Construction of a Japanese Relevance-tagged Corpus

Daisuke Kawahara
University of Tokyo
7-3-1, Hongo Bunkyo-ku,
Tokyo, 113-8656, Japan
kawahara@kc.t.u-tokyo.ac.jp

Sadao Kurohashi
University of Tokyo
& PRESTO, JST
7-3-1, Hongo Bunkyo-ku,
Tokyo, 113-8656, Japan
kuro@kc.t.u-tokyo.ac.jp

Kôiti Hasida
CARC, AIST
& CREST, JST
2-41-6, Aomi Koto-ku,
Tokyo, 135-0064, Japan
hasida.k@aist.go.jp

Abstract

This paper describes our corpus annotation project. The annotated corpus has relevance tags which consist of predicate-argument relations, relations between nouns, and coreferences. To construct this relevance-tagged corpus, we investigated a large corpus and established the specification of the annotation. This paper shows the specification and difficult tagging problems which have emerged through the annotation so far.

1 Introduction

A text has several types of relevance between words/phrases, such as predicate-argument relations, relations between nouns, and coreferences. Syntactic structure of a text indicates only a small part of them. To understand a text, it is necessary to recognize implicit relations as well as syntactically explicit relations. As a first step to recognize them by computers, we started a project which constructs a Japanese corpus marked with relevance.

To construct the corpus, we must investigate real texts and establish the specification of the corpus annotation: what expressions these relations have and how to annotate them. So far, however, these relations have not been investigated on a large scale, and existent corpora with these relations are not large or have only a small part of them(Marcus et al., 1994; Takezawa et al., 1998; Marcu et al., 1999; Poesio, 2000).

Our project utilizes the Kyoto University corpus (Kurohashi and Nagao, 1998) which consists of 40,000 syntactically tagged sentences (3500 newspaper articles; 11.4 sentences per article). Tags are assigned to words in each article.

In September 2001, we made a draft of the specification of the annotation, and started trial annotation. In the trial, we asked two annotators to tag the same articles, checked their difference, and established the specification of the annotation. The specification was getting stable when tagging of about 1,000 sentences was finished, and we started the constant annotation in February 2002. So far, 1,300 sentences have been tagged, and an annotator can annotate 12 sentences per hour.

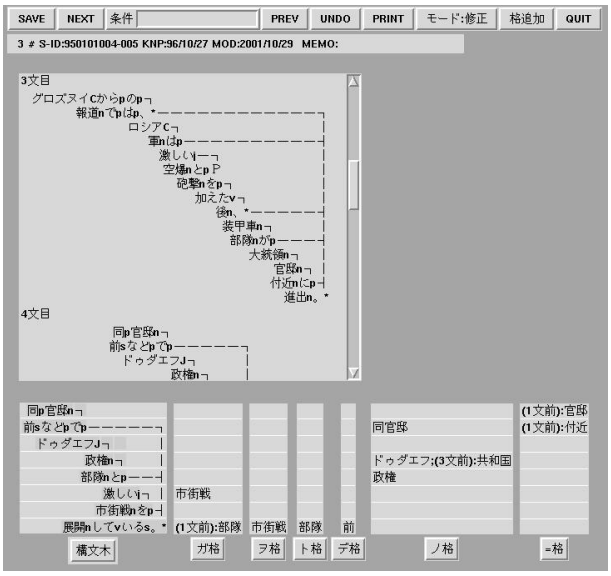

Figure 1: Annotation tool

2 Tags

We give relevance tags to words using the annotation tool which was developed through the Kyoto University corpus project (Figure 1). This annotation is performed by modifying tags that are automatically provided by our case and ellipsis analyzer. We deal with the following three classes of relevance.

2.1 Predicate-argument relations

In Japanese, postpositions function as case markers such as *ga* (nominative), *wo* (accusative), and *ni* (dative)[1]. A tag consists of an argument word and a case-marking relation (postposition itself), and is given to its predicate.

(1) *Taro* - *ga* *shimbun* - *wo* *yonda*.
Taro nom newspaper acc read

(Taro read a newspaper.)

yonda ⇐ *ga*:*Taro*
wo:*shimbun*

In this example, *Taro* and *shimbun* 'newspaper' modify *yonda* 'read', and are arguments of *yonda*. The relation between *Taro* and *yonda* is *ga* (nominative), which is indicated by the postposition following *Taro*, and the relation between *shimbun* and *yonda* is *wo* (accusative). Accordingly, the tags "*ga*:*Taro*" and "*wo*:*shimbun*" are given to *yonda*.

The important decision we have to make is whether surface cases or deep cases are used as the relations. If we use semantic relations such as deep cases, it is difficult to make the set of the relations and to select one relation for tagging, because boundaries of them are not clear. This makes the annotation task more difficult. So, our project employs surface cases as the relations.

The tags in (1) are provided correctly by our automatic analyzer. However, Japanese has two phenomena which often cause incorrect automatic analyses: disappearance of case markers and omission of arguments (zero-pronouns).

(2) *Taro* - *ga* *shimbun* - *wo* *yonda.*
Taro nom newspaper acc read

Taro - *wa* *yoku* *yomu*.
Taro TM often read

(Taro read a newspaper. Taro often reads ϕ.)

yomu ⇐ *ga*:*Taro*
wo:*shimbun*

In the second sentence, *Taro* has a case-marking relation to *yomu* 'read', but this relation is hidden by a topic marker (TM) *wa*. Since its actual case is nominative, the tag "*ga*:*Taro*" is given to *yomu*. In addition, the accusative of *yomu* is a zero-pronoun. Its referent is *shimbun* 'newspaper', so the tag "*wo*:*shimbun*" is given to *yomu*. Since the automatic analyzer possibly produces incorrect tags, it is necessary for annotators to modify incorrect ones.

Predicate-argument tags are also given to nouns which mean actions.

(3) *Kare-no* *daigaku* *nyuugaku* - *wa*
his university admission TM

yoi *news* *da.*
good be

(His admission to the university is good news.)

nyuugaku ⇐ *ga*:*Kare*
ni:*daigaku*

The noun *nyuugaku* 'admission' means an action of admitting. We assign tags to *nyuugaku* by considering it as a verb 'admit'.

2.2 Relations between nouns

Not only predicates but also nouns have some intrinsic relations with other nouns in a text. When two nouns in a text are related to each other, a tag is given to the latter noun.

(4) *Taro* - *wa* *se ga hikui.*
Taro TM short

Shikashi *imouto* - *wa* *se ga takai.*
but sister TM tall

(Taro is short. But his sister is tall.)

imouto ⇐ *no*:*Taro*

Since *imouto* 'sister' means "*Taro no imouto*" 'Taro's sister', the tag "*no*:*Taro*" is given to *imouto*, though "*Taro no*" does not appear in the sentence. In this example, *imouto* requires intrinsic relations to other nouns. This is a so-called relational noun. *no* in Japanese has many meanings, but all of them are tagged as one relation *no* for the same reason as marking with surface cases.

Not only relational nouns but also almost all of nouns have some intrinsic relations: *kuruma* 'car' and *handle*, *mado* 'window' and *curtain*. We also handle these relations.

[1] In the examples of this paper, we use the abbreviations of the cases: nom (nominative), acc (accusative), dat (dative).

2.3 Coreferences

When two nouns refer to the same entity, these two nouns are coreferential. To mark a coreference relation, "=" is used. A tag of this relation is given to the latter noun of two coreferential nouns.

(5) *Taro* - *wa* *futotteiru.* *Kare* - *wa*
Taro TM be fat he TM

itsumo *nanika* *tabeteiru.*
always something be eating

(Taro is fat. He is always eating something.)

$$Kare \Leftarrow =:Taro$$

In this example, *Kare* 'he' refers to *Taro*, and the tag "=:*Taro*" is given to *Kare*.

These coreference tags are given to not only pronouns but also definite noun phrases as follows:

(6) *Onnanoko* - *ga* *aruiteiru.*
girl nom be walking

ano *onnanoko* - *wa* *Mary* *da.*
that girl TM be

(A girl is walking. That girl is Mary.)

$$onnanoko \Leftarrow =:Onnanoko$$

When two nouns do not refer to the same entity but have an is-a or generic/non-generic relation, "≒" is used to mark this relation instead of "=".

(7) *kuruma* - *no* *hanbai-daisuu* - *wo*
car of sale acc

miruto, *jikayousya* - *wa* ...
check owner-driven car TM

(When we check the sales of cars, owner-driven cars are ...)

$$jikayousya \Leftarrow ≒:kuruma$$

Since *kuruma* 'car' and *jikayousya* 'owner-driven car' have an is-a relation, the tag "≒:*kuruma*" is given to *jikayousya*.

(8) *Chiisana* PC_1 - *ga* *ureteiru.*
small nom-CM be selling

Taro no PC_2 - *wa* *chiisai* *ga,*
Taro's TM small but

Hanako no PC_3 - *wa* *furuku-te* *ookii.*
Hanako's TM old big

(Small PCs are selling. Taro's PC is small, but Hanako's PC is old and big.)

$$PC_2 \Leftarrow ≒:PC_1$$
$$PC_3 \Leftarrow ≒:PC_1$$
$$PC_3 \Leftarrow ≒:PC_2$$

PC_1 is a generic noun, but PC_2 and PC_3 are non-generic nouns. Accordingly, the tag "≒:PC_1" is given to PC_2 and PC_3. PC_2 and PC_3 are not the same entity but are related indirectly in this text, because PC_2-PC_1 and PC_3-PC_1 are linked by "≒" relations. The tag "≒:PC_2" is given to PC_3.

3 Difficult Tagging Problems

The following is the difficult problems of the annotation.

3.1 The tagging unit of the annotation

The tagging unit is a word, but the notion of a word is not clear in Japanese. A compound noun can be one word or several words, since Japanese sentences have no word segmentation. For example, *hounichi* 'a visit to Japan' is one word in our dictionary. It is, however, also regarded as *hou* 'visit' and *nichi* 'Japan'. In our framework, the latter segmentation is better, because it is necessary to annotate the relation between these two words. For example, *nichi* refers to *nippon* 'Japan' in an article. In such cases, we modify the word segmentation of the original corpus. An annotator must consider whether a word is appropriately segmented or not at every moment of tagging.

3.2 Tags with multiple referents

There is a case that a tag has more than one referent. This is divided into two cases.

It is the first case that every referent in a tag is obligatory. When a predicate has two arguments and they are coordinate, both of them are tagged to their predicate.

(9) *Taro* - *to* *Hanako* - *ga*
and nom

gakko - *kara* *kaetta*.
school from be back

(Taro and Hanako were back from the school.)

kaetta ⇐ *ga*:*Taro*,*Hanako* [and]

In this example, since *Taro* and *Hanako* are coordinate, both of them play a role of *ga* (nominative) case of their predicate *kaetta* 'be back'. A tag which consists of both of them is given to *kaetta* with "and" flag, which means all of the elements are obligatory.

On the other hand, when two nouns are not coordinate and their predicate subordinates *to* (with) case, tags are assigned differently from the above.

(10) *Hanako* - *to* *Taro* - *ga*
with nom

kekkon-shita.
get married

(Taro got married with Hanako.)

kekkon-shita ⇐ *ga*:*Taro*
to:*Hanako*

In this example, *kekkon-shita* 'get married' subordinates *to* (with) case.

Even though two nouns,whose relation is indicated by *ka* 'or', *ya* 'or' and so on, are logically alternative, they are tagged with "and" flag.

(11) *Watashi* - *wa* *Canada* *ka*
I TM or

America - *ni* *sumitai*.
in want to live

(I want to live in Canada or America.)

sumitai ⇐ *ni*:Canada, America [and]

Canada and America is alternative in this sentence. Since both of them are arguments of their predicate *sumitai* 'want to live', their tag has "and" flag.

It is the second case of the multiple referent phenomena that any of the referents in a tag are proper. When there are more than one proper entity for a referent in a text, a tag which consists of all of them is given to their reference word with "or" flag, which means one of the elements is a referent.

(12) *Kouchi* *ken* - *no* *Hashimoto*
pref. of

chiji - *wa* *kokusekijoukou* - *wo*
governor TM nationality clause acc

teppai-suru *houshin-da*
abolish plan

(The governor of *Kouchi* pref. *Hashimoto* plans to abolish the nationality clause.)

teppai-suru ⇐ *ga*:*Kouchi ken, chiji* [or]

ga (nominative) case of *teppai-suru* 'abolish' can be understood as *Kouchi ken* '*Kouchi* pref.' or *chiji* 'governor'. A tag which consists of both of them is given to *teppai-suru* with "or" flag.

3.3 Noun-modifying clauses

When a noun is modified by a clause, the noun normally has a predicate-argument relation to its predicate.

(13) *kare* - *ga* *katta* *kuruma*
nom buy car

(the car which he bought)

katta ⇐ *ga*:*kare*
wo:*kuruma*

In this example, *kuruma* 'car' is a noun modified by a clause and is an argument of the predicate of the clause *katta* 'buy'. Since the relation between *kuruma* and *katta*, which does not appear in the sentence, is *wo* (accusative) case, the tag "*wo*:*kuruma*" is given to *katta*.

On the other hand, there is a case that the noun and its predicate do not have a predicate-argument relation.

(14) *kare* - *ga* *wairo* - *wo*
he nom bribe acc

uketotta *jijitsu*
receive fact

(the fact that he received the bribe)

uketotta ⇐ *ga* :*kare*
wo :*wairo*
non-gapping:*jijitsu*
jijitsu ⇐ content :*uketotta*

In this example, *jijitsu* 'fact' does not have a predicate-argument relation to *uketotta* 'receive'. The tag "non-gapping:*jijitsu*" is given to *uketotta*.

In addition, the predicate in a clause also has a relation to the modified noun in reverse. This relation has two types: 'content' and "*no*", which is used to mark relations between nouns. In the above example, *uketotta* has a content relation to *jijitsu*, because the clause of *uketotta* is a content clause.

Next, we show a "*no*" example.

(15) *Hanako* - *ga* *ryokou* - *ni*
nom travel acc

dekakeru *zenjitsu*
depart the day before

(the day before *Hanako* departs to travel)

dekakeru ⇐ non-gapping:*zenjitsu*
zenjitsu ⇐ *no* :*dekakeru*

zenjitsu 'the day before' has a relative relation to *dekakeru* 'depart', because the day before is relatively before the day of departure. In this case, the tag "*no*:*dekakeru*" is given to *zenjitsu*. The reason why we use "*no*" as this relation is that we can paraphrase the above expression into "*dekakeru hi no zenjitsu*" in Japanese.

The following example does not include a noun-modifying clause, but has 'content' relation.

(16) *seijika* - *ga* *wairo* - *wo*
politician nom bribe acc

uketotta. *Sono* *jijitsu* - *wa* ...
receive the fact TM

(The politician received a bribe. The fact is ...)

uketotta ⇐ *ga* :*seijika*
wo :*wairo*
jijitsu ⇐ content:*uketotta*

In this example, *jijitsu* 'fact' in the second sentence is tagged, because it refers to the first sentence.

3.4 Unspecified people

Some referents are not specific entities, but people without antecedents expressed in a text. These are tagged as "Unspecified people".

(17) *Kore* - *ga* *sekai* *saisoku no*
this nom world fastest

keisanki da - *to* *iwareteiru*.
computer that be said

(It is said that this is the fastest computer in the world.)

iwareteiru ⇐ *ni*:Unspecified people

In this example, *ga* (nominative) case of *iwareteiru* 'be said' is unspecified people.

(18) *sonoyouna* *kitei* - *wa* *nai.*
such regulation TM not

(There is not such regulation.)

kitei ⇐ *ga*:Unspecified people

There is no referent of *ga* (nominative) case of *kitei* 'regulation' in the text, and it is unspecified people.

4 Conclusion

This paper described our corpus annotation project. The corpus has relevance which consists of predicate-argument relations, relations between nouns, and coreferences. Such

linguistic/semantic annotations can be exploited to enhance NLP systems such as machine translation, information retrieval, and automatic summarization. They are useful also for end-user content, as advocated by GDA (http://i-content.org/GDA/), MPEG-7 (http://mpeg.telecomitalialab.com/), Semantic Web (http://www.semanticweb.org/), and so forth.

References

Sadao Kurohashi and Makoto Nagao. 1998. Building a Japanese parsed corpus while improving the parsing system. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, pages 719–724.

Daniel Marcu, Estibaliz Amorrortu, and Magdalena Romera. 1999. Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL'99 Workshop on Standards and Tools for Discourse Tagging*, pages 48–57.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert Macintyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the Human Language Technology Workshop*, pages 110–115.

Massimo Poesio. 2000. Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pages 211–218.

Toshiyuki Takezawa, Tsuyoshi Morimoto, and Yoshinori Sagisaka. 1998. Speech and language database for speech translation research in ATR. In *Proceedings of Oriental COCOSDA Workshop*, pages 148–155.