Using Hearst's Rules for the Automatic Acquisition of Hyponyms for Mining a Pharmaceutical Corpus

Michael P. Oakes
School of Computing and Technology
University of Sunderland
David Goldman Informatics Centre
St. Peter's Campus, Sunderland, SR6 0DD, England
michael.oakes@sunderland.ac.uk

Abstract

Fully Automatic Thesaurus Generation (ATG) seeks to generate useful thesauri by mining a corpus of raw text. A number of statistical approaches, based on term co-occurrence, exist for this, but in general they are only able to estimate the strength of the relationship between two terms, not its nature. In this paper we implement Hearst's method of discovering the hyponymy relations which are the building blocks of hierarchical thesauri. We start with the Scrip corpus of newsfeeds in the domain of psychology, and were able to discover an estimated 400 useful term relationships.

1 Introduction

A domain-specific thesaurus such as MeSH (MEDLINE) or the Derwent Drug File (DDF) gives an overview of the extent of the domain, and the categories, relations and named entities within it. They typically consist of lists of terms organised according to a semantic hierarchy. Electronic thesauri are used in document retrieval or indexing systems, for expanding queries when searching for information or the selection of a preferred form of a given search term. Experiments such as the Worm Community System have shown that the thesaurus is an excellent memory-jogging device which supports learning and serendipitous browsing. Thesauri prevent users from becoming overwhelmed by the sheer amount of available information, and the "classical vocabulary problem, which results from the diversity of expertise and backgrounds of systems users" (Chen et al., 95).

Although a number of successful commercially-available thesauri created by large teams of human experts are available, in general manual thesaurus generation is prohibitively costly. Grefenstette (94) writes that the ideal might be to use knowledge-poor approaches, starting from just the raw corpus - "if the ultimate goal of ATG (Automatic Thesaurus Generation) is the deduction of semantic relationships exclusively from free text corpora". ATG is thus an example of knowledge discovery in text databases or text data mining.

Most existing methods for automatic thesaurus generation are statistical, and rely on the co-occurrence of a pair of terms within a common "window" of text, which may be a fixed number of words, within the same syntactic clause, or within a common document in a large collection of documents. Details of such approaches were given first by Salton in 1989, and more recently by Pereira et al. (93) and Kageura et al. (00). For each word pair in the corpus vocabulary, such methods are able to generate a numeric score to estimate semantic relatedness. However, to my knowledge, such statistical methods are unable to discriminate between different types of semantic relations, such those held in the manually-produced WordNet (Fellbaum 98) thesaurus: synonym (same as), antonym (opposite to), hyponym (is_a), meronym (part of) entailment (where one entails the other, e.g. "buy" and "pay") and troponym (the two concepts which entail each other must happen at the same time).

The "thesauri" produced by purely statistical approaches tend to be networks of interrelated terms with no sense of hierarchy, whereas the most widely-used thesauri terms for narrower concepts near

the bottom and terms for broader concepts nearer the top, with all terms connected via a single generic root concept. Sometimes the hyponymy relation is described as "parent-child", where the parent is the broader term (hypernym) and the child is the narrower or more specific term (hyponym).

2 Hearst's method

In a novel alternative approach, Hearst (92) produced an automatic lexical discovery technique that uses lexico-syntactic patterns to find instances of hyponymy relations (such as "apirin is_a drug") between noun phrases identified in a free text corpus. She used a parser to identify the noun phrases (NP) in the text, but in this paper a simpler approach was taken: entities related by the hyponymy relation consisted either of single words, or of single words preceded by "the", "a" or "an". One of Hearst's lexico-syntactic patterns is given below, where NP means a noun phrase, {}* means that the enclosed sequence may repeat any number of times, {} denotes an optional sequence, and (a|b) means either a or b may occur in the sequence at that point:

such NP as {NP,} {(or|and)} NP*
as in **… works by such authors as Herrick, Goldsmith and Shakespeare.**

When a sentence containing this pattern is found, the following hyponymy relations can be inferred:

hyponym (author, Herrick)
hyponym (author, Goldsmith)
hyponym (author, Shakespeare)

This approach has the advantage over statistical methods of determining term-term relatedness, which rely on multiple occurrences of a term pair to be in proximity with one another, in that only one occurrence of the pattern need be found for the relation to be identified (Grefenstette 94).In formulating this approach, Hearst had two main motivations: to avoid the need for pre-encoded knowledge and to produce a technique which is applicable over wide range of texts. This enables a text-mining approach whereby an ontology of hyponymic relations can be derived from a corpus of raw text. The set of lexico-syntactic patterns indicating the hyponymy (the hyponym is the narrower term) were chosen to satisfy the following desiderata:

1. They occur frequently and in many text genres
2. They (almost) always indicate the relation of interest.
3. they can be recognised with little or no pre-encoded knowledge

In this paper we use Hearst's rules to discover the homonymic relations in a free-text collection of Scrip. Scrip is an electronic daily news bulletin, distributed to those working in the pharmaceutical industry. It covers a number of topics relevant to the industry, including product launches, licensing, forthcoming meetings, personnel data, announcements by regulatory authorities, company relations and clinical trials. The data set used here consisted of all issues covering the period January to March 1999, a total of 631, 269 words including HTML mark-up. Scrip is a trademark of PJP publications Ltd. (See http://www.pjpub.co.uk).

3 Experiment on *Scrip*

Hearst's lexico-syntactic rules were adapted slightly and encoded in a program written in Perl. The set of rules used in this implementation are given below, with examples of patterns found in the Scrip corpus.

NPn = the|a|an + one word

(1.1) *NP1 such as NP2*
… diseases such as hepatitis …
→ hyponym (disease, hepatitis)

(1.2) *NP1 such as NP2 (and|or) NP3*
… cities such as Beijing and Guangzhou …
→ hyponym (cities, Beijing),
hyponym (cities, Guangzhou)

(1.3) *NP1 such as NP2, NP3 (and|or) NP4*

… infections such as bronchitis, sinusitis or pneumonia …
→ hyponym (infections, bronchitis), hyponym (infections, sinusitis), hyponym (infections, pneumonia)

(2.1) *such NP1 as NP2*

Pattern (2.1) yielded no useful matches in the Scrip corpus; two spurious matches were "dispensing such gifts as a marketing tool" → hyponym (gifts, marketing), and "such factors as tenderness" → hyponym (factor, tenderness), where "factors" was too broad a term to be useful.

(3.1) *NP1 {,} (or|and) other NP2*
… vaccines, or other injectables…
→ hyponym (injectables, vaccine)

(3.2) *NP1, NP2 {,} (or|and) other NP3*
… royalties, fees, and other revenues…
→ hyponym (revenues, royalties), hyponym (revenues, fees)

(3.3) *NP1, NP2, NP3 {,} (or|and) other NP4*
… Italy, Canada, the US and other countries …
→ hyponym (countries, Italy), hyponym (countries, Canada), hyponym (countries, US)

(4.1) *NP1 {,} {including|especially} NP2*
… cytokines, including BNF …
→ hyponym (cytokines, BNF)

(4.2) *NP1 {,} {including|especially} NP2 {or|and} NP3*
… technologies including ATLAS and SCAN …
→ hyponym (technologies, ATLAS), hyponym (technologies, SCAN)

4 Discovered Relations

The broader terms most often matched by the rules are listed below in bold type, along with five examples of discovered hyponyms in italics. The figure on the left is the number of times the broader term was found to act as a hypernym in the corpus. Unless otherwise indicated, the hyponyms could all be meaningfully grouped under their parent term in a thesaurus, and thus Hearst's rules and the Scrip corpus together form the basis for a useful taxonomy of the field of pharmacology.

53 **products** - the hyponyms were a mixture of names of drugs and companies, e.g. *ACE, Astra, Calcichew, Cognex, Glaxo,* so to be useful we need a way of subdividing these.
45 **countries** - e.g. *Africa, America, Brazil, Canada, China.*
34 **diseases** - e.g. *Alzheimer, Ebola, HIV, Marburg, asthma.*
28 **areas** - (research) e.g. *ENT, asthma, biotechnology, cancer, cardiovascular.*
18 **conditions** - all hyponyms were names of body parts or diseases e.g. *acne, arthritis, balloon, colon, diabetes.*
17 **issues** - this category was too broad to be useful, containing such diverse terms as *HIV, IP, Ukraine, adherence, compulsory.*
16 **companies** e.g. *Bristol, Centeon, David, Eskom, Glaxo.*
16 **markets** - all hyponyms were names of countries, *e.g. Asia, Australia, Canada, Denmark, EC.*
13 **bodies** (regulatory) e.g. *EC, National, WHO, WTO, World.*
13 **effects** (side effects) e.g. *bleeding, dizziness, headache, hot, loss.*
11 **drugs** e.g. *ciclosporin, cocaine, corticosteroids, heroin, leukotriene.*
11 **factors** - too broad a category to be of use, containing such terms as *construction, degree, drugs, goodwill, higher, hypertension.*

Other interesting categories which were discovered included **disorders:** *Parkinson, chronic, diabetes, hepatitis, hypertension, idiopathic, sequelae, shock, stroke;* **technologies:** *ATLAS, SCAN, gene, genomics, laser, libraries, taste* and **cancers:** *breast, colon, colorectal, lung, non, ovarian, pancreatic.*

It would be difficult to print out the entire set of results in a neat, strict indented hierarchy using a recursive algorithm, as some terms had a large number of parent terms which tended not to be very descriptive (e.g. *cancer* had the hypernyms

diseases, areas, illnesses, pathology, terminal, serious, and *UK* had the hypernyms **countries, markets, member, sites, year, 1980s**). The word most commonly matching the hyponym slot was "those", suggesting that this word should be incorporated into the original definition of a noun phrase as an alternative to "the".

5 Bootstrapping Approach to Search for New Patterns

Hearst proposed, but did not implement, a bootstrapping approach to learning new lexico-syntactic patterns indicating hyponymy. The method is to gather a list of terms for which the lexical relation is known to hold, e.g. hyponym (countries, Bulgaria). The environments where this pair of terms occur syntactically near one another should be recorded, and the most common ones can then be used as patterns that indicate the relation of interest. This approach was implemented in Perl, starting with a list of all non-spurious hypernym-hyponym pairs (as edited subjectively) with a frequency of 2 or more. The most frequently occurring of these were:

6 diseases - cancer
6 countries - US
5 bodies - WHO
5 countries - UK
4 diseases - HIV
4 agreements - trade
4 countries - Canada

Whenever both the parent and child terms were found in the same sentence in the Scrip corpus, a record was kept of which term was the broader (bt) and which was the narrower (nt) and the intervening words. For example, if a sentence containing "diseases including cancer" was found "bt including nt" would be stored. The frequency of each of the stored strings was found at the end of the program, the most frequent being assumed to provide the most useful contexts. The most frequent learned contexts are listed below:

37 nt bt
20 nt and other bt
20 bt such as nt
12 bt, including nt
8 bt, including the nt
6 bt such as the nt
6 bt, such as nt
5 bt including nt
4 bt and nt
4 bt in nt
4 bt outside the nt
4 bt including the nt

This experiment was less successful in that we were only able to relearn some of the patterns originally developed by Hearst, with the exception of "bt and nt" (considered unreliable because it will pick up sibling terms rather than parent and child) and "bt in nt" and "bt outside the nt", which show how terms are related spatially rather than reveal hyponymy. Even though "nt bt" was the most common learned pattern, it is clearly not reliable to assume that every word in a text must be a hyponym of the following word.

6 Conclusion

It has been demonstated that Hearst's rules for discovering hyponymic relations for automatic thesaurus generation were highly effective when working with Scrip, a corpus of pharmaceutical newsfeeds. Altogether 1054 unique hyponymy relations were discovered, and taking the first 200 relations to be learned as a sample, 83 of the relations were deemed useful. This suggests that about 400 useful relations were learned in total. A more rigorous method of evaluating this approach would compare the relations learned with those in an existing humanly generated thesaurus in the same domain, such as the Derwent Drug File (DDF) for pharmacology. Recall would be the (number of hyponyms found both by Hearst's method and in the DDF thesaurus) divided by (the number of hyponyms in the DDF thesaurus), and Precision would be the (number of hyponyms found both by Hearst's method and in the DDF thesaurus) divided by the (number of hyponyms found by Hearst's method).

References

(Chen et al. 95) Chen, H., Yim, T., Fye D., and Schatz, B. (1995). *Automatic thesaurus generation for an electronic community system*. Journal of the American Society for Information Science 46(3): 173-195.

(DDF) Derwent Drug File, Thompson Scientific, 14 Great Queen Street, London WC2B 5DF, England.

(Fellbaum 98) Fellbaum, C. (1998). *A lexical database of English: The mother of all WordNets*. Special Issue of Computers and the Humanities, ed. P. Vossen, pp. 209-220.

(Grefenstette 94) Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Norwell MA: Kluwer Academic Publishers.

(Hearst 92) Hearst, M. A. (1992). *Automatic acquisition of hyponyms from large text corpora*. Proceedings of the 4th International Conference on Computational Linguistics (Nantes, France): COLING 1992.

(Kageura et al. 00) Kageura, K., Tsuji, T., and Aizawa, A. N. (2000). *Automatic thesaurus generation through multiple filtering*. Proceedings of the 18th International Conference on Computational Lnguistics (COLING 2000), Vol 1: 397-403.

(MEDLINE) MEDLINE, US Library of Medicine, http://medline.cos.com

(Pereira et al. 93) Pereira, F., Tishby, N. and Lee, L. (1993). *Distributional clustering of English words*. Proceedings of the Association for Computational Linguistics (ACL 93): 183-190.

(Salton 89) Salton, G. (1989). *Automatic Text Processing*. Reading MA: Addison-Wesley.