Building a Diverse Document Leads Corpus Annotated with Semantic Relations

Masatsugu Hangyo **Daisuke Kawahara** **Sadao Kurohashi**

Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, 606-8501, Japan
{hangyo,dk,kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

In these days, semantic analysis has been actively studied in natural language processing. For the study of semantic analysis, corpora with semantic annotations are essential. Although there are such corpora annotated on newspaper articles, there are various genres and styles, including linguistic expressions that are not found in newspaper articles. In this paper, we build a diverse document leads corpus annotated with semantic relations. To reduce the workload of annotators and annotate as many various documents as possible, we restrict the annotation target of each document to only the first three sentences. We have completed building a corpus of 1,000 documents and report the statistics of this corpus.

1 Introduction

In recent years, semantic analysis including predicate-argument structure analysis and anaphora resolution, has been studied as a subsequent task of syntactic parsing. Most existing studies of semantic analysis have used newspaper corpora with manual annotation. However, there are sources other than newspapers, such as encyclopedias, diaries and novels each with diverse styles in each genre. There are linguistic phenomena that rarely appear in newspapers such as requests and honorific expressions. To deal with texts that include the above phenomena, it is essential to build an annotated corpus that includes diverse-domain documents. Web pages include various genres and text styles such as news articles, encyclopedia articles, blog and business pages. Using web pages as the target documents of annotation, we build a Japanese annotated corpus that consists of various genres.

We annotate predicate-argument structures and anaphoric relations as semantic relations. We illustrate these relations and annotations in Example (1)[1]. "A←*rel*:B" represents annotating B to A with relation *rel*. In the following examples, we sometimes omit annotations that are not related to the discussion.

(1) a. 太郎は 時計を 買った。
Taro-TOP *watch*-ACC *bought.*
'Taro bought a watch.'
(買った ← GA:太郎, WO:時計)

b. 弟に それを あげた。
Little brother-DAT *it*-ACC *gave*
'He gave it to his little brother.'
(弟 ←NO:太郎
それ ←=:時計
あげた ←GA:太郎, NI:それ, WO:弟)

Predicate-argument structures express the relations between a predicate and its arguments. In Example (1a), the GA (nominative) case of 買った (bought) is 太郎 (Taro) and the WO (accusative) case of 買った is 時計 (watch). In this example, there is a topic marker (は) which hides the case relation between 太郎 and 買った. Since the hidden actual case relation is GA, we annotate which the GA case of 買った is 太郎. Such disappearances

[1] In this paper, we use the following abbreviations: NOM (nominative), ABL(ablative), ACC (accusative), DAT (dative), ALL (allative), GEN (genitive), CMI (comitative), CNJ (conjunction), INS(instrumental) and TOP (topic marker).

26th Pacific Asia Conference on Language,Information and Computation pages 535–544

of case markers occur also when a topic marker も (too) is used and when an argument is modified by the predicate.

Anaphora is a phenomenon that an expression in text (anaphor) refers to other expressions (referent). In Example (1b), それ (it) refers to 時計 (watch) in the first sentence. In Japanese, ellipses of arguments of a predicate frequently occur. They are called zero anaphora because it is considered that there exist unseen pronouns, which are called zero pronoun, in the place where the ellipsis occurs. By annotating 太郎 with the GA case of あげた (gave), we can express that there is a zero pronoun in the GA case and the referent of the zero pronoun is 太郎. Additionally, we deal with exophoric relations, whose referents do not appear in the document.

There are bridging references among the anaphoric relations. In bridging references, anaphors do not refer to referents directly but some attributes of anaphors refer to antecedents. In Example (1), we can consider that 弟 (little brother) has an attribute “big brother” that refers to 太郎. Various attributes such as hypernym-hyponym, part-whole and contrast relations refer to the referent in bridging references.

We annotate can morphological and syntactic information independently of each sentence and thus the labor of annotators increases linearly with document length. In contrast, since annotating semantic relations deals with inter-sentence relations, elements that annotators should consider increase combinationally. Therefore, if we attempt to annotate whole documents, the annotation processing time of each document becomes longer and few documents could be annotated. Since our target is building a corpus that consists of various documents, we confine the annotation target to the first several sentences. Semantic analysis systems usually use the results of previously analyzed sentences and analysis errors propagate to the following analyses. By building a corpus that consists of document leads, we expect to raise the accuracy of the analysis of both document leads and the document as a whole.

In this paper, we describe related work in Section 2. We describe the documents that the corpus consists of in Section 3 and the annotation criteria in Section 4. In Section 5 we discuss the statistics and properties of the corpus and conclude in Section 6.

2 Related Work

Existing corpora that are annotated with predicate-argument structures and anaphoric relations include the Kyoto University Text Corpus (Kawahara et al., 2002) and the Naist Text Corpus (Iida et al., 2007). These corpora are based on Mainich Newspaper articles from 1995 and annotated with predicate-argument structures and anaphoric relations. Since there are only reports and editorial articles in the newspaper, the writing styles are consistent, making it difficult to adapt a semantic analysis system based on this corpus to texts other than newspaper articles.

Corpora that consist of documents from various genres include the Balanced Corpus of Contemporary Written Japanese (BCCWJ)[2]. BCCWJ includes publications such as books and magazines and text from the Internet. BCCWJ has publications form various genres but the Internet text in BCCWJ is restricted to blogs and forums. For this reason, although company pages and other pages exist on the Internet, they are not included.

Ohara annotated predicate-argument structures defined in FrameNet to the predicates in BCCWJ (Ohara, 2011). Although the predicate-argument structures of FrameNet include the existence of zero pronoun, referents are not annotated if the referents do not exist in the same sentence. Furthermore, since anaphoric relations are not annotated, they do not annotate the inter-sentence semantic relations.

In other languages, corpora dealing with multiple genres include Z-corpus (Rello and Ilisei, 2009) and LMC (Live Memories Corpus) (Rodríguez et al., 2010). Z-corpus consists of Spanish law books, textbooks and encyclopedia articles, and they are annotated with zero anaphoric relations. They only treat zero anaphora and do not treat other anaphora and predicate-argument structures. This is because the zero anaphoric relations can be annotated independently of predicate-argument structures since the pronoun-dropping only occurs in subject in Spanish.

LMC consists of Italian wikipedia and blogs and are annotated with anaphoric relations. They deal with zero anaphora as a part of anaphora, but do not deal with predicate-argument structures. Since pronoun-dropping only occurs in subject also in Italian, they regard the predicates that contain pronoun-

[2] http://www.tokuteicorpus.jp/

Headline : *2008. 07. 10 Thursday*

(1) 気が　つけば　梅雨も
Mood-NOM *stick* *rainy season*-NOM

明けてました。
have ended.

'I think that the rainy season has ended.'

(2) 毎日　暑い　日が　続きますね。
Everyday hot day-NOM *continue.*

'It's hot every day.'

(3) 父の　手術も　終わり
Father-GEN *surgery*-NOM *finish*

少しだけ ほっとしています。
short feel easy.

'I'm feeling a little better because my father's surgery is over.'
(The rest is omitted.)

Figure 1: Example of a document whose headline does not appear in the body

dropping as anaphors.

3 Annotation Target Document

Most existing corpora annotated with semantic relations consist of newspaper articles. However, there are linguistic phenomena that rarely occur in newspaper articles, and thus we need to target various documents in order to study these phenomena. Using the web without limiting by domain, we collect various documents. To build the annotated corpus consisting of various documents, we need to reduce the workload of each document. We limit the annotating targets to the first three sentences of the document leads. The target number of documents in this corpus is 1,000 documents.

There are many inadequate documents that should not be included in the corpus in the web documents. Checking and filtering them all manually is time-consuming. The number of documents in the web is much more than the target number of documents. Therefore, we first filter out inadequate documents automatically by simple rules. Then, the remaining documents are checked manually and we only annotate the adequate documents.

Headline : 売布神社 'Mefu shrine'

(1) どもども、　森田です。
Hi, *be Morita.*

'Hi, I'm Morita.'

(2) さてさて、　前回
Now, *previous time*

中山寺に　行きましたが、　その
Nakayama temple-LOC *went but,* *that*

続きです。
continuation

'Now, this is the continuation of my previous article when I went to Nakayama temple.'
(Three sentences are ommited)

(6) この 池の　左上あたりに
This pond-GEN *upper-left*-LOC

歩いていくと、 売布神社 に
walk to *Mefu shrine*-LOC

着きます。
reach

'Walking around the upper-left of the pond, I had reached Mefu shrine.'
(The rest is omitted.)

Figure 3: Example of a document that cannot be understood without its headline

3.1 Inadequate Documents for Semantic Annotation

Language is used based on a shared situation between a speaker/writer and an audience/reader. The topic of the speech and the document has some sort of relevance to the situation.

When annotating the morphological and syntactic information, there is no need to consider this shared situation because of dealing with each sentence independently. However, in a semantic relation corpus, the shared situation must be considered. Since we deal with only text as our annotation target, documents retering to figures, tables and hyperlinks are inadequate for this corpus.

Some documents have headlines and they often have a key role to interpret the documents. However, we remove the headlines from the annotation target

Headline : 地震被害 264 億円に 県まとめ ‘The damage caused by the earthquake reached 26.4 billion yen according to Prefectural survey

(1) 岩手・宮城内陸地震の 被害は 22 日現在、
Iwate-Miyagi inland earthquake-GEN *damage*-TOP *as of 22nd,*

県災害対策本部の まとめで 264 億円に 膨らんだ。
disaster countermeasures office of prefecture-GEN *survey*-INS *26.4 billion*-ACC *swelled.*

‘According to a survey by The Disaster Countermeasures Prefectural Office, the damage to Iwate-Miyagi inland earthquake swelled to 26.4 billion as of the 22nd.’

(2) 依然として 農村、土木関係を 中心に 被害が 拡大している。
Still *farming village and construction*-ACC *focus on damage*-NOM *is increasing.*

‘The damage is still increasing with focus on farming villages and construction.’
(The rest is omitted.)

Figure 2: Example of a document that the elements of its headline appear in the first three sentences

because some of the headlines are ungrammatical sentences such as series of noun phrases. In newspaper articles, there are sentences in the leads that are abstract of the whole document and most of such documents can be understood without headlines. In web pages, some documents do not have sentences acting as an abstract and some documents cannot be understood without headlines. On the other hand, if the headlines are the date of the blog articles, the documents can be understood without headlines. We discard documents that cannot be understood without their headlines.

We automatically determine if a document has a headline. Web pages have structure information such as HTML tag, but the headlines are sometimes described by tags other than the <h> tag, which renders headlines, and there is non-headline text which are marked up with <h> tags. Therefore, we determine the headline by the content of the text. If the first sentence does not end with punctuation or ends with a noun phrase, we determine that the first sentence is the headline, otherwise we determine that the document does not have a headline. If the first sentence is the headline, we extract the following three sentences. If the first sentence is not a headline, we extract the first three sentences. We deal with these extracted sentences as our annotation target. If the document cannot be understood with only these sentences, the document is not included in the corpus. Before manual filtering, the documents which seem that they cannot be understood without the headline are removed automatically. The understandable documents are determined by the following criteria.

If no words in the headline appear in the body of the document, it is assumed that removing the headline has little influence to understand the semantic relations. For example, in Figure 1 since the headline is the date, removing the headline have no effect on understanding the document. In case of that all the words in the headline appear in the first three sentences, it would be apparent that the semantic relations can be understood without the headline. In Figure 2, the first sentence has a role as the abstract and the all content words in the title appear in the first three sentences. In this case, the document can be understood without the headline. On the other hand, if the words in the headline are only mentioned after the first three sentences, it is hard to understand the document because it is impossible to reconstruct the information in the headline from the first three sentences. In Figure 3, 売布神社 (Mefu shrine) appears in the 6th sentence. However, 売布神社 does not appear in the first three sentences, so that it is difficult to understand the context that the author was going to Mefu shrine. Therefore, if the word in the headline only appears after the first three sentences, we determine that removing the headline makes the semantic relation difficult to be understood and we remove the document from the corpus automatically.

3.2 Determination of Inadequate Document

The documents collected form the web include many unsuitable documents for annotation. We determine that the following documents are difficult to annotate and are not included in the corpus.

Need technical knowledge to understand It is difficult to annotate documents that require technical knowledge because the annotator cannot understand these documents correctly.

Discontinuous sentences Collected documents possibly contain continuous sentences that are erroneously extracted from originally separated areas in the layout. Such documents are not suitable for inter-sentential semantic annotation.

Using too much slang It is difficult to annotate text that contains too much slang.

We automatically remove the documents that have the following sentences.

- End with a noun phrase: most of such sentences are rhetorical sentences or the part of a list.
- Not end with a Japanese period: these sentences are likely to be ungrammatical such as an error of the text extraction
- More than 10 phrases: the results are often caused by morphological analysis errors.
- Contain Roman characters: these are frequently used in technical terms, acronyms or slang in Japanese, and thus they indicate that the document is domain-specific or unnatural Japanese.
- Include stop phrases shown in Table 1: these phrases are defined to eliminate input forms and automatically generated pages.

Additionally, in order to remove duplicate pages, we remove documents whose edit distance is less than 50 to another document.

ボタンを押してください
(please push the button)
自動的に移動します
(should automatically go to another page)
検索できます
(can search)

Table 1: Examples of stop phrases

4 Annotation Criteria

4.1 Types of Annotation

We annotate many types of information: morpheme, phrase, dependency, named entity, predicate-argument structure and anaphoric relation. The predicate-argument structure and anaphoric relation correspond to semantic relations. The annotations of morpheme, phrase and dependency are necessary to annotate these semantic relations in order to define the annotation unit. A named entity is not needed to annotate the semantic relations, but we annotate named entities, as they provide good clues for semantic analysis.

We annotate morpheme, phrase and dependency by the criteria of the Kyoto University Text Corpus.

We define a basic phrase, which is composed of one independent word and preceeding and following attached words, as the annotation unit for the predicate-argument structure and the anaphoric relation. We show an example of the partitions by basic phrases in Example (2). We annotate the predicate-argument structure and the anaphoric relation to each basic phrase and the arguments and the referents are selected from basic phrases. If the referent is a compound noun, we consider the head basic phrase of the compound noun as the argument and the referent. In Example (2), the referent of 党 (party) is 国民新党 (People's New Party) and thus we annotate 新党 (new party), which is the head of 国民新党, as the referent.

(2) 7月17日 国民 新党 災害
July 17th People new party disaster
対策 事務 局長と して、
countermeasures office cheaf-ABL *do*
党 を 代表して 現地へ
party-ACC *represent field*-ALL
向かいました。
went
(党 ←=:新党)

We annotate the predicate-argument structure in the same way as the Kyoto University Text Corpus. The arguments are sorted into three types. One is the argument which has dependency relation with

Author
Reader
Unspecified-Person
Unspecified-Matter
Unspecified-Situation

Table 2: Candidate referents of zero exophora

ORGNIZATION
PERSON
LOCATION
ARTIFACT
DATE
TIME
MONEY
PERCENT

Table 3: The types of Named entity

predicate, another is the argument omitted in zero anaphora and the other is the argument omitted in zero exophora. In zero anaphora and zero exophora annotation, we annotate whether zero pronoun exists and also the referent of the zero pronoun as information of the argument. We show the candidate referents of zero exophora in Table 2.

In the Kyoto University Text Corpus, GA2 case is defined for double-subject construction and they are annotated as the following example.

(3) 彼は　ビールが　飲みたい。
He-TOP *beer*-NOM *want to drink.*
'He wants to drink beer.'
(飲みたい ←GA2:彼, GA:ビール)

In Example (4), since "象が長い" (The elephant is long) is a contrived expression, 象 is not handled as the argument of GA2 case under the basis of the Kyoto University Text Corpus. In contrast, we deal with words that express a topic as the argument of GA2 case and thus annotated "GA2:象, GA:鼻" to 長い.

(4) 象は　鼻が　長い。
Elephant-TOP *trunk*-NOM *long.*
'The elephant's trunk is long'
(長い ←GA2:象, GA:鼻)

The anaphoric relations are annotated according to the criteria of the Kyoto University Text Corpus. In the Kyoto University Text Corpus, the anaphoric relations are categorized into three types. The first of these is the anaphoric relation that has a coreference relation and we annotate this relation by using "=" tag. In Example (5), 自分 (himself) and ティーンエージャー (teenager) are coreferential and we annotate "=:ティーンエージャー" to 自分.

(5) ティーンエージャーが、 懸命に
Teenager-NOM *intently*
ライトセーバーを 振り回している
Lightsaber-ACC *be swinging*
自分の　姿を　密かに
himself-GEN *figure*-ACC *secretly*
ビデオに　収めた。
video-DAT *took.*
'A teenager secretly took a video of himself intently swinging a Lightsaber.'
(自分 ←=:ティーンエージャー)

The second anaphoric relations is the bridging reference that can be expressed in the form "A の B" (B of A), and we annotate "NO:A" to B. In 相手 (opposition) of Example (6), it is possible to express "ラズナーの相手" (the opposition of Rasner) and so we annotate "NO:ラズナー" to 相手.

(6) アタマの　先発は　ラズナー、
First-GEN *starter*-TOP *Rasner,*
相手は　陽と
opposition-TOP *You*-ABL
なっています。
is
'First starter is Rasner and the opposition is You.'
(相手 ←NO:ラズナー)

The third anaphoric relations is anaphoric relations that do not have a coreference relation and the bridging reference cannot be expressed in the form, "A の B" (B of A). We annotate these with "≃." In Example (7), the hyponym of 語学 (language study) refers to 英語 (English) in the first sentence and is a bridging reference and it is impossible to express "英語の語学" (language study of English). Therefore, we annotate "≃:英語" to 語学.

(7) 英語　力を　付けたい
English power-ACC *want to acquire*
読者の　ために 毎月　様々な
reader-GEN *for* *every month varied*
学習法を　特集します。
learning method-ACC *feature.*
'We feature varied learning methods for readers who want to acquire English-language ability every month.'

語学は モチベーションが
Language study-TOP *motivation*-NOM

大事。
important.

‘Motivation is important for language study.’
(語学 ←≃:英語)

In the Kyoto University Text Corpus, the referents of anaphoric relations are confined to the expressions that are mentioned in the document itself, but we additionally annotate exophora that refer to the author and the reader. The details of this are described in Section 4.2.

We annotate named entities according to the basis of IREX[3]. Named entities are expressed by their scope and type. The types of Named entity are 8 types shown in Table 3. In Example (8), ラズナー (Rasner) is annotated with “PERSON” and ホークス (Hawks) is annotated with “ORGANIZATION.”

(8) そこで ラズナー と ホークス の
And so *Rasner*-COM *Hawks*-GEN

今季 対戦 成績を
this season match-up result-ACC

掲載します。
post.

(ラズナー ←PERSON
ホークス ←ORGNIZATION)

In the actual annotation, we first automatically annotated by the Japanese morphological analyzer JUMAN[4] and the Japanese predicate-argument structure analyzer KNP[5], and then manually modified the annotation by using the GUI tool.

4.2 Mentions of Author and Reader

The author and the reader of the document are important in discourse. Since there are phenomena that are influenced by the author/reader and the author/reader tend to be omitted, the author/reader behave differently from other discourse elements. Because of this, it is important to detect which elements are the author/reader in the document.

Because the author/reader rarely appear in context of the newspaper corpus, the author/reader have not been treated properly in existing research. However, the author/reader often appear in context of the documents other than the newspaper articles. In case of the author/reader appearing in the document, the author/reader sometimes are not mentioned explicitly. In Figure 1, the author appears in the discourse but there is no mention of the author. On the other hand, the author/reader are mentioned in the documents by various expressions other than personal pronouns. Sometimes the mentions of the author/reader are proper names or position names. In Example (9), the author is mentioned by such as こま, (Koma) which is a proper name, 主婦 (housewife) and 母 (mother), which are the position name.

(9) 東京都に 住む 「お気楽
Tokyo-metropolis-LOC *live “easygoing*

主婦」 こま です。
housewife” be Koma.

‘I am Koma, an easygoing housewife living in Tokyo metropolis.’
(主婦 ←=:Author
こま ←=:主婦)

0 才と 6 才の
0 years old-COM *6 years old*-GEN

男の子の 母 を しています。
boys-GEN *mother*-ACC *doing*

‘I am the mother of two boys who a baby and 6 years old.’
(母 ←=:主婦)

Additionally, since personal pronouns are little-used in Japanese, it is difficult to identify which element is the author/reader[6]. Therefore, identifying which elements are the author/reader requires to annotate the mentions of the author/reader expricitly.

To annotate the mentions of the author/reader in discourse, we annotate “=:Author” and “=:Reader” to the mentions of the author/reader as exophora. Assuming that the author and the reader are only one element in each document, we annotate respectively “=:Author” and “=:Reader” up to one expres-

[3] http://nlp.cs.nyu.edu/irex/NE/df990214.txt

[4] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

[5] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP

[6] In English, it can be assumed that the expression which have a coreference relation with “I” is the author.

No. of documents	1000
No. of sentences	3000
No. of morphemes	59644
No. of phrases	18905
No. of basic phrases	23938
No. of annotated basic phrases	14865

Table 4: Statistics of the corpus

	Explicit	Implicit	No appearance
Author	258	364	378
Reader	105	290	605

Table 5: Author/reader appearance in documents

sion. If the author/reader is mentioned in some expressions, which are coreferential, we annotate it to one of them. In Example (9), the three underlined parts are the author mentions and thus we annotate "=:Author" to 主婦.

In the web site of an organization such as a company, the site administrator often writes the document on behalf of the organization. In such case, we annotate the organization as the author. In Example (10), it is thought that the site administrator wrote the document to represent 神戸徳州会病院 (Kobe Tokusukai Hospital), and so 病院, which is the head of 神戸徳州会病院, is annotated with "=:Author."

(10) 神戸 徳州会 病院では 地域の
Kobe Tokusukai hospital-TOP *area*-GEN
医療 機関との 連携を
medical agency-COM *coordination*-ACC
大切にしています。
value
'Kobe Tokusukai Hospital values coordination with community medical agency.'
(病院 ←=:Author)

5 Statistics of the Corpus and Discussion

1,000 documents have been annotated by three annotators. The statistics of the annotated corpus is listed in Table 4. More than half of the basic-phrases are annotated with some relations. The corpus includes various documents such as personal web sites, news articles, publicity pages of local governments, billing pages and recipe pages. There are some documents that cannot be categorized uniquely such as publicity blog articles from companies.

The number of the documents with respect to types of the author/reader annotations are shown in Table 5. "Explicit" means that an author or a reader is mentioned explicitly and annotated. "Implicit" means that an author or a reader is not mentioned explicitly but is referred from zero pronouns as zero exophora. The remaining documents fall into "No appearance." As a result, the author appeared in the discourse on about 63% of documents and the reader appreared on about 39%. The author/reader are sometimes not mentioned explicitly though the author/reader appear in the discourse.

Word	Frequency
私 (I)	63
弊社 (our company)	12
店 (shop)	10
会 (society)	10
当社 (our company)	9
自分 (self)	8
管理人 (moderator)	5
病院 (hospital)	3
主婦 (housewife)	2
カーブス (Curves)	1
こま (Koma)	1

Table 6: Example of the mentions of authors

Word	Frequency
皆様 (you all)	28
客 (customer)	24
あなた (you)	23
方 (gentleman/lady)	9
自分 (self)	8
人 (person)	7
自身 (self)	3
読者 (reader)	1
生徒 (student)	1
贈り主 (giver)	1
市民 (citizen)	1

Table 7: Examples of the mentions of readers

The author appeared in documents 356 times and the reader appeared 134 times. The examples and their frequency are shown in Table 6 and Table 7. Among words that mention the author, 私 (I) is the most frequent expression, which appeared 63 times, but there are various words such as the position names (管理人 (moderator), and 主婦 (housewife)), the words indicating organization (店 (shop) and 病院 (hospital)) and the proper names (こま (Koma) and カーブス (Curves)). Since there are 96 words which appeared once in the whole corpus and 24 words which appeared twice, many words

	Anaphora	Exophora	Total
GA	1703	2488	4191
WO	594	100	694
NI	409	388	797
GA2	72	116	188
Total	2778	3092	5870

Table 8: Number of zero anaphora/exophora

	Author	Reader	Others	Total
GA	602	176	925	1703
WO	8	4	582	594
NI	78	44	287	409
GA2	23	8	41	72
Total	711	232	1835	2778

Table 9: Breakdown of zero anaphora

become mentions of the author depending on the context. Among words that mention the reader, the frequency of 客 (customer) is the second most frequent word after 皆様 (you all). This is because many of the web pages assuming potential readers are business pages. There are the words assuming document-specific readers such as 生徒 (student), 贈り主 (giver) and 市民 (citizen). The words that are used for both author and reader includes 自分 (self).

The numbers of the annotated zero anaphora and zero exophora are shown in Table 8. In this Table, the zero anaphora/exophora occurred most frequently in GA (nominative) case and about 60% of them are zero exophora. There is not much difference between the total of the zero anaphora and the zero exophora between WO (accusative) case and NI (dative) case, but the ratio of the zero exophora of NI case is larger than that of WO case. The breakdown of the numbers of zero anaphora is shown in Table 9 and one of zero exophora is shown in Table 10. In Table 9, “Author” and “Reader” mean that the referent of zero anaphora has a coreference relation with the author and the reader. Table 9 and Table 10 indicate that the one third of the referents of GA case are the author and the one sixth is the reader. In contrast, the reader is more than the author for the referent of zero exophora in NI case. In WO case, there are few referents that refer to the author or the reader and about 80% of the referents of zero exophora is unspecified-person and unspecified-matter.

	Anaphora	Exophora	Total
=	2201	363	2564
NO	3185	201	3386
≃	757	43	800
Total	6143	607	6750

Table 11: Number of anaphoric/exophoric relations

	Author	Reader	Others	Total
=	100	29	2072	2201
NO	256	96	2833	3185
≃	31	24	702	757
Total	387	149	5607	6143

Table 12: Breakdown of anaphoric relations

The numbers of the annotated anaphoric and exophoric relations are shown in Table 11. The breakdowns are shown in Table 12 and Table 13. Table 11 indicates that most reference relations are anaphoric relations regardless of types. Since NO relations are more than ≃, more bridging references can be rephrased as the form “A の B.”

The inter-annotator agreements are shown in Table 14 and Table 15[7]. Only the agreement of coreference, is annotated by “=,” is calculated by the MUC score (Vilain et al., 1995). For the agreement of other cases, we show only representative cases and “Total” includes cases that are omitted from the table. In Table 15, although the agreements of GA and WO are very high, that of GA2 is very low. It is because that GA2-case sometimes can be rephrased to other cases. For example, since it is possible to rephrase Example(11) to both (12) and (13), there are two annotation candidates, (11a) and (11b). We had set up a criterion that a case marker other than GA2 to which the target expression can be paraphrased is preferred to GA2. However, the judgment on such paraphrasing was not consistent between the annotators. Similarity, the judgment on paraphrasing to NO (A の B) was not stable, and this instability was a cause of the low agreement of ≃.

(11) 魚は 高くて 買えない
Fish-TOP *too expensive cannot buy*
監督。
director.
‘Fish are too expensive for the director to buy.’

a. (買えない ←GA2:監督, GA:魚)
b. (買えない ←GA:監督, WO:魚)

[7] A, B and C indicate each annotator

	Author	Reader	Unspecified- Person	Unspecified- Matter	Unspecified- Situation	Total
GA	930	637	734	95	92	2488
WO	3	9	32	52	4	100
NI	66	153	140	27	2	388
GA2	43	44	25	4	0	116
Total	1042	843	931	178	98	3092

Table 10: Breakdown of zero exophora

	Author	Reader	Unspecified- Person	Unspecified- Matter	Unspecified- Situation	Total
=	258	105	0	0	0	363
NO	95	52	28	26	0	201
≃	16	18	4	5	0	43
Total	369	175	32	31	0	607

Table 13: Breakdown of exophoric relations

A vs. B	A vs. C	B vs. C
0.709	0.770	0.691

Table 14: Agreement of coreference relations

	A vs. B	A vs. C	B vs. C
GA	0.852	0.823	0.865
WO	0.890	0.822	0.848
NI	0.726	0.729	0.766
GA2	0.593	0.385	0.296
NO	0.690	0.610	0.558
≃	0.483	0.375	0.375
Total	0.764	0.724	0.738

Table 15: Agreement of annotation

(12) 監督が 魚が 買えない。
director-NOM *fish*-NOM cannot buy.

(13) 監督が 魚を 買えない。
director-NOM *fish*-ACC cannot buy.

6 Conclusion

In this paper, we described the details of the semantically annotated corpus that consists of various documents in the web. In this corpus, we annotated predicate-argument structures and anaphoric relations as semantic annotation. We focused on the mentions of the author and the reader in the documents and annotated these mentions. In order to reduce the workload of each document, we annotated only the first three sentences. As a result, we built an annotated corpus that consists of 1,000 documents. Our corpus analysis revealed that the author and the reader appeared in many of the documents, these are mentioned in various expressions and have an important role in zero anaphora and zero exophora.

References

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a japanese text corpus with predicate-argument and coreference relations. In *Proc. of the Linguistic Annotation Workshop*, pages 132–139.

Daisuke Kawahara, Sadao Kurohashi, and Koiti Hasida. 2002. Construction of a japanese relevance-tagged corpus. In *Proc. of LREC'02*.

Kyoko Ohara. 2011. Full text annotation with japanese framenet: Study to annotation semantic frame to bccwj(in japanese). In *Proc. of the 17th Annual Meeting fo the Association for Natural Language Processing*, pages 703–704.

L. Rello and I. Ilisei. 2009. A comparative study of spanish zero pronoun distribution. In *Proc. of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL)*, pages 209–214.

Kepa Joseba Rodríguez, Francesca Delogu, Yannick Versley, Egon W. Stemle, and Massimo Poesio. 2010. Anaphoric annotation of wikipedia and blogs in the live memories corpus. In *Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proc. of the 6th conference on Message understanding*, pages 45–52.