Automatic acquisition of hyponymy-relations from HTML documents

Keiji Shinzato (210044)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 13, 2003

Keywords: Knowledge acquisition, Hypernym, Hyponym, Statistical Natural Language Processing, World Wide Web.

In this thesis, we propose a new method for acquiring hyponymy relations from the HTML documents on the World Wide Web. The hyponymy relations can play a crucial role in various natural language processing systems. A number of techniques have been developed for automatically extracting the hyponymy relations from unrestricted text corpora[4, 5, 2, 6, 7, 3, 1]. These techniques use particular lexico–syntactic patterns to extract the hyponymy relations. For example, Hearst[4] attempted acquiring hyponymy relations from Grolier's American Academic Encyclopedia by using "*NP or other NP*" pattern. However, such approaches have the following problem.

- The lexico-syntactic patterns do not appear in text corpora frequently.

For this reason, the lexico-syntactic pattern based approaches cannot acquire a large number of hyponymy relations for wide range of words or phrases from text corpora.

To avoid the above problem, we propose a new method which does not use the lexico-syntactic patterns. We expect that our procedure can be applied to a wide range of expressions for which previous approaches cannot be used. We try to acquire hyponymy relations by combining four different types of clues obtainable for wide range of words or phrases. The first type of clue is HTML tags found in normal HTML documents. The second is statistical measures such as the document frequency (*df*) and the inverse document frequency (*idf*) which are widely used in Information Retrieval. The third is semantic similarities between hypernym and hyponyms given by verb–noun co–occurrences which are obtained from the HTML documents and the newspapers. The forth is a set of heuristic rules.

We made the following three assumptions for acquiring the hyponymy relations from the HTML documents.

Assumption A The expressions in the same itemization or list in HTML documents are likely to have a common hypernym.

Assumption B Given a set of hyponyms that have a common hypernym, the hypernym appears in many documents that include the hyponyms.

Assumption C Hyponyms and their hypernym are semantically similar.

Our procedure consists of the following four steps. The first three steps are corresponding to the above three assumptions.

Step 1 Extraction of hyponym candidates from itemized expressions in HTML documents.

Step 2 Ranking of hypernym candidates by DF and IDF.

Step 3 Selection of hypernym candidates and hyponym candidates by semantic similarities between hyponym candidates and hypernym candidates

Step 4 Application of a few more heuristic rules to discard some probably wrong hypernyms.

We call a set of expressions that are expected to have a common hypernym *Hyponym Candidate Set (HCS)*. In Step 1, we obtain HCSs from a large number of HTML documents, which are downloaded from the WWW, according to the assumption 1.

In Step2, our procedure acquires a candidate of a common hypernym for an HCS, which was obtained in Step1. First we prepare two sets of documents. We randomly select a large number of HTML documents and downloaded them. We call this set of documents a *global document set*. Then, we downloaded the HTML documents that includes each elements of the HCS using the existing web search engine (We use the goo[2] search engine in our experiment.) We call this set a *local document set*. Our procedure calculates the *score* of each noun which appears in a local document set, according to the assumption 2. The score was designed so that words frequently appearing in the local document set and relatively disappearing in the global document set are highly ranked. Then our procedure selects a noun that has the largest score value as a a candidate of a common hypernym for a given HCS. We call such a candidate *hypernym candidate*.

In Step3, our procedure ranks the pairs of an HCS and their hypernym candidate, which were obtained in Step2, by using the semantic similarities between an HCS and their hypernym candidate. The final output of our procedure is the top k pairs in this ranking after some heuristic rules are applied to it in the next step. The top k pairs contains only the hypernym candidates and the HCSs that are relatively similar to each other, and we can expect that only high quality hyponymy relations are contained in the top k pairs. Our procedure calculates the similarities between an HCS and their hypernym candidate according to the assumption3. More specifically, we compute an HCS's verb–noun co–occurrences from a local document set which is constructed in Step2, and the hypernym's verb–noun co–occurrences from a large number of text corpora. Then we create two vectors based on both verb–noun co–occurrences, and we compute the similarities by the cosine measure[8]. The cosine measure is one of the most popular measures which calculate the similarities of between two vectors.

Finally, in Step 4, we apply heuristic rules to refine pairs of an HCS and their hyponym candidate which are acquired in the previous steps.

We conducted a series of experiments on the acquisition of hyponymy relations by using the proposed method and a large number of HTML documents. First we downloaded 8.7×10^6 HTML documents from the WWW and applied the Step 1 to these HTML documents. We could extract the 9.02×10^4 HCSs. We randomly selected 2,000 HCSs from among the extracted HCSs as our test set. This test set consists of about 14,000 hyponyms. We applied the remaining steps to this test set. The preci-

[2] http://www.goo.ne.jp/

sion of acquired hyponymy relations reached about 75 % for 701 hyponym candidates, which was slightly more than 5 % of all the given hyponym candidates.

In addition, we compared the followings alternative methods with our method.

Alternative1 Compute the non–null suffixes that are shared by the maximum number of hyponym candidates, and regard the longest as a hypernym candidate.

Alternative2 Extract hypernyms for hyponym candidates by looking at the captions or titles of the regions from which hyponym candidates are extracted.

Alternative3 Extract hypernyms by using lexico–syntactic patterns.

Alternative4 Combination of all the above alternatives.

The results suggest that our method can acquire the significant number of the hypernyms that these alternative cannot obtain, at least, from small document sets that we used in our experiments. Then, we can conclude the proposed method can provide the useful method for acquiring hyponymy relation .

References

[1] Maya Ando, Satoshi Sekine, and Shun Ishizaki. Automatic extraction of hyponyms from newspaper using lexicosyntactic patterns. In *IPSJ SIG Technical Report 2003-NL-157*, pp. 77–82, 2003.

[2] Sharon A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 120–126, 1999.

[3] Michael Fleischman, Eduard Hovy, and Abdessamad Echihabi. Offline strategies for online question answering: Answering questions before they are asked. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics*, pp. 1–7, 2003.

[4] Marti A. Hearst. Automatic acquistition of hyponyms from large text corpora. Technical Report S2K–92–09, 1992.

[5] Marti A. Hearst. Automated discovery of wordnet relations. In Christiane Fellbaum, editor, *WordNet: an electronic lexical database*, chapter 5, pp. 131–151. MIT Press, 1998.

[6] Kyosuke Imasumi. Automatic acqusition of hyponymy relations from coordinated noun phrases and appositions. Master's thesis, Kyushu Institute of Technology, 2001.

[7] Emmanuel Morin and Christian Jacquemin. Automatic acquisition and expansion of hypernym links. In *Computer and the Humanities 2003*, 2003. forthcoming.

[8] G. Salton and M. E. Lesk. Computer evaluation of indexing and text processing. *Journal of the ACM*, Vol. 15, No. 1, pp. 8–36, January 1968.