Statistical Analysis for Thesaurus Construction using an Encyclopedic Corpus

Yasunori OHISHI, Katunobu ITOU, Kazuya TAKEDA, Atsushi FUJII[†]

Nagoya University,
ohishi@sp.m., itou@, takeda@is.nagoya-u.ac.jp
[†] University of Tsukuba
fujii@slis.tskuba.ac.jp

Abstract

This paper proposes a discrimination method for hierarchical relations between word pairs. The method is a statistical one using an "encyclopedic corpus" extracted and organized from Web pages. In the proposed method, we use the statistical nature that hyponyms' descriptions tend to include hypernyms whereas hypernyms' descriptions do not include all of the hyponyms. Experimental results show that the method detected 61.7% of the relations in an actual thesaurus.

1. Introduction

This paper proposes a discrimination method for thesaurus construction using an "Encyclopedic Corpus." A thesaurus is useful for expanding the range of possible terms in information retrieval. However, if we try to retrieve technical information such as patents and technical articles, existent thesauri are inadequate because most technical terminology is not included in them. Though there are many methods to extract hyponyms, synonyms and hypernyms from specific expressions, such as sentences that have specific syntactic patterns ("a part of," "is a," and "such as")(Marti, 1992; Tsurumaru et al., 1991), descriptions in a dictionary(Suzuki, 2003), and a specific document structure (itemization in Web pages)(Shinzato and Torisawa, 2004), it is difficult to cover such an enormous vocabulary using these methods because less frequent words are not expected in the desired expressions. Robust estimation of the relations also requires a large variety of expressions because those expressions might include various words and viewpoints that assist with explanations.

To solve this problem, we use an Encyclopedic Corpus called the Cyclone corpus(Fujii et al., 2005), which covers an enormous vocabulary and contains many descriptions for each headword. The descriptions cover various meanings of the terms.

As an encyclopedic corpus, the Cyclone corpus is expected to be more effective for extracting hierarchical relations than common texts such as newspapers, because the contents of the Cyclone corpus have particular semantic relations between a headword and terms in its descriptions. As a corpus extracted from the Web, the Cyclone corpus covers more new terms and rarer terms due to the updating and creation of Web pages day by day. Since the quality of some documents on the Web is quite low, the Cyclone corpus was organized to exclude low-quality expressions.

2. The Cyclone Corpus

2.1. Description

The Cyclone corpus is constructed by extracting and organizing terms from 30 million pages, with the resultant corpus covering more than 700,000 terms in the encyclopedic Web-search site construction project(Fujii et al., 2005). The collection procedure for the Cyclone corpus comprises four modules: a term recognition module; a retrieval module; an extraction module; and an organization module.

The term recognition module periodically searches the Web for new morpheme sequences, which are used as target terms, while the retrieval module searches the Web for pages including a target term. The extraction module analyzes the layout (i.e., the structure of HTML tags) of the retrieved pages and identifies paragraphs that potentially describe the target term. While promising descriptions are extracted from pages resembling on-line dictionaries, descriptions can also be extracted from other types of pages, such as blogs.

The organization module classifies multiple paragraphs for the selected term into predefined domains (e.g., computers and medicine) and sorts them according to a probability score. Different word senses, which are often associated with different domains, are distinguished and high-quality descriptions are selected for each domain. The probability that paragraph d is selected as a description for domain c, $P(d|c)$, is transformed as in Eq. 1, by the Bayesian theorem

$$P(d|c) = \frac{P(c|d) \cdot P(d)}{P(c)}, \tag{1}$$

where $P(c|d)$ models the probability that d corresponds to c, and $P(d)$ models the probability that d is a description for the target term, disregarding the domain. We shall call them domain and description models, respectively. We regard $P(c)$ as a constant. The term $P(c|d)$ is modeled for 22 domains by a statistical categorization method. We decompose $P(d)$ into language, reliability, and layout properties, as shown in Eq. (2).

$$P(d) = P_L(d) \cdot P_R(d) \cdot P_S(d). \tag{2}$$

Here, $P_L(d)$, $P_R(d)$, and $P_S(d)$ respectively denote language, reliability, and layout (structure) models, respectively; where $P_L(d)$ is a trigram language model produced from a machine-readable encyclopedia.

For P_R, we implemented a software to compute PageRank, which is used in Google[1], to rate the quality of Web pages based on hyperlink information, and $P_R(d)$ is the PageRank value for the page from which d was extracted. If d is extracted from a page whose HTML layout is similar to one typically used to describe terms, $P_S(d) = 1$. Otherwise, $P_S(d) = 0.5$. The HTML layout for a page is obtained using the extraction module.

2.2. Evaluation

To evaluate the Cyclone corpus, we collected test terms from the index of a printed terminology dictionary, which lists 2,226 technical terms that frequently appear in the Information Technology Engineers Examinations. We performed two experiments using 2,074 terms for which at least one paragraph was obtained.

In the first experiment, we evaluated the effectiveness of the reliability, layout, and language models in sorting paragraphs. For each test term, paragraphs were sorted according to PageRank and at most the top 500 paragraphs were manually judged as to whether they were a correct description for the term in question. The average number of paragraphs judged per term was 141. In addition, each correct paragraph was manually annotated with one or more domains.

We used three evaluation measures. The first was mean average precision (MAP), which is a combination of recall and precision and has commonly been used to evaluate information retrieval. MAP becomes high if many correct paragraphs are sorted into high ranks for each test term. This measure is important if a user requires more than one correct description for a single term. Second, we used mean reciprocal rank (MRR), which has commonly been used to evaluate the quality of question answering. For each test term, we calculate the reciprocal of the rank at which the first correct paragraph was found. MRR is the mean of the reciprocal ranks for all test terms. This measure is important if a user requires only one correct description. Third, we used the average rank at which the first correct paragraph was found. Unlike MRR, this value, which we shall call "RANK," is in proportion to the rank. Whereas a high MAP and MRR are preferable, a low RANK value is desirable.

Table 1 shows MAP, MRR, and RANK for different combinations of the reliability (R), layout (S), language (L) models. "R" simulates a conventional search engine and is a baseline. The layout and language models were independently effective, and when used together the improvement was even greater, regardless of the evaluation measure.

In the second experiment, we used 1,472 of the test terms, to which one or more correct descriptions and domains were manually annotated, and evaluated the effectiveness of the domain model in categorizing paragraphs for these terms. We regarded only the top domain determined by the domain model as the system output. The recall and precision were 0.671 and 0.700, respectively. Thus, approximately 70% of correct descriptions can be found in correct domains.

[1] http://www.google.com/

Table 1: Effectiveness of sorting paragraphs.

	R	RS	RL	RSL
MAP	.204	.247	.410	.433
MRR	.280	.436	.595	.639
RANK	28.6	21.3	9.7	7.5

3. Hierarchical Relation Estimation Method

Descriptions of words in an encyclopedia have "directionality." For example, a description of a "lion," which is "a large animal of the cat family," includes hypernyms such as "animal" and "cat." A description of an "animal," however, may not always include a "lion" as "a living creature such as a dog or cat." Generally speaking, a collection of descriptions for a headword shares common hypernyms but may not share hyponyms, because the use of hyponyms in a description depends on the viewpoint of the explanation strategy.

Consequently, we propose a probabilistic method using the directionality of descriptions. A target function is $H(X|Y) > \epsilon$, where X is a hypernym of Y and ϵ is an arbitrary small value. We define $H(X|Y)$ as follows:

$$H(X|Y) = C(X|Y) - C(Y|X), \tag{3}$$

where $C(X|Y)$ is a probability of X, given the descriptions of Y.

For accurate calculation of $C(X|Y)$, the encyclopedic corpus is effective because it covers various descriptions from many perspectives. To calculate $C(X|Y)$, a semantic expansion technique is also used in order not to miss an indirect relationship. In the above example, although the description of "animal" does not include "lion," a hierarchical relation can be found indirectly if a description of the cat family is "the cat family includes lions and tigers." The probability over all words is defined as a square matrix:

$$A_{i,j} = P(w_i|w_j), \quad i,j = 1 \ldots m, \tag{4}$$

where m is the number of words and $P(w_i|w_j)$ is defined as the relative frequency,

$$P(w_i|w_j) = \frac{F(w_i|w_j)}{\sum_{k=1}^{m} F(w_k|w_j)}, \tag{5}$$

where $F(w_i|w_j)$ is the frequency of word i in the descriptions of word j. The nth-order probability is defined as a square matrix A^n. Finally, the expanded probability is defined as a linear combination of probabilities:

$$C = \sum_{i=1}^{N} [\alpha_i A^i], \tag{6}$$

where N is a maximum order of expansion; α_i could be determined using Linear Discrimination Analysis (LDA)(Hastie et al., 2001).

The proposed method can therefore find indirect relations. This is important for thesaurus construction, because using only direct relations is inadequate for constructing a large hierarchy.

4. Evaluation

4.1. Experimental Setup

To evaluate the proposed method's feasibility of discriminating the hierarchical relation of word pairs, we evaluated its ability to discriminate various word pairs in a technical domain. We used test terms from the computer-related domain (Sec. 2.2.).

To determine whether the relations were correct, we compared them with the results extracted from a manually constructed thesaurus(the JICST thesaurus), which includes about 43,000 words for searching scientific literature and covers 172 of the test terms.

Adding to the correct relations, we manually constructed disrelation pairs. We randomly extracted 500 pairs from the pairs of test terms, reducing this total to 497 pairs after checking whether each pair had any hierarchical relation.

Next we manually judged the correctness of the terms. On average, descriptions for each term contained 6.62 correct paragraphs, referred to as A, 10.4 correct and partially correct paragraphs, referred to as A+B, and a total of 80.7 paragraphs (including incorrect paragraphs), referred to as ALL. The evaluation of the proposed method should be referred to the result using ALL and the other test set were references.

The correct relations from the JICST thesaurus covered 136 pairs for A, 168 pairs for A+B, and 206 pairs for ALL, and the disrelation pairs covered 301 pairs for A, 366 pairs for A+B, and 497 pairs for ALL.

This paper reports only correct rates for these pairs; the JICST thesaurus may not cover all the hierarchical relations in the 2,074 word-IT vocabulary, and we had not checked the other relations that were not included the test set.

4.2. Directional Occurring Model

In the proposed method, the directional occurring model only checks whether a word pair has a hierarchical relation. However, since this model can discriminate which word is a hypernym (or hyponym), we evaluated both functions. We conducted LDA for three test sets A, A+B, and ALL, after conducting a four-fold cross-validation. For comparison, we also evaluated an exponential expansion method(Suzuki, 2003) using the following expansion,

$$C = \lim_{k \to \infty} b(aA + a^2A^2 + \cdots + a^kA^k), \quad 0 < a < 1, \quad (7)$$

where b is a normalization coefficient. We trained the weights with half the test sets and evaluated them using the other half because, in the exponential method, the best weight can only be found experimentally. The results are shown in Fig. 1.

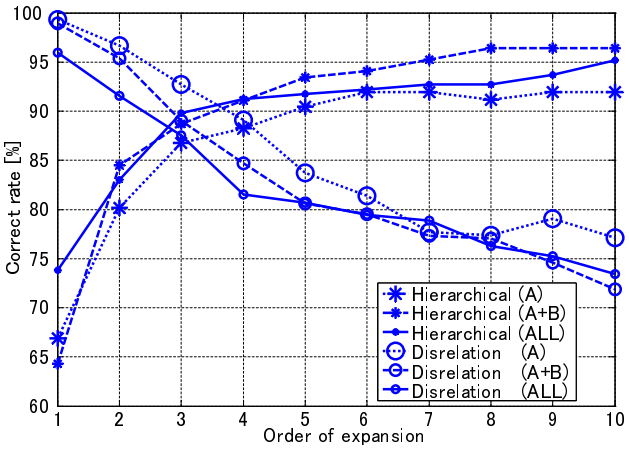

Figure 1: Correct rate for the relation check, as a function of expansion order

The correct rate of the hierarchical relation discrimination decreased as the order of expansion rose and the disrelation discrimination declined. The correct rate for the hierarchical relation was better for the A set, while that for the disrelation was better for the ALL set. These results were reasonable because the word-occurrence distribution for the disrelation pairs was as same as the general distribution, thus the distribution was spoiled by a smoothing effect of expansion. For the hierarchical relations depending on word pairs, more training data is needed to achieve a higher correct rate.

Table 2: Average performance

test set	A	A+B	ALL
proposed (%)	88.2	88.7	90.8
(order)	4	3	3
exponential (%)	88.5	88.0	86.7
(hierarchical) (%)	80.4	80.9	76.9
(disrelation) (%)	96.5	95.1	96.4
(weight)	0.6	0.5	0.3

Table 2 shows the average correct rate of the proposed method at the balance point between the hierarchical relation discrimination and the disrelation and the average correct rate of the exponential expansion method. For ALL, the proposed method achieved the highest performance of 90.7% whereas the exponential method achieved the lowest one of 86.7%. The reason is that though the exponential expansion method achieved a high correct rate for the disrelation, it achieved a low one for the hierarchical relation, because the exponential expansion method sets a higher value on the first-order matrix than the proposed method does as shown in Fig. 2, where the weights are normalized by each first-order weight. This is why it is difficult to improve the correct rate for the hierarchical relation.

Figure 3 shows the result for the proposed method using only the top correct document, where the result for A is not shown because A cannot obtain any result due to its sparseness. A+B and ALL achieved only low correct rates

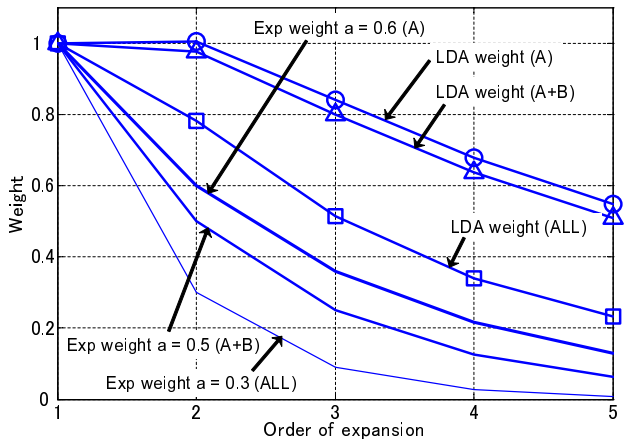

Figure 2: The weights for the expanded matrices

for the hierarchical relation whereas both of them achieved high correct rates for the disrelation.

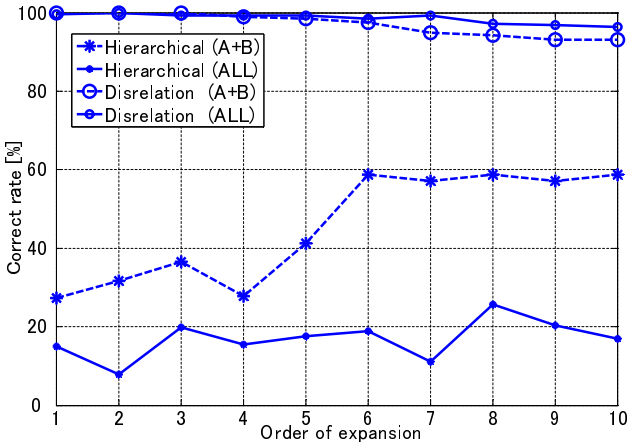

Figure 3: Correct rate for the relation check using single document, as a function of expansion order

The results reveal that multiple descriptions and correct descriptions were both effective, and that the number of descriptions compensated for the correctness of the descriptions. Therefore, the encyclopedic corpus was effective for extracting the hierarchical relation between word pairs, and the correct descriptions were effective.

Figure 4 illustrates the hierarchical discrimination performance only using the directional occurring model. Table 3 shows the highest performance for each test set. A+B achieved the highest correct rate of 66.1% at the seventh order, and ALL achieved a lower rate of 61.7%; therefore, for the hierarchical discrimination, correct documents were effective. The exponential method achieved lower correct rates of less than 60%, because also in this case the weights are too small at the higher orders of expansion.

5. Conclusion

We proposed a statistical estimation method to discriminate the hierarchical relation of a word pair using an encyclopedic corpus called the Cyclone corpus. It reflects both a

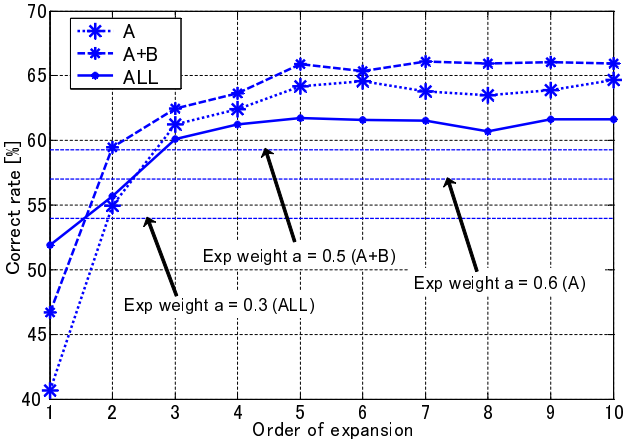

Figure 4: Correct rate for the hierarchical relation as a function of expansion order

Table 3: Highest correct rate for the hierarchical relation

test set	A	A+B	ALL
proposed(%)	64.6	66.1	61.7
(order)	6	7	5
exponential(%)	57.0	59.3	54.0

global term distribution and local statistical syntactic patterns, and was able to detect 61.7% of the relations in the IT-vocaburary section of the JICST thesaurus. The experimental results showed the Cyclone corpus to be effective for estimating hierarchical relations in the vocabulary because it contains more than 80 descriptive paragraphs on average.

6. Acknowledgment

This research was partially supported by Industrial Technology Research Grant Program in '05 from NEDO (Japan).

7. References

A. Fujii, K. Itou, and T. Ishikawa. 2005. Cyclone: An encyclopedic Web search site. In *WWW-2005*, pages 1184–1185, May.

T. Hastie, R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning -Data Mining, Inference and Prediction-*. Springer.

A. Hearst Marti. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING-1992*, August.

K. Shinzato and K. Torisawa. 2004. Extracting hyponyms of prespecified hypernyms from itemizations and headings in Web documents. In *COLING-2004*, pages 938–944, August.

S. Suzuki. 2003. Probabilistic word vector and similarity based on dictionaries. In *CICLing-2003*, pages 564–574.

H. Tsurumaru, K. Takeshita, K. Itami, T. Yanagawa, and S. Yoshida. 1991. An approach to thesaurus construction from Japanese language dictionary. *IPSJ-SIGNL*, 1991(37), May.