An Automatic Clustering of Articles Using Dictionary Definitions

Fumiyo FUKUMOTO **Yoshimi SUZUKI†**
Dept. of Electrical Engineering and Computer Science, Yamanashi University
4-3-11 Takeda, Kofu 400 Japan
{fukumoto@skye, ysuzuki@suwa†}.esi.yamanashi.ac.jp

Abstract

In this paper, we propose a statistical approach for clustering of articles using on-line dictionary definitions. One of the characteristics of our approach is that every sense of word in articles is automatically disambiguated using dictionary definitions. The other is that in order to cope with the problem of a phrasal lexicon, linking which links words with their semantically similar words in articles is introduced in our method. The results of experiments demonstrate the effectiveness of the proposed method.

1 Introduction

There has been quite a lot of research concerned with automatic clustering of articles or automatic identification of semantically similar articles(Walker, 1986), (Guthrie, 1994), (Yuasa, 1995). Most of these works deal with entirely different articles.

In general, the problem that the same word can be used differently in different subject domains is less problematic in entirely different articles, such as 'weather forecasts', 'medical reports', and 'computer manuals'. Because these articles are characterised by a larger number of different words than that of the same words. However, in texts from a restricted domain such as financial articles, e.g *Wall Street Journal* (*WSJ* in short) (Liberman, 1990), one encounters quite a large number of polysemous words. Therefore, polysemous words often hamper the precise classification of articles, each of which belongs to the restricted subject domain.

In this paper, we report an experimental study for clustering of articles by using on-line dictionary definitions and show how dictionary-definition can use effectively to classify articles, each of which belongs to the restricted subject domain. We first describe a method for disambiguating word-senses in articles based on dictionary definitions. Then, we present a method for classifying articles and finally, we report some experiments in order to show the effect of the method.

2 Related Work

One of major approaches in automatic clustering of articles is based on statistical information of words in articles. Every article is characterised by a vector, each dimension of which is associated with a specific word in articles, and every coordinate of the article is represented by term weighting. Term weighting methods have been widely studied in information retrieval research (Salton, 1983), (Jones, 1972) and some of them are used in an automatic clustering of articles. Guthrie and Yuasa used word frequencies for weighting (Guthrie, 1994), (Yuasa, 1995), and Tokunaga used weighted inverse document frequency which is a word frequency within the document divided by its frequency throughout the entire document collection (Tokunaga, 1994). The results of these methods when applied to articles' classification task, seem to show its effectiveness. However, these works do not seriously deal with the problem of polysemy.

The alternative approach is based on dictionary's information as a thesaurus. One of major problems using thesaurus categories as sense representation is a statistical sparseness for thesaurus words, since they are mostly rather uncommon words (Niwa, 1995). Yuasa reported the experimental results when using word frequencies for weighting within large documents were better results in clustering documents as those when EDR electronic dictionary as a thesaurus (Yuasa, 1995).

The technique developed by Walker also used dictionary's information and seems to cope with the discrimination of polysemy (Walker, 1986). He used the semantic codes of the *Longman Dictionary of Contemporary English* in order to determine the subject domain for a set of texts. For a given text, each word is checked against the dictionary to determine the semantic codes associated with it. By accumulating the frequencies for these senses and then ordering the list of categories in terms of frequency, the subject matter of

the text can be identified. However, as he admits, a phrasal lexicon, such as *Atlantic Seaboard, New England* gives a negative influence for clustering, since it can not be regarded as units, i.e. each word which is the element of a phrasal lexicon is assigned to each semantic code.

The approach proposed in this paper focuses on these problems, i.e. polysemy and a phrasal lexicon. Like Guthrie and Yuasa's methods, our approach adopts a vector representation, i.e. every article is characterised by a vector. However, while their approaches assign each coordinate of a vector to each word in articles, we use a word (noun) of which sense is disambiguated. Our disambiguation method of word-senses is based on Niwa's method which used the similarity between two sentences, i.e. a sentence which contains a polysemous noun and a sentence of dictionary-definition. In order to cope with Walker's problem, for the results of disambiguation technique, semantic relativeness of words are calculated, and semantically related words are grouped together.

We used *WSJ* corpus as test articles in the experiments in order to see how our method can effectively classify articles, each of which belongs to the restricted subject domain, i.e. *WSJ*.

3 Framework

3.1 Word-Sense Disambiguation

Every sense of words in articles which should be clustered is automatically disambiguated in advance. Word-sense disambiguation (WSD in short) is a serious problem for NLP, and a variety of approaches have been proposed for solving it (Brown, 1991), (Yarowsky, 1992).

Our disambiguation method is based on Niwa's method which used the similarity between a sentence containing a polysemous noun and a sentence of dictionary-definition. Let x be a polysemous noun and a sentence X be

$$X: \quad \cdots, x_{-n}, \cdots, x_{-1}, x, x_1, \cdots, x_n, \cdots$$

The vector representation of X is

$$V(X) = \sum_{i=-n}^{n} V(x_i)$$

where $V(x_i)$ is

$$V(x_i) = (Mu(x_i, o_1), \cdots, Mu(x_i, o_m))$$

Here, $Mu(x, y)$ is the value of *mutual information* proposed by (Church, 1991). o_1,$\cdots$,o_m (We call them *basic words*) are selected the 1000th most frequent words in the reference *Collins English Dictionary* (Liberman, 1990).

Let word x have senses $s1$,$s2$,$\cdots$,sp and the dictionary-definition of si be

$$Y_{si}: \quad \cdots, y_{-n}, \cdots, y_{-1}, y, y_1, \cdots, y_n, \cdots$$

The similarity of X and Y_{si} is measured by the inner product of their normalised vectors and is defined as follows:

$$Sim(X, Y_{si}) = \frac{V(X) * V(Y_{si})}{|\, V(X) \,||\, V(Y_{si}) \,|} \quad (1)$$

We infer that the sense of word x in X is si if $Sim(X, Y_{si})$ is maximum among Y_{s1},$\cdots$,Y_{sp}.

Given an article, the procedure for WSD is applied to each word (noun) in an article, i.e. the sense of each noun is estimated using formula (1) and the word is replaced by its sense. Table 1 shows sample of the results of our disambiguation method.

Table 1: The results of the WSD method

Input	A number of major airlines adopted continental airlines' $\cdots$
Output	A number5 of major airlines1 adopted continental2 airlines2 $\cdots$

In Table 1, underline signifies polysemous noun. 'Output' shows that each noun is replaced by a symbol word which corresponds to each sense of a word. We call 'Input' and 'Output' in Table 1, an *original* article and a *new* article, respectively.

Table 2: The definition of 'number'

number1:	Every number occupies a unique position in a sequence.
number2:	He was not one of our number.
number3:	A telephone number.
number4:	She was number seven in the race.
number5:	A large number of people.

Table 2 shows the definition of 'number' in the *Collins English Dictionary*. 'number1' $\sim$ 'number5' are symbol words and show different senses of 'number'.

3.2 Linking Nouns with their Semantically Similar Nouns

Our method for classification of articles uses the results of disambiguation method. The problems here are:

1. The frequency of every disambiguated noun in *new* articles is lower than that of every polysemous noun in *original* articles. For example, the frequency of 'number5' in Table 1 is lower than that of 'number'[1]. Furthermore, some nouns in articles may be semantically similar with each other. For example, 'number5' in Table 2 and 'sum4' in Table 3 are almost the same sense.
2. A phrasal lexicon which Walker suggested in his method gives a negative influence for classification.

[1] If all 'number' are used as 'number5' sense, the frequency of 'number' is the same as 'number5'.

Table 3: The definition of 'sum' in the dictionary

sum1:	The result of the addition of numbers.
sum2:	One or more columns or rows of numbers to be added.
sum3:	The limit of the first n terms of a converging infinite series as n tends to infinity.
sum4:	He borrows enormous sums.
sum5:	The essence or gist of a matter.

In order to cope with these problems, we linked nouns in *new* articles with their semantically similar nouns. The procedures for linking are the following five stages.

Stage One: Calculating *Mu*

The first stage for linking nouns with their semantically similar nouns is to calculate *Mu* between noun pair x and y in *new* articles. In order to get a reliable statistical data, we merged every *new* article into one and used it to calculate *Mu*. The results are used in the following stages.

Stage Two: Representing every noun as a vector

The goal of this stage is to represent every noun in a *new* article as a vector. Using a term weighting method, nouns in a *new* article would be represented by vector of the form

$$v = (w_1, w_2, \cdots, w_n) \quad (2)$$

where w_i is the element of a *new* article and corresponds to the weight of the noun w_i. In our method, the weight of w_i is the value of *Mu* between v and w_i which is calculated in Stage One.

Stage Three: Measuring similarity between vectors

Given a vector representation of nouns in *new* articles as in formula (2), a dissimilarity between two words (noun) v_1, v_2 in an article would be obtained by using formula (3). A dissimilarity measure is the degree of deviation of the group in an n-dimensional Euclidean space, where n is the number of nouns which co-occur with v_1 and v_2.

$$Dis(v_1, v_2) = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n}(v_{ij} - \bar{g}_j)^2}{|\bar{g}|} \quad (3)$$

$\bar{g} = (\bar{g}_1, \cdots, \bar{g}_n)$ is the centre of gravity and $|\bar{g}|$ is the length of it. A group with a smaller value of (3) is considered *semantically less deviant.*

Stage Four: Clustering method

For a set of nouns w_1, w_2, $\cdots$, w_n of a *new* article, we calculate the semantic deviation value of all possible pairs of nouns.

Table 4 shows sample of the results of nouns with their semantic deviation values.

Table 4: Pairs of nouns with $Dis(v_1,v_2)$ values

BBK		
0.125	share1	company1
0.140	giorgio	di
0.215	shares2	share2
0.262	share2	company1
...	...	...
0.345	new3	york1
...	...	...

In Table 4, 'BBK' shows the topic of the article which is tagging in the *WSJ*, i.e. 'Buybacks'. The value of Table 4 shows the semantic deviation value of two nouns[2].

The clustering algorithm is applied to the sets shown in Table 4 and produced a set of semantic clusters, which are ordered in the ascending order of their semantic deviation values. We adopted non-overlapping, group average method in our clustering technique (Jardine, 1991). The sample results of clustering is shown in Table 5.

Table 5: Clustering results of 'BBK'

0.125	[share1 company1]
0.140	[giorgio di]
0.215	[shares2 share2]
0.251	[share1 company1 share2 shares2]
...	...

The value of Table 5 shows the semantic deviation value of the cluster.

Stage Five: Linking nouns with their semantically similar nouns

We selected different 49 articles from 1988, 1989 *WSJ*, and applied to Stage One $\sim$ Four. From these results, we manually selected clusters which are judged to be semantically similar. For the selected clusters, if there is a noun which belongs to several clusters, these clusters are grouped together. As a result, each cluster is added to a sequential number. The sample of the results are shown in Table 6.

Table 6: The results of Stage Five

Seq. num	Semantically similar nouns
$word_1$:	bank3, banks3
$word_2$:	canada3, canada4
$word_3$:	American1, express1
$word_4$:	co., corp., company1 $\cdots$
$word_5$:	August, June, July, Sept. Oct. $\cdots$
...	new2 york2
...	...

[2]In Table 4, there are some nouns which are not added to the number, '1' $\sim$ '5', e.g. 'giorgio', 'di'. This shows that for these words, there is only one meaning in the dictionary.

'Seq. num' in Table 6 shows a sequential number, '$word_1$', $\cdots$,'$word_m$' which are added to the group of semantically similar nouns[3]. Table 6 shows, for example, 'new2' and 'york2' are semantically similar and form a phrasal lexicon.

3.3 Clustering of Articles

According to Table 6, frequency of every word in *new* articles is counted, i.e. if a word in a *new* article belongs to the group shown in Table 6, the word is replaced by its representative number '$word_i$' and the frequency of '$word_i$' is counted. For example, 'bank3' and 'banks3' in a *new* article are replaced by '$word_i$', and the frequency of '$word_i$' equals to the total number of frequency of 'bank3' and 'banks3'.

Using a term weighting method, articles would be represented by vectors of the form

$$A = (w_1, w_2, \cdots, w_n) \quad (4)$$

where w_i corresponds to the weight of the noun i. The weight is used to the frequency of noun. Given the vector representations of articles as in formula (4), a similarity between A_i and A_j are calculated using formula (1). The greater the value of $Sim(A_i, A_j)$ is, the more similar these two articles are. The clustering algorithm which is described in Stage Four is applied to each pair of articles, and produces a set of clusters which are ordered in the descending order of their semantic similarity values.

4 Experiments

We have conducted four experiments, i.e. 'Freq', 'Dis', 'Link', and 'Method' in order to examine how WSD method and linking words with their semantically similar words(linking method in short) affect the clustering results. 'Freq' is frequency-based experiment, i.e. we use word frequency for weighting and do not use WSD and linking methods. 'Dis' is concerned with disambiguation-based experiment, i.e. the clustering algorithm is applied to *new* articles. 'Link' is concerned with linking-based experiment, i.e. we applied linking method to *original* articles. 'Method' shows our proposed method.

4.1 Data

The training corpus we have used is the 1988, 1989 *WSJ* in ACL/DCI CD-ROM which consists of about 280,000 part-of-speech tagged sentences (Brill, 1992). From this corpus, we selected at random 49 different articles for test data, each of which consists of 3,500 sentences and has different topic name which is tagging in the *WSJ*. We classified 49 articles into eight *categories*, e.g. 'market news', 'food· restaurant', etc. The dictionary we have used is *Collins English Dictionary* in ACL/DCI CD-ROM.

In WSD method, the co-occurrence of x and y for calculating *Mu* is that the two words (x,y) appear in the training corpus in this order in a window of 100 words, i.e. x is followed by y within a 100-word distance. This is because, the larger window sizes might be considered to be useful for extracting semantic relationships between nouns. *Basic words* are selected the 1000th most frequent words in the reference *Collins English Dictionary*. The length of a sentence X which contains a polysemous noun and the length of a sentence of dictionary-definition are maximum 20 words. For each polysemous noun, we selected the first top 5 definitions in the dictionary.

In linking method, a window size of the co-occurrence of x and y for calculating *Mu* is the same as that in WSD method, i.e. a window of 100 words. We selected 969 $\sim$ 9128 different (noun, noun) pairs for each article, 377 $\sim$ 1259 different nouns on condition that frequencies and *Mu* are not low ($f(x,y) \geq 5$, $Mu(x,y) \geq 3$) to permit a reliable statistical analysis[4]. As a result of Stage Four, we manually selected clusters which are judged to be semantically similar. As a result, we selected clusters on condition that the threshold value for similarity was 0.475. For the selected clusters, if there is a noun which belongs to several clusters, these clusters are grouped together. As a result, we obtained 238 clusters in all.

4.2 Results of the experiments

The results are shown in Table 7.

Table 7: The results of the experiments

Article	Num	Freq	Link	Dis	Method
5	10	4	4	5	8
10	10	4	6	6	9
15	10	7	7	7	8
20	10	6	6	6	6
Total	40	21	23	24	31
(%)	(-)	(52.5)	(57.5)	(60.0)	(77.5)

In Table 7, 'Article' means the number of articles which are selected from test data. 'Num' means the number for each 'Article', i.e. we selected 10 sets for each 'Article'. 'Freq', 'Link', 'Dis', and 'Method' show the number of sets which are clustered correctly in each experiment.

The sample results of 'Article = 20' for each experiment is shown in Figure 1, 2, 3, and 4.

In Figure 1, 2, 3, and 4, the X-axis is the similarity value. Abbreviation words in each Figure and *categories* are shown in Table 8.

[3] In our experiments, m equals to 238.

[4] Here, $f(x,y)$ is the number of total co-occurrences of words x and y in this order in a window size of 100 words.

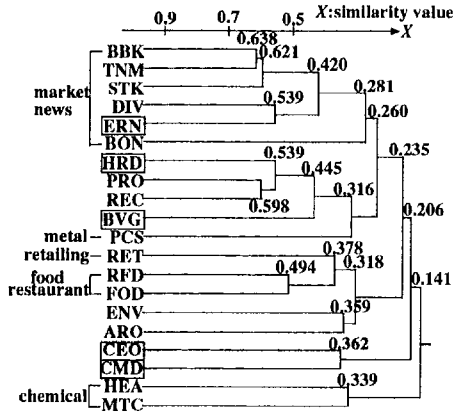

Figure 1: The results of 'Freq' experiment

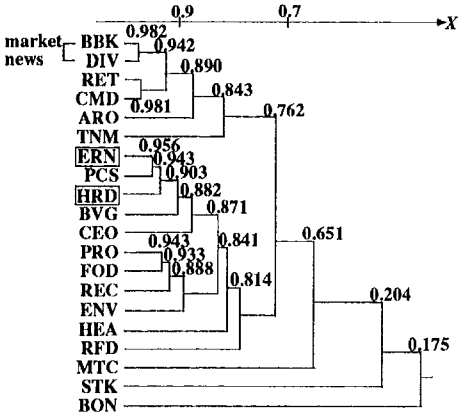

Figure 2: The results of 'Link' experiment

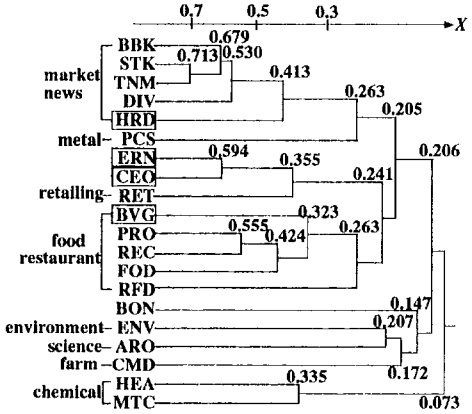

Figure 3: The results of 'Dis' experiment

5 Discussion

1. WSD method

According to Table 7, there are 24 sets which could be clustered correctly in 'Dis', while 21 sets in 'Freq'. Examining the results shown in Figure 3, 'BVG' and 'HRD' are correctly classified into 'food · restaurant' and 'market news', respectively. However, the results of 'Freq' (Figure 1) shows that they are classified incorrectly. Table 9 shows different senses of word in 'BVG', and 'HRD' which could be discriminated in 'Dis'.
In Table 9, for example, 'security' is high frequencies and used in 'being secure' sense in 'BVG' article, while 'security' is 'certificate of creditorship' sense in 'HRD'. One possible cause that the results of 'Freq' is worse than 'Dis' is that these polysemous words which are high-frequencies are not recognised polysemy in 'Freq'.

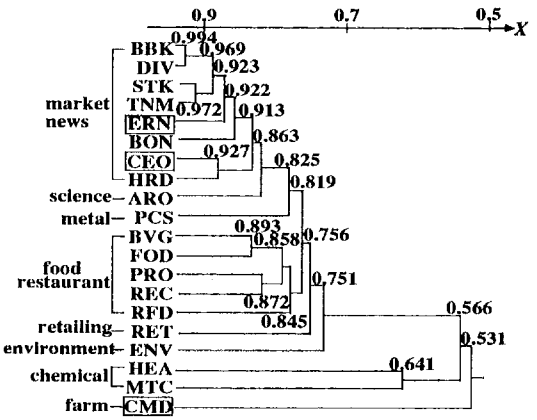

Figure 4: The results of 'Method' experiment

Table 8: Topic and *category* name

Category	Topic
market news	BBK: Buybacks
	BON: Bond Market News
	CEO: Dow Jones interview
	DIV: dividends
	ERN: earnings
	HRD: Heard on the street
	STK: stock market
	TNM: tender offers
science	ARO: aerospace
metal	PCS: precious metals, stones, gold
food restaurant	BVG: beverages
	FOD: food products
	PRO: corporate profile
	REC: recreation, entertainment
	RFD: restaurant, supermarket
retailing	RET: retailing
environment	ENV: environment
chemical	HEA: health care providers, medicine
	MTC: medical and biotechnology
farm	CMD: commodity news, farm products

2. Linking method

As shown in Table 7, there are 23 sets which could be clustered correctly in 'Link', while 21 sets in 'Freq'. For example, 'ERN' and 'HRD' are both concerned with 'market news'. In Figure 2, they are clustered with high similarity value(0.943), while in Figure 1, they are not(0.260).

Examining the results, there are 811 nouns in 'ERN' article, and 714 nouns in 'HRD', and

Table 9: Different word-senses in BVG and HRD

	BVG	HRD
security	the state of being secure	certificate of creditorship
rate	a quantity in relation	price of charge
sale	the exchange of goods	the amount of sold
stock	total goods	stock market

of these, 'shares', 'stock', and 'share' which are semantically similar are included. In linking method, there are 251 nouns in 'ERN' and 492 nouns in 'HRD' which are replaced for representative words. However, in 'Freq', each noun corresponds different coordinate, and regards to different meaning. As a result, these topics are clustered with low similarity value.

3. Our method

The results of 'Method' show that 31 out of 40 sets are classified correctly, and the percentage attained was 77.5%, while 'Freq', 'Link', and 'Dis' experiment attained 52.5%, 57.5%, 60.0%, respectively. This shows the effectiveness of our method.

In Figure 4, the articles are judged to classify into eight categories. Examining 'ERN', 'CEO' and 'CMD' in Figure 1, 'CEO' and 'CMD' are grouped together, while they have different categories with each other. On the other hand, in Figure 3, 'ERN' and 'CEO' are grouped together correctly. Examining the nouns which are belonging to 'ERN' and 'CEO', 'plant'(factory and food senses), 'oil'(petroleum and food), 'order'(command and demand), and 'interest'(debt and curiosity) which are high frequencies are correctly disambiguated. Furthermore, in Figure 4, 'ERN' and 'CEO' are classified into 'market news', and 'CMD' are classified into 'farm', correctly. For example, 'plant' which is used in 'factory' sense is linked with semantically similar words, 'manufacturing', 'factory', 'production', or 'job' etc.. In a similar way, 'plant' which is used in 'food' sense is linked with 'environment', 'forest'. As a result, the articles are classified correctly.

As shown in Table 7, there are 9 sets which could not be clustered correctly in our method. A possible improvement is that we use all definitions of words in the dictionary. We selected the first top 5 definitions in the dictionary for each noun and used them in the experiment. However, there are some words of which the meanings are not included these selected definitions. This causes the fact that it is hard to get a higher percentage of correct clustering. Another interesting possibility is to use an alternative weighting policy, such as the *widf* (*weighted inverse document frequency*) (Tokunaga, 1994). The *widf* is reported to have a marked advantage over the *idf* (*inverse document frequency*) for the text categorisation task.

6 Conclusion

We have reported an experimental study for clustering of articles by using on-line dictionary definitions and showed how dictionary-definition can use effectively to classify articles, each of which belongs to the restricted subject domain. In order to cope with the remaining problems mentioned in section 5 and apply this work to practical use, we will conduct further experiments.

References

P. F. Brown et al., 1991. Word-Sense Disambiguation Using Statistical Methods. In *Proc. of the 29th Annual Meeting of the ACL*, pp. 264-270.

E. Brill, 1992. A simple rule-based part of speech tagger. In *Proc. of the 3nd conference on applied natural language processing*, ACL, pp. 152-155. Trento, Italy, 1992.

K. W. Church et al., 1991. Using Statistics in Lexical Analysis, *Lexical acquisition: Exploiting on-line resources to build a lexicon.* (Zernik Uri (ed.)), pp. 115-164, London, Lawrence Erlbaum Associates.

L. Guthrie and E. Walker, "DOCUMENT CLASSIFICATION BY MACHINE: Theory and Practice", In *Proc. of the 15th International Conference on Computational Linguistics*, Kyoto, Japan, 1994, pp. 1059-1063

N. Jardine and R. Sibson, 1968. The construction of hierarchic and non-hierarchic classifications. In *Computer Journal*, pp. 177-184.

K. S. Jones, 1973. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28 (1973) 1, pp. 11-21.

M. Liberman, editor. 1991. *CD-ROM I*, Association for Computational Linguistics Data Collection Initiative, University of Pennsylvania.

Y. Niwa and Y. Nitta, 1995. Statistical Word Sense Disambiguation Using Dictionary Definitions In *Proc. of the Natural Language Processing Pacific Rim Symposium '95*, Seoul, Korea, pp. 665-670.

G. Salton and M. J. McGill, 1983. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

T. Tokunaga and M. Iwayama, 1994. Text Categorisation based on Weighted Inverse Document Frequency IPSJ SIG Reports, 94-NL-100, 1994.

D. Yarowsky, "Word sense disambiguation using statistical models of Roget's categories trained on large corpora", In *Proc. of the 14th International Conference on Computational Linguistics*, Nantes, France, 1992, pp. 454-460

N. Yuasa et al., 1995. Classifying Articles Using Lexical Co-occurrence in Large Document Databases In *Trans. of Information Processing Society Japan*, pp. 1819-1827, 36 (1995) 8.

D. Walker and R. Amsler, 1986. The Use of Machine-Readable Dictionaries in Sublanguage analysis, *Analyzing Language in Restricted domains*, (Grishman and Kittredge (ed.)), pp. 69-84, Lawrence Erlbaum, Hillsdale, NJ. (1987) 2.