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

Verginica Barbu Mititelu
Romanian Academy Research Institute for Artificial Intelligence

To process information you need information.
(Vossen 2003)

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 resources for their success. Ontologies make a good resource for such techniques. They have the advantage that they store **structured** information (an ontology is an explicit specification of a conceptualization). It is common sense that the better the ontology used, the better the obtained results.

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) 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 terms. But if one checks a medical ontology (such as MeSH - see http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=mesh), s/he will immediatey find the relation between the two.

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 linking them from corpora or machine-readable dictionaries (see, for instance, Maedche and Volz).

The relation organizing concepts hierarchically is called class inclusion in logics. In linguistics we speak about hyponymy[1]. Theoreticians speak of hyponyms of a superordinate (or hypernym).

There have been experiments (for a review of these works see Cederberg and Widdows 2003) in which researchers used certain patterns to extract co-occurrent hyponyms-hypernyms. These patterns were established via linguistic introspection. However, if we are interested in developing an as good as possible ontology by getting as much as possible from a corpus, then it is necessary to use an exhaustive list of such patterns.

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.

The algorithm for extracting co-occurring hyponym-hypernym patterns

For the automatic extraction of these patterns we make use of three kinds of resources: corpora, ontologies and parser. We decided to use a general corpus (British National Corpus - BNC) and a domain-specific corpus (Harrison's book, *Principles of Internal Medicine*). The reason for choosing not to work only with a general corpus, but also with one belonging to two the scientific register is that some patterns may not be instantiated in the former. What we

[1] See Lyons (1977:221) for a discussion of hyponymy in terms of class inclusion.

have in mind are the definitional structures that are less likely to appear in newspaper articles, for instance, but are quite frequent in scientific papers, especially in those works addressed to students (and Harrison's book is intended for the use of students and clinicians).

The ontologies we make use of are the Princeton WordNet (version 2.1) that is useful for the general corpus, and UMLS (Unified Medical Language System), that is used for the medical corpus.

Using a Perl script we extract from the corpora those sentences that contain words that are in a direct or indirect hyponymic relation in the corresponding ontology. Dealing only with direct hyponymy is not satisfactory, as in a text one can encounter the following string: "a disease such as hepatitis". According to UMLS *hepatitis* is a hyponym of *liver disease,* which is a hyponym of *digestive system diseases*, which is a hyponym of *diseases*. So, if we had taken into consideration only the direct hyponymy (when no node intervenes between the two), then this string would not have been identified as displaying the co-occurrence we are interested in. Ontologies vary frequently in the degree of their granularity: some of them make finer distinctions than others. In order to overcome problems appearing due to different granularity of resources, we decided to allow for the co-occurrence of indirect hyponymy as well.

The output of the algorithm so far is a list of lexical patterns that we will present in the final version of this paper, alongside with some necessary remarks on them. We are also interested in the syntactic aspect of these patterns. That is why we plan to parse the extracted sentences using Charniak's parser (Charniak 1997). In the end we can make a lexico-syntactic inventory of such patterns.

Further work

Once the sentences parsed, we would also be interested in the degree to which they are transferable into another language. The lexical translation would be eased by the existence of wordnets for various languages (see www.globalwordnet.org /gwa/wordnet_table.htm). Their link to the Interlingual Index (Vossen 1998) ensures their use as multilingual dictionaries (in which there are sense mappings from one language to the other).

We would also like to check the degree of overlap between the set of patterns in which hyponyms and hypernyms co-occur and those in which instances are involved.

The results obtained so far motivate us to run the algorithm on a semantically disambiguated corpus, such as SEMCOR (http://multisemcor.itc.it/semcor.php).

The automatic identification of patterns displaying hyponym-hypernym co-occurrence that we presented above can be adjusted to work for the extraction of patterns specific to various kinds of relations (antonymy, disease-treatment, etc.).

Bibliography:

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.

Charniak, E. (1997) Statistical Parsing with a Context-free Grammar and Word Statistics. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, AAAI Press/MIT Press, Menlo Park.

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

Maedche, A. and R. Volz, *The Ontology Extraction &Maintenance Framework Text-to-Onto*, at http://computing.breinestorm.net/ontology+learning+concepts+relations+data/

Vossen, P. (Ed.) (1998) *A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht

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