Ontology Population Using Corpus Statistics

Rogelio Nazar, Irene Renau
Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso
{rogelio.nazar,irene.renau}@ucv.cl

Abstract

This paper presents a combination of algorithms for automatic ontology building based mainly on lexical co-occurrence statistics. We populate an ontology with hypernymy links, thus we refer more specifically to a taxonomy of lexical units (nouns organized by hypernymy relations) rather than an ontology of formally defined concepts. A set of combined statistical procedures produce fragments of taxonomies from corpora that are later integrated into a unified taxonomy by a central algorithm. Our results show that with an ensemble of different components it is possible to achieve an accuracy only slightly worse than human performance. Finally, as our methods are based on quantitative linguistics, the algorithm we propose is not language specific. The language used for the experiments is, however, Spanish.

1 Introduction

The study of the vocabulary in its real context of use is currently a central part of linguistics (Kilgarriff 2007; Hanks 2013). Among other tasks in this discipline, it is of utmost importance to extract and organize vocabulary units from corpora. There is an intrinsic theoretical interest in such attempt, like the study of the laws that govern how words can be combined and classified. But there is also a practical motivation: to have the ability of transforming unstructured data into structured databases, i.e., to go from plain text to lexical databases, which can be later organized as an ontology which specifies the terminology of a domain and the conceptual relations between terms.

This paper presents a preliminary description and assessment of results of a methodology based on co-occurrence statistics to transform text into a knowledge structure, which can be later developed into a taxonomy or an ontology. The objective is to populate with lexical units the CPA Ontology (http://www.pdev.org.uk/#onto), handcrafted by Pustejovsky et al. (2004) and substantially modified later by Hanks (in process). Given a top-node ontology of around 200 lexical units (nouns) denoting the most general concepts of the language, the proposed method consists of populating this shallow ontology by means of corpus statistics. Hence, the objective is to link a noun such as "bicycle" with its hypernym, "Vehicle", and this one with "Artifact", and so on.

For a more precise definition of the terms, taxonomies and ontologies are different kinds of knowledge structures. Whereas an ontology is "a system of categories accounting for a certain vision of the world" (Maedche 1995, 11), a taxonomy can be considered a hierarchical relational structure of words. For instance, "Vehicle" can be a formally defined concept in an ontology, and "vehicle", "car" or "bicycle" words related to this concept, the first being a hypernym of the others. A hypernymy relation is a basic semantic relation between a word, the hyponym, and the word used as a descriptor to define it, the hypernym (Lyons 1977). Hypernymy provides the hierarchical structure for conceptual organization of a domain. In the following pages, we will use the term 'ontology' to refer to the most general nodes of the structure, and 'taxonomy' to refer to the connection between concepts of the ontology and words, establishing thus a difference between the ontological and the linguistic point of view.

In different ways, the paper represents an innovative way of addressing the problem. Our method is based on a combination of five different algorithms which produce raw results from corpora in the form of fragments of taxonomies, which are later compared and integrated into a single structure by a central algorithm in charge with the decision making process. The result is a tree of hyponym/hypernym relations between nouns, i.e. words rather than concepts and their formal definitions, characteristic of the linguistic view. Another novelty of the approach is that it is quantitative, thus it does not involve language or domain specific knowledge coded directly into the system. No external resources are needed apart from the analyzed corpora, a Part-of-Speech tagger and the CPA Ontology itself, which does not change because it uses English as a metalanguage. Up to now, however, experiments have been carried out only in English and more extensively in Spanish. We are starting with French and expect to continue replicating the experiment in other languages and offering the results in the accompanying website (http://www.verbario.com).

In the following sections we present a general overview of the related work and then we describe our proposal. We offer an evaluation of the results and, finally, we draft some conclusions and plans for future work.

2 Related Work on Taxonomy Building

The interest for the development of taxonomies is of course not new, as the publications on the subject span for four decades. Space limitations only allow us to offer a very brief account of the research in this field, which we organize as follows: first, some efforts to produce taxonomies by hand. Then, the literature on automatic taxonomy building from machine readable dictionaries. Finally, taxonomy extraction from corpora, on the one hand by rule-based systems and, on the other, based on quantitative analysis.

2.1 Handmade Taxonomies

There is a large body of work in handcrafted taxonomy creation. We will only focus on some of the most representative modern efforts, not 3rd century Porphyrian tree or Roget's Thesaurus. We also exclude from this account all the specialized ontologies, restricting ourselves to some of the most well-known projects devoted to general vocabulary. Among the most cited projects are FrameNet, Cyc, WordNet and the CPA Ontology. With the exception of WordNet, that also includes a Spanish version, the rest are only available in English[1].

FrameNet is aimed at the implementation of Charles Fillmore's (1976) frame semantics as a lexical database organized in conceptual structures, available at http://framenet.icsi.berkeley.edu/. The Cyc ontology, in turn, was born in 1984 not in the context of a linguistic theory but in the field of Artificial Intelligence (Lenat 1995). It is defined as an ontology of everyday common sense knowledge and is available at http://www.opencyc.org/. Another large taxonomy is WordNet (Miller 1995; Vossen 1998), originally created by psychologists but then widely used in many natural language processing tasks. WordNet, avialable at https://wordnet.princeton.edu/, is based on 'synsets', defined as sets of words that have the same sense or refer to the same concept. It can be considered a taxonomy because it includes hypernymy links. Finally, the other project that has come to our attention is the CPA Ontology, created in the context of a lexicography project (Hanks in progress). It is at the moment a shallow ontology including only the upper nodes, i.e. the most general concepts denoted by words called "semantic types" in CPA terminology: "Event", "Emotion", "Physical Object" or "Human", etc. This includes no more than 200 words hierarchically organized in hypernymy links. This top-node structure is currently being populated with lexical items by Hanks and his team. It is handmade work but built from corpus analysis, which means that categories are not assumed a priori.

An examination of these taxonomies reveals different limitations. FrameNet departs from our main interest because strictly speaking it cannot be considered a taxonomy. In the case of the Cyc ontology, the formalisms used to express the relations are too complex to be manipulated and used as a basis for this Spanish taxonomy. In the case of WordNet, its general architecture based on synsets can often be problematic because at times the words in a synset are too different from a semantic point of view. For instance, the case of the Spanish synset containing the words *animal*, *bestia*, *criatura* and *fauna*, which is equivalent to the English synset containing 'animal', 'animate_being', 'beast', 'brute', 'creature' and 'fauna'. Here, the Spanish word *pez* and its English equivalent 'fish' are correctly placed as hyponyms of 'animal', but not as hyponyms of 'beast'. Furthermore, WordNet is a top-down approach, while our interest is on the corpus-driven approach.

Overall, we decided in favor of the CPA Ontology for our taxonomy population project because its architecture, based on lexical units rather than synsets, is simple enough to be manipulated as needed.

2.2 Taxonomy Induction from Machine Readable Dictionaries

The field of automatic semantic relation extraction and, in particular, hypernymy extraction, began to develop soon after the publication of the first machine readable dictionaries in the seventies and eighties. This new resource favored the development of different methodologies to transform dictionaries made for human users into a lexical database with information stored and organized for computers (Calzolari, Pecchia, and Zampolli 1973; Calzolari 1977; Amsler 1981; Chodorow, Byrd, and Heidorn 1985; Alshawi 1989; Fox et al. 1988; Nakamura and Nagao 1988; Wilks et al. 1989; Guthrie et al. 1990; Boguraev 1991; Araujo and Pérez-Agüera 2006).

The first researchers shared the idea of taking a machine readable dictionary and study the regularities and patterns in the definitions and subsequently write a system of rules that would allow to extract hypernymy and other semantic relations between vocabulary units. Depending on the dictionary, one of these rules could be that the first noun of the definition would be the hypernym of a defined noun. However, this is not always the case, and thus one needs to develop more rules to cope with the exceptions.

2.3 Taxonomy Induction from Corpora

The Pattern-based Approaches With the advent of corpus linguistics in the nineties, researchers interested in semantic relation extraction moved on to corpus analysis but keeping the same philosophy as in the previous attempts with dictionaries, that is, elaborating rule-based systems that would search for lexico-syntactic patterns in corpora expressing the desired information. Typically, if one finds in running text a sequence such as "X is a type of Y" or "X and other (types of) Y", etc., then one would assume that any pair of nouns occupying the positions X and Y would hold a hypernymy relation (Hearst 1992; Rydin 2002; Cimiano and Völker 2005; Snow, Jurafsky, and Ng 2006; Pantel and Pennacchiotti 2006; Potrich and Pianta 2008; Auger and Barriere 2008; Aussenac-Gilles and Jacques 2008, among others).

Of course, the problem with this approach is that not always the collected patterns express the desired relations

[1] There are, however, ongoing efforts to produce a Spanish version of FrameNet as well, cf. http://sfn.uab.es/

and, in addition, many times the desired relations appear expressed in patterns that the researchers were not able to anticipate.

The Quantitative Approaches A different view on the subject is the extraction of thesauri from corpora based on distributional similarity. There are two main lines of research, one that specializes in finding semantic similarity between groups of words and the other in establishing hypernymy links between pairs of words.

In the first case, the semantic similarity between groups of words is calculated on the basis of distributional similarity, as it is considered that semantically similar words will tend to occur in similar contexts. To be semantically similar, in this case, means to be synonyms or near-synonyms or, more interestingly, words that pertain to the same semantic class, i.e., cohyponyms (Grefenstette 1994; Landauer and Dumais 1997; Schütze and Pedersen 1997; Lin 1998; Ciaramita 2002; Biemann, Bordag, and Quasthoff 2003; Alfonseca and Manandhar 2002; Pekar, Krkoska, and Staab 2004; Bullinaria 2008). This line of research is tributary to the general notion of distributional semantics initiated by Harris (1954) and developed later by many others (Sahlgren 2008; Baroni and Lenci 2010; Nazar 2010).

The second trend goes a step further than the previous notion of distributional thesauri as just clusters of similar words, and emphasizes the importance of establishing a hierarchic organization of the vocabulary, a difficult task that imposes its own challenges. As in the previous case, the data is obtained from corpora defined as document collections, the Wikipedia or the Web, but the method used is most often directed co-occurrence graphs (Woon and Madnick 2009; Wang, Barnaghi, and Bargiela 2009; Navigli, Velardi, and Faralli 2011; Nazar, Vivaldi, and Wanner 2012; Fountain and Lapata 2012; Medelyan et al. 2013; Velardi, Faralli, and Navigli 2013).

2.4 Why a New Approach

After so many publications on the subject, there continue to be attempts on taxonomy extraction, because despite the variety of ideas already proposed there is still plenty of room for improvement. The large body of bibliography appears to indicate that the field has come to a point in which the integration of different ideas is needed, i.e., an algorithm able to integrate different fragments of taxonomies.

3 Methodology: an Integration of Algorithms

The novelty of our approach lies on the modular design. Modular algorithms produce small fragments of taxonomies which are later contrasted and integrated into a larger taxonomy by a central module.

Algorithm 1 computes distributional similarity between words. Algorithm 2 calculates asymmetric relations in word co-occurrence in corpora. Algorithm 3 analyzes definitions from various dictionaries of a language and detects cases of significant definiens-definiendum co-occurrence. Algorithm 4 is a variant of 1 because it computes distributional similarity as the number of identical $n$grams (as sequences of words) that a group of words may share. Algorithm 5 is an inference engine, which tries to reason upon the results of the other algorithms and extract new hypernymy assertions. Algorithm 6, finally, is the "assembly algorithm", which is in charge of integrating the taxonomy fragments produced by all the components into a modified version of the CPA Ontology.

Experimental evaluation shows that with this method it is possible to obtain a robust homeostatic or self-regulated taxonomy, because it is based on corpus statistics and can update itself automatically. Evidently, this list of methods is not exhaustive and, as this is work in progress, we foresee the integration of other methodologies as well. Up to now we have avoided matching Hearst patterns because they are costly to develop, they are language specific and, depending on the implementation, can also be error prone, as in the case of the Text2onto software, with precision figures of 17.38% and 29.95% recall for hypernymy extraction (Cimiano and Völker 2005). We have also avoided the use of explicit semantic or grammatical knowledge, preferring a design that is self-contained and not dependent on external resources like Hearst patterns or WordNet, because this facilitates replication in other languages.

As a textual corpus for our experiments we used a collection of Spanish press articles and Wikipedia pages accumulated on a single text file of ca. a billion tokens. In the case of algorithm 3, as it is fed with a lexicographic corpus, we used online dictionaries via a web search engine.

3.1 Algorithm 1: Clustering of Nouns Based on Distributional Similarity

The first component is based on a clustering technique that produces sets of semantically related nouns on the basis of distributional similarity. It bears some resemblance with the quantitative approach of Grefenstette (1994), although aimed at cohyponyms rather than synonyms, and without grammar-specific information.

Consider, for instance, the semantic class of drinks, with elements such as "coffee", "tea", "beer", "brandy", and so on. In the case of these nouns, there is a great probability that they will co-occur with other words such as the verb "to drink" or nouns such as "glass", "cup" or "bottle". These and other shared words are the ones we used as indicators of the nouns' semantic relatedness, without POS-tag distinction.

We can represent the overlap of shared vocabulary between lexical units as a Venn diagram (figure 1). In the intersection we can observe words that are shared by the units *cerveza* (beer), *café* (coffee) and *té* (tea), e.g. *servir* (to serve), *beber* (to drink), *tomar* (to drink), *querer* (to want), etc. Of course, we also have words that are shared only by two of the units, e.g., *café* and *té* share *caliente* (hot), which does not co-occur with *cerveza*. By the same token, *cerveza* and *café* share the unit *amargo/a* (bitter), which does not co-occur with *té*.

In concrete, this component analyzes the syntagmatic context in which a word appears and extracts the vocabulary (excluding function words). It then obtains pairs of words that, following the previous examples of drinks, could be *brandy francés, bebiendo brandy, tomar brandy, brandy barato*, etc. (French brandy, drinking brandy, drink brandy,

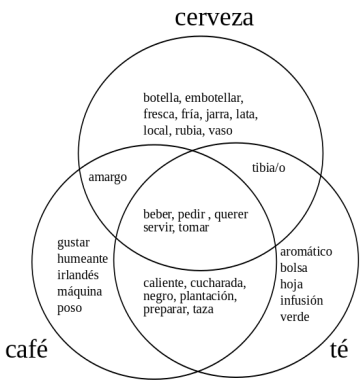

Figure 1: A Venn diagram to represent the intersection and difference between the co-occurrence sets.

cheap brandy, etc.). From these elements, a data structure is created in which each term is associated with the lexical units it co-occurs with.

Terms are then represented as co-occurrence vectors, and thus the algorithm conducts a pairwise comparison of the terms applying a similarity measure which calculates the degree of overlapping between vectors. In this case, this is calculated with the Jaccard coefficient, as suggested by Grefenstette (1994), defined as follows, where A and B are the two vectors to be compared:

$$J(A,B) = \frac{|A \cup B|}{|A \cap B|} \quad (1)$$

As it is usual in any clustering procedure, for this comparison we need a table of distances, from which we obtain a pair of units showing the greatest similarity. Hereafter, the members of this pair merge and create the first cluster, which occupies the place of both words and contains the sum of their attributes. The process is iterative, thus, another table of distances is created, but every time with one less element. This process stops when the units to cluster do not reach a similarity threshold, defined as a minimum proportion of attributes in common that a set of units must have in order to be assigned the same cluster.

Class	**Members of the cluster**
Vehicles	*carro, automóvil, coche, autobús, tranvía, carroza, carruaje, camión, jeep, camioneta*
Types of cheese	*brie, parmesano, camembert, mozzarella, gorgonzola, roquefort, gruyer*
Drinks	*chocolate, licor, chicha, cerveza, aguardiente*
Hats	*pavero, tricornio, bicornio, guarapón, canotier, calañés*
Animals	*venado, ciervo, tigre, elefante, perro, gato, puerco, cerdo, carnero, conejo, ratón, rata*

Table 1: Examples of clusters of Spanish nouns made by algorithm 1.

In order to obtain an estimation of the quality of the results produced by this single module, we manually evaluated an arbitrary selection of 145 nouns which can be classified as drinks, hats, vehicles, animals and types of cheese. Table 1 shows some examples of the clusters created from these nouns and their member elements. In this experiment, our algorithm was able to produce correct clusters in 96% of the cases, though it was only capable of classifying half of the input words (51%). Better precision was met at the expense of a considerable loss of recall. More details on this experiment will appear in Nazar & Renau (In press).

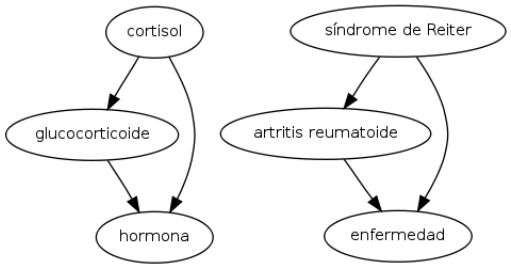

Figure 2: In algorithm 2, example of co-occurrence graphs depicting hypernymy relations. Hypernym nodes are the ones that have the largest number of incoming arrows.

3.2 Algorithm 2: Taxonomy Extraction Based on Asymmetric Word Co-occurrence

Instead of clusters of semantically-related words, as produced by the first algorithm, the second one consists of creating hyponym-hypernym pairs based on their co-occurrence patterning. Here, we define co-occurrence as a tendency of two lexical units to appear together in the same sentences, not taking into account the distance nor their order in the sentence. The main idea behind this study is that co-occurrence is asymmetric in the case of hyponym-hypernym pairs. For instance, the word *motocicleta* shows a tendency to appear in the same sentences with the word *vehículo*, but the relation is not reciprocated. The asymmetric nature of such association allows us to automatically represent hierarchical relations without resorting to external knowledge bases. The computation of these relations is produced with the help of directed graphs that express the co-occurrence relations.

The graphs shown in figure 2 illustrate this method. The arrows in the graph represent asymmetric co-occurrence relations, and the node with most incoming arrows is selected as the hypernym of the input term. Here, the input term *cortisol* tends to co-occur with *glucocoricoide* and *hormona*. In turn, *glucocoricoide* also tends to co-occur with *hormona*, but this last unit does not reciprocate the relation with neither of both. The output of the graph is read as saying that *hormona* is the hypernym of *cortisol* because it is the node with the largest number of incoming arrows. As shown by the other graph, the same pattern is exhibited by noun phrases, very common in multiword terms.

In order to test the performance of this module alone, we manually evaluated the results of an experiment with 200 Spanish nouns pertaining to the semantic classes of mammals, insects, drinks, hats, vehicles and, again, varieties of cheese. This preliminary evaluation shows that we can expect approximately a 60% chance of obtaining a correct hy-

pernym for a given noun using this algorithm in isolation. More details on this experiment can be found in Nazar & Renau (2012).

3.3 Algorithm 3: Extraction of Hypernymy Relations from Definiens-Definiendum Co-occurrence in General Dictionaries

As already mentioned in subsection 2.2, electronic dictionaries have been used in the past to extract hypernymy and other semantic relations, but in general the approach has been focused on a single dictionary, which is parsed with a rule-based system to extract the relations from the definitions. Our approach here is different because we use a set of dictionaries and infer the hypernymy relations by the frequency of co-occurrence between lexical items in the headword and in the definitions. The algorithm uses the frequency to select hypernyms from the text of the definitions, assuming that there will be some consensus among the dictionaries when selecting a given hypernym, and thus this should be the most frequent word (excluding function words). In this way we save the effort of building a set of rules for each dictionary and we make it possible to replicate the experiment in other languages.

In order to obtain an evaluation of the performance of this single module, we manually examined the results for a random sample of 150 nouns and concluded that we can expect approximately 70% chance of obtaining a correct hypernym for a given input word. More details on this experiment can be found in Renau & Nazar (2012).

3.4 Algorithm 4: Ngrams with "Asterisks"

With algorithm 4 we explored the possibility of creating clusters of words that have a tendency to occur in exactly the same positions in short sequences of words. This is why we describe this module "$n$grams with asterisks".

What we do here is to study large samples of $n$grams, defined as sequences of three to five words, and then replace one of the words inside the $n$gram with an asterisk. The goal is to record then which are the words that most frequently occur in the position of such asterisk. Normally, these words will show some kind of paradigmatic relation and therefore will have some features in common, such as the grammatical category and, in most cases, also a semantic relatedness. Consider, for instance, the case of the $n$gram 'at * airport', taken from the BNC corpus (table 2). Only a limited number of words can occur in the position of the asterisk, and these are semantically related.

*at * airport* [645]
Heathrow 56, Manchester 23, Gatwick 19, London 15, Frankfurt 15, Teesside 10, Edinburgh 8, an 7, Glasgow 7, Stansted 7, Dublin 6, Birmingham 5, Coventry 5, Aberdeen 5, ...

Table 2: Words appearing in the position of the asterisk in the sequence 'at * airport' in the BNC corpus.

To share a single $n$gram is of course not and indication of semantic similarity. But if there are words that show a tendency to appear in the same positions in a large number of different $n$grams, then one can conclude that these words are paradigmatically related. As a result, this algorithm produces clusters of words, where it can be seen that the members of each class share not only the same grammatical category but also an evident semantic relatedness. Table 3 shows some examples of the results in English. Results were virtually the same in Spanish. In this paper we are only interested in nouns, but as the table shows, the same procedure can be applied to the study of other grammatical categories.

POS-tag	**Members of the cluster**
Adverbs	*entirely, exclusively, mainly, primarily, principally*
Adjectives	*cost-effective, efficient, elaborate, professional, subtle*
Proper nouns	*Australia, Dublin, England, France, India, Janeiro, Middlesbrough, Newcastle, Sunderland, Yorkshire*
Nouns	*chess, cricket, football, golf, rugby, soccer, tennis*

Table 3: Examples of clusters in English, result of Algorithm 4.

In order to evaluate this component, we again carried out a manual examination of the results. This was done by a random sample of 30 clusters, containing 1191 words in total. We found that in 96% of the cases the clusters were consistent. This internal consistency is computed as the mean consistency of each individual cluster, with numbers for the individual clusters ranging from 80 to 100% consistency. More details on this experiment will appear in Nazar & Renau (Submitted).

3.5 Algorithm 5: Analogical Inference

We use algorithm 5 as an inference engine to analyze the results produced by the previous modules. It is the only component that is not corpus-driven, in the sense that it only analyzes the morphological and lexical features of the terms.

This component is described as an analogical inference engine because it learns to associate features of the lexical items with the category that is assigned to them by the other algorithms. In this way, if a term cannot be found in the analyzed corpora, it may still be classified by this module.

The features that are learned are both lexical and morphological. The lexical level is only useful in the case of multiword expressions, but of course these are also very frequent, especially in the case of technical or specialized domains. For instance, this algorithm first learns that the other ones are placing terms such as *síndrome de Carpenter* (Carpenter syndrome) and *síndrome de Meretoja* (Meretoja syndrome), among others, as hyponyms of *enfermedad* (disease). Thus, it learns to associate features of the terms (in the case, the sequence *síndrome de*) with a semantic class, and subsequently classify new terms such as *síndrome de Maffucci* (Maffucci syndrome) also as a disease. The same is done at the morphology level: the module learns to detect morphological similarities between the members of the same semantic class. Continuing with the same example, it can learn that very frequently the terms denoting diseases have the suffixes *-osis* or *-itis*, such as *hepatitis* or *endometriosis*. Given, thus, a new term such as *pancreatitis*, the module will recognize it

as a disease. More details on this experiment were published in Nazar et al. (2012).

3.6 Integration into a Single Taxonomy

After the implementation and experimentation with each procedure, we developed a new central or "assembly" algorithm, with the purpose of integrating the results into a single taxonomy. The task of this module is to reinforce the certainty of the results on the basis of the combined output of each module. The result is a sort of "consensus" taxonomy which, according to our preliminary experiments, is larger and more reliable than the ones produced by each module in isolation.

The integration procedure is however not straightforward, because each algorithm is of a different nature, and thus the combination of the results cannot be solved with a simple voting scheme. Algorithms 1 and 4 result in groups of semantically-similar words, while 2, 3 and 5 result in hypernymy pairs. Moreover, the desired result is to populate the already existing CPA Ontology, which, as already explained, refers to very general or abstract concepts.

Two basic operations are conducted. On the one hand, to integrate the results of modules that produce clusters of semantically similar words and, on the other hand, to link these clusters of words with a correct hypernym. This is done in sequential steps. For each noun in each cluster produced by modules 1 and 4, there (might) be a hypernym candidate provided by modules 2, 3 and 5. The result is that, for each cluster, there will be a most frequent hypernym candidate, which is thus selected as the semantic class of all the members of the cluster. As a result, a pairing of a hypernym with a group of hyponyms is obtained, e.g. *sedán*, *coche*, *limusina*, etc. are classified as hyponyms of *automóvil*. A chain of ascending hypernymy links is built until one of the semantic types of the CPA Ontology is found, and then each word is integrated in an hypernymy chain until the top node Entity:

Entity → Physical Object → Inanimate → Artifact → Machine → Vehicle → Automobile → sedan

3.7 User Interface

A first prototype is now being developed as a web demo, available at http://www.verbario.com. The taxonomy of nouns is only a module of Verbario.com, a website that is part of a wider project devoted to lexical analysis. The "Taxonomy" part shows two ways of obtaining results: the user can either introduce a noun and get the hypernymy chain or viceversa, he/she can obtain all the hyponyms of a given target noun. At the moment, more than 30,000 Spanish nouns have already been introduced and, while the system is running, more words are being added at a fast pace with a reasonably low error rate, as shown in the next section. We plan to offer regular back up files of the taxonomy in OWL format at this website.

4 Evaluation of the Results

Samples of the overall results were evaluated by a group of 8 human judges, all of them advanced graduate students in linguistics. Each one received the same instructions and a random sample of 52 nouns to evaluate. Criteria for considering a link between a word and a node as correct was that it should correspond to the hypernymy relation type. Among the words from the samples we have, for instance, *lechuga* ('lettuce'). In this case, the taxonomy offers two hypernymy chains:

Entity → Physical Object → Plant [Planta] → [Arbusto] → lechuga

Entity → Physical Object → Plant [Planta] → lechuga

For one, it states that it is a type of bush ([Arbusto]), which, in turn, is a type of plant, and so on. But a lettuce is not really a bush, thus this chain is incorrect (it cannot be analyzed as "lettuce IS A bush"). For the other, it asserts that a lettuce is a type of plant ([Planta]), which in turn is a type of physical object, and so on. This last chain is considered correct (it can be analyzed as "lettuce IS A plant").

For each noun, our human judges indicate how many hypernyms were offered by the taxonomy (two in the case of *lechuga*) and how many of them were correct: one out of two in this case. We thus calculated overall precision as the ratio of correct chains over total chains. We consider recall very difficult to calculate, because of the lack of corpus-based lexicographic material in Spanish. As a consequence, we only evaluated precision, and obtained that for a total of 763 hypernymy chains examined, 586 were found to be correct, which makes a precision of 76.80%. The standard deviation in the group of judges is 14.34. If we exclude the two judges with more extreme positive and negative scores, then mean precision rate is 77.18%, with a standard deviation of 12.82.

Regarding the control of inter-coder agreement, we included in all samples given to the judges a common group of 11 nouns, which makes 88 judgments that should ideally be identical. However, raters agreed only on 72 cases, which is still more than moderate agreement (81.8% or 63.2% if measured with a Kappa coefficient to correct for chance-related agreement). We can interpret the agreement percentage as the ceiling of the precision one can expect from this type of algorithms.

With respect to error analysis, we found out that they mostly occurred as a consequence of the polysemy of the words, a circumstance already noticed by Amsler (1981) for this kind of output. It is the case, for instance, of the word *adicción* (addiction), tagged as hyponym of *dependencia* (dependence), which is correct in principle. But then *dependencia* is only registered as a type of *construction*, according to one of the senses that the word has in Spanish.

Another frequent cause of error is confusion between hypernymy and other semantic relation. *Vitíligo*, for instance, is correctly classified as a disease in one case but incorrectly as a hyponym of *piel* (skin) in another, obviously because it is a skin-related disease and then both words tend to co-occur in the same contexts. The same happens in the case of synonyms, which are often placed incorrectly in a hypernymy relation. For instance, the word *cuchí* is a synonym and not a hyponym of *cerdo* (pig). There are also cases of meronymy, such as the word *océano* (ocean), which is

wrongly connected to *agua* (water): an ocean is made of water, but it is not a type of water.

5 Conclusions and Future Work

This paper has presented a set of combined algorithms for building a taxonomy of Spanish nouns based on procedures from quantitative linguistics. The method, based mainly on the study of co-occurrence patterns, could in principle be replicated with different languages. The precision we obtain at the moment can be improved but at the same time it is only slightly lower than the inter-coder agreement percentage. In semantic analysis, total agreement is unrealistic given the fact that even dictionaries not always agree.

For future work, we are focusing on the following aspects: on the one hand, we will try to improve precision by addressing the problem of polysemy and the confusion between synonyms and meronyms with hypernyms. We have already experimented with sense-induction algorithms which, for each noun found in a corpus, will produce a list of different senses (Nazar 2010). This algorithm can now be used to map each sense with a hypernym. Pending work also includes a detailed large scale evaluation. In this respect, we must distinguish precision of frequent nouns (e.g. *manzana* - apple, *casa* - house, etc.) from precision of very infrequent nouns (e.g. *acetábulo*, an anatomic part of a bone). Finally, we will evaluate separately how the system operates with specialized terms and with general language.

6 Acknowledgments

This paper has been made possible thanks to funding from Projects Fondecyt 11140704, lead by Irene Renau, and Fondecyt 11140686, lead by Rogelio Nazar (http://www.conicyt.cl/fondecyt). We would like to express our gratitude to the students for their participation and to the reviewers for their extended and detailed comments, which have been very helpful to improve this paper. Unfortunately, lack of time and space prevented us from introducing all of the changes that were suggested.

References

Alfonseca, E., and Manandhar, S. 2002. Extending a lexical ontology by a combination of distributional semantics signatures. In *Proc. of EKAW'02*, 1–7.

Alshawi, H. 1989. Analysing the dictionary definitions. In Boguraev, B., and Briscoe, T., eds., *Computational Lexicography for Natural Language Processing*. White Plains, NY, USA: Longman Publishing Group. 153–169.

Amsler, R. 1981. A taxonomy for English nouns and verbs. In *Proc. of 19th annual meeting on ACL (Morristown, NJ, USA)*, 133–138.

Araujo, L., and Pérez-Agüera, J. R. 2006. Enriching thesauri with hierarchical relationships by pattern matching in dictionaries. In *FinTAL*, 268–279.

Auger, A., and Barriere, C. 2008. Pattern-based approaches to semantic relation extraction - A state-of-the-art. *Terminology* 14(1):1–19.

Aussenac-Gilles, N., and Jacques, M.-P. 2008. Designing and evaluating patterns for relation acquisition from texts with Caméléon. *Terminology* 14(1):45–73.

Baroni, M., and Lenci, A. 2010. Distributional memory: A general framework for corpus-based semantics. *Comput. Linguist.* 36(4):673–721.

Biemann, C.; Bordag, S.; and Quasthoff, U. 2003. Lernen von paradigmatischen relationen auf iterierten kollokationen. In *Beiträge zum GermaNet-Workshop: Anwendungendes deutschen Wortnetzes in Theorie und Praxis*.

Boguraev, B. 1991. Building a Lexicon: The Contribution of Computers. *International Journal of Lexicography* 4(3):227–260.

Bullinaria, J. 2008. Semantic categorization using simple word co-occurrence statistics. In Baroni, M.; Evert, S.; and Lenci, A., eds., *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, 1–8.

Calzolari, N.; Pecchia, L.; and Zampolli, A. 1973. Working on the Italian machine dictionary: a semantic approach. In *Proc. of 5th Conference on Computational Linguistics (Morristown, NJ, USA)*, 49–52.

Calzolari, N. 1977. An empirical approach to circularity in dictionary definitions. *Cahiers de Lexicologie 31(2)* 118–128.

Chodorow, M.; Byrd, R.; and Heidorn, G. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proc. of the 23rd annual meeting on ACL (Chicago, Illinois, USA)*, 299–304.

Ciaramita, M. 2002. Boosting automatic lexical acquisition with morphological information. In *Proc. of the ACL-02 Workshop on Unsupervised Lexical Acquisition, ACL*, 17–25.

Cimiano, P., and Völker, J. 2005. Text2onto. In *Natural language processing and information systems*. Springer. 227–238.

Fillmore, C. J. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech* 280(1):20–32.

Fountain, T., and Lapata, M. 2012. Taxonomy induction using hierarchical random graphs. In *Proc. of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, 466–476. Stroudsburg, PA, USA: Association for Computational Linguistics.

Fox, E. A.; Nutter, J. T.; Ahlswede, T.; Evens, M.; and Markowitz, J. 1988. Building a large thesaurus for information retrieval. In *Proc. of the Second Conference on Applied Natural Language Processing*, ANLC '88, 101–108. Stroudsburg, PA, USA: Association for Computational Linguistics.

Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer, Dordrecht, The Netherlands.

Guthrie, L.; Slator, B.; Wilks, Y.; and Bruce, R. 1990. Is there content in empty heads? In *Proc. of the 13th International Conference on Computational Linguistics, COLING'90 (Helsinki, Finland)*, 138–143.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press, Cambridge, MA.

Hanks, P. in progress. Pattern Dictionary of English Verbs. http://www.pdev.org.uk/ (last access: 26/04/0215.

Harris, Z. 1954. Distributional structure. *Word* 10(23):146–162.

Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th International Conference on Computational Linguistics (Nantes, France)*, 539–545.

Kilgarriff, A. 2007. Googleology is bad science. *Comput. Linguist.* 33(1):147–151.

Landauer, T. K., and Dumais, S. T. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 211–240.

Lenat, D. B. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM* 38(11):33–38.

Lin, D. 1998. Automatic retrieval and clustering of similar words. In *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, 768–774. Stroudsburg, PA, USA: Association for Computational Linguistics.

Lyons, J. 1977. *Semantics*, volume 2. Cambridge University Press.

Maedche, A. 1995. *Ontology Learning For The Semantic Web*. Doredrecht, The Netherlands: Kluwer.

Medelyan, O.; Manion, S.; Broekstra, J.; Divoli, A.; Huang, A.; and Witten, I. H. 2013. Constructing a focused taxonomy from a document collection. In *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proc.*, 367–381.

Miller, G. A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38(11):39–41.

Nakamura, J., and Nagao, M. 1988. Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation. In *Proc. of the 12th International Conference on Computational Linguistics COLING-88 (Budapest, Hungary)*, 459–464.

Navigli, R.; Velardi, P.; and Faralli, S. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proc. of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, 1872–1877. AAAI Press.

Nazar, R., and Renau, I. 2012. A co-occurrence taxonomy from a general language corpus. In *Proc. of EURALEX*, 367–375.

Nazar, R., and Renau, I. In press. Agrupación semántica de sustantivos basada en similitud distribucional. implicaciones lexicográficas. In *Actas del V Congreso Internacional de Lexicografía Hispánica (25-27 de junio de 2012)*.

Nazar, R., and Renau, I. Submitted. Extraños-misteriosos-insondables-inescrutables son los caminos del señor: extracción de relaciones paradigmáticas mediante análisis estadístico de textos.

Nazar, R.; Vivaldi, J.; and Wanner, L. 2012. Co-occurrence graphs applied to taxonomy extraction in scientific and technical corpora. *Procesamiento del Lenguaje Natural* (49):67–74.

Nazar, R. 2010. *A Quantitative Approach to Concept Analysis*. Ph.D. Dissertation, Universitat Pompeu Fabra.

Pantel, P., and Pennacchiotti, M. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proc. of 21st International Conference on Computational Linguistics and 44th annual meeting of the ACL (Sydney, Australia)*, 113–120.

Pekar, V.; Krkoska, M.; and Staab, S. 2004. Feature weighting for co-occurrence-based classification of words. In *Proc. of the 20th International Conference on Computational Linguistics*, COLING '04. Stroudsburg, PA, USA: Association for Computational Linguistics.

Potrich, A., and Pianta, E. 2008. L-isa: Learning domain specific isa-relations from the web. In *LREC*. European Language Resources Association.

Pustejovsky, J.; Hanks, P.; and Rumshisky, A. 2004. Automated induction of sense in context. In *Proc. of the 20th International Conference on Computational Linguistics*, COLING '04. Stroudsburg, PA, USA: Association for Computational Linguistics.

Renau, I., and Nazar, R. 2012. Hypernym extraction by definiens-definiendum co-occurrence in multiple dictionaries. *Procesamiento del Lenguaje Natural* (49):83–90.

Rydin, S. 2002. Building a hyponymy lexicon with hierarchical structure. In *Proc. of the ACL-02 Workshop on Unsupervised Lexical Acquisition - Volume 9*, ULA '02, 26–33. Stroudsburg, PA, USA: Association for Computational Linguistics.

Sahlgren, M. 2008. The distributional hypothesis. *Rivista di Linguistica 20(1)* 33–53.

Schütze, H., and Pedersen, J. 1997. A co-occurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management* 33(3):307–318.

Snow, R.; Jurafsky, D.; and Ng, A. 2006. Semantic taxonomy induction from heterogeneous evidence. In *Proc. of the 21st International Conference on Computational Linguistics (Sydney, Australia)*, 801–808.

Velardi, P.; Faralli, S.; and Navigli, R. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics* 39(3):665–707.

Vossen, P. 1998. Eurowordnet: A multilingual database with lexical semantic networks. *Computers and the Humanities 32(2-3)*.

Wang, W.; Barnaghi, P.; and Bargiela, A. 2009. Probabilistic topic models for learning terminological ontologies. *IEEE Transactions on Knowledge and Data Engineering* 99(RapidPosts).

Wilks, Y.; Fass, D.; Guo, C.; McDonald, J.; Plate, T.; and Slator, B. 1989. A tractable machine dictionary as a resource for computational semantics. In *Computational Lexiography for Natural Language Processing. B. Boguraev and T. Briscoe (eds): Essex, UK: Longman*. 193–228.

Woon, W. L., and Madnick, S. 2009. Asymmetric information distances for automated taxonomy construction. *Knowl. Inf. Syst.* 21(1):91–111.