Scholars Journal of Engineering and Technology (SJET) **ISSN 2321-435X (Online)**
Sch. J. Eng. Tech., 2014; 2(2A):193-196 **ISSN 2347-9523 (Print)**

(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

Research Article

Handling Unstructured Data for Semantic Web – A Natural Language Processing Approach

Hemant Kumud
Apaji Institute, Banasthali University, Banasthali, Rajasthan, India

***Corresponding author**
Hemant Kumud
Email: hemantkumud7@gmail.com

Abstract: World Wide Web is a large repository of structured data and un-structured data. Structured data has a well-fixed structure thus is easy to analyze and use with computational machinery. On the other side handling of unstructured data is a challenging task for knowledge base. This paper describes Natural Language Processing techniques for handling unstructured data and its usefulness for semantic Web so that knowledge extracted from un-structured data can be useful for searchers ease of access and extraction of relevant documents.
Keywords: Semantic Web, NLP (Natural Language Processing), Ontology and Knowledge Base

INTRODUCTION

Web is a shared repository where data, text, images, and scanned documents are located. Data can be described in two aspects: structured and unstructured data. Structured data has a structure in terms of grammar pattern and contextual relations. The other one has no specific structure but it may have grammar. Posted queries and answers on the page, advertisements, graphics, text, emails, presentations and so forth are included in the unstructured data. Text on a web page is expressed in the form a natural language (NL). Each NL has grammar or say rules to express thoughts, so in both the cases grammar is specified with them. Moreover, unstructured data may not be fully described in relational form.

Semantic Web strategy was initially proposed by Tim Berners-Lee to add semantic information on the Web. When a user types some key words in the search engine, the results include the massive contents which are not user's required information. Keywords search techniques fail here to get required information with the colossal information. Therefore, the focus moved from original Web to the Semantic Web for fast related and precise information access. Many fields of computer science such as Data Mining, Information Retrieval, Database Management and NLP have been introduced with Semantic Web for machine supported data interpretation and process integration. Knowledge is extracted from information contained in both unstructured and structured data to form a knowledge base. Ontologies are also introduced to turn the Web into knowledge base for interpretation, integration and sharing of facts.

This paper describes how NLP layers and techniques can be involved with Semantic Web to process structured data hidden within the unstructured magnanimous data on the Web and depicting their effectiveness. In section 2 we briefly provide the amalgamated study of NLP and Semantic Web strategies also giving a brief review of the work done in the area. Section 3 describes effectiveness of NLP techniques for supporting unstructured and un-interpreted data and how deconstructed data which is extracted form unstructured Web is valuable for Semantic Web. Section 4 describes relationship between ontology modelling and NLP. Section 5 concludes the work done along with future trends.

Semantic Web and NLP

The Semantic Web is considered as an effective grounding for improving the visibility of knowledge source on the Web. However, indicates advancement to the original Web for automatic information analyzing and processing by artificial agents. The core part of Semantic Web is ontologies, which defines the relationship between related entities. For domain based ontologies, related and specific entities to the particular domain are extracted from various sources of data and then semantic data is located in a semantic store for accessing. Data representation models such as RDF(S) (Resource Description Framework/Schema) and OWL (Web Ontology Languages) recommended by W3C which convey a standardize approach for developing ontologies. W3C has standardized a layered cake of ontology languages where each layer plays a different

role and based on the top of other. Figure 1[1] depicts the Semantic Web frame work which is also referred as layered stack, RDF data model is fabricated upon the syntaxes of URIs and it includes XML syntax for describing resources. RDF uses the concept of URI-Uniform Resource Identifier) which is used to describe resources such as places, documents, graphics, etc. URI is a unique identifier for a thing.

RDF Schema and OWL (Web Ontology Language) are ontology languages which are described using RDF. RDF describes the properties of a resource whereas RDFS defines the relationship between the attributes and resources. OWL extends the RDF and RDF(S) and proposed to be used when information contained in the documents needs to be processed.

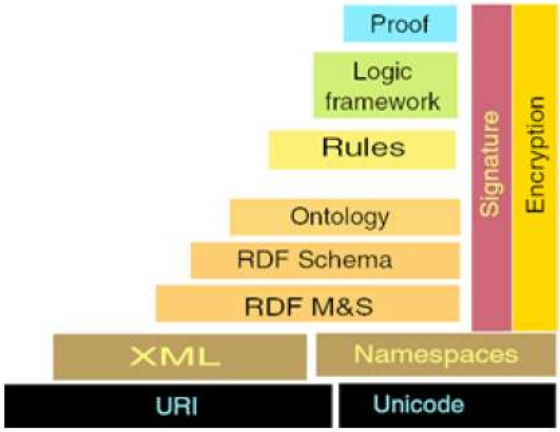

Figure 1: Semantic Web Layered Stack

Ontologies and reasoning rules are applied to reason about data and infer new information. Rules are nothing but some condition or restriction to be applied on data to draw some facts. The proof layer defines facts that an agent must be able to explain how it has came at a given conclusion. In fact, Semantic Web is like a collection of related and clustered facts. New facts are added anytime, this is a continuous process. Based on ontology descriptions, classification of ontology languages are: logical language, frame based language and graph based language. Ontologies along with Semantic Web can be mixed with growing fields of computer science where automatic extraction of knowledge is required.

NLP is a growing field of computer science. Being a branch of linguistics, sometimes it is referred as natural language understanding. Its research includes text summarization, question and answering, evaluation and ranking of machine output, word sense disambiguation, machine translation, NER (Name Entity Recognition), part of speech tagging, parsing, morph analysis, natural language generation, retrieval of information and mining of data etc. Most of the research is related to the meaning understanding and sense representation of the data. NLP techniques support in getting a structure concealed within the unstructured data on the Web for useful and meaningful information seeking rather than information extraction.

Word documents, Presentations, Images, We pages, PDFs and so forth contain text that is expressed in one of the natural languages. Text is handled by different layers of NLP. Figure 2 describes the NLP layers.

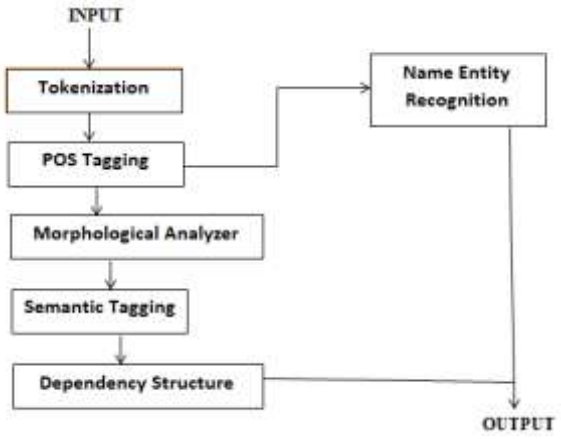

Figure 2: NLP Layers

Tokenization defines the individual words involved in the sentence or paragraph. Part of Speech Tagging describes the category and role of a word. Morphological Analyser shows the root word along with its features such as gender, case etc. Semantic Tagging refers to assigning sense to a word or phrase and can be connected with Semantic Web for semantic annotation and for the ontologies development also. Dependency phrase defines the role of the object in relation to others. NLP attempts to enable computers to make sense of human language. System machinery works on patterns for processing. Intelligent agents make use of NLP layers for fetching patterns from data. The next section includes how layers support for unstructured data.

NLP for Unstructured Data

A substantial share of information of industry, institution and individual is stored in text database which consists of collections of documents from several sources such as articles, research papers, books, digital libraries, blogs, email, messages and Web pages [2]. For finding pattern from these sources, pre-processing of the source documents required which is supported by the NLP techniques. The techniques are like stemming (finding stem) after removing suffixes and prefixes, lemmatization for replacing inflected word with its base form, Part of Speech (POS) tagging for finding grammar category of language such as noun, pronoun, adverb, adjective, proposition and etc., semantic tagging for assigning meaning based on the POS and local context of the text. Name Entities (NEs) occur densely in a Web directory, as a Web directory is a knowledge

base, which unavoidably has many proper nouns that describe entities in the world [3].

Every language has ambiguous words (word having more than one possible sense). If we refer to the word "Park", it may come in context of garden if working as Noun or it may work in terms of an activity if working as a verb (to park something). For example we usually say: "Park the car in the parking area." So meaning of the word depends on the association of words in the sentence and its contextual relation with them. Word Sense Disambiguation (WSD), a part of NLP, is used to disambiguate the sense of the word. In the above sentence "Park" is referring to a piece of ground to park something there, we come to know for this by the use of the word "car" and its POS category in the sentence. Human being quickly makes a conclusion in which context the word is being used. But how can the system make conclusion for the particular one. For this learning models are applied on systems that is achieved with the use of probability based on given evidence and then term is selected as per given probability. Semantic Tagging –Names: Recognizing names from the text, it could be a place (country, state, city etc.) name, person name, organization name, community name, date pattern, e-mail pattern, and time etc. These classification and identifications of names can be merged with information extraction which is based on ontologies.

Ontology Modelling and NLP

Ontology is a description of things that exist and how they relate to each other. It is a study of categories of things and their relation among them. For the successful understanding of natural language, there are two phases: identify the entities in the particular domain and understanding their relationship not only with the current domain but how these entities relate to each other that lie in the external domain. NLP and ontologies can be often seen contrary to each other. As above, ontology model is a classification of entities and models the relationship among those entitles while the aim of NLP is to identify the entities and understanding of relationship among those entities. Using ontologies with NLP, understanding of natural language through systems become smarter enough to make inference and respond with defined and relevant result what a user requests. When, working with domain ontology, collection and description of concepts collected for domain corpus where ontology concepts exist. The process is repeated for extending the ontology with relevant related concepts. Then a relationship between concepts is established for showing concepts dependency. Concepts relationship process is an important phase for expanding the domain store and as well as for covering at most related terminologies in the specific domain. Dependency structure, one of the NLP techniques can be applied for showing specific ontology domain concepts dependency. For example, there is a sentence which is related to computer science domain: "on-screen keyboard displays a virtual keyboard on computer-screen." The dependency structure for the sentence is shown in figure 3.

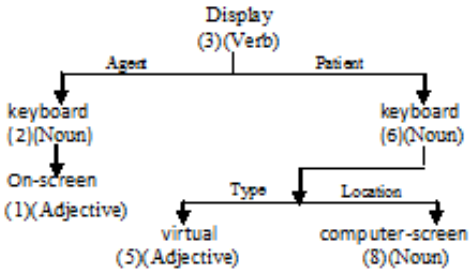

Figure 3: Dependency Structure

A dependency relation is represented by a tree structure as shown above. As figure 3 shows: "display" is the head both "keyboards" and dependents are shown at the lower ends of arrows. The numbers in the tree represent location of each word in the sentence. Using these dependencies, connections between ontology concepts can be evaluated. These are:

1: nadj (2-keyboard, 1-on-screen)
2: nsubj (3-display, 2- keyboard)
3: nobj (3-display, 6- keyboard)
4: nadj (6- keyboard, 5-virtual)
5:prep_on (6- keyboard, 8-computer-screen)

Appropriate concepts that have relevance to the domain ontology are extracted from these connectives. Here, extracted ontology concepts are:

1-on-screen keyboard
2-virtual keyboard and
3-computer-screen

The above example is for the modelling of domain ontology. Likewise, Ontology acquisition process can be applied to make ontology store huge. Hence, NLP techniques support for Semantic Web development with ontology learning and ontology mapping. Moreover, NLP technology shows a quite influential role in ontology life cycle process which in turn thrive Semantic Web.

Natural language processing within ontology development has been focused by a numbers of researchers. SOFIE , a system for ontology learning and modelling is based on natural language parsing and logical reasoning for disambiguation which is created by Suchanek et al. [4]. OntoGen [5] is a system for subject ontology construction, which uses the vector-space model for document representation and operates based on a cosine similarity between textual documents. Text2Onto [6], uses hyponym extraction on the basis of patterns and also combines machine learning

approaches with basic linguistic processing such as tokenization, lemmatization and parsing. SPRAT [7] is also based on lexico-syntactic patterns for ontology development and acquisition.

CONCLUSIONS

The goal of Semantic Web is to automate software agents for retrieving relevant and required information rather than pulling out massive unrelated data. NLP techniques with Semantic Web provide the capability of turning original web to Semantic Web while dealing with a combination of structured and unstructured data. The fluency and adequacy of results can be acquired when NLP techniques are applied, that is not possible using traditional and relational techniques.

REFERENCES

1. Berners Lee T, Hendler J, Lassila O; The semantic web. Scientific American, 2001; 30–37
2. Gharehchopogh FS, Khalifelu ZA ; Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing. Application of Information and Communication Technologies (AICT), 5th International Conference, 2011.
3. Zaihrayeu I, Sun L, Giunchiglia F, Pan W, Ju Q, Mingmin Chi, Xuanjing Huang ; From Web Directories to Ontologies: Natural Language Processing Challenges. In 6th International Semantic Web Conference (ISWC), 2007
4. Suchanek FM, Sozio M, Weikum G; SOFIE: A Self-Organizing Framework for Information Extraction. Proceedings of the 18th international conference on World Wide Web, 2007; 631-640
5. Fortuna B, Grobelnik M, Mladenić D; OntoGen: Semi-automatic Ontology Editor. Proceedings of HCI, 2007;309-318
6. Cimiano P, Volker J ; Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery. Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), 2005; 227–238
7. Maynard D, Funk A, Peters W; SPRAT: a tool for automatic semantic pattern-based ontology population. Proceedings of the International Conference for Digital Libraries and the Semantic Web, Trento, Italy, 2009.