IJIRMET ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 1, Issue 5, June 2016

DOMAIN SPECIFIC AUTOMATIC ONTOLOGY POPULATION FROM UNSTRUCTURED DOCUMENTS

[1] Sreetha S, [2] Kavitha Raju, [3] Raseek C
[1] [2]Dept of Computer Science and Engineering,Govt. Engineering College Sreekrishnapuram, Palakkad, Kerala, India
[3]Asst. Prof. Dept. Computer Science and Engineering. Govt. Engineering College Sreekrishnapuram,Palakkad, Kerala,India
[1] s.sreetha.1@gmail.com , [2] kavitharaju18@gmail.com , [3] raseek.c@gmail.com

Abstract: ***Ontology provides a shared common vocabulary for a specific domain. It is widely used as a formal structured knowledge representation tool for domain knowledge, which can be shared and reused. Ontology development is equivalent to defining a set of data and their structure for creating a variety of semantic web related tasks. The creation and population of ontology is a tedious task, if the knowledge to be encoded is to be extracted manually. This paper proposes an automatic method to populate ontology from unstructured documents by Information Extraction. Various Natural Language Processing techniques, that make use of machine learning models, are used in the realization of the Information Extraction system. It has been observed that, customizing the IE system for the underlying domain ensures improved accuracy and efficiency in information extraction. The proposed work is done in the domain "New Appointments in companies".***

Keywords: ***Information Extraction; Ontology Population; Natural Language Processing***

I. INTRODUCTION

Nowadays people use Internet to search for any information. Since the amount of data present in Internet is very huge and unstructured, it is difficult to extract information from that. Also we may get many irrelevant information. To solve this problem, the concept of semantic web has arrived. Semantic web contains linked data. It helps us to find, share, reuse and combine information in an easier way. It provides a common way of representing the data that can be understood by machines.

Information Retrieval systems returns a set of relevant documents based on the query. Keyword matching is the basic technique applied behind every IR systems. Then a ranking algorithm is also applied to those documents. Based on that rank the search engine displays the results. Whereas the Information Extraction is the process of obtaining pertinent information from the documents [4]. IE (Information Extraction) systems extracts information based on the semantics contain in the documents and query. Various Natural Language Processing techniques are required for semantic understanding.

Ontology can be considered as the backbone of semantic web. It is widely accepted that ontologies can facilitate text understanding. It is a shared common vocabulary for a specific domain. It is domain specific. Also ontology is application specific. We can create ontologies for specific applications. By creating ontology for a specific application, the process of information extraction will be easier and accurate.

II. INFORMATION EXTRACTION SYSTEM

The task of IR (Information Retrieval) system is to select from a collection of textual documents a subset which is relevant to a particular query, based on key-word search and possibly augmented by the use of a thesaurus [3]. Whereas IE systems aim to retrieve certain types of information from natural language text by processing them automatically [5]. The task of Information Extraction (IE) is to identify a predefined set of concepts in a specific domain, ignoring other irrelevant information, where a domain consists of a corpus of texts together with a clearly specified information need. In other words, IE is about deriving structured factual information from unstructured text[3]. This requires semantic understanding. Natural language processing techniques like Parts of Speech tagging, Tokenization, Co-reference resolution, Named entity recognition, Dependency Parsing is required to understand the semantics underlying in the document.

III. SEMANTIC WEB

The Semantic Web is an extension of the current Web. It is constructed by linking current Web pages to a structured data set that indicates the semantics of this linked page. A smart agent, which is able to understand this structure data set, will then be able to conduct intelligent actions and make educated decisions on a global scale [1].

In semantic web applications, ontology development and population are tasks of paramount importance. The manual performance of these tasks is laborious and therefore cost-intensive, and would profit from a maximum level of automation. For this purpose, the identification and extraction of terms that play an important role in the domain under consideration, is a vital first step [2].

Information Extraction is required to identify the domain dependent terms. Various Natural Language Processing techniques, machine learning models etc. can help to extract the required information.

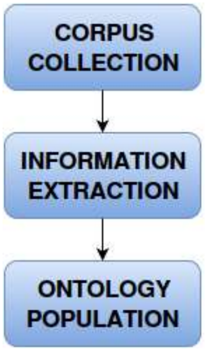

Fig. 1. Different Phases of Proposed Method

IV. PROPOSED METHOD

The Proposed method is used to populate ontology from unstructured documents. The concept of ontology is domain specific. This method is focusing on the domain “New Appointments in companies”. News articles related to this topic is the corpus. The required information has to be extracted from the unstructured text for ontology population.

The overall system consists of three phases. Corpus Collection, Information Extraction, Ontology Population. (Fig. 1.)

A. Corpus Collection

In this phase the corpus is collected. It is in this corpus, the information is extracted and populated to base ontology. Some search API can be used to collect the documents related to our domain. The efficiency of the whole system depends on the type of documents which we are collecting, that is the IR task.

B. Information Extraction

Information Extraction - IE is as a technology that can help an ontology expert during the ontology population and maintenance process [7]. IE techniques can be applied to recognize where, in the documents, concepts are instantiated by specific entities, and where important interactions are expressed by linguistic structures [9]. In this phase, the documents are processed to extract the required information. The proposed method uses machine learning techniques and natural language processing for information extraction.

C. Ontology Population

An ontology should be defined for the domain and the application. A domain expert's advice can be sought to create the base ontology. In this phase the extracted information is optimized and added to the base ontology. This enables the efficient storage and manipulation of extracted information, like querying, summarization, content generation etc.

A richer ontology can be used for a variety of semantic web related tasks such as knowledge management, information retrieval, question answering, semantic desktop applications, and so on. The detailed architecture of the proposed system is shown fig. 2.

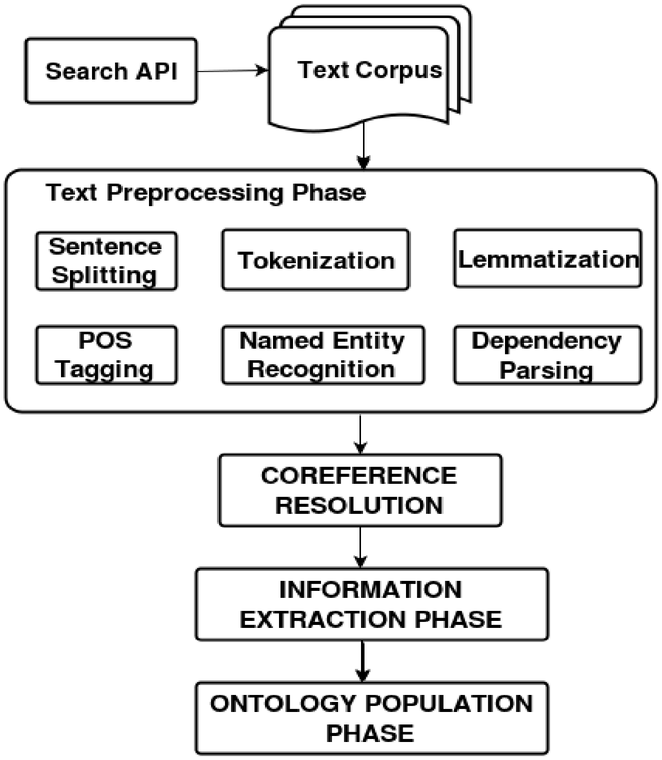

Fig 1. System Design

V. IMPLEMENTATION

The proposed work was done on the news articles about “New Appointments in companies”. A base ontology is created at first. This gives a clear idea about the structure of the ontology to be populated. Each News article is considered as an event. Since it is an appointment news, the information to be populated for an event is, “Company's name”, “Appointee's name”, “Date of appointment”, “Designation of Appointee”, “Predecessor of Appointee”. These information are extracted and populated to the ontology. Machine learning and NLP techniques are used for extracting these information from unstructured text.

A. Text Pre-processing

In the preprocessing stage, Sentence splitting, Tokenization, POS tagging, Named entity recognition, Dependency parsing are done.

Sentence splitting is the process of splitting a paragraph or document to sentences. Period (.) is used as the delimiter for sentence splitting. Many packages and models are existing, which helps to do this task. Part-of-speech tagging is the process of assigning part-of-speech to each word in the document. Tokenization is the process of splitting a document in to tokens. There are two types of tokenization - Sentence tokenization and Word tokenization. In word tokenization, token can be a word, number or punctuation mark. In sentence tokenization, the token is a sentence. Named Entity Recognition is the process of assigning predefined categories like PERSON, DATE, ORGANIZATION, LOCATION etc. to each proper noun in a document. Dependency Parsing is the process of analyzing the grammatical structure of a sentence, and it provide a representation of grammatical relations between words in a sentence. This representation can be used to extract textual relations [8].

Stanford CoreNLP models are used to pre-process the text. These results are further processed in the information extraction phase to extract the required information.

B. Co-reference Resolution

The process of identifying the reference for each pronoun in a sentence is called co-reference resolution. Consider the sentence for example

s1: “HSBC has appointed Ravi Menon as the chief executive of its asset management arm in the country”.

…

s5: “His responsibilities will include ensuring adherence to the HSBC Group's requirements and regulatory standards in the country, the statement said. ”

TABLE I. CO-REFERENCE TABLE

Sentence No	Token
1	HSBC
1	Its
2	HSBC
5	the HSBC Group 's

Here in sentence 1, 'HSBC' and 'its' refers to same entity. The system needs to identify this. Since we need proper nouns as result, each pronoun should be replaced by its corresponding proper nouns. So the Stanford models can be used to find co-reference.

The co-reference table (Table I) represents all same entities. From this output, the input corpus has to be further processed. All pronouns should be replaced by its corresponding best co-reference. Proper noun present in the co-reference table is chosen as the best co-reference. So here “HSBC” is chosen as the co-reference for the word “its”.

C. Information Extraction

It is in this phase, the system is ready to extract the information present in the text. Each news article is considered as an event. “Company's name”, “Appointee's name”, “Date of appointment”, “Designation of Appointee”, “Predecessor of that designation” - these are the information sought for an event.

First of all the promising verb should be identified. Since the system is looking for appointment news the promising verb is words like 'appointed', 'elected', 'selected' etc. and its synonyms. The synonyms of the words are identified from Wordnet. Here after we refer to this set as “promising verb”. Using Stanford's model, dependency parsing for each sentence in the corpus is identified. Using that result, the system had to extract various information which we required.

From the parsed result, the system will check if there is any promising verb present in the sentence. If yes, the required information will be present in that sentence in its subject and object positions. Consider the sentence for example, “HSBC has appointed Ravi Menon as the chief executive of its asset management arm in the country,”. Its dependency parsed result is

```
dependencies-> appointed/VBN (root)
  -> HSBC/NNP (nsubj)
  -> has/VBZ (aux)
  -> Menon/NNP (dobj)
    -> Ravi/NNP (compound)
    -> executive/NN (nmod:as)
```

```
      -> as/IN (case)
      -> the/DT (det)
      -> chief/JJ (amod)
      -> arm/NN (nmod:of)
        -> of/IN (case)
        -> its/PRP$ (nmod:poss)
        -> asset/NN (compound)
        -> management/NN (compound)
  -> country/NN (nmod:in)
    -> in/IN (case)
    -> the/DT (det)
  -> ,/, (punct)
```

From the above sentence the information to be extracted is “Company's name: HSBC”, “Appointee's name: Ravi Menon”. Since the root verb 'appointed' is a promising verb, we can find the required information in its subject and object position.

To extract the company's name, the system will first check the token present in nsubj of root. Also the NER(Named Entity Recognition) of the corresponding token. Since the NER of 'HSBC' is 'ORGANIZATION' the system can conclude it as company's name. Then the system will first check the token present in dobj of root, also its NER. Since the NER of 'Menon' is 'PERSON', the system can conclude it as appointee's name. Along with if any compound words are present, it will be concatenate with corresponding token in nsubj or dobj. The sentence structure can be different. Sometimes the appointee name will be present in nsubj. So it is decided based on the NER of the token present in nsubj and dobj. The conclusion or a rule which we can formulate from the above case is, “For a domain, there will be certain promising verbs. In this domain, if our promising verb is present in a sentence, then the information like appointee name, company's name will be present in subject or object positions of that verb”.

Designation of the appointee will be present in the sentence containing the promising verb. A list of designation is given to the system. Its synonym is also identified using Wordnet. If the system could find any of the designation in the sentence containing the promising verb, then that token is return as the designation.

To extract the date of the event, we have to take NER. The sentence which contains the promising verbs will be focused. The assumption is, the token with named entity 'DATE ' present nearby to the promising verb will the event date. Also if words like 'appointment' or its synonym is present in that sentence, then the correctness will be high.

Predecessor is the person who was in that designation previously. To extract the predecessor's name, our promising verb will be changed to verbs like ‘replaces’, 'succeeds' or words with similar meaning. The synonyms are obtained from Wordnet. The sentence containing the promising verb will have information about the predecessor in its object position. For E.g.

```
dependencies-> replaces/VBZ (root)
  -> He/PRP (nsubj)
  -> Chaddha/NNP (dobj)
```

D. Ontology Population

The ontology required for the specific domain should be defined initially. A domain expert's advice can be sought to create the base ontology. The decision about classes, object properties, data properties in the ontology should be done at the initial stage itself. This idea helps to extract proper information from the corpus. Protégé is an open source software which helps to create our base ontology.

The information collected in the information extraction phase is populated to the base ontology in this phase. Apache Jena API is used for this task [11].

VI. EVALUATION AND CONCLUSION

The performance of the proposed system is compared against human judgment. A set of corpus is given for human judges. The human judge identifies the instances and will populate a base ontology. Then the hand crafted ontology is compared with the output from the proposed system.

Precision [12] is a traditional evaluation metrics for information extraction. In the proposed system, the efficiency of ontology population directly depends on the efficiency of information extraction. So this traditional method is used to evaluate the performance. The evaluation measures were calculated and the system got a precision of 0.83.

So this paper proposes an automatic method to populate ontology from unstructured documents by Information Extraction. Various Natural Language Processing techniques.

are used to extract the information. As per the proposed method, there will be certain promising verbs for a domain. If our promising verb is present in a sentence, then the domain dependent information will be present in subject or object positions of that verb.

The efficiency of information retrieval has a direct influence on the efficiency of the proposed system. From evaluation, it is observed that the system gives a satisfiable output, which can be a deliverable in an industry.

ACKNOWLEDGMENT

I would like to express my thanks and gratitude to the guides and advisors for their valuable support.

REFERENCES

[1] Liang lu , Introduction to the semantic Web and semantic web services, Chapman & Hall/CRC, 2007

[2] Maynard, Diana, Yaoyong Li, and Wim Peters. “Nlp techniques for term extraction and ontology population.” Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. 2008.

[3] Piskorski, Jakub, and Roman Yangarber. “Information extraction: Past, present and future.” Multi-source, multilingual information extraction and summarization. Springer Berlin Heidelberg, 2013. pp.23-49.

[4] Müller, Hans-Michael, Eimear E. Kenny, and Paul W. Sternberg. "Textpresso: an ontology-based information retrieval and extraction system for biological literature." PLoS Biol 2.11 (2004): e309.

[5] Wimalasuriya, Daya C., and Dejing Dou. "Ontology-based information extraction: An introduction and a survey of current approaches." Journal of Information Science, 2010.

[6] Maedche, Alexander, Günter Neumann, and Steffen Staab. "Bootstrapping an ontology-based information extraction system." Intelligent exploration of the web. Physica-Verlag HD, 2003. pp.345-359.

[7] Celjuska, David, and Maria Vargas-Vera. "Ontosophie: A semi-automatic system for ontology population from text." Proceedings of the 3rd International Conference on Natural Language Processing (ICON). 2004.

[8] Manning, Christopher D., et al. "The Stanford CoreNLP Natural Language Processing Toolkit." ACL (System Demonstrations). 2014.

[9] Simperl, E., C. Tempich, and D. Vrandecic. "A Methodology for Ontology Learning Chapter of Book Ontology Learning and Population." IOS Press, Amsterdam (2008).

[10] Stanford CoreNLP. http://nlp.stanford.edu/software/corenlp.shtml, Visited on January 2016.

[11] Apache Jena Project. http://jena.apache.org, Visited on January 2016.

[12] Makhoul, John, et al. “Performance measures for information extraction.” Proceedings of DARPA broadcast news workshop,1999.

[13] Protégé . http://protege.stanford.edu, Visited on October 2015