ISSN 1870-4069

Proposal for Automatic Extraction of Taxonomic Relations in Domain Corpus

Hugo Raziel Lasserre Chavez, Mireya Tovar Vidal

Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación
{Hugoraziel,mireyatovar}@gmail.com

Abstract. This paper addresses the study in the Natural Language Processing area focused on the analysis and automatic extraction of taxonomic relations in domain corpus. In addition, a methodology is proposed emphasizing the automatic extraction by patterns and methods of grouping as the formal concepts analysis. This work considers the validation of the proposal through the information provided by task #13 of SemEval 2016 and by manual validation by domain experts.

Keywords: Hyponym, Hyperonym, Taxonomic extraction, taxonomy automatic extraction.

1 Introduction

Today many natural language processing applications make use of thesaurus, word lists or taxonomically classified terms used to represent concepts, such as the WordNet system [5], which serves as a dictionary of lexical knowledge for processing of the semantics of words and documents.

Building such taxonomies can be a difficult and extremely slow task. Thus, there has been a growing interest in finding methods that can learn taxonomic relationships and construct semantic hierarchies automatically [15].

Concept hierarchies (taxonomies) are important because they enable the structuring of information by categories. Relationships of "Is-a" type are an important problem in the construction of taxonomies, so the automatic acquisition of such relationships can be used to construct a taxonomy and even an ontology.

According to the Oxford dictionaries a hyponym is a word of more specific meaning than a general or superordinate term applicable to it, contrasted with hypernym which is a word with a broad meaning constituting a category into which words with more specific meanings fall. A noun is a word used to identify any of class of people, places or things, or to name a particular one of these.

This research has as its origin the task #13 of SemEval-2016 called Taxonomy Extraction Evaluation (TexEval-2) [3]. This task provides a corpus that is Wikipedia and is divided into four subtasks: Construction of taxonomy, Identification of hypernym, Multilingual taxonomy construction, Multilingual taxonomy identification.

pp. 29–39; rec. 2017-03-13; acc. 2017-05-25

Specifically, this paper addresses the first two subtasks focusing first on the identification of hyponymy through two approaches to be addressed in the following sections, identification by patters and formal concepts analysis.

The content of this paper is structured as follows, section 2 discusses the relation work of extracting taxonomic relations through patterns, formal concepts analysis among others. Section 3 mentions the proposed methodology divided into three phases, section 4 details the conclusions of this work.

2 Related Work

There is previous work in which lexical-syntactic patterns are used to identify and evaluate taxonomic and non-taxonomic relationships in domain corpus [17],[20]. As well as the identification of ontological relationships using formal concepts analysis [18], [19] and [20].

Some authors such as Pachenko et. al. [9] use a methodology based on dictionaries and use DBPedia and Wikipedia as corpus, as well as lexical resources such as Wordnet for extraction of taxonomic relationships. On the other hand, [8] uses a pattern-based technique, applying it to unstructured text and using the Wordnet lexical resource in the same way. While [12] used a standards-based methodology, in which it does not specify its corpus, but the Hearts patterns in table 1 are applied.

Table 1. Heart's lexical patterns

Pattern
A, and other B
A, or other B
A is a B
B, such as A
B, including A
B, especially A
B, particularly A
B, for example A
B, among which A

In Maitra et al [7] made a system called JUNLP which is based on two modules of detection of hyperonyms, the first deals with the semantic relations that can be found for a term and use BabelNet for the extraction of relations of hyponymy. BabelNet is a semantic network that connects concepts and named entities with a large network of semantic relationships. The second module attempts to identify subtrees present in the list of terms that may be a possible hyperonym for that term.

Pachenko et al [10] created a system called TAXI (Taxonomy Induction) for the extraction of taxonomic relations. His methodology is based on two sources of evidence, substring matches and Hearts patterns. They analyze all Wikipedia in search of the Hearts patterns and extract those relationships and make use of another corpus like GigaWord, ukWac and CommonCrawl.

Pocostales Joe [11] create a semi-supervised system called NUIG-UNL that finds hyperonym candidates for nouns by representing them as distribution vectors. This method assumes that hyperonyms can be induced by adding a compensation vector to the corresponding hyponym generated by GloVe. The vector is obtained as the average of the compensation between 200 pairs of hyponym / hyperonyms in the same space of the vector.

According to Tan et al.[16] frequently multi-word hyponym are constructions that contain another word that functions in the same way as a part of the same word. For example, "Apple pie" is essentially a "pie". This system explored the number of terms that are the same way. (multi-word) in English.

Cleuziou et al. [4] created a semi-supervised method for the acquisition of lexical taxonomies based on genetic algorithms. It is based on pre-topology theory that provides a powerful formal model of semantic relationships and transforms a list of terms into a space of structured terms in combination with different criteria of discrimination. In particular, rare but precise pieces of knowledge are used to parameterize the different criteria by defining the pre-topological term space. A structural algorithm is used to transform the pre-topological space into a lexical taxonomy.

Unlike the previously described works, the contribution of the methodology proposed in this paper will provide a list of syntactic lexical patterns that we consider can be used in different domains.

3 Proposed Methodology

The proposed methodology to be followed consists of three sequential phases which are described below and are shown graphically in Fig 1.

1. Information preprocessing: Development of an information preprocessing system that allows the processing of the corpus Wikipedia, in this phase it is also considered the preprocessing of the corpus, remove punctuation, special symbols and convert all words to lowercase.
2. Proposed models: Development and implementation of models that allow the extraction of hyponym / hyperonym taxonomic relations in domain corpus through patterns and formal concepts analysis.
3. Evaluation: The evaluation measures used in this phase are: precision and recall, in addition to manual validation by domain experts.

3.1 Phase 1- Information Preprocessing

For the first phase it is necessary to process Wikipedia in such a way that it can be understood by a computer, which is why we will use "Wikipedia Extractor" developed by Attardi [1], which is a tool that generates a flat text output of the whole database stored on Wikipedia.org. In this case each element stored in Wikipedia like the term "Benemérita Autonomous University of Puebla" in Wikipedia has as URL: https://en.wikipedia.org/wiki/Benem%C3%A9rita_Univesidad_Aut%C3%B3noma_de_Puebla . Its contents are taken as a document which is stored in XML format.

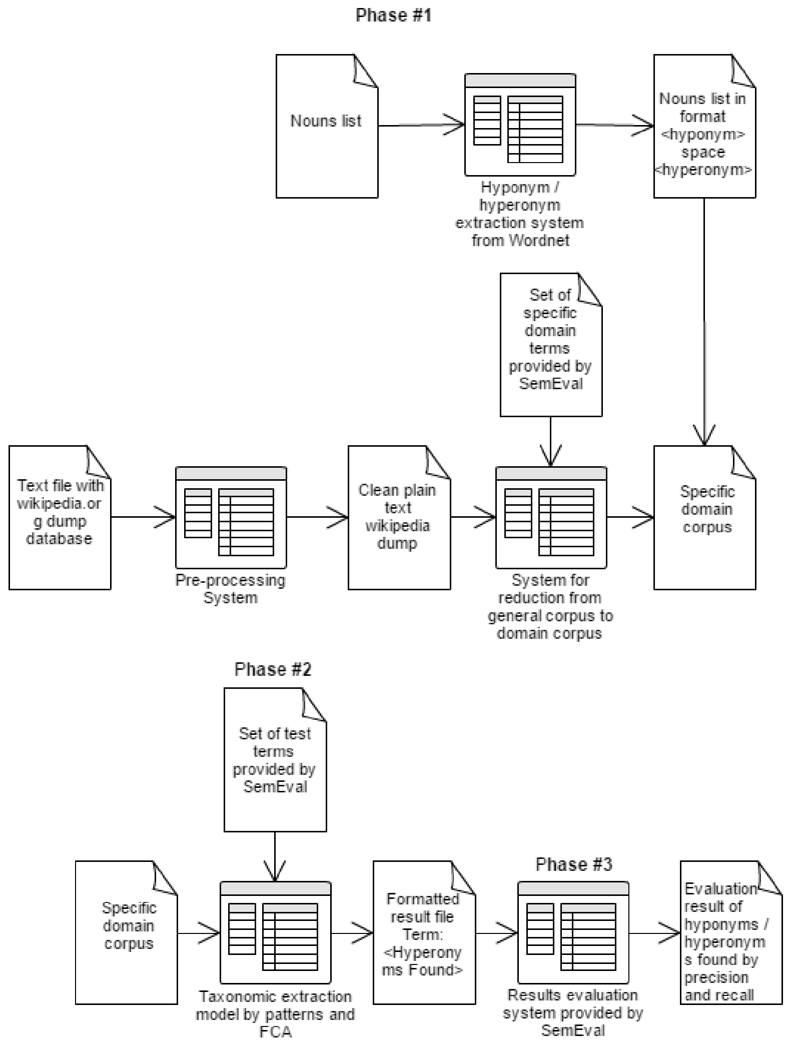
Phase #1
Nouns list
Hyponym / hyperonym extraction system from Wordnet
Nouns list in format <hyponym> space <hyperonym>
Set of specific domain terms provided by SemEval
Text file with wikipedia.org dump database
Pre-processing System
Clean plain text wikipedia dump
System for reduction from general corpus to domain corpus
Specific domain corpus
Phase #2
Set of test terms provided by SemEval
Specific domain corpus
Taxonomic extraction model by patterns and FCA
Formatted result file Term: <Hyperonyms Found>
Phase #3
Results evaluation system provided by SemEval
Evaluation result of hyponyms / hyperonyms found by precision and recall

Fig. 1. General System Methodology

The steps to follow by the information preprocessing system are as follows:

1. All corpus words will be converted to lowercase.
2. Punctuation, special characters and numbers will be removed from all documents.
3. The output will be a clean corpus.

The information retrieval system will perform these steps.

4. As input it will receive the list of nouns proposed by SemEval for a specific domain and the corpus obtained in step 3.
5. For the generation of the domain sub-corpus we will only consider the Wikipedia articles in which the nouns proposed by SemEval appear.

6. The system will return the reduced corpus for the domain of the lists of nouns that had as input.

3.2 Phase 2 – Proposed Models

In this phase, we consider two approaches for the extraction and identification of hyperonymy and hyponymy type relations: an approach based on lexical-syntactic patterns and grouping of entitites by FCA (Formal Concepts Analysis). These approaches are presented below.

3.2.1 Lexical-Syntactic Patterns

We propose a model for the extraction of this type of relations using patterns, these patterns will also be extracted automatically but filtered manually and the methodology would be the following.

1. Obtain the list of nouns of SemEval for the domain to be processed.
2. Having a list of all the nouns of the domain to be processed, request to a lexical knowledge base like Wordnet the hyperonym of each noun, with the tuple obtained of type: <extracted noun>:<hyperonym identified by Wordnet>
3. It is proposed to make a limitation of the domain corpus to documents in which the extracted noun and the hyperonym appears identified at a maximum distance of "K" words.
4. With regular expressions, it is proposed the extraction of everything that is between the extracted noun and the identified hyperonym, always remembering that what is in the middle of both should be of a window width of size "K" and this would become a candidate pattern.
5. To reduce margin of error of the patterns that may not always be correct, we will take in consideration only those candidate patterns that have more frequency in the documents and those will be validated by an expert.

An example is the following: In the list of nouns extracted from the whole subcorpus we get the word "Lion", we request WordNet its hyperonym and WordNet gives us that it is "Animal" so now we look in the corpus where the words "Lion" and "Animal" appear. We get the sentence "Lion is an Animal", we extract "is an" and it becomes a candidate pattern. If this pattern is repeated in several relationships, it is taken as an candidate pattern for the extraction of taxonomic relations, we take all patterns that repeat at least two times as a candidate pattern for a subsequent expert manual validation.

3.2.2 FCA – Formal Concepts Analysis

Formal concepts analysis (FCA) is a method of data analysis that describes the relationships between a particular set of objects and a particular set of attributes [2].

It allows data analysis methods to formally represent knowledge and produces two types of outputs for an input data. A reticle of concepts and a collection of attributes implications. The lattice is a formal collection of input data concepts order hierarchically by a sub-concept – super-concept relationship. The implication attribute describes a valid dependence of the data [18]. From a philosophical point of view a concept is the unit of thoughts composed of two parts, extension and intension, the extension covers all the objects or entities belonging to the concept, whereas the intention encompasses all attributes or properties valid for all those objects [21].

Formal concepts analysis provides a methodology for deriving a hierarchy of concepts (such an ontology) from a collection of objects and the properties they verify.

Each concept of the hierarchy obtained simulates a set of objects that share the same values for a certain set of properties or attributes [14].

For the development of this approach we propose a model that allows the construction of the hierarchy of concepts in which the collection of objects defined in the previous paragraph will be the data provided by the SemEval that is, the terms of plants and vehicles domains.

As mentioned above, each concept must meet a set of properties of attributes and with this information we can conclude that concepts share the same sets of attributes or properties and are manifesting some kind of taxonomic relationship.

Some attributes or characteristics proposed are the following: Verb and subject, Verb and object, Verbs, Prepositions, Placements, Nouns and adjectives, Punctuation marks.

3.3 Results Evaluation

Once the results of the application of the proposed models for the automatic extraction of hyponym / hyperonym relations have been obtained, it is necessary to evaluate them. As mentioned before, the evaluation is with the precision and recall measures [6]. In addition to validation by human experts.

4 Partial Results

SemEval-2016 provides test data for the evaluation of the proposal to be made by the participants. It provides concepts that deal with taxonomies already made, for example concepts of a manual taxonomy of vehicles, concepts of vehicles obtained from WordNet, concepts of plants of a manual taxonomy and concepts of plants obtained from Eurovoc and WordNet.

Since the Wikipedia database has a size of 53 GB, it is necessary to limit that corpus by means of an information retrieval system. It was decided to limit it to only the documents in which the terms that task # 13 that the SemEval-2016 provided appear and to obtain two specific domain corpus based on Wikipedia one containing the terminology of the vehicles domain and another one of the domain of plants. Both are described in table 2. On the left side we have the SemEval input nouns, total corpus lines and the total words in the plant corpus, on the right side we have the exact same parameters for the vehicle corpus.

Table 2. Properties of the two sub-corpus

Plants Corpus		Vehicle Corpus	
SemEval Input Nouns	513	SemEval Input Nouns	95
Total Corpus Lines	5000000	Total Corpus Lines	1781497
Total Corpus Words	131129942	Total Corpus Words	188093785

Table 3. Examples of hyponims / hyperonims of the SemEval proposed nouns

Plants Corpus		Vehicle Corpus	
Hyponym	Hypernym	Hyponym	Hypernym
senna	shrub	water scooter	motorboat
eelgrass	water plant	coach	car
guava	edible fruit	rocket	vehicle
eelgrass	aquatic plant	chariot	transport

Table 4. The 14 first patterns obtained by processing the plants and vehicle corpus

Plants Corpus		Vehicle Corpus	
Pattern	Frequency	Pattern	Frequency
to	1906	the	1434
and	1323	and	1493
or	563	a	665
at the	491	or	530
of	444	to	356
the	411	s	139
in the	361	by	119
a	281	and a	73
is a	244	was	68
is	173	on a	68
at	159	on the	64
and other	150	is a	57
like	130	of the	55
is a species of	84	and the	51

Table 5. Intersection of previous results with the lower frequency and some examples

Patterns Intersection		Vehicle Examples	Plant Examples
and	1295	hydrogen fuel cell-powered concept **car and sport utility** vehicle	It makes a good container **plant and ornamental tree**.
or	530	A limousine, executive **car or sport utility** vehicle is usually selected.	Some rooms were used as kitchens or pantries due to the fact of the large number of animal bones found inside, other room was used to store liquids (oil, wine or honey) in big containers or dolia and other rooms were used to store **grain or cereal** in pieces of pottery
the	411	Two would no longer be able to **lift the rocket** to launch altitude.	Lemon basil is **the** only basil used much in Indonesian cuisine
to	356	The princess had him come into the **coach to drive** back	A report by General Robert E. Lee on August 22, 1864, stated that **corn to feed** the Southern soldiers was exhausted.
is a	57	A **rocket is a pyrotechnic firework** made out of a paper	The African yam **bean is a legume** that is rich in protein and starch and an important source of calcium and amino acids.
by	40	heavy goods vehicles, and **public transport by coach** and bus	It contains the single species Eastwoodia elegans, a flower known **by** the common name yellow mock aster or yellow aster.
‘s	38	Because the rocket**'s** engine could withstand high heat	The plant**'s** flowers and fruits get set in about 10 to 11 months time followed by a maturity period of about 7–8 months and then harvested in about 18 months.
and the	35	The Inyo, as well as the express **car and the passenger car**, originally served the Virginia and Truckee Railroad in Nevada.	Predominating plants include the Moriche **Palm and the tree** "Caraipa llanorum". The dominant vegetation on the non-flooded savannas is grass.
on the	31	It is the range of 89-93% of mean state of charge which means as the **blades on the flywheel** turn, energy is being stored up between 89-93% of the given output.	Some family members use also an **oak leaf on the tree** trunk.

In the second phase of the first approach proposed for the extraction of patterns, the first step is to obtain hyponyms / hypernyms from the list of nouns proposed by semeval. Table 3 illustrates some of the results obtained by wordnet, these are extracted to have a base of patterns.

In table 4 we can observe the results obtained when executing the algorithm of the first approach. In column one and two the results are shown with the nouns of the plant domain and the remaining columns correspond to the results obtained for the domain of vehicles.

We decided to make an intersection of the previous results to get some generality of the patterns. In table 5 we show some examples of the extracted relations with some of the previous obtained patterns including the context of the full sentence, the first column is the pattern, second one is the lower frequency on both corpus, 3rd and 4th columns are vehicle examples and plant examples respectively.

The evaluation measure used is: Presicion, P = [Correct ISA] / [Sample]. Where sample is the input domain that can be plants or vehicles. In the case of the plants domain the total nouns with its hyponym / hyperonym relationship is 513, and we made a quick review of the reduced corpus and find out that we have 325 relationships found in the subcorpus so thats 58.03% of accuracy relationships found.

These are partial results as we need to make a test with the extracted general patterns to find out more relationships not limited by our first Wordnet output and check its validity with a domain expert.

5 Conclusions

As we can observe the results obtained as far as the patterns are very similar, we can find the same patterns with the execution of the first methodology in a corpus as well as in the other. It is planned to make a union of both results to join all the patterns that may be different and the patterns that are repeated add the number that were repeated in each corpus to take them to the top positions in the list of candidate patterns.

We can see that in the list there are already validated patterns by authors, such as the Hearst patterns, with finding them in the first positions we can intuit that the proposed methodology is valid and works also this work may be considered as a complement for the Hearst patterns previously described in the related work.

The Web has profoundly changed the way we communicate, do business, and do our work. You have access to millions of resources in different languages regardless of where we are today.

The problem of the Web is that the content / resources that we can find grows faster than we can classify it, thanks to the semantics in the Web, the software is able to process its content, reason with it, combine it and make logical deductions To solve everyday problems automatically.

This research proposes a system that can detect and extract hyponymic / hyperonymic taxonomic relations in flat texts (non-structurally) automatically. By ordering this hierarchical taxonomy (hierarchical), it would be classified semantically and could facilitate its access and understanding by other computer systems that require this type of information that is structured semantically.

Acknowledgements. This research work has been partially supported by VIEP-BUAP project 00478, CONACyT project 257357 and PRODEP-SEP ID 00570 (EXB-792) DSA/103.5/15/10854.

References

1. Attardi, G.: Wikipedia Extractor: A tool for extracting plain text from Wikipedia dumps. Dipartimento di Informatica Università di Pisa (2015)
2. Belohl´avek, R.: Introduction to formal context analysis. Dept of Computer Science. Palac´k y University, Olomouk., Tech. Rep. (2008)
3. Bordea, G., Lefever, E., Buitelaar, P.: Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, pp. 1081–1091 (2016)
4. Cleuziou, G., Moreno, J.: Qassit at semeval-2016 task 13: On the integration of semantic vectors in pretopological spaces for lexical taxonomy acquisition. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, pp. 1315–1319 (2016)
5. Fellbaum, C.: WordNet and wordnets. In: Brown, Keith et al. (eds.). Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier (2005)
6. Kowalski, G.: Information retrieval architecture and algorithms. Computer science, Springer-Verlag New York, Inc. (2010)
7. Maitra, P., Das, D.: Junlp at semeval-2016 task 13: A language independent approach for hypernym identification. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, pp. 1310–1314 (2016)
8. Ortega, R.: Descubrimiento Automático de Hipónimos a partir de Texto no Estructurado. Master thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica (2007)
9. Panchenko, A., Adeykin, S., Romanov, A., Romanov, P.: Extraction of Semantic Relations between Concepts with KNN Algorithms on Wikipedia. In Proceedings of Concept Discovery in Unstructured Data Workshop (CDUD) of International Conference On Formal Concept Analysis, pp 78–88 (2012)
10. Panchenko, A., Faralli, S., Ruppert, E., Remus, S., Naets, H., Fairon, V, Paolo, S., Biemann, C.: Taxi at semeval-2016 task 13: A taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, pp. 1320–1327 (2016)
11. Pocostales, J.: Nuig-unlp at semeval-2016 task 13: A simple word embedding-based approach for taxonomy extraction. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, pp. 1298–1302 (2016)
12. Rios, A., Lopez, I., Sosa, V.: Learning concept hierarchies from textual resources for ontologies construction. Expert Systems with Applications 40(15), 5907–5715 (2013)
13. Ritter, A., Soderland, S., Etzioni, O.: What Is This, Anyway: Automatic Hypernym Discovery. In: AAAI Spring Symposium: Learning by Reading and Learning to Read, pp. 88–93 (2009)

14. Sancho, F.: Análisis Formal de Conceptos. Dpto. de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Sevilla. http://www.cs.us.es/~fsancho/?e=78
15. Snow, R., Jurafsky, D., Ng, A. Y.: Learning Syntactic Patterns for Automatic Hypernym Discovery. In: Proceedings of the 17th International Conference on Neural Information Processing Systems, MIT Press, pp. 1297–1304 (2005)
16. Tan, L., Bond, F., Van, Genabith J.: Usaar at semeval-2016 task 13: Hyponym endocentricity. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, pp. 1303–1309 (2016)
17. Tovar, M., Pinto, D., Montes, A., Serna, G., Vilariño D.: Evaluación de relaciones ontológicas en corpora de dominio restringido. Computación y Sistemas 19, 135–149 (2015)
18. Tovar, M., Pinto, D., Montes, A., González, G., Vilariño, D.: Identification of ontological relations using formal concept analysis. In: Proceedings of the Ninth Latin American Workshop on Logic/Languages, Algorithms and New Methods of Reasoning, pp. 1–9 (2014)
19. Tovar, M., Pinto, D., Montes, A., González, G., Vilariño, D.: Identification of ontological relations using formal concept analysis. Engineering Letters 23(2), 72–76 (2015)
20. Tovar, M., Pinto, D., Montes, A., Serna, G., Vilariño, D.: Patterns Used to Identify Relations in Corpus Using Formal Concept Analysis. In: Carrasco, J., Martínez, J., Sossa, J., Olvera, J., Famili F. (eds) Pattern Recognition. (MCPR 2015). Lecture Notes in Computer Science 9116, pp. 236–245 (2015)
21. Wolf, E. K.: A first course in formal concept analysis. In: Faulbaum F., SoftStat'93 Advances in Statistical Software 4, pp. 429–438 (1993)