Associative Language Learning Support Applying Graph Clustering
-- for Vocabulary Learning and Improving Associative Ability

Jaeyoung Jung*, Nobuyasu Makoshi**, Hiroyuki Akama*
Department of Human System Science*, Global Scientific Information Center**
Tokyo Institute of Technology, 2-12-1, O-okayama, Meguro-ku, Tokyo, Japan
{catherina,akama@dp.hum.titech.ac.jp}, makoshi@gsic.titech.ac.jp

Abstract

In this study, we propose a new web-based associative language learning system, called ALL. It assumes that word association exercise and associative information about words would be useful for vocabulary learning and improving associative ability. For this, word association sources are used, and a graph theory and a graph clustering are employed as suitable tools for organizing the data structure and for mining meaningful information from it.

1. Introduction

As one of well-known teaching and learning strategies, associative learning has been often employed in education. What associative learning fundamentally assumes is that new concepts or knowledge can be effectively learned and recalled in the associative relations to other concepts or pre-existing knowledge.

This study proposes to take an advantage of the notion of association for language learning. Specifically, it attempts to support vocabulary learning by encouraging users to produce word association and to be exposed to a plenty of words under various associative contexts. Such information on associative relations between words is obtained as a result of analyzing the semantic structures of word association sources by graph theory concepts and graph clustering methods. In the sense that it makes the best use of associative ability in word practicing and learning, this study is assumed to present a different approach to computer-assisted vocabulary learning, unlike existent programs such as CAVOCA [1] and WUFUN [2].

Therefore, a principal focus of this study is placed on discussing about a method for constructing the database, as a result of organizing the semantic data structure and mining meaningful information by graph clustering tools (through Section 2 to 3). Then, in Section 4, the system implemented based on this database is introduced. We conclude in Section 5.

2. Semantic networks

This section introduces about the word association sources used in this study, and presents a way of constructing semantic networks for the database.

2.1. Word association sources

For constructing database, we use Edinburgh Word Association Thesaurus (hereafter, EAT) [3] and Nelson’s word association data [4], which are two representative large-scale word association data. Both data were collected through a free association task, where participants are asked to give the first word to be freely associated with presented words. The original data include the word pairs of such cue-associative response, with some additional information, such as a response time, a response order, a frequency of response, a semantic relational type of the pair.

Word association sources appear to be very useful for language learning and improving associative ability, because the various semantic, associative aspects of the words are reflected within the data together with some cultural and linguistic features. Not as learning objects, but as a catalyst, association data seem to encourage the learning motivation and to enhance the creativity.

2.2. Basic concepts from graph theory

In this study, the data are analyzed in a network form, by applying graph theory. A graph consists of nodes and edges, and degree and curvature are two basic concepts used in analyzing the structural properties of the graph. Degree of a node denotes the number of edge that a node has. Curvature of a node...
refers to a cohesiveness of a node and its neighbors, by calculating the extent of their connectedness.

The networks of linguistic data are called semantic networks, where nodes represent words and the pairing or co-occurrence relationship between words are represented by edges. In the semantic networks, the degree of a word node is a useful index to evaluate words in terms of conceptual generality and wide-ranged usability. For example, food, good, house, money, water, and work are the top six words with the highest degree values, both in the EAT and Nelson’s data. On the other hand, the curvature of a word presents a different viewpoint about a word, in terms of conceptual consistency. Usually, degree and curvature shows a reciprocal relationship, so that the words with the high degree values have a tendency of the low curvature values.

From this, Jung and Akama [5] proposed a coefficient for evaluating words in terms of usability by degree and consistency by curvature. This new index is calculated as degree x curvature (D-C), distinguishing words by characteristics. For instance, for the Nelson’s data, rather concrete daily life notions (such as intercity, hymn, flu, aspirin, and sofa) and the words representing religion, sports, music, and liquor usually have the high D-C values, while rather abstract concept words such as rebel, overwhelm, stubborn, and depletion have the low D-C values.

2.3. Compiling of datasets

Two different scaled semantic networks are constructed here, based on two datasets compiled from the EAT and the Nelson’s data. A simple combination of the two entire data generates a large-scale dataset, which seems to sufficiently cover abundant lexicon reflecting various semantic aspects that words can have. On the other hand, the other dataset is made by selectively extracting from each data the top 10% of words with the highest D-C values. Here, the lexical level appears to be controlled in terms of wide-ranged usability (that is, by high degree) and high coherence (that is, by high curvature).

The large semantic network consists of 19,933 word nodes and 137,015 edges and the small one has 2,261 word nodes and 14,902 edges. These two semantic networks are additionally restructured by graph clustering methods, because the reorganizing of the semantic network structure allows us to discover new and latent information from the networks. The details are explained in the following section.

3. Re-organizing of semantic networks

With its consistent, scalable performance, Markov Clustering (hereafter, MCL) [6] is often used for graph clustering in many areas, such as genomics, bioinformatics, linguistics, and so forth. However, some complementary processes for MCL seem to be additionally required for an optimal clustering according to data [7][8]. It also holds for our two semantic networks. The process of obtaining the adequate clustering for our data is explored here.

3.1. Clustering problem

In spite of its generally satisfying performance, several studies report that MCL often generates extremely large or small sized clusters [7][8]. In fact, when MCL was applied to our large-scale semantic network with 19,933 nodes, a plenty of small-sized clusters were generated. Particularly the clusters in size of 1 to 3 correspond even to 67% of the total. On a manual inspection, we identified that many of those small sized clusters need to be integrated further.

Such issue is referred to as a clustering problem. The size of clusters may indicate the properness of clustering results. From the sense that MCL yields a paradigmatic grouping of words, it may be difficult to regard the one-node-clusters or the clusters over a few hundred nodes as the appropriate conceptual groupings.

In order to obtain the appropriate and meaningful information by restructuring semantic networks, this clustering problem has to be solved first. The solution to this is discussed in the next.

3.2. A solution by M&A

For adjustment of the small-sized clusters, we propose a method of mergers and acquisition (M&A). Its essential idea is to find the proper clusters that absorb the extremely small ones.

The specific merging process for a cluster \(p\) which includes only one word \(q\) (cluster size is 1) goes as follows: First, M&A algorithm traces all the neighbouring words connected to word \(q\) on the semantic network (before clustered) and their clusters after MCL application. These clusters become the potential take-overs of word \(q\) in merging. Secondly, M&A algorithm investigates those candidate clusters one by one, (1) if the candidate cluster’s size is of larger than 4 and smaller than \(y\) (to prevent from generating extremely large sized clusters, a maximum size of clusters is setup), and (2) if the candidate cluster has at least two neighbors of word \(q\). If the two conditions are met, M&A algorithm permits word \(q\) to
be integrated into the candidate cluster. If not, M&A process is continued till word \( q \) finds a right cluster.

On the other hand, can the reasonable quality of MCL clustering still be preserved, even after this M&A process? This question is considered in the next.

![Figure 1. Architecture of the system](image)

### 3.3. Quality of clusters

The evaluation of clustering quality has been a significant issue. Its in-depth discussion may be beyond the range of our present study. In order to validate our new method, we just focus on a change of a relative quality of clusters before and after M&A process. For this, a new coefficient, \( \text{QualClst} \), is devised, measuring the quality of the clusters in terms of consistency and coherence. It is obtained as follows:

\[
\text{QualClst} = \text{Curvature}_{\text{principal}} + \text{LinkStrength}_{\text{inside}} + \% \text{ of words with } (\text{LinkStrength}_{\text{inside}} > \text{LinkStrength}_{\text{outside}})
\]

It is assumed here that the curvature value of the highest degree word within a cluster (\( \text{Curvature}_{\text{principal}} \)) implies the consistency of the cluster representing a paradigmatic group. The association strength imposed on the links between cue-response pairs is employed from the original word association data, for measuring the coherent solidarity within a cluster. Specifically, \( \text{LinkStrength}_{\text{inside}} \) refers to a mean value of association strength of all links within a cluster, indicating the inside cohesion. The proportion \( \% \) of words with \( \text{LinkStrength}_{\text{inside}} > \text{LinkStrength}_{\text{outside}} \) also implies the inside cohesion, but from the different respect. It measures how strongly the words within a cluster are attached to a cluster, by comparing the average association strength within a cluster \( \text{LinkStrength}_{\text{inside}} \) with the average association strength between the words and their all neighbors that are clustered outside \( \text{LinkStrength}_{\text{outside}} \).

### 3.4. Adjustment results

As mentioned above, MCL generates consistent conceptual groupings in overall, but some clusters consisting of only one or two members seem to be an obvious failure in conceptual clustering. Based on the manual inspection, we determined that the cluster size of 1 to 3 from the MCL clusters of our semantic networks need an additional treatment of M&A.

We applied the M&A process to the Nelson’s MCL clusters, to identify its performance. Through M&A, the total of 3,051 clusters is reduced to 1,100, and an average cluster size increases from 3.47 (a standard deviation, hereafter STDV=3.94) to 9.63 (STDV=8.47). In terms of the \( \text{QualClst} \) showing the change of the quality of the clusters, the old 1,001 clusters (of larger than 4 in size) before taking over any small-sized clusters have an average of 0.766 (STDV=0.429), and around 88 % of the clusters show larger \( \text{QualClst} \) values than mean-1STDV (=0.34). The new 1,100 clusters after the M&A process have an average \( \text{QualClst} \) of 1.025 (STDV=0.619) and more than 90% of the new clusters have larger \( \text{QualClst} \) values than mean-1STDV. From this result, the quality of clusters in terms of consistency and coherence seems to be preserved, or rather improved after our proposed M&A method.

Our two semantic networks constructed in Section 2 were clustered by MCL and M&A process, and then were newly structured by a graph composer, RMCL [9]. Section 4 presents a practical application of these semantic networks.

### 4. Associative language learning system

This section introduces a web-based system, implemented for vocabulary learning support based on association. Its overall design and particular tasks are explained here.

#### 4.1. Associative vocabulary learning

The new words are learnt in the several ways, simply from the definition on dictionary or through the reading contexts. For building vocabulary, various types of information about words can be also used: for example, affixes and roots, synonyms, antonyms, hyponyms, meronyms, collocation, and the like. The associative information on words can be any of these relationships.

In the system, the words are presented under the various associative contexts, and users actively participate in word association exercises. This practical environment is assumed to be effective for building
abundant vocabulary, by encouraging users to produce their own association and to be exposed to various words in a natural way. Here, associative ability and creative thinking are likely to be enhanced as well. Therefore, the system is believed to be very useful for both language users as foreign learners and also as native speakers. In this sense, our system is called Associative Language Learning, ALL.

4.2. The system, ALL

4.2.1. The database. Figure 1 presents a brief architecture of the system. For construction of the database, word association data was represented into semantic networks and additionally restructured. Through Section 2 and 3, four different semantic networks were eventually built on the database, in the different aspects of the scale (small and large datasets) and the structural type (word-basis and cluster-basis graphs). The different scale of datasets, which was for controlling the lexical level, were obtained by employing the coefficient, $D-C$ (degree x curvature). For the small datasets, the words with a rather higher usability and a higher consistency in concepts were extracted from the large-scale dataset. The different structural types of the semantic networks were built by applying graph clustering. The word-basis and the cluster-basis structures usually yield different types of word associative information: Somewhat straightforward association can be extracted from the word-basis semantic network, while the cluster-basis network provides rather extended and sometimes creative association [9][10].

4.2.2. The GUI. One of principal features of the system is its interaction with users. Instead of a simply lopsided presentation of information, the system attempts to exchange word association with users, or to encourage users to understand and produce their own association. The main interface of ALL is shown in Figure 2. It consists of three types of exercise: (1) word relay, (2) word guessing and (3) word building.

4.2.3. The Tasks. Particularly, the word relay is presented in Figure 3. The captured screen shows its actual exercise that starts from the word, brainstorm randomly selected by the system. A user types in his/her own associated word with it, idea here. Then the system takes a turn and gives belief as a response. Again, a user associates friend with it, and so on.

Figure 4 presents some interesting analyses of user’s association. The system calculates the associative distance between words of user’s association, by applying graph theory concepts. Here, the associative distance between two words denotes the number of edges to connect them. In the case that the distance is longer than two, the system provides probable link paths as in Figure 5.

In word guessing, the system induces users to guess a word in several associative contexts. In word building, it provides associations in different distances, and when users meet unknown words, the system helps them to understand unfamiliar words by additional associations or dictionary meaning.

These three tasks of ALL can be simply modeled as in Figure 1 (the lower box). One to one associative relationship in word relay, many to one in word guessing, and one to many in word building are represented.
4.2.4. The feedback. Five native speakers of English used this system, ALL, and gave the feedback on it. In the aspects of a general impression about the system and the associative information, they gave an overall positive evaluation on the system, and usually agreed with the association information provided by the system. However, an experiment for an in-depth evaluation will have to be designed and conducted to identify an effect of word learning and associative, creative thinking.

5. Conclusion

In this study, associative language learning system (ALL) was introduced, with an attempt to support vocabulary learning and creative thinking. Its datasets were compiled from word association sources based on graph theory concepts, degree and curvature. Then, the M&A process, a refinement method for reorganizing data structures, was proposed in detail. An in-depth educational and cognitive influence should be evaluated about the proposed system. Furthermore, word association data obtained from users would be probably used to evolve the semantic networks of the system.

6. References


