VOL. 10, NO. 2, FEBRUARY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

SYNONYMOUS NON-TAXONOMIC RELATIONS EXTRACTION

N.F. Nabila[1], Nurlida Basir[1] and A. Mamat[2]
[1]Fakulti Sains dan Teknologi, Universiti Sains Islam Malaysia (USIM) 71800 Nilai, N. Sembilan, Malaysia
[2]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM), Serdang, Selangor, Malaysia
E-Mail: fatin@usim.edu.my

ABSTRACT

Construction of ontology is a difficult task, expensive and time-consuming. Concept, taxonomy and non-taxonomic relations, are the three important components in the development of ontology. These three components are used to represent the whole domain texts. Currently, most of studies focused on extracting the concept, the taxonomic relationships and the non-taxonomic relationships within the scope of single sentence. In order to enrich the domain ontology, we introduced a method to extract the non-taxonomic relations by using the similarities of relations that exist in more than one sentence. The most appropriate predicate are used as a reference to relate between concepts that occur not only in the same sentence, but also in different sentences. Here, the proposed method was tested using a collection of domain texts that described electronic voting machine and are evaluated based on the standard information retrieval performance metrics, i.e. precision and recall.

Keywords: ontology, semantic, taxonomic, non-taxonomic, predicate.

INTRODUCTION

The term 'ontology' originated from philosophy which means "theory of existence". However, in computer science, ontology is a model to describe the topic area that consists of terms, the properties and the types of relationships. Nowadays, ontology becomes a popular topic of research in many areas of computer science such as artificial intelligence, information retrieval, and semantic web. Considerable efforts have been made in constructing ontologies due to its complexity and time-consuming task (Shamsfard and Barforoush, 2004).

Manual construction of ontology is a difficult task, expensive and time consuming (Shamsfard and Barforoush, 2004). Therefore, several works such as (Kavalec *et al*., 2004), (Maedche and Staab, 2000) and (Navigli, 2003) introduce an automatically or semi-automatically ontology using textual data in order to reduce the time and effort required for manual ontology construction. Even though the number of approaches for constructing ontology is increasing, most of these approaches only focuses on extracting the concept (Pantel and Lin, 2001), (Punuru and Chen, 2012), (Tomokiyo and Hurst, 2003) and taxonomic (is-a) relationships component (Caraballo, 1999), (Cimiano and Staab, 2004), (Hearst, 1992) and often neglect the importance of relationships other than is-a relation, also known as non-taxonomic relationships (Liu *et al*., 2005).

Thus, recently, several researches such as (Imsombut, 2009), (Punuru and Chen, 2007), (Villaverde *et al*, 2009) focus on non-taxonomic relations to identify non-taxonomic relationships between two concepts i.e., the relation between subject and object that occur in the same sentence. In this paper, we propose an approach to improve the extraction of non-taxonomic relation in order to enrich domain ontology from domain text. This work is a continuation of previous work (Nabila *et al*, 2011). The aim is to extract non-taxonomic relationships between concepts that occur not only in the same sentence, but also in different sentences, and thus, increase the number of relations extracted and properly represent the domain. The proposed method was tested using a collection of domain texts that described electronic voting machine and are evaluated based on the standard information retrieval performance metrics, i.e. precision and recall.

RELATED WORKS

Extracting non-taxonomic relations is one of the important tasks in the construction of ontology from texts. Most previous works have focused on extracting predicate or verb phrase that links concept as subject and concept as object in the same sentence as potential relationships. (Kavalec *et al*, 2004) identified the transaction that holds two concepts if they frequently occur within the predefined distance from the verb as verb-concept-concept (VCC (n)). (Akbik and Brob, 2009) developed Wanderlust to finds semantic relations between two entities using dependency grammar patterns. (Imsombut, 2009) used several heuristic rules to identify the subject and the object of the verbs and assume all Noun Phrases (NPs) that occur before verb are selected, as subject of the sentence and NPs that occurs after verb are selected as object of the sentence. Then, the verbs are expanded by gathering all verbs occurring between the same patterns of concept pair. In (Villaverde, 2009), to identify nouns and verb phrases, the part-of-speech (POS) tagger was applied to each sentence from the documents collection to fulfill the pattern: <term><verb><term>, where the terms are ontology concepts that appear in the same sentence with a verb link between them. (Maedche and Staab, 2000) identify relation (i.e. verb phrases) between two concepts, where the concepts are ontology concepts that appear in the same sentence. (Serra and Girardi, 2011) used an NLP approach and data mining technique to identify potential non-taxonomic relationships from textual sources. (Punuru and Chen, 2007) developed Subject-Verb-Object (SVO) Triples method to identify non-taxonomic relations

between two concepts, where the concepts must appear as the subject and the object of a sentence. They used MINIPAR dependency parser to determine the appearance of concepts. Then, the verb that occurs together with the concept pair was identified. PARNT (Serra *et al*, 2013) is a semi-automatic method to extract non-taxonomic relations from texts. Most of these works focused on extracting relationships between two concepts, i.e. subject and object that appear in the same sentence only. This thus limit the number of identified relations than what it should be, and hence, does not properly represent the domain. Our method, in contrast, is capable of identifying relations between subject and object that appear in different sentences and used same predicates or synonymous word.

EXTRACTING NON-TAXONOMIC RELATIONS

The objective of our proposed method is to improve retrieval process of non-taxonomic relations in domain-specific texts. Using this method, non-taxonomic relations that occur in those texts are extracted as many as possible. The method consists of three main steps. The first step of the method is to extract the concepts by using the pre-processing tools. Next, the concepts are classified into two lists, list of subjects and list of objects. The subjects and objects are determined based on the position of the subjects and objects occurrence in texts. The second step is to generate concept pair by using the Cartesian product between these two lists. The third step is to extract and assign potential relations for concept pairs. The relationship among these three main steps is illustrated as in Figure-1.

Extracting Concepts (subjects and Objects) from the Domain Texts

The first step is to extract the concepts i.e., the subjects and objects from texts. This step is divided into three tasks:

a) text-preprocessing,
b) formation of Predicate_subject and Predicate_object Pairs, and
c) classifying concept into subject and object.

In Task 1, the texts are split into sentences and the pre-processing tools and statistical analysis are applied to extract relevant terms from the texts. Here, term is a word of the noun that exists in domain text. Relevant terms are counted as relevant concept for the domain text.

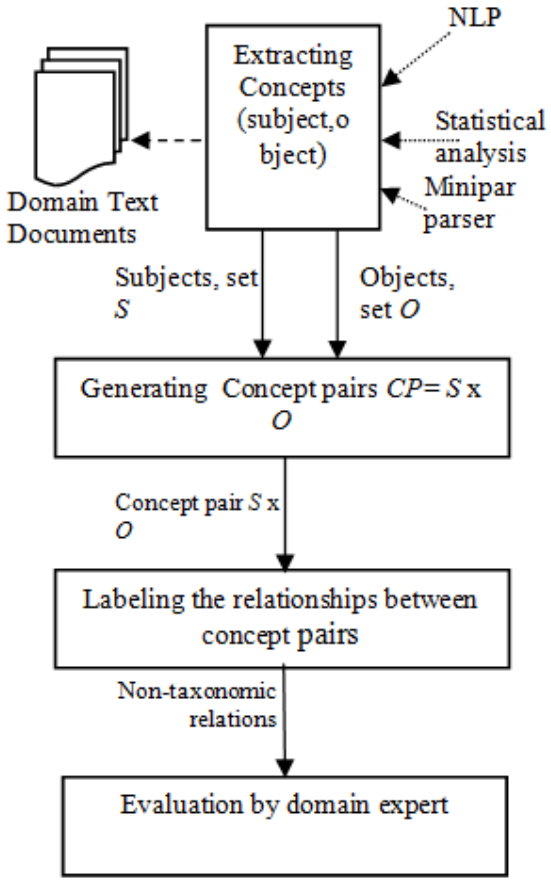

Figure-1. The main steps of proposed method.

In Task 2, the MINIPAR dependency parser (Lin, 2003) is applied to all sentences in domain text to identify the subject and the object of a sentence. For each sentence in the texts, the dependency pairs (i.e. grammatical relation between nouns (subject or object) with the predicate) are extracted and the pairs is represented in the form of (sid, *p (subj))* or (*sid, p (obj)),* where

- *sid* is the sentence identifier in which the predicate and the subject or the object appear.
- *subj* is a noun that appears as subject in a sentence.
- *obj* is a noun that appears as object in a sentence.
- *p* is a predicate that occurs together with a noun in a sentence.

All *subj*, *obj* and *dependencies pair* i.e., (*sid, p(subj)*) and (*sid, p(obj)*) that are extracted from texts are included into Predicate_Subject list, *PS,* or Predicate_Object list, *PO.*

Task 3 is to classify each relevant concepts into Subject or/and Object. In this task, the result produced from the previous two tasks, i.e. pre-processing and formation of predicate_subject or predicate_object pairs are matched. Then, the matched terms are classified into two sets, i.e. S and O.

- $S = \{s_1, s_2, ..., s_n\}$ is a set of relevant terms which appears as subject in sentences of text.
- $= \{o_1, o_2, ..., o_m\}$ is a set of relevant terms which appears as object in sentences of text.

There is a possibility that same term can be a member of both S and O.

Generating concept pairs

The second step in the proposed method is the generation of concept pairs. Here, the concept pair, *CP*, is generated from the Cartesian product of sets *S* and *O*. To prevent the existence of the taxonomic or hierarchical relation between subject and object in concept pairs, several restrictions are used. The concept pairs (*s*, *o*) are generated taking into consideration the following restrictions:

i. *s* is "not same word" with *o* or
ii. *s* is "not synonym" with *o* or
iii. *s* is "not is-a" relation with *o* or
iv. *s* is "not part-of" relation with *o*

Restriction (i) is used to avoid subject and object in concept pair from being the same word. Meanwhile, for restriction (ii), (iii) and (iv), WordNet is used to identify whether *S* has a synonym or *is-a* or *part-of* relation with O. In WordNet, hypernym refers to *is-a* relation while hyponym refers to show *part-of* relation. Concept pairs that fulfill the restrictions are considered as a valid pair and the relationships for the valid pairs are identified in the third step.

Labeling the relationship between concept pair

The third step is to extract the potential relation to label the generated concept pair. In our work, we consider three cases of predicate between subject-object pair (s, o). Case 1 is used to identify the relation between concept pair, i.e. the subject and the object that appear in the same sentence. Cases 2 and 3 are used to identify the predicate between subjects and objects that appear in different sentences. Here, we represent the list of *predicate_subject* pair and *predicate_object* pair as the following:

Let $(sid_i , p_i(s_i)) \in PS$, $(sid_j , p_j(o_j)) \in PO$, and (s_i , o_j) be the concept pair.

For example, we used six sentences extracted from voting machine domain text and expressed the sentences in the Table as shown in Table-1.

Table-1. Example of Sentences from Domain Text.

Sid	Predicate	Subject	Object
1	supply	company	
2	report	machine	ballot
3	produce		paper
4	produce	machine	
5	provide		machine
6	use		machine

We can determine the potential relation for concept pairs based on three cases as below.

Case 1: Subject, object and predicate in the same sentence.

If the predicate, subject and object appear in the same sentence, then the predicate is considered as the potential relation of the concept pairs. This idea can be expressed by the following rule,

If $((sid_i = sid_j) \&\& (p_i = p_j))$, *then* establish relation $p_i (s_i , o_j)$.

As an example, if we have a concept pair (machine, ballot), then we identified all predicate-subjects, p(machine)in the PS list and predicate-objects, p(ballot) in the PO list (refer Table-1). For instance,
(2, report (machine))
(4, produce (machine)), are in PS and
(2, report (ballot)) is in PO, hence
hence report (machine, ballot) is established.

In this example, predicate *report* exists in both PS and PO list, which is in the same sentence no 2. Therefore, the predicate *report* is considered a potential relation for concept pair *machine* and *ballot.*

Case 2: Predicate_subject and predicate_object occurs in different sentences but predicates are the same.

Case 2 is used to identify relation between concepts that appear in different sentences, but the predicates are the same. Since the predicates in PS and PO are the same, we assume the predicate can be considered as relations between concepts. In rule,

If $((sid_i \neq sid_j) \&\& (p_i = p_j))$, *then* establish relation $p_i (s_i , o_j)$.

For example, given a concept pair *(machine, paper)*, and assume predicate_subject pairs

(2, report (machine))
(6, produce (machine)) are found in PS list and
(3, produce (paper)) are in PO list.

Here, we can derive a relation of predicate *produce* with *subject* and *object* and represent as *produce (machine, paper).*

Case 3: Predicate_subject and predicate_object occurs in different sentences but predicates are synonym.

Case 3 is used to identify predicates that are synonym. In this case, since the predicates in PS and PO are synonymous, then both predicates can be considered as relations between subject and object. Any one of the predicate is chosen to represent the relationships as both are similar in meaning. In rule,

If $((sid_i \neq sid_j)$ *&&* $(p_i$ *synonym* $p_j))$, *then* establish relation $p_i\ (s_i\ ,\ o_j)$ or $p_j\ (s_i\ ,\ o_j)$.

For example, given a concept pair *(company, machine)* and assume
(1, *supply (company))* is found in PS list, and
(5, provide (machine))
(6, use (machine)) are in PO list.

Referring to WordNet, the word *supply* is synonymous with the word provide. Since the predicates are synonyms, we can use only one predicate to represent the relationships for concept pair, for example *supply (machine, trail).*

EXPERIMENTS

For conducting the experimental evaluation, we selected a set of voting machine domain texts collected from New York Times website with over 19, 111 words. Here, we compared our findings on the number of identified relations in the domain texts with the SVO method.

Based on the results (see Table-2), the SVO method generated 93 relations of which 84 are correct relations and 9 are incorrect relations. In contrast, the proposed method extracted 364 relations of which 236 are correct relations, which is higher than the SVO method. The recall value for the SVO method is 17.0% and the proposed method is 47.8%. It can be concluded that the proposed method produced better results compared to the SVO method.

Table-2. Correct and incorrect relations for the total generated relations.

Method	Total generated relations	No. of correct relations	No. of incorrect relations	Precision	Recall
SVO Method	93	84	9	90.3	17.0
Proposed Method	364	236	128	64.8	47.8

Table-3. Results for non-taxonomic relations for voting machine domain texts.

Method	No. of relation where the subject and object in		Total relations
	Same sentence	Different sentences	
SVO Method	84	0	84
Proposed Method	84	152	236

In Table-3, the proposed method has proved that there are non-taxonomic relationships between subjects and objects that occur not only in the same sentence, but also in different sentences. Based on the same domain texts, the domain expert has identified 494 valid relations. Therefore, it shows our proposed method have successfully found correct relations that are almost close to the manual domain expert evaluation. In addition, our proposed method also helps in enrichment of the domain ontology.

CONCLUSIONS

In this paper, we proposed a method for extraction of non-taxonomic relations to enrich the domain ontology from domain text. Here, the relation between subject and object that occur in different sentences are also extracted and thus increases the number of relation extracted and properly represent the domain. From the results of experiment, it is clear that the presented method was able to increase the number of relations extracted. In future, we plan to select an appropriate predicates representing a relationships for a subject-object pair.

ACKNOWLEDGEMENTS

Universiti Sains Islam Malaysia and Universiti Putra Malaysia.

REFERENCES

Akbik A. and Brob J. 2009. Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns. In WWW Workshop.

Caraballo S. A. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 120-126.

Cimiano P. and Staab S. 2004. Learning by googling. ACM SIGKDD explorations newsletter. 6(2): 24-33.

Hearst M. A. 1992. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on Computational linguistics. 2: 539-545.

Imsombut A. 2009. A statistical approach for semantic relation extraction. In Natural Language Processing, SNLP'09. Eighth International Symposium, IEEE. pp. 54-58.

Kavalec M., Maedche A. and Svátek V. 2004. Discovery of lexical entries for non-taxonomic relations in ontology learning. In SOFSEM 2004: Theory and Practice of Computer Science, Springer Berlin Heidelberg. pp. 249-256.

Lin D. 2003. Dependency-based evaluation of MINIPAR. In Treebanks, Springer Netherlands. pp. 317-329.

Liu W., Weichselbraun A., Scharl A. and Chang E. 2005. Semi-automatic ontology extension using spreading activation. Journal of Universal Knowledge Management. 1, pp. 50-58.

Maedche A. and Staab S. 2000. Semi-automatic engineering of ontologies from text. In Proceedings of the 12th international conference on software engineering and knowledge engineering. pp. 231-239.

Maedche A. and Volz R. 2001. The ontology extraction and maintenance framework Text-To-Onto. In Proc. Workshop on Integrating Data Mining and Knowledge Management, USA.

Nabila N.F, Mamat A, Azmi-Murad M.A. and Mustapha N. 2011. Enriching non-taxonomic relations extracted from domain texts. Semantic Technology and Information Retrieval International Conference. pp. 99-105.

Navigli R., Velardi P. and Gangemi A. 2003. Ontology learning and its application to automated terminology translation. Intelligent Systems, IEEE. 18(1): 22-31.

Pantel P. and Lin D. 2001. A statistical corpus-based term extractor. In Advances in Artificial Intelligence, Springer Berlin Heidelberg. pp. 36-46.

Punuru J. and Chen J. 2012. Learning non-taxonomical semantic relations from domain texts. In Journal of Intelligent Information Systems. pp. 191-207.

Punuru J. and Chen J. 2007. Extracting of non-hierarchical relations from domain texts. In Computational Intelligence and data mining, CIDM, IEEE Symposium. pp. 444-449.

Serra I., Girardi R. and Novais P. 2013. PARNT: A Statistic based Approach to Extract Non-Taxonomic Relationships of Ontologies from Text. In Information Technology: New Generations (ITNG), Tenth International Conference, IEEE. pp. 561-566.

Serra I. and Girardi R. 2011. A Process for Extracting Non-Taxonomic Relationships of Ontologies from Text Intelligent Information Management. 3(4).

Shamsfard M. and Barforoush A. A. 2004. Learning ontologies from natural language texts. International journal of human-computer studies. 60(1): 17-63.

Tomokiyo T. and Hurst M. 2003. A language model approach to keyphrase extraction. In Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment- Volume 18, Association for Computational Linguistics. pp. 33-40.

Villaverde J., Persson A., Godoy D. and Amandi A. 2009. Supporting the discovery and labeling of non-taxonomic relationships in ontology learning. Expert Systems with Applications. 36(7): 10288-10294.