STRUCTURING TAXONOMIES BY USING LINGUISTIC PATTERNS AND WORDNET ON WEB SEARCH

Ana B. Rios-Alvarado, Ivan Lopez-Arevalo and Victor Sosa-Sosa
Information Technology Laboratory, CINVESTAV, Cd. Victoria, Tamaulipas, Mexico

Keywords: Text mining, Knowledge representation.

Abstract: Finding an appropriate structure for representing the information contained in texts is not a trivial task. Ontologies provide a structural organizational knowledge to support the exchange and sharing of information. A crucial element within an ontology is the taxonomy. For building a taxonomy, the identification of hypernymy/hyponymy relations between terms is essential. Previous work have used specific lexical patterns or they have focused on identifying new patterns. Recently, the use of the Web as source of collective knowledge seems a good option for finding appropriate hypernyms. This paper introduces an approach to find hypernymy relations between terms belonging to a specific knowledge domain. This approach combines WordNet synsets and context information for building an extended query set. This query set is sent to a web search engine in order to retrieve the most representative hypernym for a term.

1 INTRODUCTION

At the beginning of the 21st century the easy way to access to digital information resources has motivated an exponential growth in the available unstructured information. This growth is not only present on web resources, but it also can be seen inside organizations, institutions, and companies. In an organization, for example, documents represent a significant source of collective expertise (*know how*). In order to store, retrieve, or infer knowledge from this information, it is necessary represent it using a conceptual structure. This can be achieved by means of taxonomies or ontologies.

An ontology can be build in a manual manner through the knowledge engineers and domain experts, resulting on long and tedious development stages, which can result in a knowledge acquisition bottleneck (Maedche and Staab, 2001). As a consequence, nowadays an important research area is *ontology learning*. Ontology learning is defined as a set of methods used for building from scratch, enriching or adapting an existing ontology in a semi-automatic fashion using heterogeneous information resources (Sánchez, 2009). The ontology learning deals with entities discovery and how such entities can be grouped, related, and subdivided according to their similarities and differences. In ontology learning, an unsupervised manner to build conceptual structures is to use text (terms) clustering techniques. Syntactic patterns or grammatical classes could, for example, be used to provide candidates for term detection. However, these approaches do not consider that words are ambiguous and sharing a semantic context. In this sense, Pantel and Lin (Pantel and Lin, 2002) provide a soft clustering algorithm called *Clustering by Committee* (CBC) which can assign words to different clusters using sets of representative elements (called *committees*) that try to discover unambiguous centroids for describing the members of a possible class. This method only creates clusters of terms, but it does not create a hierarchical structure. Cicurel *et al.* (Cicurel et al., 2007) evaluated CBC concluding that is a good technique to identify senses of words. Its disadvantage is that it requires adjust some parameters, for example the threshold between the centroid and any element for grouping. However, the use of an unsupervised learning techniques makes possible to calculate these parameters.

According to Gruber (Gruber, 1993), *"ontologies are often equated with taxonomic hierarchies of classes"*; thus, it can be said that the key component in the ontology is the taxonomy. Such taxonomies, as the main component for an ontology provide an organizational model for a domain (domain ontology), or a model suitable for specific tasks or problem solving methods (ontologies of tasks and methods) (Burgun and Bodenreider, 2001). Nevertheless, constructing

B. Rios-Alvarado A., Lopez-Arevalo I. and Sosa-Sosa V..
STRUCTURING TAXONOMIES BY USING LINGUISTIC PATTERNS AND WORDNET ON WEB SEARCH.
DOI: 10.5220/0003665902730278
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2011), pages 273-278
ISBN: 978-989-8425-80-5

taxonomy is a very hard task.

The identification of hypernymy/hyponymy relations between terms (in this work only nouns are considered as terms) is mandatory for building a taxonomy. A hyponym can be defined as: a word of more specific meaning than a general or superordinate term applicable to it. By contrast, a hypernym is a word with a broad meaning constituting a category under which more specific words fall. For example, *Mercury*, *Jupiter*, and *Mars* are hyponyms of *Planet* whereas *Planet* is a hypernym of *Mercury*, *Jupiter*, and *Mars*. Other names for the hyponym relationship are *is-a*, *parent-child*, or *broader-narrower* relationships (Cederberg and Widdows, 2003). Caraballo (Caraballo, 1999) claimed that according to WordNet, *"a word A is said to be a hypernym of a word B if native speakers of English accept the sentence B is a (kind of) A"*.

In recent years, the Web has become a source of collective knowledge, reason why it seems a good option for finding suitable hypernyms. In addition to using Web and lexical patterns, some works (Snow et al., 2005), (Ortega-Mendoza et al., 2007) identify new lexical patterns that make possible to obtain more specific hyponyms; but it is necessary rely on the known hyponymy relationships for training a classifier, which is not always possible. In this paper, an approach to find hypernym relations between terms from text belonging to domain knowledge is presented. Particularly, this approach combines WordNet synsets and contextual information for building an extended query set. With this query set, a web search is executed in order to retrieve the most representative hypernym for a term.

The rest of this document is structured as follows. In Section 2, a brief description of the related work about automatic discover of hypernyms is given. In Section 3 the approach and the method to find hypernyms are described. Later, in the Section 4, the experiments and preliminary results are presented. Finally, Section 5 gives some conclusions and the further work.

2 RELATED WORK

One of the first ideas in automatic discovering hypernyms from text was proposed by Hearst (Hearst, 1992). She proposed a method to identify a set of lexico-syntactic patterns occurring frequently in the text. Caraballo (Caraballo, 1999) proposed to automatically build a noun hierarchy from text using data on conjunctions and appositives appearing in the Wall Street Journal corpus. Both methods are limited by the number of patterns used. Pantel *et al.* (Pantel et al., 2004) showed how to learn syntactic patterns for identifying hypernym relations and binding them with clusters that were built from co-occurrence information. Blohm and Cimiano (Blohm and Cimiano, 2007) proposed a procedure to find lexico-syntactic patterns indicating hypernym relations from the Web. From this work, Ortega-Mendoza *et al.* (Ortega-Mendoza et al., 2007) and Sang (Sang, 2007) developed a method to extract hyponyms and hypernyms using lexical patterns respectively. Snow *et al.* (Snow et al., 2005) generated hypernym patterns and combined them with noun clusters to generate high-precision suggestions for unknown noun insertion into WordNet. Ritter *et al.* (Ritter et al., 2009) presented a method based on lexical patterns that find hypernyms on arbitrary noun phrases. They used a Support Vector Machine classifier to find the correct hypernyms from matches to the Hearst patterns. Most of these studies are limited due to the hand selection of pairs of terms that a hypernym relationship has, which represents the initial seed for discovering new patterns. In this sense, the automatic acquisition of terms is essential. The Schutz and Buitelaar approach (Schutz and Buitelaar, 2005) uses linguistic analysis and a predefined ontology for relation extraction with the purpose of extending domain ontology. Cimiano and Staab (Cimiano and Staab, 2004) showed that a potencial way to avoid the *knowledge acquisition bottleneck* is acquiring collective knowledge from the Web using a search engine. This idea was used by Sánchez (Sánchez, 2009), using the Web for acquiring taxonomic and non-taxonomic relationships.

3 THE METHOD

According to the ontology learning, two of the main components in an ontology are concepts and relationships. These elements should be relevant in the domain of the input corpus. This section introduces a method for extracting relevant hypernyms from the information given by specific corpus that is also complement with knowledge retrieved from the Web.

3.1 The Representation Model

Typically text is represented using the *bag of words* model. This model assumes that the order of words has no significance. However, current applications consider that a semantic representation focused on *Natural Language Processing* (NLP) has a major potential for new developments. Thus, word-context matrices and pair pattern matrices are most suitable

for measure the semantic similarity of word pairs and patterns (Turney and Pantel, 2010). In the approach presented in this paper, the proposal is to use a syntactic parser to extract the grammatical context where each word occurs. It is of special interest the focus on dependency relationships <*subject, verb*> and <*verb, object*>. With these relationships, representative pairs of words in a context (topic) are identified. The verbs are considered because they specify the interaction between two participants in an action and express their relationship (Schutz and Buitelaar, 2005). A pair-term matrix is used as representation model (see Figure 1):

Verbs	*Nouns*					
visit	Porto Novo		church			
go				safari	gallery	Los Angeles
....	.....	.....	.....	.....	.....	.....
like		cup of tea				

Figure 1: Example of pair verb-noun matrix.

	N1	N2	N3	N4	....	Nn
V2	14.0	0.0	17.0	0.0	4.0	12.0
V3	0.0	0.0	12.0	20.0	0.0	0.0
	.....	.....	.....	.....	.....	.....
Vm	0.0	7.0	0.0	0.0	0.0	11.0

Figure 2: Example of values of pair verb-noun matrix.

By means of mutual information is possible to find two related terms. The *Pointwise Mutual Information* (PMI) is the measure used for the association strength between two words (w_1, w_2). By using the Equation 1, the values of mutual information was calculated.

$$PMI(w_1, w_2) = log_2 \frac{p(w_1 AND w_2)}{p(w_1) * p(w_2)} \quad (1)$$

For each verb-noun pair (Figure 1) their PMI is calculated, thus, the pair verb-noun is mapped to numerical values as the Figure 2 shows. The representation model is obtained on the overall corpus.

3.2 Querying the Web

For obtaining close results to the domain of the input corpus, it is proposed the construction of an extended query set that considers the more representative terms in the input corpus and in the WordNet synsets. The obtained results (pages) are processed to get relevant hypernyms. In general, discovering hypernyms consists of the following phases (see Figure 3).

- Pre-processing: It is performed to identify dependencies between nouns sharing a verb in the same context. These dependencies are obtained using the Minipar[1] parser. A pair-pattern matrix is used as representation model. In the pair pattern matrix, the pairs correspond to the terms appearing in a triple term structure <*subject*> *verb* <*object*>. A noun can be a subject or an object within a sentence. The representative nouns are obtained by pairs like <*subject-verb*> and <*verb-object*>.

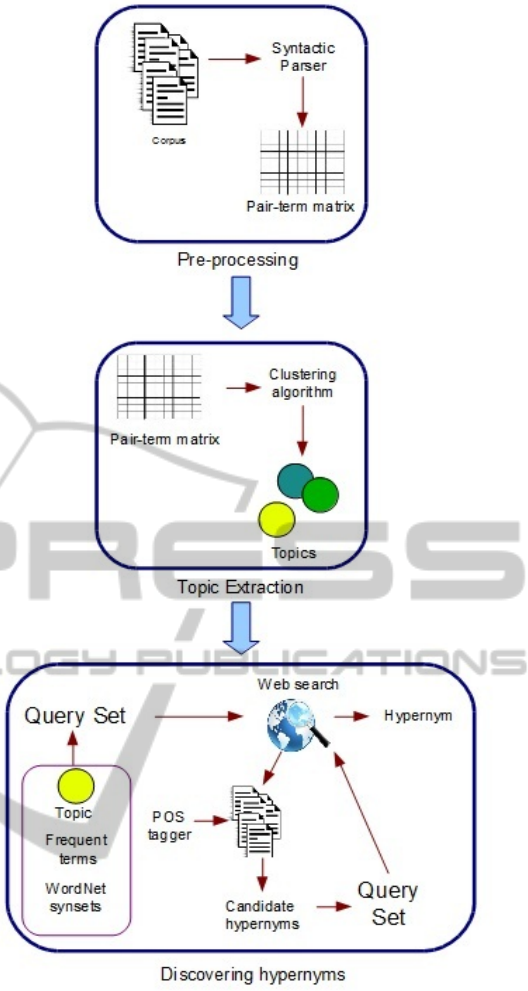

Figure 3: Method for discovering hypernyms.

- Topic extraction: The topics from the corpus are inferred using an adaptation of the CBC algorithm proposed by Pantel (Pantel, 2003).
- Discovering hypernyms: For each topic, a taxonomy is constructed. For each noun in the topic, a set of queries is generated. It it considered the following:

  1. The Hearst's patterns have shown good evidence identifying that entity *A* (noun) is a hyponym of *B*. However, Snow *et al.* (Snow et al., 2005) also identified other possible patterns as result of their method for discovering

[1] http://webdocs.cs.ualberta.ca/simlindek/minipar.htm

hypernyms (see Table 1). Both set of patterns are considered in this work.

Table 1: Lexical patterns.

Hearst's patterns	Other patterns
A, and other *B*	*B*, called *A*
A, or other *B*	*B*, particularly *A*
A is a *B*	*B*, for example *A*
B, such as *A*	*B*, among which *A*
B, including *A*	
B, especially *A*	

2. A general query on the Web like *such as* <*hyponym*> is not enough to obtain interesting and precise information. In order to get useful information, the query needs to be more specific (Sang, 2007). This is the reason why related information is added to the query: 1) contextual information and 2) supervised information. The contextual information is given to the terms with the higher frequencies in the corpus (without stopwords and after a lemmatization process). The supervised information is given to the more representative terms in the WordNet synset corresponding with the term. For extracting terms from WordNet, the gloss of the term is tagged; the words (three words) labeled as noun are considered as supervised information. If a term has more than one synset, the first synset is taken.
3. Query sets are constructed using the lexical patterns and the related information. Each query is sent to a web search engine for using the Web as a source of knowledge.
4. For each query in the hypernym query set, the *n* first pages are retrieved. The text for each n page is cleaned and parsed avoiding non-essential information (eliminating images, videos, banners, etc.). Each sentence is POS-tagged using the Stanford tagger[2], thus the lexical pattern of the query and their candidate hypernym are identified. A term is selected as hypernym if it is a noun but it is not a stopword.
5. The list of candidate hypernyms is evaluated using a new query set, where each possible hypernym will be replaced in the lexical pattern. Using its query set and the number of hits obtained in the web search each candidate hypernym (CH) is evaluated by means of the following measure to score candidate hypernym (SCH) (Cimiano and Staab, 2004) (Equation 2):

$$SCH = \frac{hits(LexicalPattern(term,CH))}{hits(CH)} \quad (2)$$

where the *LexicalPattern(term,CH)* represents a query like: <*term*>, + *and* + *other* + <*CandidateHypernym*>; *and other* corresponds to some lexical pattern. The total score for a CH is given by the sum of scores obtained for each lexical pattern. Thus, the hypernym with the highest total score in the result for the query will be the hypernym associated to the term.

4 EXPERIMENTS AND RESULTS

A sample of the Lonely Planet[3] corpus was used in the experiments. To illustrate the experiment, the term *museum* was considered. The terms with the higher frequencies in the sample corpus were: *cash*, *travel*, and *product*. The extracted words from the WordNet synset for *museum* were: *collection*, *object*, and *display*; their lexical pattern query set is shown in Table 2 and Table 3. Using the query set with only lexical patterns, the list of candidate hypernyms was: <*site, place, attraction, department of history*>. Using a query with added information, the new candidate hypernyms were: <*depository, institution*>. A new lexical pattern query set was created using each one. Then, using the number of obtained hits in the web search, the corresponding score was computed for each candidate hypernym. For example, for the term *attraction*, the obtained hits are shown in Table 4.

Table 2: Example of a web query set for term *museum* using the higher frequency terms in the Lonely Planet Corpus.

museum,+*and*+*other*+cash+travel+product
museum,+*or*+*other*+cash+travel+product
museum+*is*+*a*+cash+travel+product
such+*as*+museum+cash+travel+product
including+museum+cash+travel+product
especially+museum+cash+travel+product
called+museum+cash+travel+product
particularly+museum+cash+travel+product
for+*example*+museum+cash+travel+product
among+*which*+museum+cash+travel+product

In Table 5 can be seen that the best hypernym to *museum* is *attraction* and into the tourist context could be a good option, but it is important to note that the second best candidate is *institution*. According to different authors, the definitions of *museum* are:

...a museum is a building or institution which houses and cares for a collection of artifacts and other objects of scientific, artistic, or historical importance and makes them

[2]http://nlp.stanford.edu/software/tagger.shtml

[3]http://olc.ijs.si/lpReadme.html

Table 3: Example of a web query set for the term *museum* using WordNet synsets.

museum,+*and*+*other*+collection+object+display
museum,+*or*+*other*+collection+object+display
museum+*is*+*a*+collection+object+display
such+*as*+museum+collection+object+display
including+museum+collection+object+display
especially+museum+collection+object+display
called+museum+collection+object+display
particularly+museum+collection+object+display
for+*example*+museum+collection+object+display
among+*which*+museum+collection+object+display

Table 4: Example of the web query set for evaluating the term *attraction*.

Query	Hits
museum,+*and*+*other*+attraction	12300000
museum,+*or*+*other*+attraction	12300000
museum+*is*+*a*+attraction	26900000
attraction+*such*+*as*+museum	26900000
attraction+*including*+museum	26900000
attraction+*especially*+museum	26900000
attraction+*called*+museum	11600000
attraction+*particularly*+museum	26800000
attraction+*for*+*example*+museum	3780000
attraction+*among*+*which*+museum	12500000

Table 5: Total score of candidate hypernyms for term *museum*.

Candidate hypernym	Total score
attraction	3.74220
institution	3.65833
depository	1.50125
department of history	0.82055
place	0.21463
site	0.09794

available for public viewing through exhibits that may be permanent or temporary...[4]

Museums enable people to explore collections for inspiration, learning and enjoyment. They are institutions that collect, safeguard and make accessible artefacts and specimens, which they hold in trust for society...[5]

The museum is an empowering institution, mean to incorporate all who would become part of our shared cultural experience...[6]

According to the added information to queries, the term *institution* is a good candidate hypernym to *museum*. The Figure 4 shows the created taxonomy for the group of terms <*art, culture, library, science, book, travel*> related to *museum*. The taxonomy is constructed following next steps: pairs of terms are used for building a extended query set, thus it is sent to the web search engine. The method finds that one of two terms into the pairs is hypernym of the others. The method is repeated without the hypermym found previously. The group of terms is the result of the CBC clustering algorithm.

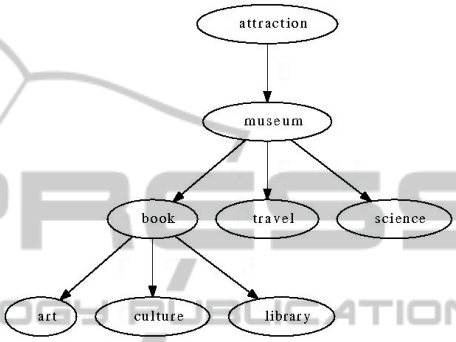

Figure 4: Taxonomy created for the group of terms related with *museum*.

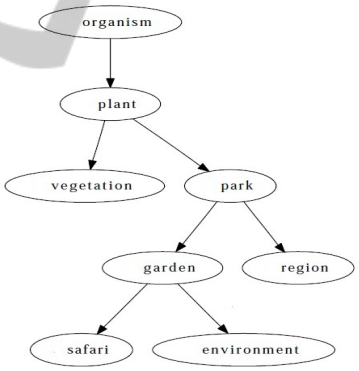

Figure 5: Taxonomy created for the group of terms related with *plant*.

Following the experiments, a query set was constructed for the term *plant* and their group of related terms using the WordNet synsets terms: *flora*, *botany*, and *organism*. In Table 6 can be seen the hypernyms obtained for each term and their appropriate WordNet hypernym for the group of terms <*plant, vegetation, park, garden, region, safari, environment*>. The found hierarchical structure is shown in Figure 5. Note that these taxonomies (Figure 4 and 5) corresponds only for the information extracted from the input corpus (Lonely Planet), they are not from the general domain, such taxonomies can be enhanced by using an additional corpus.

[4]Edward Porter Alexander, Mary Alexander. Museums in motion: an introduction to the history and functions of museums. Rowman & Littlefield, 2008 ISBN 0-7591-0509-X

[5]http://www.museumsassociation.org/about/frequently-asked-questions

[6]Mark Lilla. The Great Museum Muddle. New Republic, April 8, 1985. pp.25-29

Table 6: Hypernyms obtained and WordNet hypernym for the group of terms related with term *plant*.

Term	Hypernym obtained	WordNet hypernym
plant	organism	organism, being
park	plant	tract, piece of land
garden	park	vegetation
region	park	location
safari	garden	expedition, travel
environment	garden	geographical area
vegetation	plant	collection,aggregation

5 CONCLUSIONS

This paper describes an approach to discover hypernyms. The use of the related information in web queries seems a good approximation for narrowing the search results. This kind of queries is the most concrete and indicates that 1) there is a relation between terms and 2) the terms and their hypernym are in the same context. The method can be applied to any domain knowledge. WordNet seems to be limited because it does not nouns with more than one term and it only includes some proper nouns. The obtained results can be improved resolving ambiguous terms. Adding new lexical patterns to queries and extending the search to Frequently Questions Blogs and Wikipedia are good options to explore. The created taxonomies are consistent with the input corpus. This makes possible that taxonomies can be used on applications where the structure of corpus content is crucial. Finally, in the futher work will be considered additional experimentation and comparison with other state of art approaches.

REFERENCES

Blohm, S. and Cimiano, P. (2007). Learning Patterns from the Web-Evaluating the Evaluation Functions-Extended Abstract. *OTT'06*, 1:101.

Burgun, A. and Bodenreider, O. (2001). Aspects of the taxonomic relation in the biomedical domain. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, page 233. ACM.

Caraballo, S. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 120–126. Association for Computational Linguistics.

Cederberg, S. and Widdows, D. (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, page 118. Association for Computational Linguistics.

Cicurel, L., Bloehdorn, S., and Cimiano, P. (2007). Clustering of polysemic words. *Advances in Data Analysis*, pages 595–602.

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

Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Maedche, A. and Staab, S. (2001). Ontology learning for the semantic web. *Intelligent Systems, IEEE*, 16(2):72–79.

Ortega-Mendoza, R., Villaseñor-Pineda, L., and y Gómez, M. M. (2007). Using lexical patterns for extracting hyponyms from the web. *MICAI 2007: Advances in Artificial Intelligence*, pages 904–911.

Pantel, P. (2003). *Clustering by committee*. PhD thesis, University of Alberta.

Pantel, P. and Lin, D. (2002). Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619. ACM.

Pantel, P., Ravichandran, D., and Hovy, E. (2004). Towards terascale knowledge acquisition. In *Proceedings of the 20th international conference on Computational Linguistics*, page 771. Association for Computational Linguistics.

Ritter, A., Soderland, S., and Etzioni, O. (2009). What is this, anyway: Automatic hypernym discovery. In *Proceedings of AAAI-09 Spring Symposium on Learning by Reading and Learning to Read*, pages 88–93.

Sánchez, D. (2009). Domain ontology learning from the web. *The Knowledge Engineering Review*, 24(04):413–413.

Sang, E. (2007). Extracting hypernym pairs from the web. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 165–168. Association for Computational Linguistics.

Schutz, A. and Buitelaar, P. (2005). Relext: A tool for relation extraction from text in ontology extension. In Gil, Y., Motta, E., Benjamins, V., and Musen, M., editors, *The Semantic Web ISWC 2005*, volume 3729 of *Lecture Notes in Computer Science*, pages 593–606. Springer Berlin / Heidelberg.

Snow, R., Jurafsky, D., and Ng, A. (2005). Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems*, 17:1297–1304.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *J. Artif. Int. Res.*, 37:141–188.