Automatic Hyponym Classification

Noora Routasuo (BOTJ1050)
Graduate School of Information Sciences,
Tohoku University
Academic Advisor: Kentaro Inui

Introduction

A *hyponym* is a word or a phrase that is a *type* or a *kind* of another phrase, called its *hypernym*. For example, ‘cat’ is a hyponym of ‘animal’. The automatic acquisition of simple semantic information such as hyponyms is the basis for numerous applications, for example ontology expansion and keyword generation. My research has focused on implementing an algorithm for automatically extracting hypernym-hyponym pairs from plain unannotated English text.

The algorithm generates a list of pairs and their frequencies in a corpus. The frequencies can be used to calculate a rough confidence measure for of a given pair.

Implementation

The algorithm is based on the simple idea, originally explored by Hearst[1], that sentences with words or phrases sharing a certain semantic relationship often conform to common patterns. Terms are then found by extracting all sentences that match to a predefined pattern, and selecting among the extracted sentences using some additional information. An example of a classical pattern for finding hyponyms is ‘[hypernym] *such as* [hyponym]’. Main patterns used by my implementation are listed in Table 1.

(s1) *such as* (s2)
(s1), *including* (s2)
(s1), *particularly* (s2)
(s1), *especially* (s2)
(s1), *for example* (s2)

Table 1. Patterns used by the matcher. (s1) denotes hypernym and (s2) hyponym.

Before applying the patterns I preprocess the text with a free tool called Genia Tagger[2]. It divides the text into parts called *chunks*, which essentially define the limits of a multi-word phrase. This allows the algorithm to recognize, for example, ‘natural disaster’ as a possible match instead of just ‘disaster’. Genia Tagger also recognizes base forms of inflected words, such as plurals, which can then be removed from the output.

The algorithm could be further improved by making use of part-of-speech tagging also provided by Genia. Thus words other than nouns with possible modifiers could also be discarded from the output.

Results

I tested my algorithm on two very different corpora - the English Wikipedia[3], and the ICWSM 2009 Spinn3r Blog Dataset[4]. I used roughly half of Wikipedia, approximately 1 billion words, and a portion of 110 million words from the ICWSM dataset. The top pairs from this experiment are in Table 2.

Wikipedia	**ICWSM09**
[91] country+australia	[31] site+stumbleupon
[75] country+germany	[19] condiment+ketchup
[55] country+france	[16] development tool+compiler
[48] country+japan	[16] controversial show+south park
[45] country+india	[16] developing country+china
[43] country+canada	[16] developing country+india

Table 2. The six most frequent hypernym-hyponym pairs in each corpus and their frequencies.

The algorithm found 32 734 hypernyms and 102 201 pairs in total in Wikipedia, and 2101 hypernyms and 3946 pairs in the ICWSM dataset.

My final task is to evaluate these results by comparing them to WordNet[5] and manually hand-tagging a random sample in the final weeks of July.

Conclusions

There are several ways that this basic algorithm could be improved upon, but it is interesting to see that promising results can be obtained even by such simple methods.

The experiment also highlights the differences of the two corpora used. Formal, structured language such as used in Wikipedia may prove easier for pattern-based extraction, whereas more casual text such as that found in blogs is likely to provide very different, if sparser, results.

References

[1] 1992. Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *ACL* 1992, pages 539–545.

[2] 2007. University of Tokyo. GENIA Tagger. Available at http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

[3] 2011. Wikimedia Foundation. Wikipedia database dumps. Available at http://dumps.wikimedia.org

[4] 2009. 3rd Int'l AAAI Conference on Weblogs and Social Media. ICWSM 2009 Spinn3r Blog Dataset. Available at http://www.icwsm.org/2009/data/

[5] 2006. Princeton University. WordNet 3. Available at http://wordnet.princeton.edu/