

Title	Studies on Annotated Diverse Corpus Construction and Zero Reference Resolution in Japanese(Dissertation_全文)
Author(s)	Hangyo, Masatsugu
Citation	Kyoto University (京都大学)
Issue Date	2014-03-24
URL	http://dx.doi.org/10.14989/doctor.k18407
Right	
Type	Thesis or Dissertation
Textversion	ETD

**Studies on Annotated Diverse Corpus
Construction and Zero Reference Resolution
in Japanese**

Masatsugu Hangyo

Abstract

As the Web becomes more and more popular, Natural Language Processing (NLP) applications, such as search engines and machine translation, are being used more widely. Reliable fundamental NLP analyses are essential to improve accuracy of these NLP applications. One of analyses that have not yet achieved sufficient accuracy is zero reference resolution, which aims to detect and reconstruct omitted arguments of a predicate. By using the results of zero reference resolution, NLP applications can capture hidden relations between arguments and the predicate. Since zero references occur frequently in Japanese, zero reference resolution is a very important process in NLP applications.

Although the use of the Web has become widespread and many NLP applications are applied to Web documents, most of the previous zero reference resolution studies have focused mainly on newspaper articles. In contrast to newspaper articles, a wide variety of topics and writing styles are included in Web documents, while some linguistic phenomena with high correlations to zero references do not appear in Web documents. To apply zero reference resolution to Web documents, it is important to focus on the differences between newspaper articles and Web documents.

In this study, we consider the author and reader of a document as one of these differences. The author and reader of a newspaper are, respectively, limited to journalists and subscribers to the newspaper, and since the aim of a newspaper is to objectively report events, in which neither the journalists nor subscribers actually participate to a great extent, the author and reader hardly appear in the discourse of newspaper articles. On the other hand, Web documents are written by various authors for a wide variety of readers, and since the author describes

him/herself and reaches out to the reader in some documents, the author and reader may often appear in a discourse. Since the author and reader tend to be omitted and some linguistic phenomena, such as modality expressions and honorific expressions, are good clues for zero reference resolution about the author and reader, it is important to deal with the author and reader of a document in zero reference resolution. The author and reader frequently appear as referents of zero exophora, which is the phenomenon whereby the referent does not explicitly appear in the document. Although most previous studies ignore this phenomenon, it is essential to consider zero exophora for dealing with the author and reader.

First, we construct an annotated corpus consisting of Web documents. For annotation, we analyze various annotation issues for Web documents in terms of the author and reader of the documents. The first issue is the existence of expressions that refer to the author and reader of a document. Since these expressions are important to understand the discourse, we refer to such expressions as author and reader mentions and define criteria for them. The second issue is the ambiguity of predicate arguments. Some of the arguments of a predicate can be interpreted as either the author, reader, or an indefinite person. We classify the ambiguity expressions and define annotation criteria for them. As a result, we construct a corpus comprising 1,000 Web documents and which is annotated with semantic relations including zero reference relations.

Then, we propose a zero reference resolution model that considers zero exophora and author and reader mentions. First, the proposed model automatically detects author and reader mentions using lexico-syntactic patterns. Then, our model resolves zero references as part of predicate-argument structure analysis. The model uses information about author and reader mentions and handles zero exophora by setting pseudo entities corresponding to the author, reader, and indefinite pronouns. Experimental results show that our model is more effective than the baseline model, which does not consider the zero exophora or author and reader mentions.

Acknowledgments

先輩方に倣って、謝辞だけは日本語で書きます。

本研究を進めるにあたり、終始熱心にご指導くださいました黒橋禎夫教授に感謝いたします。先生は、いつも些細なミスばかりしている私をいつも辛抱強く指導してくださり、最後まで大変お世話になりました。

河原達也教授と西田豊明教授には、論文調査委員を引き受けていただき、有益な助言をくださったことに感謝いたします。

河原大輔准教授には研究だけでなく、英語の論文の添削などについても熱心に指導していただきました。先生のご助力なくして、本研究を進めることはできなかったと思います。

柴田知秀助教、科学技術振興機構研究員の中澤敏明さん、九州大学の村脇有吾助教、東京工業大学の笹野遼平助教に感謝いたします。先生方からは、自然言語処理に関する基礎的なことから、研究に取り組む姿勢など様々なことを教わりました。また有意義なご意見をたくさんいただきました。

クックパッド株式会社の原島純さんには、研究に関するだけでなく、日々の様々なことに相談に乗っていただきました。一年上の身近な先輩として、長い間にわたって研究室での生活において様々な面で支えていただき、本当にありがとうございました。

他にも研究室の多くの方々にご支援いただきました。皆様との議論や研究内容には日々刺激を受け、自分の研究を進める励みにもなりました。秘書の芦原裕子さんには煩雑な事務処理を円滑に進めていただきました。皆様に深く感謝いたします。

また、石川真奈見さん、二階堂奈月さん、堀内マリ香さんにはコーパス作成の作業を行なっていただいたことに感謝します。皆様との作業内容についての議論は大変有意義なもので、研究のアイデアの多くはこの議論を通じて生まれました。皆様に作成していただいたコーパスのおかげで本研究は完成できたと言っても過言で

はないと思います。

最後に、今まで支えてくれた家族に感謝して謝辞を終えたいと思います。

Contents

Abstract	i
Acknowledgments	iii
1 Introduction	1
1.1 Background	1
1.2 The Author and Reader in Japanese Zero References	4
1.2.1 Author/Reader Mentions	4
1.2.2 Zero Exophora	5
1.3 Annotated Corpus	5
1.3.1 History of Corpus Construction	6
1.3.2 Contributions of this Study	7
1.4 Zero Reference Resolution	8
1.4.1 Previous Approaches to Zero Reference Resolution	8
1.4.2 Contributions of this Study	9
1.5 Outline of this Thesis	10
2 Building a Diverse Document Leads Corpus	11
2.1 Corpus Annotated with Semantic Relations	12
2.2 Related Work	16
2.3 Annotation Target Document	18
2.3.1 Detecting Documents That Cannot Be Understood Semantic Relation With Only Raw Text	20
2.3.2 Determination of Inadequate Document	24

2.4	Annotation Criteria	25
2.4.1	Types of Annotation	25
2.4.2	Mentions of Author and Reader	30
2.4.3	Author Mention	32
2.4.4	Reader Mention	34
2.4.5	Criteria for Ambiguous Annotation	35
2.4.6	Criteria of Annotating [US-person]	35
2.4.7	Criteria of Annotating [author]	36
2.4.8	Criteria of Annotating [reader]	37
2.5	Constructed Corpus	38
2.5.1	Procedure and Setting of Annotation	39
2.5.2	Statistic of DDLC	39
2.5.3	Author/Reader Mention	42
2.5.4	Zero Reference Relation	43
2.5.5	Inter-Annotator Agreement	47
2.6	Summary of this Chapter	56
3	Author/Reader Mention Detection	57
3.1	Author/Reader Mention Detection	57
3.2	Author/Reader Detection Model	59
3.2.1	Discourse Entity	59
3.2.2	Ranking Model	60
3.2.3	Lexico-Syntactic Patterns	65
3.3	The result of Author/Reader Mention Detection	69
3.3.1	Experimental Setting	69
3.3.2	Results of Author/Reader Mention Detection	69
3.4	Summary of this Chapter	74
4	Zero Reference Resolution Model	75
4.1	Zero Reference Resolution	76
4.2	Related Work	80
4.3	Baseline Model	82
4.3.1	Feature Representation of Predicate-Argument Structure	85

4.3.2	Weight Learning	89
4.4	Proposed Zero Reference Resolution Model	92
4.4.1	Pseudo Entities and Author/Reader Mentions for Zero Ex- ophora	93
4.4.2	Feature Representation of Predicate Argument Structure	95
4.4.3	Author/Reader Mention Score	97
4.5	Experiments	97
4.5.1	Experimental Settings	97
4.5.2	Results of Zero Reference Resolution	98
4.6	Summary of This Chapter	105
5	Conclusion	107
5.1	Summary of this Research	107
5.2	Future Work	110
	Bibliography	112
	List of Publications	119

List of Figures

2.1	Example of a document whose headline do not appear in the body	20
2.2	Example of a document that the elements of its headline appear in the first three sentences	21
2.3	Example of a document which cannot be understood without its headline	22
3.1	Example of a document in which an author mention appears . . .	61
3.2	Example of a document in whose discourse an author do not appear	62
3.3	Example of a document in whose discourse an author appears but an author mention do not appear	63
3.4	Example of error of the author mention detection (1)	71
3.5	Example of error of the author mention detection (2)	72
3.6	Example of error of the reader mention detection (1)	73
4.1	Outline of zero reference resolution	83
4.2	Candidate predicate-argument structures of “紹介します” in the baseline model	86
4.3	Example of case that one case slot is assigned to multiple arguments	91
4.4	Candidate predicate-argument structures of “紹介します” in the proposed model	94
4.5	Improvement example (1)	100
4.6	Improvement example (2)	101
4.7	Example of error of the proposed model (1)	103
4.8	Example of error of the proposed model (2)	104

List of Tables

2.1	Distances between referent and zero pronoun	18
2.2	Examples of stop phrase	25
2.3	Candidate referents of zero exophora	27
2.4	The types of named entity	30
2.5	Statistics of corpus	40
2.6	Ratio of sentences in which a modality expression appear	41
2.7	Ratio of sentences in which a honorific expression appear	41
2.8	Result of manually classification of document types	42
2.9	Appearance of the author/reader in a document	43
2.10	Examples of the author mentions (excerpt)	44
2.11	Examples of the reader mentions (excerpt)	44
2.12	Number of zero references in DDLC	45
2.13	Breakdown of the numbers of zero endophora in DDLC	45
2.14	Breakdown of the numbers of zero exophora in DDLC	46
2.15	Number of zero references in KUTC	47
2.16	Breakdown of the numbers of zero exophora in KUTC	48
2.17	Number of arguments that have multiple interpretations	49
2.18	Inter-annotator agreement of the author/reader mentions	49
2.19	Agreement of predicate-argument structures for predicates	51
2.20	Agreement of predicate-argument structures for verbal nouns	51
3.1	Examples of author/reader mentions	59
3.2	Generalization type and criteria	67
3.3	First person pronoun and second person pronoun	68

3.4	Result of the author mention detection	69
3.5	Result of the reader mention detection	69
4.1	Examples of zero endophora, zero exophora and no zero reference.	77
4.2	The features for a case that is assigned to a distances entity	87
4.3	The features for a case that is not assigned to any discourse entities	89
4.4	Expressions and categories for pseudo entities	95
4.5	Results of zero endophora resolution	98
4.6	Results of zero reference resolution	99
4.7	Results of easing evaluation	102

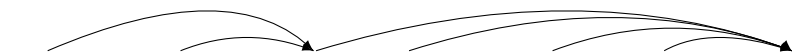
Chapter 1

Introduction

1.1 Background

In recent years, as the use of the Web has become more widespread, Natural Language Processing (NLP) has been used in a variety of situations and applications. For example, search engines are essential for efficient use of the Web, and many people use machine translation for reading foreign language Web pages. Improved accuracy of the fundamental NLP analyses leads to an improvement in the accuracy of various NLP applications. As fundamental analyses, morphological analysis, syntactic parsing, coreference resolution, and named entity recognition have achieved high accuracy. On the other hand, zero reference resolution is one of the analyses that has insufficient accuracy. Zero reference resolution is the process of reconstructing omitted arguments of a predicate.¹

(1.1) $(\phi$ ガ) パスタが 好きで 毎日 $(\phi$ ガ) $(\phi$ ヲ) 食べます。
(ϕ -NOM) pasta-NOM like everyday (ϕ -NOM) (ϕ -ACC) eat
‘Since (ϕ) likes pasta, (ϕ) eats (ϕ) every day’



For example, in Example (1.1), the topical (second nominative) argument of the predicate “好き” (like) and the nominative and accusative arguments of the predi-

¹In 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).

cate “食べます” (eat) are omitted. Zero reference resolution detects these omitted arguments and identifies the referents thereof. In this case, the referent of the accusative argument of “食べます” is “パスタ” (pasta) and the referent of the topical argument of “好き” and the nominative argument of “食べます” is the author of the text, who is not explicitly mentioned in the text. To understand text and NLP applications, it is important to organize the text in a structured representation such as a predicate-argument structure. The predicate-argument structure, which is expressed as a predicate, its arguments, and the relations between the predicate and arguments, is a minimum structure for representing an event. For example, in Example (1.1), the predicate-argument structure of “食べます” is “predicate: 食べます, NOM:[author], ACC:パスタ, TIME:毎日,” showing an agent, an object, and the time of the event. Zero reference resolution plays a very important role in capturing the predicate-argument structure. If the text has not been analyzed by zero reference resolution, only elements that have a direct dependency relation to a predicate are treated as arguments. For example, in the above example, only the TIME relation can be recognized when considering elements with a direct dependency relation only. By using the results of zero reference resolution, the missing arguments, such as the nominative and accusative arguments, can be filled in. In Japanese, since ellipsis is frequently used, zero reference resolution is very important to understand the text and NLP applications.

Zero references are categorized as either **zero endophora** or **zero exophora**. Zero endophora is the phenomenon where the referent of an omitted argument is mentioned in the document (e.g., the omission of “パスタ” in Example (1.1)). On the other hand, zero exophora is the phenomenon where the referent is not explicitly mentioned in the document (e.g., the omission of the author in Example (1.1)). In Japanese, when the referent is the author or reader of a document or an indefinite pronoun, zero exophora frequently occurs. Previous zero reference resolution studies have focused mainly on zero endophora and ignore zero exophora, as if the zero pronoun does not exist. By treating zero exophora, zero pronoun occurrences can be captured even when the referent is not explicitly mentioned in the document. It is important to deal with the zero exophora for zero referent resolution.

As use of the Web has become more widespread, communication via the Web has increased among users, and the number of available documents on the Web has grown. Therefore, many NLP applications focus on Web documents. However, in the past, studies of Japanese zero reference resolution have concentrated mainly on newspaper articles. Topics in newspaper articles are limited, and writing styles are mostly consistent. However, a great variety of topics, most of which are not treated in newspaper articles, appear in Web documents, and the writing styles are also varied. Many linguistic phenomena on the Web do not appear in newspaper articles, with some of these phenomena having high correlations with zero reference resolution. Therefore, it is difficult to simply apply a zero reference resolution system based on newspaper articles to Web documents, and it is important to study zero references focusing on the differences between newspaper articles and Web documents.

One of the differences between newspaper articles and Web documents is the existence of an author and reader of a document. Since the aim of a newspaper article is for the author (journalist) to objectively report events, most of which the author and the reader do not directly participate in, to the reader (subscriber), the author and reader hardly ever appear in the discourse of a document. On the other hand, in Web documents, since the author often describes him/herself and reaches out to the reader, the author and reader often appear in the discourse of a document. For example, in blog articles and corporate advertising sites, the author often describes events that have occurred in his/her life or activities of the corporation, while on online shopping sites, the author encourages the reader to buy commercial products. The author and reader behave characteristically in the discourse. For example, they tend to be omitted and there are linguistic phenomena with high correlations to the author and reader such as modality and honorific expressions. Since these behaviors are good clues for zero reference resolution, we propose a zero reference resolution system that focuses specifically on the author and reader of a document.

1.2 The Author and Reader in Japanese Zero References

Here we explain appearances of the author and reader. The author and reader are sometimes mentioned as personal pronouns or other expressions (e.g., “私” (I), “あなた” (you), or the name of the author). We discuss these expressions in Section 1.2.1. On the other hand, even if the author and reader are not explicitly mentioned in a document, they often have a role in the discourse. In this case, the author and reader are treated as referents of zero exophora. We explain the zero exophora in Section 1.2.2.

1.2.1 Author/Reader Mentions

The author and reader are often explicitly mentioned in a document using expressions such as personal pronouns. For example, in Example (1.2), “僕” (I) corresponds to the author and “皆さん” (you all) corresponds to the reader.

- (1.2) 僕_{author} は 京都に (僕ガ) 行こうと 思っています。
 I-TOP Kyoto-DET (I-NOM) would go thought

‘I thought I would go to Kyoto.’

皆さん_{reader} は どこに 行きたいか (皆さんガ) (僕ニ)
 you all-TOP where-DET want to go (you all-NOM) (I-DAT)

教えてください。
 let me know

‘Please let me know where you want to go.’

We call the expressions corresponding to the author and reader **author mentions** and **reader mentions**, respectively. Author and reader mentions have strong relations to some linguistic expressions such as request forms and honorific expressions. For example, in Example (1.2), the fact that the nominative case and dative case of “教えてください” (let me know), which are “皆さん” (you all) and “私” (I), respectively, have relations to “教えてください”, which is a request form. Therefore, author and reader mentions are very important clues for zero

reference resolution. In newspaper articles, even when the author and reader are explicitly mentioned, author and reader mentions are limited to a few expressions owing to the consistent writing style and can easily be detected from the lexical information. On the other hand, in Japanese Web documents, a large number of expressions can be used as author and reader mentions for the following reasons. In Japanese, personal pronouns are seldom used and the author and reader are often mentioned by name or through their role as the author and reader. Additionally, since Web documents are written by a variety of authors for various readers, names and role expressions referring to the author or reader vary greatly. Therefore, author and reader mentions cannot be easily detected from the lexical information, and other information such as syntactic and contextual information is needed to assist in their detection.

1.2.2 Zero Exophora

When all appearances of the author and reader are omitted and there are no author or reader mentions, the author and reader appear as referents of zero exophora (e.g., the omission of the author in Example (1.1)). In Japanese, since the author and reader tend to be omitted, zero exophora of the author and reader occurs frequently. In a Web corpus [6], about half the zero references are zero exophora and many of these are omissions of the author or reader. Just like author and reader mentions, the author and reader in zero exophora are related to various linguistic expressions and are essential for contextual understanding. Even when the author and reader are not explicitly mentioned, it is particularly important to deal with the zero exophora to handle the author and/or reader of a document.

1.3 Annotated Corpus

In recent years, most NLP studies have used an annotated corpus, which is a collection of documents that have been manually annotated with various pieces of information. An annotated corpus consisting of target domain documents and annotated with the gold-standard of a task is important from the following two points of view.

Evaluation experiment We can evaluate a system by comparing the annotations of the corpus with the outputs of the system. Additionally, in the evaluation of a machine learning based system, the annotated corpus can also be used as training data.

Problem analysis By analyzing the annotated corpus, various phenomena can be analyzed qualitatively and quantitatively. The results of the analysis serve to improve the system.

Since we study zero reference resolution focused on the Web, it is necessary to build a Web document corpus annotated with the zero reference information. In this section, we briefly describe the history of constructing the corpus and the contributions of the final corpus.

1.3.1 History of Corpus Construction

Research on corpus construction has a long history. It is said that the first extensive published corpus was the “Brown University Corpus of Present-Day American English.” Here we briefly present the subsequent history.

Before 1980 Raw corpora, which were not annotated with any information, were constructed (e.g., the Brown Corpus [3] and LOB Corpus [20]).

1980s Annotations of parts-of-speech and analysis of language usage using the annotations were started (e.g., the Tagged Brown Corpus [4]).

1990s The first large-scale corpus annotated with syntactic structure information (Penn-Treebank [29]) was published. Annotations of richer information (e.g., semantic and inter-sentential relations) began in the late 1990s. For example, annotations of semantic roles [2] were included and the first corpus annotated with coreferential information was constructed for shared tasks [1]. A large-scale annotated corpus based on Japanese newspaper articles (Kyoto University Text Corpus [27]) was published in the late 1990s.

2000s A corpus annotated with semantic roles [36] and another corpus annotated with coreferential information [34] were constructed. As one of the inter-sentential relations, the discourse relation was included in annotating the

Penn-Treebank by Miltsakaki et al. in 2004 [31]. Recently, an approach including many types of information as annotations in a corpus has been tackled by Hovy et al. [12].

The predicate-argument structure and coreferential relation in Japanese were annotated in the Kyoto University Text Corpus (e.g., GDA Corpus [8], Kyoto University Text Corpus version 4.0 [24] and NAIST Text Corpus [15]). Corpora based on various documents other than newspaper articles have also been constructed. For example, an annotated blog corpus was published by Hashimoto et al. in 2011 [7], while the Balanced Corpus of Contemporary Written Japanese (BCCWJ) was published by Maekawa in 2008 [28]. BCCWJ has been annotated with a FrameNet structure [35], predicate-argument structure, and coreferential relations [25].

1.3.2 Contributions of this Study

Most existing large-scale Japanese annotated corpora have been based on newspaper articles. In this study, by collecting documents from the Web, we construct a corpus that includes various expressions, many of which do not appear in newspaper articles. Limiting the annotation target to the first few sentences of each document improves the work efficiency of the annotation and contributes to the diversity of documents.

We analyze annotation issues that are not of concern in annotations for newspaper articles and define annotation criteria for the following two issues. The first issue is the existence of expressions referring to the author or reader of a document. Since these expressions, which we call **author mentions** and **reader mentions**, respectively, behave differently from other elements in the discourse, we define annotation criteria for them. The second issue is the ambiguity of zero reference annotations. In the annotation of zero reference relations, some arguments of predicates can be interpreted with multiple referents, and these ambiguities cause inconsistent annotations. In this study, we categorize ambiguous expressions and define the annotation criteria for them.

Finally, we have built an annotated corpus consisting of 1,000 documents and annotated with more than 8,000 zero reference tags.

1.4 Zero Reference Resolution

In Japanese, zero reference resolution has been widely studied with most of these studies having focused on zero endophora. This study address two issues related to dealing with the author and reader of a document: zero exophora and author and reader mentions. In this section, we briefly describe some previous studies on zero reference resolution and the contributions of our proposed model.

1.4.1 Previous Approaches to Zero Reference Resolution

Zero reference resolution can be divided into two subtasks. The first subtask is zero pronoun detection, which involves detecting the omitted argument, which is called a zero pronoun. Here we present various approaches for zero pronoun detection used in previous works.

Case frames Manually constructed case frames [17], describing what arguments a predicate has, were used for zero pronoun detection by Murata et al. in 1997 [32] and later by Seki et al. in 2002 [44]. Automatically constructed case frames [22] were used by Sasano et al. in 2008 and 2011 [42, 43].

Co-occurrence frequency The co-occurrence frequency between a candidate argument and a predicate was used for zero pronoun detection by Imamura et al. [16] and later by Hayashibe et al. [10].

The second subtask is referent identification, which is the task of identifying the referent of a zero pronoun. Here we present the various approaches for referent identification used in previous works.

Contextual and syntactic information Centering theory was proposed by Kameyama in 1986 [21]. Murata et al. in 1997 [32] proposed a rule-based referent identification model using centering theory and other syntactic and contextual rules. In recent years, these rules have been used as features in a machine learning based model [43, 16, 18]. Iida et al. [13] proposed a model that automatically acquires syntactic rules in 2006, and another model that automatically recognizes contextual rules in 2009 [14].

Selectional preference In referent identification, manually constructed case frames have been used for selectional preference by Murata et al. in 1997 [32] and Seki et al. in 2002 [44]. As with manually constructed case frames, automatically constructed case frames have also been used for selectional preference. In automatically constructed case frames, a strength of preference is defined and can be used in a probabilistic model [42] and as features in a machine learning based model [43].

1.4.2 Contributions of this Study

Automatic detection of author and reader mentions If the author and reader are explicitly mentioned in a document, it is important to deal with them differently to other discourse elements because the author and reader tend to be omitted and there are many clues for referent identification of the author and reader, such as honorific expressions and modality expressions. However, since the author and reader are mentioned in a variety of expressions in Web documents, it is difficult to detect which expressions correspond to the author and reader using only lexical information. In this study, we propose a machine learning based method that automatically detects author and reader mentions using lexico-syntactic patterns.

Treating zero exophora Most previous studies have ignored zero exophora by assuming zero pronouns do not exist in a sentence. However, such a rough approximation has impeded zero reference resolution research. Therefore, in this work, to deal with the zero exophora explicitly, we provide pseudo entities corresponding to the author, reader, and indefinite pronouns as candidate referents of zero pronouns. By dealing with the zero exophora, the existence of zero pronouns corresponds with the valency of a predicate and it is expected to improve the accuracy of machine learning based zero pronoun detection. Additionally, since it is important to capture common characteristics between the author and reader in zero exophora and the author and reader mentions, our model represents the fact that author and reader mentions have features of the author and reader in the zero exophora. Since it is useful to know that the agent of an event is the author in

the analysis of Web documents such as blog articles, treating an exophoric author is useful in such analysis.

1.5 Outline of this Thesis

The rest of this thesis is organized as follows. In Chapter 2, we describe our work on building a diverse document corpus. We first explain our annotation targets and how to collect target documents, and then present annotation criteria for Web documents. Finally, we report statistics of the constructed corpus and discuss properties of the corpus.

In Chapter 3, we present a method for automatic detection of author and reader mentions. First, we discuss the characteristics of author and reader mentions. Thereafter, we present a ranking model that trains the decision function for detecting author and reader mentions and features that are used in the ranking model. Finally, we report the results of experiments for detecting author and reader mentions.

In Chapter 4, we describe the proposed zero reference resolution model, which considers exophora and author and reader mentions. We first explain the baseline model, which considers only zero endophora, and then we present the proposed model. We discuss the experimental results, which show the effectiveness of our method.

In Chapter 5, we summarize this study and suggest areas for future work.

Chapter 2

Building a Diverse Document Leads Corpus

In this chapter, we tackle construction of annotated corpus. The annotated corpus is necessary for problem analysis and system evaluation, and it is important to build an annotated corpus that consists of target domain documents and is annotated with gold-standard of a task. Since, as described as Chapter 1, we focus on zero reference resolution for Web documents, it is necessary to construct a corpus that consists of the Web documents and annotate the corpus with various information about the zero reference information. In this chapter, we present documents that construct our corpus and annotation criteria.

The rest of this chapter is organized as follows. In Section 2.1, we explain outline of annotation information and sort issues about annotation for Web documents. In Section 2.2, we describe related works about corpus construction. In Section 2.3, we present annotation target of our corpus and how to collect the documents. In Section 2.4, we explain types of annotation and them criteria. In Section 2.5, we show statistics of the constructed corpus and discuss its properties. In Section 2.6, we present conclusion of this chapter.

2.1 Corpus Annotated with Semantic Relations

In recent years, semantic analysis has been studied as a subsequent task of syntactic parsing. There are various tasks in semantic analysis, but predicate-argument structure analysis and endophoric resolution, which are tasks clarifying the relationships between elements in a document, are the most fundamental and important tasks. In this research, we refer to these tasks as semantic relation analysis. Predicate-argument structure analysis reveals relationships between a predicate and its arguments and deals with relations that are deeper than surface dependency relations. Endophora resolution defines relationships between the expressions in a document and deals with relations between expressions that do not have dependency relations. In research on semantic relation analysis, a corpus that is manually annotated with semantic relations is necessary for evaluation and analysis.

We illustrate the semantic relations and annotations in Example (2.1), where “A ← *rel*:B” represents annotating A with B using relation *rel*. In the following examples, we sometimes omit annotations that are not related to the discussion.

(2.1) 今日は ソフマップに 行きました。

Today-TOP Sofmap-DAT went.

‘Today, I went to Sofmap.’

(行きました ← ガ:[author], ニ:ソフマップ京都)

時計を 買いたかったのですが、この 店舗は

watch-ACC want to buy this shop

扱っていませんでした。

does not deal in

‘I wanted a watch but this shop does not deal in watches.’

(買いたかった ← ガ:[author], ヲ:時計
店舗 ←=:ソフマップ京都
扱っていませんでした ← ガ:店舗, ヲ:時計)

時計を 売っている お店を コメントで 教えてください。

watch-ACC buy shop-ACC comment-INS let me know

‘Please let me know which shop sells watches.’

$$\left(\begin{array}{l} \text{時計} \leftarrow =:\text{時計} \\ \text{売っている} \leftarrow \text{ガ:お店, ヲ:時計} \\ \text{教えてください} \leftarrow \text{ガ:[author], ヲ:お店, ニ:[reader]} \end{array} \right)$$

Endophora is the phenomenon whereby an expression in the text refers to other expressions (referent). In the second sentence in Example (2.1), “店舗” (shop) refers to “ソフマップ” (Sofmap) in the first sentence. We represent the endophoric relation by annotating “店舗” with “=:ソフマップ.” The predicate-argument structure represents relations between a predicate and its arguments, and in Example (2.1), the argument of the ガ (nominative) case of “扱っていませんでした” (does not deal in) is “店舗” while the argument of the ヲ (accusative) case of “扱っていませんでした” is “時計” (watch). In this example, the explicit case marker of 店舗 is the topic marker, which hides the actual case relation between “店舗” and “扱っていませんでした.” In this example, the argument of the ヲ case, “時計,” is omitted and this omission is called a *zero reference*. In our research, we deal with a zero reference as part of the predicate-argument structure. In addition, in Japanese, zero exophora, which is the phenomenon whereby the referent of a zero pronoun is not mentioned in the document, occurs often. In Example (2.1), the argument of ガ cases of “行きました” (went) and “買いたかった” (want to buy) is the author of this document, although there are no expressions referring to the author in the document. By setting [author], [reader], [US (unspecified)-person], and others as referents of the exophora, we can annotate the predicate-argument structure including zero exophora.

In the past, annotated corpora used for Japanese semantic relation analysis were based on newspaper articles. However, there are a variety of sources other than newspaper articles, such as encyclopedias, diaries, and novels with diverse writing styles in each genre. There are linguistic phenomena that do not appear in newspaper articles such as requests and honorific expressions, and these phenomena have high correlations with semantic relations. For example, in Example (2.1), the relation that the argument of the ガ case of “買いたかった” is [author], is related to an intention expression, and relations that the argument of the ガ case of “教えてください” (let me know) is [reader], and the argument of the ニ

case is [author], are related to a request expression. Building an annotated corpus consisting of various texts and then analyzing the corpus are necessary to reveal relations between such linguistic phenomena and semantic relations. In this research, we used Web pages including news articles, blog articles, encyclopedias, business pages, and others as targets of the annotation and constructed a corpus with the various genres and writing styles annotated with semantic relations.

As explained above, phenomena that hardly appear in newspaper articles, are the annotation targets of our research. An occurrence of the author and reader in a document is one of the most important of these phenomena. Since the author and reader tend to be omitted and deeply involve modality and honorific expressions, they behave differently from other discourse elements. Since most of the content of a newspaper article consists of reporting objective facts, the author/reader of the document hardly appears in the discourse of a document with the exception of editorials. Therefore, although existing annotation bases define [author], [reader], and others as referents of exophora, specific criteria are not really discussed. On the other hand, Web pages, which are the targets of our research, contain many documents in which the author/reader appears in the discourse such as blog articles and manuals, and there are linguistic phenomena and semantic relations in these documents that cannot be assumed using existing annotation criteria. For this reason, it is important to analyze the problem of annotating documents in which the author/reader appears and to set annotation criteria.

The first problem of annotating documents in which the author/reader appears, is expressions corresponding to the author/reader in the document.

- (2.2) 僕は 京都に 行きたいのですが, 皆さんの お勧めの
 I-TOP Kyoto-DAT want to go, you all-GEN recommended
場所が あったら 教えてください。
 place-NOM there is let me know

‘I want to go to Kyoto, please let me know if there is a recommended place.’

$$\left(\begin{array}{l} \text{僕} \leftarrow =: [\text{著者}] \\ \text{皆さん} \leftarrow =: [\text{読者}] \\ \text{教えてください} \leftarrow \text{ガ:皆さん, ヲ:場所, ニ:僕} \end{array} \right)$$

In example (2.2), “僕” (I) is an expression that corresponds to author while “皆さん” (you all) corresponds to the readers. In this research, we call such expressions corresponding to the author/reader **author mentions** and **reader mentions**, respectively. Author/reader mentions behave in the same way as [author] and [reader] in the zero exophora. For example, in “教えてください” (let me know) in Example (2.2), the agent of the request expression tends to be the reader expression while the recipient of the request expression tends to be the author expression. Since authors and readers of documents that are the targets of our research, comprise various people and the author/reader is mentioned in the documents using a variety of expressions other than personal pronouns, author/reader mentions cannot easily be detected from lexical information. In this research, we annotate the author/reader mentions and the research behavior of the author/reader in a discourse.

The second problem is predicate-argument structure annotation of expressions in which the arguments are ambiguous. When describing a common occurrence in Japanese, expressions that do not clearly demonstrate an agent or a recipient are commonly used. In the annotation of newspaper articles, these expressions are annotated according to the criterion that the agent or recipient is an [US-person]. On the other hand, when the author/reader appears in the discourse, in the case of describing a common occurrence, the agent and others are often interpreted as the author/reader as well.

(2.3) ブログに 記事を 書き込んで、インターネット上で 公開する のは
 blog-DAT article-ACC post, on the Internet-LOC publish

とても 簡単です。

very easy

‘It is very easy to post blog articles and publish on the Internet.’

(公開する ← ガ:[author] ? [reader] ? [US-person], ヲ:記事)

In Example (2.3), the agent of “公開する” (publish) can be interpreted as an [US-person], because this sentence expresses a common belief. However, the sentence

can also be interpreted as an experience of the author or an act that the reader is going to do in the future. Such ambiguities cause inconsistent annotations depending on the interpretation of the annotators. In this research, we categorize ambiguous expressions and set criteria for the annotation thereof.

To deal with texts that include the above phenomena, it is important to build an annotated corpus, which includes documents from diverse domains. Web pages include various genres and text styles such as news articles, encyclopedia articles, blogs, and business pages. Using Web pages as the target documents of the annotation, we constructed a Japanese annotated corpus consisting of various genres. In contrast, since annotating semantic relations deals with inter-sentence relations, the number of elements that annotators should consider increases combinatorially. Therefore, if we wanted to annotate entire documents, the processing time for each document would increase and only a few documents would be annotated. Since our target is building a corpus that consists of a variety of documents, we confine the annotation target to the first few sentences. Since some semantic relations analysis systems use the results of previously analyzed sentences, analysis errors propagate to the subsequent analyses. By building a corpus that consists of document leads, we expect to improve analysis accuracy of both the document leads and the document as a whole.

2.2 Related Work

Existing corpora which are annotated with predicate-argument structures and endophoric relations include the Kyoto University Text Corpus [24] and the Naist Text Corpus [15]. These corpora are based on Mainich Newspaper articles from 1995 and annotated with predicate-arguments structure and endophoric relations. Since there are only reports and editorial articles in the newspaper, the writing styles are consistent, making it not possible to adapt a semantic analysis system based on this corpus to texts other than newspaper articles.

Corpora which consist of documents from various genres include the Balanced Corpus of Contemporary Written Japanese (BCCWJ)¹. BCCWJ includes publi-

¹<http://www.tokuteicorpus.jp/>

cations such as books and magazines and text from the Internet. BCCWJ has physical documents from various genres but Internet text is restricted to blogs and forums. For this reason, the company pages and other pages exist on the Internet, but not included.

Ohara [35] annotated predicate-argument structures defined in FrameNet to the predicates in BCCWJ. 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 endophoric relations are not annotated, they do not annotate the inter-sentence semantic relations. Komachi and Iida [25] have annotated predicate-argument structure and coreferential relation to BCCWJ in the same manner of NAIST Text Corpus. They applied annotation criteria for newspaper articles to BCCWJ and did not discuss differences between the newspaper articles and other document types.

In other languages, corpora dealing with multiple genres include Z-corpus [39] and LMC (Live Memories Corpus) [40]. Z-corpus consists of Spanish law books, textbooks and encyclopedia articles, and they are annotated with zero endophoric relations. They only treat zero endophora and do not treat endophora and predicate-argument structures. This is because the zero endophoric 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 endophoric relations. They deal with zero endophora as part of endophora, but do not deal with predicate-argument structures. Since pronoun-dropping only occurs in subject also in Italian, they regard the predicates which contain pronoun-dropping as endophoric expressions.

In English, some corpus annotated predicate-argument structure that treats arguments that do not have direct dependency relations to a predicate. NomBank [30] is annotated the predicate-argument structure to verbal nouns, but inter-sentential arguments are not annotated. Gerber and Chai [5] annotated inter-sentential arguments for 10 verbal nouns in NomBank. In SemEval-2010 [41], predicate-argument structure including inter-sentential arguments and null instantiations are annotated to novel text.

Table 2.1: Distances between referent and zero pronoun

	0	1	2	3	4	5 ~
Kyoto University Text Corpus	45.9%	21.7%	9.9%	5.7%	3.7%	13.1%
Web corpus of Sasano et al.	49.1%	27.7%	11.4%	5.5%	2.7%	3.6%

2.3 Annotation Target Document

Most existing Japanese corpora annotated with semantic relations consist of newspaper articles [15, 24]. However, there are linguistic phenomena which rarely occur in newspaper articles, and so we need to target various documents in order to study these phenomena. Using the Web without limiting by domain, we collect various documents. For building the annotated corpus consisting of various documents, we need to reduce the workload of each document. Therefore annotating targets are limited to the first three sentences of the document leads. 1,000 documents have been presently annotated.

The following is reason why we extracted first “three” sentences. We particularly focus on zero reference relation in semantic relations. We show locations of referents in Kyoto University Text Corpus and a Web corpus which is used by Sasano et al. [43] in Table 2.1. From this result, since about 70 % of zero reference relations appear within 1 sentence and about 80 % of zero reference relations appear within 2 sentences, we can collect various phenomena about the zero reference relation by dealing with first three sentences.

There are many inadequate documents, which should not be included in the corpus, in the web documents. The inadequate documents are classified roughly into 2 types. The first type is a document that cannot be understood from only a raw text such as a document requires specific information to Web (e.g., information of HTML or whole of site) to understand the discourse of the document. We describe treating such document in Section 2.3.1. The second type is a document whose content is difficult to be annotated such as a document written with too chatty style. We describe treating such document in Section 2.3.2.

Checking and filtering them all manually is time-consuming. The number of documents in the web is much more than the target quantity. Therefore, we filter

out the inadequate documents automatically by simple rules before checking them manually. Furthermore, the remaining documents are checked manually and we only annotate the adequate documents.

In our research, we build the corpus in the following steps.

1. Extract Japanese sentences from crawled HTML files with Kawahara et al. [22]’s method
 - (a) Detect candidate Japanese Web pages with character encoding
 - (b) Determine that a document that include post positions “が,” “を,” “に,” “は,” “の” and “で” more than 0.5% is Japanese Web page
 - (c) Split the web page into sentences by punctuations,
 tags and <p>tags
 - (d) Extract sentences that include Hiragana, Katakana and Chinese character more than 60% as Japanese sentences
2. Treat sequence of the sentences from a sentence that is initially extracted as a Japanese document
3. Automatically detect if a first sentence of the extracted document is a headline

The first sentence is the headline Extract three sentences following the headline as a target of annotation (document). Automatically detect if the three sentences can be understood without the headline.

The first sentence is not the headline Extract three sentences from the head of the document as the target.
4. Filter the extracted three sentences by simple rules (The detail is described in Section 2.3.2)
5. Manually filter the documents
6. Manually annotation

Meanwhile, we detect if the document is Japanese Web page in crawling, but do not filter by other bases such as a domain of the Web page.

Headline : 2008. 07. 10 Thursday

気が つけば 梅雨も 明けてました。

Mood-NOM stick rainy season-NOM have ended.

‘I think that the rainy season has ended.’

毎日 暑い日が 続きますね。

Everyday hot day-NOM continue.

‘It’s hot every day.’

父の 手術も 終わり、 少しでも ほっとしています。

Father-GEN surgery-NOM finish short feel easy.

‘I’m feeling a little better because my father’s surgery is over.’

(The rest is omitted.)

Figure 2.1: Example of a document whose headline do not appear in the body

2.3.1 Detecting Documents That Cannot Be Understood Semantic Relation With Only Raw Text

Language is used in speech and documents and creates 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. The situation of the Web pages corresponds to what the Web site the document is posted in and what the document is positioned as in the Web site.

When annotating for the morpheme and syntactic information, there is no need to consider this shared situation because of dealing with each sentence independently. However, in semantic relation annotation, the shared situation must be considered. Since we deal with only text as our annotation target, we include documents whose semantic relations can be understood without such situation for this corpus. For example, a news article can be realized that the document is the news article from its writing style, and in many cases, the content of the document can be understood from only its raw text. On the other hand, Since a page such as a usage note in a product introduction page is difficult to be understood without

Headline : 地震被害額 246 億円に県まとめ ‘The damage caused by the earthquake reached 26.4 billion yen according to Prefectural survey

岩手 宮城 内陸 地震の 被害 額は 22 日 現在
Iwate Miyagi inland earthquake-GEN damage amounts-TOP 22nd as of
県 災害 対策 本部の まとめで 26.4 億円に
prefecture disaster countermeasures office-GEN survey-INS 26.4 billion-ACC
膨らんだ。
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.’

依然として 農村 土木 関係を 中心に
Still farming village construction relation-ACC focus on
被害が 拡大している。
damage-NOM is increasing.

‘The damage is still increasing with focus on farming villages and construction.’

(The rest is omitted.)

Figure 2.2: Example of a document that the elements of its headline appear in the first three sentences

knowledge of the product, the page is inadequate for including this corpus. Such documents are manually removed before the annotation.

Some documents have headlines some of which have a key role in relevance to the situation. However, we remove the headlines from the annotation target because the most of the headlines are ungrammatical sentences such as series of noun phrases. In newspaper articles, there are sentences in the leads which are abstract of the whole document and most of such documents can be understood without the headlines. In the Web pages, some documents do not have sentences acting as an abstract and some documents cannot be understood without the

Headline : 売布神社 ‘Mefu shrine’

どもども、森田です。

Hi be Morita

‘Hi, I’m Morita.’

さてさて、前回 中山寺に 行きましたが、その
Now, previous time 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.’

中山寺から 西に ぶらぶらと 住宅街を
Nakayama temple-ABL west-LOC aimlessly residential area-ACC

歩いていきます。

be walking

‘I am walking to west from Nakayama temple in a residential area.’

(Three sentences are omitted)

この池の 左上あたりに 歩いていくと 売布神社に 付きます。
This pond-GEN upper-left-LOC walk to Mefu shrine-LOC reach

‘Walking around the upper-left of the pond, I had reached Mefu shrine.’

(The rest is omitted.)

Figure 2.3: Example of a document which cannot be understood without its headline

headlines. On the other hand, if the headlines are the date of the blog articles, the documents can be understood without the headlines. We do not include documents which cannot be understood without their headlines among the corpus.

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 are non-headline texts 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 following three sentences and 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 automatically removed. The understandable documents are determined by the following criteria.

In case that a content of a headline has little relevance to a content of a body text, even if the headline is removed, semantic relations of the document may be understood. For example, in Figure 2.1 the headline is the date, therefore the removing the headline has no effect on understanding. 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. In case of that all the words in the headline appear in the first three sentences, it would be appear that the semantic relation can be understood. In Figure 2.2 the first sentence has a role as the abstract and the all content words in the headline 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, the document is hard to understand because it is impossible to reconstruct the information in the headline from the first three sentences. In Figure 2.3, “壳布神社” (Mefu shrine) appears in 6th sentence. However “壳布神社” 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 from only the three sentences. 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. Thereafter, we manually confirm the remaining documents and remove the documents that cannot be understood with the extracted three sentences from the corpus.

2.3.2 Determination of Inadequate Document

The documents collected from the Web include many unsuitable documents. We determine that the following documents are difficult to annotate and do not include in the corpus.

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

Discontinuous sentences The collected documents include what is extracted the sentences which originally allocate separated place as continuous sentences. These documents cannot be annotated the inter-sentential semantic relations.

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

For removing these documents, we automatically remove the documents which 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: it is often that the sentences are ungrammatical such as the error of the text extraction
- More than 10 phrases: the results is often caused by morpheme analysis errors

Table 2.2: Examples of stop phrase

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)
ログイン
(login)
相互リンク
(mutual link)

- Contain Roman characters: these are frequently used in technical terms, acronyms or slang in Japanese, and so apply to domain-specific or unnatural Japanese
- Include stop phrases shown in Table 2.2: to eliminate input forms and automatically generated pages

Additionally, in order to remove identical pages, we remove documents whose edit distance is less than 50 to another document. The remaining inadequate documents as a result of automatically removing are manually removed before the annotation.

2.4 Annotation Criteria

2.4.1 Types of Annotation

We annotate many types of information: morpheme, phrase, dependency, named entity, predicate-argument structure and endophoric relation. The center of this research is annotation of semantic relations (predicate-argument structures and endophoric relations), but 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 essentially annotate these relations by the criteria of the Kyoto University Text Corpus [24] and IREX² and modify small partitions of these criteria. In this section, we describe the important points and modified point of these criteria.

We define basic-phrase, which is composed of one independent word and before and after attached words, as the annotation unit for the predicate-argument structure and the endophoric relation. We show an example of the partitions by the basic-phrase in Example (2.4). We annotate the predicate-argument structure and the endophoric relation to each basic-phrase and the arguments and the referents are selected from the basic-phrases. If an argument or a referent is compound noun, we consider the head basic-phrase of the compound noun as the argument or the referent. In Example (2.4), the referent of “党” (Party) is “国民新党” (People’s New Party), and so we annotate “新党” (new party), which is the head of “国民新党,” as the referent.

(2.4) 7月 17日 国民 新党 災害 対策 事務局長と
 July 17th People new party disaster countermeasures office chief-ABL

して、党-を 代表して 現地に 向かいました。
 do Party-ACC represent field-ALL went

‘On July 17th, I went to the field since I was representative of the party as the chief of the disaster countermeasures office of New People’s Party.’

(党 ←=:新党)

We annotate the predicate-argument structure in the same way of the Kyoto University Text Corpus. Cases of the arguments are defined as surface cases such as ガ, ヲ and ニ and cases that represent relations such as TIME and MODIFY, and the total number of the case types is 42. The arguments are sorted into three types. One is an argument which has dependency relation with predicate, another is an argument omitted in zero endophora and the other is an argument omitted in zero exophora. In the zero endophora and the zero exophora annotation, we annotate whether a zero pronoun exists and also a referent of the zero pronoun as information of the argument. The referents of the zero exophora are selected from

²<http://nlp.cs.nyu.edu/irex/NE/>

Table 2.3: Candidate referents of zero exophora

Referent	Example
[author]	時間とお金の関係について ([author] ガ) <u>考えてみた</u> 。 (I) thought about a relation between time and money. (考えてみた ← ガ:[author])
[reader]	コーディネートが楽しく ([reader] ガ) <u>選べます</u> 。 (You) can select coordinates delightfully. (選べます ← ガ 2:[reader], ガ:コーディネート)
[US-person]	一切の釉薬を ([US-person] ガ) <u>用いない</u> のも特徴で… (US-person) Avoiding glaze is one of features <i>cdots</i> (用いない ← ガ:[US-person], ヲ:釉薬)
[US-matter]	必ず ([US-matter] ガ) <u>削除される</u> というわけではありません。 ([US-matter]) is not always deleted. (削除される ← ガ:[US-matter], ニ:管理人)
[US-situation]	このシーズンに ([US-situation] ガ) <u>なると</u> … ([US-situation]) come this season… (なると ← ガ:[US-situation], ニ:シーズン)

candidate referents shown in Table 2.3. “US-person” refers to not only unspecified (indefinite) person but also a person that is not mentioned in a document. The predicate-argument structures are annotated to not only the predicates but also verbal nouns.

In the Kyoto University Text Corpus, a **ガ** 2 case is defined for double-subject construction and they annotate as the following example.

- (2.5) 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 ← ガ 2:彼, ガ:ビール)

In Example (2.6), since “象が長い” (The elephant is long) is contrived expression, “象” (elephant) is not handled as an argument of a **ガ** 2 case under the basis of the Kyoto University Text Corpus. In contrast, we deal with words which expresses topic as an argument of a **ガ** 2 case and so annotate “ガ 2:象, ガ:鼻” to “長い.”

- (2.6) 象は 鼻が 長い。
 Elephant-TOP trunk-NOM long
 ‘The elephant’s trunk is long’
 (長い ← ガ 2:象, ガ:鼻)

The endophoric relations are annotated according to the Kyoto University Text Corpus. In the Kyoto University Text Corpus, the endophoric relations are categorized into three types. The first of these is an endophoric relation which have coreference relation, and we annotate this relation by using “=” tag. The second of these is a bridging reference which can be expressed in the form, “A の B” (B of A), and we annotate “ノ:A” to B. The third of these is an endophoric relation which does not have a coreference relation and a bridging reference cannot be expressed in the form, “A ノ B” (B of A), and we annotate these with “≃.” The endophoric relations are not annotated to not only relations between nouns but also relations between predicates and between a noun and a predicate.

For annotating multiple arguments for a case of a predicate, Kyoto University Text Corpus defines 3 types, “AND,” “OR” and “?.” “AND” is used for an expression that annotated arguments are parallel and both of them execute such

as “A および B が V した.” In Example (2.7), since both “太郎” (Taro) and “花子” (Hanako) do “学校に行った” (went to school), arguments of a **ガ** case of “行った” are annotated with “太郎” and “花子” as “AND” relation.

- (2.7) 太郎と 花子は 学校に 行った.
 Taro and Hanako-TOP school-DAT went
 ‘Taro and Hanako went to school.’
 (行った ← ガ:太郎 AND 花子)

“OR” is used for an expression that annotated arguments are parallel and either of them execute such as “A または B が V した.” In Example (2.8), since an agent of “持っていく” (will carry) is either “太郎” or “花子,” arguments of a **ガ** case of “持っていく” are annotated with “太郎” and “花子” as “OR” relation.

- (2.8) 太郎か 花子が 持っていきます.
 Taro or Hanako-NOM will carry
 ‘Taro or Hanako carry.’
 (行った ← ガ:太郎 OR 花子)

“?” is used in cases that actual arguments cannot be identified from surface expression and context. In Example (2.9), since an agent of “撤廃する” (abolish) can be interpreted as either “高知県” (Kochi prefecture), “橋本知事” (governor Hashimoto) and [US-person] such as members of the prefecture assembly and office staff, arguments a of **ガ** case of “撤廃する” are annotated with them as “?” relation.

- (2.9) 高知県の 橋本 知事は 国籍
 Kochi prefecture-GEN Hashimoto governor-TOP nationality
 条項を 撤廃する 方針を 明らかにした。
 requirement-ACC abolish policy-ACC disclose
 ‘Hashimoto, governor of Kochi prefecture, disclosed a policy that nationality requirement will be abolished.’
 (撤廃する ← ガ:高知県 ? 橋本知事 ? [US-person])

We annotate named entities according to the basis of IREX. The named entities are expressed by their scope and type. The types of the named entity are 8 types shown in Table 2.4. In Example (2.10), “ラズナー” (Rasner) is annotated with “PERSON” and “ホークス” (Hawks) is annotated with “ORGANIZATION.”

Table 2.4: The types of named entity

ORGANIZATION
PERSON
LOCATION
ARTIFACT
DATE
TIME
MONEY
PERCENT

- (2.10) そこで ラズナーと ホークスの 今季 対戦 成績を
 And so Rasner-COM Hawks-GEN this season match-up result-ACC

掲載します。

post.

‘And so we post the this season’s scoreline between Rasner and Hawks.’

$$\left(\begin{array}{l} \text{ナズナー} \leftarrow \text{PERSON} \\ \text{ホークス} \leftarrow \text{ORGANIZATION} \end{array} \right)$$

2.4.2 Mentions of Author and Reader

The author and the reader of a document are important in the discourse. Since there are phenomena which are influenced by the author/reader and the author/reader tend to be omitted, the author/reader behave differently from other discourse elements. Existing corpora based on newspaper articles have considered the author/reader as referents of zero exophora shown in Table 2.3. However, the author/reader are sometimes explicitly mentioned in a document as author/reader mentions.

- (2.11) 私の 担当する お客様に 褒めて頂きました。
 I-GEN be in charge client-DAT receive praise

‘I received praise from a client whom I am in charge of.’

$$\left(\begin{array}{l} \text{褒めて頂きました} \leftarrow \text{ガ:私, 二:お客様} \\ \text{私} \leftarrow \text{=: [author]} \end{array} \right)$$

In example (2.11), “私” (I) is mentioned in a document as the author mention. In such case, existing corpora have treated omissions of the author/reader mentions as zero endophora in the same way of omissions of other discourse entities and have not expressly dealt with them as the author/reader. However, for researching behaviors of the author/reader in documents, it is also necessary to research behaviors of the author/reader mentions. In this research, we annotate that “私” (I) is the author mention as a coreference relation.

Because documents treated in our research are written by various authors for various readers, the author/reader mentions is mentioned not just as personal pronouns but as various expressions. In Example (2.12), the author is mentioned by such as “こま” (Koma), which is a proper representation, “主婦” (housewife) and “母” (mother), which are position names.

(2.12) 東京都に 住む 「お気楽 主婦」 こま です。
 Tokyo-metropolis-LOC live “easygoing 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 男の子の 母を しています。
 0 years old and 6 years old-GEN boys-GEN mother-ACC doing
 ‘I am the mother of two boys who a baby and 6 years old.’
 (母 ←=:主婦)

In our research, we annotate not just personal pronouns but all expressions that correspond to the author/reader of a document as the author/reader mentions.

For annotating the author/reader mentions, we annotate “=:author]” and “=:reader]” to the author/reader mentions as exophora. When the author/reader mentions are compound nouns, we annotate to a head basic-phrase of the compound noun. Assuming that the author and the reader are only one element in each document, we annotate “=:author]” and “=:reader]” to up to one expression respectively. If the author/reader is mentioned in some expressions, which are coreference, we annotate to one of them. In Example (2.12), the three underlined parts are the author mentions, and so we annotate “=:author]” to only “主婦” and “=:主婦” to “こま” and “母.”

2.4.3 Author Mention

In this section, we describe issues on the annotation of the author mentions such as expressions that refer to organization or homepage.

In homepage of the organization such as a company, it is often described that the organization has personality and animacy. In such case, actual author should be a site administrator, but we deal with the organization as the author and annotate an expression that refers to the organization as the author mention. In Example (2.13), it is thought that the site administrator wrote the document as a representative of “神戸徳洲会病院” (Kobe Tokushukai Hospital), and so “病院” (hospital), which is the head of “徳洲会病院,” is annotated with “=: [author].”

(2.13) 神戸 徳洲会 病院では 地域の 医療 機関との
Kobe Tokushukai hospital-TOP area-GEN medical agency-COM

連携を 大切にしています。
coordination-ACC value

‘Kobe Tokushukai Hospital values coordination with community medical agency.’

(病院 ←=: [author])

ご来院の 際は、是非 かかりつけの 先生の
coming to hospital-GEN when should regular doctor-GEN

紹介状を お持ち下さい。
letter of introduction take

‘When you come to the hospital, you should take a letter of introduction of a regular doctor.’

And, an expression that refers to a web site such as Example (2.14) is also treated as the author mention.

(2.14) 結婚 応援 サイトは、皆さんの 素敵な 人生の パートナー
marriage backup site-TOP, you-GEN nice life-GEN partner

探しを 応援します。
searching back up

‘Marriage backup site back up searching for your nice life partner.’

(サイト ←=: [author])

In Web pages of a shop and others, there are both an expression that refers to the shop and one that refer to a manager or staff of the shop. In such case, the author mention is annotated by judging which behaves as the author. In Example (2.15), since “タウンロフト館” (a name of a shop) behaves as the author, “=: [author]” is annotated not to “スタッフ” (staff) but to “館” (building).

- (2.15) タウン ロフト 館の 店舗 情報を お伝えします。
 Town Loft building-GEN store information will let you know
 ‘I will let you know store information of Town Loft Building.’
 (館 ←=: [author])

ご来店 予定の 際に アクセスで お困りでしたら、
 Coming to the store plan-GEN when, access-INS have trouble,
 当店 スタッフまで お気軽に ご連絡下さい。
 our shop staff-DAT feel free contact

‘If you have trouble in planning to come to the store, please feel free to contact to staffs of our shop.’

(当店 ←=: 館
 スタッフ ← ノ: 当店)

On the other hand, in Example (2.16), since “かおりん” (Kaorin), the manager, introduces the shop as the author, “=: [author]” is annotated to “かおりん.”

- (2.16) 『ソブレ』 アマゾン 店, 店長の かおりん です。
 “Sobre” Amazon shop, manager-GEN be Kaorin.
 ‘I’m Kaorin, a manager of “Sobre” on Amazon.’
 (かおりん ←=: [author])

新 商品の 情報や、 かおりん 日記を 相棒
 new item-GEN information and, Kaorin diary-ACC partner
 みかんと 一緒に 紹介します。
 Mikan-COM together introduce.

‘I introduce information of new items and Kaorin’s diary with my partner, Mikan.’

2.4.4 Reader Mention

Since documents that are treated in our corpus are collected from the Web, the documents are accessible by everybody. Hence, strictly speaking, expressions that refer to the reader are only second person pronouns. In Example (2.17), since “皆さん” (you all) is a honorific expression of a second personal pronoun, “皆さん” is annotated as the reader mention.

- (2.17) 皆さん は 初詣は どこに
 you all-TOP New Year’s first visit to a shrine-TOP where
 行かれたでしょうか？
 did went.

‘Where did you go on the first shrine visit of the New Year.’

(皆さん ←=:reader])

On the other hand, although documents can be available for inspection by everybody, many documents have targets that the author assume as the readers. In this research, we define expressions that refer to such targets as the reader mention. For example, Since Example (2.18) is a guideline for “ぼすれん登録会員” (registered member of Posuren), we treat “ぼすれん登録会員” as the reader mention and annotate “=:reader]” to “会員” (member), which is the head.

- (2.18) ぼすれん 登録 会員 が コミュニティ サービスを
 Posuren registered member-NOM community service-ACC
 ご利用いただくには、本ガイドラインの 内容を 承諾いただく
 use, this guideline-GEN content-ACC agree
 ことが 条件となります。
 that-NOM is provision.

‘A provision for a registered member of Posuren using community service is agreement to this guideline.’

(会員 ←=:reader])

On the other hand, in Example (2.19), since “写真を撮られた方” (person who take a photo) is not an expression that the author assume as the whole of readers but a part of assumed readers, “方” (person) is not treated as the reader expression.

- (2.19) 桜の 下で 写真を 撮られた 方も
 cherry tree-GEN under photo-ACC take person

多いのではないのでしょうか。

may be many

‘Many of you might take a photo under a cherry tree.’

2.4.5 Criteria for Ambiguous Annotation

In Japanese, the expression that does not specify arguments corresponding to an agent or a patient is often used. In Kyoto University Text Corpus, when candidate arguments are mentioned in a document, the expression is annotated with “?” which is described in Section 2.4. Furthermore, even when there are no candidate in expression in a document, most of arguments can be annotated as “[US-person]” in newspaper articles, which are targets of Kyoto University Text Corpus. On the other hand, when documents in whose discourse the author/reader appear such as Web documents are annotated, many arguments can be interpreted as also the author/reader.

In this research, when an argument have multiple interpretation, the argument is annotated with all candidate arguments by “?” relations. We make an annotation manual for exemplary expression that can be ambiguously interpreted and illustrate to annotators. In this section, we describe criteria for annotating [author], [reader] and [US-person] as an argument. Additionally, in following examples, we hold up instances of [author], [reader] and [US-person] as examples but the author/reader mentions, which are described in Section 2.4.2, are treated the same as [author] and [reader].

2.4.6 Criteria of Annotating [US-person]

When an event is universal or an argument refers to a person who is not explicitly mentioned in a document, [US-person] is annotated as a argument.

In Example (2.20), since this event is universal, [US-person] is annotated to a **ガ** case that corresponds to an agent of “焙煎する” (roast).

- (2.20) コーヒー 生豆とは 焙煎する 前の 裸の 状態の
 coffee raw bean-TOP before roast-GEN natural-GEN state-GEN
 豆を いい、…
 bean-ACC say …
 ‘Coffee green bean means natural bean before roasting, …’
 (焙煎する ← ガ:[US-person], ヲ:豆)

In Example (2.21), since patients of “お送りしています” are members of a mail magazine, who are not explicitly mentioned, [US-person] is annotated to 二 case.

- (2.21) メール マガジンでは お得な 情報を お送りしています。
 mail magazine-TOP saving information-ACC send to.
 是非 ご登録ください。
 must register
 ‘In e-mail magazine, we send saving information to the member. You must register.’
 (お送りしています ← ガ:[author], 二:[US-person])

2.4.7 Criteria of Annotating [author]

When an expression is interpreted that the author have an experience of an event or the description is applied to the author, [author] is annotated to the argument.

In Example (2.22), since written content is common belief and also applied to the author (railway company), we annotate not only [US-person] but also [author] to a ガ case of “整備しておかねばなりません” (need to keep up).

- (2.22) 線路は 列車の 安全を 確保し、快適な 乗り 心地を
 rail-TOP train-GEN safety-ACC ensure, comfortable ride quality-ACC
 維持する 状態に 整備しておかねばなりません。
 maintain state-DAT need to keep up
 ‘Rail need to be kept the state that the rail ensures safety of train and maintains comfortable ride.’
 (整備しておかねばなりません ← ガ:[author] ? [US-person], ヲ:線路, 二:
 状態)

In Example (2.23), since content is common belief but can be interpreted that

the author have an experience of tracing the source, we annotate [US-person] and [author] to a **ガ** case of “**辿れば**” (trace).

- (2.23) **しかし 名前からも 察する ことが できるように、 源流を**
 however from name guess that-GEN can, source-ACC

辿れば 「田楽」に 行き当たる。
 trace ”dengaku”-DAT come across

‘However as can be guessed from name, tracing source comes across “dengaku.” ’

(**辿れば** ← **ガ**:[author] ? [US-person], **ヲ**:源流)

2.4.8 Criteria of Annotating [reader]

[reader] is annotated to an expression that promotes to the reader such as request expression and an expression that recommends something to the reader. In the case of recommendation expression, the annotation is judged not only from a target predicate also from context.

In Example (2.24), since the author appeals to the reader, [reader] is annotated to a **ガ** case.

- (2.24) **メールの 際は 必ず 名前を 添えてください。**
 mail-GEN when make sure name-ACC affix

‘Please make sure to affix name to mail.’

(**添えてください** ← **ガ**:[reader])

Example (2.25) is sentences that locate in an on-line shopping site. Since someone can execute “**選択**” (select), but the whole of the page is interpreted as recommendation of the on-line shipping to the reader, [reader] and [US-person] are annotated to a **ガ** case of “**選択できます**” (can select).

- (2.25) **分割 払いなど、 多彩な お支払い 方法から**
 installments such payment, various payment from method

選択できます。 詳しくは **ガイドを ご参照ください。**
 can select. details-TOP guide-ACC please refer.

‘You can select from various payment methods such as installments payment. As for details, please refer to ’

(**選択できます** ← **ガ**:[reader] ? [US-person])

In Example (2.26), since an expression can be interpreted as recommendation to the reader, [reader] is annotated to a **ガ** case. Therefore, since the expression can be interpreted as common belief and author’s experience, [author] and [US-person] are also annotated to the **ガ** case.

- (2.26) ブログに 記事を 書き込んで、 インターネット上で 公開する のは
 blog-DAT article-ACC post, on the Internet-LOC publish
 とても 簡単です。
 very easy

‘It is very easy to post blog articles and publish on the Internet.’

(公開する ← **ガ**:[author] ? [reader] ? [US-person], **ヲ**:article)

In Example (2.27), since an expression is inducement to the reader, [reader] is annotated to a **ガ** case. Though communication through the Web site, assuming that the author looks concurrently with the reader, [author] is annotated with “AND” relation.

- (2.27) まずは 株式 市場の 分類を 見てみましょう。
 First of all, stock market-GEN classification-ACC let’s look.

‘First of all, let’s look a classification of stock markets.’

(見てみましょう ← **ガ**:[reader] AND [author])

2.5 Constructed Corpus

1,000 documents have been annotated by 3 annotators. We named the annotated corpus Diverse Document Leads Corpus (DDLCC). In this section, we describe procedure for annotation and discuss about statistics and properties of constructed corpus.

In discussion about the statistics and the properties, first, we discuss about fundamental statistics and properties, such as document types and writing styles. Next, we discuss about appearances and behaviors of the author/reader in discourses. In these discussion, we compare DDLCC to Kyoto University Text Corpus when needed. Finally, we discuss inter-annotator agreement.

2.5.1 Procedure and Setting of Annotation

In the actual annotation, we first annotated automatically by a Japanese morpheme analyzer, JUMAN³ and a Japanese dependency parser, KNP⁴, and then modified the annotation by using the GUI tool. Each document is annotated by an annotator, and then the annotation is checked and modified by another annotator. The information given to the annotators are only raw three sentences, which are a target of the annotation, and information that the texts are extracted from the Web.

The number of the annotators is three and all of them are experienced annotators. Before beginning of the annotation, we handed out manuals of Kyoto University Text Corpus⁵ and IREX⁶ and definitions and examples of author/reader mentions. After we had annotated to 1,000 documents, problems about ambiguous expressions described in Section 2.4.5 were revealed. Therefore, we discussed about criteria for annotation with the annotators and gave the result of the discussion as additional manual. We modified the 1,000 documents based on the new criteria. The annotation is now in progress with a goal of the annotating to 5,000 documents.

We will add URL information which we got a document to the constructed corpus from. Meanwhile, semantic relations that are annotated to the corpus based only on raw text, and the URL information is not essential for the semantic relation corpus.

2.5.2 Statistic of DDLC

Fundamental statistics of the constructed corpus are shown in Table 2.5. We also show statistic of Kyoto University Text Corpus (KUTC) for comparison. Since morphemes per sentence of DDLC, about 17, are less than ones of KUTC, sentences of DDLC are shorter than ones of KUTC. In DDLC, about two third of

³<http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN>

⁴<http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP>

⁵http://nlp.ist.i.kyoto-u.ac.jp/nl-resource/corpus/KyotoCorpus4.0/doc/syn_guideline.pdf
and http://nlp.ist.i.kyoto-u.ac.jp/nl-resource/corpus/KyotoCorpus4.0/doc/rel_guideline.pdf

⁶<http://nlp.cs.nyu.edu/irex/NE/df990214.txt>

Table 2.5: Statistics of corpus

	DDLC	KUTC
No. of documents	1000	567
No. of sentences	3000	4929
No. of morphemes per sentence	16.9	26.0
No. of phrases per sentence	6.3	9.9
No. of basic-phrases per sentence	8.0	13.1
No. of annotated basic-phrases per sentence	5.2	9.3

basic-phrases, which are targets of annotation, are annotated with any semantic relations.

For researching difference of writing styles, we show rates of sentences that include modalities and honorific expressions in Table 2.6 and Table 2.7. The modalities and the honorific expressions are automatically annotated by KNP. From Table 2.6, DDLC contains many modalities that have function of approach from the author to the reader, such as request, inviting, order and will. The will modalities are well contained in KUTC because many of the will modalities are used in quotes from talking. On the other hand, KUTC contains many assessment:strong and realization-evidence modalities⁷. These modalities are widely used in news reports and editorial and the differences of appearances of these modalities show difference of writing styles. From Table 2.7, in DDLC, more than 80% sentences are used any honorific expressions. Since respectful language and modest language are often used, it is conceivable that DDLC includes many documents that have consciousness of existence of the reader.

Since we collected documents from Web without limitation of domain, DDLC consists of various documents. For researching tendency of annotated documents, we manually classified annotated documents into 13 types. We show the classification result in Table 2.8. Table 2.8 shows that various documents such as company/shop pages, blog/personal pages and encyclopedia/illustration articles, are included. Additionally, one category consists of very various types of docu-

⁷An example of assessment:strong is “関係を無視した暴言と 言わざるを得ない。” and one of realization-evidence is “海部政権誕生の願望が 込められているようだ。”.

Table 2.6: Ratio of sentences in which a modality expression appear

	DDLC	KUTC
Request:A	6.82%	0.12%
Request:B	1.00%	0.47%
Inviting	1.43%	0.45%
Order	6.73%	0.20%
Will	3.04%	2.45%
Question	1.50%	1.16%
Inhibition	0.01%	0.12%
Assessment:weak	0.73%	0.35%
Assessment:strong	1.10%	1.59%
Realization-estimate	1.97%	1.61%
Realization-probability	1.07%	0.69%
Realization-evidence	1.74%	2.32%

Table 2.7: Ratio of sentences in which a honorific expression appear

	DDLC	KUTC
Polite	62.15%	1.83%
Respectful	9.60%	0.12%
Modest	11.97%	2.03%

Table 2.8: Result of manually classification of document types

Type of documents		No. of documents
Company/shop page	Item description and mail order page	165
	Introduction of the company/shop	116
	Others	133
Blog/personal page	Event of oneself	119
	Introduction of someone	114
	Introduction of the web page	27
	Others	50
Encyclopedia/illustration article		147
Searching/introduction page		40
Manual/instruction of products or applications		33
News article		32
Novel		18
Others		6

ments. For example, company/shop pages include not only sites of the corporation also various pages such as sites of schools, local governments and public institutions. Furthermore, there are documents that is difficult to be categorized to one category such as a blog page that locates in a corporation site.⁸

2.5.3 Author/Reader Mention

The numbers of the documents with respect to types of the author/reader annotations are shown in Table 2.9. “Explicit” of “Appear” means that an author or a reader is mentioned explicitly and annotated with author/reader mentions. “Implicit” of “Appear” 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 “Not appear.” As a result, the author appears in the discourse on the about 70% of documents and the reader appeared on the about 50%. The author/reader are sometimes not mentioned explicitly though the author/reader appear in the discourse.

⁸About this time, such pages are categorized to company/shop pages

Table 2.9: Appearance of the author/reader in a document

	Appear		Not appear	Total
	Explicit	Implicit		
Author	271	408	321	1000
Reader	84	417	499	1000

145 expressions and 25 expressions are used as author mentions and reader mentions respectively. The examples and their frequency are shown in Table 2.10 and Table 2.11. Here, we deal with expressions that have coreference relation with the author/reader mention as the author/reader mention.⁹ Among the author mentions, “私” (I) is the most frequently appeared expression, which appeared 56 times and is often used in blog articles. Expressions that a company refers to oneself, such as “弊社” (our company) and “当社” (our company), often appear. Additionally, there are various expressions such as position names (“管理人” (moderator), “主婦” (housewife) and “監督” (director)), words indicating organization (“協会” (association) and “病院” (hospital)) and proper representation (“真理子” (Mariko) and “ローソン” (Lawson)). Since 106 words and 24 words appear as the author mention once and twice respectively, many words become the author mentions depending on the context. Among words that mention the reader, honorific expressions of second person pronouns such as “皆様” (you all) and “皆さん” (you all) often appear. This is because that many of the web pages assuming potential readers are the business pages, and in these pages, honorific expressions are often used to the reader. Additionally, there are the words assuming document specific readers such as “生徒” (student), “ドライバー” (driver) and “市民” (citizen). Finally, “自分” (self) is used as both the author and reader mentions.

2.5.4 Zero Reference Relation

Numbers of annotated zero endophora and zero exophora are shown in Table 2.12. From this table, the zero endophora/exophora occurred most frequently in a **ガ**

⁹In example 2.12, we deal with all of “主婦”, “こま” and “母” as the author mentions

Table 2.10: Examples of the author mentions (excerpt) Table 2.11: Examples of the reader mentions (excerpt)

Author mention	Frequency
私 (I)	56
弊社 (our company)	12
店 (shop)	11
会 (society)	10
当社 (our company)	9
自分 (self)	8
当店 (our shop)	6
管理人 (moderator)	5
協会 (association)	3
病院 (hospital)	3
主婦 (housewife)	2
監督 (director)	1
ローソン (Lawson)	1
真理子 (Mariko)	1
Total	382

Reader mention	Frequency
皆様 (you all)	26
あなた (you)	23
客 (customer)	15
方 (person)	9
人 (person)	7
皆さん (you all)	7
自分 (self)	4
会員 (member)	4
自身 (self)	3
ユーザー (user)	2
ドライバー (driver)	1
生徒 (student)	1
贈り主 (giver)	1
市民 (citizen)	1
Total	107

case and about 60% of them are zero exophora. The ratios of zero exophora in a **ニ** case and a **ガ** 2 case are also high. Breakdown of numbers of zero endophora is shown in Table 2.13 and one of zero exophora is shown in Table 2.14. In Table 2.13, “Author” and “Reader” mean that referent of zero endophora is an author/reader mention or has coreference relation with the author/reader mention.¹⁰ Table 2.13 and Table 2.14 indicate that more of referents of **ガ** and **ガ** 2 cases are the author. In other words, many of agents of predicates are the author. Referents of a **ニ** case more often are the reader than ones of other cases. It is because that there are many expressions that have a role of approach from the author to the reader such as “[author] **ガ** [reader] **ニ**お勧めする” ([author] recommends to [reader]) and “[author] **ガ** [reader] **ニ**販売する” ([author] sells to [reader]).

For comparison, numbers of zero reference relations of KUTC are show in Table

¹⁰In example (2.12), cases that the referents are “主婦”, “こま” or “母” are classified to author

Table 2.12: Number of zero references in DDLC

	Zero endophora	Zero exophora	Total
ガ	1867 (37.1%)	3168 (62.9%)	5035 (100.0%)
ヲ	662 (76.9%)	199 (23.1%)	861 (100.0%)
ニ	515 (40.2%)	766 (59.8%)	1281 (100.0%)
ガ 2	107 (34.7%)	201 (65.3%)	308 (100.0%)
Others	607 (83.6%)	119 (16.4%)	726 (100.0%)
Total	3758 (45.8%)	4453 (54.2%)	8211 (100.0%)

Table 2.13: Breakdown of the numbers of zero endophora in DDLC

	Author	Reader	Others	Total
ガ	664 (35.6%)	154 (8.2%)	1049 (56.2%)	1867 (100.0%)
ヲ	12 (1.8%)	5 (0.8%)	645 (97.4%)	662 (100.0%)
ニ	93 (18.1%)	56 (10.9%)	366 (71.1%)	515 (100.0%)
ガ 2	29 (27.1%)	10 (9.3%)	68 (63.6%)	107 (100.0%)
Others	32 (5.3%)	7 (1.2%)	568 (93.6%)	607 (100.0%)
Total	830 (22.1%)	232 (6.2%)	2696 (71.7%)	3758 (100.0%)

Table 2.14: Breakdown of the numbers of zero exophora in DDL

	Author	Reader	US-person	US-matter	US-situation	Total
が	1079 (34.1%)	860 (27.1%)	1045 (33.0%)	86 (2.7%)	98 (3.1%)	3168 (100.0%)
ヲ	6 (3.0%)	16 (8.0%)	52 (26.1%)	115 (57.8%)	10 (5.0%)	199 (100.0%)
ニ	90 (11.7%)	234 (30.5%)	358 (46.7%)	71 (9.3%)	13 (1.7%)	766 (100.0%)
が 2	67 (33.3%)	69 (34.3%)	59 (29.4%)	6 (3.0%)	0 (0.0%)	201 (100.0%)
Others	19 (16.0%)	17 (14.3%)	47 (39.5%)	28 (23.5%)	8 (6.7%)	119 (100.0%)
Total	1261 (28.3%)	1196 (26.9%)	1561 (35.1%)	306 (6.9%)	129 (2.9%)	4453 (100.0%)

Table 2.15: Number of zero references in KUTC

	Zero endophora	Zero exophora	Total
ガ	7876 (76.9%)	2372 (23.1%)	10248 (100.0%)
ヲ	1529 (88.5%)	198 (11.5%)	1727 (100.0%)
ニ	1753 (70.6%)	730 (29.4%)	2483 (100.0%)
ガ 2	211 (89.8%)	24 (10.2%)	235 (100.0%)
Others	3019 (96.6%)	107 (3.4%)	3126 (100.0%)
Total	14388 (80.7%)	3431 (19.3%)	17819 (100.0%)

2.15 and breakdown of numbers of zero exophora of KUTC is shown in 2.16. Since KUTC is not annotated with author/reader mentions, breakdown of numbers of zero endophora could not be researched. According to this comparison, ratio of zero exophora in DDLC is much higher than one in KUTC and this tendency is particularly strong in ガ, ニ, ガ 2 cases. In these cases, when comparing referents of zero exophora, many of the referents in DDLC are [author] and [reader] but few of the referents in KUTC are them. The author/reader of a document hardly appear in discourse in newspaper articles but often appear in Web documents. This difference also can be seen in the referent of zero reference.

For researching ambiguous expressions, which are described in Section 2.4.5, numbers of arguments that are annotated with any of [author], [reader] or [US-person] and numbers of arguments that are annotated with a number of them are shown in Table 2.17. According to Table 2.17, about 13% of the arguments that are annotated with any of [author], [reader] or [US-person] have multiple interpretations.

2.5.5 Inter-Annotator Agreement

For researching inter-annotator agreements of author/reader mentions and predicate-argument structures, three annotators annotated common 100 documents. Morphemes, syntactic relations and coreference relations, which are required for the annotation of the author/reader mentions and the predicate-argument structures, had been preliminarily annotated upon consultation between the annotators, and then the annotators independently annotated the author/reader mentions and the

Table 2.16: Breakdown of the numbers of zero exophora in KUTC

	Author	Reader	US-person	US-matter	US-situation	Total
ガ	104 (4.4%)	3 (0.1%)	1918 (80.9%)	12 (0.5%)	335 (14.1%)	2372 (100.0%)
ヲ	0 (0.0%)	0 (0.0%)	109 (55.1%)	51 (25.8%)	38 (19.2%)	198 (100.0%)
ニ	8 (1.1%)	0 (0.0%)	704 (96.4%)	5 (0.7%)	13 (1.8%)	730 (100.0%)
ガ 2	5 (20.8%)	0 (0.0%)	19 (79.2%)	0 (0.0%)	0 (0.0%)	24 (100.0%)
Others	0 (0.0%)	0 (0.0%)	82 (76.6%)	3 (2.8%)	22 (20.6%)	107 (100.0%)
Total	117 (3.4%)	3 (0.1%)	2832 (82.5%)	71 (2.1%)	408 (11.9%)	3431 (100.0%)

Table 2.17: Number of arguments that have multiple interpretations

[author]	1646(48.8%)
[reader]	766(22.7%)
[US-person]	507(15.1%)
[author]+[reader]	27(0.8%)
[author]+[US-person]	74(2.2%)
[reader]+[US-person]	237(7.0%)
[author]+[reader]+[US-person]	111(3.3%)
Total	3368

Table 2.18: Inter-annotator agreement of the author/reader mentions

	A vs B	B vs C	C vs A	Average
Author mention	0.89	0.81	0.88	0.86
Reader mention	0.67	0.80	0.86	0.77

predicate-argument structures.

Agreements of the author/reader mentions are calculated by F1 score for dealing with annotations of one annotator as correct annotations. The results are shown in Table 2.18.

As disagreements of the author/reader mentions are checked, many of the disagreements may be considered to be mistakes of the annotators. In actual annotation, such disagreements should be removed because the annotations are checked by another annotator in all documents.

On the other hand, Example (2.28) is an example of a disagreement caused by conflict of judgments of the annotators. In this document, one annotator judged that “スタッフサービス” (staff service) is the author mention, but the others judged that there is no author mention in the document. It is thought to be the cause of the disagreement that the annotator who judged that “スタッフサービス” is the author mention recognized “スタッフサービス” as a name of a temporary manpower company but the annotators who judged no author mention recognized “スタッフサービス” as a name of temporary help service. It is difficult to judge

such expressions that can be interpreted as both a name of service and company name from only three sentences.

- (2.28) スタッフ サービス には 一般 事務だけではなく、医療 機関
 staff service-DAT general not only office work, medical agency
 専門に 派遣される スタッフ サービス メディカルも あります。
 specialty-DAT be sent staff service medical there be
 ‘There are not only general office works but also Staff Medical Service,
 which sent to only medical agency, in Staff Service.’

Example (2.29) is an error caused by inadequacy of criteria. Since it is written that “私” (I) spends a time on a monitor, this document might be a blog article that personalizes a cat and others, and actual author is estimated to the owner. Since we had not defined which discourse entity is author mention in such case, the judges are different between the annotators.

- (2.29) 台風が 通り過ぎる たびに 寒くなっていきますね。
 typhoon-NOM pass every time get cold
 ‘Every time typhoon passes, it gets cold.’

私は 暖かい 場所を 求めて 会社の 中を
 I-TOP warm place-ACC in the search company-GEN inside-ACC
 彷徨います。
 rove

‘I rove inside of the company in the search of warm place.’

今日は この モニターの 上で 過ごすことにしましょう。
 today-TOP this monitor-GEN on will spend a time

‘Today I will spend a time on this monitor.’

Similarly, since same problem might occur in novels written in the first person, we need to define the author mentions when a person who is not the author behaves as the author.

Agreement of predicate-argument structures are calculated by following equation.

$$\begin{aligned}
 F1(B; A, rel) &= \frac{2 \times Recall(B; A, rel) \times Precision(B; , rel)}{Recall(B; A, rel) + Precision(B; A, rel)} \\
 Recall(B; A, rel) &= \frac{\sum_{p \in anno-pred(A, rel)} \frac{|anno(A, rel, p) \cap anno(B, rel, p)|}{|anno(A, rel, p)|}}{|anno-pred(A, rel)|} \\
 Precision(B; A, rel) &= \frac{\sum_{p \in anno-pred(B, rel)} \frac{|anno(A, rel, p) \cap anno(B, rel, p)|}{|anno(B, rel, p)|}}{|anno-pred(B, rel)|}
 \end{aligned}$$

Here, $anno-pred(A, rel)$ means sets of basic-phrases that are annotated with rel (ガ, ヲ, ニ, ...) by annotator A and $anno(A, rel, p)$ means sets of arguments that are annotated to rel cases of a basic-phrase p by annotator A . $Recall(B; A, rel)$, $Precision(B; A, rel)$ can be said macro averages of precision and recall.

Table 2.19: Agreement of predicate-argument structures for predicates

	Overt argument	Zero endophora	Zero exophora	Total
ガ	0.92	0.57	0.71	0.87
ヲ	0.93	0.66	0.46	0.88
ニ	0.91	0.44	0.49	0.78
ガ 2	0.58	0.14	0.44	0.45
Others	0.72	0.27	0.36	0.67

Table 2.20: Agreement of predicate-argument structures for verbal nouns

	Overt argument	Zero endophora	Zero exophora	Total
ガ	0.60	0.45	0.57	0.60
ヲ	0.76	0.48	0.17	0.57
ニ	0.34	0.57	0.42	0.47
ガ 2	0.00	0.33	0.00	0.13
Others	0.52	0.38	0.28	0.49

The agreements for predicates and verbal nouns are shown in Table 2.19 and Table 2.20. Agreements of overt arguments totally tend to be higher. In particular, the agreements for ガ, ヲ and ニ cases of predicates are high because

cases are clearly specified as post positions in these cases. The agreements of zero endophora and zero exophora are almost same values and these agreements are lower than the agreements of overt arguments. Agreements for the verbal nouns tend to be lower than the agreements for the predicates.

In disagreement in predicates, there are many mismatches of cases that a predicate has. Such mismatches can be categorized into 3 types. The first type is that annotated arguments are the same but cases that the arguments are annotated to are different. In Example (2.19), a predicate-argument structure is expressible in both “春巻がくせがない” and “くせが春巻にない.”

- (2.30) くせの ない 春雨は、 サラダ・和え もの・
 peculiarity-GEN nothing gelatin noodles-TOP salad dressed food
 炒めもの・鍋物と様々な料理に 使えます。
 fry food pot food various dish-DAT can be used
 ‘Gelatin noodles, which have no peculiarity, can be used for various dishes such as salads, dressed foods, fry foods and pot foods.’
- a. (ない ← ガ:くせ, ニ:春雨)
 b. (ない ← ガ 2:春雨, ガ:くせ)

Such disagreement are often found in a ガ 2 case but found in also mismatch between a ニ case and a デ case such as Example (2.31).

- (2.31) 唐松岳に 行くつもりだったが、ライブカメラで
 Karamatsudake-DAT plan to go live camera-INS
 現地の 様子を 確認すると、もう 雨が
 actual place-GEN situation-ACC check already rain-NOM
降っている。
 have fallen
 ‘I am planning to Karamatsudake but it has been fallen by checking situation of actual place with a live camera.’
- a. (降っている ← ガ:雨, ニ:唐松岳)
 b. (降っている ← ガ:雨, デ:唐松岳)

In disagreement, we give priority to the cases other than a ガ 2 case in mismatch such as Example (2.30), and in Example (2.30), we annotated with (2.30a). In

other cases, more natural expression is decided by majority vote of the annotators and in Example (2.31), we annotated with (2.31a).

The second type is mismatches of interpretations of predicates. “イメージさせる” (evoke) of Example (2.32) take [US-person] as an argument of a 二 case when “イメージさせる” is recognized as a transitive verb. In other hand, it can be interpreted that “イメージ” is a stative verb and do not take an argument of the 二 case.

(2.32) 床板には 深い 海を イメージさせる 色合いの ガラスを
 floorboard-TOP deep sea-ACC evoke shade-GEN glass-ACC

落とし込んでおります。

be used for

‘Glass that evokes deep sea is used for floorboard.’

a. (イメージさせる ← ガ:色合い, フ:海, ニ:[US-person])

b. (イメージさせる ← ガ:色合い, フ:海)

Similarly, in Example (2.33), annotations are divided into an annotation that interprets “得られた” as a potential verb and an annotation that interprets one as a passive verb.

(2.33) ここに 今までに 得られた 資料の 一部を 公表し、広く
 here till now obtained material-GEN part-ACC publish widely

皆さまからの 資料 提供を 願っております。

from you material provision-ACC hope

‘Here, we publish a part of material that are obtained till now and we hope that you widely provide materials.’

a. (得られた ← ガ:[著者], フ:資料)

b. (得られた ← ガ:資料)

In such case, we select an annotation that has more arguments. In Example (2.32), we annotated with (2.32a), and in Example (2.33), we annotated with (2.33a).

The third type is lack of annotation of [US-person], [US-matter] and [US-situation]. Since these are arguments that explicitly mentioned in a document, even arguments whose case is an obligatory case tend to be missed. In “載せたい” (will write) of Example (2.34), a 二 case is a obligatory case and should be

annotated with [US-matter]. In annotation, one annotator did not annotate to the 二 case and this disagreement might be miss oversight of this annotator.

- (2.34) 私の作詞の 作品や 身近の 出来事や 政治
 my write lyrics work and familiar-GEN event and politics
 経済の 事を 載せたいと思います。
 economics-GEN matter-ACC will write think

‘I think that I will write about my lyrics, familiar events, politics and economics.’

(載せたい ← ガ:私, フ:作品 AND 出来事 AND 事, ニ:[US-matter])

Checked by multiple annotators, such mistakes can be modified.

Disagreements of annotation to ambiguous expression, which are defined in this research, are rarely founded. Example (2.35) is an example of disagreements caused by difficulty of judgment from only context. “判断する” (diagnose) is annotated with (2.35a) if an annotator interprets this document as meaning that the author recommends “サイコロジカルライン” (psychological line) to the reader. In the other hand, in case of that the annotator interprets the document as just description of “サイコロジカルライン”, if the annotator recognizes that “サイコロジカルライン” is the method used by “投資家” (investor) and researchers of investment ([US-person]), the annotator annotates with (2.35b), and if the annotator recognizes that “サイコロジカルライン” is the method used by only the researchers, the annotator annotates with (2.35c). It is difficult to detect which interpretation is correct from three sentences, which are annotation targets of this research. About this time, in such case, all arguments that can be interpreted are annotated and we annotated “判断する” with (2.35d). On the other hand, when whole document is annotated, since the interpretation is uniquely decided from a content of following sentences, there might be no problem in such cases.

- (2.35) サイコロジカルとは、日本語に 訳すと 『心理的』という
 psychological-TOP Japanese-DAT translate psychological
 意味です。
 means

“psychological” is translated as psychological in Japanese.’

サイコロジカル ラインは、投資 家 心理に 基づいて、
 psychological line-TOP investment -er mind-DAT based on
 買われすぎか 売られすぎかを 判断する 時に 利用します。
 over-buying over-selling-ACC diagnose when use

‘Psychological line is used when diagnosing either over-buying or over-selling based on minds of investor.’

直近 1 2 日間で、終値が 前日の 株価を
 last 12 days closing price-NOM yesterday stock price-ACC
 上回った 確率を 示すのが 一般的です。
 exceed probability-ACC show general

‘The psychological line generally show a probability that closing price exceeded yesterday’s stock price.’

- a. (判断する ←[author] ? [reader] ? [US-person])
- b. (判断する ←[US-person] ? 投資家)
- c. (判断する ←[US-person])
- d. (判断する ←[author] ? [reader] ? [US-person] ? 投資家)

Agreements of verbal nouns are lower than ones of predicates. It is because that nouns are annotated with the predicate-argument structures only when the nouns are verbal noun but the basis of the verbal nouns are different between the annotators. In Example (2.36), one annotator judge that “付け合わせ” (garnish) is the verbal noun and annotated with (2.36a), but the other annotators annotated “付け合わせ” with (2.36b), which is an annotation for non-verbal nouns. In such cases, we select the annotation that have more arguments and annotated with (2.36a).

(2.36) 我々 日本人は、生の キャベツの 千切りを トンカツの
 we Japanese-TOP raw cabbage-GEN julienne-ACC pork cutlet-GEN
 付け合わせ に している。
 garnish-DAT do

‘We Japanese use julienne raw cabbages for a garnish of a pork cutlet.’

- a. (付け合わせ ← ガ:日本人, ヲ:千切り, ニ:トンカツ)
- b. (付け合わせ ← ノ:トンカツ)

2.6 Summary of this Chapter

In this chapter, we described the details of the semantically annotated corpus that consists of various documents in the web. In this corpus, we annotated with 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 which consists of 1000 documents. When we analyzed the corpus, we revealed that the author and the reader appeared in many of the documents, these are mentioned in various expressions and these have important role in zero anaphora and zero exophora.

Chapter 3

Author/Reader Mention Detection

In this chapter, we focus author/reader mention detection. In Chapter 2, we defined the author/reader mentions. Since very various expressions are used as the author/reader mentions, it is difficult to detect the author/reader mentions from only lexical information. In this work, we propose a learning-to-rank based author/reader mentions model by using lexico-syntactic patterns as features.

The rest of this chapter is organized as follows. In Section 3.1, we sort issues about the author/reader mentions. In Section 3.2, we present the author/reader mention detection model. In Section 3.3, we explain results of the author/reader mention detection. In Section 3.4, we present conclusion of this chapter.

3.1 Author/Reader Mention Detection

We defined expressions that refer to the author/reader of a document as author/reader mentions in Section 2.4.2. The author/reader tends to be omitted but there are many clues for referent identification of the author/reader such as honorific expressions and modality expressions. Therefore, it is important to deal with the author/reader explicitly in referent identification.

The author/reader is mentioned using a variety of expressions such as personal pronouns, proper expressions, and role expressions.

- (3.1) こんにちは、企画チームの 梅辻 *author* です。
 Hello project team-GEN am Umetsuji
 ‘Hello, I’m Umetsuji on the project team.’
- (3.2) 問題が あれば 管理人 *author* まで お知らせください。
 problem-NOM exist to moderator let me know
 ‘Please let me know if there are any problems.’
- (3.3) お客様 *reader* の アドレスに メールが 自動 返信されます。
 customer-GEN address-DAT mail-NOM automatically be sent
 ‘A reply is automatically sent to the customer’s address.’

In example (3.1), the author is mentioned as “梅辻” (Umetsuji), which is the name of the author, and in example (3.2), the author is mentioned as “管理人” (moderator), which expresses the role of the author. Likewise, the reader is sometimes mentioned as “お客様” (customer) as in Example (3.3). Additionally, since Web documents, which are the target of our study, are freely written and posted, the documents are written by various authors for a wide variety of readers. We show examples of author/reader mentions in Table 3.1. On the other hand, the expressions shown in Table 3.1 are sometimes not used as author/reader mentions. In Example (3.4), since it would appear that “お客様” (customer) refers to a particular customer that differs from the customers assumed to be the readers of this document, “お客様” is not a reader mention in this document.

- (3.4) 先月、 お部屋の リフォームを された お客様 の
 last month room-GEN renovate-ACC did customer-GEN
 例を 紹介します。
 example-ACC will introduce
 ‘I will introduce an example of a customer who renovated a room last month.’

In English and other languages, author/reader mentions can be detected from coreference information because it can be assumed that an expression with a coreference relation with a first or second person pronoun is an author/reader mention. However, since the author/reader tends to be omitted and personal pronouns are rarely used in Japanese, it is difficult to detect author/reader mentions from coreference information.

Table 3.1: Examples of author/reader mentions

	Author	Reader
Personal pronoun	私 (I), 我々 (we), 弊社 (our company)	あなた (you), 皆様 (you all)
Proper expression	ふうこ (Fuko), 畑中 (Hatanaka), インプラントジャパン (Implant Japan)	-
Role expression	管理人 (moderator), 外務副大臣 (vice minister of foreign relations), 調査会社 (research company)	ユーザー (user), お客様 (customer), 会員 (member)

For the above reasons, it is difficult to detect which discourse entity is the author/reader mention from lexical information of the entities. In this study, author/reader mentions are detected from lexico-syntactic (LS) patterns in the document. We use a learning-to-rank [11, 19] algorithm to detect author/reader mentions using the LS patterns as features.

3.2 Author/Reader Detection Model

In this section, we describe a learning-to-rank based author/reader mention detection model. In Section 3.2.1, we describe a discourse entity, which is a unit that we treat the author/reader mention. In Section 3.2.2, we present a method that makes ranking datas for learning-to-rank. In Section 3.2.3, we explain lexico-syntactic patterns that are used in the learning-to-rank.

3.2.1 Discourse Entity

As described in Section 2.4.2, in Diverse Document Leads Corpus (DDLC), the author/reader mentions are annotated to basic-phrases, and when the expressions that have coreferential relations are the author/reader mentions, only one of them are annotated with the author/reader mention. Therefore, we set a unit called **discourse entity**, which is what mentions in a coreference chain are bound into

and we treat the discourse entities as candidate author/reader mentions. For example, in Figure 3.1, since “米子タウンホテル” (Yonago Town Hotel) in the first sentence and “ホテル” (hotel) in the second sentence have a coreferential relation, we treat them as one discourse entity and this discourse entity as the author mention.

3.2.2 Ranking Model

We use a learning-to-rank method for detecting author/reader mentions. This method learns the ranking that entities of the author/reader mentions have a higher rank than other discourse entities. For example, in the author mention detection in Figure 3.1, we make a ranking data that discourse entity (1) has a higher rank than other discourse entities. Then, in author/reader mention detection, we estimate that a discourse entity that is given the highest rank by the learned discriminant function is the author/reader mention. Also, we only deal with discourse entities that satisfy one of the following conditions as the candidate author/reader mentions.

- JUMAN category of a content morpheme is “Person”, “Organization” or “Location”
- A discourse entity is a part of named entity
- Morphemes of the discourse entity include “方” or “人”

Here, it is an important point that there are no author/reader mentions in some documents. The documents in which the author/reader mentions do not appear are classified into two types. The first type is a document that the author/reader do not appear in the discourse of the document such as Figure 3.2. The second type is a document that the author/reader appear in the discourse but all of their mentions are omitted. For example, in Figure 3.3, the author appears in the discourse (e.g. the topical argument of “気がつけば” (think)) but is not mentioned explicitly. We introduce two pseudo entities corresponding to these types.

The first pseudo entity “no author/reader mention (not appear in discourse)” represents the document that the author/reader do not appear in the discourse.

米子 タウン ホテル_{author} は 米子 駅の 正面に
 Yonago town hotel-TOP Yonago station-GEN in front of-LOC
 ございます。
 be
 ‘Yonago Town Hotel is in front of Yonago station.’

駐車 場も 完備しており、ビジネスに 観光に 大変
 parking area-TOP available business-DAT sightseeing-DAT very
 便利な 立地の ホテルです。
 convenient location-GEN hotel
 ‘Since parking area is available, the location of the hotel is very convenient
 for business and sightseeing.’

米子に お越しの際は ぜひ ご利用下さい。
 Yonago-DAT come when-TOP please use
 ‘When you come Yonago, please use the hotel.’

— Candidate discourse entities for author/reader mentions —

(1){ 米子タウンホテル, ホテル }, (2){ 米子駅 }, (3){ 真正面 }, (4){ 駐車場 }, (5){ 立地 }, (6){ 米子 }

Figure 3.1: Example of a document in which an author mention appears

公共 事業の 削減で、地方 経済は 製造 業など
 public enterprise-GEN reduce local economy-TOP manufacturing business
 誘致 企業への 依存 度を 強めてきた。
 invite company depend degree-ACC emphasize

‘Because of reduction of public enterprise, local economies have emphasized
 dependency of invited companies such as manufacturing business.’

このため、 世界 的な 金融 危機による 減産が
 because of this world -like financial crisis reduction of product-NOM
 大きな ダメージに なった。
 big damage-DAT become

‘Because of this, reduction of product due to global financial crisis became
 big damage’

さらに、 アジアなど 新興 国 市場の 台頭や
 additionally Asia emerging country market-GEN rise of
 円高に 伴う 工場の 海外 シフトに
 strong yen-DAT caused by factory-GEN oversea shifting-DAT
 苦しんでいる。
 be suffer

‘Additionally, they are suffer the shifting to oversea caused by rise of emerging
 country market such as Asia and strong yes.’

— Candidate discourse entities for author/reader mentions —

(1){ 地方 }, (2){ 世界的 }, (3){ アジア }, (4){ 新興国 }, (5){ 工場 }, (6){
 海外 }

Figure 3.2: Example of a document in whose discourse an author do not appear

気が つけば 梅雨も 明けてました。
 Mood-NOM stick rainy season-NOM have ended.

‘I think that the rainy season has ended.’

毎日 暑い日が 続きますね。
 Everyday hot day-NOM continue.

‘It’s hot every day.’

父の 手術も 終わり、少しだけ ほんとはしています。
 Father-GEN surgery-NOM finish short feel easy.

‘I’m feeling a little better because my father’s surgery is over.’

— Candidate discourse entities for author/reader mentions —
 (1){父}

Figure 3.3: Example of a document in whose discourse an author appears but an author mention do not appear

It is considered that the document that the author/reader do not appear in has characteristics of writing style such that honorific expressions and request expressions are rarely used. This pseudo entity is represented as a document vector that consists of LS pattern features of the whole document, which reflect a writing style of a document.

The second pseudo entity “no author/reader mention (appear in discourse as omission)” represents the document in which all mentions of the author/reader are omitted and this pseudo entity is represented as 0 vector. Since a decision score of this pseudo entity is always 0, a discourse entity whose score is lower than the score of this pseudo entity can be treated as a negative example in a binary classification.

We describe a method of making a ranking data for each document. We make the ranking data of each document for the author mention and the reader mention using by the following methods, and then all of the ranking data are merged in the author mention and reader mention respectively and the merged datas are fed into the learning-to-rank model.

When there are the author/reader mentions in a document, we make ranking data where the discourse entity of the author/reader mention has a higher rank than other discourse entities and “no author/reader mention” pseudo entities. For example, we make following ranking data for the author estimation to Figure 3.1.

$$\begin{aligned} (1) &> (2) = (3) = \dots = (6) \\ &= \text{“no author mention (not appear in discourse)”} \\ &= \text{“no author mention (appear in discourse as omission)”} \end{aligned}$$

When the author/reader do not appear in the discourse, we make ranking data where “no author/reader mention (not appear in discourse)” has a higher rank than all discourse entities and “no author/reader mention (appear in discourse as omission)”. For example, we make following ranking data for the author estimation to Figure 3.2.

$$\begin{aligned} \text{“no author mention (not appear in discourse)”} &> (1) = (2) = \dots = (6) \\ &= \text{“no author mention (appear in discourse as omission)”} \end{aligned}$$

When the author/reader appear in the discourse but all mentions are omitted, we make ranking data where “no author/reader mention (appear in discourse as omission)” has a higher rank than all discourse entities and “no author/reader mention (not appear in discourse)”. For example, we make following ranking data for the author estimation to Figure 3.3.

$$\begin{aligned} \text{“no author mention (appear in discourse as omission)”} &> (1) \\ &= \text{“no author mention (not appear in discourse)”} \end{aligned}$$

We judge that the author/reader appear in the discourse if the author/reader appear as a referent of zero reference in gold-standard predicate-argument structures. For example, in Figure 3.3, since the author appear as referents of a **ガ** 2 case of “**気がつけば**” (think) and a **ガ** case of “**ほっとしています**” (feel easy), we can realize that the author appears in the discourse of this document. Meanwhile, this information is used only when making training data and is not used when author/reader mention detection for test data.

For the author/reader mention detection, by using learned decision function, we calculate scores of all discourse entities and the pseudo entities and select the discourse entity with the highest score to the author/reader mention. If any “no author/reader mention” have the highest score, we decide that there are no author/reader mentions in the document.

3.2.3 Lexico-Syntactic Patterns

Each discourse entity is represented as lexico-syntactic (LS) patterns of the discourse entity, its parent and their dependency relation. Here, we consider basic-phrase and clause as an unit for dealing with the LS patterns. It is because that the discourse entity is treating by the basic-phrase, but information of the clause that the basic-phrase belongs to is important.

The LS patterns that represent the discourse entity are what the basic-phrases/clauses of the discourse entity, the basic-phrases/clauses of parent of the discourse entity and their dependency relations are generalized on various levels (types) described below. When a discourse entity is mentioned multiple times, the LS patterns of all mentions are used as the features of the discourse entities. Since

expressions about self-introduction are often used in the first sentences, discriminating the first sentences from other sentences might be effective. Therefore, we deal with the LS patterns that appear in the first sentences as also added features. For example, the features corresponding to discourse entity (1) in Figure 3.1 are what the following elements are generalized.

“basic-phrase:ホテルは”, “parent-basic-phrase:ございます。”, “dependency-basic-phrase:ホテルは → ございます。”, “clause:米子タウンホテルは”, “parent-clause:ございます。”, “dependency-clause:米子タウンホテルは → ございます。”, “basic-phrase:ホテルです。”, “clause:ホテルです。”, “1st-basic-phrase:ホテルは”, “1st-parent-basic-phrase:ございます。”, “1st-dependency-basic-phrase:ホテルは → ございます。”, “1st-clause:米子タウンホテルは”, “1st-parent-clause:ございます。”, “1st-dependency-clause:米子タウンホテルは → ございます。”

Table 3.2 lists generalization types. On the *morphemeA* type, we make a basic-phrase/clause LS pattern by generalizing each morpheme. Meanwhile, only content words are generalized and function words are not generalized on <POS> type.

On the *morphemeB* type, only content words are generalized in the same manner as the <POS> type. When a content word do not have generalized representation on a type (e.g., nominal noun on the <named entity>), the content word is generalized on the <POS> type.

On the *morphemeC* type, words are basically generalized in the same manner as the *morphemeB* type and sometimes are generalized across words. In generalization on <thesaurus> type, when a compound noun that consists of morphemes in the basic-phrase/clause is registered in the thesaurus, the morphemes are generalized as entry of the compound noun. For example, although “ゴルフ場” (golf course) is composed of two morphemes, since the thesaurus has entry “ゴルフ場 = land-use”, “ゴルフ場” is generalized to “land-use”. In the generalization on the <named entity> (NE) type, named entity information that is given to each morpheme is attached with position in a named entity such as “NName:head”, “NName:middle”, “NName:tail” and “NName:single”. In the generalization

Table 3.2: Generalization type and criteria

Type	Generalization unit	Criteria	Example
<word>	<i>morphemeA</i>	No generalizing	basic-phrase:ホテルは → basic-phrase(word):ホテル+は
<original form>	<i>morphemeA</i>	Original form of a morpheme	basic-phrase:ホテルです → basic-phrase(original form):ホテル+だ
<representative form>	<i>morphemeA</i>	JUMAN representative form of a morpheme	basic-phrase:いただきます → basic-phrase(representative form):頂く+ます
<POS>	<i>morphemeA</i>	POS and conjugation of a morpheme	basic-phrase:行きます → basic-phrase(POS):verb(continuative)+ます
<category>(CT)	<i>morphemeB</i>	JUMAN category of a morpheme	basic-phrase:僕の → basic-phrase(CT):CT-Person+の
<first person pronoun>(FPP)	<i>morphemeB</i>	first personal pronoun shown in Table 3.3	basic-phrase:僕の → basic-phrase(FPP):FPP+の
<second person pronoun>(SPP)	<i>morphemeB</i>	second personal pronoun shown in Table 3.3	basic-phrase:皆様には → basic-phrase(SPP):SPP+に+は
<named entity>(NE)	<i>morphemeC</i>	Named entity type	basic-phrase:ヤフーは → basic-phrase(NE):NE-ORG:single+は
<thesaurus>	<i>morphemeC</i>	category of thesaurus [33]	文節:ゴルフ場の → 文節 (thesaurus):land-use+の
<attached word>	<i>basic-phrase/clause</i>	sequence of attached word	basic-phrase:弊社では → basic-phrase(attached word):で+は
<modality>	<i>basic-phrase/clause</i>	Modality given by KNP	clause:紹介していきたい → clause(modality):will
<honorific>	<i>basic-phrase/clause</i>	Honorific expression given by KNP	basic-phrase:お過ごしください → basic-phrase(honorific):respectful

Table 3.3: First person pronoun and second person pronoun

First person pronoun	私 (I), 我々 (we), 僕 (I), 俺 (I), 弊社 (our company), 当社 (our company)
Second person pronoun	あなた (you), 皆様 (you all), 皆さん (you all)

for the clause, the sequence of these are coordinated and generalized as “NE-name”. For example, since “ヤフージャパン株式会社” (Yahoo Japan Corporation) is given the named entity information as “ORGANIZATION:head + ORGANIZATION:middle + ORGANIZATION:middle + ORGANIZATION:tail”, “ヤフージャパン株式会社” is generalized to “ORGANIZATION”.

In these generalization based on morphemes, each morpheme is generalized, and then what each of generalized expressions in a basic-phrase/clause are jointed is LS patterns for the basic-phrase/clause. For example, when “basic-phrase:僕は” is generalized on the <category> (CT) type, “僕” is generalized to “CT-Person” and “は” is not generalized because “は” is a function word. Then, “basic-phrase(category):CT-Person+は”, which is what they are jointed, is generalized expression of this basic-phrase on the <category> type.

We also use generalizations of individual morphemes use as the LS patterns. For example, when “basic-phrase:僕は” is generalized on the <category> type, we use “basic-phrase-morpheme:CT:Person” as a feature in addition to “basic-phrase(category): CT-Person + は.”

On the *basic-phrase/clause* type, each basic-phrase/clause is generalized according to information assigned to the basic-phrase/clause. Therefore the information of the morphemes is not used as the feature.

For “no author/reader mention (not appear in discourse)” instance, the above features of all mentions, including verbs and adjectives, and their dependencies in the document are gathered and used as the features representing the instance.

Table 3.4: Result of the author mention detection

		System output		
		Exist		None
		Correct	Wrong	
Gold	Exist	138	9	124
-standard	None	-	38	691

Table 3.5: Result of the reader mention detection

		System output		
		Exist		None
		Correct	Wrong	
Gold	Exist	50	2	32
-standard	None	-	30	886

3.3 The result of Author/Reader Mention Detection

3.3.1 Experimental Setting

We used 1,000 documents from DDLC and performed 5-fold cross-validation. 271 documents are annotated with author mentions and 84 documents are annotated with reader mentions. We used gold-standard (manually annotated) morphemes, named entities, dependency structures and coreference relations to focus on the author/reader detection. We used SVM^{rank1} for the learning-to-rank method of the author/reader detection. The categories of words are given by the morphological analyzer JUMAN². Predicate features (e.g., honorific expressions, modality) are given by the syntactic parser KNP.³

3.3.2 Results of Author/Reader Mention Detection

We show results of author and reader mention detection in Table 3.4 and Table 3.5. In these tables, “exist” indicates numbers of documents in which the author/reader mentions are manually annotated or our system estimated that

¹http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

²<http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN>

³<http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP>

some discourse entities are author/reader mentions. “None” indicates numbers of documents in which the author/reader mentions are not annotated or our system estimated that there are no author/reader mentions. From these results, the author/reader mentions including “none” can be predicted to accuracies of approximately 80%. On the other hand, the recalls are not particularly high: the recall of author is 0.51 (138/271) and the recall of reader is 0.60 (50/84). This is because documents in which the author/reader mention does not appear are more than the ones in which the author/reader mention appears, and the system preferred to output “no author/reader mention” as results of training.

We show examples of error in Figure 3.4, Figure 3.5 and Figure 3.6. In Figure 3.4, there is no author mention in the corpus, but our system estimated that “山形県山上市” is the author mention. Information of named entity is an important clue for the author mention detection, and named entity that appears in the first sentence is peculiarly tends to be the author mention because the named entity in the first sentences is often mentioned in self-introduction (e.g., The first sentence of Figure 3.1). Additionally, writing style of this document has property of documents that the author appears in the discourse. Because of these reasons, it is considered that our system incorrectly estimated that “山形県山上市” is the author mention.

In Figure 3.5, the author mention is “ジュエリー工房” in the corpus, but our system estimated that there is no author mention. In this example, the expressions used for LS patterns are only “ジュエリー工房だから”, “実現します。” and there are few clues. Additionally, because “ジュエリー工房” is not named entity, it is difficult to estimated that “ジュエリー工房” is the author mention.

In Figure 3.6, there is no reader mention in the corpus, but our system estimated that the reader mention is “お客様”. In this example, “お客様” refers to not the reader of this document but to questioners. Even in such cases, honorific expressions such as “様” and “頂きます” are often used. Since these expressions are often used for also the reader, it would appear that our system estimated that the reader mention is “お客様”, which is a target of these expressions.

フットサル 東北 大会の ために 山形県
 futsal Tohoku tournament-GEN for Yamagata prefecture

山上市に 来ています。
 Yamagami city-DAT come

‘We are coming to Yamagami city, Yamagata prefecture, for a futsal tournament in Tohoku area’

今回の 試合は 3 年生 最後の 大会ですので 全力で
 this match-TOP third-year student last tournament with full effort

戦ってきます。
 compete

‘Because this match is last tournament for third-year students, we will compete with full effort.’

試合 結果など 随時 ブログで 更新します。
 outcome game anytime blog-INS update

‘We will update the blog about outcomes of the game anytime.’

	Corpus	Estimation
Author mention	None	山形県山上市
Reader mention	None	None

Figure 3.4: Example of error of the author mention detection (1)

特許 取得の 手に 馴染んで 装着しやすい オリジナル
 patent acquisition-GEN hand-DAT fit easy to slip original

結婚 指輪が 人気です。
 wedding ring-NOM popular

‘A patented original wedding ring, which fit comfortably in hands and is easy to slip, is popular.’

リングの サイズ 直し/ ピアス/ ネックレス/ ペンダントなどの 宝石の
 ring-GEN size adjust earring necklace pendant -GEN jewel-GEN

修理・加工も 承ります。
 repair process happy to

‘We are happy to adjust ring size and repair and process a jewel of an earring, necklace and pendant.’

デザインから お渡しまで 一貫した ジュエリー 工房だから 高
 from design to delivery consistently jewel because workshop high

品質で 低 価格が 実現します。
 quality low price-GEN achieve

‘We achieve high quality and low price because we are the jewel workshop that consistently carries out from design to delivery.’

	Corpus	Estimation
Author mention	ジュエリー工房	None
Reader mention	None	None

Figure 3.5: Example of error of the author mention detection (2)

メイク ブラシを ご購入頂いた お客様から 「お手入れは どう
makeup brush-ACC bought from customer care-TOP how

すればいいの?」とお問い合わせを 頂きます。

to do inquiry-ACC receive

‘We have received an inquiry, ‘how to care’ from customers who bought the makeup brush.’

そこで 今日から 出来る 簡単な 「日常のお手入れ方法」と
Therefore beginning today can easy daily care method and

「クリーニング 方法」を ご紹介したいと 思います。

cleaning method-ACC introduce will

‘Therefore, I will introduce easy ‘daily care method’ and ‘cleaning method’ that you can begin today’

「クリーニング」は 念のため 2つの 洗浄 方法を
cleaning-TOP just in case two cleaning method-ACC

ご案内しましたので 順番に チェックして下さいね。

introduced in order please check

‘We introduced two cleaning method just in case, please check them in order.’

	Corpus	Estimation
Author mention	None	None
Reader mention	None	お客様

Figure 3.6: Example of error of the reader mention detection (1)

3.4 Summary of this Chapter

In this chapter, we described the author/reader mention detection model. We use a learning-to-rank algorithm to represent relations among the author/reader mention, other discourse entity and absence of the author/reader mentions. We represent each discourse entity as collection of lexico-syntactic patterns. In experiments, our proposed model detects the author/reader mentions in high precisions but low recalls.

Chapter 4

Zero Reference Resolution Considering Exophora and Author/Reader Mentions

In this chapter, we address zero reference resolution. As discussed above chapters, in the zero reference resolution for Web documents, zero exophora and author/reader mentions should be considered, but most of previous studies have not seriously treated them. It would be appear that treating the zero exophora and the author/reader mentions improves both of two subtasks of the zero reference resolution: zero pronoun detection and referent identification. In this chapter, we propose zero reference resolution model that considers the zero exophora and the author/reader mentions.

The rest of this chapter is organized as follows. In Section 4.1, we sort issues about the zero exophora and the author/reader mentions, and then, in Section 4.2, we explain related works about the zero reference resolution. In Section 4.3, we present a baseline model that does not treat the zero exophora and the author/reader mentions. In Section 4.4, we describe the proposed model that considers the zero exophora and the author/reader mentions. In Section 4.5, we report experimental results and discuss about the result. In Section 4.6, we present conclusion of this chapter.

4.1 Zero Reference Resolution

Zero reference resolution is the task of detecting and identifying omitted arguments of a predicate. Since arguments are often omitted in Japanese, zero reference resolution is essential in a wide range of Japanese NLP applications such as information retrieval and machine translation.

- (4.1) パスタが 好きで 毎日 (ϕ ガ) (ϕ ヲ) 食べます。
 pasta-NOM like everyday (ϕ -NOM) (ϕ -ACC) eat
 (Liking pasta, (ϕ) eats (ϕ) every day)

For example, in Example (4.1), ガ (nominative) and ヲ (accusative) arguments of the predicate “食べます” (eat) are omitted.¹ An omitted argument is called a zero pronoun. Here the zero pronoun of the ヲ case refers to “パスタ” (pasta), which is mentioned in this document, while the zero pronoun of the ガ case refers to the author of this document, who is not mentioned explicitly. A zero reference where the referent is mentioned in the document, such as the omission of “パスタ” in Example (4.1), is referred to as **zero endophora**, which was the main focus of previous studies. On the other hand, a zero reference where the referent is not mentioned explicitly in the document, such as the omission of the author in Example (4.1), is called **zero exophora**. Zero exophora often occurs in Japanese when the referent is an author or reader of a document or an indefinite pronoun such as in Example (4.2).

- (4.2) 最近 は パソコンで 動画を ([*unspecified:person*]ガ) 見れる。
 recently PC-INS movie-ACC ([*unspecified:person*]-NOM) can watch
 ‘Recently, (people) can watch movies on a PC.’

In the past, studies of Japanese zero reference resolution have focused mainly on a newspaper article corpus annotated with zero reference relations. Since the aim of newspaper articles is for the author to objectively report events to the reader, the author and reader hardly ever appear in the discourse of a document. On the other hand, in recent years, communication via the Web has become very active, and NLP for Web documents has become increasingly important. In

¹In the following examples, omitted arguments are placed in parentheses and referents not mentioned explicitly are placed in square brackets.

	Zero pronoun	Referent in a document	Example
(a) Zero endophora	Exist	Exist	僕はカフェが好きで毎日 (カフェニ)通っている。 (I like cafes and go (to a cafe) everyday.)
(b) Zero exophora	Exist	Do not exist	私がメリットを ([reader] ニ)説明させていただきます。 (I would like to explain the advantage (to [reader]).)
(c) No zero reference	Do not exist	Do not exist	あなたはリラックスタイムが (×ニ)過ごせる。 (You can have a relaxing time.) *There is no dative case.

Table 4.1: Examples of zero endophora, zero exophora and no zero reference.

Web text, since the author often describes him/herself and reaches out to the reader, the author and reader often appear in the discourse of a document. For example, in blog articles and corporate advertising sites, the author often describes personal events and corporate activities, while on online shopping sites, the author encourages the reader to buy commercial products. Therefore, inevitably many zero references about the author and reader occur, in which many zero exophoric relations are included. In the Web corpus [6], about half the zero references are zero exophora. Hence, in zero reference resolution of Web documents, it is particularly important to deal with the zero exophora.

Most previous studies have ignored zero exophora by assuming zero pronouns do not exist in a sentence. However, such a rough approximation has impeded zero reference resolution research. In this work, to deal with zero exophora explicitly, we provide pseudo entities such as [author], [reader] and [unspecified:person] as candidate referents of zero pronouns. When the case of an argument does not have a superficial argument (e.g., ガ, ヲ, and ニ cases in Example (4.1) and ガ and ニ cases in Example (4.2)), the case is sorted into three types as shown in

Table 4.1. By dealing with the zero exophora, even when there is no referent in the document, it is possible to deal with the phenomenon that the case of a predicate has a zero pronoun as an argument. Thus, the existence of zero pronouns comes to agree with the valency of a predicate and this is expected to improve the accuracy of machine learning based zero pronoun detection.

If a predicate has a zero pronoun as an argument, a referent of the argument is identified. In referent identification, selectional preferences of a predicate [42, 43, 16, 10] and contextual information [13, 14] have been widely used. In addition, in this work, information of the author/reader of a document is used in referent identification. The author and reader of a document have not been used as contextual clues because these rarely appear in the discourse in corpora based on newspaper articles, which were the main target of previous studies. Although the author/reader tends to be omitted, there are many clues for referent identification of the author/reader such as honorific expressions and modality expressions. Therefore, it is important to deal with the author/reader of a document explicitly in referent identification.

Additionally, the author/reader can appear not only as the exophora but also as the endophora.

- (4.3) 私_{author} は もともと アウトドア 派では なかったので、東京に
 I-TOP originally outdoors interest not Tokyo-LOC
 いた 頃も キャンプに 行ったことはありませんでした。
 live when camping-DAT had not gone
 ‘Since I originally did not like the outdoors, even when I lived in Tokyo, I never went camping.’
- (4.4) あなた_{reader} は 今 ある 情報か 資料を 送って、
 you-TOP now exist information or document-ACC send
 アドバイザーからの 質問に 答えるだけ。
 of adviser-GEN question-DAT only answer
 ‘All you need to do is send existing information or a document and answer the questions of the adviser.’

In Example (4.3), “私” (I), which is explicitly mentioned in the document, is the author of the document, and in Example (4.4), “あなた” (you) is the reader. As

explained in the previous chapters, in this study, we call these expressions, which refer to the author and reader, **author mentions** and **reader mentions**, respectively, and treat them explicitly to improve the performance of zero reference resolution. Since the author/reader is mentioned using a variety of expressions besides personal pronouns in Japanese, it is difficult to detect author/reader mentions based merely on lexical information. In this work, we automatically detect author/reader mentions using the method described in Chapter 3.

Once author/reader mentions have been detected, their information is useful for referent identification. Author/reader mentions have the property of a discourse element mentioned in the document and the property of a zero exophoric author/reader.

(4.5) 僕_{author} は 京都に (僕ガ) 行こうと 思っています。
I-TOP Kyoto-DET (I-NOM) will go thought

(I thought I would go to Kyoto.)

皆さん_{reader} は どこに 行きたいか (皆さんガ) (僕二)
you all-TOP where-DET want to go you all-NOM I-DAT
教えてください。
let me know

(Please let me know where you want to go.)

In the first sentence of Example (4.5), the referent of the zero pronoun of the **ガ** case of “行こう” (will go) can be estimated from the contextual clue that “僕” (I) is the topic of the sentence and syntactic clues that “僕” (I) depends on “思っています” (thought) over the predicate “行こう” (will go).² Such contextual clues are available only for discourse entities that are mentioned explicitly. On the other hand, in the second sentence, since “教えてください” (let me know) is a request form, it can be assumed that the referent of the zero pronoun of the **二** case is “僕” (I), which is the author, and the referent of the zero pronoun of the **ガ** case is “皆様” (you all), which are the readers. Clues such as request forms, honorific expressions, and modality expressions are available for both the

²Since “僕” (I) depends on “思っています” (thought), the relation between “僕” (I) and “行こう” (will go) is the zero reference.

author and reader. Additionally, these clues, which are specific to the author and reader, are also available for the author and reader in the zero exophora. In this work, to represent this aspect of author/reader mentions, both the endophora and exophora features are allocated to them.

4.2 Related Work

Several approaches to Japanese zero reference resolution have been proposed and many of them have focused on zero endophora.

Some of zero reference resolution systems have addressed only referent identification assuming that zero pronouns are known. Iida et al. [13] proposed a zero reference resolution model that uses the syntactic relations between a zero pronoun and a candidate referent as a feature. They deal with zero exophora by judging that a zero pronoun does not have anaphoricity. It can be said that this study has distinguished between “(a) Zero endophora” and “(b) Zero exophora” in Table 4.1 but has not treated “(c) No zero reference.” Isozaki et al. [18] proposed a referent identification model that has used a learning-to-rank algorithm. Zero pronouns dealt with this study are limited to zero pronouns whose referents appear in a document. In other words, this study has treated only “(a) Zero endophora” in Table 4.1.

Zero reference resolution has been often tackled as a part of predicate-argument structure analysis. Taira et al. [45], Imamura et al. [16] and Hayashibe et al. [10] have addressed the predicate-argument structure analysis independently for each case. Taira et al. [45] proposed a predicate-argument structure analysis model using decision lists. They treated words in a document as arguments. Imamura et al. [16] proposed a predicate-argument structure analysis model based on a log-linear model that simultaneously conducts zero endophora resolution. They assumed a particular candidate referent, NULL, and when the analyzer selected this referent, the analyzer outputs “zero exophora or no zero pronoun.” Hayashibe et al. [10] proposed a predicate-argument structure analysis model based on a tournament model using features such as co-occurrence between a predicate and an argument. This study also has not distinguished zero exophora and absence of a

zero pronoun, and has tackled only 力 (nominative) case. Sasano et al. [42, 43] proposed a predicate-argument analysis model that comprehensively analyzes all cases of a predicate. They proposed a probabilistic predicate-argument structure analysis model including zero endophora resolution by using wide-coverage case frames constructed from a web corpus in 2008 [42]. They extended the probabilistic model by focusing on zero endophora in 2011 [43]. Their model is based on a log-linear model that uses case frame information and the location of a candidate referent as features. In their work, zero exophora is not treated and they assumed that a zero pronoun is absent when there is no referent in a document. It can be said that these studies have not distinguished between “(b) Zero exophora” and “(c) No zero reference.”

Taira et al. [46] and Hattori and Harada [9] addressed zero exophora. They have treated zero reference resolution including the zero exophora as a part of predicate-argument analysis for newspaper articles. However, they have not treated relations between the author/reader in zero exophora and the author/reader in zero endophora (author/reader mention in our study). Additionally, Taira et al. reported that there are few zero exophoric relations in a newspaper corpus.

For languages other than Japanese, zero pronoun resolution methods have been proposed for Chinese, Portuguese, Spanish and other languages. In Chinese, Kong and Zhou [26] proposed tree-kernel based models for three subtasks: zero pronoun detection, anaphoricity decision and referent selection. In Portuguese and Spanish, only a subject word is omitted and zero pronoun resolution has been tackled as a part of coreference resolution. Poesio et al. [37] and Rello et al. [38] detected omitted subjects and made a decision whether the omitted subject has anaphoricity or not as preprocessing of coreference resolution systems.

In English, semantic role labeling, which is similar to zero reference resolution, has been tackled. Gerber and Chai [5] annotated semantic role including implicit arguments, which do not have dependency relation in for predicates, for frequent nominal predicates and built automatically identification system of the implicit arguments. Ruppenhofer et al. [41] treated omitted arguments as a part of semantic role labeling and distinguished between arguments whose referents are identified (Definite Null Instance) and arguments whose referents are not identified

(Indefinite Null Instance).

4.3 Baseline Model

In this section, we describe a baseline zero reference resolution system. In our model, the zero reference resolution is conducted as a part of predicate-argument structure (PAS) analysis for each predicate. The PAS analysis for each predicate can capture relations between a predicate and more than one argument. For example, in zero reference resolution of a **ガ** case in “(不動産屋**ガ**) 物件を紹介する” ((estate agent) introduce properties), it is a good clue that an argument of a **ヲ** case is “物件” (properties).

The PAS consists of a case frame and an alignment between case slots and referents. The case frames are constructed for each meaning of a predicate. Each case frame describes surface cases that each predicate has (case slot) and words that can fill each case slot (example). In this study, the case frames are constructed from 6.9 billion Web sentences by using Kawahara et al. [22]’s method. We show the examples of constructed case frames in Figure 4.1.³

In our model, we treat referents of zero pronouns using a unit called **discourse entity**, which is what mentions in a coreference chain are bound into. In Figure 4.1, we treat “僕” (I) and “自分” (oneself), which are in a coreference chain, as one discourse entity. Similarly, “ラーメン屋₁” (noodle shop), “その店” (that shop) and “ラーメン屋₂” are treated as one discourse entity. In Figure 4.1, the discourse entity (a), which corresponds to “僕,” is selected for the referent of a **ガ** case of the predicate “紹介します” (will introduce).

We show example of the PAS analysis in Figure 4.1.⁴ In this example, “紹介する (1)” and “紹介する (2)” are case frames corresponding to each meaning of “紹介する”. “紹介する (1)” is selected from case frames that correspond to “紹介します,” and a **ガ** case, a **ヲ** case and a TIME case of the case frame are respectively

³<TIME> means what time expressions such as “今日” (today) and “3時” (3 o’clock) are generalized into.

⁴Referents of each case slot are actually selected from discourse entities but are attached with a representative word for illustration. “Null” indicates that a case slot is not assigned to any discourse entities.

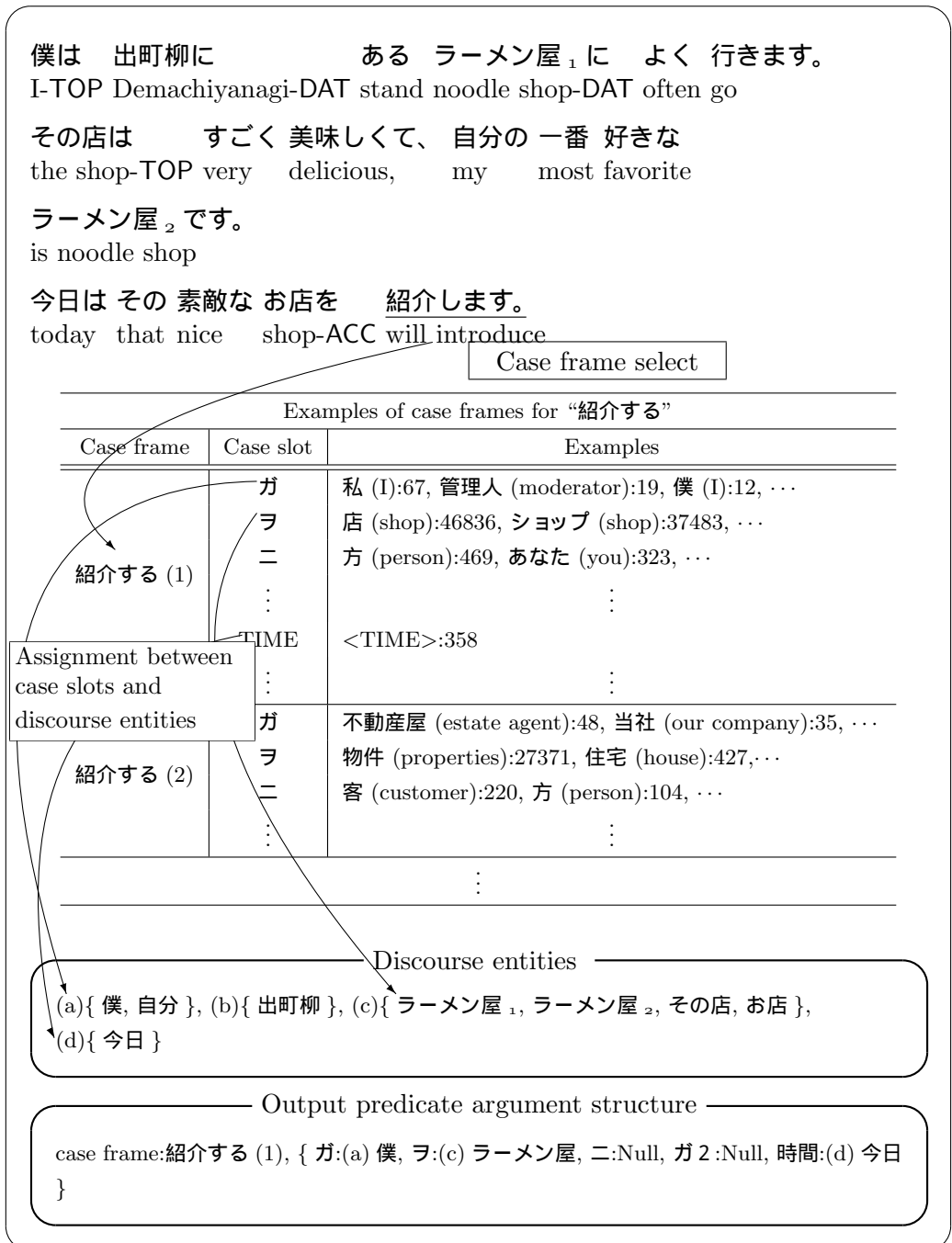


Figure 4.1: Outline of zero reference resolution

assigned to discourse entity (a), (c) and (d), and other cases are not assigned to any discourse entities.

The baseline model does not treat zero exophora as the previous studies. The baseline model analyzes a document in the following procedure in the same way as a previous study [43].⁵

1. Parse the input document and recognize named entities.
2. Resolve coreferential relations and set discourse entities.
3. Analyze the PAS for each predicate using the following steps:
 - (a) Generate candidate PASs.
 - i. Select one case frame from ones that correspond to the target predicate.
 - ii. Assign discourse entities that have a dependency relation with the target predicate to case slots of the case frame.
 - iii. Assign the remaining case slots to the remaining discourse entities.
 - (b) Calculate the score of each PAS and select the PAS with the highest score.

We illustrate the PAS analysis in Step 3. In Step 3a, possible combinations of a case frame (*cf*) and an alignment (*a*) between case slots and discourse entities are listed. First, one case frame is selected from case frames for the predicate (in Step 3(a)i). Next, overt arguments, which have dependency relations with the predicate, are aligned to case slots of the case frame (in Step 3(a)ii). Finally, each of zero pronouns of remaining case slots is assigned to a discourse entity or is not assigned to any discourse entity (in Step 3(a)iii). In this step, because of a heuristic that a discourse entity is not assigned to multiple cases of a predicate [32], discourse entities that have been already assigned to any case slots are not assigned to zero pronouns. A case slot whose zero pronoun is not assigned to any

⁵For learning, the previous study used a log-linear model, but we use a learning-to-rank model. In our preliminary experiment of the baseline model, there is little difference between the results of these methods.

discourse entity corresponds to the case that does not have a zero pronoun. In Figure 4.2, we show examples of candidate PASs. Since alignments between case slots and discourse entities of the PAS [1-2] and [2-2] are the same but their case frames are different, we deal with them as discrete PASs. In this case, however, the results of zero reference resolution are the same.

We represent each PAS as a feature vector, which is described in Section 4.3.1, and calculate a score of each PAS with the learned weights. Finally, the system outputs the PAS with the highest score.

4.3.1 Feature Representation of Predicate-Argument Structure

In this section, we illustrate a feature vector that represents a PAS. When text t and target predicate p are given and PAS (cf, a) is chosen, we represent a feature vector of the PAS as $\phi(cf, a, p, t)$. $\phi(cf, a, p, t)$ consists of a feature vector $\phi_{overt-PAS}(cf, a_{overt}, p, t)$ and feature vectors $\phi(cf, c \leftarrow e, p, t)$. Where $\phi_{overt-PAS}(cf, a_{overt}, p, t)$ corresponds to overt alignment a_{overt} , which is alignment between case slots and overt (not omitted) arguments, and $\phi(cf, c \leftarrow e, p, t)$ represents that a case slot c is assigned to a discourse entity e . Specifically, the $\phi(cf, a, p, t)$ is represented as the following equation.

$$\begin{aligned} \phi(cf, a, p, t) = & (\phi_{overt-PAS}(cf, a_{overt}, p, t), \\ & \phi_{case}(cf, ガ \leftarrow eガ, p, t), \phi_{case}(cf, ヲ \leftarrow eヲ, p, t), \\ & \phi_{case}(cf, ニ \leftarrow eニ, p, t), \phi_{case}(cf, ガ2 \leftarrow eガ2, p, t)) \end{aligned} \quad (4.1)$$

Each feature vector $\phi(cf, c \leftarrow e, p, t)$ consists of $\phi_A(cf, c \leftarrow e, p, t)$ and $\phi_{NA}(cf, c \leftarrow \text{Null}, p, t)$. $\phi_A(cf, c \leftarrow e, p, t)$ becomes active when the case slot c is assigned to the discourse entity e and $\phi_{NA}(cf, c \leftarrow \text{Null}, p, t)$ becomes active when the case slot c is not assigned to any discourse entities. When a case slot is assigned to an overt entity, $\phi(cf, c \leftarrow e, p, t)$ is set to a zero vector. For example, the feature vector $\phi(\text{紹介する}(2), \{ガ : (a) 僕, ヲ : (c) ラーメン屋, ニ : \text{Null}, ガ2 : \text{Null}, 時間 : (d) 今日\})$, which represents the PAS [2-2] in Figure 4.2,

- [1-1] case frame:紹介する (1), { ガ:Null, ヲ:(c) ラーメン屋, ニ:Null, ガ 2 :Null, TIME:(d) 今日 }
- [1-2] case frame:紹介する (1), { ガ:(a) 僕, ヲ:(c) ラーメン屋, ニ:Null, ガ 2 :Null, TIME:(d) 今日 }
- [1-3] case frame:紹介する (1), { ガ:(b) 出町柳, ヲ:(c) ラーメン屋, ニ:Null, ガ 2 :Null, TIME:(d) 今日 }
- [1-4] case frame:紹介する (1), { ガ:Null, ヲ:(c) ラーメン屋, ニ:(a) 僕, ガ 2 :Null, TIME:(d) 今日 }
- [1-5] case frame:紹介する (1), { ガ:(a) 僕, ヲ:(c) ラーメン屋, ニ:(b) 出町柳, ガ 2 :Null, TIME:(d) 今日 }
- ⋮
- [2-1] case frame:紹介する (2), { ガ:Null, ヲ:(c) ラーメン屋, ニ:Null, ガ 2 :Null, TIME:(d) 今日 }
- [2-2] case frame:紹介する (2), { ガ:(a) 僕, ヲ:(c) ラーメン屋, ニ:Null, ガ 2 :Null, TIME:(d) 今日 }
- [2-3] case frame:紹介する (2), { ガ:(b) 出町柳, ヲ:(c) ラーメン屋, ニ:Null, ガ 2 :Null, TIME:(d) 今日 }
- [2-4] case frame:紹介する (2), { ガ:Null, ヲ:(c) ラーメン屋, ニ:(a) 僕, ガ 2 :Null, TIME:(d) 今日 }
- [2-5] case frame:紹介する (2), { ガ:(a) 僕, ヲ:(c) ラーメン屋, ニ:(b) 出町柳, ガ 2 :Null, TIME:(d) 今日 }
- ⋮

Figure 4.2: Candidate predicate-argument structures of “紹介します” in the base-line model

Table 4.2: The features for a case that is assigned to a distances entity

Type	Value	Description
Case frame	Log	Probabilities that {words, categories and named entity types} of e is assigned to c of cf
	Log	Generative probabilities of {words, categories and named entity types} of e
	Log	PMIs between {words, categories and named entity types} of e and c of cf
	Log	Max of PMIs between {words, categories and named entity types} of e and c of cf
	Log	Probability that c of cf is assigned to any words
	Log	Ratio of examples of c to ones of cf
	Binary	c of cf is {adjacent and obligate} case
Predicate	Binary	Modality types of p
	Binary	Honorific expressions of p
	Binary	Tenses of p
	Binary	p is potential form
	Binary	Modifier of p (predicate, noun and end of sentence)
	Binary	p is {dynamic and stative} verb
Context	Binary	Named entity types of e
	Integer	Number of mentions about e in t
	Integer	Number of mentions about e {before and after} p in t
	Binary	e is mentioned with post position “は” in a target sentence
	Binary	Sentence distances between e and p
	Binary	Location categories of e [43]
	Binary	e is mentioned at head of a target sentence
	Binary	e is mentioned with post position {“は” and “か”} at head of a target sentence
	Binary	e is mentioned at head of the first sentence
	Binary	e is mentioned with post position “は” at head of the first sentence
	Binary	e is mentioned at end of the first sentence
	Binary	e is mentioned with copula at end of the first sentence
	Binary	e is mentioned with noun phrase stop at end of the first sentence
Binary	Saliency score of e is larger than 1 [43]	
Others	Binary	c is assigned

is the following.⁶

$$\begin{aligned}
\phi(\text{紹介する (2)}, \{ \text{ガ} : (a) \text{ 僕}, \text{ヲ} : (c) \text{ ラーメン屋}, \text{ニ} : \text{Null}, \text{ガ} 2 : \text{Null}, \text{時間} : (d) \text{ 今日} \}) = \\
& (\phi_{\text{overt-PAS}}(\text{紹介する (2)}, \{ \text{ガ} : \text{Null}, \text{ヲ} : (d) \text{ ラーメン屋}, \text{ニ} : \text{Null}, \\
& \qquad \qquad \qquad \text{ガ} 2 : \text{Null}, \text{時間} : (c) \text{ 今日} \}), \qquad \qquad \qquad (4.2) \\
& \phi_A(\text{紹介する (2)}, \text{ガ} \leftarrow (a) \text{ 僕}), \qquad \qquad \qquad \mathbf{0}_{\phi_{NA}}, \\
& \mathbf{0}_{\phi_A}, \qquad \qquad \qquad \mathbf{0}_{\phi_{NA}}, \\
& \mathbf{0}_{\phi_A}, \qquad \qquad \qquad \phi_{NA}(\text{紹介する (2)}, \text{ニ} \leftarrow \text{Null}), \\
& \mathbf{0}_{\phi_A}, \qquad \qquad \qquad \phi_{NA}(\text{紹介する (2)}, \text{ガ} 2 \leftarrow \text{Null}))
\end{aligned}$$

We present the details of $\phi_{\text{overt-PAS}}(cf, a, p, t)$, $\phi_A(cf, c \leftarrow e, p, t)$ and $\phi_{NA}(cf, c \leftarrow \text{Null}, p, t)$. We use a score of the probabilistic PAS analysis [23] to $\phi_{\text{overt-PAS}}(cf, a_{\text{overt}}, p, t)$. We list the features of $\phi_A(cf, c \leftarrow e, p, t)$ in Table 4.2.⁷ “Case frame” features are informations from the case frames. When e is mentioned more than once, the largest value of values that correspond to each mention is used for the value of each feature. For example, we think about feature, probability that a discourse entity e is assigned to a case c of a case frame cf , in the ガ case of Equation (4.2). In above example, the discourse entity (a), which is assigned to the ガ case, is twice mentioned as “僕” and “自分.” Therefore, we calculate each probability that “僕” and “自分” is assigned to the ガ case of the “紹介する (2)”, and the highest probability is treated as the probability that the discourse entity (a) is assigned to the ガ case of the “紹介する (2).” “Predicate” features are informations that are given by Japanese dependency parser KNP.⁸ “Context” features are the informations that e appears in context, and when e appears more than once, all of the appearances are used for the features. A feature that c is assigned to any discourse entities is a feature that controls tendency that c is assigned. We list the features of $\phi_{NA}(cf, c \leftarrow \text{Null}, p, t)$ in Table 4.3. Since there

⁶In the following example, p and t are sometimes omitted, and $\mathbf{0}_{\phi}$ is 0 vector that has the same dimension as ϕ .

⁷In “value” column, “Log” means that logarithmic value is used for the feature. “Binary” means that binary representation of multi value is used for the feature. “Int” is used just value for the feature.

⁸<http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP>

Table 4.3: The features for a case that is not assigned to any discourse entities

Type	Value	Description
Case frame	Log	Probability that c of cf is not assigned
	Log	Ratio of number of examples of c to ones of cf
	Binary	c of cf is {adjacent and obligate} case

is not assigned e in $\phi_{NA}(cf, c \leftarrow Null, p, t)$, only the “Case frame” features are used for $\phi_{NA}(cf, c \leftarrow Null, p, t)$.

4.3.2 Weight Learning

In the previous section, we defined the feature vector $\phi(cf, a, p, t)$, which represents a PAS. In this section, we illustrate a learning method of a weight vector corresponding to the feature vector. The weight vector is learned by using a learning-to-rank algorithm [11, 19].

We make a ranking data of each predicate by the following method. Then all of the ranking data are merged, and the merged data is fed into the learning-to-rank algorithm. If correct PAS were defined uniquely, we should make ranking data that the correct PAS has higher rank than other candidate PASs. However, there are the following two problems.

The first problem is that there are predicates whose case is annotated with more than one argument in gold-standard corpus. For example, “焼いている” (bake) in Figure 4.3 is annotated with { ガ:Null, ヲ:(b) ケーキ+(c) クッキー, ニ:Null, ガ2:Null, TIME:(d) 毎週 }, and the ヲ case is annotated with two arguments. On the other hand, described above, the proposed method assigns one case to only one discourse entity. Therefore, when cases are annotated with multiple discourse entities, we treat alignments that the cases are assigned to one of the annotated discourse entities as correct alignments. For example, in Figure 4.3, { ガ:Null, ヲ:(b) ケーキ, ニ:Null, ガ2:Null, TIME:(d) 毎週 } and { ガ:Null,

ヲ:(c) クッキー, ニ:Null, ガ2:Null, TIME:(d) 毎週 } are treated as the correct alignments. And, the set of the correct alignments is defined as (a^*_1, \dots, a^*_N) .

The second problem is that case frames are not annotated in a corpus. Since the case frames are constructed for each meaning, some of them are unsuitable for a usage of a predicate in a context (e.g., idiomatic usage). If training data includes PASs (cf, a^*) whose cf is such case frame as correct instances, these are harmful for training. Hence, we treat a case frame cf^* which is selected by a heuristic method as a correct case frame and remove (cf, a^*) which has other cf .

In particular, we make a ranking data for learning in each target predicate p in the following steps.

1. List possible PASs (cf, a) for predicate p .
2. For correct alignments a^*_1, \dots, a^*_N
 - (a) Calculate a probabilistic zero reference resolution score [42] for each PAS (cf, a^*_i) and define the PAS with highest score as (cf^*_i, a^*_i) .
 - (b) Remove (cf, a^*_i) except (cf^*_i, a^*_i) from the learning instance.
3. Make a ranking data that $(cf^*_1, a^*_1), \dots, (cf^*_N, a^*_N)$ have higher ranks than other (cf, a) .

In the above steps, we make the ranking data for each predicate and use ranking data collected from all target predicates as training data.

We illustrate this method with a concrete example of Figure 4.3. Firstly, the [1-1], \dots , [2-1], \dots are listed as candidate PASs for “焼いている” (in Step 1). In these PASs, PASs (cf, a^*_i) , whose alignments are correct, are [1-2] and [2-2], which correspond to { ガ:Null, ヲ:(b) ケーキ, ニ:Null, ガ2:Null, 時間:(d) 毎週 } and [1-3] and [2-3], which correspond to { ガ:Null, ヲ:(c) クッキー, ニ:Null, ガ2:Null, 時間:(d) 毎週 }. Then, we calculate the probabilistic zero reference resolution scores of [1-2], [2-2], [1-3] and [2-3], and we assume that score of [1-2] is larger than one of [2-2] and one of [1-3] is larger than one of [2-3]. In this case, (cf, a^*_i) are [1-2] and [1-3] (in Step 2a). Then, [2-2] and [2-3] are removed from training instances (in Step 2b). Finally, we make the ranking data, $[1-2] = [1-3] > [1-1] = [1-4] = \dots = [2-1] = [2-4] = \dots$, for the training

チョコのケーキやクッキーが好きで、毎週 焼いている。

— Discourse entities —

(a){ チョコ }, (b){ ケーキ }, (c){ クッキー }, (d){ 毎週 }

— Candidate predicate-argument structures —

- [1-1] case frame:焼く (1), { ガ:Null, ヲ:Null, ニ:Null, ガ 2 :Null, TIME:(d) 毎週 }
- [1-2] case frame:焼く (1), { ガ:Null, ヲ:(b) ケーキ, ニ:Null, ガ 2 :Null, TIME:(d) 毎週 }
}
- [1-3] case frame:焼く (1), { ガ:Null, ヲ:(c) クッキー, ニ:Null, ガ 2 :Null, TIME:(d) 毎週 }
}
- [1-4] case frame:焼く (1), { ガ:(b) ケーキ, ヲ:(c) クッキー, ニ:Null, ガ 2 :Null, TIME:(d) 毎週 }
毎週 }
⋮
- [2-1] case frame:焼く (2), { ガ:Null, ヲ:Null, ニ:Null, ガ 2 :Null, TIME:(d) 毎週 }
- [2-2] case frame:焼く (2), { ガ:Null, ヲ:(b) ケーキ, ニ:Null, ガ 2 :Null, TIME:(d) 毎週 }
}
- [2-3] case frame:焼く (2), { ガ:Null, ヲ:(c) クッキー, ニ:Null, ガ 2 :Null, TIME:(d) 毎週 }
}
- [2-4] case frame:焼く (2), { ガ:(b) ケーキ, ヲ:(c) クッキー, ニ:Null, ガ 2 :Null, TIME:(d) 毎週 }
毎週 }
⋮

Figure 4.3: Example of case that one case slot is assigned to multiple arguments

data of “*焼いている*” (in Step 3). We make ranking datas for each predicate and a weight vector is learned by using the training data that is what the ranking datas are merged.

4.4 Zero Reference Resolution Considering Exophora and Author/Reader Mentions

In this section, we describe a zero reference resolution model that considers zero exophora and author/reader mentions. The proposed model resolves zero reference as a part of PAS analysis based on a baseline model.

The proposed model analyzes the PASs in the following steps:

1. Parse an input document and recognize named entities.
2. Resolve coreferential relations and set discourse entities.
3. Detect author/reader mentions of the document.
4. Set pseudo entities from the estimated author/reader mentions.
5. Analyze a predicate-argument structure for each predicate using the following steps:
 - (a) Generate candidate predicate-argument structures.
 - i. Select one case frame from ones that correspond to the target predicate.
 - ii. Assign words that have a dependency relation with the target predicate to case slots of the case frame.
 - iii. Assign the remaining case slots to the remaining discourse entities.
 - (b) Calculate the score of each predicate-argument structure and select the one with the highest score.¹

Differences from the baseline model are the estimation of the author/reader mentions in Step 3 and setting of pseudo entities in Step 4.

4.4.1 Pseudo Entities and Author/Reader Mentions for Zero Exophora

In the baseline model, referents of zero pronouns are selected from discourse entities, which correspond to zero endophora. The proposed model assumes pseudo entities([author], [reader], [US:person] (unspecified:person) and [US:others] (unspecified:others)⁹) to deal with zero exophora. In Example 4.6, a ガ case and a ニ case of “説明します” (will introduce) are respectively assigned to [author] and [reader].

- (4.6) 今日はお得なポイントカードについて ([author] ガ) ([reader] ニ)
 today value point card-DAT about [author]-NOM [reader]-DAT
説明します。
 will introduce

‘Today, I will introduce about a value point card.’

We add these pseudo entities to candidate referents, and when these pseudo entities are selected as a referent of a case, we deal with the case as zero exophora.

When author/reader mentions appear in a document, the author/reader pseudo entities raise an issue. In Example (4.7), a referent of a zero pronoun ϕ can be interpreted as both “私” (I) and [author].

- (4.7) 肩こりや腰痛で来院された患者さんに 対し、
 stiff shoulders backache-INS come to hospital patient-DAT for
私_{author} は脈を診ることにしています。
 I-TOP pulse feel

‘For a patient who comes to the hospital cause of stiff shoulders and backache, I feel to the pulse.’

それは心臓の状態を (ϕ ガ) 診ているだけではなく、
 because hart-GEN condition-ACC (ϕ -NOM) examine
 身体全体のバランスを (ϕ ガ) 診たいからです。
 body entire-GEN balance-ACC (ϕ -NOM) want to examine

⁹We merge [US:matter] and [US:situation] because of the small amount of [US:situation] in a corpus.

- [1-1] case frame:紹介する (1), { ガ:Null, ヲ:(c) ラーメン屋, ニ:Null, ガ 2 :Null, TIME:(d) 今日 }
- [1-2] case frame:紹介する (1), { ガ:(a) 僕, ヲ:(c) ラーメン屋, ニ:Null, ガ 2 :Null, TIME:(d) 今日 }
- [1-3] case frame:紹介する (1), { ガ:[reader], ヲ:(c) ラーメン屋, ニ:Null, ガ 2 :Null, TIME:(d) 今日 }
- [1-4] case frame:紹介する (1), { ガ:[US-person], ヲ:(c) ラーメン屋, ニ:Null, ガ 2 :Null, TIME:(d) 今日 }
- [1-5] case frame:紹介する (1), { ガ:(a) 僕, ヲ:(c) ラーメン屋, ニ:[reader], ガ 2 :Null, TIME:(d) 今日 }
- [1-6] case frame:紹介する (1), { ガ:[reader], ヲ:(c) ラーメン屋, ニ:(a) 僕, ガ 2 :Null, TIME:(d) 今日 }
- ⋮
- [2-1] case frame:紹介する (2), { ガ:Null, ヲ:(c) ラーメン屋, ニ:Null, ガ 2 :Null, TIME:(d) 今日 }
- [2-2] case frame:紹介する (2), { ガ:(a) 僕, ヲ:(c) ラーメン屋, ニ:Null, ガ 2 :Null, TIME:(d) 今日 }
- ⋮

Figure 4.4: Candidate predicate-argument structures of “紹介します” in the proposed model

‘It is because that I want to examine not only condition of a hart also balance of entire body.’

In this work, to remove such ambiguities, the author/reader mentions are given priority over [author] and [reader]. In this example, a referent of the zero pronoun is “私.” In analyzing process, when there are the author/reader mentions, [author] and [reader] are not assigned to any cases as referents. For example, in Figure 4.1, since there is an author mention, “僕,” [author] is not assigned to any cases. Candidate PASs of “紹介します” in Figure 4.1 are shown in Figure 4.4.

Table 4.4: Expressions and categories for pseudo entities

	Expressions	Categories
author	私 (I), 我々 (we), 俺 (I), 僕 (I), 当社 (our company), 弊社 (our company), 当店 (our shop)	PERSON, ORGANIZATION
reader	あなた (you), 客 (customer), 君 (you), 皆様 (you all), 皆さん (you all), 方 (person), 方々 (people)	PERSON
US:person	人 (person), 人々 (people)	PERSON
US:others	もの (thing), 状況 (situation)	all categories except PERSON and ORGANIZATION

Meanwhile, the author/reader mentions behave similarly to the [author] and [reader] pseudo entities in the discourse.¹⁰ Therefore, we discriminate the author/reader mentions from other discourse entities and give the features to have behavior of the [author] and [reader] pseudo entities. The details are described in Section 4.4.2.

4.4.2 Feature Representation of Predicate Argument Structure

In the same way as the baseline model, the proposed model represents a PAS as a feature vector that consists of a feature vector $\phi_{overt-PAS}(cf, a, p, t)$ and feature vectors $\phi_{case}(cf, c \leftarrow e, p, t)$, which consist of $\phi_A(cf, c \leftarrow e, p, t)$ and $\phi_{NA}(cf, c \leftarrow Null, p, t)$. The difference from the baseline model is a composition of $\phi_A(cf, c \leftarrow e, p, t)$.¹¹ In the proposed model, each $\phi_A(cf, c \leftarrow e)$ is composed of vectors, $\phi_{discourse}(cf, c \leftarrow e)$, $\phi_{[author]}(cf, c \leftarrow e)$, $\phi_{[reader]}(cf, c \leftarrow e)$, $\phi_{[US:person]}(cf, c \leftarrow e)$, $\phi_{[US:others]}(cf, c \leftarrow e)$ and $\phi_{max}(cf, c \leftarrow e)$. Their contents and dimensions are the same and similar to $\phi_A(cf, c \leftarrow e)$ of the baseline model the except for addition of a few features described in section 4.4.3.

¹⁰For example, both the author mention and [author] tend to be an agent of a modest expression.

¹¹In the following equations, p and t are omitted.

$\phi_{discourse}$ corresponds to discourse entities, which are mentioned explicitly in a document, and becomes active when e is a discourse entity including the author/reader mentions. $\phi_{discourse}$ is almost the same as ϕ_A of the baseline model and the difference is explained in section 4.4.3. $\phi_{[author]}$ and $\phi_{[reader]}$ become active when e is [author]/[reader] or the discourse entity corresponding to the author/reader mention. In particular, when e is the discourse entity corresponding to the author/reader mention, both $\phi_{discourse}$ and $\phi_{[author]}/\phi_{[reader]}$ become active. This representation gives the author/reader mentions the properties of the discourse entity and the author/reader. $\phi_{[US:person]}$ and $\phi_{[US:others]}$ become active when e is [US:person] and [US:others]. Because $\phi_{[author]}$, $\phi_{[reader]}$, $\phi_{[US:person]}$ and $\phi_{[US:others]}$ correspond to the pseudo entities, which are not mentioned explicitly, we cannot use word information such as expressions and categories. We assume that the pseudo entities have expressions and categories shown in Table 4.4 and use these to calculate case frame features. Finally, ϕ_{max} consists of the highest value of correspondent feature of the above feature vectors.

We explain each case of $\phi_A(cf, c \leftarrow e, p, t)$ of candidate PAS [1-5] in Figure 4.4.

$$\begin{aligned} \phi_A(cf, \text{力} \leftarrow (a) \text{僕}, p, t) = & (\phi_{discourse}(cf, \text{力} \leftarrow (a) \text{僕}, p, t), \\ & \phi_{[author]}(cf, \text{力} \leftarrow (a) \text{僕}, p, t), \\ & \mathbf{0}_{\phi_{[reader]}}, \mathbf{0}_{\phi_{[US:person]}}, \mathbf{0}_{\phi_{[US:others]}}), \\ & \max(\phi_{mentioned}(cf, \text{力} \leftarrow (a) \text{僕}, p, t), \\ & \phi_{[author]}(cf, \text{力} \leftarrow (a) \text{僕}, y, p, t)) \end{aligned}$$

In a 力 case, since “僕” is mentioned explicitly in the document, $\phi_{discourse}(cf, \text{力} \leftarrow (a) \text{僕}, p, t)$ becomes active, and since the discourse entity (a) corresponds to the author mention, also $\phi_{[author]}(cf, \text{力} \leftarrow (a) \text{僕}, p, t)$ becomes active. $\phi_{[reader]}(cf, \text{力} \leftarrow e, p, t)$, $\phi_{[US:person]}(cf, \text{力} \leftarrow e, p, t)$, and $\phi_{[US:others]}(cf, \text{力} \leftarrow e, p, t)$ do not become active and are set 0 vectors. Each value in $\phi_{max}(cf, \text{力} \leftarrow e, p, t)$ is larger value of corresponding features of $\phi_{discourse}(cf, \text{力} \leftarrow (a) \text{僕}, p, t)$ and $\phi_{[author]}(cf, \text{力} \leftarrow (a) \text{僕}, p, t)$.

$$\begin{aligned} \phi_A(cf, \equiv \leftarrow [reader], p, t) = & (\mathbf{0}_{\phi_{mentioned}}, \mathbf{0}_{\phi_{[author]}}, \\ & \phi_{[reader]}(cf, \equiv \leftarrow [reader], p, t), \mathbf{0}_{\phi_{[US:person]}}, \\ & \mathbf{0}_{\phi_{[US:others]}}, \\ & \max(\phi_{[reader]}(cf, \equiv \leftarrow [reader], p, t)) \end{aligned}$$

Since a \equiv case is assigned to [reader], which is not explicitly mentioned, only $\phi_{[reader]}(cf, \equiv \leftarrow [reader], p, t)$ become active and $\phi_{max}(cf, \equiv \leftarrow [reader], p, t)$ is same as $\phi_{[reader]}(cf, \equiv \leftarrow [reader], p, t)$. Since a \exists case is assigned to “ラ－メソノ屋”, which has direct dependency relation to the predicate, both $\phi_A(cf, \exists \leftarrow e, p, t)$ and $\phi_{NA}(cf, \exists \leftarrow \text{Null}, p, t)$ are set 0 vectors as with the baseline model. Since a \forall case is not assigned to any discourse entity, $\phi_{NA}(cf, \forall \leftarrow \text{Null}, p, t)$ become active and $\phi_A(cf, \forall \leftarrow e, p, t)$ is set 0 vector as with the baseline model.

4.4.3 Author/Reader Mention Score

We add author/reader mention score features to feature vector $\phi_A(cf, c \leftarrow e, p, t)$ described in Table 4.2. The author/reader mention scores are the discriminant function scores of the author/reader mention detection. When e is the author/reader mention, we set the author/reader mention score to the feature.

4.5 Experiments

4.5.1 Experimental Settings

We used 1,000 documents from DDLC and performed 5-fold cross-validation. 1,539 cases and 2,072 cases are annotated with zero endophora and zero exophora respectively in these documents. 271 documents are annotated with author mentions and 84 documents are annotated with reader mentions. We used gold-standard (manually annotated) morphemes, named entities, dependency structures and coreference relations to focus on author/reader detection and zero reference resolution. We used SVM^{rank12} for the learning-to-rank method of the

¹²http://www.cs.cornell.edu/people/tj/svm-light/svm_rank.html

Table 4.5: Results of zero endophora resolution

	Recall	Precision	F1
Baseline	0.270	0.370	0.312
Proposed model (estimate)	0.298	0.447	0.357
Proposed model (gold-standard)	0.411	0.536	0.465

author/reader detection and the PAS analysis. The categories of words are given by the morphological analyzer JUMAN¹³. Predicate features (e.g., honorific expressions, modality) are given by the syntactic parser KNP.¹⁴

We compared three model, “Baseline”, “Proposed model (estimate)” and “Proposed model (gold-standard).” “Baseline” is a model that does not consider author/reader mentions and zero exophora, described in Section 4.3. “Proposed model (estimate)” is a proposed model, described in 4.4, that estimated the author/reader mentions by the method shown in Chapter 3 and “Proposed model (gold-standard)” the proposed model that is given the author/reader mentions of gold-standard from the corpus. Outputs are evaluated by each case of each predicate, and when a case is annotated with multiple arguments, we deal with an output that matches one of the arguments as correct.

4.5.2 Results of Zero Reference Resolution

We show the results of zero reference resolution in Table 4.5 and Table 4.6. The difference between the baseline and the proposed model is statistically significant ($p < 0.05$) from the McNemar’s test. In Table 4.5, we evaluate only the zero endophora for comparison to the baseline model, which deals with only the zero endophora.

From Table 4.5, considering the zero exophora and the author/reader mentions improves accuracy of zero endophora resolution as well as zero reference resolution including zero exophora. In the evaluation of zero endophora, the proposed model

¹³<http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN>

¹⁴<http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP>

Table 4.6: Results of zero reference resolution

	Recall	Precision	F1
Baseline	0.115	0.370	0.176
Proposed model (estimate)	0.356	0.458	0.401
Proposed model (gold-standard)	0.423	0.535	0.472

that used the estimated author/reader mentions improves both the recall and the precision compared with the baseline model. The reasons for the improvement of the precision could be the following two reasons. The first reason is that when an obligatory case is assigned, it is not necessary to forcibly select a referent from a document and it is possible to select the referent from pseudo discourse entities. The second reason is that informations about the author/reader mentions, such as honorific expression, improve referent identification about the author/reader mentions. For example, in Figure 4.5, a **ガ** case is assigned to “あなた” (you) in the baseline model. In the proposed model, since the **ガ** case is assigned to [author], the precision increases even just the evaluation of the zero endophora. And a **ニ** case, which is a recipient of a honorific expression, can be assigned to “あなた”, which is the reader mention.

The reason for the improvement of the recall could be the following reason. The baseline model learns that it is not necessary to assigned a case slot even if the case is an obligatory case because the baseline model recognizes the zero-exophora as absence of a zero pronoun. On the other hand, The proposed model learns that the obligatory case should be assigned to any discourse entity. For example, in Figure 4.6, the baseline model did not assigned **ガ** case, which is the obligatory case, to any discourse entity. On the other hand, the proposed model could assigned a **ガ** case to “神さま” (god).

From Table 4.5 and Table 4.6, the proposed model given the gold-standard author/reader mentions achieves extraordinarily high accuracies. This result indicates that improvement of the author/reader mention detection improves the accuracy of zero reference resolution in the proposed model.

フレッツ光を はじめるために 押さえておきたい ポイントを
 FLET'S Hikari-ACC for start should hold point-ACC

ご案内します。
 will introduce

‘I will introduce a point that you should hold to start FLET’S Hikari.’

ご利用 中の 回線から、 フレッツ光へ 乗り換える 方法を
 use while from line, to FLET'S Hikari switch method

ご案内します。
 will introduce

‘I will introduce a method that you switch from using line to FLET’S Hikari’

あなたに ぴったりな プロバイダー 選びを お手伝いします。
 you-DAT suit provider choosing will help

‘I will help choosing a suit provider for you.’

	Author mention	Reader mention	Predicate argument structure
Corpus	None	あなた	([author] ガ) 方法ヲ (あなたニ) ご案内します
Baseline	-	-	(あなたガ) 方法を (Null ニ) ご案内します
Proposed model (estimate)	None	あなた	([author] ガ) 方法ヲ (あなたニ) ご案内します
Proposed model (gold-standard)	None	あなた	([author] ガ) 方法ヲ (あなたニ) ご案内します

Figure 4.5: Improvement example (1)

むかし むかし、この世界を つくった インディアンの 神さまが 旅に
old old, this world-ACC created Indian-GEN god-NOM journey

でました。

went

‘long, long ago, the Indian’s god, who had created this world, went on a journey.’

そして、雪が いっぱい つもった 村に きました。

Then, snow-NOM much covered village came

‘Then, he came to a village which is covered in much snow.’

村に はいると、おばあさんが ないています。

village-DAT go into, old woman-NOM was crying

‘When he went into the village, an old woman was crying.’

	Author mention	Reader mention	Predicate argument structure
Corpus	None	None	(神さまガ) (Null ニ) 村ニ きました
Baseline	-	-	(Null ガ) (Null ニ) 村ニ きました
Proposed model (estimate)	None	None	(神さまガ) (Null ニ) 村ニ きました
Proposed model (gold-standard)	None	None	(神さまガ) (Null ニ) 村ニ きました

Figure 4.6: Improvement example (2)

Table 4.7: Results of easing evaluation

	Recall	Precision	F1
Zero endophora	0.406	0.524	0.457
Zero reference	0.402	0.518	0.453

Examples of wrong analyses of the proposed model are Figure 4.7 and Figure 4.8. In Figure 4.7, the proposed model assigned [US-person] to a **ガ** case of “**捻出しなければならない**” (should raise). It is because that there are few examples that “**国**” (nation) is assigned to the **ガ** case in “**捻出する**” (raise).

An error in Figure 4.8 is caused by an error of author mention detection. Since an author mention of this document is “**領事館**” (consulate), a **ガ** case of “**開設しました**” (established) should be assigned to “**領事館.**” The proposed model estimated that there is no author mention and assigned the **ガ** case to [author]. However, it is said that such errors are not exact errors because the proposed model can estimate a referent is author of a document. Then, we evaluated by dealing with errors that assigned cases that should be assigned to author/reader mentions to [author]/[reader] as correct and show the results in Table 4.7. Comparing Table 4.7, Table 4.5 and Table 4.6, accuracy is greatly improved in easing evaluation. From this result, it can be said that many of errors of the proposed model (estimate) are caused by author/reader mentions detection in the same as Figure 4.8. On the other hand, even when evaluation basis is eased, the accuracies of the proposed model (estimate) are lower than the accuracies of the proposed model (gold-standard). It is because the following reason. When a referent is an author/reader mention, $\phi_{mentioned}$ and $\phi_{[author]}$ or $\phi_{[reader]}$ become active. On the other hand, when a referent is [author]/[reader], only $\phi_{[author]}$ or $\phi_{[reader]}$ becomes active. From this result, it is effective that the author/reader mentions are given properties both of a mentioned discourse entity and of a pseudo entity such as the proposed method.

最も 事業 仕分けが 必要なのは「国」です。

most work review-NOM need-TOP nation

‘What most need work review is the nation.’

国の 事業には、省庁 縦 割りや 前例
nation-GEN work-TOP government office vertical division precedent

踏襲 主義などの 弊害により、まだまだ 無駄があります。
following principle-GEN malady ever waste there are

‘There are ever wastes in the nation works by malady such as vertical division of government offices and principle of following precedent.’

これらを 少しでも 減らして、必要な 事業に
these-ACC as much as possible reduce necessary work-DAT

資金を 捻出しなければなりません。
capital-ACC should raise

‘Reducing these as much as possible, the nation should raise capital for necessary works.’

	Author mention	Reader mention	Predicate argument structure
Corpus	None	None	(国ガ) 資金ヲ 事業ニ 捻出しなければなりません
Baseline	-	-	(国ガ) 資金ヲ 事業ニ 捻出しなければなりません
Proposed model (estimate)	None	None	([不特定:人ガ]) 事業ニ 資金ヲ 捻出しなければなりません
Proposed model (gold-standard)	None	None	([不特定:人ガ]) 事業ニ 資金ヲ 捻出しなければなりません

Figure 4.7: Example of error of the proposed model (1)

在 福岡 モンゴル 国 名誉 領事 館の 公式
 in Fukuoka Mongol country honorary consul place-GEN homepage-ACC
 ホームページを 開設しました。
 established

‘Homepage of honorary consulate of Mongolia in Fukuoka has been established.’

当 名誉 領事 館の 概要、九州 沖縄・ モンゴル
 this honorary consul place-GEN outline Kyusyu Okinawa Mongol
 友好 協会の イベント 情報や 活動 状況などを
 friendship association-GEN event informations status activity-ACC
 掲載してまいります。
 will publish

‘We will publish outline of this honorary consulate and event informations and activities of Kyusyu-Okinawa-Mongol friendship association.’

(The third sentence is omitted.)

	Author mention	Reader mention	Predicate argument structure
Corpus	領事館	None	(領事館ガ) ホームページヲ (Null 二) 開設しました
Baseline	-	-	(友好協会ガ) ホームページヲ (Null 二) 開設しました
Proposed model (estimate)	None	None	([author] ガ) ホームページヲ (Null 二) 開設しました
Proposed model (gold-standard)	領事館	None	(領事館ガ) ホームページヲ (Null 二) 開設しました

Figure 4.8: Example of error of the proposed model (2)

4.6 Summary of This Chapter

This chapter presented a zero reference resolution model considering exophora and author/reader mentions. First, we presented a baseline model, which treats only zero endophora. Our model resolved zero reference resolution as a part of predicate-argument structure analysis and treated referents using a unit, discourse entity. Our model represented predicate-argument structures as feature vectors and learned a weight vector corresponding to the feature vector by using a learning-to-rank algorithm. A proposed model detected author/reader mentions as preprocessing and assumed pseudo entities that correspond to zero exophora. The proposed model gave the author/reader mentions properties of the author/reader and discourse entities. In the experiments, our proposed model achieves higher accuracy than the baseline model.

Chapter 5

Conclusion

5.1 Summary of this Research

As the use of the Web increases, NLP applications are being more widely used in a variety of situations. Improved accuracy of fundamental NLP analysis techniques is necessary to improve the performance of these applications. In this study, we focused on zero reference resolution, as one of the fundamental NLP analysis techniques. In previous zero reference resolution studies the main target was newspaper articles. To apply zero reference resolution to Web documents, we noticed that the behavior of the author and reader of a document differs greatly in newspaper articles and Web documents. We categorized appearances of the author and reader in a document as one of two types: author/reader mentions or zero exophora. Author/reader mentions are expressions that refer to the author/reader of a document, and in Japanese, various expressions are used to express author/reader mentions. On the other hand, when author/reader mentions do not appear in a document, but the author/reader has a role in the discourse, the author/reader appears as a referent of zero exophora. In this study, we focused particularly on these phenomena and showed their importance and effectiveness in zero reference resolution.

Annotated Corpus Construction

In Chapter 2, we addressed building a corpus consisting of Web documents and annotated with semantic relations including zero reference relations. By collecting various documents from the Web and automatically and manually filtering them, we gathered documents suitable for annotation of semantic relations, including inter-sentential information. Through annotation, we identified problems inherent in the annotation of Web documents, in the discourse of which the author and reader appear often. The first problem is the annotation of author/reader mentions. A vast number of varied expressions, including not only personal pronouns, but also names and roles, are used to refer to the author/reader of a document. We defined all of these expressions as author/reader mentions. On organization Web sites, the organization often behaves as if it has animacy and a personality, and in such cases, we dealt with the organization as the author. The second problem is the ambiguity of predicate arguments. If the author and reader appear in the discourse, certain arguments can be annotated with either the author, reader, or an indefinite person. Having categorized ambiguous arguments and defined criteria for them, we annotated 1,000 documents based on the criteria. Results of manually categorizing the documents showed that the annotated documents comprise vastly different documents, such as blog articles, online shopping sites, and encyclopedia articles. We researched the ratio of modality and honorific expressions, which reflects the writing style of a document, and the results showed that properties of our corpus differ from those of a newspaper article corpus. In many of the documents, the author and reader appear in the discourse, and in some cases they are mentioned as author/reader mentions. From the results of annotating author and reader mentions, we found that many expressions are used as author and reader mentions and some of these are peculiar to particular types of document. From the results of annotating predicate-argument structures, about half the zero reference relations are zero exophora and about 10 % of the arguments whose referents are either the author, reader, or an indefinite person are ambiguous. Finally, we calculated the inter-annotator agreement. Since the agreement is reasonably high and many of the disagreements are caused by annotator errors, the defined criteria provide consistent annotations.

Author and Reader Mention Detection

In Chapter 3, we focused on author and reader mentions. Since Web documents are written by various authors for a variety of readers, many expressions are used as author and reader mentions. We proposed an automatic author and reader mention detection model. We used a learning-to-rank algorithm to model relations among the author/reader mentions, other discourse entities, and the absence of author/reader mentions, which is represented as two pseudo entities corresponding to two types of absence. The first type is a document in which the author/reader mention does not appear, but the author/reader appears as a referent of zero exophora. The second type is a document in which the author/reader mention does not appear and the author/reader does not have a role in the discourse. Each discourse entity is represented as a lexico-syntactic pattern, which is used to generalize the discourse entity and its parent according to the various types. Experimental results show that our model detects author/reader mentions with high precision but low recall. Based on error analysis, named entity information and honorific expressions strongly affect author/reader detection.

Zero Reference Resolution Considering Zero Exophora and Author and Reader Mentions

In Chapter 4, we proposed a zero reference resolution model considering exophora and author/reader mentions. First, we presented a baseline model that only considers zero endophora. Our model deals with zero reference resolution as part of predicate-argument structure analysis. The baseline model treats only explicitly mentioned discourse entities as candidate referents. By adding pseudo entities, corresponding to referents of zero exophora as candidate referents, the proposed model also considers the zero exophora. Our model allocates common features to the author and reader in zero exophora and author and reader mentions and represents the particular properties that the author and reader have. Our model learns a weight vector corresponding to a feature vector representing a predicate-argument structure by using a learning-to-rank algorithm. Experimental results show the efficiency of our proposed method in the evaluation of both zero en-

dophora and all zero references.

5.2 Future Work

In this study, we constructed an annotated corpus based on Web documents and proposed a zero reference resolution model dealing with zero exophora and author and reader mentions. However, there are some problems that need to be addressed in future work.

Annotated Corpus Construction

In the constructed corpus, we only annotated the first three sentences in order to reduce the workload per sentence and retain a variety of documents. However, some linguistic phenomena do not appear in the leading parts of documents. For example, conclusions only appear at the end of paragraphs and documents. To deal with these phenomena, a corpus that is annotated over the whole document is needed; however, the computational cost of inter-sentential annotation increases exponentially with the length of the document. To annotate the whole document, it is necessary to develop an efficient annotation scheme. For example, the n -best results of an automatic analysis could be shown to the annotators as major candidates.

Since we focused on the author and reader of a document, we defined author and reader mentions and annotated these. However, there are many other relations between the author and reader such as possessions and membership of the organization of the author. These relations play an important role in the same way as author and reader mentions. For example, possessions of the author tend to be introduced in a document. For the above reason, it is important to categorize these relations and include annotations thereof in the corpus.

Zero Reference Resolution

We proposed a zero reference resolution model that considers zero exophora and author and reader mentions. As preprocessing for zero reference resolution, we automatically detected author and reader mentions. However, author and reader

detection recall is low; if author and reader mentions can be detected precisely, zero reference resolution can achieve higher accuracy. For the above reason, it is important to improve author/reader mention detection. In this study, we used only information of the body text; however, in an actual analysis of Web documents, other information, such as HTML tags, meta data, and URLs, can also be used. This kind of information should be beneficial in author/reader mention detection. For example, the author name is directly given in the meta data, and Web sites in the “.com” and “.ac” domain are aimed at customers and students, respectively.

In zero reference resolution, we used both contextual and syntactic information. However, there are many relations between predicate-argument structures, such as causal relationships. Since these relations are essential for capturing the context of a document, it is also important to use these relations as features.

Bibliography

- [1] Association for Computational Linguistics. *MUC6 '95: Proceedings of the 6th Conference on Message Understanding*, Stroudsburg, PA, USA, 1995. Association for Computational Linguistics.
- [2] C. F. Baker, C. J. Fillmore, and J. B. Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.
- [3] W. N. Francis. A standard corpus of edited present-day american english. *College English*, 26(4):267–273, 1965.
- [4] W. N. Francis, H. Kucera, and A. W. Mackie. *FREQUENCY ANALYSIS OF ENGLISH USAGE: LEXICON AND GRAMMAR*. Houghton Mifflin, 1982.
- [5] M. Gerber and J. Chai. Beyond nombank: A study of implicit arguments for nominal predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1583–1592, Uppsala, Sweden, July 2010. Association for Computational Linguistics.
- [6] M. Hangyo, D. Kawahara, and S. Kurohashi. Building a diverse document leads corpus annotated with semantic relations. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 535–544, Bali, Indonesia, November 2012. Faculty of Computer Science, Universitas Indonesia.

- [7] C. Hashimoto, S. Kurohashi, D. Kawahara, K. Shinzato, and M. Nagata. Construction of a blog corpus with syntactic, anaphoric and sentiment annotations. *Journal of natural language processing*, 18(2):175–201, 2011. (in Japanese).
- [8] K. Hasida. *Global Document Annotation (GDA)*, 2002. (in Japanese).
- [9] M. Hattori and M. Harada. Anaphoric analysis system anasys -zero-anaphora resolution based on naive bayes method for semantic features gained from sentence semantic analysis-. In *IEICE Technical Report NLC-2013-NLC2013-46*, pages 7–12, December 2013.
- [10] Y. Hayashibe, M. Komachi, and Y. Matsumoto. Japanese predicate argument structure analysis exploiting argument position and type. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 201–209, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.
- [11] R. Herbrich, T. Graepel, P. Bollmann-Sdorra, and K. Obermayer. Learning preference relations for information retrieval. In *ICML-98 Workshop: text categorization and machine learning*, pages 80–84, 1998.
- [12] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics, 2006.
- [13] R. Iida, K. Inui, and Y. Matsumoto. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 625–632, Sydney, Australia, July 2006. Association for Computational Linguistics.
- [14] R. Iida, K. Inui, and Y. Matsumoto. Capturing salience with a trainable cache model for zero-anaphora resolution. In *Proceedings of the Joint Conference of*

- the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 647–655, Suntec, Singapore, August 2009. Association for Computational Linguistics.
- [15] R. Iida, M. Komachi, K. Inui, and Y. Matsumoto. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proc. of the Linguistic Annotation Workshop*, pages 132–139, 2007.
- [16] K. Imamura, K. Saito, and T. Izumi. Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 85–88, Suntec, Singapore, August 2009. Association for Computational Linguistics.
- [17] Information-technology Promotion Agency, editor. *IPA Lexicon of the Japanese language for computers (Basic Verbs)*. Information-technology Promotion Agency, 1987. (in Japanese).
- [18] H. Isozaki, H. Kazawa, and T. Hirao. Japanese zero pronoun resolution based on lexicographical ordering of penalties. *Journal of Information Processing*, 47(7):2279–2294, jul 2006. (in Japanese).
- [19] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.
- [20] S. Johansson, G. Leech, and H. Goodluck. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Department of English, University of Oslo, 1978.
- [21] M. Kameyama. A property-sharing constraint in centering. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 200–206, New York, New York, USA, July 1986. Association for Computational Linguistics.
- [22] D. Kawahara and S. Kurohashi. Case frame compilation from the web using high-performance computing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 67–73, 2006.

- [23] D. Kawahara and S. Kurohashi. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 176–183, New York City, USA, June 2006. Association for Computational Linguistics.
- [24] D. Kawahara, S. Kurohashi, and K. Hasida. Construction of a japanese relevance-tagged corpus. In *Proceedings of The Third International Conference on Language Resources Evaluation*, May 2002.
- [25] M. Komachi and R. Iida. Annotating predicate-argument structure and anaphoric relations to bccwj. In *Processing of Workshop of Japanese Corpus 2010*, pages 325–330, March 2011. (in Japanese).
- [26] F. Kong and G. Zhou. A tree kernel-based unified framework for chinese zero anaphora resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 882–891, Cambridge, MA, October 2010. Association for Computational Linguistics.
- [27] S. Kurohashi and M. Nagao. 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, 1998.
- [28] K. Maekawa. Balanced corpus of contemporary written japanese. In *IJCNLP*, pages 101–102, 2008.
- [29] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330, 1993.
- [30] A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. Annotating noun argument structure for nombank. In *LREC*, volume 4, pages 803–806, 2004.
- [31] E. Miltsakaki, R. Prasad, A. K. Joshi, and B. L. Webber. The penn discourse treebank. In *Proc. of the Forth International Conference on Language Resources and Evaluation (LREC'04)*, 2004.

- [32] M. Murata and M. Nagao. An estimate of referents of pronouns in japanese sentences using examples and surface expressions. *Journal of natural language processing*, 4(1):87–109, 1997. (in Japanese).
- [33] National Institution for Japanese Language and Linguistics, editor. *Word List by Semantic Principles, Revised and Enlarged Edition*. Dainippon tosho, 2004. (in Japanese).
- [34] NIST. *Automatic Content Extraction 2000*, 2000.
- [35] K. Ohara. Full text annotation with japanese framenet: Study to annotation semantic frame to bccwj(in japanese). In *Proceedings of the 17th Annual Meeting fo the Association for Natural Language Processing*, pages 703–704, 2011. (in Japanese).
- [36] M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.
- [37] M. Poesio, O. Uryupina, and Y. Versley. Creating a coreference resolution system for italian. In N. C. C. Chair, K. Choukri, B. Maegaard, J. Mariani, J. Odiijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).
- [38] L. Rello, R. Baeza-Yates, and R. Mitkov. Elliphant: Improved automatic detection of zero subjects and impersonal constructions in spanish. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 706–715. Association for Computational Linguistics, 2012.
- [39] L. Rello and I. Ilisei. 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, 2009.

- [40] K. J. Rodríguez, F. Delogu, Y. Versley, E. W. Stemle, and M. Poesio. 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)*, 2010.
- [41] J. Ruppenhofer, C. Sporleder, R. Morante, C. Baker, and M. Palmer. Semeval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 45–50, Uppsala, Sweden, July 2010. Association for Computational Linguistics.
- [42] R. Sasano, D. Kawahara, and S. Kurohashi. A fully-lexicalized probabilistic model for japanese zero anaphora resolution. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 769–776, Manchester, UK, August 2008. Coling 2008 Organizing Committee.
- [43] R. Sasano and S. Kurohashi. A discriminative approach to japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 758–766, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.
- [44] K. Seki, A. Fujii, and T. Ishikawa. A probabilistic method for analyzing japanese anaphora integrating zero pronoun detection and resolution. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.
- [45] H. Taira, S. Fujita, and M. Nagata. A japanese predicate argument structure analysis using decision lists. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 523–532, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.
- [46] H. Taira and M. Nagata. A study on japanese ellipsis analysis with predicate argument structure analysis. In *Processing of 19th Annual Meeting fo the Association for Natural Language Processing*, pages 106–109, 3 2013. (in Japanese).

List of Major Publications

- [1] Masatsugu Hangyo, Daisuke Kawahara and Sadao Kurohashi. Building a Diverse Document Leads Corpus Annotated with Semantic Relations. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 535–544, 2012.
- [2] Masatsugu Hangyo, Daisuke Kawahara and Sadao Kurohashi. Japanese Zero Reference Resolution Considering Exophora and Author/Reader Mentions. In *Proceedings of EMNLP 2013: Conference on Empirical Methods in Natural Language Processing*, pages 924–934, 2013.
- [3] Masatsugu Hangyo, Daisuke Kawahara and Sadao Kurohashi. Building and Analyzing a Diverse Document Leads Corpus Annotated with Semantic Relations. In *Journal of Natural Language Processing*, 21(2), 2014. (in Japanese). (to appear).
- [4] Masatsugu Hangyo, Daisuke Kawahara and Sadao Kurohashi. Japanese Zero Reference Resolution Considering Exophora and Author/Reader Mentions. In *Journal of Natural Language Processing*, 2014. (in Japanese). (submitted).

List of Other Publications

- [1] Masatsugu Hangyo and Sadao Kurohashi. Progressive Interpretation of Definitions of Dictionary Based on Structured Attribute Representation. In *Proceedings of The 17th Annual Meeting of The Association for Natural Language Processing*, pages 21–24, 2011. (in Japanese).
- [2] Masatsugu Hangyo, Daisuke Kawahara and Sadao Kurohashi. Building Diverse Document Leads Corpus Annotated with Semantic Relations. In *IPSJ SIG Technical Reports 2012-NL-206*, 2012. (in Japanese).
- [3] Masatsugu Hangyo, Daisuke Kawahara and Sadao Kurohashi. Japanese Zero Reference Resolution Considering Exophora and Author/Reader Mentions. In *Proceedings of The 20th Annual Meeting of The Association for Natural Language Processing*, 2014 (in Japanese). (submitted).