Automatic Knowledge Acquisition for Highly Reliable Internet Search Engine

Keiji Shinzato Kentaro Torisawa
{skeiji,torisawa}@jaist.ac.jp
School of Information Science, JAIST

1 Aim

The goal of our project is to become able to automatically acquire various semantic relations between two words such as hyponymy relations and object-attribute relations from HTML documents on the WWW. The acquired relations will be used to enhance the reliability of search engines for WWW. We think a highly reliable internet search engine must have two properties. The first property is that the engine can provide a wide range of HTML documents that are related to given keywords. The second is that the system can prevent users from missing useful web pages. We expect that the semantic relations acquired by our methods to acquire semantic relations will be useful in developing the search engines that have the properties.

2 Approach

The search engine we are going to develop consists of two steps. In the first step, the system finds a hypernym of a word in a user's query. Here, a hypernym of a word X is another word Y such that the expression "X is a kind of Y" is acceptable. If a word in the query has two or more hypernyms, the system requests the user to select correct one. The obtained hypernym is used to collect a wide range of HTML documents that are related to the query. More precisely, the hypernym is added to the query and the search engine provides a set of documents that are related to the *expanded* query. Previous studies showed such a process, which is called a *query expansion* in general, can contribute to improvement of the search performance.

In the second step, the system ranks the results obtained in the first step so that important pages are highly ranked. This is to prevent users from missing useful web pages. In this step, we use the semantic relations between an object and the attributes that are used to describe the object. In general, when we explain something in detail, we have to describe it by using its attributes. For instance, when we describe a PC in detail, we have to talk about its manufacturer, processor, memory size, and O.S., which can be all regarded as attributes of PCs. We assume that useful web pages describing an object are likely to include a large number of such attributes of the object. The system ranks search results so that the web pages that include many attributes are likely to occupy a high position in the ranking, and we expect that this prevents us from missing important web pages.

Thus, we expect that semantic relations between words, such as hyponymy relations and object-attribute relations, are very useful knowledge to build highly reliable internet search engines. As previous attempts to acquire the relations automatically, many methods that rely on particular linguistic patterns, such as "*NP such as NP*", have been proposed. But it is difficult to acquire a wide range of the relations using the patterns because the frequencies of use for such linguistic patterns are relatively low.

Our approach relies on HTML tags and statistical measures such as document frequency and inverse document frequency instead of the linguistic patterns mentioned above. We observed that such information can provide clues to find semantic relations for a wider range of words or expressions than that of the expressions that conventional linguistic pattern based methods are applicable, because of the recent *explosion* of HTML documents on the WWW.

3 Progress of This Year

In this year, we have developed methods to acquire hyponymy relations and semantically coherent word classes from HTML documents on the WWW. First, we explain a hyponymy relation acquisition method. As mentioned, we do not use the

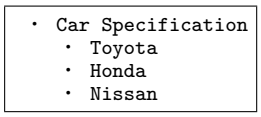

Figure 1 An example of itemization

linguistic patterns which were used in previous attempts. Instead, we used following assumptions to acquire a wide range of hyponymy relations.

Assumption A Expressions included in the same itemization or listing in an HTML document 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 hypernyms are semantically similar.

Consider the itemization in Fig. 1. We extract the set of expressions, $\{Toyota, Honda, Nissan\}$ from it. According to Assumption A, we can regard these expressions as candidates of hyponyms that have a common hypernym such as "*company.*" We then try to find a common hypernym for the expressions. We expect that the hypernym "*company*" appears in many documents including "*Toyota*", "*Honda*" and "*Nissan.*" We designed the score such that nouns that appear only in the documents that include one expression in the itemization have large score value, according to Assumption B, and regard the noun that has the largest score value as a common hypernym for each expression. Finally, we compute the similarity between a set of expressions in an itemization and its hypernym and then produce the pairs that have strong similarity as hyponymy relations. Pairs that do not have strong similarity, such as the pair of "*price*" and $\{Toyota, Honda, Nissan\}$, are discarded in our method according to Assumption C.

We then have tested the effectiveness of the method through a series of experiments in which we used HTML documents downloaded from actual web sites. The method could acquire correct hyponymy relations with precision of 75% when we regard top 5% of all the given hyponym candidates as output. We observed that the method can find a significant number of hypernyms that alternative hypernym acquisition procedures cannot acquire.

Next, we have extended our method for acquiring hyponymys of prespecified hypernyms. To acquire such hyponyms, we assumed that a heading of an itemization is likely to include a hypernym and used a sort of association strength between them as a clue of acquiring hyponymy relations. We used previous three assumptions to compute the association strength. The experiment showed that the extended method can acquire correct hyponymy relations with precision of 70% when we regard top 5% of all the given hyponym candidates as output. Although this method indicated lower precision than previous our method, the method could acquire multi-word hypernyms such as "*fortune telling web sites*" that previous our method could not acquire.

Finally, we have developed a method to extract semantically coherent itemizations from HTML documents. we expected that the method can speed up our hyponymy relation acquisition methods by using it as a filtering procedure.

4 Future Direction

We will try to improve the performance of our method as the first future work. One of the possible approaches will be to combine our method with the linguistic pattern based approaches. The second future work is to develop a method for acquiring object-attribute relations.

Publication List

Refereed Papers

1. Keiji Shinzato and Kentaro Torisawa, Acquiring Hyponymy Relations from Web Documents, In *Proceedings of Human Language Technology conference/North American chapter of the Association for Computational Linguistics annual meeting(HLT–NAACL04)*, pp.73–80, Boston, May, 2004.
2. Keiji Shinzato and Kentaro Torisawa, Extracting Hyponyms of Prespecified Hypernyms from Itemizations and Headings in Web Documents, In *Proceedings of the 20th International Conference on Computational Linguistics (COLING04)*, pp.938–944, Geneva, Aug, 2004.
3. Keiji Shinzato and Kentaro Torisawa, Automatic acquisition of hyponymy relations from HTML documents, Journal of Natural Language Processing, volume 12, Number 1, pp.125–pp.150, 2005.

Others

1. Keiji Shinzato and Kentaro Torisawa, Automatic acquisition of hyponyms based on headings and itemizations in HTML documents, IPSJ SIG Technical Reports NL–163, pp.29–pp.36, Tokyo, Sep, 2004. (in Japanese)
2. Keiji Shinzato and Kentaro Torisawa, Tangokurasu no kantanna tsukurikata, In *Proceedings of The 11th Annual Meeting of The Association for Natural Language Processing*, Kagawa, Mar, 2005. (in Japanese) (to appear)