Extraction of Attribute Concepts from Japanese Adjectives

Kyoko Kanzaki[1], Francis Bond[1], Noriko Tomuro[2], Hitoshi Isahara[1]
[1]National Institute of Information and Communications Technology,
3-5 Hikari-dai, Seika-cho, Souraku-gun, Kyoto, Japan
[2] School of Computer Science, Telecommunications and Information Systems
DePaul University, Chicago, IL 60604, U.S.A.
E-mail: {kanzaki, bond, isahara}@nict.go.jp tomuro@cs.depaul.edu

Abstract

We describe various syntactic and semantic conditions for finding abstract nouns which refer to concepts of adjectives from a text, in an attempt to explore the creation of a thesaurus from text. Depending on usages, six kinds of syntactic patterns are shown. In the syntactic and semantic conditions an omission of an abstract noun is mainly used, but in addition, various linguistic clues are needed. We then compare our results with synsets of Japanese WordNet. From a viewpoint of Japanese WordNet, the degree of agreement of "Attribute" between our data and Japanese WordNet was 22%. On the other hand, the total number of differences of obtained abstract nouns was 267. From a viewpoint of our data, the degree of agreement of abstract nouns between our data and Japanese WordNet was 54%.

1. Introduction

A thesaurus is a classification of semantically categorized words organized by concepts. Several kinds of thesauri exist, such as Princeton's WordNet in English and the EDR concept dictionary in Japanese and English. However, Princeton's WordNet and WordNets in various languages are particularly weak regarding the relations among adjectival concepts. If a thesaurus could be built from corpora, then lexical meaning could be observed objectively, and lexical semantics in a thesaurus could be described more fully.

To construct a thesaurus from corpora, this paper addresses the following questions:

1) How to extract attribute concepts of adjectives like [feeling], [temperature], and [color] from corpora; and
2) How to automatically organize attribute concepts obtained in question 1) by establishing similarity relations and hierarchical relations among attribute concepts.

We focus mainly on question 1). The statistical extraction of semantic relations of words, such as hypernym-hyponym relations or part-whole relations, from corpora using syntactic patterns has been done in previous work (Hearst 1992, Caraballo 1999, Berland and Charniak 2000, Pantel and Ravichandran 2004, etc.). However, as for the detection of a hypernym concept (or a category) of an adjective, there has been relatively little research.

We describe various syntactic and semantic conditions for detecting abstract nouns referring to the attribute concept of adjectives from text, and then try to compare our results with Japanese Wordnet, though it is not completed yet.

2. Concepts in Linguistic Ontology

When organizing concepts in a thesaurus, all concepts are related to each other in a hierarchical relation. Below is an example of the hierarchical relation of superordinate concepts for the word "rabbit" from WordNet:

- [entity]
- [physical entity]
  - [object, physical object]
    - [whole, unit]
      - [living thing, animate thing]
        - [organism, being]
          - [chordate]
            - [vertebrate, craniate]
              - [mammal, mammalian]
                - [placental, placental mammal, …]
                - [lagomorph, gnawing mammal]
                - [leporid, leporid mammal]

Each word list in [ ] is a concept node that categorizes words, and in each concept node synonyms are classified as instances. For example, in the [leporid, leporid mammal] node, the synonyms "rabbits" and "hares" are classified as instances. In the [lagomorph, gnawing mammal] node, synonyms referring to [lagomorph, gnawing mammal, etc.], such as "mice," "rabbits" and "hares", are classified as instances. The following figure shows part of the structure of hierarchical concepts and their instances. { } is a group of instances.

→ [placental mammal]
{monkeys, lions, mice, rabbits, hares, …}
→ [lagomorph, gnawing mammal]
{mice, rabbits, hares, …}
→ [leporid, leporid mammal]
{rabbits, hares, …}

In our work, the abstract noun that we extracted is considered as a concept and an adjective is considered as an instance belonging to the concept.

3. Linguistic Analysis of Attribute Concepts and their Adjectives

3.1 Semantic functions of Japanese abstract nouns

The types of semantic functions of abstract nouns are shown below.

1) Abstract nouns encapsulating various concrete instances in nominal concept:
   (1) the problem of destruction of the ozone layer
   (2) {cheerful/dark/vivid, …} atmosphere
2) Abstract nouns which have the grammatical function of a "particle" and add subtle meanings to adjectives:
   (3) *hodo* (the ADJ_er, …, the ADJ_er, …), …
   Denryu ga ookii hodo, jikai ga kyoryoku ni naru
   (The higher the current, the stronger the magnetic field)
3) Particle representing a mood
   (4) *atari* (around, "…, say, …"), …
   Hagire no ii atari ga ninki no himitsu
   (liveliness) (, say,) subjectM (popularity) (secret)
   (The secret of his popularity may be that he is, say, lively.)
4) Pronouns with anaphoric function, such as "you," "he" and "it," etc.

The word "problem" in example 1 of item 1 is like "a shell noun" suggested by Schmid (2000), which allows speakers to encapsulate these complex pieces of information in a "temporary" nominal concept. However, we consider an example like "atmosphere" in example 2 of item 1, which encapsulates these complex pieces of information in a "lexical" nominal concept.
Some abstract nouns can represent a small range of connotation and a wider range of denotation. Since this type of abstract noun is an expression of formalization, the meaning of an abstract noun is more transparent. An abstract noun sometimes refers to an upper-level concept (Takahashi 1975, Nemoto 1969). Some examples are given below.
In the following example "*jotai* (condition)" can be omitted:

(5) *kokkai wa fuanteina joutai da.*
(national assembly) (unstable) (condition) (is an auxiliary indicating predication)
The national assembly is now in an unstable condition.

In this case, "*jotai* (condition)" can be omitted if the usage of the adjective "*fuanteina*" is changed from adnominal usage to predicative usage, that is, "*fuanteina*" (in adnominal usage) → "*fuanteida*" (in predicative usage).

(5') → (*kokkai wa*) *fuanteida*.
(be unstable)
(The national assembly is unstable.)

In this example, "*jotai* (condition)" is transparent and an upper-level concept of "*fuanteina* (unstable)" which implies the meaning of "*jotai* (condition)."
We analyze linguistic conditions and rules for the extraction of "instance – its concept" relations between adjectives and nouns by using the linguistic clue of "omission" according to the syntactic usage of an adjective. The syntactic usages that we investigated are shown in section 3.2.

3.2. Semantic functions of Japanese abstract nouns

3.2.1. Syntactic patterns composed from adjectives and nouns

An adjective has three kinds of usage in Japanese: a predicative usage, an adnominal usage, and an adverbial usage. In the following pattern, "N," "X," "Adj," and "V" refer to "noun 1," "noun 2," "adjective," and "verb" respectively. "X (noun 2)" is not only abstract nouns but also nouns appearing at X in the syntactic pattern. The Japanese syntactic patterns are literally translated into English patterns.

Predicative usage:

1-1) N *wa* Adj ← N *wa* X *ga* Adj
(N is Adj ← N (topic M) X is Adj)

(6) *Kono baggu wa takai*
← *Kono baggu wa nedan ga takai*
(This bag is expensive.
← This bag (topic marker) the price is expensive.

1-2) N *wa* Adj ← N *no* X *wa* Adj
(N is Adj ← X of N is Adj)

(7) *Yagi wa otonashii*
← *Yagi no seishitsu wa otonashii*
(The goat is gentle.
← The nature of the goat is gentle.)

1-3) N *wa* Adj ← N *wa* Adj X *da*
(N is Adj ← N is Adj X)

(8) *Kyodai wa ken'aku da*
← *Kyoudai wa ken'aku na naka da*
(The brothers are abrupt.
← The brothers have an abrupt relationship.)

Adnominal usage:

Adj + N ← Adj X *no* N
(Adj N ← N of Adj X)

(9) *akai tulip* ← *akai iro no tulip*
(a red tulip ← a tulip of red color)

Adverbial usage:

3-1) N *wa* Adj *ku/ni naru*
(Adj in adverbial usage)
←N *wa* Adj + X *ni naru*
(Adj in adnominal usage)
(N becomes Adj-ly ← N becomes Adj X)

(10) *Shorui wa boudaini naru*
(documents) (huge_adverbial) (become)
← *Shorui wa boudaina ryou ni naru*
(documents) (huge_adnominal) (quantity) (become)
(The document becomes huge.
← The document becomes a huge amount.)

3-2) N *wa* Adj *ku/ni* V ← Adj + X *de* V
N V Adj-ly ← N V in/from/with/… Adj X

(11) *rekishiteki ni kangaeru*
(historically) (consider)
← *rekishitekina kanten de kangaeruto*
(from historical viewpoint) (consider)
(Consider it historically. ← Consider from a historical viewpoint.)

3.2.2. Collected examples

First, we collected examples from 10 years of newspapers according to the syntactic pattern shown in section 3.2.1. Then we detected the linguistic clues to find abstract nouns which refer to upper-level concepts of adjectives. The number of examples we gathered based on each kind of syntactic pattern is shown below.

Predicative:

1-1) N wa X ga Adj (N (topic M) X is Adj):
24057 examples
1-2) N *no* X *ga* Adj(X of N is Adj):
61210 examples
1-3) N *wa* Adj X *da* (N is Adj X):
98 examples

Adnominal usage:

Adj X *no* N (N of Adj X):
48624 examples

Adverbial usage:

3-1) N *wa* Adj + X *ni naru* (N becomes Adj X):
700 examples
3-2) Adj + X *de* V (N V in/from/with Adj X):
52 examples

From the examples extracted from 10 years of newspapers, the number of examples in predicative and adnominal usages is larger than those in adverbial usage. Especially, in "N *no* X *ga* Adj (X of N is Adj)" (1-2 in predicative usage) and "Adj X *no* N (N of Adj X)", a lot of examples were extracted from newspapers.

Among the above examples, we observed examples of the following in this paper:

(1-2) "N no X ga Adj," (X of N is Adj)
(2) "Adj X no N," (N of Adj X)
(3-1) N *wa* Adj + X *ni naru* (N becomes Adj + X)
(3-2) Adj + X *de* V (N V in/from/with … Adj X)

3.2.3 Linguistic clues

1) Predicative and adnominal usage

Rules to detect nouns which refer to upper-level concepts of adjectives are mostly applied to both predicative and adnominal usage.

Rule 1) Omission of "X"

Here is an example of being able to omit "X" without changing its meaning.

(12) The area of this room is large. = This room is large.
X N Adj = N Adj

Even if "X" is omitted, "N wa Adj (N is Adj)" or "Adj N (adnominal usage)" is completed, inferring that the meaning of the "X" Japanese pattern is as follows:

N *no* X *ga* Adj = N *wa* Adj
(X of N is Adj ← N is Adj)

If an example fulfills this rule, we choose the example. After choosing data, we check chosen examples by using some rules for excluding unsuitable examples.
The rule for exclusion of unsuitable examples is as follows.

Rule 2) (For adnominal usage)

In the "Adj X no N (N of Adj X)" pattern, if "X no N (N of X)" is completed without an adjective, the example is excluded, because, if X is a transparent word, "X no N (N of X)" is not completed without concrete expression.

Rule 3) Change "X" into "because of X"

(13) The lens of my glasses is thick.
X N Adj
(14) The depth of this hole is shallow.
X N Adj

When we compare the semantic relations between "X" and an adjective in examples 14 and 15 (that is, the relation between "lens" and "thick", and the relation between "depth" and "shallow"), the distance of the semantic relation between N and an adjective is different. The relation between "depth" and "deep" seems to be a "concept and its instance" relation, on the other hand, the relation of "lens" and "thick" seems to be related to each other via an intermediate concept like "thickness."
In our data, examples like the above example are not chosen. In this case, we can change "The lens of my glasses are thick" into

"My glasses are thick because of the lens."
N Adj X

On other hand, we cannot change "The depth of this hole is shallow" into

"This hole is shallow because of the depth."
N is Adj X

In the case of the above examples, we will choose "The depth of this hole is deep" instead of "The lens of my glasses is thick."

In Japanese,

(15) *Watashi no megane wa renzu ga buatsui.*
My glasses (Topic) the lens (is) thick.
Watashi no megane ga buatsui no wa renzu no tameda.
My glasses are thick (Topic) because of the lens.

Rule 4) Exclude (1) ISA relation between N and X, or (2) N is apposition for X, or (3) N is the concrete content of X.

A Japanese example is shown below:

(16) *Kono sakura no ki wa furui.*
This cherry (of) tree is old.
N X Adj
This tree of cherry is old.

In this example, the relation between "N" and "X" is an ISA relation, that is, "the cherry is a tree." Therefore, "X " is not the upper-level concept of an adjective but that of "N." In this example, "tree" is not an upper-level concept of "old" but that of "cherry."

Rule 5) Basically, if X is an action noun, we do not choose it.

Rule 6) (for adnominal usage) Exclude the pattern in which N refers to an abstract noun like a particle.

(17) *kibishii kun'ren no naka*, …..
hard training in
Adj X (of) N
In the hard training, …..

In Japanese we can say both "kibishii naka (in hard(-state), N + Adj)" and "kibishii kun'ren no naka (in hard training, N + Adj no X)." In this case there is no clue to decide whether "X" is an abstract concept of an adjective.

2) Adverbial usage

Rule 1) Omission of "X"
An example of adverbial usage is shown as follows.

(18) He negated with a strong tone.
V Adj X
= He negated strongly.
V Adj-ly

In Japanese

(3-1) N *wa* Adj + X *ni naru* = N wa Adj ku/ni naru
(N becomes Adj + X) = (N becomes Adj-*ly*)
(3-2) Adj + X *de* V = Adj ku/de V
(V with/in Adj + X) = (Adj-*ly* V)

The rule for excluding unsuitable examples is as follows:

Rule 2) (for Adj + X *de* V) "X de V"
If "X *de* V" can be completed semantically, it is excluded. If "X" is an abstract concept of an adjective, "X *de* V" is not completed because the transparent word "X" needs concrete expression.

Rule 3) Exclude (1) ISA relation between N and X, or (2) N is apposition for X, or (3) N is the concrete content of X.

Rule 4) Exclude a time expression and a number of times.

Rule 5) Exclude the "(N wa Adj) X ni naru."

(19) *Houan no kaketsu* *wa*
The approval of the bill (Topic)
N
kon'nan na mitoshi *ni naru*
difficult expectation become
Adj X
The approval of the bill is expected to be difficult.

In this example, "N *wa* Adj (The approval of the bill is difficult)" represents the content of "X" (in this example "X" refers to "expectation"). Therefore, "X" is not an abstract concept of an adjective but the encapsulating expression of "the approval of the bill is difficult (N is Adj)."

Rule 6) (for Adj + X de V) Exclude V in the quotation, conjunction particle, case of an implement, case of a location.

4. Comparison of our extracted data with Japanese WordNet

Currently, we obtained the following examples.

(1-2) "N no X ga Adj," (X of N is Adj)
481 examples (data being processed)/ 61210 examples
(2) "Adj X no N," (N of Adj X)
2020 /total 48624 examples
(including Adj which is a part of a phrase)
1716/48624
(without Adj which is a part of a phrase)
(3-1) N *wa* Adj + X *ni naru* (N becomes Adj + X)
123/700 examples
(3-2) Adj + X *de* V (N V in/from/with … Adj X)
5/52 examples

The total number of examples we extracted was 2325. Differences of abstract nouns were 267 words, and differences of adjectives were 900 words.
We compare our extracted data with the domain of adjectives in Japanese WordNet First version of Japanese WordNet started from translating WordNet 3.0. Based on the translation of about 5000 core synsets created automatically using multiple existing thesauri, it has been modified and extended manually (Isahara et.al 2008; Bond et.al 2008).
In the structure of WordNet, there are "synset (a group of synonyms)" and "pointer" indicating a relation with another synset. "Pointer" indicates "Derivationally related form," "Also see," "Antonym," "Attribute," and "similar to," "Pertainym" and so on. Each pointer links words to related synonyms.
In the adjective domain of Wordnet, the pointer "Attribute" is linking an adjective to synonyms representing the upper-level concept of adjectives. For example, "high" is related to "[degree][grade][level]" via the pointer "Attribute."
First of all we investigated how the extracted 267 abstract nouns are covered with synonyms of "Attribute" in WordNet.
The number of "Attribute" of adjectives in WordNet is 639.
Among them, 146 abstract nouns corresponding to "Attribute" were obtained from our data. From a viewpoint of Japanese WordNet, the degree of agreement of "Attribute" between our data and Japanese WordNet was 22%. On the other hands differences of our obtained abstract nouns were 267 words. From a viewpoint of our data, the degree of agreement of abstract nouns between them was 54%.

Some attribute concepts which do not appear in Japanese WordNet but do exist in our data are, for examples, "*nedan* (price)," "*kimochi* (feeling)", "*joukyou* (situation)" and so on. Japanese WordNet has not

completed yet, therefore there may be differences of expression between our data and Japanese WordNet. Then the upper-level concept of adjectives has not translated into Japanese yet.

In our data, "*hiroi* (large)" and "*semai* (small, limited)" are related to "*menseki* (area)" or "*han'i* (range)," but, in this Japanese WordNet, they are related to only "*ookisa* (size)." The Japanese words need to be carefully checked by humans.

5. Application of Linguistic Analysis: Compilation of 3-dimensional semantic map

In a future work, we will distribute nouns extracted in Section 3 using Kohonen's Self-Organizing map (1995). We are now trying to distribute abstract concepts of adjectives in similarity and hierarchical relations on the map by introducing the Top node, which means the most abstract concept. Our first experiment is shown below.

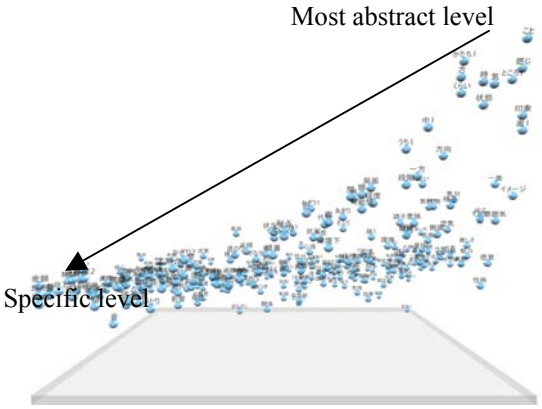

Fig. 1. Distribution of concepts on self-organizing map

This work is still underway, but the map shows the possibilities of constructing objective thesauri from corpora. We compared the upper level of "warm" concepts in our results with those of WordNet.

Upper level concepts of "warm" in our results:
[TOP]
→[feeling]
→[state/situation] → [temperature/climate]
→[feeling]/[emotion]→[attitude]→
[characteristics of something/someone]
Those in WordNet are as follows:
→ [emotionalism/emotionality]
→[temperature]

WordNet defines the attribute concepts of "warm" as "emotionalism/emotionality" and "temperature." In our data "characteristics" is a hyponym of [feeling/emotion]→[attitude], and "temperature" is a hyponym of [feeling]→[state/situation].
In the future we will consider whether or not the taxonomic structure of concepts of adjectives extracted from corpora will be helpful for verifying existing thesauri.

References

Matthew Berland, and Eugene Charniak. (2000). Finding Parts in Very Large Corpora. In *Proceedings of 38th Annual Meetings of the Association for Computational Linguistics (ACL)*: 57–64.

Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, Kiyotaka Uchimoto (2008), Bootstrapping a WordNet using multiple existing WordNets. In *Proceedings of 6th edition of the Language Resources and Evaluation Conference* (*LREC*2008) *will appear*.

Sharon A. Caraballo. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meetings of the Association for Computational Linguistics (ACL)*: 120–126.

Marti A. Hearst. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of 14th International Conference on Computational Linguistics (COLING)*: 539–545.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, Kyoko Kanzaki, (2008), Development of the Japanese WordNet. In *Proceedings of 6th edition of the Language Resources and Evaluation Conference* (*LREC*2008) *will appear.*

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)

Takahashi, Takahashi.(1975) *Various Phases Related to the Part-Whole Relationship Investigated in the Sentence*, Studies in the Japanese language 103, The Society of Japanese Linguistics, 1–16 (in Japanese).

Patrick Pantel and Deepak Ravichandran. (2004). Automatically Labeling Semantic Classes. In *Proceedings of Human Language Technology / North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*: 321–328.

Schmid, Hans-Jorg. (2000): English Abstract Nouns as Conceptual Shells, Mouton de Gruyter.