrel and it is a SWRL based language for ontological queries based on OWL.[12] It performs like Structured Query Language and provides SQL operation in order to obtain data from the knowledge base. Basically, there are two features in SQWRL Language; they are Code Operator and Collection Operator. On one hand, code operator is a language which uses a SWRL rule proform as a plan specification and replaces the rule consequently by selecting and formatting operations. SQWRL uses SWRL’s internal provision as an extension point. The main key operator is sqwrl:select. It takes up one or many arguments that are considered as the variables for pattern specification of queries and build a relation using those parameters. On the other hand, Collection Operator provides some advance features of querying process, SQWRL has range of operators that provide higher grouping, gathering functionality and limited forms of inversion as failure and so on.[12]. As SQWRL using the SWRL’s built-in facility and those built-in describe a set of operators which are later used to build the retrieval specifications. Hence, existingSWRL, are used to build and edit the SQWRL queries in addition to that serialized mechanism are also used, by using it the queries are stored in the ontology.[13]

**SEMANTIC DESCRIPTIONS OF NL QUERIES**

Some conventional way to express the semantic NL queries is framed, FOPL (first order predicate logic), semantic trees etc… In our proposed system, NL queries are described based on the semantic trees. Semantic trees will follow the structure, e.g. Utilizing the architecture for adding the protocols (constrains) to the semantic trees. As we discussed earlier, in the Semantic Web, the role of ontology is to store the domain information in the Knowledge Base (KB) whereas it serves as a tool. However, here the glossary (Annotation) of NL queries based on ontology has an indirect connection with the ontology for acquiring the field information in the Knowledge Base (KB). This clearly means that our glossary of NL queries is fully independent to a particular Knowledge Base (KB). Hence ontologies are used exclusively for defining the NL query semantics.

**Ontology For NL Query Semantics**

Figure 3 shows the field independent Natural Language Ontology based on queries. Each and every spell serves as an occurrence of Word Ontology section and also few grammatical keys are added to it. hasNextWord Property is used to link all Word instance. The sentence Class instance, will cover-up the entire NL query. Named Entity is supported by Ontology.

Through startWithWord, an instance of Named Entity is linked to its Word content and its length is stored in a property called numberOfWords. In general named entities are represented as an expression which carries unique meaning like cities, people and countries. In our proposed system, it is step beyond the traditional way, by extending the usage of the named entity to all words which contains important semantic information about NL queries. Hence, named entity is presented as both “traditional” like accommodation type, place,
facilities etc... and “specific domain knowledge & Key phrase” like relative location and price. Taxonomy which is formed by named entity is also shown in the Fig.3. For some particular domain, this taxonomy is lengthened as ontology. Triples are created for each and every query. A free and open source ontology editor tool is used to develop the ontology with knowledge, i.e. Protégé  In which Natural Language query ontologies are represented in OWL format.

NL Query Corpus Annotation
Two independent annotators are annotated to the NL query corpus. For this assignment, we created an annotation tool which provides equivalence partitioning between Utility and Robust nature. The complexity of the triples and sentence ontology were hidden by the editors. [2].

SEMANTIC ANALYSIS OF NL QUERIES
The goal of understanding the Natural Language component is to develop a semantic briefing of formerly hidden Natural Language queries. Our Natural Language Understanding (NLU) element is developed on the basis of supervised learning and statistical model. More over the newly created procession appliances are also used by this element. Each user query is pre-processed using tokenizer, parts-of-speech tagger etc…. [14]. In query pre-processing a complex task is in recognition of Named Entity. To achieve this we propose three Named Entity Recognition (NER) approaches.

- **Maximum Entropy NER**, which is used to find some general entities like place and city.
- **LINGVOParse** which is based on handwritten grammars, is used to find the Named Entity of difficult structure like Date, Currency or Number.
- **String Similarity Matching** is a different NER approach which is used to deal with named entities to identify the NER which is not recognized in previous NER’s.

All these are generally domain-dependent entities. **Ontology NER** is one of the main features in our proposed system in which NLU unit is autonomous for back-end Knowledge Base (KB).

Few related works [1], [2] uses KB as an important source for identifying several domain entities The vocabulary based NER is another approach for named item candidates which use occurrences from the Knowledge Base (KB).

SEMANTIC INTERPRETATION SEARCH
Having found the annotation of Natural query which is done with its meaningful description, it must be disturbed so as to find the best result. Some memory is independent of our semantic interpretation and to fulfill the search request in Knowledge Base (KB) it must transform into ontology query language (SPARQL). Hence, the annotation of NL query is given as an input and a SPARQL query at the back-end KB is produced as an output.

A set of field oriented existing manual criteria are there for the transformation. Each logic processes a specific NL query annotation and eventually a SPARQL is generated.
MATCHING ENTITIES WITH ONTOLOGY INSTANCE

Few titled entities of NL query are identical to an ontology instance like City, Place etc. Based on the string similarity the mapping is done between named entity and KB instance. We are supposed to build a small subgroup of thought instance combined with entitled entities to select the best
set of character metrics. Approximately 50 pairs are made in our proposed system and also closeness of several types of string metrics is measured. 96% closeness is attained by the Jaro-Winkler distance which is the best accuracy ever. [15]. Three errors may occur under this technique. First, titled entity is declared in a manner in which the correlation amongst the KB instances and the words must be very less. Second, the best entitled entity has no identical instance in KB. Third, uncertainty within the KB, i.e. Two identical occurrences have two identical descriptions.

INTERPRETATION OF NAMED ENTITIES WITH NUMBERS AND DATES

Entity types namely the Number and Date are not linked with any ontology. As we discussed earlier LINGVOPraser tool is used to identify the named entity types and it also holds the internal support for semantic perception of named entities. This tool makes use of manual, semantic grammar for the converting into a computer based sketch which allows integrating and resulting in a SPARQL query. A set of accommodation instance is the result of the semantic search and the system will splash the entire fetched accommodation instance in a list along with its properties.

EVALUATION

The above figure represents the Home Screen of the Semantic Web Search Using Natural Language

This figure(s) represents, each and every users are provided with a search box and through which user can express in their own words and look for appropriate result. User’s are free to give the query in either single or multiple words as well as a sentence. As shown in the figure the user is asked to type the query and on clicking “search” button, the appropriate result is fetched and displayed.

The above figure shows that, an appropriated result is given back by the system to the user

This article can be downloaded from http://www.ijerst.com/International-Conference-on-ICARSM-2015.php#1
which matches the user’s query. Only the accurate result has been fetched out and all the result displayed to the user will match the user requirement.

This final figure let you know that, the user is given with the accurate result which matches the user query. The Admin who extracted the data from a web page is similar to this figure and so the user query matches with the data set fed by the admin and the result is displayed to the User.

CONCLUSION

This concluding section draws the conclusion of this article and as well as open issues in this system. The Semantic Web Search using Natural Language is very complex and at the same time challenge task to do. To make this system in a real-world this could be deployed for commercial use and lots of theoretical as well as practical problems must be resolved. At present we examine those issues and we will propose solution for some of them.

Here we are discussing about the Performance Issues. Performance of ontology reasoning restrict from testing our system in the real Web environment which is one of the most important issues. Approximately 15 minutes are required to answer the single SPARQL query and we found this to happen because of the size of the KB. However, on the core of our knowledge, we cannot improve this without altering the back-end Semantic Web tool. Some practical issues related to KB are discussed here. Often Structured Data’s are missing out which turns to most important problem i.e. Data fetching from the source web site is mostly insufficient and it is not wise to verify and validate the data mutually. This problem often occurs in the system that uses the public Web source and this will populate the KB as well. Another problem is, there is no consistent source data. To overcome these problems, a possible solution is to do Post-Processing.

FUTURE WORKS

There are four areas which are worth to explore the future, they are:

1. Larger NL query corpus. The major limitation in our system is the small-scale corpus. Thus, it is good to gain more data by deploying the system in testing server and hence accessible by the public.

2. Accommodating full text search. Based on some research, both structured search and full text search are integrated which is promising future task.

3. Better Named Entity Recognition. The final output clearly shows that NER element is most important during the syntactic search. Better NER will result in revised performance of the same and most correct or accurate search results will obtain.

4. Performance Improvement. The performance of this system can be improved by substituting
the back end with more reliable and flexible tools or a relational database can also be used.

REFERENCES


11. Asunción Gómez-Pérez and Oscar Corcho, Ontology Languages for the Semantic Web, Universidad Politécnica de Madrid.

12. SQWRL: http://protege.cim3.net/cgi-bin/wiki.pl?SQWRL

13. Martin O’Connor and Amar Das, SQWRL: a Query Language for OWL; Stanford center for Biomedical Informatics Research, Stanford, CA, 94305

