Analysis of the Semantic Network Structure of Japanese Word Associations: An Investigation of Clustering Granularity with Two Extracted Sub-Networks

Maki Miyake[1] **and Terry Joyce**[2]
(1) Graduate School of Language and Culture,Osaka University, 1-8 Machikaneyama-cho, Toyonaka-shi, Osaka, 560-0043, Japan
(2) School of Global Studies, Tama University, 802 Engyo, Fujisawa, Kanagawa 252-0805, Japan

Abstract

This paper presents some network analyses of a large-scale semantic network representation of the Japanese Word Association Database (JWAD). Version 1 of the JWAD consists of the word association responses made to a selection of approximately 2,100 basic Japanese kanji and words from up to 50 respondents. Graph representation and graph theory techniques are particularly promising methods for detecting and perceiving the intricate patterns of connectivity within large-scale linguistic knowledge resources. This paper focuses on the structure of the JWAD association network representation from the perspectives of two important statistical features; namely, the distribution in node connections and the clustering coefficient as an index of the interconnectivity strength between neighboring nodes. The developed association network is shown to exhibit scale-free characteristics and a pattern of sparse connectivity. The Recurrent Markov Clustering (RMCL) method of graph clustering is also applied to the JWAD representation. The RMCL improves on van Dongen's Markov Clustering algorithm by making it possible to adjust the proportions of cluster sizes, thereby providing greater control over the sizes of concept domains by modifying graph granularity. RMCL clustering yields structurally simpler network representations which can be utilized in hierarchically organizing semantic spaces. Accordingly, the technique is especially useful for the visualization of large-scale linguistic resources.

1. Introduction

In striving to extend our understanding of lexical knowledge, various disciplines within cognitive science, including psychology and computational linguistics, are seeking to unravel the rich networks of associations that connect words together. Key methodologies for that enterprise are the techniques of graph representation and their analysis that allow us to discern the patterns of connectivity within large-scale resources of linguistic knowledge and to perceive the inherent relationships between words and word groups. Although some studies have shown reasonable degrees of success in applying variations of the multidimensional space model, such as Latent Semantic Analysis and multidimensional scaling, in the analyses of texts, graph theory and network analysis methodologies are undoubtedly well suited for detecting and perceiving patterns of connectivity within large-scale resources of association knowledge and for tracing out the inherent relationships between words and word groups. For instance, a number of studies have recently applied graph theory approaches, as alternatives to computational methods based on word frequencies, in investigating various aspects of linguistic knowledge resources, such as employing graph clustering techniques to detect lexical ambiguities and to acquire semantic classes in the study by Dorow, Widdows, Ling, Eckmann, Sergi and Moses (2005). More relevant to the present paper are two studies conducted by

pp. 273–280 © 2008 The Organizing Committee of the International Meeting of the Psychometrics Society

Steyvers, Shiffrin, and Nelson (2005), and Steyvers, and Tenenbaum (2005) that directly address word association knowledge. Both studies utilize one of the largest databases of word associations compiled for American English by Nelson, McEvoy, and Schreiber (1998). Steyvers, and Tenenbaum (2005) observed, for instance, interesting similarities between three semantic networks in terms of their scale-free patterns of connectivity and small-world structures, when they applied graph theory and network analysis techniques in investigating the networks'structural features. In a similar spirit, this paper applies a range of network analyses investigating the structural characteristics of an association network representation of the large-scale Japanese Word Association Database (JWAD), under ongoing construction by Joyce (2005); Joyce (2006); Joyce (2007).

After briefly introducing the JWAD and the association network representation created from it in the Section 2, Section 3 presents the results from some basic statistical analyses concerning the structural characteristics of the network, such as degree distributions and average clustering coefficient distributions. The network is also clustered by applying the Recurrent Markov clustering (RMCL). Section 4 reports on Markov clustering (MCL) applied with different parameters for clustering granularity to two sub-networks that were extracted from the JWAD network with different clustering coefficients as a threshold value. Finally, the two measures of modularity and the F measure are employed in assessing the quality of the divisions within the network.

2. Japanese Word Association Database

This section outlines the ongoing development of a semantic network representation of Japanese word associations. After briefly noting some existing word association norms as frames of reference for the Japanese Word Association Database (JWAD) project (Joyce (2005); Joyce (2006)), the JWAD and its semantic network representation are introduced.

2.1. Existing word association norms

Although comprehensive word association norm data has been available for some time for English (see Moss and Older (1996) for British English and Nelson, McEvoy, and Schreiber (1998) for American English), a comparably large-scale database is currently being constructed for Japanese (Joyce (2005); Joyce (2006); Joyce (2007)). In sharp contrast to the early survey by Umemoto (1969), which gathered free associations from 1,000 university students for a very small set of 210 words, and the data collected by Okamoto and Ishizaki (2001), which includes 10 responses for 1,656 nouns, the JWAD survey list of 5,000 basic Japanese kanji and words is clearly far more extensive. One particular important feature of the JWAD is that the data consists solely of free word association responses.

2.2. Questionnaire surveys

The majority of the word association responses for JWAD have come from two surveys, in which questionnaires were administered to 1,481 native Japanese university students (929 males and 552 females; average age = 19.03, SD = 0.97). In both surveys, a questionnaire consisted of 100 items, and the free word association task required participants to look at each printed item and write down the first semantically-related Japanese word that came to mind. The first survey collected up to 50 responses for a random sample of approximately 2,000 items, while the second survey collected at least ten responses for the remaining 3,000 items. In order to collect the large-scale quantities of association responses required in the ongoing construction of the JWAD, a web-based version of the free word association survey has also been developed(http://nerva.dp.hum.titech.ac.jp/terry/index.jsp).

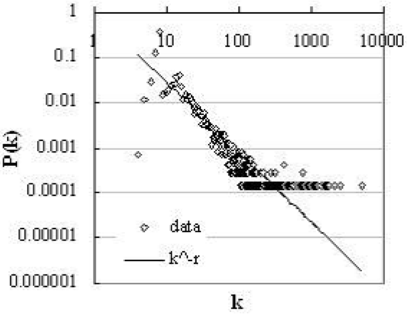

Fig. 1 Degree distribution.

2.3. Semantic network representation

As already noted, the JWAD is the core component in a project to investigate lexical knowledge in Japanese by mapping out Japanese word associations (Joyce (2005); Joyce (2006); Joyce (2007)). Version 1 of the JWAD consists of the word association responses collected from up to 50 respondents for a list of 2,099 items, which was randomly extracted from the present project corpus of 5,000 basic Japanese kanji and words. Because the JWAD employs the free word association task in collecting association responses, the JWAD data more faithfully reflects the rich and diverse nature of word associations. In order to construct the association network representation from the JWAD database, only response words with a response frequency of two or more were extracted, which resulted in a network graph consisting of 8,970 words.

3. Analysis of JWAD Network Structures

Graph representation and the techniques of graph theory and network analysis offer very effective ways of detecting and investigating the intricate patterns of connectivity that exist within large-scale linguistic knowledge resources. Steyvers, and Tenenbaum (2005) have, for example, conducted a particularly interesting study that examined the structural features of three semantic networks based on Nelson, McEvoy, and Schreiber (1998) word association database, WordNet (Fellbaum, 1998), and Roget's thesaurus (Roget (1991)), respectively. Through their calculations for a range of statistical features, including the average shortest paths, diameters, clustering coefficients, and degree distributions, Steyvers and Tenenbaum found striking similarities between the three networks in terms of their scale-free patterns of connectivity and small-world structures. In analyzing the characteristics of the semantic network representation of the JWAD, this present study also calculates its degree distribution and the clustering coefficient, as an index of the interconnectivity strength between neighboring nodes within the graph. This study also reports on the application of some graph clustering techniques to the constructed JWAD association network representation.

3.1. Degree distribution

Based on their computations of degree distributions, Barabasi and Albert (1999) have argued that in the case of network structures the degree distribution P(k) should correspond to a power law, which can be expressed as $P(k) \approx k^{-\gamma}$.

As Figure 1 illustrates, the degree distribution (P(k)) of word occurrences in the JWAD network conforms to a power-law, for which the best fit power function has an r exponent value of 2.1. The average degree value of 3.3 (0.03%) for

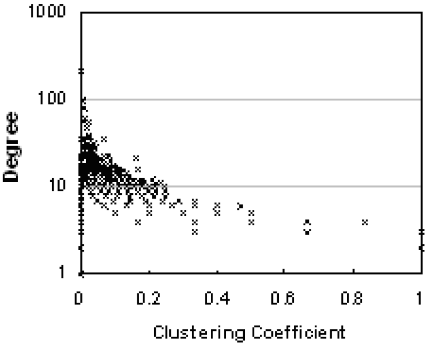

Fig. 2 Clustering coefficient vs. degree.

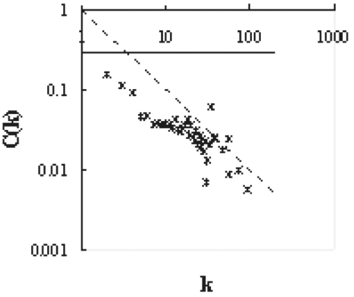

Fig. 3 Clustering coefficient distribution.

the complete semantic network of 8,970 nodes clearly indicates that the network exhibits a pattern of sparse connectivity; in other words, that it possesses the characteristics of a scale-free network.

3.2. Clustering coefficient

In a study of social networks looking at the probabilities that an acquaintance of an acquaintance is also your acquaintance, Watts and Strogatz (1998) have proposed the notion of clustering coefficient as an appropriate index of the degree of connections between nodes. In this study, we define the clustering coefficient of n nodes as:

$$C(n) = \frac{number\ of\ links\ among\ n's\ neighbors}{N(n)(N(n)-1)/2}$$

where N(n) represents the number of adjacent nodes. Accordingly, a clustering coefficient is a value between 0-1. Figure 2 plots clustering coefficients as a function of degree. The average clustering coefficient is 0.03, indicating that the complete network basically consists of many star graphs connected together. The clustering coefficient for 7,363 nodes (82% of the total) is 0. There are 148 nodes that have a clustering coefficient value of 1, which indicates that each node connects to only a few other nodes and that these together form small complete graphs.

Moreover, Ravasz and Barabasi (2003) advocate a similar notion of clustering coefficient dependence on node degree, based on the hierarchical model of $C(k) \approx k^{-\beta}$ where β is the hierarchical exponent. The results of scaling C(k) with k for the JWAD network are presented in Figure 3. The solid lines in the figure correspond to the average clustering coefficient. As the JWAD network appears to conform to a power law, we may assume that the network possess an intrinsic hierarchy.

3.3. Recurrent Markov Clustering

While Markov Clustering (MCL) is widely recognized as an effective method for grouping related items within large and sparsely connected data structures, particularly for large-scale corpora (Dorow, Widdows, Ling, Eckmann, Sergi and Moses (2005), Steyvers, and Tenenbaum (2005)), a serious shortcoming with the method is the fact that it is impossible to control for the distribution in the sizes of clusters that are generated. Recently, Jung, Miyake, and Akama (2006) have proposed an improvement to the basic MCL method called Recurrent Markov Clustering (RMCL), which offers some control over cluster sizes through the adjustment of graph granularity. The recurrent process is essentially achieved by incorporating feedback data about the states of overlapping clusters prior to the

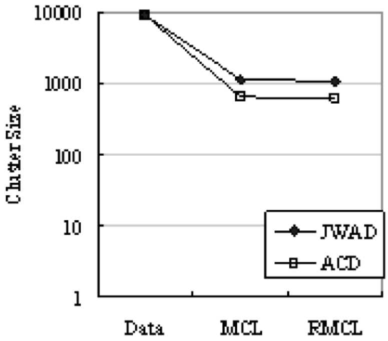

Fig. 4 Cluster sizes results for MCL and RMCL.

final MCL output stage. The reverse tracing procedure, which is a key feature of the RMCL, makes it possible to generate a virtual adjacency matrix for non-overlapping clusters from the convergent state in the MCL process. The resultant condensed matrix represents a simpler graph that can highlight the conceptual structures that underlie similar words. In this study, a value of 1.5 is adopted for the inflation parameter. Accordingly, these clustering stages were used in the RMCL process. Thus, the MCL yielded 1,144 hard clusters (average cluster size = 5.5, SD = 7.2), while the RMCL yielded 1,084 clusters, where the average number of cluster components was 1.1 with an SD of 0.28. Figure 4 presents the MCL and the RMCL cluster sizes for the JWAD networks, illustrating the downsizing transitions that occurred during the graph clustering process.

4. Graph Clustering for Two Sub-Networks

The fact that 82% of all the nodes have a clustering coefficient of 0 highlights the sparseness of the network. This low level of connectivity is undoubtedly due to the fact that the JWAD survey corpus has been compiled to be representative of basic Japanese vocabulary, and so it includes items from a wide range of semantic categories. Accordingly, in order to sample words with higher levels of connectivity, two sub-networks are created by adjusting the clustering coefficient. We briefly present and discuss the results of applying the clustering methods to these two sub-network representations.

4.1. Applied method

Using the clustering coefficient as a threshold, the JWAD network was reduced into two sub-networks. For one of the sub-networks the clustering coefficient was set to a value of 0.03, which is approximately the average for all words, while for other the value was set to 0.1 to investigate the relationships between the words. These settings resulted in one sub-network consisting of 1,305 words that had clustering coefficients of 0.03 or greater (C0.03) and one sub-network consisting of 455 words with clustering coefficients of 0.1 or greater (C0.1). The C0.03 network consists of 237 cue words (52% of the total), while the C0.1 has 847 cues (65%). As these sub-networks are relatively small in size, only the MCL algorithm was applied to them, but with a range of different inflation parameters that influence the clustering granularity of the process. If the value of r is set high, then the resultant clusters will tend to be small in size. While this parameter is typically set as r = 2, Vechthomova, Vechthomova, Gfeller and Chappelier (2005) claim that a value of 1.6 is reasonable in their study to create a dictionary of French synonyms. In order to optimize the inflation parameter, we employ Newman and

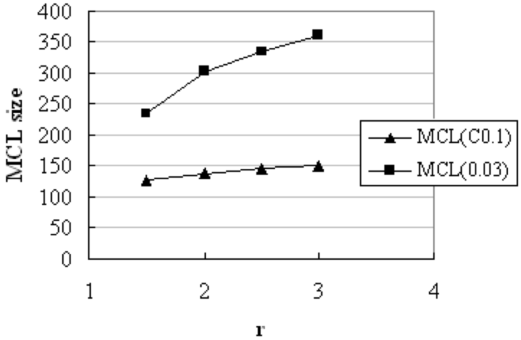

Fig. 5 MCL sizes as a function of r.

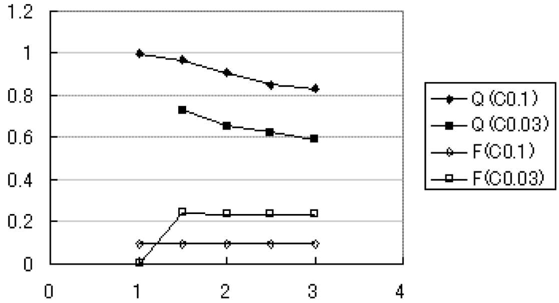

Fig. 6 Q & F values as a function of r.

Girvan (2004) notion of modularity, which is particularly useful in assessing the quality of divisions within a network. Modularity Q highlights the differences in edge distributions between graphs with meaningful partitions and random graphs under identical vertices conditions (numbers and sum of their degrees). In scale-free networks, high Q values are rare, with the values usually settling within a range of between 0.3 and 0.7. Moreover, to consider the recall rates of nodes, the F measure is also employed to optimize the selection of the most appropriate results, because the precision rate, P, always depends on a trade-off relationship between Modularity Q and the recall rate R. For the two JWAD sub-networks, the MCL was processed with r values ranging from 1.5 to 3, with the two measures of modularity and the F measure being calculated for each result.

4.2. MCL clustering results

The MCL process generated a nearly-idempotent stochastic matrix at around the 15th clustering stage. Figure 5 plots the MCL cluster sizes as a function of the inflation parameter r ranging from 1.5 to 3. In the case of the C0.03 sub-network, taking r = 1.5 as the smallest value, the results yielded the relatively low number of 235 MCL clusters with a high standard deviation (SD) of 3.53, although there was a series of 361 smaller clusters (SD=1.81) when r = 3. Figure 6 plots modularity and the F measure as a function of r, and indicates there are no peaks in the Q value. This finding means that the smaller the value of r is, the greater the value of Q is.

4.3. Discussion

As Figure 6 shows, Modularity Q reaches a maximum value of 0.96 under the conditions of r = 1.5 and a threshold of c = 0.1. However, the number of words included within the clusters was only 455, which represents just 0.5% of the total. The Q values for the C0.1 sub-network are much higher than for the C0.03 sub-network, while, conversely, the F measurers for the C0.03 sub-network are greater than for the C0.1 sub-network. Combining these results for Q and F, an r value of 1.5 in the case of the C0.03 sub-network can be regarded as being an appropriate value for this parameter. These results demonstrate the effectiveness of combining the modularity measurement and the F measure to control cluster sizes. The combination of the RMCL graph clustering method and the modularity measurement provides even greater control over cluster sizes. According to Newman and Girvan (2004), modularity is a particularly useful index for assessing the quality of divisions within a network. Accordingly, Miyake and Joyce (2007) have proposed the combination of the RMCL clustering algorithm and this modularity index in order to optimize the inflation parameter within the clustering stages of the RMCL process.

5. Conclusions

This paper has reported on the statistical features of a semantic network representation of Japanese word associations and the application of graph clustering to that graph. After outlining the continuing construction of the large-scale JWAD, the paper analyzed the characteristics of the initial JWAD network representation. Calculated degree distributions for the network indicate that it has a scale-free organization and a pattern of sparse connectivity. Employing sub-networks extracted from the JWAD network according to different clustering coefficients, this paper has also proposed the application of the modularity measurement to the optimization of the r parameter in the MCL, which influences clustering granularity and so provides greater control over cluster sizes. As the JWAD expands, we plan to continually apply and develop these graph theory approaches in mapping out the growth of the JWAD semantic network. One particularly promising technique in that respect is in utilizing the clustering methodology for the automatic construction of condensed network representations as a means of highlighting the structures within hierarchically-organized semantic spaces and of visualizing large-scale linguistic knowledge resources.

Acknowlegement

This research has been supported by the COE21-LKR. The authors would like to express their thanks to Prof. Furui, Prof. Akama, and Ms. Jung. The authors have been supported by Grants-in-Aid for Scientific Research from the Japanese Society for the Promotion of Science: Research project number 19700238 for the first author and number 18500200 for the second author.

References

Barabasi, A. L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science, 286*, 509–512.

Dorow, B., Widdows, D., Ling, K., Eckmann, J., Sergi, D., & Moses, E. (2005). Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination. *MEANING-2005.*

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Joyce, T. (2005). Constructing a Large-scale Database of Japanese Word Associations. In K. Tamaoka (Ed.), *Corpus Studies on Japanese Kanji, (Glottometrics 10).* Tokyo and Lüdenschied: RAM-Verlag, pp. 82–98.

Joyce, T. (2006). Mapping Word Knowledge in Japanese: Constructing and Utilizing a Large-scale Database of Japanese Word Associations. *LKR2006*, 155–158.

Joyce, T. (2007). Mapping Word Knowledge in Japanese: Coding Japanese Word Associations. *LKR2007*, 233–238.

Jung, J., Miyake, M., & Akama, H. (2006). Recurrent Markov Cluster (RMCL) Algorithm for the Refinement of the Semantic Network. *LREC2006*, 1428–1432.

Miyake, M., & Joyce, T. (2007). Mapping out a Semantic Network of Japanese Word Associations through a Combination of Recurrent Markov Clustering and Modularity. *LTC2007*, 114–118.

Moss, H., & Older, L. (1996). *Birkbeck Word Association Norms.* Hove: Psychological Press.

Nelson, D. L., McEvoy, C., & Schreiber, T. A. (1998). The University of South Florida Word Association, Rhyme, and Word Fragment Norms. *http://www.usf.edu/FreeAssociation*.

Newman, M. E., & Girvan, M. (2004). Finding and Evaluating Community Structure in Networks. *Phys. Rev., E69*, 026113.

Okamoto, J., & Ishizaki, S. (2001). Associative Concept Dictionary and its Comparison with Electronic Concept Dictionaries. *PACLING2001*, 214–220.

Ravasz, E., & Barabasi, A. L. (2003). Hierarchical Organization in Complex Networks. *Physical Rev. E, 67*, 026112.

Roget, P. M. (1991). Roget's Thesaurus of English Words and Phrases. *http://www.gutenberg.org/etext/10681.*

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2005). Word Association Spaces for Predicting Semantic Similarity Effects in Episodic Memory. In A. F. Healy (Ed.), *Experimental Cognitive Psychology and its Applications. (Decade of Behavior).* Washington, D.C.: APA.

Steyvers, M., & Tenenbaum, J. B. (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cog. Sci., 29*, 41–78.

Umemoto, T. (1969). *Table of Association Norms: Based on the Free Associations of 1,000 University Students. (in Japanese).* Tokyo: Tokyo Daigaku Shuppankai.

van Dongen, S. (2000). *Graph Clustering by Flow Simulation.* Ph.D. thesis: University of Utrecht.

Vechthomova, O., Gfeller, D., Chappelier, J.-C., & De Los Rios, P. (2005). Synonym Dictionary Improvement through Markov Clustering and Clustering Stability. *International Symposium on Applied Stochastic Models and Data Analysis*, 106–113.

Watts, D., & Strogatz, S. (1998). Collective Dynamics of 'Small-world' Networks. *Nature, 393*, 440–442.