The "Close-Distant" Relation of Adjectival Concepts Based on Self-Organizing Map

Kyoko Kanzaki, Hitoshi Isahara
National Institute of Information and
Communications Technology
3-5, Hikaridai, Seikacho,
Sorakugun, Kyoto, 619-0289,
Japan
{kanzaki,isahara}@nict.go.jp

Noriko Tomuro
School of Computer Science, Telecommunications and Information Systems
DePaul University
Chicago, IL 60604
U.S.A
tomuro@cs.depaul.edu

Abstract

In this paper we aim to detect some aspects of adjectival meanings. Concepts of adjectives are distributed by SOM (Self-Organizing map) whose feature vectors are calculated by MI (Mutual Information). For the SOM obtained, we make tight clusters from map nodes, calculated by cosine. In addition, the number of tight clusters obtained by cosine was increased using map nodes and Japanese thesaurus. As a result, the number of extended clusters of concepts was 149 clusters. From the map, we found 8 adjectival clusters in super-ordinate level and some tendencies of similar and dissimilar clusters.

1 Introduction

This paper aims to find a diversity range of adjectival meanings from a coordinate map in which "close-distant" relationships between adjectival classes is reflected. In related research over adjectives, Alonge et.al (2000), Solar (2003), Marrafa and Mendes (2006) suggested that WordNet and EuroWordNet lack sufficient adjectival classes and semantic relations, and extended the resources over such relations.

For the sake of identifying the diversity of adjectival meanings, it is necessary to analyze adjectival semantics via "close-distant" relationships extracted from texts. In our work on extracting adjective semantics, we consider abstract nouns as semantic proxies of adjectives. For the clustering method, we utilized a self-organizing map (SOM) based on a neural network model (Kohonen, 1997). One of the features of SOM is that it assigns words coordinates, allowing for the possibility of visualizing word similarity. SOM has two advantages for our task. One is that we can utilize the map nodes of words to locate members of clusters that clustering methods have failed to classify. The other is that the map shows the relative relations of whole clusters of adjectival concepts. By observing such a map in which the relations of clusters are reflected, we can analyze the diversity of adjectival meaning.

2 Abstract Nouns that Categorize Adjectives

Collocations between adjectives and nouns in “concrete value and its concept” relations can be used to represent adjectival semantics. Nemoto (1969) indicated that expressions such as “*iro ga akai* (the color is red)” and “*hayasa ga hayai* (literally, the speed is fast)” are a kind of tautology. Some studies have suggested that some abstract nouns collocating with adjectives are hypernymic concepts (or concepts) of those adjectives, and that some semantic relations between abstract nouns and adjectives represent a kind of repetition of meaning.

This paper defines such abstract nouns as the semantic categorization of an adjective (or an adjectival concept).

The data for this study was obtained by extracting adjectives co-occurring with abstract nouns in 100 novels, 100 essays, and 42 years of newspaper articles.

We extracted the abstract nouns according to the procedure described by Kanzaki et.al (2006). Here, they evaluated the category labels of adjectives obtained by the proposed procedure and found that for 63% of the adjectives, the ex-

© 2008. Licensed under the *Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported* license (http://creativecommons.org/licenses/by-nc-sa/3.0/). Some rights reserved.

tracted categories were found to be appropriate. We constructed a list as follows:

Abstract Nouns:
Adjectives modifying abstract nouns
KIMOCHI (feeling):
ureshii (glad), *kanashii* (sad),
shiawasena (happy) …

In this list, “*KIMOCHI* (feeling)” is defined by “*ureshii* (glad), *kanashii* (sad), and *shiawasena* (happy)”, for example. Here, each abstract noun conveys the concept or hypernym of the given adjectives.

Next we classify these abstract nouns based on their co-occurring adjectives using SOM.

3. A Map of Adjective Semantics

3.1 Input Data

In our SOM, we use adjectives which occur more than four times in our corpus. The number of such adjectives was 2374. Then we identified 361 abstract nouns that co-occurred with four or more of the adjectives. The maximum number of co-occurring adjectives for a given abstract noun in the corpus was 1,594.

In the data, each abstract noun was defined by a feature vector, in the form of noun co-occurrences represented by *pointwise mutual information* (Manning and Schutze, 1999). Mutual information (MI) is an information theoric measure and has been used in many NLP tasks, including clustering words (e.g. Lin and Pantel, 2002).

3.2 SOM

Kohonen’s self-organizing map (SOM) is an unsupervised learning method, where input instances are projected onto a grid/map of nodes arranged in an *n*-dimensional space. Input instances are usually high-dimensional data, while the map is usually two-dimensional (i.e., $n = 2$). Thus, SOM essentially reduces the dimensionality of the data, and can be used as an effective tool for data visualization – projecting complex, high-dimensional data onto a low-dimensional map. SOM can also be utilized for clustering. Each node in a map represents a cluster and is associated with a reference vector of *m*-dimensions, where *m* is the dimension of the input instances. During learning, input instances are mapped to a map node whose (current) reference vector is the closest to the instance vector (where SOM uses Euclidean distance as the measure of similarity by default), and the reference vectors are gradually smoothed so that the differences between the reference vector and the instance vectors mapped to the node are minimized. This way, instances mapped to the same node form a cluster, and the reference vector essentially corresponds to the centroid of the cluster.

SOM maps are self-organizing in the sense that input instances that are similar are gradually pulled closer during learning and assigned to nodes that are topographically close to one another on the map. The mapping from input instances to map nodes is one-to-one (i.e., one instance is assigned to exactly one node), but from map nodes to instances, the mapping is one-to-many (i.e., one map node is assigned to zero, one, or more instances).

The input data was the set of 361 abstract nouns defined by the 2,374 co-occurring adjectives, as described in the previous section. These abstract nouns were distributed visually on the 2-dimensional map based on co-occurring adjectives. This map is a “map of adjective semantics” because the abstract nouns are identified as proxies for adjective semantics.

As mentioned before, similar words are located in neighboring nodes on the 2-dimensional map. The next step is to identify similar clusters on the map.

4. Clusters of Adjective Semantics

4.1 Tight Clusters from the Map Nodes

In SOMs, each node represents a cluster, i.e. a set of nouns assigned to the same node. These nouns are very similar and can be considered to be synonyms. However, nouns that are similar might map to different nodes because the algorithm’s self-organization is sensitive to the parameter settings. To account for this, and also to obtain a more (coarse-grained) qualitative description of the map, tight clusters—clusters of map nodes whose reference vectors are significantly close—were extracted. All groupings of map nodes whose average cosine coefficient between the reference vectors in the group was greater than 0.96 were extracted (Salton and McGill, 1983).

4.2 Result

The total number of clusters was 213. Excluding singleton clusters, the number of clustes was 81. 229 concepts were classified into 81 clusters, with 132 concepts not classified into any cluster.

In order to evaluate the quality of the conceptual classification, we utilized the "*Bunruigoihyou*" Japanese thesaurus (National Institute of Japanese Language, 1964). In "*Bunruigoihyou*," each category is assigned a 5-digit category number, with close numbers indicating similar categories.

Among the 81 with two or more concepts, the number of clusters containing words with the same class was 36. That is, for 44% of the clusters, the constituent nouns had the same "*Bunruigoihyou*" class label. The ratio of concept agreement between "*Bunruigiohyou*" and our obtained clusters was found to be 20.87/81=0.25. We also compared tight clusters by performing hierarchical clustering with the *k*-means algorithm.

The results of the hierarchical clustering were as follows:

1) The rate of clusters agreeing with "*Bunruigoihyou*": 30/96 = 0.31

2) The average rate of agreement for each tight cluster: 21.07/96 = 0.21

In the case of *k*-means:

3)The rate of clusters agreeing with "*Bunruigoihyou*": 33/143 = 0.23

4) The average rate of agreement for each tight cluster: 28.37/143 = 0.198

From these results, we can observe that clusters obtained with cosine similarity agree more with the Japanese thesaurus than the other two methods. Therefore, in terms of quality, clusters obtained by cosine similarity seem to be superior to the others.

4.3 Using the Position of Map Nodes

However, even for the result obtained with cosine similarity, 132 concepts were not classified into any clusters. Additionally, the clusters appear to be overly fine grained: most tight clusters include 1, 2 or 3 concepts. In order to find similar concepts that cosine similarity failed to cluster together, we used the position information of the map nodes.

After we plotted clusters obtained by cosine similarity on the map, we checked for singleton concepts located near a cluster which are members of the same "*Bunruigoihyou*" class. Also, we checked to see if concepts in clusters located at neighboring nodes could be clustered together using the category numbers of "*Bunruigoihyou.* "

By extending the clusters, we generated a total of 149 clusters, including 68 with two or more elements and 81 singleton clusters.

5. Interpreting the Adjectival Clusters

In our final map, 361 concepts were distributed based on 2374 adjectives into 149 clusters. Among the 149 clusters, 68 contained two or more concepts.

5.1 "Close-Distant" Relations of Clusters and Adjectives

In the final map, clusters at the superordinate level are located around the center of the map. Upper level concepts tend to agree with clusters in "Bunruigoihyou." For examples, "image and impression," "situation and state", "feeling and mood" are located around the center of the map.

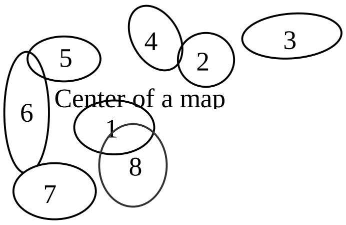

Fig7. Cluster 7 on the map

Cluster1 (Center of the map): *koto* (matter), *in'shou* (impression), *men* (side of something or someone), and *kankaku* (sense/feeling)

Cluster2: *seishitsu* (characteristics of someone/something), *yousou* (aspect)

Cluster3: *kanten* (viewpoint), tachiba (standpoint), *bun'ya* (domain)

Cluster4: *taido* (attitude), *yarikata* (way of doing)

Cluster5: *gaikan*, *gaiken*, *sugata* (outlook and appearance of someone/something)

Cluster6: *fun'iki*, *kuuki*, *kehai* (atmosphere)

Cluster7: *kimochi*, *kanji* (feeling)

Cluster8: *joutai* (state), *joukyou* (situation)

In our experiment, at the top level, adjectival concepts seem to be divided into 8 basic clusters. From the distribution of the map, we find "close-distant" relationships between clusters, that is clusters located far from each other tend to be semantically disparate. In terms of adjective semantics, the semantic relationship between "*kimochi, kanji* (feeling)" (Cluster7) and "*seishitsu* (characteristics of someone/something), *yousou* (aspect)" are distant.

However, "*kimochi, kanji* (feeling)" (Cluster7) has a close relation to "*fun'iki, kuuki, kehai* (atmosphere) " (Cluster6) and also "*joutai* (state), *joukyou* (situation)" (Cluster8).

1. In our experiment, 77 adjectives belonged to one or two clusters. Though there is the possibility of data sparseness, there is also the possibility that the meanings of these adjectives are specific. Examples of adjectives belonging to specific clusters are as follows:

Adjectives in distant relationships;
- Clusters 2: *keisandakai* (seeing everything in terms of money), *ken'meina* (wise), …
- Cluster 7: *akkenai* (disappointing/easily), *kiyasui* (feel at home),…

Adjectives in close relationships;
- Cluster 6: *ayashigena* (fishy)
- Cluster7: *akkenai* (disappointing /easily), *kiyasui* (feel at home)
- Cluster8: *meihakuna* (obvious), *omoshiroi* (interesting), *makkurana* (dark)

Japanese adjectives are often said to represent "kanjou (mental state)", "joutai (state)," "seisitsu (characteristics)" and "teido (degree)", in addition to "positive/negative image." In our experiment, the SOM unearthed not only these adjectival meanings, but also "inshou (impression)", "taido (attitude)", "kanten (viewpoint)" and "sugata (outlook)", which seem to be discriminative meanings of adjectives.

6. Future work

We classified 361 concepts based on 2374 adjectives using a self-organizing map. Since the SOM shows the distribution visually, it provides not only clusters of adjectives but also "close-distant" relationships between clusters. As a result, adjectival concepts at the superordinate level are divided into 8 main clusters. The results not only verify previous work but also suggest new discriminative adjective classes. One of the advantages of SOM is that it presents its outputs visually. As a result, we can explore "close- distant" relationships between clusters, and analyze the meaning of each. In addition to increasing the range of adjectival classes and improving our method, our method provides the means to analyze concepts which did not agree with those in existing thesauri such as "Bunruigoihyou", the EDR dictionary or Japanese Word Net.

References

Alonge, Antonietta., Francesca Bertagna, Nicoletta Calzolari, Andriana Roventini and Antonio Zampolli. 2000. Encoding Information on Adjectives in a Lexical-semantic Net for Computational Applications, *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics(NAACL-00)* :42-49

Kyoko Kanzaki,Qing Ma, Eiko Yamamoto and Hitoshi Isahara, 2006, Semantic Analysis of Abstract Nouns to Compile a Thesaurus of Adjectives, *In Proceedings of The International Conference on Language Resources and Evaluation (LREC-06)*

Kohonen, Teuvo. 1997. *Self-Organizing Maps, Second Edition*, Springer.

Lin, Dekang., and Patrick Pantel. 2002. Concept Discovery from Text, *Proceedings of the 19th International Conference on Computational Linguistics(COLING-02)*: 768-774

Manning, Christopher D., and Hinrich Shütze. 1999. *Foundations of Statistical Natural language Processing*, The MIT Press.

Marrafa, Palmira., and Sara Mendes. 2006. Modeling Adjectives in Computational Relational Lexica, *Proceedings of the COLING/ACL2006*:555-562

National Institute for Japanese Language. 1964. *Bunruigoihyou* (Word List by Semantic Principles).

Nemoto, Kesao. 1969. The combination of the noun with "ga-Case" and the adjective, *Language research 2 for the computer*, National Language Research Institute: 63-73 (in Japanese)

Salton, Gerard., and Michael J. McGill. 1983. Introduction to Modern Information Retrieval. McGraw Hill.

Solar, Clara. 2003. Extension of Spanish WordNet, Proceedings of the third International WordNet Conference(GWC-06):213-219