Extracting Word Sets with Non-Taxonomical Relation

Eiko Yamamoto **Hitoshi Isahara**
Computational Linguistics Group
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{eiko, isahara}@nict.go.jp

Abstract

At least two kinds of relations exist among related words: taxonomical relations and thematic relations. Both relations identify related words useful to language understanding and generation, information retrieval, and so on. However, although words with taxonomical relations are easy to identify from linguistic resources such as dictionaries and thesauri, words with thematic relations are difficult to identify because they are rarely maintained in linguistic resources. In this paper, we sought to extract thematically (non-taxonomically) related word sets among words in documents by employing case-marking particles derived from syntactic analysis. We then verified the usefulness of word sets with non-taxonomical relation that seems to be a thematic relation for information retrieval.

1. Introduction

Related word sets are useful linguistic resources for language understanding and generation, information retrieval, and so on. In previous research on natural language processing, many methodologies for extracting various relations from corpora have been developed, such as the "is-a" relation (Hearst 1992), "part-of" relation (Berland and Charniak 1999), causal relation (Girju 2003), and entailment relation (Geffet and Dagan 2005).

Related words can be used to support retrieval in order to lead users to high-quality information. One simple method is to provide additional words related to the key words users have input, such as an input support function within the Google search engine. What kind of relation between the key words that have been input and the additional word is effective for information retrieval?

As for the relations among words, at least two kinds of relations exist: the taxonomical relation and the thematic relation. The former is a relation representing the physical resemblance among objects, which is typically a semantic relation such as a hierarchal, synonymic, or antonymic relation; the latter is a relation between objects through a thematic scene, such as "milk" and "cow" as recollected in the scene "milking a cow," and "milk" and "baby," as recollected in the scene "giving baby milk," which include causal relation and entailment relation. Wisniewski and Bassok (1999) showed that both relations are important in recognizing those objects. However, while taxonomical relations are comparatively easy to identify from linguistic resources such as dictionaries and thesauri, thematic relations are difficult to identify because they are rarely maintained in linguistic resources.

In this paper, we sought to extract word sets with a thematic relation from documents by employing case-marking particles derived from syntactic analysis. We then verified the usefulness of word sets with non-taxonomical relation that seems to be a thematic relation for information retrieval.

2. Method

In order to derive word sets that direct users to obtain information, we applied a method based on the Complementary Similarity Measure (CSM), which can determine a relation between two words in a corpus by estimating inclusive relations between two vectors representing each appearance pattern for each words (Yamamoto *et al.* 2005).

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 141–144,
Prague, June 2007.

We first extracted word pairs having an inclusive relation between the words by calculating the CSM values. Extracted word pairs are expressed by a tuple <w_i, w_j>, where $CSM(\boldsymbol{V_i}, \boldsymbol{V_j})$ is greater than $CSM(\boldsymbol{V_j}, \boldsymbol{V_i})$ when words w_i and w_j have each appearance pattern represented by each binary vector $\boldsymbol{V_i}$ and $\boldsymbol{V_j}$. Then, we connected word pairs with CSM values greater than a certain threshold and constructed word sets. A feature of the CSM-based method is that it can extract not only pairs of related words but also sets of related words because it connects tuples consistently.

Suppose we have <A, B>, <B, C>, <Z, B>, <C, D>, <C, E>, and <C, F> in the order of their CSM values, which are greater than the threshold. For example, let <B, C> be an initial word set {B, C}. First, we find the tuple with the greatest CSM value among the tuples in which the word C at the tail of the current word set is the left word, and connect the right word behind C. In this example, word "D" is connected to {B, C} because <C, D> has the greatest CSM value among the three tuples <C, D>, <C, E>, and <C, F>, making the current word set {B, C, D}. This process is repeated until no tuples exist. Next, we find the tuple with the greatest CSM value among the tuples in which the word B at the head of the current word set is the right word, and connect the left word before B. This process is repeated until no tuples exist. In this example, we obtain the word set {A, B, C, D}.

Finally, we removed ones with a taxonomical relation by using thesaurus. The rest of the word sets have a non-taxonomical relation — including a thematic relation — among the words. We then extracted those word sets that do not agree with the thesaurus as word sets with a thematic relation.

3. Experiment

In our experiment, we used domain-specific Japanese documents within the medical domain (225,402 sentences, 10,144 pages, 37MB) gathered from the Web pages of a medical school and the 2005 Medical Subject Headings (MeSH) thesaurus[1]. Recently, there has been a study on query expansion with this thesaurus as domain information (Friberg 2007).

We extracted word sets by utilizing inclusive relations of the appearance pattern between words based on a modified/modifier relationship in documents. The Japanese language has case-marking particles that indicate the semantic relation between two elements in a dependency relation. Then, we collected from documents dependency relations matching the following five patterns; "A <*no* (of)> B," "P <*wo* (object)> V," "Q <*ga* (subject)> V," "R <*ni* (dative)> V," and "S <*ha* (topic)> V," where A, B, P, Q, R, and S are nouns, V is a verb, and <*X*> is a case-marking particle. From such collected dependency relations, we compiled the following types of experimental data; ***NN-data*** based on co-occurrence between nouns for each sentence, ***NV-data*** based on a dependency relation between noun and verb for each case-marking particle <*wo*>, <*ga*>, <*ni*>, and <*ha*>, and ***SO-data*** based on a collocation between subject and object that depends on the same verb V as the subject. These data are represented with a binary vector which corresponds to the appearance pattern of a noun and these vectors are used as arguments of CSM.

We translated descriptors in the MeSH thesaurus into Japanese and used them as Japanese medical terms. The number of terms appearing in this experiment is 2,557 among them. We constructed word sets consisting of these medical terms. Then, we chose 977 word sets consisting of three or more terms from them, and removed word sets with a taxonomical relation from them with the MeSH thesaurus in order to obtain the rest 847 word sets as word sets with a thematic relation.

4. Verification

In verifying the capability of our word sets to retrieve Web pages, we examined whether they could help limit the search results to more informative Web pages with Google as a search engine.

We assume that addition of suitable key words to the query reduces the number of pages retrieved and the remaining pages are informative pages. Based on this assumption, we examined the decrease of the retrieved pages by additional key words and the contents of the retrieved pages in order to verify the availability of our word sets.

Among 847 word sets, we used 294 word sets in which one of the terms is classified into one category and the rest are classified into another.

[1] The U.S. National Library of Medicine created, maintains, and provides the MeSH® thesaurus.

ovary - spleen - palpation (NN)
variation - cross reactions - outbreaks - secretion (Wo)
bleeding - pyrexia - hematuria - consciousness disorder - vertigo - high blood pressure (Ga)
space flight - insemination - immunity (Ni)
cough - fetus - bronchiolitis obliterans organizing pneumonia (Ha)
latency period - erythrocyte - hepatic cell (SO)

Figure 1. Examples of word sets used to verify.

Figure 1 shows examples of the word sets, where terms in a different category are underlined.

In retrieving Web pages for verification, we input the terms composed of these word sets into the search engine. We created three types of search terms from the word set we extracted. Suppose the extracted word set is $\{X_1, ..., X_n, Y\}$, where X_i is classified into one category and Y is classified into another. The first type uses all terms except the one classified into a category different from the others: $\{X_1, ..., X_n\}$ removing Y. The second type uses all terms except the one in the same category as the rest: $\{X_1, ..., X_{k-1}, X_{k+1}, ..., X_n\}$ removing X_k from Type 1. In our experiment, we removed the term X_k with the highest or lowest frequency among X_i. The third type uses terms in Type 2 and Y: $\{X_1, ..., X_{k-1}, X_{k+1}, ..., X_n, Y\}$.

In other words, when we consider the terms in Type 2 as base key words, the terms in Type 1 are key words with the addition of one term having the highest or lowest frequency among the terms in the same category; i.e., the additional term has a feature related to frequency in the documents and is taxonomically related to other terms. The terms in Type 3 are key words with the addition of one term in a category different from those of the other component terms; i.e., the additional term seems to be thematically related — at least non-taxonomically related — to other terms.

First, we quantitatively compared the retrieval results. We used the estimated number of pages retrieved by Google's search engine. Suppose that we first input Type 2 as key words into Google, did not satisfy the result extracted, and added one word to the previous key words. We then sought to determine whether to use Type 1 or Type 3 to obtain more suitable results. The results are shown in Figures 2 and 3, which include the results for the highest frequency and the lowest frequency, respectively. In these figures, the horizontal axis is the number of pages retrieved with Type 2 and the vertical axis is the number of pages retrieved when

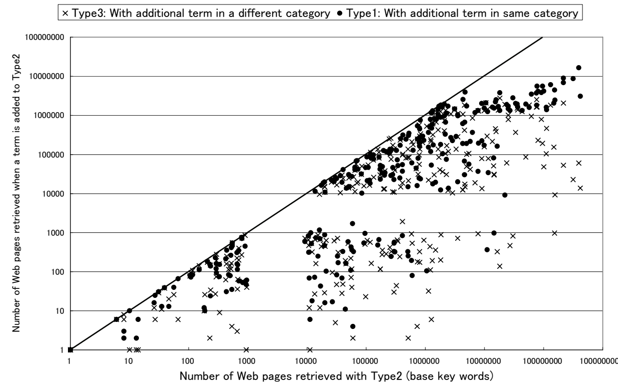

Figure 2. Fluctuation of number of pages retrieved (with the high frequency term).

Type of Data	NN	NV			
		Wo	Ga	Ni	Ha
Word sets for verification	175	43	23	13	26
Cases in which Type 3 defeated Type 1 in retrieval	108	37	15	12	18

Table 1. Number of cases in which Type 3 defeated Type 1 with the high frequency term.

a certain term is added to Type 2. The circles (•) show the retrieval results with additional key word related taxonomically (Type 1). The crosses (×) show the results with additional key word related non-taxonomically (Type 3). The diagonal line shows that adding one term to the base key words does not affect the number of Web pages retrieved.

In Figure 2, most crosses fall further below the line. This graph indicates that when searching by Google, adding a search term related non-taxonomically tends to make a bigger difference than adding a term related taxonomically and with high frequency. This means that adding a term related non-taxonomically to the other terms is crucial to retrieving informative pages; that is, such terms are informative terms themselves. Table 1 shows the number of cases in which term in different category decreases the number of hit pages more than high frequency term. By this table, we found that most of the additional terms with high frequency contributed less than additional terms related non-taxonomically to decreasing the number of Web pages retrieved. This means that, in comparison to the high frequency terms, which might not be so informative in themselves, the terms in the other category — related non-taxonomically — are effective for retrieving useful Web pages.

In Figure 3, most circles fall further below the line, in contrast to Figure 2. This indicates that

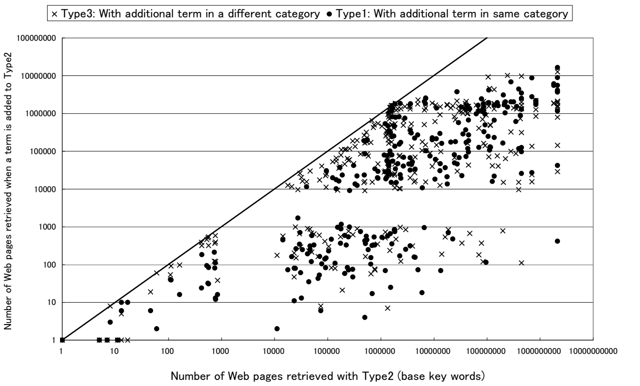

Figure 3. Fluctuation of number of pages retrieved (with the low frequency term).

Type of Data	NN	NV			
		Wo	Ga	Ni	Ha
Word sets for verification	175	43	23	13	26
Cases in which Type 3 defeated Type 1 in retrieval	61	18	7	6	13

Table 2. Number of cases in which Type 3 defeated Type 1 with the low frequency term.

adding a term related taxonomically and with low frequency tends to make a bigger difference than adding a term with high frequency. Certainly, additional terms with low frequency would be informative terms, even though they are related taxonomically, because they may be rare terms on the Web and therefore the number of pages containing the term would be small. Table 2 shows the number of cases in which term in different category decreases the number of hit pages more than low frequency term. In comparing these numbers, we found that the additional term with low frequency helped to reduce the number of Web pages retrieved, making no effort to determine the kind of relation the term had with the other terms. Thus, the terms with low frequencies are quantitatively effective when used for retrieval. However, if we compare the results retrieved with Type 1 search terms and Type 3 search terms, it is clear that big differences exist between them.

For example, consider "latency period - erythrocyte - hepatic cell" obtained from SO-data in Figure 1. "Latency period" is classified into a category different from the other terms and "hepatic cell" has the lowest frequency in this word set. When we used all the three terms, we obtained pages related to "malaria" at the top of the results and the title of the top page was "What is malaria?" in Japanese. With "latency period" and "erythrocyte," we again obtained the same page at the top, although it was not at the top when we used "erythrocyte" and "hepatic cell" which have a taxonomical relation.

As we showed above, the terms with thematic relations with other search terms are effective at directing users to informative pages. Quantitatively, terms with a high frequency are not effective at reducing the number of pages retrieved; qualitatively, low frequency terms may not effective to direct users to informative pages. We will continue our research in order to extract terms in thematic relation more accurately and verify the usefulness of them more quantitatively and qualitatively.

5. Conclusion

We sought to extract word sets with a thematic relation from documents by employing case-marking particles derived from syntactic analysis. We compared the results retrieved with terms related only taxonomically and the results retrieved with terms that included a term related non-taxonomically to the other terms. As a result, we found adding term which is thematically related to terms that have already been input as key words is effective at retrieving informative pages.

References

Berland, M. and Charniak, E. 1999. Finding parts in very large corpora, In *Proceedings of ACL 99*, 57–64.

Friberg, K. 2007. Query expansion using domain information in compounds, In *Proceedings of NAACL-HLT 2007 Doctoral Consortium*, 1–4.

Geffet, M. and Dagan, I. 2005. The distribution inclusion hypotheses and lexical entailment. In *Proceedings of ACL 2005*, 107–114.

Girju, R. 2003. Automatic detection of causal relations for question answering. In *Proceedings of ACL Workshop on Multilingual summarization and question answering*, 76–114.

Hearst, M. A. 1992, Automatic acquisition of hyponyms from large text corpora, In *Proceedings of Coling 92*, 539–545.

Wisniewski, E. J. and Bassok. M. 1999. What makes a man similar to a tie? *Cognitive Psychology*, 39: 208–238.

Yamamoto, E., Kanzaki, K., and Isahara, H. 2005. Extraction of hierarchies based on inclusion of co-occurring words with frequency information. In *Proceedings of IJCAI 2005*, 1166–1172.