Acquiring Ontological Knowledge from Query Logs

Satoshi Sekine
New York University
715 Broadway, 7th FL
New York, NY 10003 USA
+1-212-998-3175
sekine@cs.nyu.edu

Hisami Suzuki
Microsoft Research
One Microsoft Way
Redmond, WA 98052 USA
+1-425-703-2564
hisamis@microsoft.com

ABSTRACT
We present a method for acquiring ontological knowledge using search query logs. We first use query logs to identify important contexts associated with terms belonging to a semantic category; we then use these contexts to harvest new words belonging to this category. Our evaluation on selected categories indicates that the method works very well to help harvesting terms, achieving 85% to 95% accuracy in categorizing newly acquired terms.

Categories and Subject Descriptors
I.2.6, I.2.7 [**Learning & Natural Language Processing**]: *Knowledge Acquisition, Text Analysis.*

General Terms
Experimentation, Human Factors, Languages

Keywords
Named Entity, Ontology, Knowledge Acquisition, Query Logs.

1. INTRODUCTION
When people inquire about something (an object or an event), they typically want to know about its specific properties. For example, when people inquire about an award, they are usually interested in the nominees or winners. In this paper, we will describe a method for building up ontological knowledge about semantic categories and their properties using search query logs. Query logs are a very useful resource for this task, and present unique advantages over other knowledge resources, such as knowledge bases created by human labor (e.g. Wikipedia or Open Mind) or unanalyzed web data. Most importantly, query logs directly reflect people's search interests. They are also expressed in an extremely succinct manner, typically by a sequence of keywords, which has little syntactic structure. This makes query log data less noisy and easier to process than text

In order to acquire semantic knowledge from query logs, we focus on Named Entity (NE) categories like 'people', 'location', and 'organization', because it is well-known that such entities account a large potion of queries. While most work on NE recognition focuses on just 7-8 categories, we believe that a much larger and richer set of categories is needed for the actual search applications, which have to handle a wide variety of topics. Sekine et al. [1][2] describe an extended NE, which was designed to cover various topics. In this paper, we use this extended NE categories as our semantic categories.

WWW 2007, May 8–12, 2007, Banff, Canada.
ACM 978-1-59593-654-7/07/0005.

2. DATA
In this experiment, we start with 1.6 billion search queries. We used only ascii queries consisting of multiple keywords, which resulted in 750 million queries. This is equivalent to about 100 years' worth of newspaper text, assuming that we can extract four NE collocations from one sentence on average.

Our semantic categories, i.e. Sekine et al's Extended NE, include about 200 categories, with subcategories such as GPE (geo-political entities) and international region (e.g. continent, area) for location, as well as new categories such as product names (with subcategories such as food, clothing, car), event names, natural objects, colors, award names, and so on. The list contains about 250,000 entries, and was created by hand using text corpora, existing dictionaries and encyclopedias. For example, the category "award" has 641 entries, some of which are shown in Figure 1.

100 greatest Britons, 100 worst Britons, aaass/orbis books prize for polish studies, abel prize, academy award, academy awards, acm turing award, agatha award, Agatha awards, air medal, akutagawa prize

Figure 1. Examples of Award entries

3. ALGORITHM

3.1 Extraction of Typical Context
The list of NEs is matched against the query logs, and frequencies are counted in order to identify typical contexts. Figure 2 shows an example: it shows that "the academy awards" appears 202 times, "academy awards winners" appears 86 times, and so on (# indicates where the NE occurs in a query). We refer to the pattern in the rightmost column of Figure 2 as the *context* of a category.

202	academy+awards	the+#
86	academy+awards	#+winners
76	academy+awards	#+history
74	academy+awards	#+nominations

Figure 2. Examples of contexts for awards

The most frequent contexts for each category can also be very general ones which appear very often regardless of the category, such as "the+#", "#+pictures" and "#+history". We want to penalize these so that we can focus more on the contexts which are good discriminators for a particular category. This is a well-studied task of co-occurrence normalization, for which various solutions have been proposed, including TFIDF, mutual information and chi-square tests. However, as none of these metrics yielded satisfactory results on our task, we devised a new formula based on the type frequency of a context normalized by the total frequency of that context in the entire dataset. The

normalization factor is also standardized by the same factor on the category, which was estimated based on the top 1,000 most frequent entities. The most typical contexts identified by the scoring function are shown in Figure 3. It shows that these contexts include the category properties relevant for web search.

$$Score(c) = f\_type\{c\} * \log(g(c) / C)$$
$$g(c) = f\_type\{c\}/F\_inst\{c\}$$
$$C = f\_type\{ctop1000\} / F\_inst\{ctop1000\}$$

F_type: Frequency of context c in the category
F_inst: Frequency of context c in the entire data
ctop1000: 100 most frequent contexts

#+winners, #+nominees, #+nominations, #+winner, #+award, who+won+#, winners+of+#, list+of+#+winners, winners+of+the+#

Figure 3. Nine highest scoring contexts for the Awards category

3.2 Finding New Named Entities

The list of NEs is inevitably incomplete, as new terms are constantly being created. We find, however, that the acquired contexts and query logs are great resources for finding additional entities for a category. The basic idea is that terms which appear with the typical contexts but which are not already category members might be good additions. In our experiment, we used the 20 highest-scoring contexts and ranked new candidates by the number of distinct contexts with which they co-occurred. Figure 4 shows some results.

AWARD: golden globes, grammys, golden globe, kentucky derby, daytime emmy, sag, sag awards, american idol, daytime, emmys
BIRD: cardinal, eagle, bird, penguin, hawk

Figure 4. New words found for categories Award and Bird

The results look very promising. A new award name such as "American Idol" was found. For the Bird category, the original NE dictionary included specific names such as "Northern Cardinal" or "Red-capped Cardinal", but was missing general terms such as "cardinal" or "eagle". The proposed method was able to fill in this gap.

4. EVALUATION

Evaluating typical context directly is inevitably subjective and hence very difficult. In contrast, newly found NEs are much easier to evaluate, because human judges are asked only whether an NE belongs to a particular category or not. Also, this evaluation is a good indicator of the reliability of the scoring function, as the candidates are ranked based on the frequency of co-occurrence with highest scoring contexts. Figure 5 shows the evaluation results on the categories Book, Award and Bird. The evaluation was conducted on three groups of newly acquired words, each containing 20 words: given the 20 highest scoring contexts, the *Top* group contains words that occurred with these contexts most frequently; the *Middle* group with 6 of these contexts, and the *Bottom* group with 3. These newly found words are checked to see if they exist in Wikipedia; for books we also checked to see if they are in the top 10 Google results and in amazon.com. The figure shows a strong correlation between the number of co-occurrence of contexts and the accuracy of new word categorization, demonstrating that the scoring function works very well.

5. DISCUSSION

The method described here is related to bootstrapping methods to populate terms from small number of seeds [3][4]. However, these methods are vulnerable in that once a wrong name is learned, the process goes astray and will incorrectly assign words to categories. This problem is solved here by using a very large quantify of query logs and a larger list of named entities, as no iteration is necessary due to large and reliable term lists.

There are many other ways to use query logs as a resource for knowledge acquisition. One of the most important directions is to design categories. This must be left to future work, but the result reported here suggests that this line of research is promising.

6. ACKNOWLEDGMENTS

This research was conducted while the first author was at Microsoft Research as a visiting researcher. We would like to thank Bill Dolan, Mark Johnson, and the colleagues at Microsoft Research for making the study possible. Also, we would like to thank Prof. Ralph Grishman for his generous help.

7. REFERENCES

[1] Sekine, S, Nobata, C. "Definition, Dictionary and Tagger for Extended Named Entities" LREC 2004.

[2] Extended Named Entity definition and resources http://nlp.cs.nyu.edu/ene

[3] Brin, S. "Extracting Patterns and Relations from the World Wide Web", Workshop on the Web and Database 1998.

[4] Collins, M and Singer, Y. "Unsupervised Models for Named Entity Classification". EMNLP and VLC 1999.

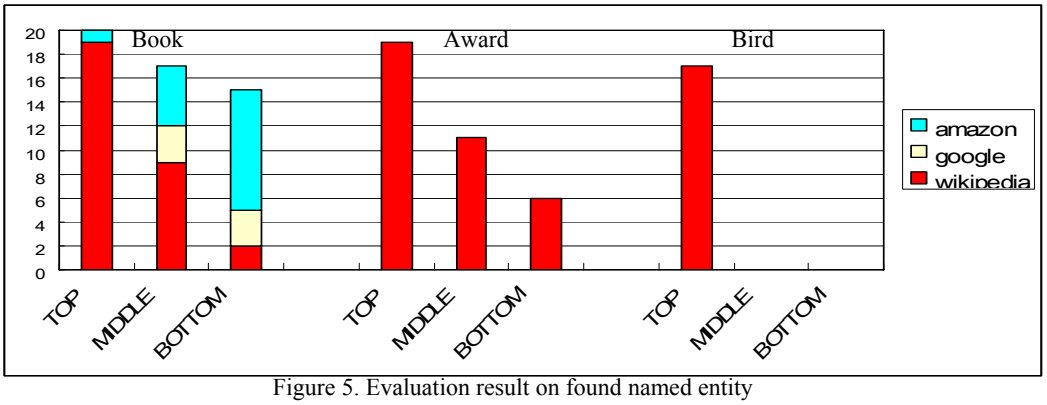

Figure 5. Evaluation result on found named entity