Preliminary Study on an Ontology Learning from Textual Data

[1]Saidah Saad, [2]Naomi Salim
Department of Information System,
Faculty of Computer Science and Information System
Universiti Teknologi Malaysia, Skudai, Johor
saidah@ftsm.ukm.my[1], naomie@fsksm.utm.my[2]

Abstract

Natural language understanding is needed to intelligently handle the large volumes of information that is explosive growth over the last decade on the WWW. Ontologies may help with analyzing and understanding text where ontology provides a capability to represent objects, concepts and other entities and the relationships between them. Ontologies may be used as a tool for finding possible meanings of words in text, and meaning of text in general. Now, much of this ontology development has been directed towards extraction from textual data as human language is a primary mode of knowledge transfer.

The aim of this paper is to give a general overview and preliminary study on some of the ontology learning from text that plays a prominent role on the knowledge retrieval and how the ontological semantic can be improved through the adoption of semantic web technology.

Keyword: knowledge management, ontological semantic, ontology learning, NLP,

1.0 Introduction

Traditional search engines return ranked retrieval lists that offer little or no information on the semantic relationships between documents. Knowledge workers consequently spend a substantial amount of their time browsing and reading to find out how documents are related to one another and where each falls into the overall structure of the problem domain. Yet only when knowledge workers begin to locate the similarities and differences between pieces of information do they move into an essential part of their work — building relationships to create new knowledge.

According to J Davies et. al (2005), current knowledge management systems have significant weaknesses such as

- Existing keyword-based searching information,
- Human browsing and reading is required to extract relevant information from information sources but fail to integrate information distributed over different sources.
- Maintaining weakly structured text sources is a difficult and time-consuming activity when such sources become large. Keeping such collections consistent, correct, and up-to-date requires mechanised representations of semantics that help to detect anomalies.
- Automatic document generation enable adaptive Web sites that are dynamically reconfigured according to user profiles or other aspects of relevance. Generation of semistructured information presentations from semistructured data requires a machine-accessible representation of the semantics of these information sources.

The solution is to move from a documentcentric view of information retrieval to a knowledgecentric view, wherein tools are not returning ranked lists of documents to the user, but, instead, attempt to provide them with the specific information they need perhaps gathered from multiple documents. The use of ontologies and supporting tools offer an opportunity to significantly improve knowledge management capabilities in large organizations.

2.0 Literature Review

Ontology, by its most cited definition in AI, is a shared, formal conceptualization of a domain (Gruber, 1993). As this definition suggests, ontologies differ from data models in two significant aspects (M.Peter and A.Hans, 2003).

i. Ontologies build upon a shared understanding within a community. This understanding represents an agreement over the concepts and their relationships that are present in a domain.
ii. Ontologies use machine processable, logic-based representations that allow computers to manipulate ontologies. This includes transferring ontologies among computers, storing ontologies, checking the consistency of ontologies, reasoning about ontologies etc.

With the support of the ontology, both user and system can communicate with each other by the shared and common understanding of a domain and they serve as reusable vocabularies that can be shared by human as well as computer system (U.Jan et al. 2003).

There are many ontology learning from text applications that have been published and presented in various domain. P.Cimiano and J.Volker (2005) present Text2Onto, a framework for ontology learning from textual resources. Alani et al. (2003) present Artequakt that automatically extracts knowledge from unstructured document about artists from the Web on an ontology. A. Schutz and P.Buitelaar (2005) describe a system call RelExt that is capable of automatically identifying highly relevant triples (pairs of concepts connected by a relation) over concepts from an existing ontology and much more.

Common to all these technologies (Nirenburg, S. and V. Raskin. 2004) such as ontology extraction, ontology learning and ontology annotation tool are that they use existing descriptions to create ontologies. In general, the more structured and formal the existing description, the higher the quality of the resulting ontology.

One of the theory has been apply in ontology learning from textual data is call ontological semantic. Ontological semantics is a theory of meaning in natural language and an approach to natural language processing (NLP) which uses a constructed world model, or ontology, as the central resource for extracting and representing meaning of natural language texts, reasoning about knowledge derived from texts, as well as generating natural language texts based on representations of their meaning (Raskin and Nirenburg, 2004).

Figure 1, illustrates the interactions among the data, the processors and the static knowledge sources in ontological semantic.

In ontology learning from text, P.Cimiano divided it into six layers of the different subtasks of learning ontology (refer figure 2) (Buitelaar. P et. al. 2005. It consists:

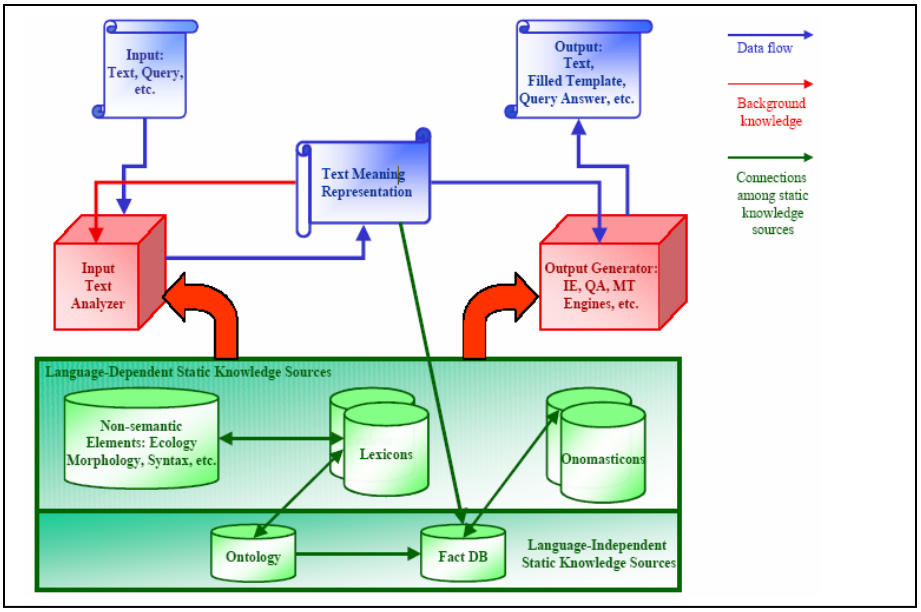

Figure 1. The Data, the Processors and the Static Knowledge Sources in

i. Term Extraction

Term extraction is a prerequisite for all aspects of ontology learning from text. Terms are linguistic realizations of domain-specific concepts. The task here is to find a set relevant terms or sign for concept and relation. Term can be a single word or multi-word compound relevant for the domain in question as a term (P.Cimiano, 2006).

The technique that commanly use in term extraction such as statistical anlysis (co-occurence analysis extract from corpus, comparison of frequencies between domain and general corpora) , extract pattern (Adjective-Noun, Noun-Noun, Adj-Noun-Noun, ..), linguistic methods (linguistic analysis such as part-of-speech, tagging, morphological anlysis ext), term disambiguation & compositional interpretation. The hybrid methods can be used by using linguistic rules to extract term candidate and statistical method for pre and post filtering.

ii. (Multilingual) Synonyms

Identify terms that share (some) semantic, i.e. potentially refer to the same concept. It can be synonym (within languages) or it is a translation (between languages). The techniques used for extraction of synonym or translations are classification (example by extending WordNet where with SynSets corresponding to classes) and clustering where it's cluster according to similar distribution.

iii. Concept Extraction

It was terms that indicate a concept within the domain population. Concept formation should ideally provide an intension definition of concepts, their extension and the lexical which are used to refer to them (P.Cimiano, 2006).

iv. Taxonomy Extraction

Taxonomy is frequently hierarchical in tree structure classifications for a given set of objects. The most common relationship is 'is a'. One of the examples of taxonomy extraction present by A.Popescu, A.Yates & O.Etzioni (2004) illustrate in table 1. There are several technique proposed in the past research that is listed down by Buitelaar. P et. al. (2005).

v. Relation Extraction

It is a non-taxonomic ontological relation.

vi. Rules and Axiom

The task is to learn which concepts, relations or pairs of concepts the axioms in the system apply to (P.Cimiano, 2006). These axioms can be represented for example using first-order logic.The extraction of general axioms is probably the least researched area in the context of ontology learning (P.Cimiano, 2006).

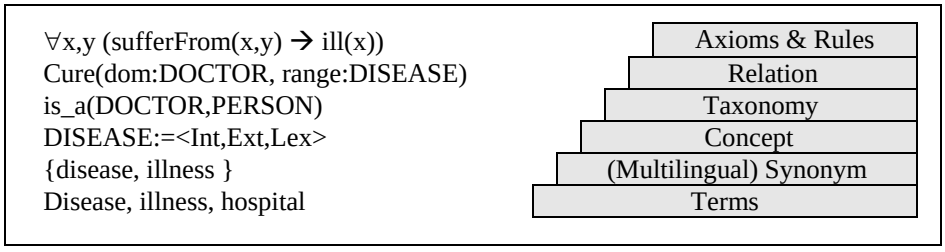

Figure 2 Ontology learning layer cake

Source : P.Buitelaar et. al., Ontology Learning from Text. Tutorial at ECML/PKDD, Oct. 2005, Porto, Portugal.

Pattern	Extraction
C_1 {",} "such as" CN	isA(CN;C_1)
"such" C_1 "as" CN	isA(CN;C_1)
CN {",} "and other" C_1	isA(CN;C_1)
CN {",} "or other" C_1	isA(CN;C_1)
C_1 {",} "including" CN	isA(CN;C_1)
C_1 {",} "especially" CN	isA(CN;C_1)
C_1 "and" CN	isA(CN; class(C_1))
C_1 {",} C_2 {",} "and" CN	isA(CN; class(C_1))

Table 1: Rules for Taxonomy Extraction.

In most cases the layers conceptually built one upon another in the sense that the processes within higher layers rely on the output of processes situated at lower layer. Figure 2 include concrete example from the domain of medical at the left of each layer. Within the terminology acquisition step, we would find relevant term such as disease, illness, hospital. At synonym discovery step, we might group together disease and illness as in certain context they are sunonym. This group of synonym might then provide the lexicon for the concept disease with an intension and extension. The extension might for example be spesified as '*health status; when something is wrong with a bodily function'*. Futher more, we might learn relation together with their domain such as the *cure* relation between *doctor* and *disease* where *doctor* is a *person* in taxonomy layer that can be find before. Finally, we also might derive more complex relationships between concepts and relations in the form of axioms.

3.0 Discussion

Ontologies are widely used in knowledge management and it construction has been addressed in several research activities that already mention a few before. In recent years, two approaches have been concerned to solve the ontology engineering problems (M. Shamsfard & A.A. Barforoush, 2004).

i. Development of methods, methodologies, tools and algorithms to map, merge and alignment of existing ontologies
ii. Development of methods, methodologies, tools and algorithms to acquire and learn ontologies (semi) automatically.

In this paper, a writes focused on the second approach in order to get an overview of current research and technique in the field of ontology learning that already proposed in that area.

Based on Table 2, it shows the summary on research based on ontology learning that already done based on five layer introduce by Buitelaar. P et al (2005), we can made a conclusion that there are many of the research on learning methods consist of term extraction, synonym extraction, taxonomy extraction and relation extraction or hybrid methods on the technique. There are a few researches recently done on concept extraction and on axiom and rules in context of ontology learning from textual data.

In order to development and use of ontologies , there are a few problem according to M. Shamsfard & A.A. Barforoush (2004).

i. Lacking of standards to integrate or reuse existing ontologies
ii. Using fixed categories based on a single viewpoint
iii. Absence of full automatic knowledge acquisition methods.

4.0 Further Work

Automatic learning of ontologies is a solution to ontology creation and management problem (M. Shamsfard & A.A. Barforoush , 2004). And in order to get more semantic rich, enhancement of the NLP component consisting of language expansion (expanding the grammar and the linguistic knowledge with axiom) to cover wider range of sentences. For further research are towards combining different ontology learning paradigms via a machine learning approach, much further research is needed in this direction to unveil the full potential of such combination-based approach (P.Cimiano, 2006).

5.0 Conclusion

This paper has given an overview on ontology learning from the textual data and research that already been developed. It been describing based on five layers of the different subtasks of learning ontology.

Reference

1. Ana-Maria Popescu, Alexander Yates and Oren Etzioni (2004), Proceedings of the AAAI 2004 Workshop on Adaptive Text Extraction and Mining. San Jose, California

2. A. Cucchiarelli, R. Navigli, F. Neri, P. Velardi (2004). Automatic Generation of Glosses in the OntoLearn System, Proc. of 4th International Conference on Language Resources and Evaluation (LREC 2004), Libsoa, 26-28th May, 2004.

3. A. Kayed & R.M. Colomb (2001). Re-engineering Approach to Build Domain Ontologies, the First Asia-Pacific Conference on Web Intelligence (WI-2001), Springer, Lecture Notes in Computer Science (2198).

4. A. Schutz and P.Buitelaar (2005). RelExt: A Tool for Relation Extraction from Text in Ontology Extension. Proc. of the 4th International Semantic Web Conference, Galway, Ireland.

5. Brasethvik and Gulla (2001). Natural language analysis for semantic document modeling. Data & Knowledge Engineering. Volume 38, Issue 1 , July 2001, Pages 45-62.

6. Chang-Shing Lee; Zhi-Wei Jian; Lin-Kai Huang (2005). A fuzzy ontology and its application to news summarization, Systems, Man and Cybernetics, Part B, IEEE Transactions on Volume 35, Issue 5, Oct. 2005 Page(s): 859 - 880

7. J Davies, R.Studer, Y.Sure, P.W. Warren (2005). Next Generation Knowledge Management. BT Technology Journal, Vol 23 No.3, July 2005. Page(s): 175-190.

8. M.Peter and A.Hans, 2003. Analysis of the State-of-the-Art in Ontology-based Knowledge Management. Semantic Web and Peer-to-Peer Project Report.

9. M. Shamsfard, A.A. & Barforoush (2004). International Journal of Human-Computer Studies 60 (2004) 17–63

10. Nirenburg, S. and V. Raskin. 2004. Ontological Semantics. MIT Press.Cambridge, MA.

11. P.Buitelaar, P.Cimiano, M.Grobelnik, M.Sintek (2005). Ontology Learning from Text. Tutorial at ECML/PKDD, Oct. 2005, Porto, Portugal.

12. P.Cimiano (2006). Ontology Lerning and Population From Text: Algorithms, Evaluation and application. PhD Thesis, University of Karlsruhe, German.

13. P.Cimiano and J.Volker (2005). Text2Onto: A Framework for Ontology Learning and Data-Driven Change Discovery.

14. T. R. Gruber. (1993). A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199-220.

15. V.Raskin, and S.Nirenburg (2004). Ontological Semantics. MIT Press, Cambridge, MA.

Table 2: The summary on research based on six layers of the different subtasks of learning ontology

	TERM	SYNONYM	CONCEPTS	TAXONOMY	RELATION	RULES AND AXIOM
Text2Onto (P.Cimiano and J.Volker 2005)	Basic linguistic parsing,	WordNet	Concept extraction	Taxonomy extraction	Relation extraction	
Brasethvik and Gulla (2001)	Term extraction		Concept extraction	Taxonomy extraction	Relation extraction	
Artequakt (Alani 2003)	Term extraction, syntactic pattern-based extraction rules	WordNet, a general-purpose lexical database, and GATE			Relation extraction	
OntoLT (Buitelaar. P et al 2004)	Term extraction, Linguistic parsing	Thesaurus extraction		Taxonomy extraction, document clustering	Relation extraction	
Fuzzy Ontology (Lee et al. 2005)	Term extraction		Fuzzy Concept		Fuzzy Relation	
Y.L.Chi (2006)	Term extraction		Concept extraction			
TextToOnto (Maedche and Staab, 2000)	Linguistic parsing	WordNet	Concept extraction	Taxonomy extraction	Relation extraction	
Hasti (M. Shamsfard & A.A. Barforoush 2004) – Persian Language	linguistic parsing		Concept extraction	Taxonomy extraction	Relation extraction	Axiom
A. Kayed & R.M. Colomb (2001)	Term extraction, loosely in NL		Conceptual graph concept			
OntoLearn (A. Cucchiarelli, R. Navigli, F. Neri, P. Velardi, 2004).	Term extraction, Semantic interpretation	WordNet	Concept extraction		Relation extraction	
RelExt (A. Schutz and P.Buitelaar 2005)	Term extraction		Concept extraction	Taxonomy extraction	Relation extraction	