2006－DBS－140 (II) (55)
2006／7／13

Word Space:

A New Approach to Describe Word Meanings

Hiroshi OHNISHI†, Yuichi YAGUCHI†, Kenta YAMAKI††, Ryuichi OKA†, and Keitaro NARUSE†

† University of Aizu, 90, Kamiiawase, Tsuruga, Ikkimachi, Aizuwakamatsu, Fukushima, 965–8580 Japan
E-mail: ††{m5101114,m5101127,oka,naruse}@u-aizu.ac.jp, ††drgarugari@gmail.com

Abstract The purpose of this research is to acquire new knowledge about the meanings of words by arranging the words in space. We allocated the words and displayed their relationship with each other in a three-dimensional space with a method called Associated Keyword Space (ASKS). In the experiment, the data obtained by ASKS were compared with Thesaurus.com and WordNet, conventional methods of meaning description. The result of the experiment showed two important points. First, the strength of each meaning of a verb with polysemy was expressed by the visual relationship with and distance to associated words. Second, polysemy of the verb was expressed by the extension of the synonyms.

Key words Nonlinear Multidimensional Scaling, Association Words, Word Similarity

1. Introduction

Internet web pages contain a large number of sentences. Written content such as news, weblogs, reports and other information are proliferating because Internet technology allows people more easily to write opinion and discussion on web pages. In addition, a large number of historical records have been recently archived, and users want to retrieve this content more quickly and correctly. Thus, a retrieval system needs to use natural language analysis to understand and use the context of sentences to increase retrieval quality.

In the field of language analysis, the importance of computer analysis is increasing because such analysis can extract new knowledge from a large amount of data, and this knowledge exposes people to novel ideas [5] [8]. There are many heuristic language analysis and computer language analysis methods. Many language analysis methods use a special-purpose corpus [10] [2] or a general corpus such as the British National Corpus [11] and the American National Corpus [1]. Corpus-based language analysis attempts to find grammatical knowledge [3] [9] and usage of words derived from word co-occurrence [13].

On the other hand, word-cluster-based language analysis is also important in natural language analysis. A thesaurus [7] is one of the results of word-cluster-based language analysis. Words in the same class are transposable with each other and have the same conceptual meaning. Thus, if the word classes have a characteristic usage in a different context, there is the possibility of finding a hierarchy of word classes from usage because it can analyze the lexis from usage and co-occurrence rules. WordNet [4] is a type of word clustering that has proved most effective for finding the meaning of words. WordNet describes the relations between words as a network, with words as nodes and the arcs between nodes as representing the relationships between words.

However, these methods have some problems. First, the semantic interpretation of each word is ambiguous. Second, even if the distance between two words is shown by the number of arcs connecting them, the validity of this measure is not persuasive. In addition, words in the sentence have different contextual meanings even if the words are semantically transposable. WordNet and Thesaurus could not reflect them, but a word corpus may.

This paper proposes a new meaning description method that expresses the semantic distance between words by arranging the words in space. This method, which is a sort of mining method, is called the Associated Keyword Space (ASKS) [12]. ASKS creates a multidimensional space, called an association space, and words are embedded in this space according to data obtained from affinity between words. Our tool [14] shows the three-dimensional positions of the words within this space by visualizing the association space.

The main contribution of this research is that we describe the stronger meanings and the polysemy of the verbs visually. ASKS shows the following features. Synonyms that are located nearer in the association space have a stronger relationship to each other with respect to word usage. The spread of the synonyms shows word ambiguity.

The rest of this paper is organized as follows: section 2 explains ASKS, which is our proposed method; section 3 expounds on our experiment and shows the result of comparing ASKS with other methods; section 4 discusses the knowledge obtained; section 5 presents our future work; and section 6 concludes the paper.

2. Associated Keyword Space (ASKS)

Quantification Method Type IV (Q-IV) [6] is well known as a linear multidimensional scaling (MDS) method. ASKS is a nonlinear version of MDS and is strong for noisy data [12]. This section explains what ASKS is and how to calculate it.

2.1 Distance Measure of ASKS

Let N denote the spatial dimension of an allocated object. Each

object is numbered by i and its location is defined by x_i. The distance is measured by the following formula F:

$$d_{ij} = -F(x_j - x_i). \tag{1}$$

F has a parameter a and is defined as:

$$F(k) = \begin{cases} |k|^2 & (|k| < a) \\ 2a|k| - a^2 & (|k| \geqq a). \end{cases} \tag{2}$$

Figure 1 shows a plot of this function.

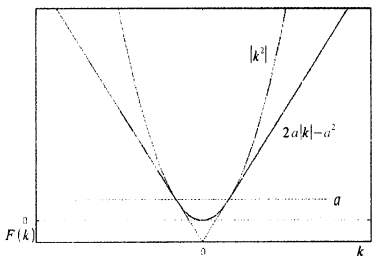

Figure 1 Nonlinear function used in ASKS

2.2 Uniformalization of the Distribution in ASKS

Three kinds of constraints on the distribution of objects are defined to decide the amount of space allocated to similar objects in distinguishable clusters.

(1) Make the original point the center of gravity for the samples.

(2) Obtain covariance matrices such that dispersion in any direction creates the same value.

(3) Uniformalize the samples in a radial direction.

Figure 2 shows the method for uniformalization of the distribution in the super-sphere.

Uniformalization is useful for clustering of noisy data that otherwise tends to distribute connections too evenly across the data.

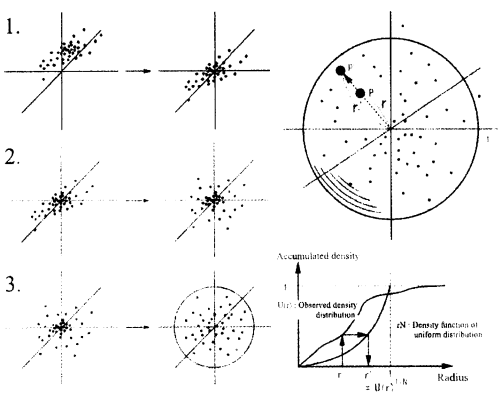

Figure 2 Uniformalization types used in ASKS

2.3 Iterative Solution of Nonlinear Optimization

The criterion function of ASKS is as follows.

$$J(x_1, x_2, \ldots, x_n) = \sum_i \sum_j \{-M_{ij} F(x_j - x_i)\} \rightarrow \max \tag{3}$$

M_{ij} is an affinity (a nonnegative value) between objects i and j. M_{ij} is calculated from the co-occurrence of objects i and j. The partial derivative of J with respect to x_i gives the following formula for determining the values of x_i that maximize J:

$$\frac{\partial}{\partial x_i} \sum_i \sum_j \{-M_{ij} F(x_j - x_i)\} \equiv 0, \tag{4}$$

$$\sum_j M_{ij} F'(x_j - x_i) \equiv 0. \tag{5}$$

The derivative of F is:

$$F'(k) = \begin{cases} 2k & (|k| < a) \\ 2a\frac{k}{|k|} & (|k| \geqq a). \end{cases} \tag{6}$$

Next, defining D by:

$$D(k) = \begin{cases} 2 & (|k| < a) \\ \frac{2a}{|k|} & (|k| \geqq a), \end{cases} \tag{7}$$

we procure the following expression:

$$F'(x_j - x_i) = D(x_j - x_i)(x_j - x_i). \tag{8}$$

The following iterative computation converges to the solution x_i.

$$x_i^{t+1} = \frac{\sum_j M_{ij} D(x_j^{(t)} - x_i^{(t)}) x_j^{(t)}}{\sum_j M_{ij} D(x_j^{(t)})} \tag{9}$$

The three constraints must be enforced in every step of the iterative computation for all variables x_i $(i = 1, 2, \ldots, n)$.

2.4 Comparison of Q-IV and ASKS

The effectiveness of ASKS is shown by comparing its performance with Q-IV.

Assuming that 1,000,000 objects were to be clustered into categories (C) of 100, 1,000, and 10,000 objects, we generated a set of affinity data between objects $M_{ij} (1 \leqq i \leqq C, 1 \leqq j \leqq C)$, where each M_{ij} took a value of 1 or 0 corresponding to whether objects i and j belonged to the same category or not. We counted the number of the former case, (N_i), and the latter case, (N_o), and then we defined R_i as the sum of the affinities in a class for the former case and R_o as the sum of the affinities between classes for the other case. If objects i and j belonged to the same category, then M_{ij} was set to $M_{ij} = 1$ with a probability of R_i/N_i, and the other values of M_{ij} were set to $M_{ij} = 0$. If these belonged to different categories, in the same way, the value of M_{ij} was set to $M_{ij} = 1$ according to R_o/N_o. The ratio of R_o/R_i expressed the level of noise, where a value of zero denoted no noise, and larger values (which could be > 1.0) denoted a high level of noise.

Both methods were applied to the case of 1,000 categories. The clustering results for the Q-IV approach are shown in Figure 3. The Q-IV method is characterized using a linear optimization and standard distributions of the various noise levels. Only 20,000 objects belonging to 20 categories are plotted for the purpose of visualization. The results for the ASKS method under the same conditions

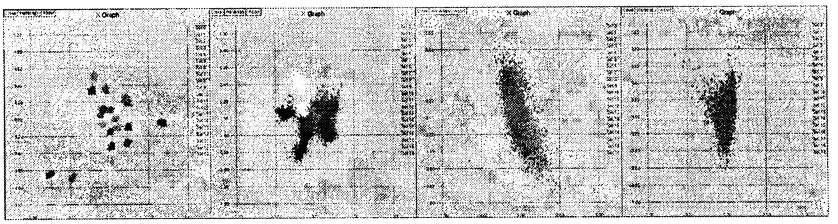

Figure 3 Allocation of items by Q-IV. Noise level (Ro/Ri) [left = 0.01, 0.1, 1.0, right = 100.0]

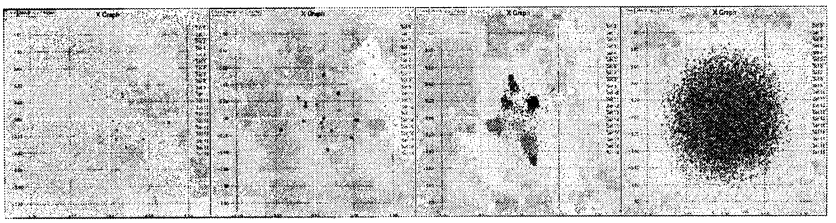

Figure 4 Allocation of items by ASKS. Noise level (Ro/Ri) [left = 0.01, 0.1, 1.0, right = 100.0]

are shown in Figure 4. The ASKS method is characterized using a nonlinear optimization and a uniform distribution of the various noise levels. The results show the superiority of the ASKS technique over the Q-IV approach, because ASKS can gather objects belonging to the same category in a more compact space and can distinguish categories at higher noise values.

3. Experiment

Three steps are necessary to extract knowledge using ASKS:

(1) Make an affinity matrix from the co-occurrences between each word.

(2) Use the ASKS algorithm to allocate each word to a point in association space.

(3) Search features in the relationships between words and compare these with conventional dictionaries.

In this research, verbs are the targets of analysis because meanings of verbs are changed a lot by surrounding words. The data used were obtained from 40,008 sentences in journal articles about human interfaces containing 29,918 unique words.

3.1 Calculation of the Affinity Matrix

A major problem for this research is determining which calculation formula for M_{ij} (3) is best.

The distance between each word in a sentence is denoted by R. For example, in the sentence "I have a pen", the R between "I" and "pen" is 3. The frequency that the distance of word i to word j is R in all sentences is defined by C_{ijR}. Let L denote the maximum distance R over which to calculate the affinity. We experimented using four different formulas for M_{ij}, listed below, with $L = 8$.

$$M_{ij} = \sum_{k=1}^{L} C_{ijk} \tag{10}$$

$$M_{ij} = \sum_{k=1}^{L} C_{ijk}/k \tag{11}$$

$$M_{ij} = \sum_{k=1}^{L} C_{ijk}/k^2 \tag{12}$$

$$M_{ij} = \sum_{k=1}^{L} C_{ijk}((L+1)-k) \tag{13}$$

One day was taken to obtain all the affinities between each word by these calculations in this experiment.

3.2 Visualization of the Word Relationships

We developed a visualization tool named Visualize ASKS to project the data onto a three-dimensional association space. Using this tool, we can understand the positional relations of words visually and intuitively, as shown in Figure 5(a). Visualize ASKS can represent the density around each word and the relational tree by drawing a link path (Figure 5(b)).

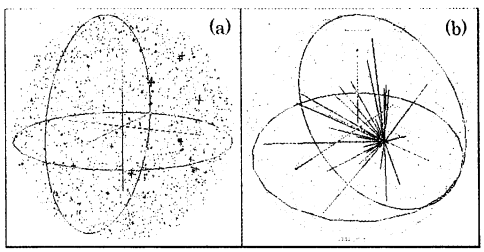

Figure 5 Visualization tool: (a) global view, (b)link path of a word

3.3 Evaluation of Thesaurus Data by ASKS

The aim here is to examine how synonyms are arranged in the neighborhood of each verb. The synonym data were acquired from Thesaurus.com (*http://thesaurus.reference.com/*).

Table 1 shows the rate that synonyms exist within a neighborhood of radius r in association space. The synonyms do not seem to be simply correlated with the neighborhood structure because the percentages are not very high.

Table 1 Percentage of synonyms included in the neighborhood.

Formula	Content rate
$r = 0.1$	
$M_{ij} = \sum_{k=1}^{L} C_{ijk}$	2.096%
$M_{ij} = \sum_{k=1}^{L} C_{ijk}/k$	2.140%
$M_{ij} = \sum_{k=1}^{L} C_{ijk}/k^2$	1.933%
$M_{ij} = \sum_{k=1}^{L} C_{ijk}((L+1)-k)$	2.216%
$r = 0.05$	
$M_{ij} = \sum_{k=1}^{L} C_{ijk}$	2.339%
$M_{ij} = \sum_{k=1}^{L} C_{ijk}/k$	3.030%
$M_{ij} = \sum_{k=1}^{L} C_{ijk}/k^2$	2.551%
$M_{ij} = \sum_{k=1}^{L} C_{ijk}((L+1)-k)$	3.010%
$r = 0.01$	
$M_{ij} = \sum_{k=1}^{L} C_{ijk}$	4.819%
$M_{ij} = \sum_{k=1}^{L} C_{ijk}/k$	5.263%
$M_{ij} = \sum_{k=1}^{L} C_{ijk}/k^2$	3.896%
$M_{ij} = \sum_{k=1}^{L} C_{ijk}((L+1)-k)$	7.042%

3.4 Comparison between ASKS and WordNet

How the synonyms classified according to WordNet are arranged in Association Space is shown here. The synonyms of some words in WordNet are shown in Table 4.. Data that apply the formula (13) are used for comparison because these data have the highest number of synonyms when compared with the Thesaurus, as shown in Table 1.

Figure 6 shows synonyms of the verb "use" (Table 4.(a)) and its relational path in association space. We can see that some synonyms are at a position far away from the verb "use", and are distributed over a large range in Figure 6.

Then, we examined the words surrounding each synonym. Figure 7 shows that the verb "exercise", which is a synonym of the verb "practice" (Table 4.(b)), was arranged near "practice". In Figure 8, the noun "cash", which means money, was seen in a close vicinity of the verb "expend" with the meaning of "pay out" in Table 4.(c), though these synonyms were not seen in the neighborhood. Unfortunately, no interesting word is in the neighborhood of the verb "employ", and the synonyms are located a long way away, as shown in Figure 9.

Furthermore, the arrangement of the synonyms of the verb "comprehend" is shown in Figure 10. Two synonyms, "cover" and "perceive", of the verb "comprehend", were seen to be very near, although the meaning of "comprehend" as the synonym of "cover" is different to that of the synonym of "perceive", as shown in Table 4.(e).

3.5 Other Features

Two interesting tendencies were found in association space. One is that the verbs that mean repeating something tend to be arranged in close vicinity of the original word, as shown in Figure 11 . Another is that a proper noun, such as a person's name, is inclined to settle in one place, as shown in Figure 12.

4. Discussion

Looking at the relation of the verb "use" to its synonyms, we can consider that the verb "use" frequently acts as a synonym for the verbs "utilize", "utilise", and "apply", which are located nearer than the verbs "practice" and "expend". Thus, it seems that the meaning

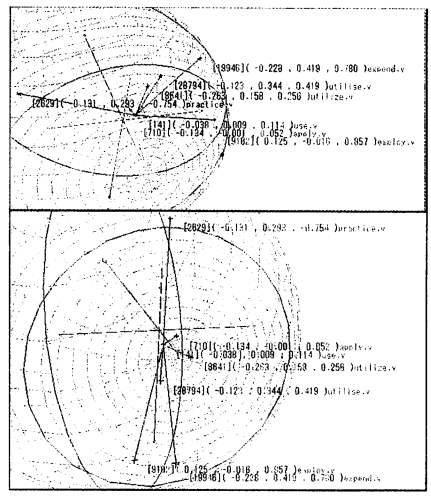

Figure 6 Synonyms of the verb "use".

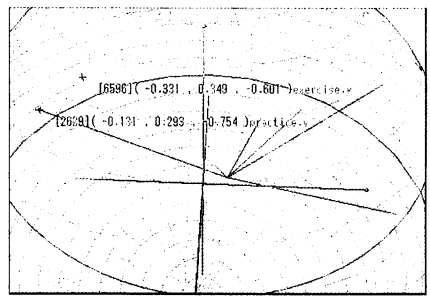

Figure 7 The words "practice" and "exercise" in association space.

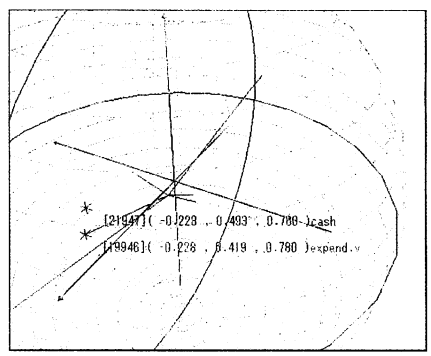

Figure 8 Surrounding of the verb "expend".

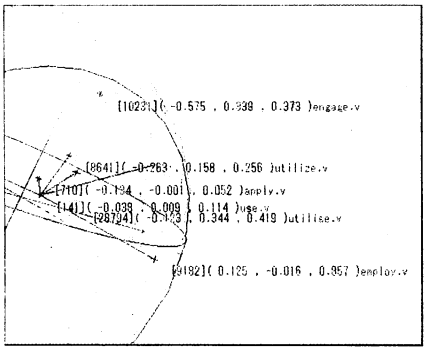

Figure 9 Synonyms of the verb "employ"

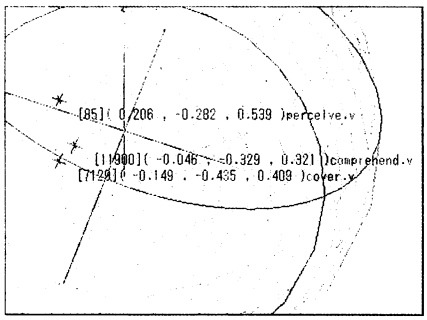

Figure 10 Synonyms of the verb "comprehend"

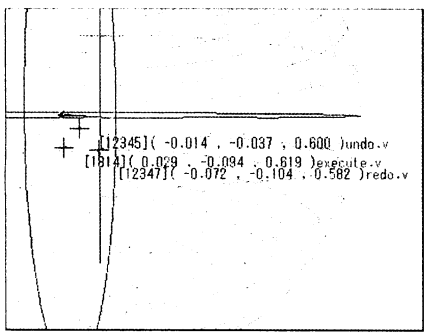

Figure 11 Surrounds of the verb "excuse".

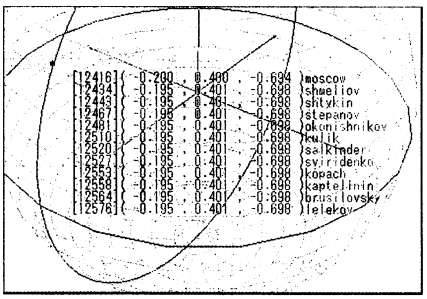

Figure 12 Person's name in association space.

of "put into service" is strong with the verb "use". Moreover, we can assume that the meaning of the verb "practice" as a synonym of "exercise" is stronger than as a synonym of "use", and the meaning of the verb "expend" as "pay out" is stronger than as a synonym of "use". Using ASKS, the strength of each meaning of a verb with polysemy can be distinguished visually by observing the relationship of the verb with its synonyms.

The extension of the synonyms of the verb "comprehend" in Figure 10 was smaller than that of the verb "use". In addition, the words where the meaning is uniquely decided, like proper nouns, tend to gather in a tight space, as shown in Figure 12. Here, when the meaning space of a word is defined by an area that includes the word and its synonyms, we can consider that the ambiguity of the word is described visually by the shape and size of the meaning space. According to this idea, the ambiguity of the verb "use" is larger than that of the verb "comprehend". WordNet actually showed that the verb "use" has more meanings than the verb "comprehend", as shown in Table 4..

5. Future Work

We have identified two directions for future work. The first is to improve the quality of the meaning description method, and the second is to evaluate differences in word arrangement in association space using different input databases.

For the first direction, three methods have been devised: expanding the input database, searching for a better computational method with the affinity matrix, and the further analysis of the meaning of positional relationships in association space. We would like to use at least 100,000 or more sentences in our next experiment.

In the other research direction, the evaluation of the differences of the positions of words having the same meaning using texts of different fields or different languages is interesting. We expect that features of the input text will be found.

6. Conclusion

This paper describes novel knowledge about words allocated in a three-dimensional space using ASKS.

In this experiment, two interesting features were found by projecting the relationships between synonyms of the verbs in WordNet to an association space. Synonyms that were near a verb and synonyms that were far away were both seen. When investigating the surroundings of the synonyms, words that were associated with other meanings of the synonyms existed near the other synonyms or class words.

Table 2 Synonyms in WordNet

Word (verb)	Group of the Synonyms	Meaning
(a) use	use, utilize, utilise, apply, employ	put into service, make work or employ
	use, habituate	take or consume
	use	seek or achieve an end by using to one's advantage
	use, expand	use up, consume fully
	practice, apply, use	avail oneself to
	use	habitually do something
(b) practice	drill, exercise, practice, practise	learn by repetition
	practice, apply, use	avail oneself to
	practice, practise, exercise, do	carry out or practise; as of jobs and professions
	rehearse, practise, practice	engage in a rehearsal
(c) expend	use, expend	use up, consume fully
	spend, expend, drop	pay out
(d) employ	use, utilize, utilise, apply, employ	put into service
	hire, engage, employ	engage or hire for work
(e) comprehend	grok, comprehend, savvy, dig, grasp, compass, apprehend	get the meaning of something
	perceive, comprehend	to become aware of through the senses
	embrace, encompass, comprehend, cover	include in scope
(f) engage	prosecute, engage, pursue	carry out or participate in an activity
	absorb, engross, engage, occupy	engage or engross wholly
	hire, engage, employ	engage or hire for work
	engage	ask to represent; of legal counsel
	betroth, engage, affiance, plight	give to in marriage
	engage	get caught
	engage, wage	carry on
	engage, enlist	hire for work or assistance
	lease, rent, hire, charter, engage, take	engage for service under a term of contract
	engage, mesh, lock, operate	keep engaged

As a result, we were able to determine which meaning of the verb was stronger by visual inspection. Furthermore, the polysemy of a verb was able to be expressed visually by defining the extensions of the synonyms as the meaning space. This means that we succeeded in expressing the knowledge of the words visually.

We hope to mine new knowledge of words from association spaces in future experiments.

7. Acknowledgment

We wish to thank Mr. Matsumura (President of Mediadrive, Inc.) and Mr. Mitsumori (Engineer at Mediadrive, Inc.) for their cooperation with this experiment.

References

[1] American National Corpus Project, "American national corpus," 2005. url; http://americannationalcorpus.org/.

[2] G. Barnbrook, Language and Computers: A Practical Introduction to the Computer Analysis of Language, Edinburgh Textbooks in Empirical Linguistics, Columbia University Press, New York, NY, 1996.

[3] J. Carroll, G. Minnen, and T. Briscoe, "Corpus annotation for parser evaluation," Proc. of the EACL workshop on LINC, June 1999.

[4] C. Fellbaum, ed., WordNet: An Electronic Lexical Database, Language, speech, and communication series, The MIT Press, Cambridge, MA, 1998.

[5] R. Gaizauskas, "Evaluation in language and speech technology," Computer Speech & Language, vol.12, no.4, pp.249–262, October 1998.

[6] T. Komazawa and C. Hayashi, Quantification Theory and Data Processing, Asakura Shoten, Shinjuku, Tokyo, Japan, 1986. (in Japanese).

[7] Lexico Publishing Group, LLC, "Thesaurus.com," 2006. url; http://thesaurus.reference.com/.

[8] R. Lopes-Cozar, A.J. Rubio, P. Garcia, and J.C. Segura, "A spoken dialogue system based on a dialogue corpus analysis," In proceedings of the LREC'98, pp.55–58, May 1998.

[9] M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz, "Buiding a large annotated corpus of english: The penn treebank," Computational Linguistics, vol.19, no.2, pp.313–330, June 1993.

[10] T. McEnery and A. Wilson, Corpus Linguistics, Edinburgh Textbooks in Empirical Linguistics, Columbia University Press, New York, NY, 1996.

[11] Oxford University Computing Services, "British national corpus," February 2006. url; http://natcorp.ox.ac.uk/index.html.

[12] H. Takahashi and R. Oka, "Self-organization an associated keyword space for text retrieval," Proceedings of WMSCI2001, pp.302–307, July 2001.

[13] J. Xu and W.B. Croft, "Corpus-based stemming using cooccurrence of word variants," ACM Trans. on Information System, vol.16, no.1, pp.61–81, January 1998.

[14] Y. Yaguchi, H. Ohnishi, S. Mori, K. Naruse, R. Oka, and H. Takahashi, "A mining method for linked web pages using associated keyword space," Proceedings of SAINT2006, pp.268–276, January 2006.