Ontology-based Information Extraction: An Overview and a Study of different Approaches

Ritesh Shah
Ph.D. Research Scholar
Mewar University, Rajasthan
India

Suresh Jain
Director, Shushila Devi Bansal
College of Technology Indore
India

ABSTRACT
Information Extraction is a process to retrieve Information from Natural Language Text or unstructured text by automated process. OBIE[1] (Ontology based Information Extraction) is one of the most emerging subfields of Information Extraction. Where Ontology is a formal and explicit specification of conceptualization which plays a crucial role in the process of Information Extraction[2]. Ontology has potential to support to build a Semantic Web and also plays a vital role in the Knowledge Representation. This paper attempts to survey some relevant research in the field of Ontology Based Information Extraction.

Keywords
Ontology, Information Extraction, Semantic Web, Knowledge Base

1. INTRODUCTION
In the present scenario, the WWW is the most popular and interactive medium to distribute information. The WWW is source of huge amount of divers and dynamic Information. People use the Internet to search the requested information. But most of the time, they gets lots of irrelevant and unimportant web documents. Therefore, there are requirement of the efficient and relevant information from Unstructured or Semi-Structured information present in WWW. So various Ontology Based Modals are proposed by researchers from last decade.

In this survey paper, first describe the Ontology Based Information Extraction System.

Generally Ontologies are domain specific , it means that different domains are having different Ontologies, which it shows the relationship between different classes and entities. Also Ontologies are application dependent , It can create specific ontology for specific application. For examples in Recruitment ontology, qualification , age and experience of Candidate(Concept) can be used as a guide in the process of information extraction. Ontologies -hierarchical representation which shows the classes and sub-classes relationship of components of any domain specific concept which assist to extract the information on the basis of specified concepts and also building and updating Ontologies . Here the Output of the process (OBIE) is also in the form of Ontoloies (OWL/ RDF).

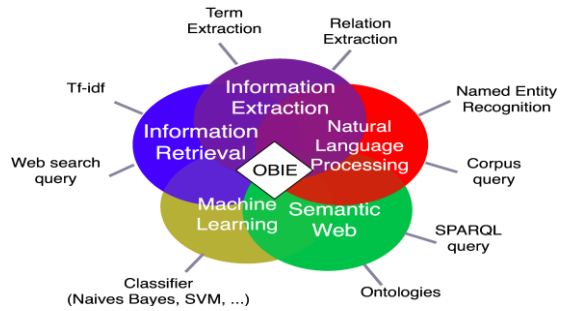

Figure 1: Ontology Based Information Extraction

Information Extraction employed various algorithms and method for information retrieval. In OBIE , there is not any specific algorithm/ method . Ontologies are used just for guiding these algorithms for efficient and relevant information extraction.

In this section provided the general introduction. Rest of the paper attempt to review the work done in the field of OBIE, based on various literature of these fields. This section discussed the general functionality and architecture of OBIE system.

2. ONTOLOGY BASED INFROMATION EXTRACTION
This section concentrating the key features of OBIE system and find out the factors on that make OBIE system different from general IE system. Generate Semantic Annotation from unstructured or semi unstructured text: OBIE system is subfield of Information Extraction, which is generally observe as a subfield of NLP. It is reasonable to limit the input to Natural processing text

2.1 Ontology Learning and ontology population:
Ontology learning approaches built an ontology by processing unstructured and semi-structured text. The term was originally used by Maedche and Staab [3]. Cimiano[4] describes ontology learning as acquisition of a domain model from data. It is divided into following steps:

- Extraction of domain terminology and synonyms from a corpus of documents (e.g., {city}, {country, nation})
- Identifying of main concepts on the basis of the detected relevant terms and the classes of synonyms

  (e.g., c := country := nation)
- Structuring of concepts into taxonomy (e.g., capital < city)

- Learning of non-taxonomic relations between concepts

  (e.g., is capital(city, country))

- Structuring of relations into hierarchy

  (e.g., is_capital(city, country) î located_incity,

  country)

- Learning the adoption of axiomatic definitions of and between concepts (e.g., disjoint, equivalence) and relations (e.g., cardinalities, transitive, reflexive, symmetric)

- Learning general axioms

  (e.g., currentProject(person, project) ^ member(project, organization) → member(person, organization)).Ontology population approaches identify entities in text that are related to a pre-defined domain ontology. Cimiano [2006] outlines three major task topics in populating ontologies:

- Learning instances of concepts in the domain ontology (e.g., concrete cities, persons, or locations). This task is similar to a Named Entity Recognition (NER).

- Learning instances of formalized relations between two or more instances of concepts (e.g., instantiations of the relationship foaf:member between an instance of foaf:Group and any kind of foaf:Agent such as a foaf:Person.)

- Semantically annotating entity references with instantiations of relations or concepts in a domain ontology.

Wimalasuriya and Dou [1] describe a close relation between OBIE and the Semantic Web. OBIE systems generate semantic content which is known as Semantic Annotation for the Web pages. Semantic agents can directly process semantic content for Information Retrieval.

Information Extraction guided by Ontologies: Wimalasuriya and Dou [1]

"*We believe that "guide" is a suitable word to describe the interaction between the ontology and the information extraction process in an OBIE system: in all OBIE systems, the information extraction process is guided by the ontology to extract things such as classes, properties and instances. This means that no new information extraction method is invented but an existing method is oriented to identify the components of ontology".*

OBIE system is a automatic metadata generation [4] and closely related to Semantic Web[1].

Fig 2 shows a generalized architecture of OBIE system by Wimalasuriya and Dejing Dou.

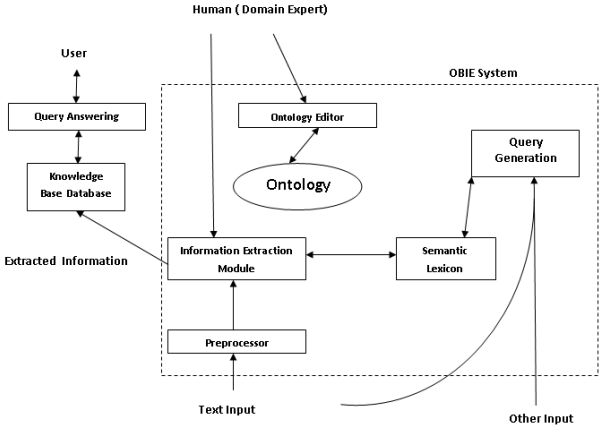

Figure 2: OBIE system, by Wimalasuriya and Dou[1]

Above figure present combinations of Ontology with IE system. It distinguish between Knowledge base and an ontology. Information Extractors populates the Knowledge base and is queried by users. The beauty of this architecture is that Human intervention as a Domain expert is allowed to manipulate the internal extraction logic. The information extractors in this work utilize RDF graphs from Semantic Web sources, extract information from text, and return extracted information as RDF graphs. SPARQL queries provide means for additional filtering mechanisms. The role of semantic lexicons is fulfilled by applying linguistic resources such as text segmenters, POS taggers, and text chunkers.

3. DIFFERENT APPROACHES OF ONTOLOGY BASED INFORMATION EXTRACTION

Some input documents (text , HTML pages , XML with forms of tables and database) requirement for the preprocessing (like lexical analysis , syntactic analysis , semantic analysis) . Then move toward the extraction process which may be rule based, Machine Learning system, hybrid system . Use of the ontology in the Information Extraction process for

1. The domain entities and their variation (Synonyms , co – references)
2. Design a conceptual hierarchies by IE system for generating its extraction rules.
3. Concept properties and relation between properties(NER) , guide the IE extraction process system
4. Relation between concept and also related to their NER

The output of this Information Extraction process to be in form of Semantic Annotation, filled template , populated / Improved ontology. Bootstrapping approach[13] is involved to increase the IE performance.

3.1 Main IE approaches employed by the OBIE systems

1. Embley et al. [6] provided a use case framework for converting data-rich unstructured documents into structure documents for populating domain ontology.
2. Bontcheva et al.,[7] enables the use of ontologies in IE by providing OntoGazetteers and OntoRootGazetteer. OntoGazetteers allow a manual

mapping between gazetteer lists to ontology classes. OntoRootGazetteer analyze existing concept labels in ontologies with tokenizers, POS taggers, and stemmers in order to recognize these labels in text sources in restricted language.

3. Li and Bontcheva [8], introduces the Hierarchical Learning approach for IE , which uses the target ontology as an essential part of the Extraction process, by considering relations between concepts.
4. Paul Buitelaara, Philipp Cimianob,_, Anette Frankc, Matthias Hartungc, Stefania Racioppaa [9] propose SOBA(SmartWeb Ontology Based Annotation) is a component for OBIE system. It extract the information from heterogeneous sources like tabular structures , text , image caption in a Semantically integrated (link) way.
5. B. Endres-Niggemeyer[10]. An interacting IE modules uses a distributed semantic agent as a component in IE architecture. The advantage of the Distributed Ontology based Architecture proposed easy tracking of decisions and explanations to users. New agent can be integrated easily so that the agent community learns.
6. H. Cunningham, K. Bontcheva[11], V. Tablan, and D. Maynard, using regulation expression to extract the Information. This approach uses the Natural Language Processing framework and provide platform to employ linguistics rules for building the knowledge.
7. Mendes et al. [12] DBpedia Spotlight a tool enables users to link text documents to the Linked Open Data Cloud through the DBpedia Interlinking hub.

4. CONCLUSION

In this paper, we are review the some of the Ontology Based Information Extraction method. Ontology Based Information Extraction method only guides the system that how to pull out efficient and relevant information using the Information Extraction methods. There are several direction for future work with OBIE System like improving the efficiency of IE process to improve the precision and recall. Generating the semantic contents for the Semantic Web is one of the major factors that make OBIE an interesting research field. OBIE system can be used for identified the semantic content for semantic web and also implemented Ontology Based web service for result. Most of the OBIE system uses single Domain specific ontology. However , there is no rule to not to use multiple Ontologies[14].

5. REFERENCES

[1] D. C. Wimalasuriya and D. Dou. Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 2010, 36(3): 306.

[2] T. R. Gruber: A translation approach to portable ontology specifications, *Knowlegde Acquisition 5-199-220, 1993*

[3] A. Maedche and S. Staab. Ontology Learning for the Semantic Web. IEEE Intelligent Systems, 16:72{79, March 2001. ISSN 1541-1672. 66.

[4] P. Cimiano. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387306323. 15, 66, 67, 72

[5] B. Popov, A. Kiryakov, D. Ognyano_, D. Manov, and A. Kirilov. KIM { a semantic platform for information extraction and retrieval. Natural Language Engineering, 10 (3-4):375{392, 2004. 65, 68

[6] D. W. Embley, D. M. Campbell, R. D. Smith, and S. W. Liddle. Ontology-based extraction and structuring of information from data-rich unstructured documents. In CIKM '98: Proc. of the 7th international conference on Information and knowledge management, pages 52{59, New York, NY, USA, 1998. ACM. ISBN 1-58113-061-966,69

[7] K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham. Evolving GATE to meetnew challenges in language engineering. Natural Language Engineering, 10(3{4):349{373, 2004. ISSN 1351-3249. 38, 66, 69

[8] K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham. Evolving GATE to meetnew challenges in language engineering. Natural Language Engineering, 2004. ISSN 1351-3249. 38, 66, 69

[9] P. Buitelaar, P. Cimiano, S. Racioppa, and M. Siegel. Ontology-based Information Extraction with SOBA. In Proceeding of LREC, Genoa, Italy, 5 2006. 69

[10] B. Endres-Niggemeyer. Ontology-based information extraction in agents' hands. In Proceedings 1st International and KI-08 Workshop on Ontology-based Information Extraction Systems, volume 400, pages 15{21. DFKI, 2008. 70

[11] H. Cunningham, K. Bontcheva, V. Tablan, and D. Maynard, *General Architecture for Text Engineering (GATE) (2003)*. Available at: http://www.gate.ac.uk (accessed 25 June 2009)

[12] P. Mendes, M. Jakob, A. Garcia-Silva, and C. Bizer. Dbpedia spotlight: Shedding light on the web of documents. In Proceedings of the 7th International Conference on Semantic Systems (I-Semantics), September 2011. 70, 137, 197, 222

[13] Alexander Maedche2, G¨unter Neumann1, Steffen Staab *Bootstrapping an Ontology-Based Information Extraction System*. In: Studies in Fuzziness and Soft Computing, IntelligentExploration of the Web, Springer, 2002

[14] D.C. Wimalasuriya and D. Dou, Using multiple ontologies in information extraction. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, (ACM, New York,2009)