Development of Meaning Structure by Usage-based Word Relationships

Takashi Hashimoto
*
Lab. for Information Synthesis
Brain Science Institute, RIKEN
Hirosawa 2-1, Wako, Saitama, 351-01, JAPAN

Abstract

Development of meaning structure is studied from a usage-based viewpoint by a constructive approach. The meaning structure is represented by relationships between words. A word's relationship to other words, which represents meanings of the word, is derived by analyzing similarity of the word's usage in sentences. Words make clusters according to their similarity. The word clusters are classified into several types and show some remarkable dynamics. The clusters and their dynamics are studied by means of hierarchical cluster analysis and principal coordinate analysis.

Keywords: Usage-based viewpoint; Constructive approach; Evolutionary linguistics; Word similarity; Word clusters

1 introduction

We propose *usage-based viewpoint* as that meanings of words should be discussed in terms of how language is used [1]. Since interrelationships among words are employed as a representation of meanings of words, the point of view means that a word's relation to other words should be derived by analyzing the word's usage in the language.

Constructive approaches are highly advantageous for understanding dynamically complex systems [2]. These approaches are also useful for studying *evolutionary linguistics* which is a new candidate for potentially clarifying the origins and evolution of language [3]. It is important to note that language are typically expressed as such dynamically complex systems as emergence, self-organization, collective behavior, clustering, diversification, hierarchy formation, and so on. It is also important to point out that language systems must have both adaptability and stability. If a language is too rigid, its users will not be able to formulate new expressions to describe diverse experiences, and if it is too unstable, no communication will be possible at all. Such dynamical stability and adaptability is often seen also in complex systems.

We have proposed some models for evolutionary linguistics by constructive approach [4, 5, 6]. In [6], we have tried to catch dynamical aspects of meaning structurization and discuss its relevance with linguistic categorization. In this paper, the model and results of [6] are summarized and hierarchical cluster analysis and principal coordinate analysis for the results are reported.

2 Word Similarity

In our model, an agents has generative grammar system and word similarity matrix. The agents try to recognize given sentences by means of their own grammar and articulate words from successfully recognized sentences. Word similarity matrix is updated by each recognition of a sentence. The grammar of agents is small at the beginning of simulation and modified according to its usage in the recognition processes. For the detail of the model, refer to [6].

Word similarity matrix is the representation of relationships among words which is calculated by Karov and Edelman's algorithm of words similarity formulae [7] with some revisions. A key concept in this definition is the mutual dependency between words and sentences. That is, similar words are used in similar sentences and similar sentences are composed of similar words.

The similarities between words and between sentences are respectively defined by the following formulae:

$$sim_{n+1}(w_i, w_j) =$$

*E-mail address: takashi@brain.riken.go.jp

$$\begin{cases} \sum_{s \ni w_i} weight(s, w_i) aff_n(s, w_j) & \text{if } i \neq j \ , \\ 1.0 & \text{if } i = j \ , \end{cases} \quad (1)$$

$$sim_{n+1}(s_i, s_j) = \begin{cases} \sum_{w \in s_i} weight(w, s_i) aff_n(w, s_j) & \text{if } i \neq j \ , \\ 1.0 & \text{if } i = j \ , \end{cases} \quad (2)$$

$$aff_n(s, w) = \sum_{s' \ni w} weight(s', w) sim_n(s, s') \ , \quad (3)$$

$$aff_n(w, s) = \sum_{w' \in s} weight(w', s) sim_n(w, w') \ , \quad (4)$$

$$weight(s, w) = \frac{factor(s, w)}{\sum_{s' \ni w} factor(s', w)} \ , \quad (5)$$

$$factor(s, w) = \frac{p(s)}{\#(s, w)} \ , \quad (6)$$

$$weight(w, s) = \frac{factor(w, s)}{\sum_{w' \in s} factor(w', s)} \ , \quad (7)$$

and

$$factor(w, s) = \frac{1}{p(w) lg(s)} \ . \quad (8)$$

In the above formulae, a suffix n indicates the times of iterations, $w \in s$ means words included in a sentence s, and $s \ni w$ means sentences including a word w. The functions $weight(s, w)$ and $weight(w, s)$ are normalizing factors that define contribution from the appearance frequency and length of each word and sentence to affinity and similarity. In Eqs. (6) and (8), $p(w)$ and $p(s)$ are the appearance frequencies of a word w and a sentence s, respectively; $lg(s)$ is the length of a sentence s, which is defined by the number of words included in the sentence; and $\#(s, w)$ is the number of appearances of a sentence s including a word w.

At the initial iteration step ($n = 0$), similarity with word itself ($sim_0(w_i, w_i)$) is 1.0, the others are 0.0. Word-sentence affinity (Eq. (4)) at $n = 0$ is calculated from this initial word similarity matrix. Then, these four formulae are iteratively calculated as Eqs. (2) $\rightarrow$ (3) $\rightarrow$ (1) $\rightarrow$ (4) at each successful recognition of a sentence by the agents.

3 Word Cluster

Words are clustered in word similarity space according to having or not having similarity with each other. By using principal coordinate analysis, which is one method of the multi-dimensional scaling [8], the clusters of words are clearly seen in two dimensional space. Fig.1 is an example of scatter diagram constructed from the result of the analysis. The first two axes are used as x and y axes respectively to draw the scatter diagram. We can see three clusters of words in this figure. At top-right and top-left in the diagram there are several words relatively closer relationships, namely similar usage in sentences. Words in the third cluster at bottom-center are solitary words. By this analysis any solitary words are grouped into one cluster (completely same position), since they share the same relation, that is, they have no relationships with any other words.

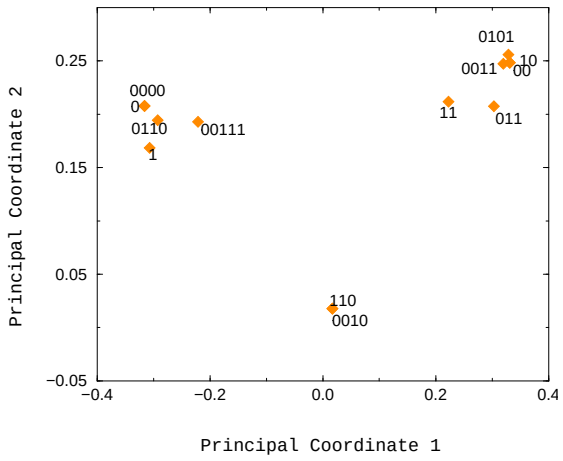


Figure 1: An example of principal coordinate analysis of word similarity matrix $sim(w_i, w_j)$.

Various shapes of cluster structure appear, which depends on the initial grammar of the agents. We have classified the structures in the space of similarity into six types according to their shapes as follows[6]: (a) *solitary word* – a word has no relation to other words; (b) *flat cluster* – words in a cluster have almost identical similarities with each other; (c) *gradual cluster* – similarity between words in a cluster depends on words and gradually changes; (d) *two-peak cluster* – words are in a cluster but there are two peaks of similarity; (e) *sub-clustering* – a cluster has a stepwise structure. (f) Words form plural and unrelated clusters.

We use hierarchical cluster analysis [8] to study the nature of the clusters. The results are shown in Fig.2. Sectioning a dendrogram parallel to the vertical axis at any similarity level yields a partition of words into clusters. Hereafter, sectioning level is denoted by $\widehat{sim}$. For example, sectioning the dendrogram in Fig.2(d) at

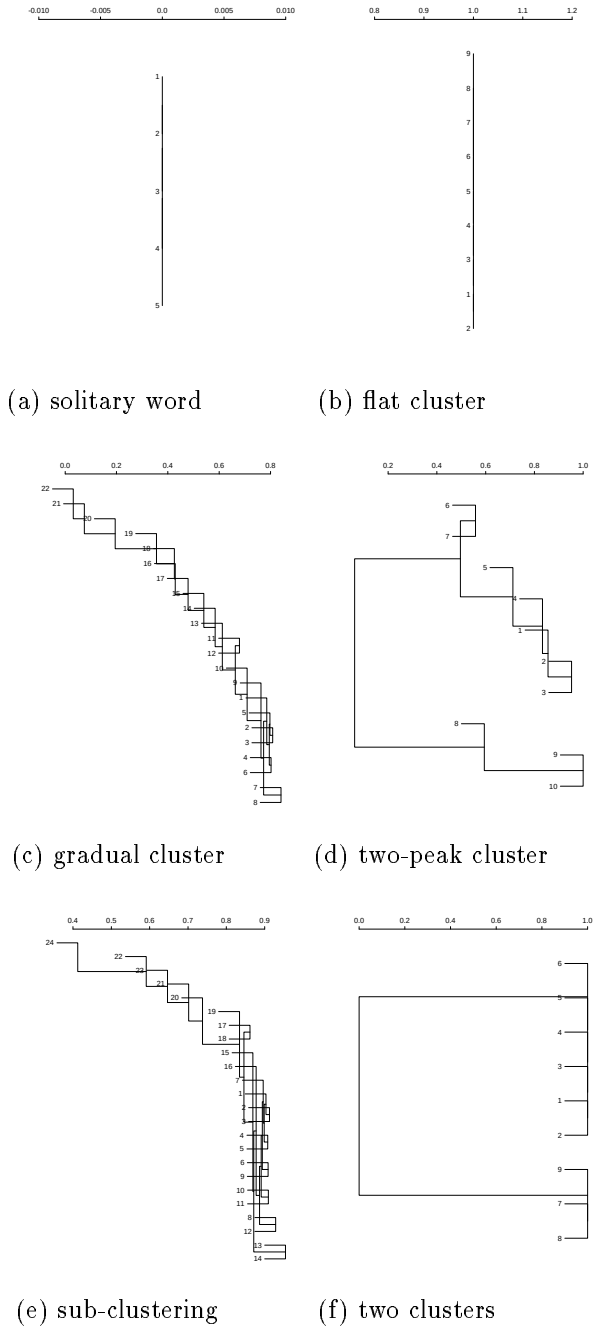
$\widehat{sim} = 0.65$ yields the clusters (1-5), 6, 7, 8, (9,10). If we decrease the sectioning level, some clusters are combined into one cluster, and vice versa. For instance, by sectioning the dendrogram in the above example at $\widehat{sim} = 0.58$ clusters 8 and (9,10) are combined into one cluster, and sectioning at $\widehat{sim} = 0.75$ tell word 5 as the different group apart from the cluster (1-4).

The findings from this analysis are summarized as follows: (a) solitary word. Since each word has no similarity with each others, all words are aligned at the line where similarity equals to zero; (b) flat cluster. All words have almost identical similarity with the all other words in the cluster. So, they are aligned at the line where similarity equals to 1.0; (c) gradual cluster. Words are connected to a cluster one after another. In the case of this type, there is no appropriate sectioning point because of the successive connection and the gradual change of similarity; (d) two-peak cluster. Sectioning the dendrogram between $0.05 \leq \widehat{sim} < 0.5$ yields two clusters. Words in each cluster are excluded successively with increasing the sectioning point from $\widehat{sim} = 0.5$. From this analysis, we should say that the two-peak cluster is combined version of two gradual clusters. The two gradual clusters are seen as one cluster when the sectioning point comes below $\widehat{sim} = 0.05$; (e) sub-clustering. The dendrogram shows two parts, dense ($0.8 < sim$) and gradual changing part ($sim \leq 0.76$). The dense part makes almost flat cluster structure and the other part makes the other group. (f) two clusters. We can figure out two independent clusters by sectioning any level between $0.0 < \widehat{sim} < 1.0$;

4 Development of Clusters

A general scenario of the development of structure in the similarity space is the following[6]. At first an agent can recognize only a one-word sentence. It develops the ability to recognize several sentences, but these all are one-word sentences, and therefore there are several solitary words in the similarity space. Then it becomes able to articulate plural words from sentences, which forms relations between words. Eventually words form gradual clusters. In the course of development, some remarkable dynamics of clusters such as boundary expansion, clusters merging, and structural change from gradual to flat cluster are observed. Parallel to this development, syntactic structure also develops from sequential to branch and to loop structures.

In Fig.3, we represent the process of cluster merging by means of principal coordinate analysis [8]. At $n = 10$ there are three clusters in the scatter diagram. Words



(a) solitary word (b) flat cluster

(c) gradual cluster (d) two-peak cluster

(e) sub-clustering (f) two clusters

Figure 2: Examples of structures of word similarity matrices analyzed by hierarchical cluster analysis. The horizontal axis shows similarity between words. Each leaf of branch means each word. Numbers put beside the leaves are indices for words. We have six distinct types of structures. (a) Solitary word. (b) Flat cluster. (c) Gradual cluster. (d) Two-peak cluster. (e) Sub-clustering. (f) Two clusters.

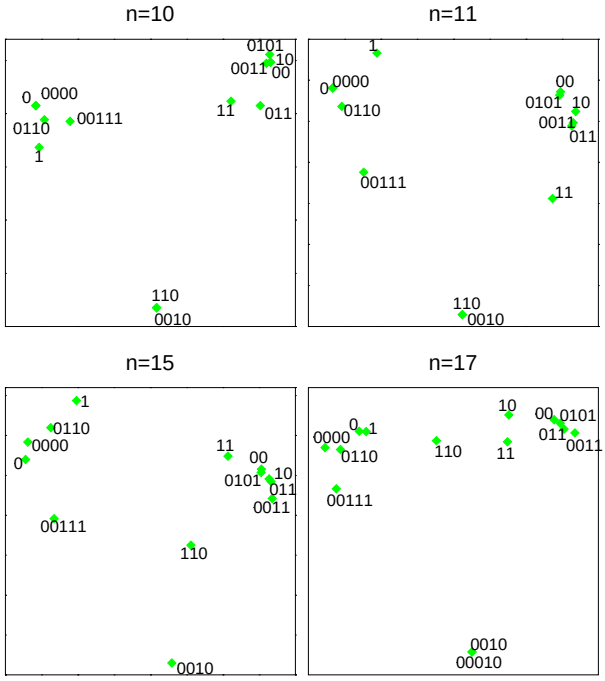


Figure 3: Dynamics of word clusters on merging. These are scatter diagrams from principal coordinate analysis of word-similarity matrix $sim_n(w_i, w_j)$. n is the number of iterations. Since we are interested only in the relative distances among words, value of x and y axis is omitted.

in the cluster at bottom-center are solitary words. Two clusters at top-left and top-right are gradual cluster. But from this diagram gradualness of the clusters is not clearly seen. Words in a gradual cluster typically form like a line. At $n = 11$, the two clusters strengthen the gradualness so words in the clusters become to be lined up. At $n = 15$, one of the solitary word, 110, is related with the top-right cluster, then the boundary of the cluster expands. At this time, the two clusters at top of the diagram have small relationships, but from this analysis such feature can not be detected. In the next figure, $n = 17$, these two clusters become one two-peak cluster. We find that word 110 play like a cramp between two cores.

5 Summary

We have proposed an evaluation of meaning representation by relationship among words based on the similarity of language usages from the viewpoint of usage-based meaning. We studied structurization of word cluster in word similarity space by means of an artificial agent with a grammar system. These clusters, classified into six types, are analyzed by hierarchical method of cluster analysis. Each type shows characteristic feature also by this analysis. One of the remarkable dynamics of the structures, cluster merging, is studied by principal coordinate analysis. Development of two clusters from small to gradual ones and mergence to one two-peak cluster can be clearly seen.

Acknowledgments

I am grateful to Mr. Iba for his help on statistical analysis. This work is supported by Special Postdoctoral Researchers Program at Institute of Physical and Chemical Research (RIKEN).

References

[1] Wittgenstein, L., *Philosophische Untersuchungen*, Basil Blackwell, 1953.

[2] Kaneko, K. and Tsuda, I., Constructive complexity and artificial reality: an introduction, *Physica*, **D75**, 1–10, 1994.

[3] Steels, L., Synthesising the origins of language and meaning using co-evolution, self-organisation and level formation, in *Evolution of Human Language*, Hurford, J (ed.), Edinburgh Univ. Press, Edinburgh, 1997.

[4] Hashimoto, T. and Ikegami, T., Evolution of Symbolic Grammar Systems, *Advances in Artificial Life*, F. Morán et al. (eds.), Springer, Belrin, 812–823, 1995.

[5] Hashimoto, T. and Ikegami, T., Emergence of net-grammar in communicating agents, *BioSystems*, **38**,1–14, 1996.

[6] Hashimoto, T., Usage-based Structuralization of Relationships between Words, *Fourth European Conference on Artificial Life* Phil Husbands and Inman Harvey (Eds.), MIT Press, Cambridge, MA pp. 483–492, 1997.

[7] Karov, Y. and Edelman, S., Similarity-based word sense disambiguation, Technical Report of Weizmann Institute, CS-TR 96-06, 1996.

[8] Gordon, A. D., *Classification*, London, Chapman and Hall, 1981.