IDENTIFYING THE SEMANTIC RELATIONS ON UNSTRUCTURED DATA

Chien D C Ta[1] and Tuoi Phan Thi[1]

[1]Faculty of Computer Science and Engineering,
HoChiMinh City University of Technology

Abstract

Ontologisms have been applied to many applications in recent years, especially on Sematic Web, Information Retrieval, Information Extraction, and Question and Answer. The purpose of domain-specific ontology is to get rid of conceptual and terminological confusion. It accomplishes this by specifying a set of generic concepts that characterizes the domain as well as their definitions and interrelationships. This paper will describe some algorithms for identifying semantic relations and constructing an Information Technology Ontology, while extracting the concepts and objects from different sources. The Ontology is constructed based on three main resources: ACM, Wikipedia and unstructured files from ACM Digital Library. Our algorithms are combined of Natural Language Processing and Machine Learning. We use Natural Language Processing tools, such as OpenNLP, Stanford Lexical Dependency Parser in order to explore sentences. We then extract these sentences based on English pattern in order to build training set. We use a random sample among 245 categories of ACM to evaluate our results. Results generated show that our system yields superior performance.

Keywords

Domain ontology, Knowledge based system, Semantic Relation, Information Extraction.

1. Introduction

The methods of human computer interaction on Internet as with Google Search, Information Extraction, Question and Answer have become important tools in modern society. End users can use these systems with various purposes, such as querying information, learning, E-commerce, etc. However, the precision of these systems is always a big issue that the research must solve. In order to achieve high precision, the systems should consider semantic of sentences. Previous research has been done in the realm of semantic relation detection but it still remains challenging to this day. V. Malase et al [1] use lexical- syntactic patterns to detect semantic relations between the main terms of definition in order to help terminologist build structured terminology following these relations. With the dramatic increase of data on the Internet, identifying semantic relations plays an important role in semantic-oriented applications.

Our goal is to automatically identify some of the semantic relations that might be found in domain-specific corpora. Here, we will describe methods of identifying semantic relations. For this purpose, we will combine Natural Language Processing and Matching Learning. We will also define some English patterns and semantic roles to identify semantic relations in 2000 text files from ACM Digital Library. We then propose a method to identify synonyms, hyponyms, and hypernyms of instances in domain-specific ontology. Finally, we will use three measures: Precision, Recall and F-Measure in order to evaluate these methods. The evaluation results shown afterward will prove the effectiveness of the proposed mythology.

DOI : 10.5121/ijist.2014.4301

2. Related Work

Information extraction is an important research topic in Natural language Processing (NLP) [2]. It tries to find semantic relations, relevant information from the large amount of text documents and on the World Wide Web. Y. Jie et al [3] focused on semantic rules to build Extraction system from LIDAR (Light Detection and Ranging). F. Gomez et al [4] built a semantic interpreter to assign meaning to the grammatical relations of the sentences when they constructed a knowledge base about a given topic. K. Kongkachandra et al [5] proposed semantic based key-phrase recovery for domain-independent key-phrase extraction. In this method, he added a key-phrase recovery function as a post process of the conventional key-phrase extractors in order to reconsider the failed key phrases by semantic matching based on sentence meaning. Z.Goudong et al [6] proposed novel tree kernel-based method with rich syntactic and semantic information for the extraction of semantic relations between named entities. A.B. Abacha et al [7] built a platform MeTAE (Medical Texts Annotation and Exploration). This system allows the extracting and annotating of Medical entities and relationships from Medical text. He relied on linguistic patterns to detect the semantic relations in medical text files. A.D.S Jayatilaka et al [8] constructed ontology from Web pages. He introduced web usage patterns as a novel source of semantics in ontology learning. The proposed methodology combines web content mining with web usage mining in the knowledge extraction process. H. Li et al [9] extract semantic relations between Chinese named entities based on semantic features and the Vector Space Model (VSM).

Those research attempts were meant to identify semantic relations from web documents or text files in order to develop ontology. They either used natural language processing techniques, the statistical, or the machine learning approach in the ontology learning process. Our research combines Natural Language Processing and Matching Learning to identify semantic relations on unstructured data.

3. Algorithms for Identifying Semantic Relations

3.1 Information Technology Ontology (ITO)

Domain Ontology includes concepts, attributes and events. Ontology is a tuple O = (C, I, R, T, V, $\leq$, $\perp$, $\in$, =) [10], where:
C is the set of concepts, I is the set of individuals including attributes and events.
R is the set of relations; T is the set of data types.
V is the set of values (C, I, R, T, V being pair wise disjoint)
$\leq$ is a relation on$(C\times C) \cup (R\times R) \cup (T\times T)$ called specialization.
$\perp$ is a relation on$(C\times C) \cup (R\times R) \cup (T\times T)$ called exclusion. $\in$ is a relation over $(I\times C) \cup (V\times T)$ called instantiation, = is a relation over $I\times P\times (I\cup V)$ called assignment.
We propose an ontology structure, as shown in Figure.1

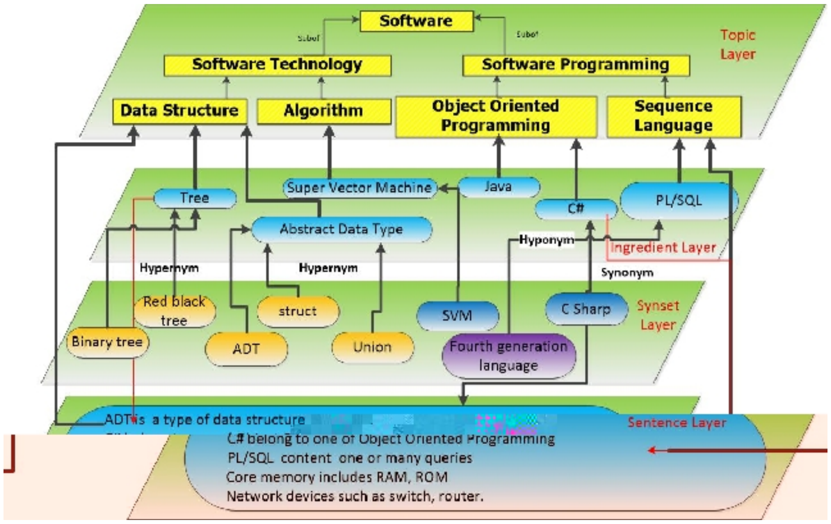

Figure1. Structure of Information Technology Ontology

There are four layers in this Ontology

Topic layer includes 245 categories in Information Technology area from ACM Category [11]. Ingredient layer includes many instances in Information Technology area from difference resources such as Wikipedia and unstructured text files.

Synset layer is a set of synonyms, hyponyms, and hypernyms of instances from ingredient layer. We will describe it in detail in next session.

The last layer is known as sentence layer. We will introduce it in detail in next session.

3.2 Algorithm for Identifying Synonyms, Hyponyms and Hypernyms Relations

The Synset layer in this ontology includes synonyms, hyponyms and hypernyms of instances belonging to an ingredient layer. In order to find a set of synonyms, hyponyms and hypernyms of instances from ingredient layer, we use WordNet. Similarly to Wikipedia, WordNet is an ontology that includes many fields and languages. However, we will only focus on English language. Our proposed algorithm is as follows.

```
   Procedure Find_out_Synset ()
     While (instance of Ingredient layer is not null)
     Begin
   Synonym_List =QueryIntoWordNet (instance)
         If (Synset_List is not null)
     Link (instance to Synonym, hyponym, hypernym)
         End if
     End
   End While
End Pro
```

Table 1 represents results after applying the algorithm for some instances.

Table 1. Set of Synonym, Hyponym and Hypernym corresponding with instances

Instances from Ingredient layer	Synonyms	Hyponyms	Hypernyms
NLP	Natural Language Processing		Informatics, information processing
Data structure		Hierarchical structure	Organization, system
Computer Network		Internet, intranet, WAN	Electronic network
RAM	Random Access Memory	Core memory	Volatile storage

From Table 1, we can identify some semantic relations between an instance of Ingredient layer with its synonyms, hyponyms and hypernyms, such as

- NLP is a Natural Language Processing
- NLP such as Informatics, information processing
- Hierarchical structure includes Data structure
- Data structure such as organization, system
- RAM is random access memory
- Core memory includes RAM

3.3 Algorithm for Identifying Semantic Relations based on Syntax Patterns and Linguistic

To enrich the sentence layer of ontology, we use English syntax patterns to extract sentences related to instances of ingredient layer. This process is implemented during the extraction of instances. For this purpose, we use OpenNLP tool to recognize sentences. Then, we use a Stanford Lexical Dependency Parser (SLDP) to refine them. SLDP can output typed dependency tree to present a grammatical relationship between keywords in the sentence. After that, we apply linguistic markers to the sentences to recognize their semantic meaning. Our proposed model is shown in Figure. 2.

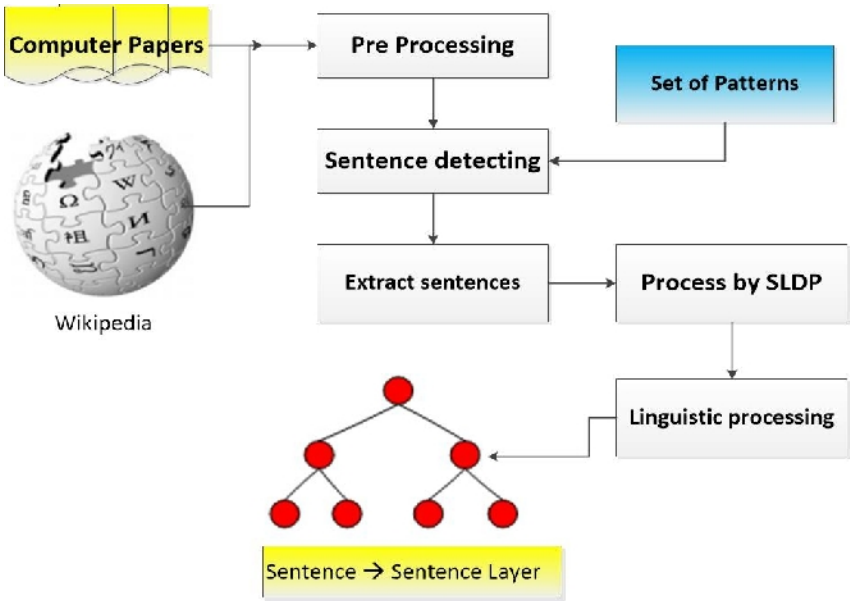

Figure2. Identifying Semantic Relations based on Pattern and Linguistic

Our training corpus (2000 papers) for testing is focused on the field of Information Technology. Electronic documents were automatically collected from ACM digital library. We designed eleven English patterns to select sentences identified by OpenNLP. The English patterns are shown in table 2.

Table 2. English Patterns used to extract sentences

No	Pattern	Example
1	S + V	Computer is broken
2	S + V + Object	He was mowing the lawn
3	S + V + Adjective	The girl was tall
4	S + V + Indirect Object	The woman went to the house
5	S + V + Direct Object	The man hit the ball
6	S + V + Complement	Some students in the class are engineers
7	S + V + Prepositional Phrase	The cat waited for its owner yesterday
8	S +V+ Indirect Object+ Direct Object	Granny left Gary all of her money
9	S + V + Direct Object + Adjective	The Jury found the defendant guilty
10	S + V + Direct Object + NP	The Jury found the defendant guilty
11	S + V + Direct Object + Complement	The class picked Susieclass representative.

After extracting sentences based on the English patterns, we refine them by SLDP. With SLDP, we eliminate the unnecessary words in a sentence. Below are some examples of such sentences.

Table 3. Examples of eliminating unnecessary words

Before applying SLDP	After applying SLDP
COBOL is not popular programming language in recent years.	COBOL is not popular programming language.
Oracle database is one of the Relational Database Management System.	Oracle is Relational Database Management System.
In my opinion Java Language is difficult to program	Java Language is difficult to program
C Sharp in Microsoft Dot Net is also an Object Oriented programming language	C Sharp is Object Oriented programming language

In our ontology, most of the concepts have semantic relations between each other. In order to identify those semantic relations, we rely on linguistic roles (linguistic marker), which are defined as follows:

IS-A: this generic-specific relation reflects hierarchical inheritance in network of concepts. All entities are categorized as instances of a particular class. Class can become instances of a particular class. Thus, any concepts can be linked to its immediate super ordinate concept. For example, Random Access Memory (concept) is a core memory (concept) in computer.

PART-OF: this relation also reflects the hierarchical structure of the domain. This relation directly refers to the parts of each concept in sentence. For example, Random Access Memory (ROM) (concept) is part of memory (concept)

MADE-OF: this relation links to concepts, which made of material concepts. For example, Integrated Circuit (IC) Chip can be made of a semiconductor material.

DELIMITED-BY: this relation marks the boundaries, dividing one concept from another. This is a domain-specific relation, mainly for the concepts, which are belonged to different topic in the field of Information technology. This relation is usually represented by a number of verbs, such as include, delimit, limit, circumscribe, restrict, etc. For example, the processing of a computer is restricted by CPU, RAM.

TAKES-PLACE-IN: this relation describes the context of processes, which are related to spatial and temporal dimensions. A number of verbs represent this relation, such as happen, occur, take place in, etc. For example, In order to tackle the conflict of process, time scheduling takes place in the Operating System.

ATTRIBUTE-OF: this relation is only useful for concepts designated by specialized adjectives, such as strong, powerful, etc., or nouns that define the properties of other concepts. For example, these router devices are powerful and useful in a network.

RESULT-OF: this relation is relevant to either processes or entities that are derived from other processes. For example, as a result of the inconsistency, this file is considered corrupted.

AFFECTS: this relation, along with RESULT-OF, is a crucial semantic relation in the knowledge base since both can relate all kinds of concepts in ontology.

CAUSES: this relation is directly relevant to the processes or concepts that are derived from other processes. On the contrary of RESULT-OF, this relation usually represents negative meaning. These linguistic roles help us identify semantic relations between keywords in a sentence. Therefore, we can recognize sentence meaning and exactly categorize them. The semantic relations are represented, as shown in Figure. 3.

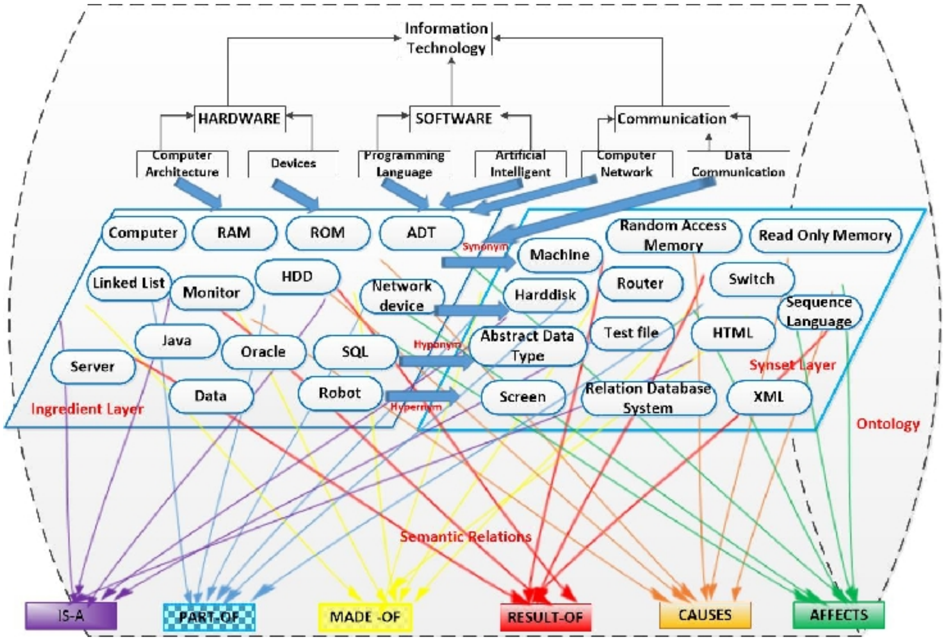

Figure 3. Semantic relations in Information Technology Ontology

4. Experimental Results

Over 2000 text documents from the ACM Digital Library were used for testing purposes. These documents with different topics belonged to the field of Information technology. In order to easily process the documents, we grouped them by category before extracting instances and concepts. In our paper, we chose randomsample corpora among 245 categories from ACM.

The system’s performance was calculatedby using three measures: Precision, Recall and F-measure. They are calculated by each category in domain-specific ontology as below:

$$P(C_i) = \frac{Correct(Ci)}{Correct(Ci) + Bad(Ci)}$$

$$RC_i = \frac{Correct(Ci)}{Correct(Ci) + Missing(Ci)}$$

$$\mathrm{F - Mesure} = 2\frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Where Ci represents a category in ITO and correct, bad, missing represented the number of correct, wrong, missing, respectively.

Experiment results are shown in table 4, 5, 6, 7, and 8.

Table 4. Experiments results on instances of ingredient layer

Category	Qualityof instances	Precision (%)	Recall (%)	F-Measure (%)
Application (AP_inst)	3672	79.26	76.51	77.86
Artificial Intelligent (AI_inst)	5714	82.94	78.92	80.88
Logic Design (LD_inst)	4644	82.18	80.06	81.11
Operating System (OS_inst)	6785	84.47	81.37	82.89
Process Management (PM_inst)	3056	76.53	72.51	74.47
Software (Soft_inst)	4249	81.64	79.62	80.62

Table 5. Experiments results on set of synonyms

Category	Quality of synonym	Precision (%)	Recall (%)	F-Measure (%)
Application (AP_syno)	524	79.26	76.51	77.86
Artificial Intelligent (AI_syno)	689	94.41	88.15	91.17
Logic Design (LD_syno)	472	92.24	84.27	88.08
Operating System (OS_syno)	861	96.18	91.58	93.82
Process Management (PM_syno)	517	93.25	86.16	89.56
Software (SW_syno)	583	94.26	89.04	91.57

Table 6.Experiments results on set of Hyponyms

Category	Qualityof Hyponym	Precision (%)	Recall (%)	F-Measure (%)
Application (AP_hypo)	714	89.38	76.51	82.45
ArtificialIntelligent (AI_hypo)	837	96.14	88.29	92.04
Logic Design (LD_hypo)	718	87.54	84.26	85.86
Operating System (OS_hypo)	972	96.82	91.42	94.04
Process Management (PM_hypo)	728	88.31	85.15	86.70
Software (SW_hypo)	646	85.64	81.04	83.28

Table 7. Experiments results on set of Hypernyms

Category	**Qualityof Hypernym**	**Precision (%)**	**Recall (%)**	**F-Measure (%)**
Application (AP_hype)	916	79.26	76.51	77.86
Artificial Intelligent (AI_hype)	1321	92.41	91.17	91.79
Logic Design (LD_hype)	954	84.62	79.37	81.91
Operating System (OS_hype)	1413	95.04	96.81	95.92
Process Management (PM_hype)	834	82.31	84.55	83.41
Software (SW_hype)	893	85.48	80.19	82.75

Table 8. Experiments results on set of Sentences

Category	**Qualityof Sentences**	**Precision (%)**	**Recall (%)**	**F-Measure (%)**
Application (AP_sent)	683	83.26	80.75	81.98
Artificial Intelligent (AI_sent)	972	87.93	85.64	86.77
Logic Design (LD_sent)	589	81.52	79.18	80.33
Operating System (OS_sent)	1026	92.41	88.32	90.32
Process Management (PM_sent)	647	82.83	79.17	80.96
Software (SW_sent)	762	86.74	82.14	84.38

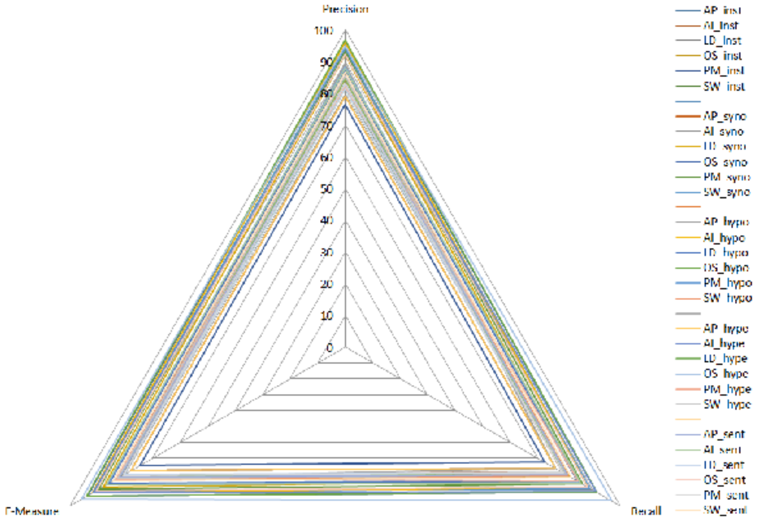

Figure 4. Experiment result evaluation of instances, synonyms, hyponyms and hypernymsbased on Precision, Recall and F-Measures.

5. Conclusions

Our experiment tried to identify semantic relations between nouns or noun phrases in unstructured files based on linguistic roles in order to build domain ontology on Information Technology. After usingan OpenNLP tool to recognize words and sentences, we applied English patterns defined by us to extract them. We then refined them using the SDLC tool. Therefore, our experimental results have a high precision and high recall. In order to identify semantic relations, we applied linguistic roles. The semantic roles directly link to instances in the layers of domain ontology. Also, we propose an algorithm to identify semantic relations based on synonyms, hyponyms and hypernyms from instances of the ingredient layer. Overall scores are computed based on three measures: Precision, Recall ad F-Measure. Efforts must also be invested in order to reduce the overall processing time of the system.

In future works, we will focus on building an Information Extraction system based on this ontology to solve a number of problems, such as ontology learning and Question and Answer.

REFERENCES

[1] V. MalaSe et al , "Detecting semantic relations between terms in definitions," in The 3rd International Workshop on Computational Terminology - CompuTerm 2004 , 2004.

[2] G. Zhou et al, "Tree Kernel-Based Semantic Relation Extraction with Rich Syntactic and Semantic Information," Information Sciences, vol. 180, no. 8, pp. 1313 - 1325, 2010.

[3] Y. Jie et al, "Building Extraction from LIDAR based Semantic Analysis," Geo-Spatial Information Science, vol. 9, no. 4, Sep. 2006.

[4] F. Gomez et al, "Semantic interpretation and knowledge extraction," Knowledge-Based Systems, vol. 20, no. 1, pp. 51 - 60, July 2006.

[5] G.Kongkachandra et al, "Abductive Reasoning for Keyword Recovering in Semantic-based Key-word Extraction," in The Fifth International Conference on Information Technology: New Generations - IEEE, 2008, pp. 714 - 719.

[6] Z. Goudong et al, "Tree kernel-based semantic relation extraction with rich syntactic and semantic information," Information Sciences, vol. 180, no. 8, pp. 1313 - 1325, Dec. 2009.

[7] A.B Abacha et al, "Automatic Extraction of Semantic Relations between Medical Entities- a rule based approach," Journal of Biomedical Semantics, 2011.

[8] A.D.S Jayatilaka, "Knowledge Extraction for Semantic Web Using Web Mining ," in The International Conference on Advances in ICT for Emerging Regions (ICTer 2011) - IEEE, 2011, pp. 89 - 94.

[9] H. Li et al, "A Relation Extraction Method of Chinese Named Entities based on Location and Semantic Features," Applied Intelligence, vol. 18, no. 1, pp. 1- 14, May 2012.

[10] J. Euzenat et al, Ontology Matching.: Springer, 2007.

[11] ACM. [Online].http://www.acm.org/about/class/ccs98-html.

AUTHORS

Tuoi Phan Thi, Professor. : She is a Professor at the Faculty of Computer Science and Engineering, HCMC University of Technology, Vietnam. She obtained her Ph.D. in Computer Science from Charles University, Czech Republic, in 1985. Her research interests are compiler, information retrieval, natural language processing. She has been the Chief Investigator of national key projects and published many papers in international journals and conference proceedings in those areas.

Chien D C Ta, Mr .: He is Ph.D. Student at the Faculty of Computer Science and Engineering, HCMC University of Technology, Vietnam. His research interests include Natural Language Processing and its applications, Information Extraction.