Automatic Extraction of Patterns Displaying Hyponym-Hypernym Co-Occurrence from Corpora

Verginica Barbu Mititelu
Romanian Academy Research Institute for Artificial Intelligence
13, Calea 13 Septembrie, Bucharest 050711, Romania
vergi@racai.ro

Abstract

Many of the tasks in computational linguistics, such as information retrieval, document classification, automatic summaries, word sense disambiguation, resolving prepositional phrase attachment, etc. (see Vossen 2003 for a presentation of the uses of various ontologies in solving different tasks in Natural Language Processing) need good ontologies for their success. The manual development of an ontology requires considerable time and money investments. An alternative way for their development is to extract the relevant content from (domain-specific) corpora. A prerequisite in such an experiment is the inventory of patterns which allow for the instantiation in text of the taxonomic relation organizing the ontology. We ran an experiment in which we identify such patterns in corpora and classify them from a lexical point of view. Another resource on which we rely is WordNet, whose already encoded hyponymy relations help us to identify the patterns in which they occur in corpus.

Introduction

Querying large corpora (and even the web) for extracting necessary information may need resources specific to the domain to which the query belongs. A general linguistic ontology such as WordNet (Fellbaum 1998) may prove insufficient when trying to establish the relation between, for instance, *chronic hepatitis* and *toxic hepatitis*. WordNet 2.1 does not record either of the medical terms. But if one checks a medical ontology (such as MeSH[1]), s/he will immediately find the relation between the two (this relation is called co-hyponymy in linguistics: Lyons 1977).

However, not all domains benefit of an ontology. One way of creating such a resource is to simply develop it by hand (which is a very time- and money-consuming method, in spite of the accuracy obtained). Another way is to create it by extracting its concepts and the relations between them from corpora or machine readable dictionaries (ideally dictionaries specific to the respective domain, not general language dictionaries).

The relation organizing concepts hierarchically is called class inclusion in logics. In linguistics we speak about hyponymy[2]. Theoreticians speak of hyponyms of a superordinate (or of a hypernym).

There have been experiments (for a review of these works see Cederberg and Widdows 2003) in which researchers used certain patterns to extract hyponyms and hypernyms co-occurrent in the same syntactic unit (a sentence). These patterns were not established via a comprehensive method. However, if we are interested in automatically developing an as good as possible ontology by getting as much as possible from a corpus or at least making use of a way to help us develop the resource, then it is necessary to use an exhaustive list of such patterns. Moreover, the degree of reliability of these patterns has to be as high as possible.

Our aim is precisely that of identifying the (possibly exhaustive) inventory of patterns that allow for the co-occurrence of hyponyms and hypernyms in corpora and of establishing their specificity to hyponymy: when discussing about hyponymy patterns, different authors give examples such as:

[1] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=mesh
[2] See Lyons (1977:221) for a discussion of hyponymy in terms of class inclusion.

(1) ... most European countries, especially France, England, and Spain. (Hearst 1998)
(2) Even then, we would trail behind other European Community members, such as Germany, France and Italy... (BNC) (Cederberg and Widdows 2003)

However, such examples are not appropriate, because *France*, *England*, *Spain*, in the first one, and G*ermany*, *France*, *Italy*, in the second, are not hyponyms of *countries* and *members*, respectively, but they are instances of the concepts lexicalized by the respective words. So, the patterns in the two examples are not specific only to hyponymy.

Our article is organized as follows: in the next section we present work that is related to ours. Section 2 contains the description of the experiment we have undergone and whose results and their interpretation are presented in section 3, while the conclusions and the future work we aim to conduct will close the article.

1. Related work

The first article in which a method for discovering patterns that allow for the co-occurrence of hyponyms and hypernyms in a corpus belongs to Hearst (1992). As she clearly points out, the idea is not entirely new: on the one hand, automatic extraction of taxonomic relations had been done using a Machine Readable Dictionary (Alshawi 1987, Markowitz et al. 1986, Jensen and Binot 1987, Nakamura and Nagao 1988); on the other hand, those working with text corpora (Coates-Stephens 1991, Velardi and Pazienza 1989, Brent 1991, Smadja and McKeown 1990, Calzolari and Bindi 1990) were interested in using patterns for extracting other types of information: semantic description of proper nouns, case roles assignment, verb subcategorization frame recognition, collocation acquisition and, respectively, prepositional complementation relations, modification relations, and significant compounds. More recent works with corpora for extracting taxonomic relation belong to Caraballo (2001), Widdows (2003).

Hearst (1998) takes over the (1992) experiment without modifying either the number and structure of patterns or the main lines of the algorithm for identifying pairs of hyponym(s)-hypernym. Her algorithm has the following steps:

- decide on a lexical-semantic relation of interest;
- decide a list of word pairs from WordNet in which this relation is known to hold;
- extract sentences from a large corpus in which these terms both occur, and record the lexical and syntactic context;
- find the communalities among these contexts and hypothesize that the common ones yield patterns that indicate the relation of interest.

Alfonseca and Manandhar (2001) apply Hearst's algorithm in a broader way with the final aim of improving a system that classifies unknown concepts in the WordNet ontology. The steps of their algorithm are the following:

- for each WordNet synset, a query is automatically constructed for Altavista Internet search engine and a set of documents is collected that contain the words in that synset;
- the documents are processed (tokenization, sentences splitting, POS tagging, stemming, NP chunking);
- select from these documents the sentences that contain both any of the synset words and any of the hypernym's words;
- extract the hyponymy patterns from the sentences using first order logic predicates and prune the low frequency ones.

As their primary aim was not to make an inventory of the hyponymy patterns, the article does not contain a list of the lexical-syntactic patterns identified using the described algorithm. Only four such patterns are provided as examples. (As noted in our Introduction, the examples are chosen such that, in fact, they exemplify the co-occurrence of an instance

and the class it belongs to.) From their article it is not obvious either how long the distance between the hyponym and the co-occurrent hypernym is; more precisely, we do not know if they allow only for direct hyponymy or also for indirect hyponymy (of various distance between the synsets to which the two words belong).

2. The algorithm for extracting co-occurring hyponym-hypernym patterns

2.1. Resources

For the experiment we present in this article we make use of two valuable linguistic resources: WordNet 2.1 and British National Corpus (BNC).

We chose BNC because it is a corpus representative for the general use of English. Although very rich in synsets, WordNet is also representative for the general language rather than being appropriate when working with a corpus belonging to a well-defined domain (WordNet contains terms from various domains but not to such a great extent that could make it useful as an ontology for a certain domain).

2.2. WordNet Characteristics

WordNet[3] (Fellbaum 1998) is a semantic network in which English words belonging to the open classes (nouns, verbs, adjectives and adverbs) are organized according to the semantic and lexical relations they establish for each of their senses: more precisely, nouns are organized in hyponymic and meronymic hierarchies, verbs in hyponymic, troponymic and lexical implication hierarchies, adjectives are organized in clusters whose head is an antonymic pair of adjectives for which the similar adjectives are encoded, and adverbs have no organization; if the case, their antonyms are encoded. Each node of the semantic network contains a set of synonyms (thus the name *synset* for the content of each node) to which literals belong with only one of their meanings.

2.3. Processing BNC

BNC was released in SGML format (see *BNC Users Reference Guide*, www.natcorp.ox.ac.uk/World/HTML). We found that converting it in XML format makes it easier for us to process (XML has strict rules of well-formedness, unlike SGML which allows for minimizations: in BNC the following minimizations were used: final tag suppression, attribute name omission). For this conversion we used the Windows version of the tool SP 1.3.4 developed by James Clark (www.jclark.com/sp/). However, we ran the experiment on only one file from the entire BNC. The motivation behind this decision can be found in **3.** below.

2.3. The Algorithm

The aim of our experiment is to extract from the corpus those sentences in which at least one hyponym is co-occurrent with one of its direct or indirect hypernyms. In order to achieve this aim, we followed the following steps (implemented in a Perl scrip):

i. For each sentence in the corpus extract the occurring nouns and verbs. We are not interested in adjectives and adverbs because these are not represented in

[3] http://wordnet.princeton.edu/

WordNet as linked by the hyponymy relation. Those cases when a certain word is annotated with two different parts of speech (ex.: <w type="nn1-vvb" teiform="w">dwarf </w> is annotated both as noun and verb) are considered by our algorithm twice: once as a noun and once as a verb.

ii. For each noun and verb, respectively, check if any of its hypernyms appear in the current sentence. The hypernyms are extracted from WordNet 2.1 (using the query *hypes* which returns both hypernyms and classes to which instances belong). If such a co-occurrence is encountered, then the respective sentence is extracted in a different file (depending on the part of speech of the target words).

iii. Group the extracted sentences according to the lexical similarity of the context between hyponym and hypernym.

So, we do not impose in our algorithm the length of the chain between the hyponym and its co-occurring hypernym (although we restrict the search only to the limit of a sentence), as, on the one hand, WordNet contains also some artificial nodes in its organization (see for instance the synset {change of magnitude}), and, on the other hand, because it is not a rare case to encounter structures in which a hyponym is co-occurrent not with its direct hypernym, but with an indirect one.

We also allow for the co-occurrence of the hyponym and hypernym in irrespective order. This is ensured by the fact that for each noun and verb we check if any of the co-occurrent nouns or verbs is its hypernym.

3. Results and Interpretations

We ran the Perl script on a BNC file and, initially grouped the results according to the syntagmatic distance between the co-occurrent hyponym and hypernym: the syntagmatic distance varies from 0 to 312 in the case of nouns and from 0 to 300 in that of verbs. The sentences in which the distance is 0 display the auto-hyponymy phenomenon existing in WordNet: a synset containing a literal with sense number *i* appears as hyponym of a synset containing the same literal with sense number *j*:

(3) In the experience of friends who *canvass* for the Labour party, old, white, middle-class men are the rudest.

Here is the fragment of WordNet verb hierarchy motivating this sentence extraction[4]:

poll, *canvass*, canvas -- (get the opinions (of people) by asking specific questions)
=> survey -- (make a survey of; for statistical purposes)
=> analyze, analyse, study, examine, *canvass*, canvas -- (consider in detail and subject to an analysis in order to discover essential features or meaning; "analyze a sonnet by Shakespeare"; "analyze the evidence in a criminal trial"; "analyze your real motives")

When the syntagmatic distance is 1 the hyponym appears next to its hypernym and, most of the times (and the syntactic relation between the two is specification). There is no lexical element between the hyponym and hypernym in such cases. However, our algorithm simply cannot catch any interesting context at the left of the first word of interest and at the right of the second (see, as an example, the underlined part in (4)):

[4] The sense with which *canvass* is used in (3) is none of the two senses extracted here from WordNet. See below the discussion about the semantic annotation of the corpus.

(4) Indeed, unlike those other forms of *discrimination*, *ageism* has yet to attract the attention of policy makers and the public, so deeply engrained is it in our thoughts and actions.

An analysis of the extracted sentences from the point of view of the syntagmatic distance motivated us to choose as relevant material for further study those sentences with a distance from 2 to 5. It is true that interesting sentences can also be found in cases of longer distances: for instance, sentences with large enumerations:

(5) apples, pears, oranges, bananas, grapefruit, lemons, limes, tomatoes, pineapple, avocado, guava, passion and other exotic fruits, mango, nectarines, apricots, dried fruit, berries

We grouped the structures of a syntagmatic distance from 2 to 5 according to their lexical similarity, especially their identity, but without disregarding the partial similarity. Besides structures completely irrelevant, we found the following patterns that we consider relevant for the co-occurrence of interest for us:

in_particular#r: countries, in particular Japan
particularly#r: food (particularly chocolate
particularly#r the#at0: waters, particularly the reservoirs
particularly#r and#cjc: handicrafts, particularly knitting and sewing
including#prp: fish, including hake
including#prp the#at0: activity, including the provision
especially#r: fruits, especially citrus
especially#r the#at0: animals, especially the reptiles
such_as#prp: alkanes such as methane
such_as#prp the#at0: country such as the US
such_as#prp a#at0: time such as a day
except#cjs: cities (except Birmingham
except#cjs-prp: animal except Homo sapiens
as#cjs-prp: ventures as undertaking
notably#r: sports, notably rugby
usually#r: material (usually cedarwood
mostly#r: people , mostly women
mainly#r: fruits , mainly raspberries
like#prp: exercise (like jogging
like#prp other#a: Glasgow, like other cities
as#cjs: creatures as bees
as#cjs-prp a#at0: English , as a subject
as#r: teeth as fangs
even#r: names , even nicknames
in_common_with#prp other#a: Christianity , in common with other religions
as_well_as#cjc the#at0: cohesion , as well as the development
other_than#prp the#at0: person other than the candidates
not#xx0 least#dt0: countries , not least Germany
but#cjc not#xx0: fluids but not coffee
for_example#r: subjects: for example history
for_example#r the#at0: process , for example, the use
e.g.#r: vegetables , e.g. carrots

eg#r: metals , eg sodium
i.e.#r: institutions , i.e. banks
ie#r: products (ie goods
another#dt0: weight: another property
an#at0: Heparin (an anticoagulant
a#at0: Fibrinogen : a substance
kind#n of#prf: speech , a kind of monologue
call#v: granite , called batholiths
and#cjc: woodpeckers and birds
and#cjc other#a: aspirin and other drugs
and#cjc sometimes#r other#a: rats and sometimes other creatures
and#cjc many#dt0 other#a: salamanders and many other animal
and#cjc in#prp other#a: Germany and in other countries
or#cjc: inlets or fjords
or#cjc other#a: penne or other pasta
or#cjc any#dt0 other#a: dog , or any other animal
of#prf: state of excitement
be#v: Carbohydrates are compounds
be#v another#dt0: levels is another device
be#v the#at0: psychology is the science
be#v the#at0 only#a: Leicestershire is the only county
be#v an#at0: VAT is an tax
be#v a#at0: Canada was a land
which#dtq be#v the#at0: nature which was the creation

These structures can be grouped according to the order of occurrence of the words in hyponymy relation:

- hypernym-hyponym structures: *in_particular, particularly (the), including (the), especially (the), notably, for_example (the), as, such_as (the/a), not least* şi altele.
- hyponym-hypernym structures: *and other, or other, be another, a kind of, a/an, another, like other*.

As one can easily notice, some structures, although slightly different, are, in fact, identical: the difference between them is due to the syntactic realization of the second element of the pair: definite-indefinite (*food particularly chocolate – waters particularly the reservoirs*) or to the occurrence of an adjunct (*aspirin and other drugs – rats and sometimes other creatures*). Thus, we can say that the lexical perspective on the pattern is not enough and it should be completed with the syntactic one. There are cases when the lexical element is not specific to the co-occurrence of our interest here, but its appearance in a certain syntactic configuration makes it a signal of such an occurrence: see *usually* in an appositive position. Syntactic analysis is also helpful in cases as *Heparin: an anticoagulant, cricket: a game*, etc.
Some of the extracted patterns apply only to some vocabulary areas: *state of excitement, city of Danzig*. In other cases, patterns allow for the co-occurrence of words that are not in hyponymic relation for the senses with which they co-occur in the respective context: *study of Astronomy (study* here: "applying the mind to learning and understanding a subject (especially by reading)"; *study* as hyperonym of *Astronomy*: "a branch of knowledge").

Some patterns are surprising: *and* and *or*: see, for instance, *woodpeckers and birds*. However, the larger context explains the structure:

(6) *The trail takes you through both old and new plantations - look out for* ***woodpeckers and birds such as long tailed tits and wrens*** *which share their habitat*

Some of the patterns extracted from corpus allow for both the hyponymy and instance-of relations:
countries, in particular Japan
continent (including Antarctica
cities, except Birmingham
countries, e.g. USA
city of Danzig
etc.

Among the extracted patterns, there are three that are specific rather to co-hyponyms:

rather than#prp: management rather than administration
as opposed to#prp: capsules (as opposed to tablets
turn#n into#prp: process turns into action

Another pattern seems specific to synonyms:

also#r know#v as#prp: aconite , also known as monkshood

Further Work

The obvious next step of our experiment is to test the degree of relevance of the above-enumerated patterns for the instantiation of the hyponymy relation. For that we can investigate a comprehensive set of examples: either sentences extracted from a corpus, or snippets returned by Google search engine.

Given the different characteristics of the texts belonging to various language registers, we can try running the experiment on a corpus belonging to a certain domain. There may be structures (e.g. definitional structures) that are poorly represented in newspaper articles. Moreover, for the further aim of extracting an ontology from corpora, the patterns specific to the scientific domain are more relevant.

As hinted above, in order to get better results with our algorithm, it is helpful that the corpus should be semantically disambiguated. One such corpus is SemCor[5]. Thus, we would avoid dealing with the case when the two marked words are not hyponyms for the senses with which they co-occur in the respective sentence:

(7) *This includes a reminder that any person caught swearing must be made to pay for it.* (*reminder* here: "a message that helps you remember something"; *reminder* as a hyponym of *person*: "someone who gives a warning so that a mistake can be avoided")

A syntactically annotated corpus would also prove a good test bed for running an experiment, as it can offer the possibility of grouping syntactically similar structures under the same umbrella, instead of treating them as different (see definite-indefinite structures, with or without adjuncts, with or without modifying adjectives, etc.).

As shown in (4) above, the analysis needs be extended to the left and right context of the structure in which the hyponym-hypernym pair occurs, because sometimes the structure in between the two members of the pair is not relevant enough (see the *woodpeckers and birds* example above).

[5] http://multisemcor.itc.it/semcor.php

References:

Alfonseca, E. and S. Manandhar. 2001. Improving an Ontology Refinement Method with Hyponymy Patterns. In *Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain, pages 235-239.

Alshawi, H. 1987. Processing Dictionary Definitions with Phrasal Pattern Hierarchies. *American Journal of Computational Linguistics*, 13 (3): 195-202.

Brent, M. R. 1991. Automatic Acquisition of Subcategorization Frames from Untagged, Free-text Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 209-214.

Calzolari, N. and R. Bindi. 1990. Acquisition of Lexical Information from A Large Textual Italian Corpus. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, Helsinki, pages 54-59.

Caraballo, S. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *37th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, pp. 120-126.

Cederberg, S. and D. Widdows. 2003. *Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction.* In Conference on Natural Language Learning (CoNLL-2003), Edmonton, Canada, pp. 111-118.

Coates-Stephens, S. 1991. Coping with Lexical Inadequacy - the Automatic Acquisition of Proper Nouns from New Text. In *The Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora*, Oxford, pages 154-169.

Fellbaum, Ch. (Ed.) 1998. *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge MA.

Hearst, M. A. 1992. Automated Acquisition of Hyponyms from Large text Corpora. In *Proceedings in the Fourteenth International Conference on Computational Linguistics*, Nantes, France.

Hearst, M. A. 1998. *Automated Discovery of WordNet relations*. In Fellbaum (Ed.) (1998), p.131-151.

Jensen, K. and J-L. Binot. 1987. Disambiguating prepositional phrase attachments by using on-line dictionary definitions, *American Journal of Computational Linguistics*, 13 (3): 251-260.

Lyons, J. 1977. *Semantics*, Cambridge University Press, volume 1.

Markowitz, J., T. Ahlswede, M. Evens. 1986. Semantically Significant Patterns in Dictionary Definitions. In *Procedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 112-119.

Nakamura, J. and M. Nagao. 1988. Extraction of Semantic Information from an Ordinary English Dictionary and Its Evaluation, In *Proceedings of the Twelfth International Conference on Computational Linguistics*, Budapest, pages 459-464.

Smadja, F.A. and K.R. McKeown. 1990. Automatically Extracting and Representing Collocations for Language Generation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 252-259.

Velardi, P. and M. T. Pazienza. 1989. Computer Aided Interpretation of Lexical Cooccurrences. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 185-192.

Vossen, P. 2003. *Ontologies*. In R. Mitkov (Ed.) *The Oxford Handbook in Computational Linguistics*, Oxford University Press.

Widdows, D. 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of HLT/NAACL 2003*, Edmonton, Canada, pages 276-283.