AUTOMATIC NOUN CLASSIFICATION BY USING JAPANESE-ENGLISH WORD PAIRS*

Naomi Inoue
KDD R & D Laboratories
2-1-5 Ohara, Kamifukuoka-shi Saitama 356, Japan
inoue@kddlab.kddlabs.cp.jp

ABSTRACT

This paper describes a method of classifying semantically similar nouns. The approach is based on the "distributional hypothesis". Our approach is characterized by distinguishing among senses of the same word in order to resolve the "polysemy" issue. The classification result demonstrates that our approach is successful.

1. INTRODUCTION

Sets of semantically similar words are very useful in natural language processing. The general approach toward classifying words is to use semantic categories, for example the thesaurus. The "is-a" relation is connected between words and categories. However, it is not easy to acquire the "is-a" connection by hand, and it becomes expensive.

Approaches toward automatically classifying words using existing dictionaries were therefore attempted[Chodorow] [Tsurumaru] [Nakamura]. These approaches are partially successful. However, there is a fatal problem in these approaches, namely, existing dictionaries, particularly Japanese dictionaries, are not assembled on the basis of semantic hierarchy.

On the other hand, approaches toward automatically classifying words by using a large-scale corpus have also been attempted[Shirai][Hindle]. They seem to be based on the idea that semantically similar words appear in similar environments. This idea is derived from Harris's "distributional hypothesis"[Harris] in linguistics. Focusing on nouns, the idea claims that each noun is characterized by verbs with which it occurs, and also that nouns are similar to the extent that they share verbs. These automatic classification approaches are also partially successful. However, Hindle says that there is a number of issues to be confronted. The most important issue is that of "polysemy". In Hindle's experiment, two senses of "table", that is to say "table under which one can hide" and "table which can be commuted or memorized", are conflated in the set of words similar to "table". His result shows that senses of the word must be distinguished before classification.

(1)I sit on the table.
(2)I sit on the chair.
(3)I fill in the table.
(4)I fill in the list.

For example, the above sentences may appear in the corpus. In sentences (1) and (2), "table" and "chair" share the same verb "sit on". In sentences (3) and (4), "table" and "list" share the same verb "fill in". However, "table" is used in two different senses. Unless they are distinguished before classification, "table", "chair" and "list" may be put into the same category because "chair" and "list" share the same verbs which are associated with "table". It is thus necessary to distinguish the senses of "table" before automatic classification. Moreover, when the corpus is not sufficiently large, this must be performed for verbs as well as nouns. In the following Japanese sentences, the Japanese verb "聞く" is used in different senses. One is

* This study was done during the author's stay at ATR Interpreting Telephony Research Laboratories.

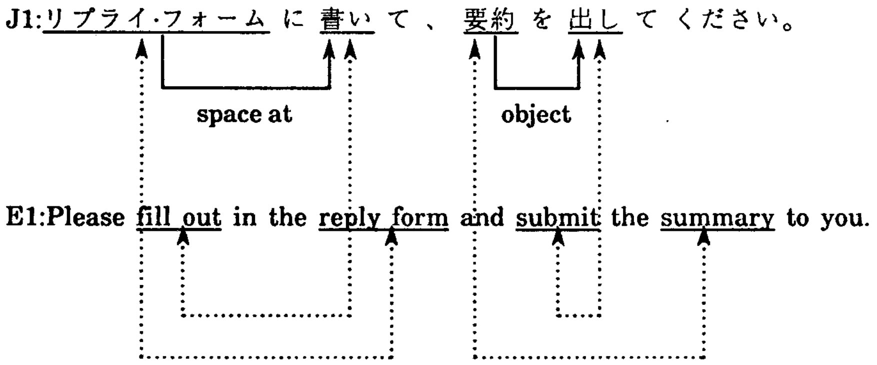
J1:リプライ·フォーム に 書い て 、 要約 を 出し て ください。
space at
object
E1:Please fill out in the reply form and submit the summary to you.

Figure 1 An example of deep semantic relations and the correspondence

"to request information from someone". The other is "to give attention in hearing". Japanese words " 名 前(name)" and " 音 楽 (music)" share the same verb " 聞 く ". Using the small corpus, " 名 前(name)" and " 音 楽 (music)" may be classified into the same category because they share the same verb, though not the same sense, on relatively frequent.

(5)名前を聞く
(6)音楽を聞く

This paper describes an approach to automatically classify the Japanese nouns. Our approach is characterized by distinguishing among senses of the same word by using Japanese-English word pairs extracted from a bilingual database. We suppose here that some senses of Japanese words are distinguished when Japanese sentences are translated into another language. For example, The following Japanese sentences (7),(8) are translated into English sentences (9),(10), respectively.

(7)彼が手紙を出す
(8)彼が本を出す
(9)He sends a letter.
(10)He publishes a book.

The Japanese word " 出 す" has at least two senses. One is "to cause to go or be taken to a place" and the other is "to have printed and put on sale". In the above example, the Japanese word " 出 す" corresponds to "send" from sentences (7) and (9). The Japanese word " 出 す" also corresponds to "publish" from sentences (8) and (10). That is to say, the Japanese word " 出 す" is translated into different English words according to the sense. This example shows that it may be possible to distinguish among senses of the same word by using words from another language. We used Japanese-English word pairs, for example, " 出 す-send" and " 出 す-publish", as senses of Japanese words.

In this paper, these word pairs are acquired from ATR's large scale database.

2. CONTENT OF THE DATABASE

ATR has constructed a large-scale database which is collected from simulated telephone and keyboard conversations [Ehara]. The sentences collected in Japanese are manually translated into English. We obtain a bilingual database. The database is called the ATR Dialogue Database(ADD). ATR aims to build ADD to one million words covering two tasks. One task is dialogues between secretaries and participants of international conferences. The other is dialogues between travel agents and customers. Collected Japanese and English sentences are morphologically analyzed. Japanese sentences are also dependency analyzed and given deep semantic relations. We use 63 deep semantic cases[Inoue]. Correspondences of Japanese and English are made by several linguistic units, for example words, sentences and so on.

Figure 1 shows an example of deep semantic relations and correspondences of Japanese and English words. The sentence is already morphologically analyzed. The solid line shows deep semantic relations. The Japanese nouns "リプライフォーム" and "要

約" modify the Japanese verbs "書い" and "出し", respectively. The semantic relations are "space at" and "object", which are almost equal to "locative" and "objective" of Fillmore's deep case[Fillmore]. The dotted line shows the word correspondence between Japanese and English. The Japanese words "リプライフォーム", "書い", "要約" and "出し" correspond to the English words "reply form", "fill out", "summary" and "submit", respectively. Here, "書い" and "出し" are conjugations of "書く" and "出す", respectively. However, it is possible to extract semantic relations and word correspondence in dictionary form, because ADD includes the dictionary forms.

3. CLASSIFICATION OF NOUNS

3.1 Using Data

We automatically extracted from ADD not only deep semantic relations between Japanese nouns and verbs but also the English word which corresponds to the Japanese word. We used telephone dialogues between secretaries and participants because the scale of analyzed words was largest. Table 1 shows the current number of analyzed words.

Table 1 Analyzed words counts of ADD

Media	Task	Words
Telephone	Conference Travel	139,774 11,709
Keyboard	Conference Travel	64,059 0

Figure 2 shows an example of the data extracted from ADD. Each field is delimited by the delimiter "|". The first field is the dialogue identification number in which the semantic relation appears. The second and the third fields are Japanese nouns and their corresponding English words. The next 2 fields are Japanese verbs and their corresponding English words. The last is the semantic relations between nouns and verbs.

Moreover, we automatically acquired word pairs from the data shown in Figure 2. Different senses of nouns appear far less frequently than those of verbs because the database is restricted to a specific task. In this experiment, only word pairs of verbs are used. Figure 3 shows deep semantic relations between nouns and word pairs of verbs. The last field is raw frequency of co-occurrence. We used the data shown in Figure 3 for noun classification.

```
1|登録費|registration fee|出す|pay|object
15|要約|summary|出す|send|object
157|プロシーディング|proceeding|出す
|issue|object
4|会議|conference|有る|be held|object
8|質問|question|有る|have|object
3|バス|bus|乗る|take|object
180|新聞|newspaper|乗る|see|space at
```

Figure 2 An example of data extracted from ADD

The experiment is done for a sample of 138 nouns which are included in the 500 most frequent words. The 500 most frequent words cover 90% of words accumulated in the telephone dialogue. Those nouns appear more frequently than 9 in ADD.

```
登録費|出す-pay|object|1
要約|出す-send|object|2
プロシーディング|出す-issue|object|2
会議|有る-be held|object|6
質問|有る-have|object|7
バス|乗る-take|object|1
新聞|乗る-see|space at|1
```

Figure 3 An example of semantic relations of nouns and word pairs

3.2 Semantic Distance of Nouns

Our classification approach is based on the "distributional hypothesis". Based on this semantic theory, nouns are similar to the extent that they share verb senses. The aim of this paper is to show the efficiency of using the word pair as the word sense. We therefore used the following expression(1), which was already defined by Shirai[Shirai] as the distance between two words. The

$$d(a,b) = 1 - \frac{\sum\sum_{v \in V, r \in R} \Phi(M(a,v,r), M(b,v,r))}{\sum\sum_{v \in V, r \in R} (M(a,v,r) + M(b,v,r))} \quad (1)$$

Here, a,b : noun (a,b ∈ N)
r : semantic relation
v : verb senses
N : the set of nouns
V : the set of verb senses
R : the set of semantic relations
M(a,v,r) : the frequency of the semantic relation r between a and v

$$\Phi(x,y) = \begin{cases} x + y & (x>0, y>0) \\ 0 & (x=0 \text{ or } y=0) \end{cases}$$

second term of the expression can show the semantic similarity between two nouns, because it is the ratio of the verb senses with which both nouns (a and b) occur and all the verb senses with which each noun (a or b) occurs. The distance is normalized from 0.0 to 1.0. If one noun (a) shares all verb senses with the other noun (b) and the frequency is also same, the distance is 0.0. If one noun (a) shares no verb senses with the other noun (b), the distance is 1.0.

3.3 Classification Method

For the classification, we adopted cluster analysis which is one of the approaches in multivariant analysis. Cluster analysis is generally used in various fields, for example biology, psychology, etc.. Some hierarchical clustering methods, for example the nearest neighbor method, the centroid method, etc., have been studied. It has been proved that the centroid method can avoid the chain effect. The chain effect is an undesirable phenomenon in which the nearest unit is not always classified into a cluster and more distant units are chained into a cluster. The centroid method is a method in which the cluster is characterized by the centroid of categorized units. In the following section, the result obtained by the centroid method is shown.

4.EXPERIMENT

4.1 Clustering Result

All 138 nouns are hierarchically classified. However, only some subsets of the whole hierarchy are shown, as space is limited. In Figure 4, we can see that semantically similar nouns, which may be defined as "things made from paper", are grouped together. The X-axis is the semantic distance defined before. Figure 5 shows another subset. All nouns in Figure 5, "決定(decision)", "発表(presentation)", "スピーチ(speech)" and "話(talk)", have an active concept like verbs. Subsets of nouns shown in Figures 4 and 5 are fairly coherent. However, all subsets of nouns are not coherent. In Figure 6, "スライド(slide)", "原稿(draft)", "会場(conference site)", "8日(8th)" and "駅(station)" are grouped together. The semantic distances are 0.67, 0.6, 0.7 and 0.8. The distance is upset when "会場(conference site)" is attached to the cluster containing "スライド(slide)" and "原稿(draft)". This is one characteristic of the centroid method. However, this seems to result in a semantically less similar cluster. The word pairs of verbs, the deep semantic relations and the frequency are shown in Table 2. After "スライド(slide)" and "原稿(draft)" are grouped into a cluster, the cluster and "会場(conference site)" share two word pairs, "使う-use" and "成る-be". "成る-be" contributes more largely to attach "会場(conference site)" to the cluster than "使う-use" because the frequency of co-occurrence is greater. In this sample, "成る-be" occurs with more nouns than "使う-use". It shows that "成る-be" is less important in characterizing nouns though the raw frequency of co-occurrence is greater. It is therefore necessary to develop a means of not relying on the raw frequency of co-occurrence, in order to make the clustering result more accurate. This is left to further study.

4.2 Estimation of the Result

All nouns are hierarchically classified, but some semantically separated clusters are acquired if the threshold is used.

It is possible to compare clusters derived from this experiment with semantic categories which are used in our automatic interpreting telephony system. We used expression (2), which was defined by Goodman and Kruskal[Goodman], in order to objectively compare them.

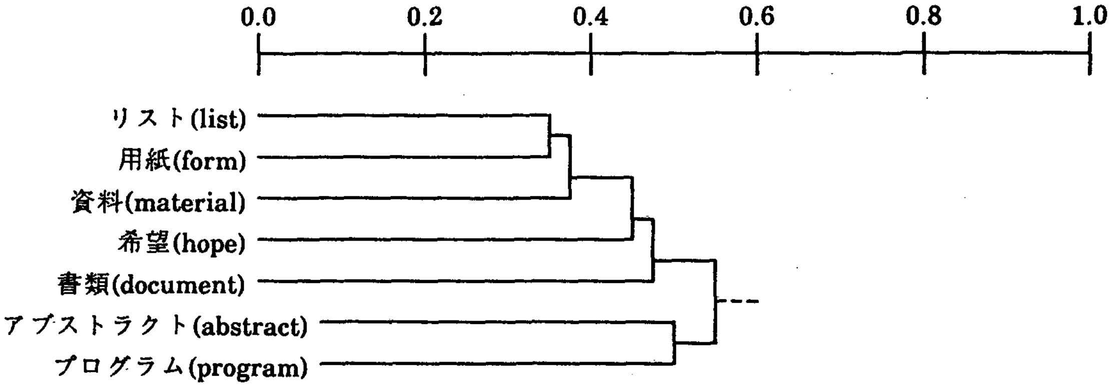

Figure 4 An example of the classification of nouns

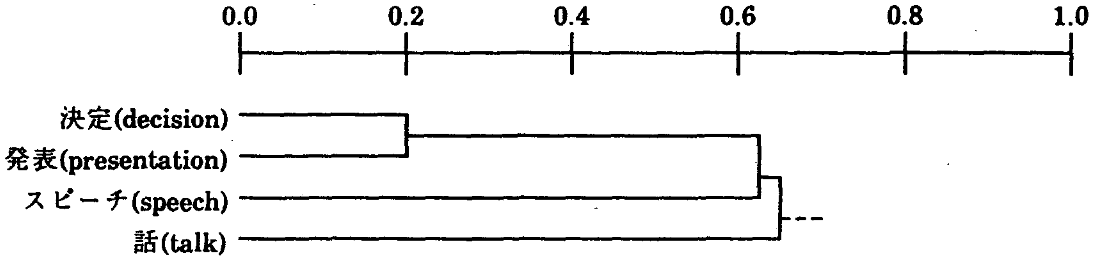

Figure 5 Another example of the classification of nouns

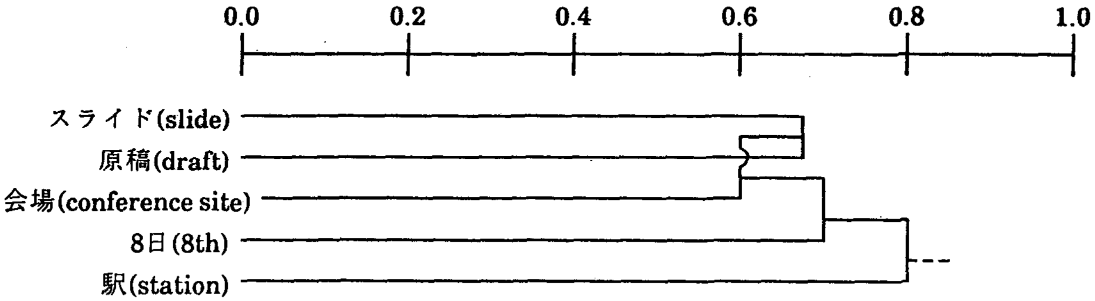

Figure 6 Another example of the classification of nouns

Table 2 A subset of semantically similar nouns

noun	*word pairs of verb*	*deep case*	*frequency*
スライド(slide)	する-make	goal	1
	作る-make	object	1
	使う-use	object	1
原稿(draft)	作る-make	object	1
	成る-be	object	1
	待つ-look forward to	object	1
会場(conference site)	掛る-take	condition	1
	伺う-get	space to	1
	使う-use	object	1
	出来る-can	space at	1
	言う-say	space at	1
	成る-be	object	2
8日(8th)	終る-end	time	2
	成る-be	object	1
	聞く-guess	content	1
駅(station)	掛る-take	condition	1
	御座います-there be	space from	1

$$P = \frac{P_1 - P_2}{P_1} \qquad (2)$$

Here, $P_1 = 1 - f_{\cdot m}$

$$P_2 = \sum_{i=1}^{p} f_{i\cdot}(1 - f_{im_i}/f_{i\cdot})$$

$f_{\cdot m} = \max\{f_{\cdot 1}, f_{\cdot 2}, \cdots, f_{\cdot q}\}$

$f_{am_a} = \max\{f_{a1}, f_{a2}, \cdots, f_{aq}\}$

$f_{ij} = n_{ij}/n$

$f_{\cdot j} = n_{\cdot j}/n$

A : a set of clusters which are automatically obtained.

B : a set of clusters which are used in our interpreting telephony system.

p : the number of clusters of a set A

q : the number of clusters of a set B

n_{ij} : the number of nouns which are included in both the ith cluster of A and the jth cluster of B

$n_{\cdot j}$: the number of nouns which are included in the jth cluster of B

n : all nouns which are included in A or B

They proposed that one set of clusters, called 'A', can be estimated to the extent that 'A' associates with the other set of clusters, called 'B'. In figure 7, two results are shown. One (solid line) is the result of using the word pair to distinguish among senses of the same verb. The other (dotted line) is the result of using the verb form itself. The X-axis is the number of classified nouns and the Y-axis is the value derived from the above expression.Figure 7 shows that it is better to use word pairs of verbs than not use them, when fewer than about 30 nouns are classified. However, both are almost the same, when more than about 30 nouns are classified. The result proves that the distinction of senses of verbs is successful when only a few nouns are classified.

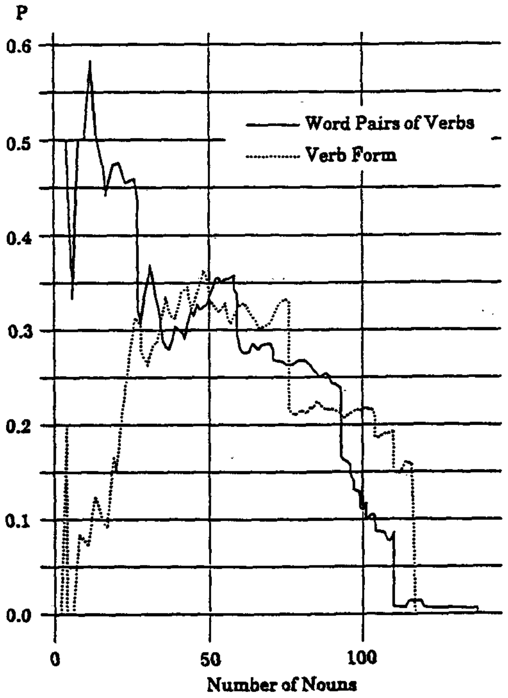

Figure 7 Estimation result

5. CONCLUSION

Using word pairs of Japanese and English to distinguish among senses of the same verb, we have shown that using word pairs to classify nouns is better than not using word pairs, when only a few nouns are classified. However, this experiment did not succeed for a sufficient number of nouns for two reasons. One is that the raw co-occurrent frequency is used to calculate the semantic distance. The other is that the sample size is too small. It is thus necessary to resolve the following issues to make the classification result more accurate.

(1)to develop a means of using the frequency normalized by expected word pairs.
(2)to estimate an adequate sample size.

In this experiment, we acquired word pairs and semantic relations from our database. However, they are made by hand. It is also preferable to develop a method of automatically acquiring them from the bilingual text database.

Moreover, we want to apply the hierarchically classified result to the translated word selection problem in Machine translation.

ACKNOWLEDGEMENTS

The author is deeply grateful to Dr. Akira Kurematsu, President of ATR Interpreting Telephony Research Laboratories, Dr. Toshiyuki Takezawa and other members of the Knowledge & Data Base Department for their encouragement, during the author's stay at ATR Interpreting Telephony Research Laboratories.

REFERENCES

[Chodorow] Chodorow, M. S., et al. "Extracting Semantic Hierarchies from a Large On-line Dictionary.", Proceedings of the 23rd Annual Meeting of the ACL, 1985.

[Ehara] Ehara, T., et al. "ATR Dialogue Database", Proceedings of ICSLP, 1990.

[Fillmore] Fillmore, C. J. "The case for case", in E. Bach & Harms (Eds.) Universals in linguistic theory, 1968.

[Goodman] Goodman, L. A., and Kruskal W.H. "Measures of Association for Cross Classifications", J. Amer. Statist. Assoc. 49, 1954.

[Harris] Harris, Z. S. "Mathematical Structures of Language", a Wiley-Interscience Publication.

[Hindle] Hindle, D. "Noun Classification from Predicate-Argument Structures", Proceedings of 28th Annual Meeting of the ACL, 1990.

[Inoue] Inoue, N., et al. "Semantic Relations in ATR Linguistic Database" (in Japanese), ATR Technical Report TR-I-0029, 1988.

[Nakamura] Nakamura, J., et al. "Automatic Analysis of Semantic Relation between English Nouns by an Ordinal English Dictionary" (in Japanese), the Institute of Electronics, Information and Communication Engineers, Technical Report, NLC-86, 1986.

[Shirai] Shirai K., et al. "Database Formulation and Learning Procedure for Kakariuke Dependency Analysis" (in Japanese), Transactions of Information Processing Society of Japan, Vol.26, No.4, 1985.

[Tsurumaru] Tsurumaru H., et al. "Automatic Extraction of Hierarchical Structure of Words from Definition Sentences" (in Japanese), the Information Processing Society of Japan, Sig. Notes, 87-NL-64, 1987.