2-4 Acquisition of Taxonomic Relations among Words from Huge Corpora and its Application

KANZAKI Kyoko, YAMAMOTO Eiko, and ISAHARA Hitoshi

Thesaurus is very important lexical knowledge for our inference activity. However, we have only thesaurus compiled by human because we didn't have huge corpora and the algorism to organize concepts using such corpora.

For sake of a verification of an existing thesaurus made by human, we automatically extract lexical knowledge from huge corpora. In our method, we extracted attribute concepts whose instances are adjectives from corpora and calculated similarity relations by Self-Organizing Map and hypernym-hyponym relations by Complimentary Similarity Measures. As a result, we constructed the taxonomic relations of attribute concepts of adjectives on a map. Also we applied our methods to extract related word sets which can be useful for retrieval support. Concretely, in order to extract word sets with thematic relation, we extract related word sets with non-taxonomical relation. Then, we verified the effectiveness of such word sets as key words for information retrieval.

Keywords
Thesaurus, Taxonomic relation, Self-organizing map, Thematic relation, Retrieval support

1 Introduction

The importance of lexical systematization

The goals of this research are to automatically acquire system of Japanese lexical concepts from a corpus and to demonstrate the effectiveness of the lexical relations thus extracted for use in applications such as information retrieval.

Dictionaries systematizing lexical semantic relations form a very important foundation for the extraction of needed information, making use of efficient induction by computer. Specifically, a dictionary structuring lexical semantic relations is one that structures a wide range of information such as similarity relations and hierarchical relations between words, relations between parts and the whole, and ownership relations. Armed with this sort of information, it becomes possible to use a given word as a clue to discover related words. For example, the word “automobile” is related to the following types of information:

Hypernym relation: vehicle, etc.
Hyponym relation:
compact car, full-sized car, standard car
Further hyponym relations:
Toyota, Nissan, BMW, etc.
Similarity relations:
train, bicycle, airplane, boat, etc.
Part relations:
tires, steering wheel, engine, doors, etc.

Given the presence of the foregoing types of information, when a person says “I want to buy a car” we can imagine candidates in hyponym relations to “car” concerning what the speaker may wish to purchase. Likewise, when somebody says “How about going by

car?" we can think of other means of transportation in similarity relations to "car", by thinking "Wouldn't another means of transportation be better?" Further, when a person says "My car broke down" we can naturally wonder which part might have broken.

Being able to structure the semantic relations between words in this way is a crucial basic capacity for computers, just as this ability is fundamental to efficient induction by human beings.

What is a thesaurus?

A thesaurus refers to a dictionary that systematizes the relations between the meanings or concepts of words.

Semantic relations aren't limited to nouns; verbs and adjectives bear these relations as well. For example, "red" and "white" can be grouped together in a similarity relation by their shared nature as colors, while "big" and "small" can be grouped together in a similarity relation by their shared nature as sizes.

In general, groups defined in terms of comprehension (i.e., with natures shared among their constituent members) are referred to as concepts, while those defined in terms of extension (i.e., as groups of their constituent members) are referred to as categories[1].

In defining red and white by their shared nature, we can consider color to be a concept. Likewise, for big and small size can be considered a concept. Viewed another way, red and white can be considered categories of color and big and small categories of size.

A thesaurus, which systematizes relations between concepts such as color and size using taxonomic relations (taxonomy), is used in computer processing that handles words not as superficial strings of text but by using their meanings.

Thesauri to date

Large-scale thesauri have already been constructed in the field of natural-language processing. Examples include dictionaries such as the EDR Electronic Dictionary dictionaries distributed by NICT, the Word List by Semantic Principles from the National Institute for Japanese Language, and NTT's Nihongo Goi Taikei.

These thesauri, which serve to structure Japanese-language lexicons, have been constructed by a number of people over several years. As they grew larger they came to include some questionable content, requiring verification and correction. However, with the exception of relatively simple errors that can be corrected automatically, in their current state it is not possible to make modifications, corrections, or other changes involving the structures of the thesauri themselves.

Our research

Language-processing technologies have advanced in recent years with the increasing ability to access large volumes of usable text data. Accordingly, we are now experimenting with automatic conceptual systematization using actual texts. The ability to extract conceptual systems automatically from actual large-scale text data would in theory enable verification of large-scale thesauri traditionally created by hand and to consider sections requiring correction and similar issues. Further, the refinement of methods of automatic extraction from a corpus would enable the extraction of linguistic knowledge from unfamiliar large-scale compilations of data as well. For example, it would even be possible to automatically construct thesauri of specialized terms — structuralizing such terms in fields such as medicine, the life sciences, law, and patents, fields in which the demand for such thesauri are increasing every year. We believe that constructing thesauri of such terms could assist in extracting needed information from large-scale electronic documents in these and other specialized fields. Our current experiments examine applications in the area of medical terminology.

2 Self-organizing map in a neural-network model, adopting directed measures of similarity

We are currently examining full conceptual structuring based on adjectival concepts

extracted from text, for the purpose of verifying existing linguistic resources consisting of thesauri constructed by hand. This structuring method makes use of self-organizing maps (SOMs) [2] to extract a thesaurus automatically from a large-scale corpora. In other words, using the examples given above, we are attempting to extract from a corpus the hypernym concept "color" for the adjectives "red" and "white" and the hypernym concept "size" for "large" and "small" and then automatically structuring terms such as color and size that express the adjectival concepts. In describing the formulas used in this approach, for convenience the adjectival concepts such as color and size extracted from the data are referred to simply as "words".

Under our approach, in encoding the data input to the SOMs, we first calculate similarity measures to derive directionality — such as the hypernym-hyponym relations — with respect to the semantic distances between two words. As a result, self-organization on the maps distributes not just similarity relations between concepts but also hypernym-hyponym relations.

2.1 Input data

In order to extract from a corpus abstract nouns that can be used to categorize adjectives, we searched the corpus for semantic relations between nouns used to categorize adjectives and collected data on these nouns [3][4]. The approach we took was to use the sentence pattern "X toiu Y", a pattern by which X categorizes Y [5], as a hint in extracting from the corpus patterns in which X was the adjective and Y the abstract noun. From this data, to some degree we chose corresponding adjectives and nouns by hand.

As concept names for adjectives, we extracted abstract nouns Y from *Mainichi Shimbun* newspaper articles for the two-year period 1994 to 1995. We searched for sample usages of adjectives and adjective verbs co-occurring with the abstract nouns from 11 years' worth of *Mainichi Shimbun* articles, 10 years of *Nihon Keizai Shimbun* articles, seven years of *Nikkei Sangyo Shimbun*, *Nikkei Kinyu Shimbun*, and *Nikkei Ryutsu Shimbun* articles, 14 years of *Yomiuri Shimbun* articles, and 100 essays and 100 novels.

We extracted 365 abstract nouns and 10,525 different adjectives, from a grand total of 35,173 extracted words. The largest number of co-occurrences was 1,594, involving situational words. Sample data is shown below.

[Ex.]

Thoughts : Happy, fun, sad, etc.
Feelings : Fun, happy, joyful, etc.
Perspectives : Medical, historical, scientific, etc.

2.2 Encoding the input data

We encoded the data input to the SOMs [6] as described below.

In general, we assumed that maps would be constructed for a number of ω types of nouns w_i ($i = 1, \ldots, \omega$). As a specific example, consider mapping data input for thoughts (e.g., happy, proud, sad, etc.). In this case, the noun wi is defined in terms of a set with co-occurring adjectives, as shown below:

$$w_i = \{a_1^{(i)}, a_2^{(i)}, \ldots, a_{\alpha i}^{(i)}\}$$

In this case, $a_j^{(i)}$ represents the adjective numbered j co-occurring with w_i, while αi represents the number of adjectives co-occurring with w_i. We used correlated coding to encode these adjectives. Correlated coding seeks to reflect the semantic correlations (or semantic distances) between nouns.

Each individual d_{ij} represents the relation between the two nouns w_i and w_j. When considering other nouns for reference purposes, it is not possible to reflect the relations between these two nouns themselves or between these two nouns and other nouns solely as a collection of these d_{ij} values. Instead, these values represent limited semantic relations. However, by creating a grid as shown in Table 1 based on these individual limited semantic distances, one can see that each row consists of limited semantic distance from w-1 nouns, not including the correlation between identical nouns. In other words, each row can be considered to

Table 1 Correlative matrix of nouns

	w_1	w_2	...	w_ω
w_1	d_{11}	d_{12}	...	$d_{1\omega}$
w_2	d_{21}	d_{22}	...	$d_{2\omega}$
⋮	⋱			
w_ω	$d_{\omega 1}$	$d_{\omega 2}$	...	$d_{\omega\omega}$

reflect the semantic relation between a given noun and all other nouns.

Using the correlated coding method proposed here thus enables us to encode multidimensional vectors as shown below, using this grid for each noun w_i.

$$V(w_i) = [d_{i1}, d_{i2}, \ldots, d_{i\omega}]^T$$

Above, $V(w_i)$ represents input to the SOM. Self-organization of this multidimensional vector results in an expression of the two-dimensional space actualized in the semantic relationship among words.

2.3 A complementary similarity measure for deriving hypernym and hyponym relations between two words

For d_{ij}, the semantic distance between two words, we used a complementary similarity measure effective for deriving hypernym-hyponym relations between two words[7].

Assume we have at present two abstract nouns F and T, defined as sets with co-occurring adjectives. In our data, the characteristic vectors of F and T correspond to the expression of the state of appearance of their co-occurring adjectives as 0 or 1.

The above example is expressed as shown below:

$$\vec{F} = (f_1, f_2, \ldots, f_i, \ldots, f_n)(f_i = 0 or 1)$$
$$\vec{T} = (t_1, t_2, \ldots, t_i, \ldots t_n)(t_i 0 or 1)$$

Next, the complementary similarity measure is expressed using the following formula:

$$Sc(\vec{F}, \vec{T}) = \frac{ad - bc}{\sqrt{(a+c)(b+d)}}$$

Here, "a" represents the number of adjectives co-occurring with both F and T, "b" the number of adjectives co-occurring with F but not with T, "c" the number of adjectives co-occurring with T but not with F, and "d" the number of adjectives co-occurring with neither F nor T. If T is fully a subset of F, then c = 0. Likewise, if F is fully a subset of T, then b = 0, and thus bc = 0. Since the degree of complementary similarity measure is the difference between matching (ad) and non-matching (bc) data, the degree of similarity between two words in an inclusion relation will be high.

Another characteristic of this measure is the asymmetrical relation between the degree of complementary similarity measure from F to T and from T to F. In the degree of complementary similarity measure seen from F to T, b represents the number of adjectives appearing with F only, while c represents the number of adjectives appearing with T only. Likewise, in the degree of complementary similarity measure as seen from T to F, b represents the number of adjectives appearing with T only and c represents the number of adjectives appearing with F only. A look at the denominator in the formula shows that when calculating the degree of similarity in either direction between F and T, the relative size of the values entered in place of b and c will reverse, resulting in asymmetrical degrees of similarity.

We used the complementary similarity measure value between two words as input data for self-organization, calculating values in the correlation by entering these in place of semantic distance d_{ij}, as discussed in section **2.2** above.

2.4 Constructing hierarchical relations for concepts

Based on the results obtained from the complementary similarity measure, we arranged all words in hierarchical relations from highest to lowest[17] and plotted these relations on the map. In doing so, we employed the following steps:

(1) We linked words A and B to higher levels

of similarity, indicating inclusion relations. In doing so, we assumed word A was the hypernym and word B the hyponym.

(2) First, we repeated connection in a downward (rearward) direction using A-B as the starting point, using word B as the hypernym and searching for word Y, the highest-valued hyponym, and then connecting word Y behind B. Next, with word A as the hyponym, we searched for word X, the highest-valued hypernym, and connected it in front of word A. We repeated upward (forward) connection in the same manner, with A-B as the starting point. At the same time, we made sure to save the hypernym-hyponym relations. When a hypernym-hyponym relation broke down, we did not connect the corresponding words. We constructed a single hierarchy in this manner.

(3) We merged short hierarchies contained fully within long hierarchies. When two hierarchies differed by only one word, if the complementary similarity measure between the two differing words showed a hypernym-hyponym relation, we combined these in accordance with this relation.

(4) Last, we combined the highest-ranking "situations" on each hierarchy. A situation can be considered to be the most abstract concept, capable of co-occurring with all adjectives. To save computation time, we combined situations with the final highest ranking on each hierarchy. Thus in the end we were able to extract hierarchies consisting of abstract nouns, with situations as the highest-ranking corresponding concepts.

3 SOM accounting for hierarchical relations of adjective abstract concepts

We constructed a SOM reflecting the hierarchical structure derived through the steps described above. The concept behind this structure involves ranking the abstract level at the top, with specific nouns distributed at lower levels. Figure 1 shows the concept hierarchy related to emotions on the map. At right is a sample concept hierarchy of adjectives, displaying states such as natures and conditions in addition to emotions.

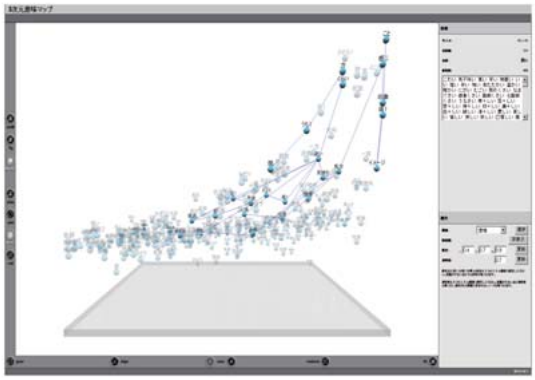

Fig. 1 *SOM of adjective-attribute concepts reflecting hierarchical relations*

The hierarchy ranks concepts from highest to lowest.

Kind

Situation – screen 1 – image – impression – thought – feeling – emotion – affection – consideration

Open, stubborn, advanced, bright, sincere

Situation – screen 1 – personality – temperament – disposition – nature

Difficult, dominant, threatening, desperate

Situation – condition – state of affairs – prospects

Ex *Conceptual hierarchy of emotions from the perspective of adjectives*

4 Comparison of automatically constructed hierarchy and EDR electronic dictionary hierarchy

We automatically constructed three hierarchies using the methods — (1) complementary similarity measure, (2) overlap coefficient, (3) complementary similarity measure made by taking into consideration the frequency. We

compared each of the hierarchies with that of the EDR dictionary constructed by hand. We used a total of 20 test subjects, consisting of linguists, of others involved in the field of natural-language processing, and dictionary editors. We conducted psychological experiments using Scheffe's method of paired comparisons[8]. As a result, we determined that 43% of the hierarchy constructed automatically using all three methods was either more appropriate than the EDR hierarchy or no less appropriate, when tested at a significance level of 1% using the t statistic. In addition, our assessment shows that the hierarchies created individually using each method were not more appropriate than the EDR hierarchy. In the future, we plan to contribute further to thesaurus verification by refining methods of structuring taxonomic relations of abstract concepts related to adjectives, to construct automatically structures appropriate for taxonomic relations.

Section **5** below covers research conducted by applying our approach to constructing hierarchical relations between words to specialized terminology.

5 Applying automatic hierarchical-structuring method to extracting sets of related words

Next, we will consider the automatic hierarchical structuring method[17] shown in Section **2.4** above. In addition to hierarchical relations, it is expected that collections of related words extracted from a corpus will prove beneficial in uses such as language processing, language generation, and information retrieval. Today, while many methods are being developed to extract from corpora various relations between words[9]-[13], methods of learning patterns to extract relations have also been proposed[14][15]. Collections of related words are useful in information retrieval, guiding users to informative pages. It is conceivable that words related to those input by a user could also be provided, as with Google's retrieval-support functions. However, it might also be possible to help the user arrive at the informative pages by showing how these other words are related to the input words. To this end, we attempted to extract sets of related words from documents by applying automatic hierarchy-structuring methods, from the viewpoint of assessing the effectiveness of the extracted word sets as keywords in information retrieval. We also analyzed the effectiveness of these sets of related words as keywords.

5.1 Relations between words

In supporting retrieval, in what kinds of relations should added keywords be positioned in order to achieve effective results? At the very least, words relate to each other in two ways: through taxonomical relations and through thematic relations. It has been reported that these relations are important in recognizing and understanding the relations between words[16].

Taxonomical relations refer to relations indicating the taxonomy between the attributes of a concept. Examples include relations between words like "horse", "cow", and "animal". This category of taxonomical relations includes semantic relations such as synonym relations, antonym relations, and hierarchical relations. On the other hand, thematic relations refer to relations connecting concepts by thematic situations. Examples include situations brought to mind by words, such as "milking a cow" for the words "cow" and "milk" and "feeding a baby" for the words "baby" and "milk", or relations connecting concepts in such situations. Thematic relations include associated relations, causal relations, and entailment relations.

Related words added to support retrieval often use words in a taxonomical relationship to the keywords entered, for the purpose of replacing these entered words with better keywords. These related words are covered directly by existing dictionaries and thesauri. A primary factor in using these words is the relative

ease of obtaining and using them. However, sometimes the search results are not narrowed down in a useful way, but instead include unintended items. Use of thematic relations promises to provide the user with new information by narrowing down search results to those that are more relevant, through the addition of related words based on relations between words connected with the content in the pages. With the foregoing aim in mind, this study focuses on thematic relations. We extracted sets of related words considered to bear thematic relationships and investigated the effectiveness of the terms making up these word sets in supporting retrieval.

5.2 Extraction method

To extract word sets in thematic relation, we took the following steps: 1) preparation of experimental data by collecting dependency relations from the documents, 2) extraction of sets of related words using automatic hierarchical structuring, and 3) selection of word sets in non-taxonomic relationships, using a thesaurus.

5.2.1 Collecting co-occurrence relations

We parsed the collection of documents to collect from each sentence dependency relations corresponding to the following patterns: "B <of> A", "(conduct) V <on> P", "Q <does> V", "(conduct) V <on> R", and "S <is> V". Here, <X> represents a particle, A, B, P, Q, R, and S are nouns, and V is a verb. From the relations collected, we prepared three types of data: specifically, NN data based on co-occurrence relations between nouns, NV data based on dependencies between nouns and verbs (for each particle), and SO data based on relations between subjects and objects.

5.2.2 Extracting sets of related words

We will expand the automatic hierarchical-structuring method proposed in this paper to extract sets of related words. This method estimates relations between words for each pair of words from inclusive relations of individual co-occurring words and appearance patterns. Since the extraction of semantic relations between words described through the previous section was intended to extract hierarchical structures, the co-occurring words used were limited to hyponyms of each word. In this section, instead of limiting co-occurrence relations to hierarchical ones, we will address dependency relations for individual data elements as described above. This will allow us not only to extract hierarchical structures but also sets of words bearing other relations.

5.2.3 Selecting sets of related words in thematic relations

Last, we will extract collections of related words in thematic relations by using a thesaurus to remove sets of related words in taxonomic relations from the extracted word sets. Since in general the words included in a thesaurus are distributed in ways that express their taxonomic relations, sets of related words in taxonomic relations are classified into the same categories in a thesaurus. In other words, if a set of related words matches the classification in a thesaurus, then we can interpret this to mean that the words making up this word set are in a taxonomic relation. In accordance with this thinking, we will remove sets of related words matching the thesaurus and extract the remaining sets of related words in non-taxonomic relations as word sets in thematic relations.

5.3 Experiment

In this experiment, we used a collection of documents (10,144 pages, 225,402 sentences) limited to the medical domain. In Japanese-language analysis, we did not use tools such as medical or other specialized dictionaries. The data prepared from the sets of related words collected from this group of documents totaled 225,402 cases of NN data, 20,234 cases of NV data for "wo cases", 15,924 cases of NV data for "ga cases", 14,215 cases of NV data for "ni cases", 15,896 cases of NV data for undefined cases, and 4,437 cases of SO data. We used the Medical Subject Headings (MeSH®) thesaurus, using Japanese translations of its headwords and synonyms included as cross references as the medical terms of which the

extracted set of related words would consist. Of these, 2,557 words appeared in our experimental data.

Figure 2 shows some of the sets of related words extracted. Of the extracted word sets, those consisting of three or more terms were subject to the next selection process.

5.4 Analysis

We used Google searches to investigate the effectiveness of the extracted sets of related words in searching — in other words, to check that search results were limited to useful web pages. We chose sets of related words consisting of terms divided into two categories so that only one term was classified into a category different from that of the other words. Of the 847 sets of related words extracted, 294 fit this description. If we depict the sets of related words subject to this investigation as $\{X_1, X_2, ..., X_n, Y\}$, the X_i terms represent terms classified into the same category, while Y represents a term classified into a category different from that of the X_i terms. In this step, we created the following three types of search keywords from each of these sets of related words:

- Type 1: $\{X_1, X_2, ..., X_n\}$, not including Y, which was classified into a different category
- Type 2: $\{X_1, X_2, ..., X_{k-1}, X_{k+1}, ..., Xn\}$, not including Y or one term X_k in the same category
- Type 3: $\{X_1, X_2, ..., X_{k-1}, X_{k+1}, ..., X_n, Y\}$, not including one term X_k in the same category

ovary - spleen - palpation
neonate - patent ductus arteriosus - necrotizing enterocolitis
secretion - gastric acid - gastric mucosa - duodenal ulcer
skin - atopic dermatitis - herpes viruses - antiviral drugs
skin - abdomen - cervix - cavitas oris - chest
fatigue - uterine muscle - pregnancy toxemia
water - oxygen - hydrogen - hydrogen ion
fatigue – stress – duodenal ulcer
latency period - erythrocyte - hepatic cell
snow - school - gas
variation - death - limb
hospitalist - corneal opacities - triazolam
cross reaction - apoptoses - injuries
research - survey - altered taste - rice
environment - state interest - water - meat - diarrhea
rights - energy generating resources - cordia - education – deforestation

Fig.2 *Some sets of related words extracted*

These three types are based on the keywords of Type 2 — i.e., the keywords input originally — . Type 1 is a set of keywords including added keywords classified into the same category as in Type 2. The added keywords are characterized by their frequencies within the documents used in this study (i.e., high or low frequency) and are terms related taxonomically to the terms in Type 2. Type 3 is a set of keywords with terms added that are in different categories from Type 2. These added terms can be considered to be in thematic — or non-taxonomical — relations with the terms in Type 2.

First of all, we will use the Google search engine's estimated number of page hits to compare search results quantitatively. Specifically, based on the number of page hits obtained using Type 2 terms, we will compare numbers of page hits obtained using Type 1 and Type 3 terms — each of which adds one term to the Type 2 terms. Figures 3 and 4 show the results of comparing numbers of page hits related to high and low frequency for each. In these figures, the horizontal axis shows the number of page hits using the base (Type 2) keywords, while the vertical axis represents the number of page hits when adding one term to the base keywords (Type 1 or Type 3). In these figures, "o" represents the number of page hits when adding a term in the same category (Type 1) and "x" represents the number of page hits when adding a term in a different category (Type 3). The diagonal line represents a hypothetical case in which adding a term to Type 2 had no effect on number of page hits.

Figure 3 shows that most x's are substantially below the diagonal line. This illustrates the tendency for the addition of non-taxonomically related terms in different categories to reduce the number of page hits, compared to the addition of high-frequency taxonomically related terms in the same category. We can therefore conclude that adding non-taxonomically related terms is effective in quantitative terms for useful page searching, and that such non-taxonomically related terms are more

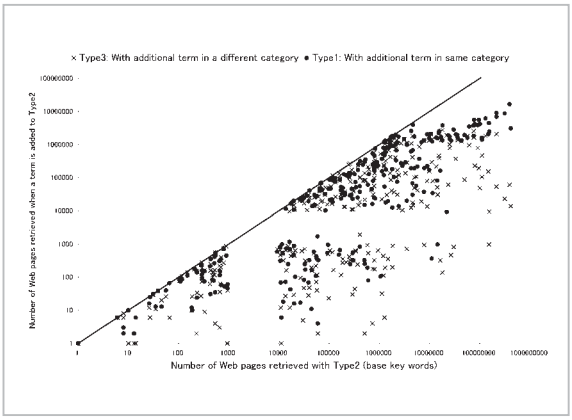

Fig.3 *Variation in number of page hits when adding high-frequency terms and terms in different categories*

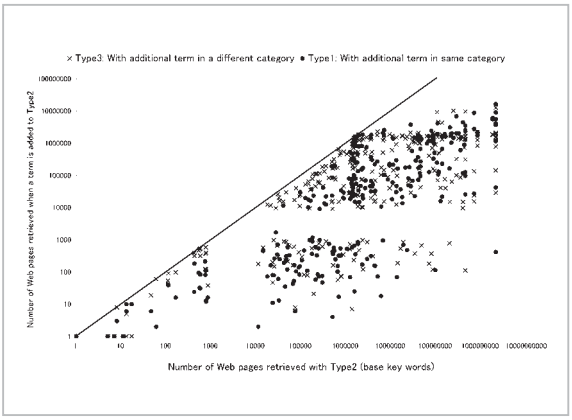

Fig.4 *Variation in number of page hits when adding low-frequency terms and terms in different categories*

beneficial as terms than high-frequency taxonomically related terms. In contrast to Fig. 3, Fig. 4 shows that most o's are substantially below the diagonal line. A look at these groups of related words shows that in most cases the added taxonomically related terms had the lowest frequencies in their groups of related words. This illustrates the tendency for the addition of low-frequency terms to reduce the number of page hits, compared to the addition of non-taxonomically related terms. In fact, low-frequency terms are uncommonly used even on the Internet, so one could predict that the number of web pages including such terms would itself be small. For this reason, adding low-frequency terms is quantitatively effective in terms of search results, regardless of the type of relation among the terms. However, examination of the content of results when adding non-taxonomically related terms and low-frequency taxonomically related terms shows a considerable difference between the two approaches.

As one example, consider the following set of related words obtained from SO data: "latency period — erythrocyte — hepatic cell". In this example, "latency period" is a term grouped in the MeSH thesaurus under a category that is different from that of the other terms, while "hepatic cell" is a low-frequency term grouped in the same category as the remaining term, "erythrocyte". When using all of the terms making up this group of related words as keywords (in Japanese), the top search result is a Japanese-language page whose title translates to "What is Malaria?" When using "latency period" and "erythrocyte" (Type 3), the same page is the top search result. However, when using "erythrocyte" and "hepatic cell" (Type 1), although this page was in the top-ten search results, it was not number one. As another example, consider the following set of related words obtained from NN data: "ovary — spleen — palpation". In this example, "palpation" is a term grouped in the MeSH thesaurus under a category different from that of the other terms. When using all of the terms making up this set of related words as keywords (again, in Japanese), the search results in a Japanese-language page containing the following sentence: "Conditions of the ovary and the spleen can be diagnosed using palpation". From this result, we can interpret this set of related words to bear a causal relation. These results therefore suggest that this set of related words correctly defines the user's intention, enabling retrieval of the relevant web pages.

In this experiment, terms in non-taxonomic relations to other terms were effective in limiting search results to informative pages. At the same time, in comparison with results for non-taxonomically related terms, for taxonomically related terms high-frequency terms were not quantitatively effective, and low-frequency terms did not show any qualitatively

significant tendencies. As an initial experiment, this study was limited to a single domain. Topics of interest for the future will include the extension of research to extract sets of related words featuring more accurate thematic relations and verifying the usefulness of sets of related words in more quantitative and qualitative ways.

References

1 T. Kawahara, "Structure and Processing of Concept", Journal of the Japanese Society for Artificial Intelligence, Vol.16, No.3, 435-440, 2001.

2 Kohonen, T., "Self-organizing maps 2nd Edition", Springer, Berlin, 1997.

3 K. Nemoto, "The combination of the noun with "ga-Case" and the adjective", Language research 2 for the computer, National Language Research Institute, 63-73, 1969.

4 T. Takahashi, "A various phase related to the part-whole relation investigated in the sentence", Studies in the Japanese language 103, The society of Japanese Linguistics, 1-16, 1975.

5 T. Masuoka, "Connected forms of phrases modifying their head nouns - About phrases representing concrete contents of its head nouns", Yukinori Tanabe (ed), Expression of adnominal modifications in Japanese, Kuroshio, 1994.

6 Q. Ma, K. Kanzaki, M. Murata, K. Uchimoto, and H. Isahara, 2000. "Self-Organization Semantic Maps of Japanese Noun in Terms of Adnominal Constituents", In Proceedings of IJCNN'2000, Como, Italy, Vol.6, 91-96.

7 Yamamoto, E. and Umemura, K. 2002. "A Similarity Measure for estimation of One-to-Many Relationship in Corpus", Journal of Natural Language Processing, 45-75.

8 Scheffe H., "An analysis of variance for paired comparison". Journal of the American Statistical Association, 47, 381-400, 1952.

9 M. Geffet and I. Dagan, "The Distribution Inclusion Hypotheses and Lexical Entailment", Proceedings of ACL 2005, pp.107-114, 2005.

10 R. Girju, "Automatic Detection of Causal Relations for Question Answering", Proceedings of ACL Workshop on Multilingual summarization and question answering, pp.76-83, 2003.

11 R. Girju, A. Badulescu, and D. Moldovan, "Automatic Discovery of Part-Whole Relations", Computational Linguistics, 32(1), pp. 83-135, 2006.

12 M. A. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora", Proceedings of Coling 92, pp.539-545, 1992.

13 I. Szpektor, H. Tanev, I. Dagan, and B. Coppola, "Scaling Web-based Acquisition of Entailment Relations", Proceedings of EMNLP 2004, 2004.

14 D. Ravichanfran and E. H. Hovy, "Learning Surface Text Patterns for A Question Answering System", Proceedings of ACL 2002, pp.41-47, 2002.

15 P. Pantel and M. Pennacchiotti, "Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations", Proceedings of ACL 2006, pp.113-120, 2006.

16 E. J. Wisniewski and M. Bassok, "What makes a man similar to a tie?", Cognitive Psychology, 39, pp.208-238, 1999.

17 E. Yamamoto, K. Kanzaki, and H. Isahara, "Extraction of Hierarchies based on Inclusion of Co-occurring Words with Frequency Information", IJCAI 2005, pp.1166-1172, 2005.

***KANZAKI Kyoko**, Ph.D.*

Researcher, Computational Linguistic Group, Knowledge Creating Communication Research Center (former: Researcher, Computational Linguistic Group, Keihanna Human Info-Communication Research Center, Information and Communications Department)

Natural Language Processing

***ISAHARA Hitoshi**, Dr. Eng.*

Group leader, Computational Linguistic Group, Knowledge Creating Communication Research Center (former: Group leader, Computational Linguistic Group, Keihanna Human Info-Communication Research Center, Information and Communications Department)

Natural Language Processing

***YAMAMOTO Eiko**, Dr. Eng.*

Guest Researcher, Computational Linguistic Group, Knowledge Creating Communication Research Center (former: Limited-Term Researcher, Computational Linguistic Group, Keihanna Human Info-Communication Research Center, Information and Communications Department)

Natural Language Processing