Extraction and Organization of Attribute Concepts of Adjectives from Corpora

Kyoko Kanzaki\textsuperscript{1}, Noriko Tomuro,\textsuperscript{2} and Hitoshi Isahara\textsuperscript{1}

\textsuperscript{1}. National Institute of Information and Communications Technology
\textsuperscript{2}. DePaul University
1. Introduction

• Task 1

To detect attribute concepts which combine adjectives and their head nouns from corpora

red, color, flower,
kind, attitude, person,
characteristics

Purpose 1: From our results, we want to:

1) verify attribute concepts of adjectives in existing thesauri

2) clarify which attribute concepts we can know automatically and what kind of information we have to think about intuitively
Task 2:

If we can obtain attribute concepts of adjectives from corpora, how are they related to each other?

We detect the structure of attribute concepts of adjectives like a thesaurus of adjectives.

The most abstract level

The specific level
Purpose 2: From the structure of attribute concepts that we obtain, we want to:

1) verify existing thesauri and linguistic analysis by comparing our results with those we obtained from corpora

2) apply the structure to define semantic relations between adjectives and their head nouns
About Task 1 (extracting attribute concepts of adjectives)

We utilized syntactic and semantic relations between adjectives and their head nouns in order to extract attribute concepts of adjectives from corpora.

About Task 2 (finding structure of attribute concepts of adjectives)

We calculated similarity and hypernym/hyponym relations (a taxonomic relation) among attribute concepts based on their instances, adjectives.

→ We used Kohonnen’s Self-Organizing Map in Neural Network model.
2. How to Extract Abstract Concepts of Adjectives from Corpora
In order to extract “Instance-Attribute” relations from corpora, we conducted the following process:

(1) Extract “A to iu B”
   (English: B that/of A)
   A is a concrete expression of B

   e.g.) the city of Tokyo (apposition)
       the story that I visited Thailand…

(2) Extract adjectives co-occurring with B without a “to iu” expression

   Adj 1 + B , Adj 2 + B , Adj 3 + B, …
(3) Manually extract “attribute concept and its instance” relations between adjectives and nouns from data gathered in step 2.

——— Linguistic clues———

1) Some “Adj+Noun” patterns cannot be paraphrased into a predicative relation in Japanese.

```
tanoshii  omoi  →  * omoi ga tanoshii
(happy) (feeling) →  feeling is happy.
```
(2) Some Adjectives in “Adj + Noun” pattern can be paraphrased into a predicative position without the Japanese determiner.

\[
takai \quad ondo \quad \Rightarrow \quad ondo \quad ga \quad takai
\]

(high) (temperature) (temperature) Subject-marker (high)

a high temperature \Rightarrow a temperature is high

direct relation

\[
takai \quad \text{(high)} \quad \text{temperature}
\]

c.f.
\[
ookii \quad hito \quad \Rightarrow \quad sono \quad hito \quad wa \quad ookii
\]

(large) (person) \Rightarrow the/that person topic-marker large.

a large person \Rightarrow the person is large

Indirect relation

\[
ookii \quad \text{(large)} \quad \text{size} \quad \text{man}
\]

age

nationality
We made a list of collocations from the corpora.

**KIMOCHI** (feeling): *ureshii* (glad), *kanashii* (sad), *shiawasena* (happy) ...

**OMOI** (thought): *ureshii* (glad), *tanoshii* (pleased), *hokorashii* (proud)...

**KANTEN** (viewpoint): *igakutekina* (medical), *rekishitekina* (historical) ...

We extracted these collocations from 100 novels, 100 essays, and 42 years’ worth of newspaper articles. We have now gathered about 365 such abstract nouns.
In the next step, we compare attribute concepts that we extracted with those in the EDR lexicon in order to investigate whether or not the attribute concepts (abstract nouns) that we extracted are feasible for adopting as attribute concepts of adjectives.
3. Verification: Comparison of concepts obtained from linguistic clues with concepts in the EDR lexicon

• 3.1. Procedures

We prepared sets of groups of adjectives extracted using linguistic clues.

Participants

Five participants were chosen from a group of linguists, persons engaged in Japanese education, and NLP researchers.

1) We showed participants a group of adjectives and asked them to determine its category by themselves.
Example)

Question:
Please define the category of the adjectives shown below.

Adjectives: 歯切れの悪い (inarticulate), 舌たらずな (babble), とげとげしい (acrid), 熱っぽい (vehement), 口幅ったい (shooting off one’s mouth), いやみな (bitter)

Your category: ***

2) For each group of adjectives, participants judged the suitability of three different categories corresponding to the group of adjectives: (1) our extracted category, (2) a category in the EDR lexicon, and (3) their own category created in procedure 1. They also considered whether or not the group of adjectives needs a category.
Example)

Question:
Please choose the most suitable category from the three.
Note that the third category is the category you defined.

Adjectives: hagireno warui (inarticulate), shitatarazuna (babble), togetogeshii (acrid), netsuppoi (vehement), kuchihabattai (shoot off one’s mouth), iyamina (bitterness)

(1) 言い方 (way of speaking)
(2) 態度や性格の値 (a value of one’s attitude or character)
(3) 言い方 (way of speaking)
3.2. Experimental Results

Fig. 1. Suitability of categories among the three methods.
Fig. 2 Breakdown of the suitability of categories among the three methods.
3.3 Considerations

1) We found that the correspondence between a group of adjectives and their category name was more suitable in our method than in the EDR lexicon made by humans.

We considered two reasons why our attribute concepts are better than those in the EDR lexicon:

1-1) The definitions of concepts in EDR are too detailed, therefore, it is sometimes rather difficult to understand them.

For the adjectives: *ken’akuna* (stormy), *shinmitsuna* (cordial/close), *shitashii* (close/familiar), *kiyasui* (friendly)

Our attribute concept: *naka* (relationship)

Concepts in EDR: *Hitodoushino shinritekina chikasa* (nearness of the mental distance between persons)
1-2) There are no categories in the EDR lexicon

For adjectives “komayakana (paying close attention to), saishin’na (solicitous), shinsetsuna (hospitality), yasashii (tender)”

   Our attribute concept: hairyo (care),

   Concept in EDR: NONE

According to examinees’ comments, in cases in which a category was not found in EDR, most examinees considered that a category was needed for the group of adjectives.

2) Other comments revealed that it was sometimes difficult for examinees to determine the most suitable category name because the category names were too abstract (both in our method and EDR).
In the next section, we aim to distribute these categories, that is, the attribute concepts that we extracted, on a map by using the self-organizing map (SOM) in a neural network model.
4. Self-Organizing Map (SOM)

4.1. Algorithm of SOM

Input data; if the noun co-occurs with the adjective, the value is 1

<table>
<thead>
<tr>
<th></th>
<th>happy</th>
<th>sad</th>
<th>academic</th>
<th>warm</th>
<th>....</th>
</tr>
</thead>
<tbody>
<tr>
<td>kimochi (feeling)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>omoi (thought)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>kanten (viewpoint)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Input data

kimochi (feeling) = \{1,1,0,1, 0,0,0…..\}
SOM Algorithm in Euclidean Space

\[ m_i = [a_1, a_2, a_3, \ldots, a_j] \]

SOM(Kohonen)
4.2 Introducing the the “Top” node for SOM

The “Top” node is fictitiously made as the topmost abstract concept.

All vectors of TOP node are assigned to value “1.”
(Top = \{1,1,1,1,1,……\})

and fixed in the center node of the map during learning of SOM
(utilizing the option available in SOM_PAK).

Therefore, attribute concepts would be laid out in a general
(most abstract) - to-specific order, moving out from the center
in all directions.
5. Tight Clusters – Similarity Relations

Extracted from SOM –

In the SOM map, if attribute concepts are assigned to nodes close to each other, they are very similar and can be considered to be synonyms to each other.

For indicating similarity between attribute concepts objectively we extracted **tight clusters**: (clusters of map nodes whose reference vectors are significantly close.)

We extracted all groupings of map nodes whose average cosine coefficient (Salton and McGill, 1983) between the reference vectors in the group was greater than 0.96.
Results

We obtained 88 clusters extracted from SOM nodes; the other 81 abstract nouns did not make clusters.

Examples of clusters:

- *jikan* (time), *jikoku* (time)
- *kishitsu* (temperament), *kishou* (nature),
- *kifuu* (trait), *hitogara* (character)
- *kouka* (effect), *eikyou* (influence)
- *keijou* (shape), *katachi* (form), *keitai* (form)

etc.
Evaluation

We compared our clusters of abstract nouns with categories of nouns in “Bunruigo ihyou,” one of the most popular Japanese thesauri made by humans.

We investigated whether or not abstract nouns classified into the same cluster are members of the same category in “Bunruigo ihyou.”

We found that in 52% of clusters generated from SOM nodes, sets of synonyms corresponding to the category numbers of “Bunruigo ihyou” were included, while in 47% of clusters, sets of synonyms did not correspond to the category numbers of “Bunruigo ihyou.”
Among clusters in which abstract nouns did not correspond to category numbers in “Bunruigoihyou,” some sets of abstract nouns seemed to be closely related to each other though they are not registered as synonyms in “Bunruigoihyou.”

For example, the sets “kokoro (heart)” and “akarusa (literally, brightness)” should belong to similar categories, but “akarusa,” which can mean “cheerfulness,” is not registered in “Bunruigo ihyou.”
6. Analysis: Relative relations of attribute concepts

- Temperature
- "Warm" clusters
- State/aspect
- Feeling/emotion
- Attitude
- Characteristics of something
- "Square" clusters
Consideration

1) The distribution of “warm” concepts on the map is an exact fit for the intuition of linguists and lexicographers. For example,

WordNet:  
“warm” \(\leftarrow\) emotionalism/emotionality
\(\leftarrow\) temperature

EDR lexicon:
“warm” \(\leftarrow\) state (or aspect)
\(\leftarrow\) characteristics of things

Our result based on co-occurring adjectives
“warm” \(\leftarrow\) state / aspect ________ (subordinate level)
\(\leftarrow\) feeling / emotion ________ (temperature)
\(\leftarrow\) characteristics of things
2) We can obtain various synonyms of “warm” related to its attribute concept.

*Atatakai kotoba*

(warm) (words)

In one of our tight clusters of attribute concepts, there are “attitude, way of speaking, way of doing …”. Semantic relations between “warm” and “words” may be related to those attributes. If so, the synonyms of “atatakai (warm)” in a viewpoint of “a way of speaking” are “kibishii (hard), tsumetai (unkind), gouman’na (imperious), son’daina (haughty), kan’daina (generous), …” and so on.
3) One of advantage of knowing a hypernym (or hyponym) is that it is possible for a computer to partially infer an indirect semantic relation between an adjective and a concrete noun.

*Watashi wa atatakai futon ni haitta.*

(warm) (blanket)

I went under a warm blanket

“Warm” is an instance of an attribute concept “temperature,” however, “blanket” does not have an attribute concept relating to temperature in itself, therefore the semantic relations between “warm” and “blanket” is indirect.
7. Conclusion

1) We discussed how to extract attribute concepts of adjectives, and manually evaluated the capability of extracting “attribute concept and its instance” relations.

2) Finally, we classified attribute concepts based on adjectives, and showed the resulting map generated by SOM.
From our result in experiment 1, we wanted to:
1) verify attribute concepts of adjectives in existing thesauri
2) clarify which attribute concepts we can know automatically and what kind of information we have to think about intuitively.

We have not yet sufficiently achieved these two goals, but we were able to provide some considerations for item 1.
For item 1), we found that the correspondence between a group of adjectives and their category name was more suitable in our method than in the EDR lexicon. One of the reasons we considered is that the explanations in EDR are too detailed, on the other hand, a definition by using abstract noun is simple and easy to understand.

For item 2), we have not conducted enough experiments to be able to provide conclusive results, but we’ll continue to conduct research and analysis.
From our results in experiment 2, we wanted to:

1) verify existing thesauri and linguistic analysis by comparing our results that we obtained from corpora

2) apply the structure to define semantic relations between adjectives and their head nouns

Again…

We have not reached any conclusions, but we did give you some possibilities.
For item 1) 

We compared our clusters of abstract nouns with categories of nouns in “Bunruigoihyou,” one of the most popular Japanese thesauri made by humans. In 52% of the clusters generated from SOM nodes, sets of synonyms corresponding to category numbers of “Bunruigoihyou” were included.

We instantiated attribute concepts of “warm” and showed that the distribution of “warm” concepts on the map were a fit for the intuition of linguists and lexicographers.
For item 2)

We showed the possibility of using concepts inherited from super-ordinate or subordinate concepts for the description of semantic relations between adjectives and their head nouns. We instantiated “warm blanket,” whose relation is indirect.

Cheerful room

- feeling
- atmosphere
- cheerful
- size
- atmosphere
- extent
- room
Thank you for your attention.

Kyoto in Japan on April