Ontology Based Webpage Understanding for Information Extraction

A.Medhini
Dvr & Dr. Hs Mic College Of Technology
Kanchikacherla,Krishna
medhini.abbireddy@gmail.com

Asst Prof. N.Ashok
Dvr & Dr. Hs Mic College Of Technology
Kanchikacherla,Krishna
Nutalapati.Ashok@gmail.com

Abstract: Traditional methods on Information Extraction (IE) have focused on the use of supervised learning techniques such as hidden Markov models, self-supervised methods, rule learning, and conditional random fields (CRF). WebNLP framework is based on CRF's and markov models. These techniques learn a language model or a set of rules from a set of hand-tagged training documents and then apply the model or rules to new texts. Models learned in this manner are effective on documents similar to the set of training documents, but extract quite poorly when applied to documents with a different genre or style which is usually found on web. As a result, this approach has difficulty scaling to the Web due to the diversity of text styles and genres on the Web and the prohibitive cost of creating an equally diverse set of hand tagged documents. In this paper we propose an adaptive IE system which uses Ontology Based Information Extraction techniques that extracts all relations by learning a set of lexico-syntactic patterns unlike WebNLP. It permits greater machine interpretability of content than that supported by XML, RDF and RDF Schema (RDF-S), by providing additional vocabulary along with a formal semantics. So, ontologies represent an ideal knowledge background in which to base text understanding and enable the extraction of relevant information.

I INTRODUCTION

Information Extraction (IE) is the process of automatic extraction of information about prespecified types of events, entities or relationships from text such as newswire articles or Web pages. A lot of work have been done on named entity recognition, a basic task of IE, which aims to classify the proper nouns and/or numerical information in documents. Actually most IE tasks can be viewed as the task of recognising some information entities from the text. IE can be useful in many applications, such as information gathering in a variety of domains, automatic annotations of web pages for semantic web, and knowledge management.

Traditional methods on IE have focused on the use of supervised learning techniques such as hidden Markov models (Freitag and McCallum 1999; Skounakis, Craven et al. 2003), self-supervised methods (Etzioni, Cafarella et al. 2005), rule learning (Soderland 1999), and conditional random fields (McCallum 2003). These techniques learn a language model or a set of rules from a set of hand-tagged training documents and then apply the model or rules to new texts. Models learned in this manner are effective on documents similar to the set of training documents, but extract quite poorly when applied to documents with a different genre or style. As a result, this approach has difficulty scaling to the Web due to the diversity of text styles and genres on the Web and the prohibitive cost of creating an equally diverse set of hand tagged documents.

IE's ultimate goal, which is the detection and extraction of relevant information from textual documents, depends on proper understanding of text resources. Rule based IE systems are limited by the rigidity and ad-hoc nature of the manually composed extraction rules. As a result, they present a very limited semantic background. Information extraction (IE) aims to retrieve certain types of information from natural language text by processing them automatically. For example, an IE system might retrieve information about geopolitical indicators of countries from a set of web pages while ignoring other types of information.

In this paper, Ontology-based information extraction was proposed which has recently emerged as a subfield of information extraction. Here, ontologies - which provide formal and explicit specifications of conceptualizations - play a crucial role in the IE process. Because of the use of ontologies, this field is related to knowledge representation and has the potential to assist the development of the Semantic Web.

Ontologies are designed for being used in applications that need to process the content of information, as well as to reason about it, instead of just presenting information to humans. They permit greater machine interpretability of content than that supported by XML, RDF and RDF Schema (RDF-S), by providing additional vocabulary along with a formal semantics. So, ontologies represent an ideal knowledge background in which to base text understanding and enable the extraction of relevant information. This may enable the development of more flexible and adaptive IE systems than those relying on manually composed extraction rules (both based on linguistic constructions or document structure).

II RELATED WORK

Webpage understanding plays an important role in information retrieval from the Web. There are two main branches of work for webpage understanding: template-dependent approaches and template-independent approaches.

Template-dependent approaches (i.e., wrapper-based approaches) can generate wrappers either with supervision or without supervision. The supervised approaches take in some manually labeled web pages and learn some extraction rules (i.e., wrappers) based on the labeling results. Unsupervised approaches do not need labeled training samples. They first automatically discover clusters of the web pages and then produce wrappers from the clustered web pages. No matter how the wrappers are generated, they can only work on the web pages generated by the same template. Therefore, they are not suitable for general purpose webpage understanding. In contrast, template-independent approaches can process various pages from different templates.

However, most of the methods in the literature can only handle some special kinds of pages or specific tasks such as object block (i.e., data record) detection. For example, can only segment list pages can only detect the main block in the page. Another method, segments data on list pages using the information contained in their detail pages. The need of detail pages is a limitation because automatically identifying links that point to detail pages is nontrivial and there are also many pages that do not have detail pages behind them. Zhai and Liu proposed an algorithm to extract structured data from list pages. The method consists of two steps. It first identifies individual records based on visual information and a tree matching method. Then a partial tree alignment technique is used to align and extract data items from the identified records. Song et al. define the block importance estimation as a learning problem. First, they use the Vision-based Page Segmentation (VIPS) algorithm to partition a webpage into semantic blocks with a hierarchy structure. Then, spatial features (such as position and size) and content features (such as the number of images and links) are extracted to construct a feature vector for each block. Based on these features, learning algorithms, such as SVM and neural network, are applied to train various block importance models.

III ONTOLOGY EXPLOITATION FOR IE

IE and ontologies are involved in two main and related tasks :

- Ontology is used for Information Extraction: IE needs ontologies as part of the understanding process for extracting the relevant information;
- Information Extraction is used for populating and enhancing the ontology: texts are useful sources of knowledge to design and enrich ontologies.

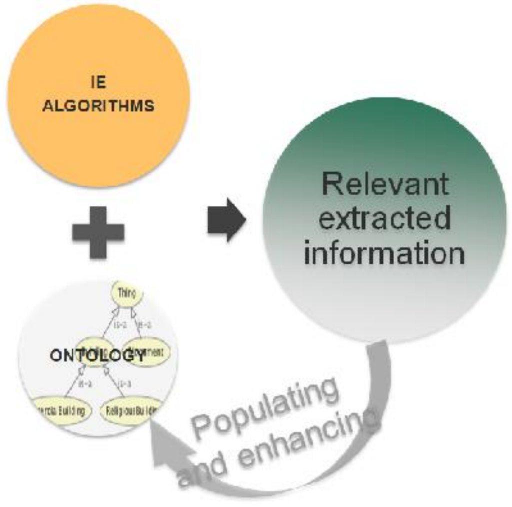

Fig 1: Ontology exploitation for IE

These two tasks, as can be seen in Figure 1, can be combined in a cyclic process: ontologies are used for interpreting the text at the right level for IE and IE extracts new knowledge from text, to be integrated in the ontology. An ontology identifies the entities that exist in a given domain and specifies their essential properties. It does not describe the spurious properties of these entities. The goal of IE is to extract factual knowledge to instantiate one or several predefined forms.

Whether one wants to use ontological knowledge to interpret natural language or to exploit written documents to create or update ontologies, in any case, the ontology has to be connected to linguistic phenomena. The complexity of the linguistic anchoring of ontological knowledge is well known. A concept can be expressed by different terms and many words are ambiguous. Rhetoric, such as lexicalized metonymies or elisions, introduces conceptual shortcuts at the linguistic level and must be elicited to be interpreted into domain knowledge.

A) Sets of entities

Recognizing and classifying named entities in texts requires knowledge on the domain entities. Specialized lexical or keyword lists are commonly used to identify the referential entities in documents. Three main objectives of these specialized lexicons can be distinguished: semantic tagging, naming normalization and linguistic normalization.

- Semantic tagging. List of entities are used to tag the text entities with the relevant semantic information. In the ontology or lexicon, an entity is described by its type (the semantic class to which it belongs, here PERSON) and by the list of the various textual forms (typographical variants, abbreviations, synonyms) that may refer to it.
- Naming normalization. As a by-effect, these resources are also used for normalization purposes. This avoids rule overfitting by enabling specific rules to be abstracted.
- Linguistic normalization. Beyond typographical normalization, the semantic tagging of entities contributes to sentence normalization at a linguistic level. It solves some syntactic ambiguities.

IV Ontology-based Information Extraction

We consider ontology-based IE systems as those approaches relying on predefined ontologies in one or several stages of the extraction process. Those approaches are document driven: they start from a particular document (or set of documents) and they try to identify entities found in that context, trying to annotate them according to the input ontology. So, on the contrary to plain IE systems, ontology-based ones are able to specify their output in terms of a pre-existing formal ontology. These systems almost always use a domain-specific ontology in their operation, but we consider a system to be domain-independent if it can operate without modification on ontologies covering a wide range of domains. So, the problem is very similar to semantic annotation. Annotations represent a specific sort of metadata that provides references between entities appearing in resources and domain concepts modelled in an ontology. Semantic annotation is one fundamental pillar of the Semantic Web making it possible for Web-based tools to understand and satisfy the requests of people and machines to exploit Web content.

we refer to both semantic annotation and ontology-based IE indistinctly. The basic idea of the techniques in this category is to focus the processing on the ontology basic elements (classes, relations), leveraging this knowledge to find resources that can be analysed to obtain useful information (in most cases, instances of the ontology classes). This method presents some benefits:

- Focusing on the ontology components seems a natural way to exploit all kinds of ontological data.
- These systems can consider a huge amount of different resources (e.g. the Web), and are not constrained by a limited corpus of documents.
- The systems concentrate all their resources on searching directly for information related to the ontology components, rather than having to analyse a potentially large number of documents that do not contain interesting information.

We distinguish between two types of matching: Direct Matching and Semantic Matching. In this initial step, the system tries to find a direct match between the potential subsumers of a named entities and the ontology classes. The semantic matching step

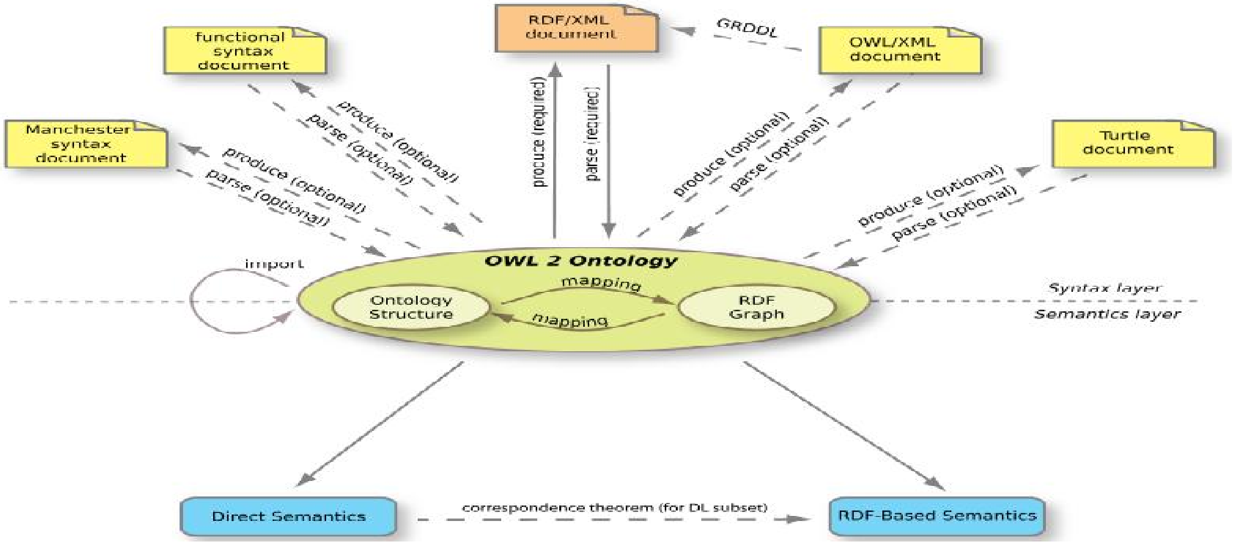

Fig 2: Structure of OWL 2

is performed when the direct matching has not produced any result.

V Technique for Ontology based Information Extraction

The OWL 2 Web Ontology Language, informally OWL 2, is an ontology language for the Semantic Web with formally defined meaning. OWL 2 ontologies provide classes, properties, individuals, and data values and are stored as Semantic Web documents. OWL 2 ontologies can be used along with information written in RDF(resource description framework), and OWL 2 ontologies themselves are primarily exchanged as RDF documents.RDF purpose is to provide structure for describing identified things.RDF provides flexibility and scalability among data relationships. By using XML, OWL information can easily be exchanged between different types of computers using different types of operating system and application languages.

OWL 2 adds new functionality with respect to OWL 1. Some of the new features are :syntactic sugar (e.g., disjoint union of classes) ; expressivity; keys; propertychains; richer datatypes, data ranges; qualified cardinality restrictions; asymmetric, reflexive, and disjoint properties; and enhanced annotation capabilities .

The conceptual structure of OWL 2 ontologies as in Fig 2, is defined in the OWL 2 Structural Specification document. This document uses UML to define the structural elements available in OWL 2, explaining their roles and functionalities in abstract terms and without reference to any particular syntax. It also defines the functional-style syntax, which closely follows the structural specification and allows OWL 2 ontologies to be written in a compact form. Any OWL 2 ontology can also be viewed as an RDF graph. The relationship between these two views is specified by the Mapping to RDF Graphs document, which defines a mapping from the structural form to the RDF graph form, and vice Versa. The OWL 2 Quick Reference Guide provides a simple overview of these two views of OWL 2, laid out side by side. "OWL 2 Full" is used informally to refer to RDF graphs considered as OWL 2 ontologies and interpreted using the RDF-Based Semantics.

The OWL 2 specification identifies several profiles- In logic, profiles are usually called fragments or sublanguages. OWL 2 provides three profiles- OWL 2 EL, OWL 2 QL, and OWL 2 RL- each of which provides different expressive power and targets different application scenarios. The OWL 2 profiles are defined by placing restrictions on the Functional-Style Syntax of OWL 2. An ontology written in any of these profiles is a valid OWL 2 ontology. Therefore, the semantics of the OWL 2 profiles is given by the direct model theoretic semantics of OWL 2. Ontology modelers who want to ensure that their ontologies are in a certain profile can use these restrictions as a guide; further more , tool developers can easily use the corresponding grammars to create tools for checking which profile an ontology belongs to.

VI PERFORMANCE

In general, the use of statistical measures (e.g. co-occurrence measures) in knowledge related tasks for inferring the degree of relationship between concepts is a very common technique when processing unstructured text. However, statistical techniques typically suffer from the sparse data problem (i.e. the fact that data available on words of interest may not be indicative of their meaning). So, they perform poorly when the words are relatively rare, due to the scarcity of data. This problem can be addressed by using lexical databases or with a combination of statistics and lexical information. However, the analysis of such an enormous repository for extracting candidate concepts and/or statistics is, in most cases, impracticable. Here is where the use of lightweight techniques that can scale well with high amounts of information, in combination with the statistical information obtained directly from the Web, can represent a good deal.

An ontology is defined as a formal, explicit specification of a shared conceptualization. Conceptualization refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concepts used, and the constraints of their use, are explicitly defined. Formal refers to the fact that the ontology should be machine-readable. Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted by a group.

VII CONCLUSION

Webpage understanding plays an important role in Web search and mining. It contains two main tasks, i.e., page structure understanding and natural language understanding. However, little work has been done toward an integrated statistical model for understanding webpage structures and processing natural language sentences within the HTML elements.

In this paper, Proposes to use Ontology Based Information Extraction techniques that extracts all relations of a web page by learning a set of lexico-syntactic patterns. Here, ontologies - which provide formal and explicit specifications of conceptualizations - play a crucial role in the IE process. Because of the use of ontologies, this field is related to knowledge representation and has the potential to assist the development of the Semantic Web.Ontology's are implemented using the Web Semantic representation language, OWL 2. OWL2 hasthe following benefits: Syntactic sugar (e.g., disjoint union of classes) , Expressivity , Keys, Property chains, Richer , datatypes, data ranges ,Qualified cardinality restrictions, Asymmetric, reflexive, and disjoint , properties , Enhanced annotation capabilities.Faster and better performance when compared to WebNLP frameworkbased on both CRF's and markov models. ontologies are used to drive the extraction process indicating the concepts that we want to extract from an analysed entity in a particular domain.

VIII REFERENCES

[1] J. Cowie and W. Lehnert, "Information Extraction," Comm. ACM, vol. 39, no. 1, pp. 80-91, 1996.

[2] C. Cardie, "Empirical Methods in Information Extraction," AI Magazine, vol. 18, no. 4, pp. 65-80, 1997.

[3] R. Baumgartner, S. Flesca, and G. Gottlob, "Visual Web Information Extraction with Lixto," Proc. Conf. Very Large Data Bases (VLDB), pp. 119-128, 2001.

[4] A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. ACM SIGMOD, pp. 337-348, 2003.

[5] Agirre, E., O. Ansa, et al. (2000). Enriching very large ontologies using the www. Proceedings of the Ontology Learning Workshop, ECAI.

[6] Ahmad, K., M. Tariq, et al. (2003). Corpus-Based Thesaurus Construction for Image Retrieval in Specialist Domains. Advances in Information Retrieval. F. Sebastiani, Springer Berlin / Heidelberg. ***2633:*** *76-76.*

[7] Alfonseca, E. and S. Manandhar (2002). Improving an ontology refinement method with hyponymy patterns. 3rd International Conference on Language Resources and Evaluation, LREC 2002, Las Palmas, Spain.

[8] O. Etzioni, M.J. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates, "Unsupervised Named- Entity Extraction from the Web: An Experimental Study," Artificial Intelligence, vol. 165, no. 1, pp. 91-134, 2005.

[9] Brill, E. (2003). Processing Natural Language without Natural Language Processing. 4th International Conference on Computational Linguistics and Intelligent Text Processing and Computational Linguistics, CICLing 2003, Mexico City, Mexico, Springer Berlin / Heidelberg pp. 360-369.

[10] Buitelaar, P., P. Cimiano, et al. (2008). "Ontology-based information extraction and integration from heterogeneous data sources." International Journal of Human-Computer Studies ***66****(11): 759 - 788.*