Japanese Anaphora Resolution Based on Automatically Acquired World Knowledge
(自動獲得した世界知識に基づく日本語照応解析)

笹野遼平

Japanese Anaphora Resolution Based on Automatically Acquired World Knowledge (自動獲得した世界知識に基づく日本語照応解析)

Ryohei Sasano

December 2008

Abstract

What is represented in natural language text has originally a network structure, in which several mentions refer to the same entities, and several entities have tight relations with each other. However, due to the linear constraints of text, most of them are not obviously expressed in the normal form of text; thus automatic recognition of such relations is considered to be an essential step in natural language understanding.

This thesis focuses on anaphora resolution in Japanese text. Anaphora resolution is the task to recognize anaphoric relations in text, which include anaphoric relations between coreferential mentions, zero anaphoric relations, and bridging relations. Since there is no syntactic dependency relation between an anaphor and its antecedent, few grammatical clues can be used to resolve anaphoric relations. Therefore, several kinds of knowledge concerning anaphoric relations are necessary to resolve them, and we first acquire such knowledge from very large corpora and some dictionaries.

To resolve anaphoric relations with high accuracy, fundamental analyses, such as word segmentation, part-of-speech tagging, and named entities (NE) recognition, are also important. A typical model for understanding Japanese texts first segments input sentences into word sequences, assigns part-of-speech tags, recognizes NEs, and then recognizes syntactic structure and case structure. As a consequence of these analyses, relations between expressions that have syntactic dependency relations are recognized. In succession to these analyses, anaphora resolution is conducted. Therefore, in order to construct high performance anaphora resolution system, it is important to conduct these former analyses with high accuracy. Especially, the NE recognition result heavily affects anaphora resolution performance; to recognize NEs with very high accuracy is considered to be essential for anaphora resolution. Therefore, in this thesis, we also propose high performance NE recognition system that utilizes non-local information.

As for resolving anaphora relations, we propose integrated anaphora resolution system, which includes coreference resolution, zero anaphora resolution, and bridging reference resolution. Our system first recognizes coreference relations by using automatically acquired knowl-

edge of nominal relations. Then, our system resolves zero anaphora and bridging reference simultaneously by using probabilistic model based on automatically acquired case frames.

In Chapter 2, we describe how to acquire world knowledge for anaphora resolution automatically. We first acquire knowledge of synonyms, which is useful for coreference resolution, from a large raw corpus and dictionary definition sentences. Secondly, we construct wide-coverage case frames from modifier-head examples in the resulting parses of large corpora. We take a gradual approach that begins to acquire basic case frames and gradually acquires richer ones by doing both case frame acquisition and text understanding one after another. In addition, in order to deal with data sparseness problem, we generalize the examples of case slots. Finally, we construct nominal case frames, which describe indispensable entities of nouns and utilized for bridging reference resolution. The point of the construction method is the integrated use of a dictionary and example phrases from large corpora.

Then, we report an attempt to improve the NE recognition performance. The state-of-the-art NE recognizer has achieved reasonable performance. However, since NEs can be an important clue for anaphora resolution, more accurate NE recognition systems are considered to benefit the performance of anaphora resolution. In Chapter 3, we propose an NE recognition system that uses non-local information. While conventional Japanese NE recognition systems have been often performed immediately after morphological analysis and rely only upon local context, our system performs after structural analyses and uses four types of non-local information: cache features, coreference relations, syntactic features, and case frame features, which are obtained from structural analyses. We evaluated our approach on CRL NE data and obtained a higher F-measure than existing approaches that do not use non-local information. We also conducted experiments on IREX NE data and an NE-annotated web corpus, and confirmed that non-local information improves the performance of NE recognition.

In Chapter 4, we present a knowledge-rich approach to Japanese coreference resolution, which resolves anaphoric relations between coreferential mentions that are not omitted. In Japanese, since pronouns are often omitted, proper noun coreference and common noun coreference occupy a central position in coreference relations. To resolve such types of coreference, knowledge of synonyms is considered to be useful; thus we utilize automatically acquired knowledge of synonyms in coreference resolution. Furthermore, to boost the performance of coreference resolution, we integrate a primitive bridging reference resolver into coreference resolver. The experimental results show that utilization of knowledge of synonyms and bridging reference resolver boosted the performance of coreference resolution.

In Chapter 5, we propose a probabilistic model for Japanese zero anaphora and bridging reference resolution. By using the results of coreference resolution, this model first recognizes discourse entities and links all mentions to them. Zero pronouns are then detected by case structure analysis based on automatically constructed case frames; their appropriate antecedents are selected from the discourse entities with high salience scores. In this model, case structure and zero anaphoric relations are simultaneously determined based on probabilistic evaluation metrics that uses case frames and several preferences on the relation between a zero pronoun and an antecedent.

We report the effect of corpus size on case frame acquisition for discourse analysis in Chapter 6. For this study, case frames were constructed from corpora of six different sizes ranging from 1.6 million to 1.6 billion sentences. These case frames were then applied to syntactic and case structure analysis, and zero anaphora resolution. Better results were obtained by using case frames constructed from larger corpora; the performance was not saturated even with a corpus size of 1.6 billion sentences.

In Chapter 7, we summarize this thesis, provide concluding remarks, and outline the areas for future work.

Acknowledgments

I would like to express my gratitude to Professor Sadao Kurohashi (currently at Kyoto University) for his supervision of this thesis and for his constructive suggestions and continuous encouragement throughout this work.

I would like to express my sincere appreciation to Professor Kenjiro Taura for his valuable comments. I am also grateful to my thesis committee members: Mitsuru Ishizuka, Masaru Kitsuregawa, Jun Adachi, Jun'ichi Tsujii of The University of Tokyo, for their valuable suggestions and comments about my thesis research.

I owe a great deal to all previous and current members of Kurohashi Laboratory. Especially, I would like to thank Dr. Daisuke Kawahara (currently at NICT), Dr. Tomohide Shibata (currently at Kyoto University), Dr Masashi Okamoto (currently at Tokyo University of Technology) and Dr. Keiji Shinzato (currently at Kyoto University). I would like to thank Nobuko Iiyama and Yuko Ashihara for making administrative matters run smoothly.

I would like to thank many people involved in the corpus annotation project. I gratefully acknowledge tireless labor from annotators: Manami Ishikawa, Natsuki Nikaido, and Marika Horiuchi.

I am grateful to the members of Professor Taura's Laboratory. Especially, I would like to thank Hideo Saito. I would like to thank Eri Watanabe, Shihomi Shimomura, Anzu Uchida and Yoshiko Kuroda for making administrative matters run smoothly.

Finally, I would like to thank my family and friends for their continuous supports and encouragements during and before this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Anaphora Resolution in Text Understanding

What is represented in natural language text has originally a network structure, in which several mentions refer to the same entities, and several entities have tight relations with each other. However, due to the linear constraints of text, most of them are not obviously expressed in the normal form of text; thus automatic recognition of such relations is considered to be an essential step in natural language understanding.

Anaphora resolution is one of the important subtasks to recognize such relations, which recognizes anaphoric relations in text. This thesis focuses on anaphora resolution in Japanese text, in which a lot of anaphoric relations are included. Figure 1.1 shows examples of anaphoric relations in Japanese text. In Figure 1.1, the solid lines mean syntactic dependency relations; the broken lines mean anaphoric relations. There are three types of anaphoric relations in this example.

The first type is anaphoric relation between coreferential mentions. For example, the two mentions of "*kakaku*" (price) in Figure 1.1 refer to the same entity and have anaphoric relation. To recognize such relations is called coreference resolution.

The second type is zero anaphoric relation. In Japanese, anaphors are often omitted; such omissions are called zero pronouns and such anaphora is called zero anaphora. For example, "*ga*" (nominative) case of "*hanbai*" (sell) in Figure 1.1 is omitted and the zero pronoun refers to "Toyota." The relation between the zero pronoun and "Toyota" is called a zero anaphoric relation; to recognize such relations is called zero anaphora resolution.

The last type is bridging relation. Some nouns strongly imply the necessity of certain arguments. For example, "*kakaku*" (price) in Figure 1.1 means "Prius-*no kakaku*" (the price of

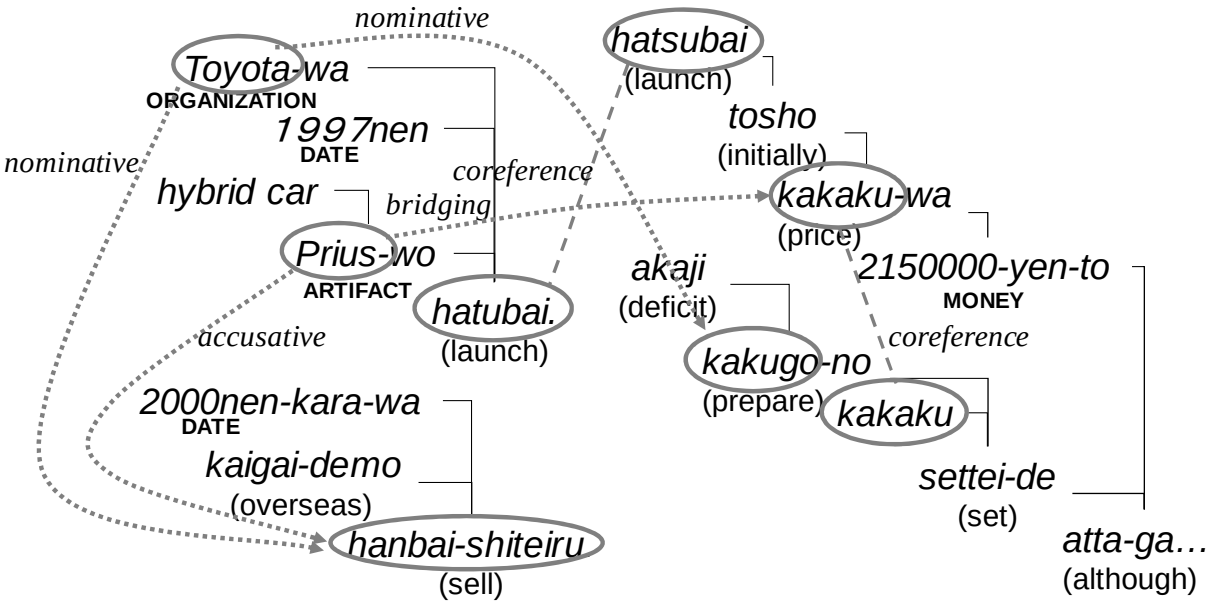


Figure 1.1: Anaphoric Relations in Japanese Text.

Prius) and indirectly refers to "Prius." The connection between "*kakaku*" (price) and "Prius" is called a bridging relation; to recognize such relations is called bridging reference resolution.

Each type of anaphora resolution plays an important role in text understanding; in this thesis, we aim to resolve all types of anaphoric relations: anaphoric relations between coreferential mentions, zero anaphoric relations, and bridging relations.

1.2 Toward Anaphora Resolution

Since there is no syntactic dependency relation between an anaphor and its antecedent, few grammatical clues can be used to resolve anaphoric relations. Therefore, several kinds of knowledge concerning anaphoric relations are necessary to resolve them. For examples, knowledge of synonyms is essential for recognizing coreference relations between paraphrased mentions; case frames, which describe what kinds of cases each predicate has and what kinds of nouns can fill these case slots, are essential for zero anaphora resolution; nominal case frames, which describe indispensable entities of nouns, are essential for bridging reference resolution.

There have been some studies that have tried to elaborate these kinds of knowledge by hand, but the problem is their coverage. That is to say, it is difficult to make wide-coverage knowledge

manually, because language is composed of an enormous number of content words. Moreover, there are technical terms or jargon for every domain, and new words are coined every day. In this thesis, by acquiring automatically from very large raw corpus and several dictionaries, we overcome such data sparseness problem.

We first acquire knowledge of synonyms, which is utilized for coreference resolution, from a large raw corpus and dictionary definition sentences. Secondly, we construct case frames from modifier-head examples in the parses of large corpora. We take a gradual approach that begins to acquire basic case frames and gradually acquires richer ones by doing both case frame acquisition and text understanding one after another. In addition, in order to deal with data sparseness problem, we generalize the examples of case slots. Finally, we construct nominal case frames, which describe indispensable entities of nouns and utilized for bridging reference resolution. The point of the construction method is the integrated use of a dictionary and example phrases from large corpora.

To resolve anaphoric relations with high accuracy, fundamental analyses, such as word segmentation, part-of-speech tagging, and named entity (NE) recognition, are important. Figure 1.2 shows a typical model for understanding Japanese texts. This model first conducts morphological analysis, that is, segments input sentences into word sequences and assigns part-of-speech, next recognizes named entities, and then recognizes syntactic and case structure. As a consequence of these analyses, relations between expressions that have syntactic dependency relations are recognized. In succession to these analyses, anaphora resolution is conducted. Most previous work concerning Japanese text understanding, such as named entity recognition [1–5], coreference resolution [6,7], and zero anaphora resolution [8–10], presupposed such processing order.

Therefore, in order to construct high performance anaphora resolution system, it is important to conduct these former analyses with high accuracy. Especially, the NE recognition result heavily affects anaphora resolution performance; to recognize NEs with high accuracy is considered to be essential for anaphora resolution. Hence, we also try to build high performance NE recognition system. In this thesis, we improve NE recognition system by using non-local information.

In addition, these analyses are considered to depend on each other, and should not be resolved separately; thus this thesis proposes an integrated model for text understanding. In this model, NE recognizer utilizes the primitive syntactic and case structure analysis and the result of primitive coreference resolution; coreference resolution system utilizes the result of primitive

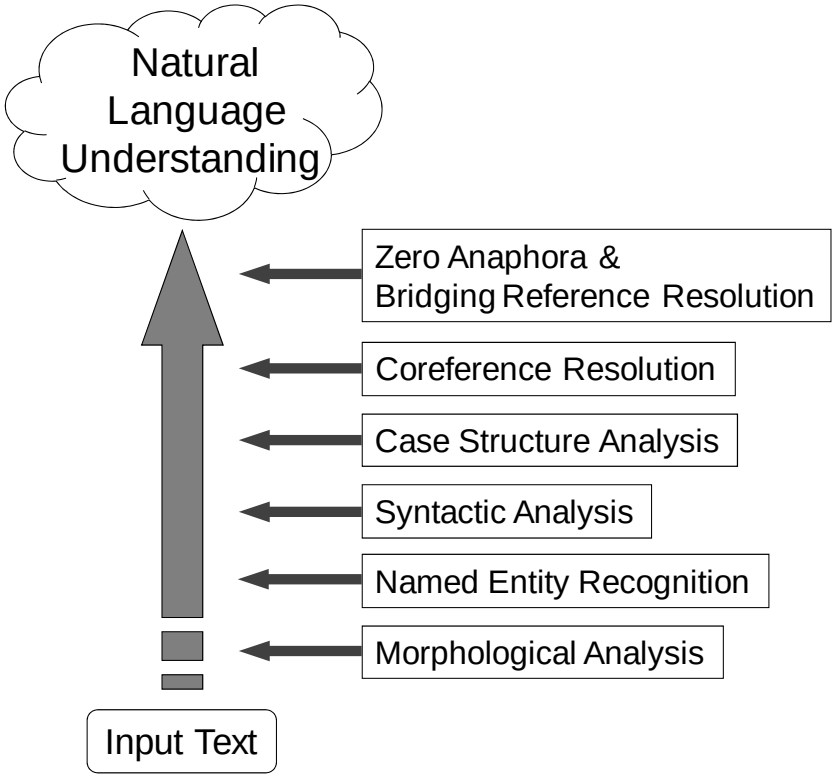


Figure 1.2: A Typical Model for Understanding Japanese Texts.

bridging reference resolver. As for resolving zero anaphoric and bridging relations, we build an integrated anaphora resolution system, which includes coreference resolution, zero anaphora resolution, and bridging reference resolution. This system first recognizes coreference relations by using automatically acquired knowledge of nominal relations, and then resolves zero anaphora and bridging reference simultaneously by using probabilistic model based on automatically acquired case frames.

1.3 Our Model for Anaphora Resolution

As mentioned above, in our proposed model, NE recognizer utilizes the primitive syntactic and case structure analysis and the result of primitive coreference resolution; coreference resolution system utilizes the result of primitive bridging reference resolver; zero anaphora resolution and bridging reference resolution are conducted simultaneously. In addition, we apply the probabilistic model for Japanese syntactic and case structure analysis proposed by Kawahara and Kurohashi [11], which resolves syntactic and case structure simultaneously. Consequently, outline of our proposed model for anaphora resolution is as follows:

1. Segment input sentences into word sequences and assign part-of-speech tags (**morphological analysis**).
2. Conduct a primitive **NE recognition**.
3. Parse an input text using the Japanese parser KNP [12] (**syntactic analysis**).
4. Analyze **case structure** using the method proposed by Kawahara and Kurohashi [13].
5. Resolve **bridging reference** using primitive bridging reference resolver.
6. Conduct **coreference resolution** using the result of primitive bridging reference resolution (**Feedback**: **bridging reference resolution**).
7. **Recognize NEs** using non-local information (**Feedback**: **syntactic analysis, case structure analysis, coreference resolution**).
8. Resolve **syntactic** and **case structure** simultaneously by using Kawahara's method [11].
9. Resolve **case structure**, **zero anaphoric relation** and **Bridging reference**, simultaneously.

Figure 1.3 shows the overview of this model.

1.4 Contribution of this Work

There are not a lot of previous works for Japanese anaphora resolution. Especially for bridging reference resolution, there are only a few previous works, and none of them handle whole bridging reference. All previous works for Japanese anaphora resolution concentrated upon only one anaphoric relation type. Thus, we can say that this is the first work of integrated anaphora resolution including coreference resolution, zero anaphora resolution, and bridging reference resolution.

In addition, our NE recognition system achieved state-of-the-art performance against both CRL NE data and IREX test data. This result shows the usefulness of non-local information that obtained from structural analyses for NE recognition in Japanese, and the effectiveness integrated analyses.

This thesis also reports the effect of corpus size on case frame acquisition for discourse analysis in Japanese. As a result of several discourse analyses using, we confirm that better

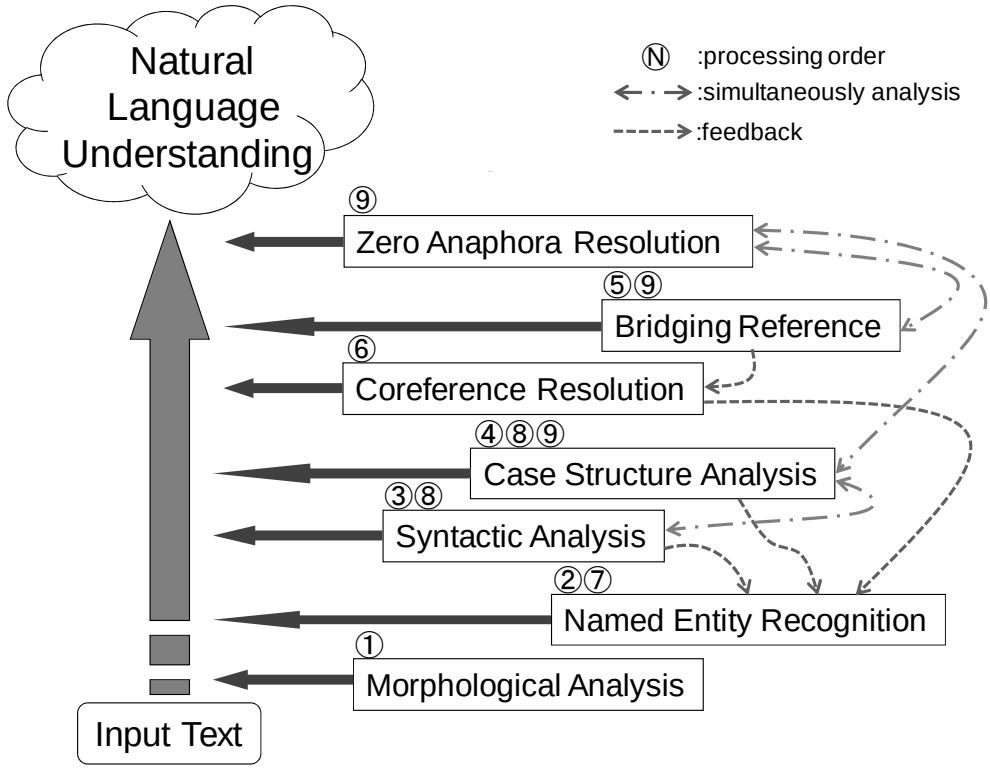


Figure 1.3: Overview of Proposed Model for Text Understanding.

results were obtained by using case frames constructed from larger corpora; the performance was not saturated even with a corpus size of approximately 100 billion words.

1.5 Outline of this Thesis

Chapter 2 describes how to acquire world knowledge automatically from very large corpora. We first acquire knowledge of synonyms, which is useful for coreference resolution, from a large raw corpus and dictionary definition sentences. Secondly, we construct wide-coverage case frames from modifier-head examples in the resulting parses of large corpora. We take a gradual approach that begins to acquire basic case frames and gradually acquires richer ones by doing both case frame acquisition and text understanding one after another. In addition, in order to deal with data sparseness problem, we generalize the examples of case slots. Finally, we construct nominal case frames, which describe indispensable entities of nouns and utilized for bridging reference resolution. The point of the construction method is the integrated use of a dictionary and example phrases from large corpora.

Chapter 3 describes the NE recognition (NER) system that uses non-local information. While conventional Japanese NER systems have been often performed immediately after morphological analysis and rely only on local context, our NER system performs after structural analyses, and uses four types of non-local information: cache features, coreference relations, syntactic features and caseframe features, which are obtained from structural analyses. We evaluated our approach on CRL NE data and obtained a higher F-measure than existing approaches that do not use non-local information. We also conducted experiments on IREX NE data and an NE-annotated web corpus and confirmed that non-local information improves the performance of NER.

Chapter 4 presents a knowledge-rich approach to Japanese coreference resolution, which resolves anaphoric relations between coreferential mentions that are not omitted. In Japanese, since pronouns are often omitted, proper noun coreference and common noun coreference occupy a central position in coreference relations. To resolve such types of coreference, knowledge of synonyms is considered to be useful; thus we utilize automatically acquired knowledge of synonyms in coreference resolution. Furthermore, to boost the performance of coreference resolution, we integrate a primitive bridging reference resolver into coreference resolver. The experimental results show that utilization of knowledge of synonyms and bridging reference resolver boosted the performance of coreference resolution.

Chapter 5 presents a probabilistic model for Japanese zero anaphora and bridging reference resolution. By using the results of coreference resolution, this model first recognizes discourse entities and links all mentions to them. Zero pronouns are then detected by case structure analysis based on automatically constructed case frames; their appropriate antecedents are selected from the discourse entities with high salience scores. In this model, case structure and zero anaphoric relations are simultaneously determined based on probabilistic evaluation metrics that uses case frames and several preferences on the relation between a zero pronoun and an antecedent.

Chapter 6 reports the effect of corpus size on case frame acquisition for discourse analysis. For this study, a Japanese corpus consisting of up to approximately 100 billion words was collected from the Web, and case frames were constructed from corpora of six different sizes ranging from 1.6 million to 1.6 billion sentences. These case frames were then applied to syntactic and case structure analysis, and zero anaphora resolution. Better results were obtained by using case frames constructed from larger corpora; the performance was not saturated even with a corpus size of approximately 100 billion words.

Chapter 7 provides concluding remarks, summaries the thesis, and outlines the areas for future work.

Chapter 2

Knowledge Acquisition for Anaphora Resolution

2.1 Introduction

In order to resolve anaphoric relations with high accuracy, several kinds of world knowledge are essential. For instance, knowledge of synonyms is essential for recognizing coreference relations between paraphrased mentions; case frames, which describe what kinds of cases each predicate has and what kinds of nouns can fill these case slots, are essential for zero anaphora resolution; nominal case frames, which describe indispensable entities of nouns, are essential for bridging reference resolution.

Therefore, we acquire such knowledge in advance of anaphora resolution. In this chapter, we illustrate what kinds of world knowledge are required for anaphora resolution and how to acquire such knowledge.

2.2 Synonym Extraction

2.2.1 Overview of Synonym Extraction for Coreference Resolution

It is difficult to recognize coreference relations between absolutely different expressions without knowledge of synonyms. To construct a high performance coreference resolver, we acquire knowledge of synonyms in advance. Note that, in this thesis, synonyms include acronyms and abbreviations.

As resources for synonym extraction, we use parenthesis expressions in raw corpus, and dictionary definition sentences. The characteristic of synonym extraction from parenthesis expressions is the ability to respond to new words. However, very familiar synonyms, such as *US*

Table 2.1: Thresholds for Synonym Extraction.

	type	threshold	other constraints
1	One consists of alphabets and the other does not	2	none
2	One consists of Japanese letters *katakana* and the other does not	5	none
3	One consists of Chinese characters and the other is the abbreviation [1]	1	difference of length > 2
4	others	200	each frequency > 8

and *America*, are considered not to be extracted from parenthesis expressions. Thus, in order to extract such very familiar synonyms, we also extract synonyms from dictionaries for humans.

2.2.2 Synonym Extraction from Parenthesis Expressions

When unfamiliar synonymous expressions are used for the first time in text, the information is often written in text by using parenthesis. In example (2.1), "KEDO," a synonym of "*Chosen Hanto Enerugi Kaihatu Kiko*" (Korean Peninsula Energy Development Organization), is written in the following parenthesis. Therefore, we first considered to extract synonyms from parenthesis expressions that appeared in raw corpus.

(2.1) *Suzuki Chosen Hanto Energy Kaihatu Kiko* (KEDO)
Suzuki Korean Peninsula Energy Development Organization (KEDO)
taishi-ga yutai-shita.
ambassador retire

(The Korean Peninsula Energy Development Organization (KEDO) Ambassador Suzuki retired.)

Parenthesis is not always used to indicate synonym. For example, parenthesis is sometimes used to indicate attribution of preceding noun phrases such as age, affiliation and birthplace. Thus, one problem is how to extract parenthesis pairs that indicate synonym. In addiction, even if parenthesis is used to indicate synonym, it is not so easy to discriminate the synonymous part from the preceding noun phrases.

[1] One expression must include all Chinese characters included in the other expression.

Table 2.2: The Result of Synonym Extraction from Parenthesis Expressions.

type	number	accuracy	examples
1	1,572	99.5%	*kokunai sou-seisan* = GDP domestic gross product GDP *Kita taiseiyo joyaku kiko* = NATO North Atlantic Treaty Organization NATO *Europe rengo* = EU European Union EU
2	727	98.5%	*jugyo keikaku* = *syllabus* class plan syllabus *kinyu hasei shohin* = *derivative* financial instrument derivative *shien kigyo* = *sponsor* support company sponsor
3	239	98.7%	*Gakushu kenkyuu sha* = *Gakken* study pursuit corporation = Gakken *Nihon kogyo ginko* = *Kogin* Japan industrial bank = Kogin
4	110	96.4%	*ushi kaimenjou noushou* = *kyogyubyo* bovine spongiform encephalopathy mad cow disease *Myanmar* = *Burma* Myanmar Burma
sum	2,648	99.0%	

In order to deal with these problem, we make an assumption that a parenthesis expression "*A(B)*" indicate synonym, if the reverse pair "*B(A)*" can also appeared in corpus, which we call two-way pair, and the product of frequencies of the parenthesis expressions is high. Note that we consider several preceding noun phrase candidates "*A*." According to this assumption, we extract synonym pairs from parenthesis expressions as follows:

1. Count the frequency of parenthesis expression "*A(B)*"and the frequency of parenthesis expression "*B(A)*," and calculate the product of them.

2. If the product exceeds the thresholds, the pair A and B is judged as a synonym pair.

Table 2.1 shows the thresholds set by observing the products of randomly selected 100 pairs, which are set in order not extract any incorrect synonym pairs.

We extracted synonym pairs from Japanese newspaper articles in 26 years (12 years of *Mainichi* newspaper and 14 years of *Yomiuri* newspaper). There are about 10 million parenthesis expressions, 110 thousand unique parenthesis expressions and 5,800 two-way parenthesis

expression pairs in the newspaper articles. Table 2.2 shows the result of extraction. The accuracy is evaluated by using randomly selected synonym pairs. We use 200 pairs for type 1 and 2, and use all extracted pairs for type 3 and 4.

We acquired 2,648 synonym pairs. About 99% of the extracted synonym pairs were correct. This is because we set the threshold not to extract incorrect synonym pairs. Comparing with previous work [14], we can confirm that our approach extracts adequate amount of synonym pairs with high accuracy by using large amounts of corpus.

2.2.3 Synonym Extraction from Dictionary

Secondly, in order to extract very familiar synonyms, we use definition sentences of dictionaries for humans. There have been many previous studies that tried to extract synonyms from dictionaries [15, 16], and most of them are tried to extract as many synonyms as possible. However, our purpose of synonym extraction from dictionaries is to acquire synonyms that are not extracted from parenthesis expressions and useful for coreference resolution. Therefore, we used a simple and strict rule that might extract only very familiar synonyms from dictionaries. The following process is carried out for each dictionary entry *A*.

1. If the definition sentence ends with "*no ryaku*" (abbreviation of) or "*no koto*" (synonym of), we extract the rest of the sentence as a synonym candidate *B*; otherwise extract whole the sentence as *B*.

2. If *B* itself is an entry of dictionaries or enclosed by angle brackets, the pair of *A* and *B* is judged as a synonym pair.

We extracted synonyms from *Reikai Shougaku Kokugojiten* [17] and *Iwanami Kokugo Jiten* [18]. They have about 30 thousand entries and 60 thousand entries, respectively. As a result, we extracted 150 synonym pairs from dictionary definition sentences. Table 2.3 shows examples of extracted synonym pairs.

Only 6 synonym pairs extracted from dictionary definition sentences overlapped with the synonym pairs extracted from parenthesis expressions. Therefore, it is reasonable to suppose that we extract very familiar synonyms from definition sentences that were not extracted from parenthesis expressions in raw corpus. As a whole, we acquired 2,792 synonym pairs from both raw corpus and dictionary definition sentences.

Table 2.3: Examples of Extracted Synonyms from Dictionaries.

type of definition sentence	examples entry	extracted synonym
. . .-no ryaku	*fukei* policewoman	*fujin keikan* woman cop
	Niti JP	*Nihon* Japan
. . .-no koto	*Chuugoku* China	*Chuuka Jinmin Kyowakoku* the People's Republic of China
	Bei US	*America* America
others	*Chokou* Yanzi Jiang	*Yousukou* Chang Jiang
	Japan Japan	*Nihon* Nippon

2.3 Construction of Case Frame

2.3.1 Overview of Case Frame Construction

The case frames are useful knowledge for syntactic analysis or parsing for Japanese text, and essential knowledge for zero anaphora resolution, especially for zero pronoun detection.

Some research institutes have constructed Japanese case frames manually [19–21]. However, it is quite expensive, and almost impossible to construct wide-coverage case frames by hand. For acquiring wide-coverage case frames, Kawahara and Kurohashi proposed a method for constructing case frames from large corpora [22]. We basically follow their method for case frame construction. In addition, we propose a method for generalizing case slot examples in order to alleviate the example sparseness problem. In this section, we outline the method for constructing case frames and show how to generalize the case slot examples.

2.3.2 Basic Method

The biggest problem in automatic case frame construction is verb sense ambiguity. Verbs which have different meanings should have different case frames, but it is hard to disambiguate verb senses precisely. To deal with this problem, predicate-argument examples that are collected from a large corpus are distinguished by coupling a verb and its closest case component. That is, examples are not distinguished by verbs (e.g. "*tsumu*" (load/accumulate)), but by couples

Table 2.4: Examples of Constructed Case Frames.

	case slot	examples	generalized examples with rate
tsumu (1) (load)	*ga* (nominative)	he, driver, friend, $\cdots$	[CT:PERSON]:0.45, [NE:PERSON]:0.08, $\cdots$
	wo (accusative)	baggage, luggage, hay, $\cdots$	[CT:ARTIFACT]:0.31, $\cdots$
	ni (dative)	car, truck, vessel, seat, $\cdots$	[CT:VEHICLE]:0.32, $\cdots$
tsumu (2) (accumulate)	*ga* (nominative)	player, children, party, $\cdots$	[CT:PERSON]:0.40, [NE:PERSON]:0.12, $\cdots$
	wo (accusative)	experience, knowledge, $\cdots$	[CT:ABSTRACT]:0.47, $\cdots$
$\vdots$	$\vdots$		$\vdots$
hanbai (1) (sell)	*ga* (nominative)	company, Microsoft, $\cdots$	[NE:ORG.]:0.16, [CT:ORG.]:0.13, $\cdots$
	wo (accusative)	goods, product, ticket, $\cdots$	[CT:ARTIFACT]:0.40, [CT:FOOD]:0.07, $\cdots$
	ni (dative)	customer, company, $\cdots$	[CT:PERSON]:0.28, $\cdots$
	de (locative)	shop, bookstore, site $\cdots$	[CT:FACILITY]:0.40, [CT:LOCATION]:0.39, $\cdots$
$\vdots$	$\vdots$		$\vdots$

(e.g. "*nimotsu-wo tsumu*" (load baggage) and "*keiken-wo tsumu*" (accumulate experience)).

This process makes separate case frames which have almost the same meaning or usage. For example, "*nimotsu-wo tsumu*" (load baggage) and "*busshi-wo tsumu*" (load supply) are similar, but have separate case frames. To cope with this problem, the case frames are clustered.

To sum up, the procedure for the automatic case frame construction is as follows.

1. A large raw corpus is parsed by the Japanese parser, KNP [12], and reliable predicate-argument examples are extracted from the parse results.

2. The extracted examples are bundled according to the verb and its closest case component.

3. The case frames are clustered using a similarity measure function, resulting in the final case frames. The similarity is calculated using a Japanese thesaurus [23], and its maximum score is 1.0. The details of the similarity measure function are described in [24].

First, modifier-head examples that had no syntactic ambiguity were extracted; they were disambiguated by coupling a predicate and its closest case component. In order to remove inappropriate modifier-head examples, the threshold α was introduced; only modifier-head examples that appeared no less than α times in the corpora were used.

The basic case frames were then clustered to merge similar case frames. For example, since *nimotsu-wo tsumu* (load baggage) and *busshi-wo tsumu* (load supplies) were similar, they were merged. Table 2.4 shows some of case frame examples.

2.3.3 Generalization of Examples

When using hand-crafted case frames, the data sparseness problem is serious; by using case frames automatically constructed from a large corpus, it was alleviated to some extent but not eliminated. For instance, there are thousands of named entities (NEs) that cannot be covered intrinsically. To deal with this problem, generalized examples were given for the case slots. Kawahara and Kurohashi also gave generalized examples but only for a few types [22]. In this thesis, case slot examples were generalized based upon common noun categories and NE classes.

First, the categories created by the Japanese morphological analyzer JUMAN[2] were added to each case slot. In JUMAN, about 20 categories have been defined and tagged to common nouns. For example, *ringo* (apple), *inu* (dog) and *byoin* (hospital) are tagged as FOOD, ANIMAL and FACILITY, respectively. For each category, the ratio of the categorized example among all case slot examples was calculated, and added to the case slot (e.g. [CT:FOOD]:0.07).

The NEs were also generalized. First, the NEs in the source corpus were recognized by using the NE recognizer that we will mention in Chapter 3; case frames were then constructed using the NE-recognized corpus. Similar to the categories, for each NE class, the NE ratio to all the case slot examples was calculated and added to the case slot (e.g. [NE:PERSON]:0.12). The generalized examples are also included in Table 2.4.

2.3.4 Case Frame Construction Using the Web

Using this gradual procedure, we constructed case frames from the web corpus [22]. For this studies, approximately 6 billion Japanese sentences consisting of approximately 100 billion words were acquired from 100 million Japanese web pages. After discarding duplicate sentences, which may have been extracted from mirror sites, a corpus was acquired comprising of 1.6 billion (1.6G) unique Japanese sentences consisting of approximately 25 billion words. Case frames were constructed from this corpus. As the threshold $\alpha = 10$ that introduced in Section 2.3.2, we set $\alpha = 10$.

Completing the case frame construction took about two weeks using 300 CPUs. As a result, we acquired about 1.6 million case frames. The number of unique predicates was 65,679, the average number of case frames for a predicate was 25.3, the average number of unique examples for a case slot was 9.64, and the average number of the kinds of generalized examples was 0.84.

[2]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

In Chapter 6, we will investigate the effect of corpus size on discourse analysis. Case frames were constructed from corpora of different sizes; and more detailed statistics of constructed case frames will be shown in Chapter 6.

2.4 Construction of Nominal Case Frame

2.4.1 Overview of Nominal Case Frame Construction

What is represented in a text has originally a network structure, in which several concepts have tight relations with each other. However, because of the linear constraint of texts, most of them disappear in the normal form of texts. One of such latent relationship is bridging reference (also called indirect anaphora or functional anaphora), such as the following examples.

(2.2) *Ken-wo katta.* *Kakaku-wa 20 doru datta.*
ticket bought price dollars was
(I bought a ticket. The price was 20 dollars.)

(2.3) Ie-ga atta. Yane-wa siro-katta.
house roof white
(There was a house. The roof was white.)

Here, “the price” means “the price of a ticket” and “the roof” means “the roof of a house,” and the reference of “the price” to “a ticket” and the reference of “the roof” to “a house” are called bridging reference.

Most nouns have their indispensable or requisite entities: “price” is a price of some goods or service, “roof” is a roof of some building, “coach” is a coach of some sport, and “virus” is a virus causing some disease. The relation between a noun and its indispensable entity is parallel to that between a verb and its arguments or obligatory cases. In this thesis, we call indispensable entities of nouns *obligatory cases*. Bridging reference resolution needs a comprehensive information or dictionary of obligatory cases of nouns. We call such information as nominal case frames, and construct them using raw corpus and several hand-crafted knowledge.

As mentioned in Section 2.3, in case of verbs, case markers such as *ga*, *wo*, and *ni* in Japanese, or as syntactic structures such as subject/object/PP in English can be utilized as a strong clue to distinguish several obligatory cases and adjuncts (and adverbs), which makes it feasible to construct case frames from large corpora automatically [24, 25].

On the other hand, in case of nouns, obligatory cases of noun N_h appear, in most cases, in the single form of noun phrase "N_h of N_m" in English, or "N_m *no* N_h" in Japanese. This single form can express several obligatory cases, and furthermore optional cases, such as "*rugby no coach*" (obligatory case concerning what sport), "*club no coach*" (obligatory case concerning which institution), and "*kyonen* 'last year' *no coach*" (optional case). Therefore, the key issue to construct nominal case frames is to analyze "N_h of N_m" or "N_m *no* N_h" phrases to distinguish obligatory case examples and others.

Work which addressed bridging reference in English texts so far restricts relationships to a small, relatively well-defined set, mainly part-of relation like the above example (2.2.4.1), and utilized hand-crafted heuristic rules or hand-crafted lexical knowledge such as WordNet [26–28]. Poesio et al. proposed a method of acquiring lexical knowledge from "N_h of N_m" phrases, but again concentrated on part-of relation [29].

In case of Japanese, Murata et al. proposed a method of utilizing "N_m *no* N_h" phrases as primitive nominal case frames for indirect anaphora, or bridging reference resolution of diverse relationships [30]. However, they basically used all "N_m *no* N_h" phrases from corpora, just excluding some pre-fixed stop words. They confessed that an accurate analysis of "N_m *no* N_h" phrases is necessary for acquiring better nominal case frames and the further improvement of bridging reference resolution.

In this thesis, following the work by Kurohashi and Sakai [31] for analysis of "N_m *no* N_h," we propose a method to construct Japanese nominal case frames from large corpora, based on an accurate analysis of "N_m *no* N_h" phrases using an ordinary dictionary and a thesaurus.

2.4.2 Semantic Analysis of Japanese Noun Phrases N_m *no* N_h

In many cases, obligatory cases of nouns are described in an ordinary dictionary for human being. For example, a Japanese dictionary for children, *Reikai Shougaku Kokugojiten*, or RSK [17], gives the definitions of the word *coach* and *virus* as follows:

coach a person who teaches technique in some sport

virus a living thing even smaller than bacteria which causes infectious disease like influenza

Note that, although our method handles Japanese noun phrases by using Japanese definition sentences, in this thesis we use their English translations for the explanation. In some sense, the essential point of our method is language-independent.

Based on such an observation, Kurohashi and Sakai [31] proposed a semantic analysis method of "N_m *no* N_h," consisting of the two modules: dictionary-based analysis (abbreviated to DBA hereafter) and semantic feature-based analysis (abbreviated to SBA hereafter). This section briefly introduces their method.

Semantic Feature Dictionary

We first briefly introduce NTT Semantic Feature Dictionary. NTT Semantic Feature Dictionary consists of a semantic feature tree, whose 3,000 nodes are semantic features, and a nominal dictionary containing about 300,000 nouns, each of which is given one or more appropriate semantic features.

The main purpose of using this dictionary is to calculate the similarity between two words. Suppose the word x and y have a semantic feature s_x and s_y, respectively, their depth is d_x and d_y in the semantic tree, and the depth of their lowest (most specific) common node is d_c, the similarity between x and y, $sim(x, y)$, is calculated as follows:

$$sim(x, y) = (d_c \times 2)/(d_x + d_y).$$

If s_x and s_y are the same, the similarity is 1.0, the maximum score based on this criteria.

Dictionary-based analysis

Obligatory case information of nouns in an ordinary dictionary can be utilized to solve the difficult problem in the semantic analysis of "N_m *no* N_h" phrases. In other words, we can say the problem disappears.

For example, "*rugby no coach*" can be interpreted by the definition of *coach* as follows: the dictionary describes that the noun *coach* has an obligatory case *sport*, and the phrase "*rugby no coach*" specifies that the *sport* is *rugby*. That is, the interpretation of the phrase can be regarded as matching *rugby* in the phrase to *some sport* in the definition sentence of *coach*. "*Kaze* 'cold' *no virus*" is also easily interpreted based on the definition of *virus*, linking *kaze* 'cold' to *infectious disease*.

Dictionary-based analysis (DBA) tries to find a correspondence between N_m and an obligatory case of N_h by utilizing RSK and NTT Semantic Feature Dictionary, by the following process:

1. Look up N_h in RSK and obtain the definition sentences of N_h.

Table 2.5: Examples of Rules for Semantic Feature-Based Analysis.

N_m:HUMAN, N_h:RELATIVE → <obligatory case(relative)>	e.g. *kare* 'he' *no oba* 'aunt'
N_m:HUMAN, N_h:HUMAN → <modification(apposition)>	e.g. *gakusei* 'student' *no kare* 'he'
N_m:ORGANIZATION, N_h:HUMAN → <belonging>	e.g. *gakkou* 'school' *no seito* 'student'
N_m:AGENT, N_h:EVENT → <agent>	e.g. *watashi* 'I' *no chousa* 'study'
N_m:MATERIAL, N_h:CONCRETE → <modification(material)>	e.g. *ki* 'wood' *no hako* 'box'
N_m:TIME, N_h:$*$ → <time>	e.g. *aki* 'autumn' *no hatake* 'field'
N_m:COLOR, QUANTITY, or FIGURE, N_h:$*$ → <modification>	e.g. *gray no seihuku* 'uniform'
N_m:$*$, N_h:QUANTITY → <obligatory case(attribute)>	e.g. *hei* 'wall' *no takasa* 'height'
N_m:$*$, N_h:POSITION → <obligatory case(position)>	e.g. *tsukue* 'desk' *no migi* 'right'
N_m:AGENT, N_h:$*$ → <possession>	e.g. *watashi* 'I' *no kuruma* 'car'
N_m:PLACE or POSITION, N_h:$*$ → <place>	e.g. *Kyoto no mise* 'store'

'$*$' meets any noun.

2. For each word *w* in the definition sentences other than the genus words, do the following steps:

   2.1. When *w* is a noun which shows an obligatory case explicitly, like *kotogara* 'thing', *monogoto* 'matter', *nanika* 'something', and N_m does not have a semantic feature of HUMAN or TIME, give 0.8 to their correspondence [3].

   2.2. When *w* is other noun, calculate the similarity between N_m and *w* by using NTT Semantic Feature Dictionary, and give the similarity score to their correspondence.

3. Finally, if the best correspondence score is 0.75 or more, DBA outputs the best correspondence, which can be an obligatory case of the input; if not, DBA outputs nothing.

In case of the phrase "*rugby no coach*," "technique" and "sport" in the definition sentences are checked: the similarity between "technique" and "rugby" is calculated to be 0.21, and then the similarity between "sport" and "rugby" is calculated to be 1.0. Therefore, DBA outputs "sport."

Semantic feature-based analysis

Since diverse relations in "N_m *no* N_h" are handled by DBA, the remaining relations can be detected by simple rules checking the semantic features of N_m and/or N_h.

[3]For the present, parameters in the algorithm were given empirically, not optimized by a learning method.

Table 2.6: Preliminary Case Frames for *hisashi* 'eaves/visor.'

DBA result	
1. A roof that stick out above the window of a house.	
[house]	hall:2, balcony:1, building:1, $\cdots$
[window]	window:2, ceiling:1, counter:1, $\cdots$
2. The fore piece of a cap.	
[cap]	cap:8, helmet:1, $\cdots$
SBA result	
<place>	parking:3, store:3, shop:2, $\cdots$
<modification>	concrete:1, metal:1, silver:1, $\cdots$
No semantic analysis result	
<other>	part:1, light:1, phone:1, $\cdots$

Table 2.5 shows examples of the rules. For example, the rule 1 means that if N_m has a semantic feature HUMAN and N_h RELATIVE, <obligatory case> relation is assigned to the phrase. The rules 1, 2, 8 and 9 are for certain obligatory cases. We use these rules because these relations can be analyzed more accurately by using explicit semantic features, rather than based on a dictionary.

Integration of two analyses

Usually, either DBA or SBA outputs some relation. When both DBA and SBA output some relations, the results are integrated. Basically, if DBA correspondence score is higher than 0.8, DBA result is selected; if not, SBA result is selected. In rare cases, neither analysis outputs any relations, which means analysis failure.

2.4.3 Automatic Construction of Nominal Case Frames

Collection and analysis of N_m *no* N_h

Syntactically unambiguous noun phrases "N_m *no* N_h" are collected from the automatic parse results of large corpora, and they are analyzed using the method described in the previous section.

By just collecting the analysis results of each head word N_h, we can obtain its preliminary case frames. Table 2.6 shows an example of preliminary case frames for *hisashi* 'eaves/visor.'

The upper part of the table shows the results by DBA. The line starting with "[house]" denotes a group of analysis results corresponding to the word "house" in the first definition sentence. For example, "hall *no* hisashi" occurs twice in the corpora, and they were analyzed by DBA to correspond to "house."

The middle part of the table shows the results by SBA. Noun phrases that have no semantic analysis result (analysis failure) are bundled and named <other>, as shown in the last part of the table.

A case frame should be constructed for each meaning (definition) of N_h, and groups starting with "[. . .]" or "<. . . >" in Table 2.6 are possible case slots. The problem is how to arrange the analysis results of DBA and SBA and how to distinguish obligatory cases and others. The following sections explain how to handle these problems.

Case slot clustering

One obligatory case might be separated in preliminary case frames, since the definition sentence is sometimes too specific or too detailed. For example, in the case of *hisashi* 'eaves/visor' in Table 2.6, [house], [window], and <place> have very similar examples that mean building or part of building. Therefore, case slots are merged if similarity of two case slots is more than 0.5 (case slots in different definition sentences are not merged in any case). Similarity of two case slots is the average of top 25% similarities of all possible pairs of examples.

In the case of Table 2.6, the similarity between [house] and [window] is 0.80, and that between [house] and <place> is 0.67, so that these three case slots are merged into one case slot.

Obligatory case selection

Preliminary case frames contain both obligatory cases and optional cases for the head word. Since we can expect that an obligatory case co-occurs with the head word in the form of noun phrase frequently, we can take frequent case slots as obligatory case of the head word.

However, we have to be careful to set up the thresholds of frequency proportion, because case slots detected by DBA or <obligatory case> by SBA are more likely to be obligatory; on the other hand case slots of <modification> or <time> should be always optional. Considering these tendencies and observing the proportions of about 30 nouns, we set thresholds for

Table 2.7: Threshold to Select Obligatory Slots.

type of case slots	threshold of proportion	
analyzed by DBA	0.42% (1/240)	if corresponded to the first word in definition sentences
analyzed by DBA	1.25% (1/80)	otherwise
<obligatory case>	5.0% (1/20)	
<belonging>	1.7% (1/60)	
<possessive>	5.0% (1/20)	
<agent>	1.0% (1/100)	
<place>	10% (1/10)	
<other>	10% (1/10)	
<modification>	not used	
<time>	not used	

obligatory cases as shown in Table 2.7. In this table, the proportion is calculated as follows:

$$proportion = \frac{C(\text{“}N_m \boldsymbol{no} N_h'')}{C(\text{“}N_h'')}$$

($C(N)$ means the frequency of N in the corpus.)

In the case of *hisashi* ‘eaves/visor’ in Table 2.6, [house-window]-<place> slot and [cap] slot are chosen as the obligatory cases.

Case frame construction for each meaning

Case slots that are derived from each definition sentence constitute a case frame.

If a case slot of <obligatory case> by SBA or <other> is not merged into case slots in definition sentences, it can be considered that it indicates a meaning of N_h which is not covered in the dictionary. Therefore, such a case slot is eligible to constitute an independent case frame.

On the other hand, when other case slots by SBA such as <belonging> and <possessive> are remaining, we have to treat them differently. The reason why they are remaining is that they are not always described in the definition sentences, but their frequent occurrences indicate they are obligatory cases. Therefore, we add these case slots to the case frames derived from definition sentences.

Table 2.8 shows several examples of resultant case frames. *Hyoujou* ‘expression’ has a case frame containing two case slots. *Hisashi* ‘eaves/visor’ has two case frames according to the two definition sentences. In case of *hikidashi* ‘drawer,’ the first case frame corresponds

Table 2.8: Examples of Nominal Case Frames.

	case slot	examples
hisashi:1 'eaves/visor'	(the edges of a roof that stick out above the window of a house etc.)	
	[house, window]	parking, store, hall, $\cdots$
hisashi:2 'eaves/visor'	(the fore piece of a cap.)	
	[cap]	cap, helmet, $\cdots$
hyoujou 'expression'	(to express one's feelings on the face or by gestures.)	
	[one]	people, person, citizen, $\cdots$
	[feelings]	relief, margin, $\cdots$
hikidashi:1 'drawer'	(a boxlike container in a desk or a chest.)	
	[desk, chest]	desk, chest, dresser, $\cdots$
hikidashi:2 'drawer'	<other>	credit, fund, saving, $\cdots$
coach	(a person who teaches technique in some sport.)	
	[sport]	baseball, swimming, $\cdots$
	<belonging>	team, club, $\cdots$
kabushiki 'stock'	(the total value of a company's shares.)	
	[company]	company, corporation, $\cdots$

to the definition given in the dictionary, and the second case frame was constructed from the <other> case slot, which is actually another sense of *hikidashi*, missed in the dictionary. In case of *coach*, <possessive> is added to the case frame which was made from the definition, producing a reasonable case frame for the word.

Point of nominal case frame construction

The point of our method is the integrated use of a dictionary and example phrases from large corpora. Although dictionary definition sentences are informative resource to indicate obligatory cases of nouns, it is difficult to do indirect anaphora resolution by using a dictionary as it is, because all nouns in a definition sentence are not an obligatory case, and only the frequency information of noun phrases tells us which is the obligatory case. Furthermore, sometimes a definition is too specific or detailed, and the example phrases can adjust it properly, as in the example of *hisashi* in Table 2.6.

On the other hand, a simple method that just collects and clusters "N_m *no* N_h" phrases (based on some similarity measure of nouns) cannot construct comprehensive nominal case frames, because of polysemy and multiple obligatory cases. We can see that dictionary definition can guide the clustering properly even for such difficult cases.

Case frame for compound nouns and generalization

In this thesis, we attempt to construct case frames not only for nouns but also for compound nouns. The indispensable entities of a compound noun are not always included in the indispensable entities of composing nouns of the compound noun, and case frames for compound nouns are considered to be useful for such compound nouns. For example, while the information about destination is considered to be indispensable for the compound noun "saishu bus" (last bus) in (2.4), it is not for the noun "bus" (bus).

(2.4) Kiyomizu iki-no saishu bus.
bound last bus
(The last bus bound for Kiyomizu.)

In addition, case slot examples of nominal case frames were also generalized based upon common noun categories and NE classes in the same way as case frames of verbs.

2.4.4 Nominal Case Frame Construct Using Web

We constructed nominal case frames from the same 1.6 billion sentences as is used for case frame construction for verbs. In this corpus, there were about 390 million noun phrases "N_m *no* N_h," about 100 million unique noun phrases, and about 17 million unique head nouns. There are about 4.07 million head nouns that appear more than 10 times in the corpus, and we use only such head nouns.

The resultant nominal case frames consisted of about 564,000 nouns including compound nouns. The average number of case frames for a noun was 1.0031, and the average number of case slots for a case frame was 1.0101. However, these statistics were differs with the frequency of the noun. Therefore, we investigated the statistics of constructed nominal case frames for each group classified by the frequency of the nouns, 1-100, 101-1,000, 1,001-10,000, 10,001-100,000 and the others. Table 2.9 shows the result.

As for the 10,000 most frequently appeared nouns, which occupy about 70% of all noun appearance, the average number of case frames for a noun was 1.11, and the average number of case slots for a case frame is 1.17.

In order to evaluate the resultant case frames, we randomly selected 100 nouns from the 10,000 most frequent nouns, and created gold standard case frames for the nouns by hand. Table 2.10 shows the selected nouns. For each noun, case frames were given if the noun was

Table 2.9: The Details of Constructed Nominal Case Frames.

Rank of frequency of nouns	Proportion of nouns that have case frames	Average number of case slots for a noun that have case frames	Average number of case slots for a case frame	Proportion of the frequency against all nouns
1-100	0.560	1.34	1.07	0.173
101-1000	0.688	1.17	1.16	0.256
1001-10000	0.517	1.11	1.17	0.270
10001-100000	0.148	1.05	1.13	0.176
100001-	0.137	1.0009	1.0053	0.125
all (1-4,074,038)	0.139	1.0031	1.0101	1.000

considered to have any indispensable entity, and for each case frame, obligatory case slots were given manually. As a result, 70 case frames were created that had 82 case slots, that is, 58 case frames had only one case slot, the other 12 case frames had two case slots. 30 nouns had no case frames.

We first evaluated each noun. For nouns that have case frames, we regard as correct result if the correspond case frame were created. For nouns that have no case frame, we regard as correct result if the target noun's case frame was not created. The accuracy in Table 2.11 shows the result. We obtained proper result for 70 nouns. Then, we evaluated automatically constructed case slots for these selected nouns. The evaluation result is also shown in Table 2.11: the system output 70 case slots, and out of them, 62 case frames were judged as correct. The F-measure was 0.81. Since the boundary between indispensable case and optional case of a noun is not obvious, this score is considered to be reasonable.

2.5 Summary of this Chapter

In this chapter, we first proposed a method for acquiring synonym knowledge. As resources for synonym extraction, we use parenthesis expressions in raw corpus, and dictionary definition sentences. We extracted synonyms using Japanese newspaper articles in 26 years, and *Reikai Shougaku Kokugojiten* [17] and *Iwanami Kokugo Jiten* [18]. As a result, we acquired 2,792 synonym pairs with very high precision.

Then we constructed case frames for verbs by using Kawahara and Kurohashi's method [22] from 1.6 billion sentences acquired from 100 million Japanese web pages. As a result, we acquired about 1.6 million case frames for about 65 thousand verbs.

Table 2.10: Randomly Selected 100 Nouns.

case frame	nouns
2 case slots	*senshu* (player), *kojo* (factory), *mitsumori* (estimate), *seizo* (manufacture), *mastsu* (friction), *ureshisa* (joy), *shohosen* (formula), *soshitsu* (potential), *kokoroatari* (guess), *shirushi* (mark), *ondo* (initiative), *sekimu* (obligation)
1 case slot	*kakunin* (confirmation), *houkosei* (directivity), *sankasha* (participant), *kumiawase* (combination), *kakuho* (securement), *keijo* (form), *yoji* (engagement), *kanshoku* (touch), *henshu* edting, *bengoshi* (lawyer), *kinchokan* (tension), *kujo* (complant), *yubinkyoku* (post office), *shijutu* (execution), *gai* (harm), *reigai* (exception), *anteisei* (stability), *parameta* (parameter), *hirune* (nap), *danna* (husband), *koraboreshon* (collaboration), *shatai* (body), *yusen juni* (order of priority), *senryoku* (military strength), *byoshitsu* (sickroom), *naitei* (informal decision), *seme* (discipline), *kubiwa* (collar), *genryo* (wight-loss), *regura* (regular), *taiso* (exercise), *sougei* (pickup), *nukege* (fallen hair), *naigai* (within and without), *bougyo* (defense), *saiken* (reconstruction), *roukyuka* (obsolescence), *zensoku* (asthma), *gakucho* (president), *shokai-bun* (introduction), *todoke-saki* (destination), *koujou-shin* (aspiration), *handanryoku* (discretion), *fuyu-yasumi* (winter vacation), *shi* (master), *todome* (kibosh), *omutsu* (diaper), *yuuretsu* (superiority or inferiority), *nafuda* (name plate), *hon-keiyaku* (formal ocntract), *kikendo* (risk), *kouryakubon* (book for conquest), *genryu* (headwater), *haka-mairi* (a visit to a grave), *seikatsu-rizumu* (rhythm for life), *yaruki* (enthusiasm), *kyosei-chiryo* (remedy), *ni-bu* (second part)
no case frame	*kohi* (cafe), *Hokkaido* (Hokkaido), *youfuku* (dress), *kenbi-kyo* (microscope), *Tokyo-eki* (Tokyo Station), *maccha* (ceremonial tea), *homu stei* (homestay), *kaisou* (seaweed), *heichi* (flat land), *Uki* (Uki), *BMW* (BMW), *shoku-pan* (bread), *chusho-kigyo* (smaller enterprises), *miso* (miso), *asahi* (the rising sun), *iPod* (iPod), *fuben* (inconvenience), *honken* (the matter in hand), *bukkyo* (Buddhism), *wake* (division), *koresuterouru* (cholesterol), *tihou-jichitai* (local auhtority), *ON* (ON), *futugou* (inconvenience), *kobara* (stomach), *shishitsu* (lipid), *kawa* (river), *barentain-dei* (Saint Valentine's Day), *kawara* (riverbank), *kouseido* (high precision)

Table 2.11: Evaluation Result of Case Frames.

accuracy	precision	recall	F-measure
70/100 (0.70)	62/70 (0.89)	62/84 (0.74)	0.81

Lastly, we proposed an automatic construction method of nominal case frames. This method is based on semantic analysis of noun phrases "N_m *no* N_h" (N_h of N_m). The point of our method is the integrated use of a dictionary and example phrases from large corpora. We constructed nominal case frames from 1.6 billion sentences, and acquired about 566 thousand nominal case frames for about 564 thousand nouns.

Chapter 3

Named Entity Recognition Using Non-Local Information

3.1 Introduction

Named entity recognition (NER) is the task of identifying and classifying phrases into certain classes of named entities (NEs), such as names of persons, organizations and locations, expressions of times, quantities, etc. [32–34]. We can say NER is a fundamental task for several natural language processing areas, including machine translation, information retrieval and anaphora resolution, and NER system with high previous and accuracy could benefit the performance of anaphora resolution.j In this chapter, we describe named entity recognition using non-local information.

Japanese texts, which we focus on, are written without using blank spaces. Therefore, Japanese NER has tight relation with morphological analysis, and thus it has been often performed immediately after morphological analysis [3,5]. However, such approaches rely only on local context. The Japanese NER system proposed by Nakano and Hirai [35], which achieved the highest F-measure among conventional systems, introduced the *bunsetsu*[1] feature in order to consider wider context, but considers only adjacent *bunsetsus*.

On the other hand, as for English or Chinese, various NER systems have explored global information and reported their effectiveness. For examples, Malouf [36] and Chieu and Ng [37] utilized information about features assigned to other instances of the same token. Ji and Grishman [38] used the information obtained from coreference analysis for NER. Mohit and Hwa [39] used syntactic features in building a semi-supervised NE tagger. Finkel et al. [40]

[1] *Bunsetsu* is a commonly used linguistic unit in Japanese, consisting of one or more adjacent content words and zero or more following functional words.

Table 3.1: Definition of NE in IREX.

NE class	Examples
ORGANIZATION	NHK Symphony Orchestra
PERSON	Kawasaki Kenjiro
LOCATION	Rome, Sinuiju
ARTIFACT	Nobel Prize
DATE	July 17, April this year
TIME	twelve o'clock noon
MONEY	sixty thousand dollars
PERCENT	20%, thirty percents

and Krishnan and Manning [41] proposed NER systems that use non-local information, such as label consistency.

In this chapter, we describe a Japanese NER system that uses global information obtained from several structural analyses. To be more specific, our system is based on SVM, recognizes NEs after syntactic, case and coreference analyses and uses information obtained from these analyses and the NER results for the previous context, integrally. At this point, it is true that NER results are useful for syntactic, case and coreference analyses, and thus these analyses and NER should be performed in a complementary way. However, since we focus on NER, we recognize NE after these structural analyses.

3.2 Japanese NER Task

A common standard definition for Japanese NER task is provided by IREX workshop [33]. IREX defined eight NE classes as shown in Table 3.1. Compared with the MUC-6 NE task definition [34], the NE class "ARTIFACT," which contains book titles, laws, brand names and so on, is added.

NER task can be defined as a chunking problem to identify token sequences that compose NEs. The chunking problem is solved by annotating chunk tags to tokens. Five chunk tag sets, IOB1, IOB2, IOE1, IOE2 and IOBES are commonly used [42]. In this thesis, we use the IOBES model, in which "S" denotes a chunk itself, and "B," "I" and "E" denote the beginning, intermediate and end parts of a chunk. If a token does not belong to any named entity, it is tagged as "O." Since IREX defined eight NE classes, tokens are classified into 33 (= $8 \times 4 + 1$) NE tags. For example, NE tags are assigned as following:

(3.1) *Kotoshi* *4* *gatsu* *Roma* *ni itta.*
this year April Rome to went
B-DATE I-DATE E-DATE S-LOCATION O O
(ϕ went to Rome on April this year.)

3.3 Motivation for Our Approach

Our NER system utilizes non-local information. In this section, we describe the motivation for our approach.

High-performance Japanese NER systems are often based on supervised learning, and most of them use only local features, such as features obtained from the target token, two preceding tokens and two succeeding tokens. However, in some cases, NEs cannot be recognized by using only local features.

For example, while "*Kawasaki*" in the second sentence of (3.2) is the name of a person, "*Kawasaki*" in the second sentence of (3.3) is the name of a soccer team. However, the second sentences of (3.2) and (3.3) are exactly the same, and thus it is impossible to correctly distinguish these NE classes by only using information obtained from the second sentences.

(3.2) *Kachi-ha senpatsu-no* ***Kawasaki*** *Kenjiro.*
winner starter
Kawasaki-ha genzai 4 shou 3 pai.
now won lost
(The winning pitcher is the starter Kenjiro **Kawasaki**. Kawasaki has won 4 and lost 3.)

(3.3) *Dai 10 setsu-wa* ***Kawasaki*** *Frontale-to taisen.*
the round against
Kawasaki-ha genzai 4 shou 3 pai.
now won lost
(The 10th round is against **Kawasaki** Frontale. Kawasaki has won 4 and lost 3.)

In order to recognize these NE classes, it is essential to use the information obtained from the previous context. Therefore, we utilize information obtained from the analysis of the previous context: **cache feature** and **coreference relation**.

For another example, "*Shingishu*" in (3.4) is the name of city in North Korea. The most important clue for recognizing "*Shingishu*" as "LOCATION" may be the information obtained from the head verb, "*wataru* (get across)."

(3.4) *Shingishu-kara Ouryokko-wo* ***wataru***.
Sinuiju from Amnokkang get across
(ϕ **gets across** the Amnokkang River from Sinuiju.)

However, since the dictionary for morphological analysis has no entry "*Shingishu*," "*Shingishu*" is analyzed as consisting of three morphemes: "*shin*," "*gi*" and "*shu*"; and when using only local features, the word "*wataru*" is not taken into consideration because there are more than two morphemes between "*shu*" and "*wataru*." In order to deal with such problem, we use the information obtained from the head verb: **syntactic feature** and **case frame feature**.

3.4 NER Using Non-local Information

3.4.1 Outline of Our NER System

Our NER system performs the chunking process based on morpheme units because character-based methods do not outperform morpheme-based methods [3] and are not suitable for considering wider context.

A wide variety of trainable models have been applied to Japanese NER task, including maximum entropy (ME) models [1], support vector machines (SVM) [5, 35] and conditional random fields (CRF) [4]. Since, for Japanese NER, SVM-based systems achieved higher F-measure than the other systems, our system applies SVMs. Isozaki and Kazawa [2] proposed an SVM-based NER system with Viterbi search, which outperforms an SVM-based NER system with sequential determination, and our system basically follows this system. Our NER system consists of the following four steps:

1. Morphological analysis
2. Syntactic, case and coreference analyses
3. Feature extraction for chunking
4. SVM and Viterbi search based chunking

The following sections describe each of these steps in detail.

Input sentence:

Gai	*mu*	*sho*	*no*	*shin*	*Bei*	*ha*	.
foreign affairs		ministry	in	pro	America	group	

(Pro-America group in the Ministry of Foreign Affairs.)

Output of JUMAN:

Gaimu	*sho*	*no*	*shin*	*Bei*	*ha*	.
noun	noun	particle	noun	noun	noun	

Output of ChaSen:

Gaimusho	*no*	*shin-Bei*	*ha*	.
noun	particle	noun	noun	

Figure 3.1: Example of Morphological Analyses.

3.4.2 Morphological Analysis

While most existing Japanese NER systems use ChaSen [43] as a morphological analyzer, our NER system uses a Japanese morphological analyzer JUMAN [44] because of the following two reasons.

First, JUMAN tends to segment a sentence into smaller morphemes than ChaSen, and this is a good tendency for morpheme-based NER systems because the boundary contradictions between morphological analysis and NEs are considered to be reduced. Figure 3.1 shows an example of the outputs of JUMAN and ChaSen. Although both analyses are reasonable, JUMAN divided "*Gaimusho*" and "*shin-Bei*" into two morphemes, while ChaSen left them as a single morpheme. Second, JUMAN adds categories to some morphemes, which can be utilized for NER. In JUMAN, about thirty categories are defined and tagged to about one fifth of morphemes. For example, "*ringo* (apple)," "*inu* (dog)" and "*byoin* (hospital)" are tagged as "FOOD," "ANIMAL" and "FACILITY," respectively.

In addition, in order to cope with the boundary contradictions between the morphological analysis and NEs, after morphological analysis by JUMAN, some morphemes are divided into two or three morphemes by using simple rules made from the learning data.

3.4.3 Syntactic, Case and Coreference Analyses

syntactic analysis Syntactic analysis is performed by using the Japanese parser KNP [12]. KNP employs some heuristic rules to determine the head of a modifier.

case analysis Case analysis is performed by using the system proposed by Kawahara and Kurohashi [24]. This system uses Japanese case frames that are automatically constructed from a large corpus. To utilize case analysis for NER, we constructed case frames that include NE labels in advance. We explain details in Section 3.4.4. The case analysis is applied to each predicate in an input sentence. For details see [24].

coreference analysis Coreference analysis is performed by using the coreference analyzer described in Chapter 4. As will be mentioned in Section 3.4.4, our NER system uses coreference relations only when coreferential expressions do not share the same morphemes. Basically, such coreference relations are recognized by using automatically acquired synonym knowledge.

3.4.4 Feature Extraction

Basic Features

As basic features for chunking, our NER system uses the morpheme itself, character type, part of speech (POS) tag. category of the morpheme if it exists, and the head morpheme of the bunsetsu.

As character types, we defined seven types: "*kanji*," "*hiragana*," "*katakana*," "*kanji* with *hiragana*," "punctuation mark," "alphabet" and "digit." As for POS tag, more than one POS feature are extracted if the target morpheme has POS ambiguity. In addition, besides POS tag obtained by JUMAN, our system also uses POS tag obtained from Japanese morphological analyzer MeCab[2] that uses IPADIC as a word dictionary [45]. The JUMAN dictionary has few named entity entries; thus our system supplements the lack of lexical knowledge by using MeCab.

As categories, we use the categories that Japanese morphological analyzer JUMAN[3] adds to common nouns. In JUMAN, about twenty categories are defined and tagged to common nouns. For example, "*ringo* (apple)," "*inu* (dog)" and "*byoin* (hospital)" are tagged as "FOOD," "ANIMAL" and "FACILITY," respectively.

Non-local Features

Our NER system uses three types of non-local features: cache features, syntactic features and case frame features, and one rule that reflects coreference relations. Although the coreference

[2]http://mecab.sourceforge.jp/

[3]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

relations are not used as features but as a rule, we also describe how to use the coreference relations in this section.

cache feature If the same morpheme appears multiple times in a single document, in most cases the NE tags of these morphemes have some relation to each other, and the NER results for previous parts of the document can be a clue for the analysis for following parts.

We consider the examples (3.2) and (3.3) again. Although the second sentences of (3.2) and (3.3) are exactly the same, we can recognize "*Kawasaki*" in the second sentence of (3.2) is "S-PERSON" and "*Kawasaki*" in the second sentence of (3.3) is "S-ORGANIZATION" by reading the first sentences.

In order to utilize the information obtained from previous parts of the document, our system uses the NER results for previous parts of the document as features, called cache features. When analyzing (3.2), our system uses the outputs of NE recognizer for "*Kawasaki*" in the first sentence as a feature for "*Kawasaki*" in the second sentence. For simplicity, our system uses correct NE tags when training. That is, as a feature for "*Kawasaki*" in the second sentence of (3.2), the correct feature "B-PERSON" is always added when training, not always added when analyzing.

coreference rule Coreference relation can be a clue for NER. This clue is considered by using cache features to a certain extent. However, if the same morpheme is not used, cache features cannot work.

For example, "*NHK kokyo gakudan*" and "*N-kyo*" in (5) have coreference relation, but they do not share the same morpheme.

(3.5) ***NHK kokyo*** ***gakudan****-no ongaku kantoku-ni shuunin.*
symphony orchestra musical director became

N-kyo-to *kyoen-shite* *irai ... .*
perform together since

(He became musical director of the **NHK Symphony Orchestra**.
Since performing together with N-kyo ... .)

In this case, "*NHK kokyo gakudan*" can easily be recognized as "ORGANIZATION," because it ends with "*kokyo gakudan* (symphony orchestra)." Meanwhile, "*N-kyo*," the abbreviation of

"*NHK kokyo gakudan*," cannot easily be recognized as "ORGANIZATION."

Therefore, our system uses a heuristic rule that if a morpheme sequence is analyzed to be coreferential to a previous morpheme sequence that is recognized as an NE class, the latter morpheme sequence is recognized as the same NE class. Since this heuristic rule is introduced in order to utilize the coreference relation that is not reflected by cache features, our system applies this rule only when coreferential expressions do not have any morphemes in common.

syntactic feature As mentioned in Section 3.3, our system utilizes the information obtained from the head verb. As syntactic features, our system uses the head verb itself and the surface case of the *bunsetsu* that includes the target morpheme.

For the morpheme "*shin*" in example (3.4), the head verb "*wataru* (get across)" and the surface case "*kara* (from)" are added as syntactic features.

case frame feature Syntactic features cannot work if the head verb does not appear in the training data. To overcome this data sparseness problem, case frame features are introduced.

For example, although the head verb "*haken* (dispatch)" can be a clue for recognizing "*ICAO*" in (3.6) as "ORGANIZATION," syntactic features cannot work if "*haken* (dispatch)" did not appear in the training data.

(3.6) *ICAO-ha genchi-ni senmonka-wo* ***haken****-shita.*
scene to expert dispatched
(*ICAO* **dispatched** experts to the scene)

However, this clue can be utilized if there is knowledge that the "*ga* (nominative)" case of "*haken* (dispatch)" is often assigned by "ORGANIZATION."

Therefore, we construct case frames that include NE labels in advance. Case frames describe what kinds of cases each verb has and what kinds of nouns can fill a case slot, and is explained details in Section 2.3. We construct case frames from about five hundred million sentences. We first recognize NEs appearing in the sentences by using a primitive NER system that uses only local features, and then construct the case frames from the NE-recognized sentences. To be more specific, if one tenth of the examples of a case are classified as a certain NE class, the corresponding label is attached to the case. Table 3.2 shows the constructed case frame of "*haken* (dispatch)." In the "*ga* (nominative)" case, the NE labels, "ORGANIZATION" and "LOCATION" are attached.

Table 3.2: Case Frame of "*haken* (dispatch)."

case	examples
ga (nominative)	Japan:23, party:13, country:12, government:7, company6, ward:6, corps:5, UN:4, US:4, Korea:4, team:4, . . . (ORGANIZATION, LOCATION)
wo (accusative)	party:1249, him:1017, soldier:932, official:906, company:214, instructor:823, expert:799, helper:694, staff:398, army:347, . . .
ni (locative)	Iraq:700, on-the-scene:576, abroad:335, home:172, Japan:171, Indirect Ocean:142, scene:141, China:125, . . . (LOCATION)
kara (from)	Japan:61, party:11, company:9, prefecture:8, city:7, center:7, ministry:7, town:6, group:6, . . . (LOCATION, ORGANIZATION)
e (to)	Iraq:219, abroad:93, city:88, company:68, on-the-scene:66, Japan:47, area:43, China:34, . . . (LOCATION)
	⋮

We then explain how to utilize these case frames. Our system first performs case analysis, and uses as case frame features the NE labels attached in the case to which the target morpheme is assigned. For instance, by the case analyzer, the postpositional particle "*-ha*" in (3.6) is recognized as meaning nominative and "*ICAO*" is assigned to the "*ga* (nominative)" case of the case frame of "*haken* (dispatch)."Therefore, the case frame features, "ORGANIZATION" and "LOCATION" are added to the features for the morpheme "*ICAO*."

3.4.5 SVM and Viterbi Search Based Chunking

In order to utilize cache features obtained from the previous parts of the same sentence, our system determines NE tags clause by clause. The features extracted from two preceding morphemes and two succeeding morphemes are also used for chunking a target morpheme. Since SVM can solve only a two-class problem, we have to extend a binary classifier SVM to n-class classifier. Here, we employ the one versus rest method, in which we prepared n binary classifiers and each classifier is trained to distinguish a class from the rest of the classes.

To consider consistency of NE tags in a clause, our system uses Viterbi search with some constraints such as a "B-DATE" must be followed by "I-DATE" or "E-DATE." Since SVMs do

Table 3.3: Experimental Results (F-measure).

	CRL		IREX		WEB	
baseline	88.63		85.47		68.98	
+ cache feature	88.81	+0.18*	85.94	+0.47	69.67	+0.69*
+ coreference rule	88.68	+0.05	86.52	+1.05***	69.17	+0.19
+ syntactic feature	88.80	+0.17*	85.77	+0.30	70.25	+1.27**
+ case frame feature	88.57	−0.06	85.51	+0.04	70.12	+1.14*
use all no-local information	89.21	+0.58***	86.98	+2.25***	**71.27**	+2.05***
(without case frame feature)	(89.09	+0.46***)				
+ thesaurus	89.40	+0.77***	**87.72**	+2.25***	71.03	+2.05***
(without case frame feature)	(**89.43**	+0.80***)				

significant at the 0.1 level:*, 0.01 level:**, 0.001 level:***

not output probabilities, our system uses the SVM+sigmoid method [46]. That is, a sigmoid function:

$$s(x) = \frac{1}{(1 + exp(-\beta x))} \tag{i}$$

is applied to map the output of SVM to a probability-like value. Our system determines NE tags by using these probability-like values. Our system is trained by TinySVM-0.09[4] with C = 0.1 and uses a fixed value β = 10. This process is almost the same as the process proposed by Isozaki and Kazawa and for details see [2].

3.5 Experiments

3.5.1 Experimental Setting

For training, we use CRL NE data, which was prepared for IREX. CRL NE data has 18,677 NEs on 1,174 articles in Mainichi Newspaper.

For evaluation, we use three data: CRL NE data, IREX's formal test data called GENERAL and WEB NE data. When using CRL NE data for evaluation, we perform five-fold cross-validation. IREX test data has 1,510 NEs in 71 articles from Mainichi Newspaper. Although both CRL NE data and IREX test data use Mainichi Newspaper, these formats are not the same. For example, CRL NE data removes parenthesis expressions, but IREX test data does not. WEB NE data, which we annotated NEs on corpus collected from the Web, has 1,686 NEs

[4]http://chasen.org/ taku/software/TinySVM/

in 354 articles. Although the domain of the web corpus differs from that of CRL NE data, the format of the web corpus is the same as CRL NE data format.

3.5.2 Experiments and Discussion

To confirm the effect of each feature, we conducted experiments on seven conditions as follows:

1. Use only basic features (baseline)
2. Add cache features to baseline
3. Add the coreference rule to baseline
4. Add parent features to baseline
5. Add case frame features to baseline
6. Use all non-local information
7. Use all non-local information and thesaurus

Since Asahara and Matsumoto [3] and Nakano and Hirai [35] reported the performance of NER system was improved by using a thesaurus, we also conducted experiment in which semantic classes obtained from a Japanese thesaurus "*Bunrui Goi Hyo*" [47] were added to the SVM features. Table 3.3 shows the experimental results.

In experiments using case frame features, case frames that constructed from about five hundred million sentences were applied. Since these case frames constructed by using a primitive NER system learned from CRL NE data, the experiments that used case frame features on CRL data were not strictly open experiments. Therefore, for CRL data, we also conducted experiments on conditions of using all non-local information except case frame features.

To judge the statistical significance of the differences between the performance of the baseline system and that of the others, we conducted a McNemar-like test. First, we extract the outputs that differ between the baseline method and the target method. Then, we count the number of the outputs that only baseline method is correct and that only target method is correct. Here, we assume that these outputs have the binomial distribution and apply binomial test. As significance level, we use 0.1 level, 0.01 level and 0.001 level. The results of the significance tests are also shown in Table 3.3.

Table 3.4: Experimental Results of Each NE Types When Using Baseline Features.

	CRL		IREX		WEB	
	recall	precision	recall	precision	recall	precision
ORGANIZATION	81.07	85.90	73.13	82.24	43.82	57.07
PERSON	88.02	91.03	88.46	90.88	69.30	69.72
LOCATION	89.64	91.07	86.68	87.32	73.28	76.40
ARTIFACT	43.51	68.86	35.42	34.00	21.92	55.17
DATE	94.11	94.09	93.46	95.29	94.01	89.00
TIME	90.24	92.07	94.44	92.73	57.50	76.67
MONEY	92.82	97.31	100.00	100.00	89.74	89.74
PERCENT	97.36	97.96	100.00	95.45	80.00	84.21
ALL	86.91	90.42	83.07	87.03	64.62	73.97
F-measure	88.63		85.47		68.98	

Table 3.5: Experimental Results of Each NE Types When Using All Information.

	CRL		IREX		WEB	
	recall	precision	recall	precision	recall	precision
ORGANIZATION	82.40	86.89	79.22	81.25	44.57	68.79
PERSON	88.72	92.86	93.49	92.94	68.39	77.05
LOCATION	90.15	91.56	86.44	89.03	75.65	76.97
ARTIFACT	44.18	69.18	35.42	34.00	23.29	69.86
DATE	93.61	94.19	92.31	94.49	93.31	90.14
TIME	89.64	92.40	94.44	98.08	52.50	77.78
MONEY	92.82	97.84	100.00	100.00	89.74	92.11
PERCENT	95.73	97.31	100.00	95.45	80.00	66.67
ALL	87.34	91.15	86.29	87.69	65.16	78.65
F-measure	89.21		86.98		71.27	

When comparing the performance between data sets, we can say that the performance for WEB NE data is much worse than the others. This may be because the domain of the WEB corpus differs from that of CRL NE data, which is also reported in [48].

Table 3.4 shows the detail of the experimental results when using only baseline features and Table 3.5 shows the detail of the experimental results when using all non-local information. We confirm that non-local features mainly increased the performance for proper nouns, such as names of persons and organizations.

Table 3.6: Comparison with Previous Work (F-measure).

	CRL cross validation	IREX test data	Learning Method	Analysis Units	Features
Utsuro et al. [1]		84.07*	ME	morpheme	stacked generalizer
Isozaki & Kazawa [2]	86.77	85.10	SVM+Viterbi	morpheme	basic features
Asahara & Matsumoto [3]	87.21		SVM	character	+thesaurus
Fukuoka [4]	87.71		Semi-Markov CRF	character	basic features
Yamada [5]	88.33		SVM+Shift-Reduce	morpheme	+bunsetsu features
Kazama & Torisawa [49]	88.93		CRF	character	inducing gazetteer by large-scale clustering
Nakano & Hirai [35]	89.03		SVM	character	+bunsetsu features +thesaurus
Fukushima et al. [50]	89.29		SVM	character	+bunsetsu +entity list extracted from Web
Our system	**89.43**	**87.72**	SVM+Viterbi	morpheme	+non-local information +thesaurus

*Not considering NEs whose segmentation boundaries are not in morphemes.

As for the differences in the same data set, cache features and syntactic features improve the performance not dramatically but consistently and independently from the data set. The coreference rule also improves the performance for all data sets, but especially for IREX test data. This may be because IREX test data does not remove parenthesis expressions, and thus there are a many coreferential expressions in the data. Case frame features improve the performance for WEB NE data, but do not contribute to the performance for CRL NE data and IREX test data. This result shows that case frame features are highly generalized features and effective for data of different domain. On the other hand, thesaurus features improve the performance for CRL NE data and IREX test data, but worsen the performance for WEB NE data. The main cause for this may be overfitting to the domain of the training data.

By using all non-local information, the performance is significantly improved for all data sets, and thus we can say that the non-local information improves the performance of NER.

3.5.3 Comparison with Previous Work

Table 3.6 shows the comparison with previous work for CRL NE data and IREX test data. Since there was not any previous work that using WWW data, the performance for WWW data was not compared. Our system outperforms all other systems including Fukushima et al. [50] and Isozaki and Kazawa [2], which achieved highest F-measure for CRL NE data and IREX test

Table 3.7: Contribution of Each Feature to the Baseline Model.

Condition	CRL	
Baseline	88.63	
−categories	88.08	−0.55
−head morpheme	88.48	−0.15
−MeCab	87.34	−1.29
Without these three information	86.65	−1.98

data respectively, and thus we can confirm the effectiveness of our approach.

Factors responsible for the high performance of our system include not only the non-local features but also the relatively high baseline performance. Especially, against CRL NE data the F-measure of our system was 1.8 points higher than that of Isozaki and Kazawa's system that is basis of our system.

The high baseline performance is considered to be caused by following three factors:

1. Using the categories that JUMAN adds to common nouns.

2. Using the head morpheme of the bunsetsu.

3. Using plural Japanese morphological analyzer, Juman and MeCab.

In order to confirm the effects of these factors, we conducted several experiments against CRL NE data and investigated how the F-measure changed without categories, head morphemes and/or MeCab. 3.7 shows the experimental results. We can confirm that categories and MeCab chiefly contribute to high baseline performance.

In previous work, Kazama and Torisawa [49] and Fukushima et al. [50] shows that the performance of NER can be improved by using information extracted from large-scale Web text. By using such information besides non-local information, the performance of our NER system is considered to be improved.

3.6 Summary of this Chapter

In this chapter, we presented an approach that uses non-local information for Japanese NER. We use a Support Vector Machine (SVM) based NER system as a baseline system, and introduced four types of non-local information to them: cache features, coreference rules, syntactic features and case frame features.

We evaluated our approach on CRL NE data and obtained a higher F-measure than existing approaches that do not use non-local information. We also conducted experiments on IREX NE data and an NE-annotated web corpus and confirmed that non-local information improves the performance of NER. As a consequence, the performance of NER was improved by using non-local information and our approach achieved a higher F-measure than existing approaches.

Chapter 4

Coreference Resolution Using Knowledge of Nominal Relations

4.1 Introduction

In text, expressions that refer to the same entity are repeatedly used. Coreference resolution, which recognizes such expressions, is an important technique for natural language processing. This chapter focuses on coreference resolution for Japanese text.

In Japanese, since pronouns are often omitted, most anaphors are represented as proper noun phrases or common noun phrases. To resolve coreference for such language, string matching technique is useful, because an anaphor and its antecedent often share strings [51]. Learning-based coreference approaches, which have been intensively studied in recent years [7, 52–54], use string matching as features for learning. However, in some cases, coreferential expressions share no string, and string matching technique cannot be applied.

Resolving such coreference relations requires knowledge that these two expressions share the same meaning. Thus, as described in Chapter 2, we first acquire knowledge of synonyms from large raw corpus and dictionary definition sentences, and then utilize the synonyms to coreference resolution.

Our target language Japanese also has a characteristic that it has no article. Articles can be a clue for anaphoricity determination, so this characteristic makes anaphoricity determination difficult. In this thesis, to boost the performance of coreference resolution, we integrate bridging reference resolution system that uses automatically constructed nominal case frames into coreference resolver. Roughly speaking, we consider modified NPs are not anaphoric. But if an NP have a bridging relation, it is considered as anaphoric.

The rest of this chapter is organized as follows. In Section 4.2, we present basic strategy

for coreference resolution and how to use the extracted synonyms and the result of bridging reference resolution for coreference resolution. We show the experimental results on news paper articles in Section 4.3, and compare our approaches with some related works in Section 4.4. Lastly, we provide summaries our approaches for coreference resolution in Section 4.5.

4.2 Strategy for Coreference Resolution

We use a coreference resolver based on string machines as a basic strategy, and propose a method to improve the coreference resolution using knowledge of synonyms and bridging reference resolution.

4.2.1 Basic Strategy for Coreference Resolution

The outline of our coreference resolver is as follows:

1. Parse input sentences by using a Japanese parser and recognize named entity.
2. Consider each subsequence of a noun phrase as a possible anaphor if it meets "*Condition 1*."
3. For each anaphor:
   (a) From the position of the anaphor to the beginning of document, consider each noun sequence as antecedent candidate.
   (b) If the anaphor and the antecedent candidate meet "*Condition 2*," judge as coreferential expressions and move to next anaphor.

"*Condition 1*" and "*Condition 2*" are varied between methods. "*Condition 1*" judges the anaphoricity of the subsequence, and "*Condition 2*" judges whether target anaphor and the antecedent candidate are coreferential or not. We use KNP [12] as a Japanese parser. To recognize named entity, we apply the NE recognizer described in Chapter 3.

Determination of Markables

The first step of coreference resolution is to identify the markables. Markables are noun phrases that related to coreference. We consider how to deal with compound nouns. Most previous

work on coreference resolution for Japanese texts focused only on the whole compound noun. However, such approaches cannot deal with following example:

(4.1) *Lifestyle-no chosa-wo jisshi-shita. Chosa naiyo-wa . . .*
lifestyle investigation conduct investigation content
(ϕ conducted an investigation. The content of the investigation was . . .)

In this example, the second "*chosa*" (investigation) that is contained in a compound noun "*chosa naiyo*" refers to the preceding "*chosa*." To deal with such a coreference relation, we consider every subsequence of a compound noun as a markable, that is, we consider "*chosa naiyo*," "*chosa*" and "*naiyo*" as a markable for *chosa naiyo*. Provided that the same shall not apply to to named entities, because substrings of named entities are scarcely considered to be a markable. Named entities are not divided and handled as a whole.

Baseline Methods

We consider 3 baseline methods. In all of these methods, "*Condition 2*" is true when the anaphor exactly matches the antecedent candidate. Only "*Condition 1*" (i.e. anaphoricity determination) varies among these 3 baselines.

In a primitive baseline (*baseline 1*), "*Condition 1*" is always true, that is, every noun sequence is considered an anaphor. For a bit more sophisticated baselines (*baseline 2* and *baseline 3*), we assume that a modified noun phrase is not anaphoric.

(4.2) a. *Uno shusho-wa Doitsu-ni totyaku-shita. Shusho-wa kuukou-de . . .*
Uno prime minister Germany arrived prime minister airport
(Prime minister Uno arrived in Germany. At the airport the minister . . .)

b. *Uno shusho-wa Doitsu-ni totyaku-shita. Asu Doitsu Shusho-tono. . .*
Uno prime minister Germany arrived Tomorrow German prime minister
(Prime minister Uno arrived in Germany. Tomorrow, with German prime minister . . .)

In example (4.2a), "*shusho*" (prime minister) in the first and second sentence refer to the same entity, but not in example (4.2b). This is because the second "*shusho*" in (4.2b) is modified by "*Doitsu*" (German), and this "*shusho*" is turned out to be a person other than "*Uno shusho*."

Table 4.1: *Condition 1* for Each Baseline.

	Condition 1
baseline 1	always true
baseline 2	true when the noun sequence is not modified by its preceding nouns in the same phrase
baseline 3	true when the noun sequence has no modifier

We consider that a partial noun sequence of a compound noun is modified by its preceding nouns in the compound noun. For example, for the compound noun "XY," "Y" is considered to be modified by "X," and thus "Y" is regarded as non-anaphoric (in this case, noun sequences "XY" and "X" are regarded as anaphoric).

In both *Baseline 2* and *baseline 3*, modified noun phrases are considered non-anaphoric. These two methods differ in the scope of the considered modifier. In *baseline 2*, "*Condition 1*" is true when the noun sequence is not modified by its preceding nouns in the same noun phrase. On the other hand, in *baseline 3*, "*Condition 1*" is true only when the noun sequence do not have any modifier including clausal modifier and adjective modifier. Table 4.1 show the "*Condition 1*" for each baseline.

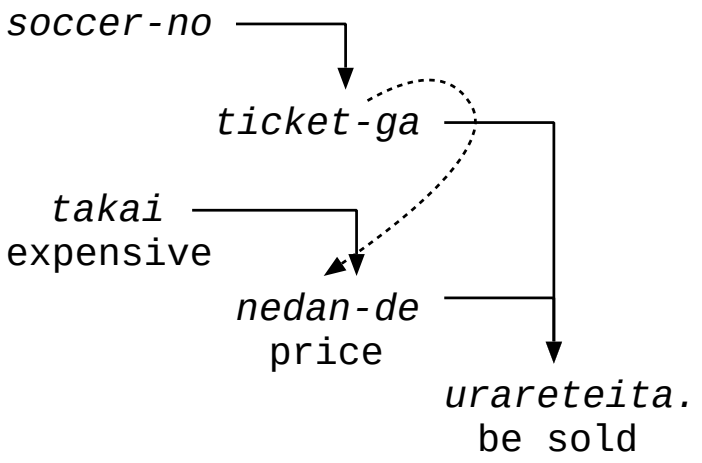
4.2.2 Utilization of Knowledge of Synonyms

The basic strategy for determining a coreference relation is based on precise string matching between an anaphor and its antecedent candidate. We also make use of knowledge of synonyms to resolve a coreference relation that cannot be recognized by string matching.

In this system, "*Condition 2*" is true not only when the anaphor exactly matches the antecedent candidate, but also when the anaphor is a synonym of the antecedent candidate.

4.2.3 Utilization of Bridging Reference Resolution

We then explain how to use the bridging reference resolution to coreference resolution. As mentioned, we do not consider a modified NP anaphoric in *baseline 2* and *baseline 3*. However, in some cases, a modified NP can be anaphoric. To deal with such cases, if two NPs share strings and have a bridging relation to the same entity, we consider the latter NP is anaphoric and has coreference relation to the former NP.



	case slot	examples	result
ticket	[theater,transport]	stage, game,$\cdots$	*soccer*
nedan	[things]	thing, ticket,$\cdots$	*ticket*

ticket a printed piece of paper which shows that you have paid to enter a theater or use a transport
nedan the amount of money for which things are sold or bought

Figure 4.1: Analysis Process of Bridging Reference Resolution.

Preliminary Bridging Relation Resolution System

In order to recognize bridging reference, we built a preliminary bridging reference resolution system based on the nominal case frames that were constructed in Section 2.4.

An input sentence is parsed using the Japanese parser, KNP [12]. Then, from the beginning of the sentence, each noun x is analyzed. When x has more than one case frame, the process of antecedent estimation (stated in the next paragraph) is performed for each case frame, and the case frame with the highest similarity score (described below) and assignments of antecedents to the case frame are selected as a final result.

For each case slot of the target case frame of x, its antecedent is estimated. A possible antecedent y in the target sentence and the previous two sentences is checked. This is done one by one, from the syntactically closer y. If the similarity of y to the case slot is equal to or greater than a threshold α (currently 0.95), it is assigned to the case slot. The similarity between y and a case slot is defined as the highest similarity between y and an example in the case slot.

(4.3) *Soccer-no ticket-ga takai nedan-de urareteita.*
soccer ticket expensive price sold
(The ticket was sold at expensive price.)

For instance, let us consider the sentence (4.3). Figure 4.1 shows the analysis process. *Soccer*, at the beginning of the sentence, is first analyzed. Since *soccer* has no case frame, *soccer* is considered to have no obligatory case. For the second noun *ticket*, *soccer*, which is a nominal modifier of *ticket*, is examined in advance. The similarity between *soccer* and the examples of the case slot [theater, transport] exceeds the threshold α, and *soccer* is assigned to [theater, transport]. Lastly, for *nedan* 'price,' its possible antecedents are *ticket* and *soccer*. *ticket*, which is the closest from *nedan*, is checked first. The similarity between *ticket* and the examples of the case slot [things] exceeds the threshold α, and *ticket* is judged as the antecedent of *nedan*.

Coreference Resolution using Bridging Reference Resolution

Then we illustrate how to use the bridging reference resolution in coreference resolution using following example:

(4.4) *Murayama shusho-wa nento-no kisha kaiken-de shokan-wo*
Murayama prime minister beginning of year press conference impressions

happyo-shita. Nento shokan-no yoshi-wa ika-no tori.
express beginning of year impressions point as follows

(Prime Minister Murayama expressed his impressions at the press conference of the beginning of the year. The point of the impressions is as follows.)

In example (4.4), the second "*shokan*" (impression) is modified by "*nento*" (beginning of year) and is not considered anaphoric in *baseline 2* or *baseline 3* method. However, "*shokan*" (impression) has a case frame named "AGENT" as shown in Table 4.2, and its bridging relation to "*shusho*" (prime minister) is recognized (i.e. the system recognize that the impression is the impression of the prime minister). Accordingly, the second "*shokan*" is considered anaphoric and the coreference relation between the first and the second "*shokan*" is recognized.

In the methods using bridging reference resolution, "*Condition 1*" is also true when the anaphor has a bridging relation, and then "*Condition 2*" is true only when the anaphor and its antecedent candidate have the same referent of bridging.

Table 4.2: Nominal Case Frame of "*shokan*" (impression).

case frame	examples	:	frequency
AGENT	"*watashi*" (I)	:	24
	"*chiji*" (governor)	:	16
	"*sori*" (prime minister)	:	3
	"*hissha*" (writer)	:	2
	...	:	...

Table 4.3: Nominal Case Frame of "*kekka*" (result).

case frame	examples	:	rate
"*koto*	"*chosa*" (investigation)	:	7648
(something)	"*senkyo*" (election)	:	1346
	"*enquête*" (questionnaire)	:	734
	"*jikken*" (experiment)	:	442
	...	:	...

"*koto*" = "*Aru koto-ga moto-ni natte okotta kotogara.*"
(a consequence, issue, or outcome of something)

As another example, although the second "*kekka*" (result) in example (4.5) is modified by "enquêtet" and is not considered anaphoric in *baseline 2* or *baseline 3* method, bridging reference resolver recognizes the two "*kekka*" refer to the same entity "enquête" by using the nominal case frame of "kekka" (result) shown in Table 4.3. Therefore, by considering bridging reference resolution, the system can recognize the coreference relation between the first and the second "*result*."

(4.5) 2006 FIFA *world cup-no yushokoku yosou enquête-wo okonatta.*
2006 FIFA world cup winner expectation questionnaire conducted

Kekka-wa Brazil-ga top-datta. Kuwasii enquête kekka-wa HP-de.
result Brazil top detail questionnaire result web page

(The expectation questionnaire about 2006 FIFA world cup win ner was conducted. The top of the questionnaire result was Brazil. The detail of the result appeared in web page.)

Table 4.4: Experimental Results of Coreference Resolution.

	Kyoto Corpus			Web corpus		
Condition	Precision	Recall	F-measure	Precision	Recall	F-measure
baseline1:	57.4 (2251/3925)	78.4 (2251/2870)	66.3	56.2 (585/1041)	82.1 (585/717)	66.6
baseline2:	72.2 (2191/3033)	76.3 (2191/2870)	74.2	68.6 (583/850)	81.3 (583/717)	74.4
+ bridging	72.0 (2209/3068)	77.0 (2209/2870)	74.4	68.1 (586/861)	81.7 (586/717)	74.3
+ synonym	72.6 (2239/3086)	78.0 (2239/2870)	75.2	68.7 (586/853)	81.7 (586/717)	74.6
+ bridging + synonym	72.3 (2257/3121)	**78.6** (2257/2870)	**75.3**	68.2 (589/864)	**82.1** (589/717)	74.5
baseline3:	78.1 (1946/2492)	67.8 (1946/2870)	72.6	82.5 (515/624)	71.8 (515/717)	76.8
+ bridging	77.6 (2004/2583)	69.8 (2004/2870)	73.5	80.5 (532/661)	74.2 (532/717)	77.2
+ synonym	**78.4** (1995/2544)	69.5 (1995/2870)	73.7	**82.6** (518/627)	72.2 (518/717)	77.1
+ bridging + synonym	77.9 (2052/2634)	71.5 (2052/2870)	74.6	80.6 (535/664)	74.6 (535/717)	**77.5**

4.3 Experiments

4.3.1 Experimental Setting

We conducted experiments on the Kyoto Corpus Version 4.0 [55]. In the corpus, coreference relations are manually annotated on the articles of *Mainichi* newspaper. We used 322 articles, which comprise 2098 sentences. These sentences have 2872 coreference tags that match our coreference criteria. In addition, so as to confirm the effectiveness of our method on web text, we created an anaphoric relation-tagged corpus consisting of 186 web documents, and also conducted experiments on the corpus. In this corpus, there are 979 sentences and 717 coreference tags. Each document consisting of this corpus includes no more than 10 sentences.

We used 3 baseline methods, *baseline 1*, *baseline 2* and *baseline 3*. In addition, for *baseline 2* and *baseline 3*, we also conducted experiments with knowledge of synonyms and/or bridging reference resolution. Thus, all in all we conducted experiments in 9 different conditions.

Table 4.5: Coreference Relations Newly Generated by Using Knowledge of Synonyms.

True positive

- *Nihon-ni umareta zai-nichi kankoku chosenjin-wo koumuin-toshite haijo-suru . . .*
  Japan born in Japan Koreans civil servant exclude
  (To exclude Koreans in Japan who was born in Japan from civil servant . . .)
- *Ippou Chosen-minshushugi-jinmin-kyowakoku-wa . . . Kita-chosen-no . . .*
  on the other hand the Democratic People's Republic of Korea North Korea
  (On the other hand, the Democratic People's Republic of Korea . . . North Korea's . . .)

False positive

- *Eisei-tuushin-wa . . . . genzai 5 channel-no BS-ga*
  satellite communications at present
  (Satellite communications are . . . . BS, which has 5 channels at present, . . .)

4.3.2 Experiments and Discussion

Table 4.4 shows the results of coreference resolution.*Baseline 1* achieve high recall but lowest precision and F-measure. We can say that considering modified NPs as non-anaphoric improves F-measure. We can also say that the condition used in *baseline 2*, "*Condition 1*" is true when the noun sequence is not modified by its preceding nouns in the same phrase, achieve best performance. Furthermore, using knowledge of synonyms and the result of bridging reference resolution improves F-measure and the usefulness of them is confirmed, but the effect is limited.

Table 4.5 shows examples of coreference relations newly generated by using knowledge of synonyms, which is the difference between baseline2 and baseline2 with knowledge of synonyms on both Kyoto Corpus and web corpus. There were 51 true positives and only 5 false positives. 21 true positives out of 51 were generated owing to the synonyms extracted from the dictionaries; thus we can confirm that although there were not so many synonyms extracted from the dictionaries, they contribute the performance of coreference resolution.

Table 4.6 shows examples of coreference relations newly generated by using bridging reference resolution, which is the difference between baseline3 and baseline3 with bridging reference resolution on both Kyoto Corpus and web corpus. There were 75 true positives and 53 false positives. Most false positives was caused by errors in bridging reference resolution.

Table 4.6: Coreference Relations Newly Generated by Considering Bridging Reference.

True positive

- ... *jisshi-sareta* *zen-giin* *anketo-de* ... *kekka-wa* ... *anketo-kekka-wo* ...
  asked all representative questionnaire result result of questionnaire
  (Representative questionnaire that was asked ... The result was ... The result of the questionnaire was ...)

False positive

- *Dokuji* *kouho* *youritsu-e muke* ... *"renraku kyogikai"-wo* ... *yuryoku kouho to* ...
  unique candidate support counsil strong candidate
  (To support unique candidate ... the counsil ... with strong candidate ...)

Table 4.7: Recall for Each Coreference Type.

relations between anaphor & antecedent	recall
1. anaphor's string is contained in antecedent's string	86.9 (192/221)
2. anaphor and its antecedent have a synonymous relation	71.4 (5/7)
3. other coreference types	0.0 (0/22)
sum	76.1 (197/250)

To investigate recall for several coreference types, we randomly selected 250 coreference tags from the Kyoto Corpus and evaluated the result of coreference resolution using *baseline 2* method with knowledge of synonyms and bridging reference resolution. Table 4.7 shows the recall for each coreference type.

The coreference relations that can be recognized by string matching were well recognized. The relations that need knowledge of synonyms to recognize were also well recognize and we can say that the coverage of the automatically acquired synonyms is not too small for resolving coreference relations between synonymous expressions. The other types of coreference relations, such as relations between hypernym and hyponym, cannot recognize fundamentally by our proposed method. To resolve such relations is our future work.

In order to investigate the cause of erroneous system outputs, we classify erroneous system

Table 4.8: Error Analysis of Erroneous System Outputs.

error type	num
The anaphor and antecedent candidate refer to another entities	52
The possible anaphor is a general noun and not anaphoric	32
The antecedent candidate is a general noun and not anaphoric	7
others	9
sum	100

outputs into 4 categories. Table 4.8 shows the classified error types of randomly selected 100 erroneous system outputs of *baseline 2* method with knowledge of synonyms and bridging reference resolution. Major erroneous system outputs were caused by two reasons:

1. *Baseline 2* method does not consider clausal or adjective modifiers.

2. Our system does not consider the generic usage of nouns.

In example (4.6), though the second "*jishin*" (earthquake) does not have coreference relation to "*Sanriku Harukaoki Jishin*," our system judges the two "*jishin*" refer to the same entity because our system does not consider the modifiers "*yoshin-to mirareru*" (thought to be an aftershock).

(4.6) *Sanriku Harukaoki Jishin-no yoshin-to mirareru jishin-ga hassei-shita.*
Far-off Sanriku Earthquake aftershock thought earthquake occurred
(An earthquake thought to be an aftershock of Far-off Sanriku Earthquake occurred.)

In example (4.7), although the second "wine" is used in generic usage, our system considers the second "wine" have coreference relation to "French wine" because our system does not consider generic usage of nouns.

(4.7) *Kare-wa* France-*no* wine-*ga suki-de kare-no ie-niwa* wine cellar-*ga aru.*
he French wine like his house wine cellar have
(He likes French wine and has wine cellar in his house.)

Table 4.9: Comparison with Previous Work.

	precision	recall	F-measure
Murata and Nagao [6]	78.7 (89/113)	77.3 (89/115)	78.1
Iida et al. [7]	76.7 (582/759)	65.9 (582/883)	70.9
Proposed:			
Kyoto Corpus	72.3 (2257/3121)	78.6 (2257/2870)	75.3
Web corpus	80.6 (535/664)	74.6 (535/717)	77.5

Table 4.9 shows the comparison with previous work and our proposed method. The details of the previous work will be given in next section. Since they used different data set and coreference criteria for experiments, these scores are not comparable as-is. However, taking into consideration Murata and Nagao uses small and supposedly easy corpus, we can say that our proposed method achieved enough performance. Though these scores are not comparable as-is, rule-based methods outperformed learning-based methods in Japanese. This may be because recognizing most of coreference relations does not need complicated rules.

4.4 Related Work

Most previous works on coreference resolution can be divided into two types. One is rule-based approach. Zhou and Su [56] proposed rule-base approach for English coreference resolution. They divided coreference relations into 7 types and created rules for each type. As for Japanese, Murata and Nagao proposed a rule-based coreference resolution method for determining the referents of noun phrases in Japanese sentences by using referential properties, modifiers and possessors [6]. As a result of experiments, they obtained a precision of 78.7% and a recall of 77.3%. Their method performed relatively well. This may be because their experiments were constructed on small and supposedly easy corpus. Half of their corpus was occupied by fairy tale that was supposed to be easy to analyze.

The other approach is learning-based approach. There are plenty of learning-based coreference resolution systems. Sonn et al. [52] built a decision tree classifier to label pairs of mentions as coreferent or not. Using their classifier, they would build up coreference chains, where each

mention was linked up with the most recent previous mention that the classifier labeled as coreferent, if such a mention existed. Transitive closure in this model was done implicitly. Much work that followed improved upon this strategy, by improving the features, changing mention links to be to the most likely antecedent rather than the most recent positively labeled antecedent (Ng and Cardie [57]), and the type of classifier (Denis and Baldridge [58]). As for Japanese, Iida et al. proposed a machine learning approach for coreference resolution for Japanese [7]. Their process is similar to the model proposed by Ng and Cardie [59]. As a result of experiments on Japanese newspaper articles, they obtained a precision of 76.7% and a recall of 65.9%.

These studies adopted the mention-pair model, which recasts coreference resolution to a binary classification problem of determining whether or not two mentions in a document are co-referring. Although having achieved reasonable success, the mention-pair model has a limitation that information beyond mention pairs is ignored for training and testing. As an individual mention usually lacks adequate descriptive information of the referred entity, it is often difficult to judge whether or not two mentions are talking about the same entity simply from the pair alone.

To deal with this problem, some work adopted the entity-mention model. Luo et al. [53] proposed a system that performs coreference resolution by doing search in a large space of entities. They trained a classifier that can determine the likelihood that an active mention should belong to an entity. The entity-level features were calculated with an “ Any-X ” strategy: an entity mention pair would be assigned a feature X, if any mention in the entity has the feature X with the active mention. Culotta et al. [60] presented a system which used an online learning approach to train a classifier to judge whether two entities are coreferential or not. The features describing the relationships between two entities were obtained based on the information of every possible pair of mentions from the two entities. Different from Luo et al. [53], the entity-level features were computed using a “ Most-X ”strategy, that is, two given entities would have a feature X, if most of the mention pairs from the two entities have the feature X. Yang et al. [61] presented an expressive entity-mention model that performed coreference resolution at an entity level and adopted the Inductive Logic Programming (ILP) algorithm, which provides a relational way to organize different knowledge of entities and mentions.

While successful, supervised learning approaches require labeled training data, consisting of mention pairs and the correct decisions for them. This limits their applicability. To overcome this limits, recent years have seen unsupervised approaches. Unsupervised approaches are attractive due to the availability of large quantities of unlabeled text. Haghighi and Klein [62]

proposed a unsupervised coreference resolution system using a nonparametric Bayesian model based on hierarchical Dirichlet processes. At the heart of their system is a mixture model with a few linguistically motivated features such as head words, entity properties and salience. Poon and Domingos [63] presented unsupervised approach that perform joint inference across mentions, in contrast to the pairwise classification typically used in supervised methods, and by using Markov logic as a representation language, which enables us to easily express relations like apposition and predicate nominals.

4.5 Summary of this Chapter

We have described a knowledge-rich approach to Japanese coreference resolution. We first proposed a method for acquiring knowledge of synonyms from large raw corpus and definition sentences of dictionaries for humans. Second, we proposed a method for improving coreference resolution by using the automatically acquired synonyms and the result of bridging reference resolution. Using the acquired synonyms and the result of bridging reference resolution boosted the performance of coreference resolution and the effectiveness of our integrated method is confirmed.

Chapter 5

Probabilistic Model for Zero Anaphora and Bridging Reference Resolution

5.1 Introduction

Anaphora resolution is one of the most important techniques in discourse analysis. In English, definite noun phrases such as *the company* and overt pronouns such as *he* are anaphors that refer to preceding entities (antecedents). On the other hand, in Japanese, anaphors are often omitted and these omissions are called *zero pronouns*, and such anaphora is called *zero anaphora*. In this Chapter, we first focus on zero anaphora resolution of Japanese web corpus; and propose a probabilistic model for zero anaphora resolution that utilizes case frames.

Next, we try to extend this probabilistic model to bridging reference resolution. As is mentioned before, bridging reference (also called indirect anaphora, or functional anaphora) is a phenomenon that an anaphoric expression refers to a discourse entity that wasn't mentioned before but is somehow related to a discourse entity that already has. In order to recognize such reference, we constructed nominal case frames in Section 2.4, which describe such knowledge as which nouns can indirectly refer to a discourse entity and what discourse entity can be referred by the noun. We call the indispensable entities of nouns *obligatory cases*, and by considering the relation between a noun and its indispensable entity is parallel to that between a verb and its arguments or obligatory cases, we extend the probabilistic model for zero anaphora resolution to bridging referred resolution.

Zero anaphora resolution can be divided into two phases. The first phase is zero pronoun detection and the second phase is zero pronoun resolution. Zero pronoun resolution is similar to coreference resolution and pronoun resolution, which have been studied for many years (e.g. Soon et al. [52]; Mitkov [64]; Ng [54]). Isozaki and Hirao [8] and Iida et al. [10] focused

on zero pronoun resolution assuming perfect pre-detection of zero pronouns. However, we consider that zero pronoun detection and resolution have a tight relation and should not be handled independently. Our proposed model aims not only to resolve zero pronouns but to detect zero pronouns.

Zero pronouns are not expressed in a text and have to be detected prior to identifying their antecedents. Seki et al. [65] proposed a probabilistic model for zero pronoun detection and resolution that uses hand-crafted case frames. In order to alleviate the sparseness of hand-crafted case frames, Kawahara and Kurohashi [9] introduced wide-coverage case frames to zero pronoun detection that are automatically constructed from a large corpus. They use the case frames as selectional restriction for zero pronoun resolution, but do not utilize the frequency of each example of case slots. However, since the frequency is shown to be a good clue for syntactic and case structure analysis [11], we consider the frequency also can benefit zero pronoun detection. Therefore we propose a probabilistic model for zero anaphora resolution that utilizes case frames. This model directly considers the frequency and estimates case assignments for overt case components and antecedents of zero pronouns simultaneously.

In addition, our model directly links each zero pronoun to an entity, while most existing models link it to a certain mention of an entity. In our model, mentions and zero pronouns are treated similarly and all of them are linked to corresponding entities. In this point, our model is similar to the coreference model proposed by Luo [66] and that proposed by Yang et al. [61]. Due to this characteristic, our model can utilize information beyond a mention and easily consider salience (the importance of an entity).

5.2 Anaphora Resolution Model

In this section, we describe a probabilistic model for Japanese zero anaphora resolution that utilizes case frames.

5.2.1 Overview

The outline of our model is as follows:

1. Parse an input text using the Japanese parser KNP [12] and recognize NEs using the NE recognizer described in Chapter 3.
2. Conduct coreference resolution by using coreference resolution described in Chapter 4 and link each mention to an entity or create new entity.

3. For each sentence, from the end of the sentence, analyze each predicate by the following steps:

   (a) Select a case frame temporarily.

   (b) Consider all possible correspondence between each input case component and a case slot of the selected case frame.

   (c) Regard case slots that have no correspondence as zero pronoun candidates.

   (d) Consider all possible correspondence between each zero pronoun candidate and an existing entity of high salience.

   (e) For each possible case frame, estimate each correspondence probabilistically, and select the most likely case frame and correspondence.

In this thesis, we concentrate upon three case slots for zero anaphora resolution: "*ga* (nominative)," "*wo* (accusative)" and "*ni* (dative)," which cover about 90% of zero anaphora. Morphological analysis, NE recognition, syntactic analysis and coreference resolution are conducted as pre-processes for zero anaphora resolution. Therefore, the model has already recognized existing entities before zero anaphora resolution.

For example, let us consider the following text:

(5.1) Toyota-*wa* 1997-*nen* hybrid car Prius-*wo* *hatsubai*(launch). 2000-*nen-karaha* *kaigai* (overseas)-*demo* *hanbai*(sell)-shiteiru.
(Toyota launched the hybrid car Prius in 1997. ϕ_1 started selling ϕ_2 overseas in 2000.)

Figure 5.1 shows the analysis process for this text. There are four mentions, Toyota, 1997-*nen* (year 1997), hybrid car, and Prius in the first sentence, and the two mentions, *hybrid car* and *Prius*, appear in apposition. Thus, after the pre-processes, three entities, {Toyota}, {1997-*nen*} and {hybrid-car, Prius}, are created.

Then, case structure analysis for the predicate *hatsubai* (launch) is conducted. First, one of the case frames of *hatsubai* (launch) is temporarily selected and each input case component is assigned to an appropriate case slot. For instance, case component *Toyota* is assigned to *ga* case slot and *Prius* is assigned to *wo* case slot. Note that since there are some non case-making postpositions in Japanese, such as "*wa*" and "*mo*," several correspondences can be considered. In this case, though there is a mention *hybrid-car* that is not a case component of *hatsubai* (launch) by itself, it refers to the same entity as *Prius* refers, a relatively low salience is given

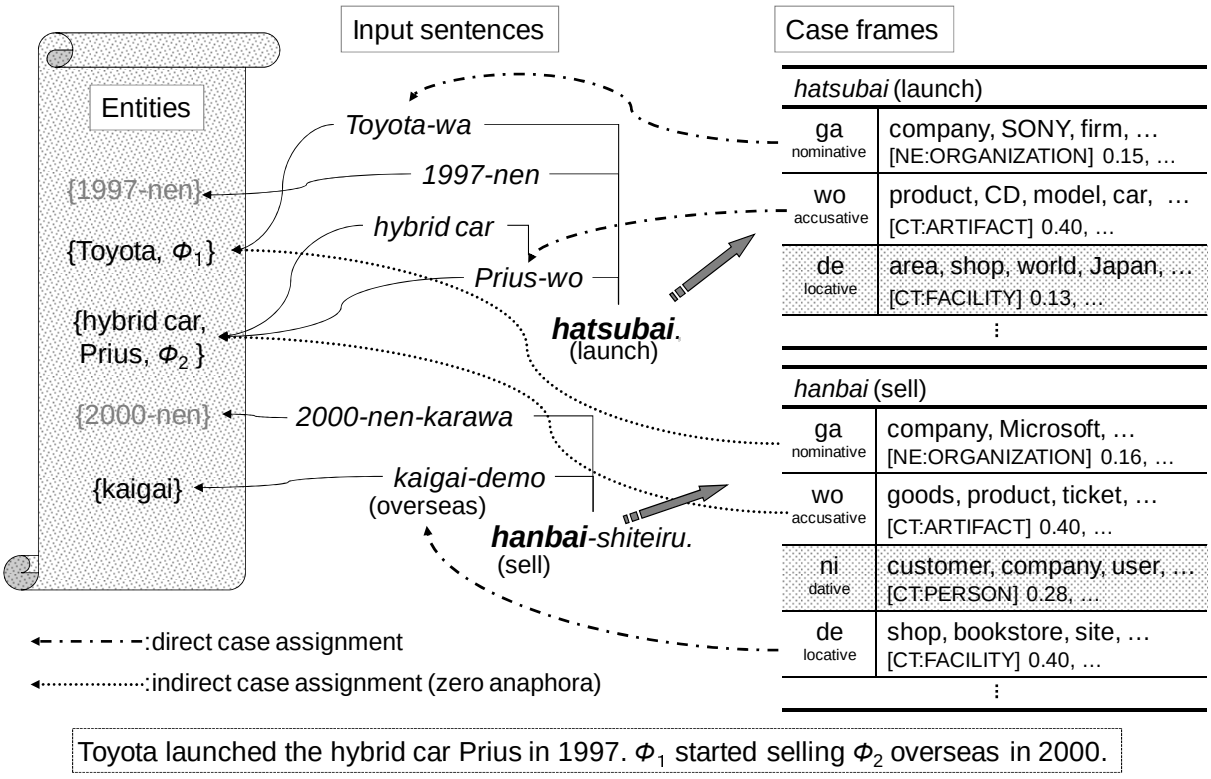


Figure 5.1: An Example of Case Assignment CA_k.

to the entity {1997-*nen*} by the rules described in Section 5.2.4. Thus, there is no entity of high salience that is not linked to *hatsubai* (launch), and no further analysis is conducted.

Now, let us consider the second sentence. Two mentions *kaigai* (overseas) and *2000-nen* (year 2000) appear and new entities {*kaigai*} and {2000-*nen*} are created. Then, case structure analysis for the predicate *hanbai* (sell) is conducted. There is only one overt case component *kaigai* (overseas), and it is assigned to a case slot of the selected case frame of *hanbai* (sell). For instance, the case frame *hanbai*(1) in Table 5.1 is selected and *kaigai* (overseas) is assigned to *de* (locative) case slot. In this case, the remaining case slots *ga*, *wo* and *ni* are considered as zero pronouns, and all possible correspondences between zero pronouns and remaining entities of high salience are considered. As a result of probabilistic estimation, the entity {Toyota} is assigned to *ga* case, the entity {hybrid-car, Prius} is assigned to *wo* case and no entity is assigned to *ni* case.

Now, we show how to estimate the correspondence probabilistically in the next section.

Table 5.1: Examples of Constructed Case Frames (identical to Figure 2.4).

	case slot	examples	generalized examples with rate
tsumu (1) (load)	*ga* (nominative)	he, driver, friend, $\cdots$	[CT:PERSON]:0.45, [NE:PERSON]:0.08, $\cdots$
	wo (accusative)	baggage, luggage, hay, $\cdots$	[CT:ARTIFACT]:0.31, $\cdots$
	ni (dative)	car, truck, vessel, seat, $\cdots$	[CT:VEHICLE]:0.32, $\cdots$
tsumu (2) (accumulate)	*ga* (nominative)	player, children, party, $\cdots$	[CT:PERSON]:0.40, [NE:PERSON]:0.12, $\cdots$
	wo (accusative)	experience, knowledge, $\cdots$	[CT:ABSTRACT]:0.47, $\cdots$
$\vdots$	$\vdots$		$\vdots$
hanbai (1) (sell)	*ga* (nominative)	company, Microsoft, $\cdots$	[NE:ORG.]:0.16, [CT:ORG.]:0.13, $\cdots$
	wo (accusative)	goods, product, ticket, $\cdots$	[CT:ARTIFACT]:0.40, [CT:FOOD]:0.07, $\cdots$
	ni (dative)	customer, company, $\cdots$	[CT:PERSON]:0.28, $\cdots$
	de (locative)	shop, bookstore, site $\cdots$	[CT:FACILITY]:0.40, [CT:LOCATION]:0.39, $\cdots$
$\vdots$	$\vdots$		$\vdots$

5.2.2 Probabilistic Model for Zero Anaphora Resolution

The proposed model gives a probability to each possible case frame CF and case assignment CA when target predicate v, input case components ICC and existing entities ENT are given. It also outputs the case frame and case assignment that have the highest probability. That is to say, our model selects the case frame CF_{best} and the case assignment CA_{best} that maximize the probability $P(CF, CA|v, ICC, ENT)$:

$$(CF_{best}, CA_{best}) = \underset{CF,CA}{\operatorname{argmax}} P(CF, CA|v, ICC, ENT) \quad \text{(i)}$$

Though case assignment CA usually represents correspondences between input case components and case slots, in our model it also represents correspondences between antecedents of zero pronouns and case slots. Hereafter, we call the former direct case assignment (DCA) and the latter indirect case assignment (ICA). Then, we transform $P(CF_l, CA_k|v, ICC, ENT)$ as follows:

$$\begin{aligned}
&P(CF_l, CA_k|v, ICC, ENT) \\
=&P(CF_l|v, ICC, ENT) \times P(DCA_k|v, ICC, ENT, CF_l) \\
&\times P(ICA_k|v, ICC, ENT, CF_l, DCA_k) \\
\approx&P(CF_l|v, ICC) \times P(DCA_k|ICC, CF_l) \times P(ICA_k|ENT, CF_l, DCA_k) && \text{(ii)} \\
=&P(CF_l|v) \times P(DCA_k, ICC|CF_l)/P(ICC|v) \times P(ICA_k|ENT, CF_l, DCA_k) && \text{(iii)}
\end{aligned}$$

$$\Big(\because P(CF_l|v, ICC) = \frac{P(CF_l, ICC|v)}{P(ICC|v)} = \frac{P(ICC|CF_l, v) \cdot P(CF_l|v)}{P(ICC|v)} = \frac{P(ICC|CF_l) \cdot P(CF_l|v)}{P(ICC|v)},$$

($\because CF_l$ contains the information about v.)

$$P(DCA_k|ICC, CF_l) = \frac{P(DCA_k, ICC|CF_l)}{P(ICC|CF_l)} \Big)$$

Equation (ii) is derived because we assume that the case frame CF_l and direct case assignment DCA_k are independent of existing entities ENT, and indirect case assignment ICA_k is independent of input case components ICC.

Because $P(ICC|v)$ is constant, we can say that our model selects the case frame CF_{best} and the direct case assignment DCA_{best} and indirect case assignment ICA_{best} that maximize the probability $P(CF, DCA, ICA|v, ICC, ENT)$:

$$(CF_{best}, DCA_{best}, ICA_{best}) = \underset{CF,DCA,ICA}{\operatorname{argmax}} \Big(P(CF|v) \times P(DCA, ICC|CF) \times P(ICA|ENT, CF, DCA) \Big) \quad \text{(iv)}$$

The probability $P(CF_l|v)$, called *generative probability of a case frame*, is estimated from case structure analysis of a large raw corpus. The following subsections illustrate how to calculate $P(DCA_k, ICC|CF_l)$ and $P(ICA_k|ENT, CF_l, DCA_k)$.

Generative Probability of Direct Case Assignment

For estimation of *generative probability of direct case assignment* $P(DCA_k, ICC|CF_l)$, we follow Kawahara and Kurohashi's [11] method. They decompose $P(DCA_k, ICC|CF_l)$ into the following product depending on whether a case slot s_j is filled with an input case component or vacant:

$$\begin{aligned} P(DCA_k, ICC|CF_l) = & \prod_{s_j:A(s_j)=1} P(A(s_j) = 1, n_j, c_j|CF_l, s_j) \times \prod_{s_j:A(s_j)=0} P(A(s_j) = 0|CF_l, s_j) \\ = & \prod_{s_j:A(s_j)=1} \Big\{ P(A(s_j) = 1|CF_l, s_j) \times P(n_j, c_j|CF_l, s_j, A(s_j) = 1) \Big\} \\ & \times \prod_{s_j:A(s_j)=0} P(A(s_j) = 0|CF_l, s_j) \end{aligned} \quad \text{(v)}$$

where the function $A(s_j)$ returns 1 if a case slot s_j is filled with an input case component; otherwise 0, n_j denotes the content part of the case component, and c_j denotes the surface case of the case component.

The probabilities $P(A(s_j) = 1|CF_l, s_j)$ and $P(A(s_j) = 0|CF_l, s_j)$ are called *generative probability of a case slot*, and estimated from case structure analysis of a large raw corpus as well as *generative probability of a case frame*. The probability $P(n_j, c_j|CF_l, s_j, A(s_j) = 1)$ is called *generative probability of a case component* and estimated as follows:

$$\begin{aligned} &P(n_j, c_j|CF_l, s_j, A(s_j) = 1) \\ \approx &P(n_j|CF_l, s_j, A(s_j) = 1) \times P(c_j|s_j, A(s_j) = 1) \end{aligned} \tag{vi}$$

$P(n_j|CF_l, s_j, A(s_j) = 1)$ means the generative probability of a content part n_j from a case slot s_j in a case frame CF_l, and estimated by using the frequency of a case slot example in the automatically constructed case frames. $P(c_j|s_j, A(s_j) = 1)$ is approximated by $P(c_j|case\_type\_of(s_j), A(s_j) = 1)$ and estimated from the web corpus in which the relationship between a surface case marker and a case slot is annotated by hand.

Probability of Indirect Case Assignment

To estimate *probability of indirect case assignment* $P(ICA_k|ENT, CF_l, DCA_k)$ we also decompose it into the following product depending on whether a case slot s_j is filled with an entity ent_j or vacant:

$$\begin{aligned} P(ICA_k|ENT, CF_l, DCA_k) = &\prod_{s_j:A'(s_j)=1} P(A'(s_j) = 1, ent_j|ENT, CF_l, s_j) \\ &\times \prod_{s_j:A'(s_j)=0} P(A'(s_j) = 0|ENT, CF_l, s_j) \end{aligned} \tag{vii}$$

where the function $A'(s_j)$ returns 1 if a case slot s_j is filled with an entity ent_j; otherwise 0. Note that we only consider case slots *ga*, *wo* and *ni* that is not filled with an input case component. We approximate $P(A'(s_j) = 1, ent_j|ENT, CF_l, s_j)$ and $P(A'(s_j) = 0|ENT, CF_l, s_j)$ as follows:

$$\begin{aligned} &P(A'(s_j) = 1, ent_j|ENT, CF_l, s_j) \\ &\quad \approx P(A'(s_j) = 1, ent_j|ent_j, CF_l, s_j) = P(A'(s_j) = 1|ent_j, CF_l, s_j) \end{aligned} \tag{viii}$$

$$\begin{aligned} &P(A'(s_j) = 0|ENT, CF_l, s_j) \\ &\quad \approx P(A'(s_j) = 0|CF_l, s_j) \approx P(A(s_j) = 0|CF_l, s_j) \end{aligned} \tag{ix}$$

Table 5.2: Location Classes of Antecedents.

intra-sentence:(overt case component (case))
L_{1c}: overt case components of parent predicate of V_z
L_{2c}: overt case components of child predicate of V_z
L_{3c}: overt case components of parent predicate of V_z" (parallel)
L_{4c}: overt case components of child predicate of V_z (parallel)
L_{5c}: overt case components of parent predicate of parent noun phrase of V_z
L_{6c}: overt case components of parent predicate of parent predicate of V_z
intra-sentence:(omitted case component (zero))
L_{1z}: omitted case components of parent predicate of V_z
L_{2z}: omitted case components of child predicate of V_z
L_{3z}: omitted case components of parent predicate of V_z" (parallel)
L_{4z}: omitted case components of child predicate of V_z (parallel)
L_{5z}: omitted case components of parent predicate of parent noun phrase of V_z
L_{6z}: omitted case components of parent predicate of parent predicate of V_z
intra-sentence:(other noun phrases)
L_7 : other noun phrases with topic marker
L_8 : other noun phrases preceding V_z
L_9 : other noun phrases following V_z
inter-sentence:
L_{10}: noun phrases with topic marker in 1 sentence before
L_{11}: the last noun phrases in 1 sentence before
L_{12}: other noun phrases in 1 sentence before with topic marker
L_{13}: noun phrases in 2 sentences before
L_{14}: noun phrases in 3 sentences before
L_{15}: noun phrases in more than 3 sentences before

Equation (viii) is derived because we assume $P(A'(s_j) = 1|CF_l, s_j)$ is independent of existing entities that are not assigned to s_j. Equation (ix) is derived because we assume $P(A'(s_j)=0)$ is independent of ENT and the generative probability of a case slot are the same for overt case and omitted case. Note that $P(A'(s_j) = 0|CF_l, s_j)$ is the probability that a case slot has no correspondence after zero anaphora resolution and $P(A'(s_j)=0|CF_l, s_j)$ is the probability that a case slot has no correspondence after case structure analysis. Although they have similar forms, there is no guarantee that their values are similar; and we can say this is a bit rough approximation.

Let us consider the probability $P(A'(s_j)= 1|ent_j, CF_l, s_j)$. We decompose ent_j into content part n_{j_m}, surface case c_{j_n} and location class l_{j_n}. Location classes denote the locational relations between zero pronouns and their antecedents. We defined 21 location classes as described in Table 5.2. In Table 5.2, V_z means a predicate that has a zero pronoun. Note that we also con-

sider the locations of zero pronouns. Now we roughly approximate $P(A'(s_j)=1|ent_j, CF_l, s_j)$ as follows:

$$P(A'(s_j) = 1|ent_j, CF_l, s_j)$$

$$=P(A'(s_j) = 1|n_{j_m}, c_{j_n}, l_{j_n}, CF_l, s_j) = \frac{P(A'(s_j) = 1, n_{j_m}, c_{j_n}, l_{j_n}|CF_l, s_j,)}{P(n_{j_m}, c_{j_n}, l_{j_n}|CF_l, s_j)}$$

$$=\frac{P(n_{j_m}, c_{j_n}, l_{j_n}|CF_l, s_j, A'(s_j) = 1) \times P(A'(s_j) = 1|CF_l, s_j)}{P(n_{j_m}, c_{j_n}, l_{j_n}|CF_l, s_j)}$$

$$\approx\frac{P(n_{j_m}|CF_l, s_j, A'(s_j) = 1)}{P(n_{j_m}|CF_l, s_j)} \times \frac{P(c_{j_n}|CF_l, s_j, A'(s_j) = 1)}{P(c_{j_n}|CF_l, s_j)} \quad \text{(x)}$$
$$\times \frac{P(l_{j_n}|CF_l, s_j, A'(s_j) = 1)}{P(l_{j_n}|CF_l, s_j)} \times P(A'(s_j) = 1|CF_l, s_j)$$

$$\approx\frac{P(n_{j_m}|CF_l, s_j, A'(s_j) = 1)}{P(n_{j_m})} \times \frac{P(c_{j_n}|case\_type\_of(s_j), A'(s_j) = 1)}{P(c_{j_n})} \quad \text{(xi)}$$
$$\times P(A'(s_j) = 1|l_{j_n}, case\_type\_of(s_j))$$

$$\Big(\because \frac{P(l_{j_n}|CF_l, s_j, A'(s_j) = 1)}{P(l_{j_n}|CF_l, s_j)} \times P(A'(s_j) = 1|CF_l, s_j)$$

$$= \frac{P(A'(s_j) = 1, l_{j_n}|CF_l, s_j)}{P(l_{j_n}|CF_l, s_j)} = P(A'(s_j) = 1|CF_l, l_{j_n}, s_j) \Big)$$

Note that because ent_j is often mentioned more than one time, there are several combinations of content part n_{j_m}, surface case c_{j_n} and location class l_{j_n} candidates. We select the pair of m and n with the highest probability.

Equation (x) is derived because we assume n_{j_m}, c_{j_n} and l_{j_n} are independent of each other. Equation (xi) is derived because we approximate $P(A'(s_j) = 1|CF_l, l_{j_n}, s_j)$ as $P(A'(s_j) = 1|l_{j_n}, case\_type\_of(s_j))$, and assume $P(n_{j_m})$ and $P(c_{j_n})$ are independent of CF_l and s_j. These approximation is also a bit rough, that is, $P(n_{j_m})$ and $P(c_{j_n})$ tend to be somewhat smaller than $P(n_{j_m}|CF_l, s_j)$ and $P(c_{j_n}|CF_l, s_j)$ and equation (xi) often becomes large.

The first term of equation (xi) represents how likely an entity that contains n_{j_m} as a content part is considered to be an antecedent, the second term represents how likely an entity that

contains c_{j_n} as a surface case is considered to be an antecedent, and the third term gives the probability that an entity that appears in location class l_{j_n} is an antecedent. The probabilities $P(n_{j_m})$ and $P(c_{j_n})$ are estimated from a large raw corpus. The probabilities $P(c_{j_n}|case\_type\_of(s_j))$ and $P(A'(s_j)=1|l_{j_n}case\_type\_of(s_j))$, which we call *location probabilities*, are estimated from the web corpus in which the relationship between an antecedent of a zero pronoun and a case slot, and the relationship between its surface case marker and a case slot are annotated by hand. Then, let us consider the probability $P(n_{j_m}|CF_l, s_j, A'(s_j) = 1)$ in the next section.

Probability of Component Part of Zero Pronoun

$P(n_{j_m}|CF_l, s_j, A'(s_j)=1)$ is similar to $P(n_j|CF_l, s_j, A(s_j)=1)$ and can be estimated approximately from case frames using the frequencies of case slot examples. However, while $A'(s_j)=1$ means s_j is not filled with input case component but filled with an entity as the result of zero anaphora resolution, case frames are constructed by extracting only the input case components. Therefore, the content part of a zero pronoun n_{j_m} is often not included in the case slot examples. To cope with this problem, we utilize generalized examples introduced in Section 2.3.3.

When one mention of an entity is tagged any category or recognized as an NE, we also use the category or the NE class as the content part of the entity. For example, if an entity {Prius} is recognized as an artifact name and assigned to *wo* case of the case frame *hanbai*(1) in Table 5.1, the system also calculates the following equation and uses the higher value:

$$\frac{P(NE{:}ARTIFACT|\textit{hanbai}(1), \textit{wo}, A'(\textit{wo}) = 1)}{P(NE : ARTIFACT)},$$

$$besides \quad \frac{P(Prius|\textit{hanbai}(1), \textit{wo}, A'(\textit{wo}) = 1)}{P(Prius)}.$$

5.2.3 Extension to Bridging Reference Resolution

In this section, we extend this probabilistic model to bridging reference resolution. In this model, bridging reference is regarded as a kind of zero anaphora, that is, the relation between a noun and its obligatory cases is considered to be parallel to that between a verb and its arguments or obligatory cases. Omitted obligatory case components are considered to be zero pronouns and resolved by the same process as zero anaphora resolution.

Unlike zero anaphora resolution, however, the anaphoricity of target noun phrases has to be determined before bridging reference resolution. Here, we use the same criterion as the anaphoricity determination for coreference resolution described in Chapter 4. We only consider such noun phrases anaphoric that have no modifier, that is, if a noun phrase with no modifier have case frames, the model considers the case slots as zero pronouns and resolves the omitted discourse entities.

Note that, while case frames for verbs describe both obligatory and optional cases, nominal case frames describe only obligatory cases. Therefore, we consider all case slots of nominal case as the target of bridging reference resolution.

5.2.4 Introduction of Salience Score

Previous works reported the usefulness of salience for anaphora resolution [64, 67]. In order to consider salience of an entity, we introduce salience score, which is calculated by the following set of simple rules:

- $+2$: mentioned with topical marker "*wa*," or at the end of a sentence.
- $+1$: mentioned without topical marker "*wa*."
- $+1$: assigned to a zero pronoun.
- $\times 0.5$ (="decay rate") : beginning of each sentence.

For example, we consider the salience score of the entity {Toyota} in the following text (identical to (5.1)):

(5.2) Toyota-*wa* 1997-*nen* hybrid car Prius-*wo hatsubai*(launch). 2000-*nen-karaha kaigai* (overseas)-*demo hanbai*(sell)-shiteiru.
(Toyota launched the hybrid car Prius in 1997. ϕ_1 started selling ϕ_2 overseas in 2000.)

In the first sentence, since {Toyota} is mentioned with topical marker "*wa*," the salience score is 2. At the beginning of the second sentence it becomes 1.0, and after assigned to the zero pronoun of "*hanbai*" it becomes 2.0. In this thesis, we use the salience score not as a probabilistic clue but as a filter to consider the target entity as a possible antecedent. When we use the salience score, we only consider the entities that have a salience score no less than 1.

5.3 Experiments

5.3.1 Experimental Setting

We created an anaphoric relation-tagged corpus consisting of 186 web documents (979 sentences) as mentioned in Section 4.3.1. In this corpus, all predicate-argument relations and relations between nouns are tagged. We show some examples:

(5.3) *Taro-ga shimbun-wo yonda. Taro-wa yoku yomu.*
newspaper read often read
TAG: *yonda* ⇐ *ga*:*Taro*, *wo*:*shimbun*,
yomu ⇐ *ga*:*Taro*,***wo*:*shimbun***
(Taro read a newspaper. Taro often reads ϕ.)

For the predicate "*yonda* (read)," "*Taro*" is tagged as nominative (*ga*) case component and "*shimbun*" is tagged as accusative (*wo*) case component. For the predicate "*yomu* (reads)," "*Taro*" is tagged as nominative (*ga*) case component and "*shimbun*" is tagged as omitted accusative (*wo*) case component, and such an omitted case component is the target of zero anaphora resolution (indicated in bold).

(5.4) *Taro-no imouto.* TAG: *imouto* ⇐ *no*:*Taro* (indispensable)
sister
(Taro's sister.)

(5.5) *Taro-wa imouto-to yatte-kita.* TAG: *imouto* ⇐ ***no*:*Taro*** **(indispensable)**
sister with came.
(Taro came with ϕ's sister.)

(5.6) *Kôen-niwa benchi-ga atta.* TAG: *benchi* ⇐ *no*:*kôen* (optional)
park was
(There was a bench in the park.)

(5.7) *Ki-no ita.* TAG: *ita* ⇐ *no*:*ki* (modifier)
wood board
(Board of wood.)

As for relations between nouns, both overt and implicit relations are tagged with the Japanese case marker "*no*" (of). In addition, relations between nouns are classified into three categories:

Table 5.3: Zero Anaphora and Bridging Reference Relations in Annotated Corpus.

Case		Zero Anaphora *ga* (nominative)	*wo* (accusative)	*ni* (dative)	Bridging Reference *no* (of)	All
all		142	48	43	63	296
Sentence Number	1st	12	1	5	5	23
	2nd	29	13	8	8	58
	3rd	25	8	6	12	51
	4th	26	9	10	14	59
	5th	12	3	1	4	20
	6th	5	4	7	7	23
	7th	10	5	3	4	22
	8th	11	1	0	4	16
	9th	7	1	2	3	13
	10th	5	3	1	2	11
Difference of Sentences between Anaphora and Antecedent	0	61	26	25	31	143
	1	31	15	11	14	71
	2	14	2	4	6	26
	3	9	2	1	6	28
	4	7	0	0	2	9
	5	2	1	1	2	6
	6	3	0	1	0	4
	7	3	2	0	1	6
	8	2	0	0	1	3

indispensable, optional, and modifier. Since it is not always obvious whether the relations are indispensable or not, borderline relations between indispensable and modifier are tagged optional. We consider only the implicit relations that are tagged indispensable as the target of bridging reference resolution. We do not evaluate the relations tagged optional, that is, if the system outputs such relations as bridging reference relations, we consider the outputs as neither positive or negative.

We used the first 51 documents for test and used the other 135 documents for calculating several probabilities. In the 51 test documents, 233 zero anaphora and 63 bridging reference relations were tagged between one of the mentions of the antecedent and the target predicate that had the zero pronoun. Table 5.3 shows the details of the corpus. Note that each document consisting of this corpus includes no more than 10 sentences. 48% (143/296) of zero pronouns refer to mentions of the same sentence, and 91% (268/296) of zero pronouns refer to the mentions that appear within 3 sentences.

Each parameter for proposed model was estimated using maximum likelihood from the data described in Table 5.4. The case frames for verbs and nominal case frames were automatically

Table 5.4: Data for Parameter Estimation.

probability	data
$P(n_j)$	raw corpus
$P(c_j)$	raw corpus
$P(c_j \mid case\_type\_of(s_j), A(s_j)=1)$	tagged corpus
$P(c_j \mid case\_type\_of(s_j), A'(s_j)=1)$	tagged corpus
$P(n_j \mid CF_l, s_j, A(s_j)=1)$	case frames
$P(n_j \mid CF_l, s_j, A'(s_j)=1)$	case frames
$P(CF_l \mid v_i)$	case structure analysis of raw corpus
$P(A(s_j)=\{0,1\} \mid CF_l, s_j)$	case structure analysis of raw corpus
$P(A'(s_j)=1 \mid l_j, case\_type\_of(s_j))$	tagged corpus

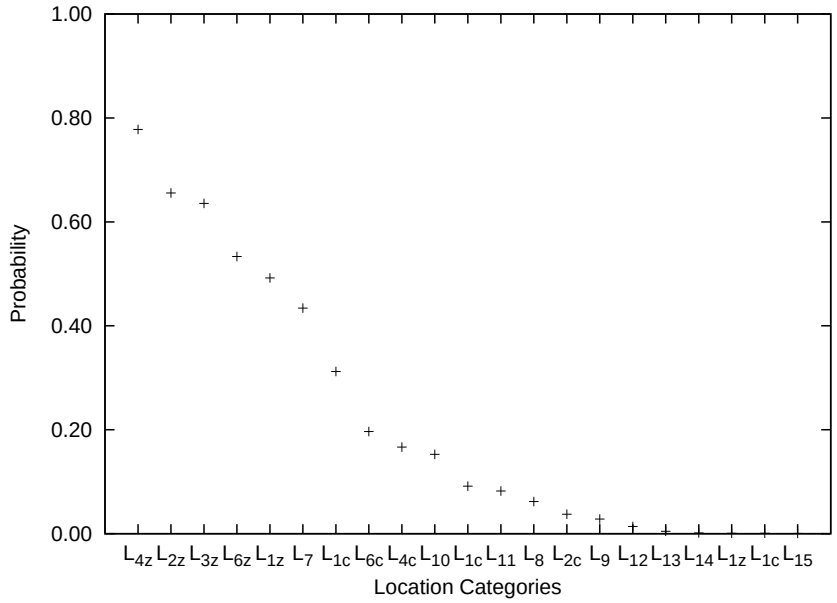


Figure 5.2: Location Probabilities for *ga* (nominative) Case.

constructed from web corpus comprising 1.6 billion sentences. The case structure analysis was conducted on 50 million sentences in the web corpus, and $P(n_j)$ and $P(c_j)$ were calculated from the same 50 million sentences. Figure 5.2 - 5.5 show the location probability for *ga* (nominative), *wo* (accusative), *ni* (dative) case, and bridging reference, respectively.

In order to concentrate on zero anaphora and bridging anaphora resolution, we used the correct morphemes, named entities, syntactic structures and coreferential relations that were annotated by hand. Since correct coreferential relations were given, the number of created entities was the same between the gold standard and the system output because zero anaphora resolution did not create new entities.

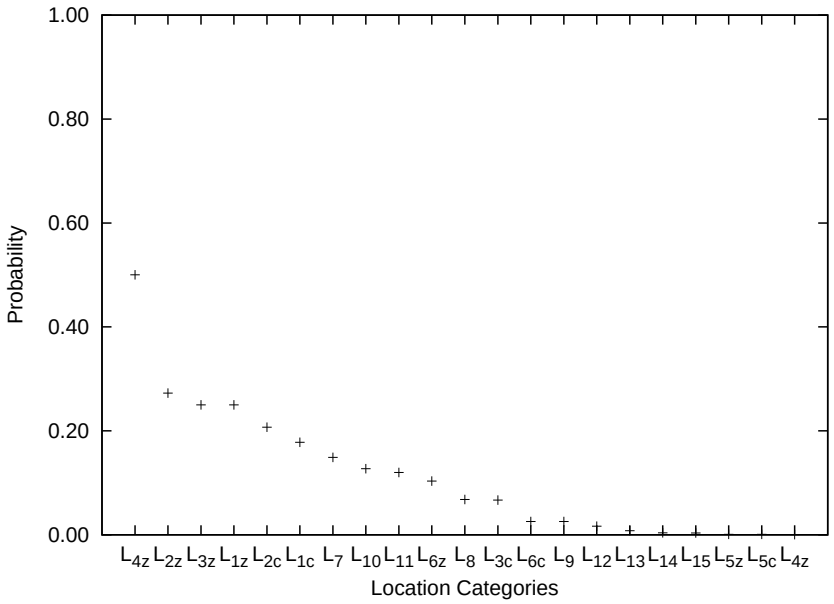


Figure 5.3: Location Probabilities for *wo* (accusative) Case.

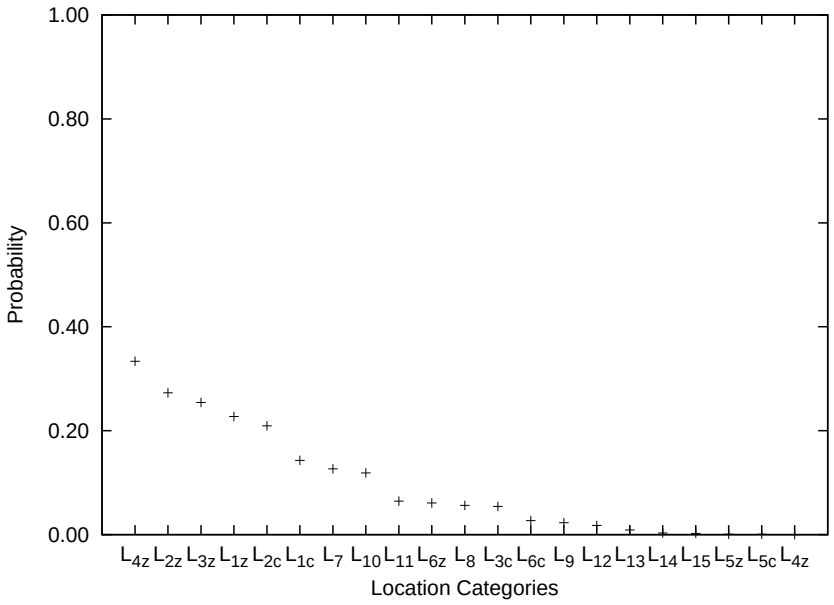


Figure 5.4: Location Probabilities for *ni* (dative) Case.

5.3.2 Experiments

We conducted experiments of zero anaphora resolution. Table 5.5 shows the experimental results when resolving zero anaphora and bridging reference simultaneously. While the performance of bridging reference resolution was not so high, the performance of zero anaphora was

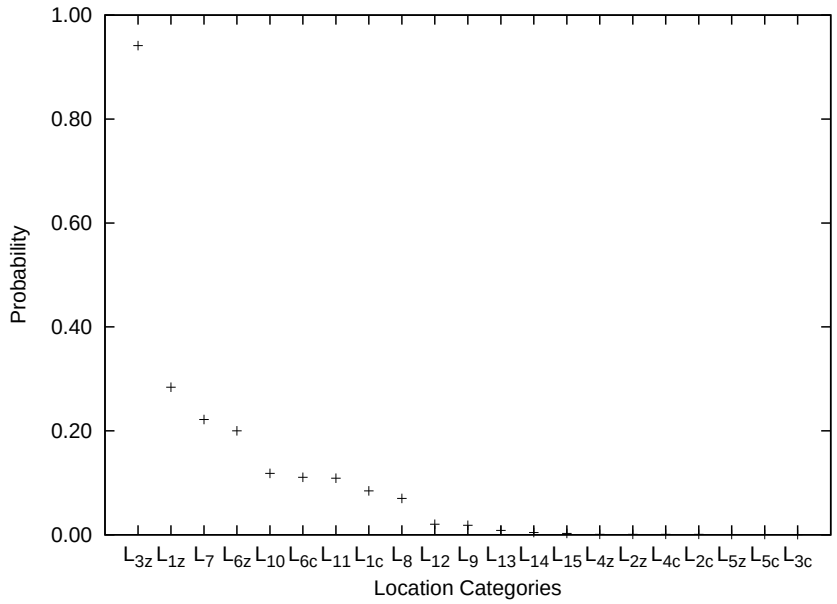


Figure 5.5: Location Probabilities for Bridging Reference.

Table 5.5: Experimental Results of Zero Anaphora and Bridging Reference Resolution.

	Precision	Recall	F-measure
Zero Anaphora	0.395 (92/233)	0.423 (92/214)	0.412
ga (nominative)	0.493 (70/142)	0.556 (70/126)	0.522
wo (accusative)	0.250 (12/48)	0.343 (12/35)	0.289
ni (dative)	0.233 (10/43)	0.189 (10/53)	0.208
Bridging Reference	0.365 (23/63)	0.291 (23/79)	0.324
Total	0.388 (115/296)	0.392 (115/293)	0.390

Table 5.6: Experimental Results of Anaphora Resolution Resolving Separately.

	Recall	Precision	F-measure
Zero Anaphora	0.395 (92/233)	0.442 (92/208)	0.417
Bridging Reference	0.159 (10/63)	0.222 (10/51)	0.175

reasonable. Especially, the resolution for *ga* case achieved an F-measure of 52.3%. Table 5.6 shows the performance when resolving zero anaphora and bridging reference separately. We can confirm the performance of bridging reference resolution is much improved by simultaneously resolving it with zero anaphora resolution. This is because there are not a lot of bridging relations in text and the salience score can not be estimated properly without zero anaphora

Table 5.7: Zero Anaphora and Bridging Reference Resolution Under Several Conditions.

CT	NE	SS	Recall	Precision	F-measure
√	√	√	**0.388 (115/296)**	**0.392 (115/293)**	**0.390**
√		√	0.318 (94/296)	0.332 (94/283)	0.325
	√	√	0.260 (77/296)	0.367 (77/210)	0.304
		√	0.243 (72/296)	0.356 (72/202)	0.289
√	√		0.338 (100/296)	0.178 (100/561)	0.233

resolution. On the other hand, the performance of zero anaphora resolution became worse by resolving simultaneously with bridging reference resolving. However, the difference was insignificant.

In order to confirm the effectiveness of generalized examples of case frames and salience score, we also conducted experiments under several conditions. The results are shown in Table 5.7, in which CT means generalized categories, NE means generalized NEs and SS means salience score. Without using any generalized examples, the F-measure is about 10% lower than the method using generalized examples, and we can confirm the effectiveness of the generalized examples. While generalized categories much improved the F-measure, generalized NEs contribute little. This may be because the NE rate is smaller than common noun rate, and so the effect is limited. We also confirmed that the salience score filter improved F-measure. Moreover, by using salience score filter, the zero anaphora resolution becomes about 2.0 times faster. This is because the system can avoid checking entities with low salience as antecedent candidates.

There are several major causes that led to analysis errors as follows:

Case Frame Sparseness When appropriate case frames were not constructed, the system cannot resolve zero anaphora. As will be discussed in Chapter 6, by using larger corpus to construct case frames, this problem can be alleviated to some extent. However, since the case frames were constructed from only overt predicate argument pairs, some frequently omitted case components are not described in the case frames; thus our model cannot avoid this sparseness problem.

Unknown word By using generalized examples, our model dealt with such antecedents as was not described in case frames. However, unknown words except NEs cannot be generalized, and thus our model cannot recognized such antecedents.

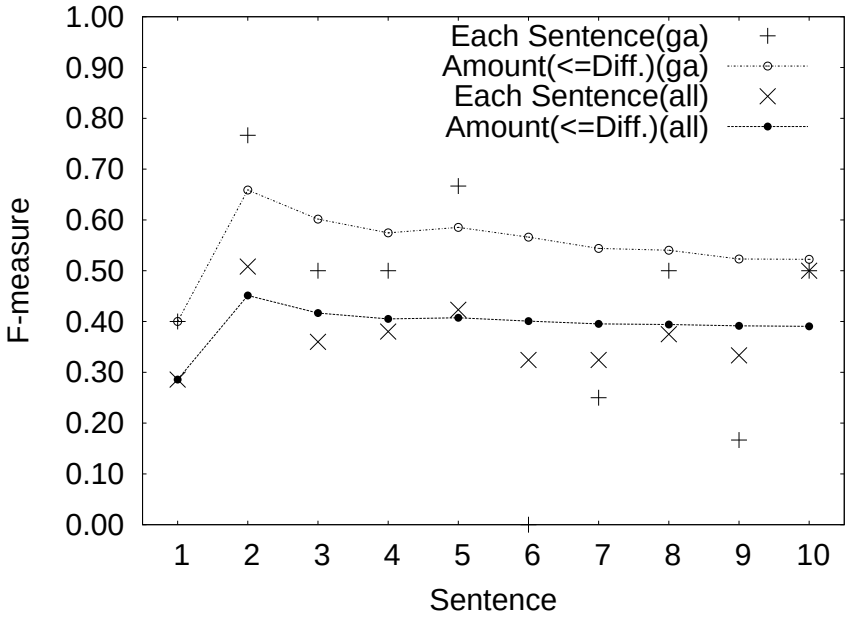


Figure 5.6: F-measure for Each Sentence Number.

Lack of the recognition of discourse By introducing salience score of discourse entities, our model considers global discourse to some extent. However, there is much more information concerning discourse structure, such as paragraph boundaries, and some zero anaphoric relations are hard to resolve without deeper recognition of discourse structure.

Lack of consideration of the existence of writer or speaker In some texts, such as essays and monologues, most of zero pronouns refer to the author of the sentences or speaker of the utterance, which are not written in the text obviously. However, since our model does not consider such entities, our model often outputs erroneous antecedents for such zero pronouns.

(5.8) *Gokiburi-ga tonde koshi-wo nukashita.*
black beetle flew be quite unmanned by the sight
(A black beetle flew up; ϕ was quite unmanned by the sight.)

For example, the *ga* (nominative) case of "*koshi-wo nukasu*" (be quite unmanned by the sight) in (5.8) should be assigned to the writer. However, since our model does not consider the existence of the writer, it tries to assign the case to a discourse entity mentioned in the text such as "*gokiburi*" (black beetle).

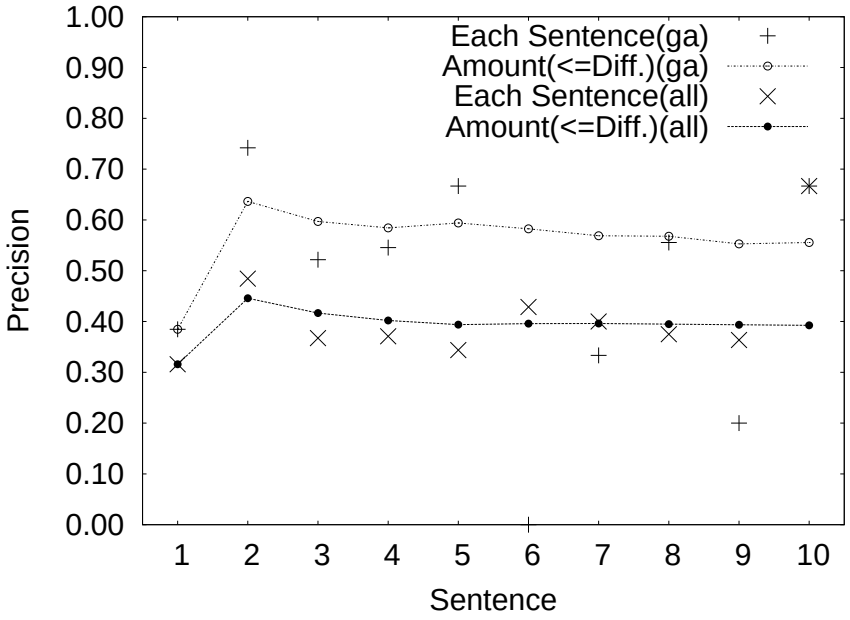


Figure 5.7: Precision for Each Sentence Number.

5.3.3 Discussion

Generally speaking, in longer texts, more discourse entities appear and zero anaphora resolution becomes more difficult. In order to confirm the relation between the performance of anaphora resolution and the length of texts, we investigated the F-measure for each sentence number. Figure 5.6 shows the results.

The F-measure of second sentence was highest. Although there are more possible antecedents when analyzing second sentence than analyzing first sentence, the F-measure of second sentence was higher than that of first sentence. This may be because zero anaphora in the second sentence is the most typical as shown in Table 5.3 and easy to resolve.

The F-measure for first two sentences was 45% for all, 66% for *ga* case. In some natural language tasks, such as recognizing textual entailment, query analysis in information retrieval and question answering, the texts to be analysed are often short, and we can say that our system is practicable for such tasks.

When considering the practicability for more wide-ranging tasks, more accurate system is required. Thus, hereafter, we focus on the precision of the system. In order to extract reliable system outputs, we first investigated the relation between the precision of anaphora resolution and the sentence number. Figure 5.7 shows the results. As well as F-measure, the precision

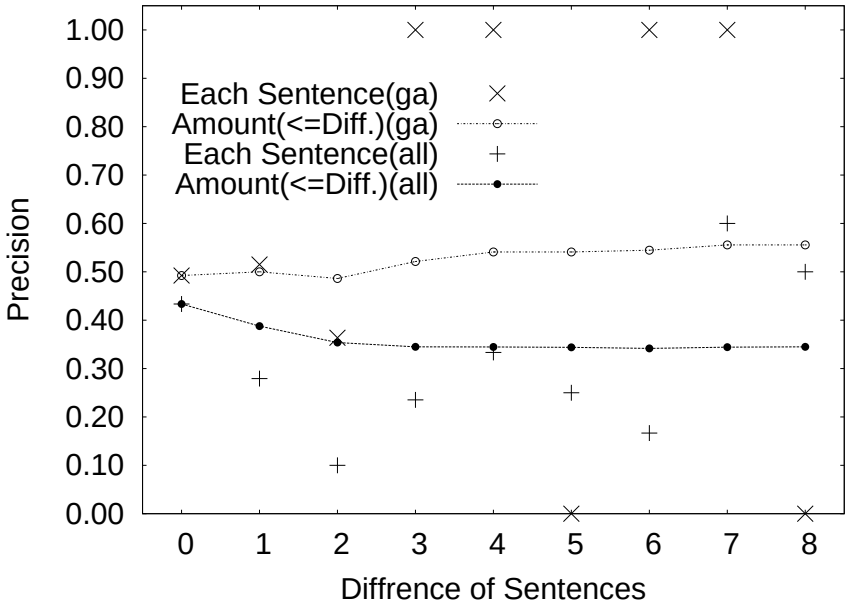


Figure 5.8: Precision Classified by the Distance between the Anaphor and its Antecedent.

of second sentence was high, and the precision for first two sentences was 45% for all, 64% for *ga* case. However, we can confirm little correlation between the precision and the sentence number.

We also investigated the precision of system outputs classified by the distance between the anaphor and its antecedent. Figure 5.8 shows the results. As for difference of 0, that is, intra-sentence anaphora, the precision was 43% for all, 49% for *ga* case. There was little correlation between the precision and the distance, though we expected to obtain higher precision for closer reference.

Next, in order to extract reliable outputs, we introduced parameter α, which corrected $P(A'(s_j)= 1|ent_j, CF_l, s_j)$ in the probability that a case slot s_j is assigned to an entity ent_j. When using this parameter, the probability $P(A'(s_j)= 1|ent_j, CF_l, s_j)$ is multiplied by the parameter α. For examples, if using 1/2 as parameter α, the probability $P(A'(s_j)= 1|ent_j, CF_l, s_j)$ is reduced in half and fewer zero anaphoric relations are outputted.

Figure 5.9 shows the results. We can confirm that by using smaller parameter α, the outputted zero anaphoric relations became more reliable. When we used the parameter of 1/16, the system achieved the precision of 70%. However, by using smaller parameter α, the recall became small, and the highest F-measure was obtained when we do not use parameter α, that is, $\alpha = 1$.

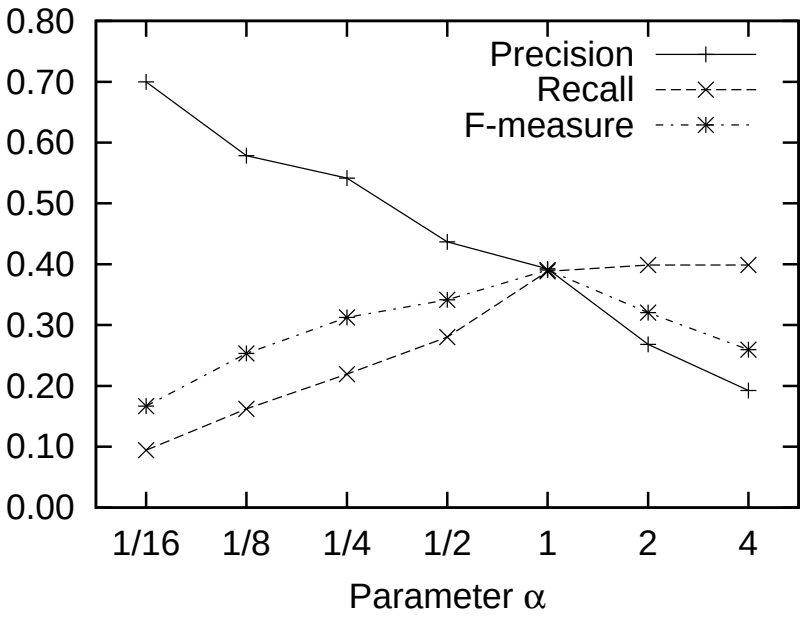


Figure 5.9: Trade-off between Recall and Precision.

Finally, we varied the decay rate of salience score that is introduced in Section 5.2.4. Figure 5.10 shows the results. Although we expected to obtain higher precision with small decay rate, the highest precision was achieved by the default decay rate 0.5. When we used the decay rate smaller than 0.5, the recall score became worse significantly. This may be because with the decay rate smaller than 0.5 the system can hardly recognize the inter-sentence anaphora.

5.3.4 Comparison with Previous Work

We now compare our model with some previous works for zero anaphora resolution.

Seki et al. [65] achieved a precision of 48.9%, a recall of 88.2%, and an F-measure of 62.9% for zero pronoun detection, and an accuracy of 54.0% for antecedent estimation on 30 newspaper articles, that is, they achieved an F-measure of about 34% for whole zero pronoun resolution.

Kawahara and Kurohashi's model [9] achieved almost an F-measure of 50% against newspaper articles. However, as a result of our experiment against web documents, the F-measure was only about 20%. This may be because anaphoric relations in web documents were not so clear as those in newspaper articles and more difficult to recognize.

Iida et al. proposed an machine learning model for zero pronoun resolution [10,68]. In their experiments on Japanese news paper articles, they used correct zero pronouns and concentrated

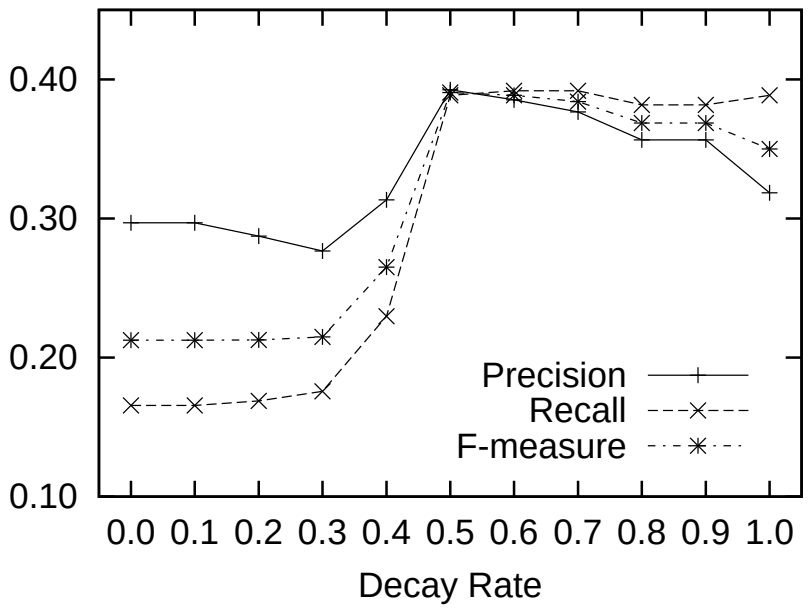


Figure 5.10: The Performance Under Several Decay Rates of Salience.

on *ga* (nominative) case. Their model achieved the accuracy of 51.0%.

It is difficult to directly compare their results with ours due to the difference of the corpus, but our method achieved an F-measure of 41.2% for all cases and an F-measure of 52.2% for *ga* (nominative) case, and we can confirm that our model achieves reasonable performance considering the task difficulty.

As for bridging reference in Japanese text, there is very few previous work. Murata et al. proposed heuristic rule based approach for bridging reference [30]. They obtained a recall of 63% and a precision of 68%. However, we cannot compare our results to theirs, since their experiments were conducted on fairy tales and editorials whose details were not opened.

5.4 Summary of this Chapter

In this chapter, we proposed a probabilistic model for Japanese zero anaphora and bridging reference resolution. First, this model recognizes discourse entities and links all mentions to them. Zero pronouns are then detected by case structure analysis based on automatically constructed case frames. Their appropriate antecedents are selected from the entities with high salience scores, based on the case frames and several preferences on the relation between a zero pronoun and an antecedent. Case structure and zero anaphora relation are simultaneously de-

termined based on probabilistic evaluation metrics. By using automatically constructed wide-coverage case frames that include generalized examples and introducing salience score filter, our model achieves reasonable performance against web corpus. As future work, we can think of integrating this model to a lexicalized probabilistic model for Japanese syntactic and case structure analysis [11].

Chapter 6

The Effect of Corpus Size on Case Frame Construction for Discourse Analysis

6.1 Introduction

Very large corpora obtained from the Web have been successfully utilized for many natural language processing (NLP) applications, such as prepositional phrase (PP) attachment, other-anaphora resolution, spelling correction, confusable word set disambiguation and machine translation [69–73].

Most of the previous work utilized only the surface information of the corpora, such as n-grams, co-occurrence counts, and simple surface syntax. This may have been because these studies did not require structured knowledge, and for such studies, the size of currently available corpora is considered to have been almost enough. For instance, while Brants et al. [73] reported that translation quality continued to improve with increasing corpus size for training language models at even size of 2 trillion tokens, the increase became small at the corpus size of larger than 30 billion tokens.

However, for more complex NLP tasks, such as case structure analysis and zero anaphora resolution, it is necessary to obtain more structured knowledge, such as case frames, which describe the cases for each predicate and the types of nouns that can fill a case slot. Note that case frames offer not only the knowledge of the relationships between a predicate and its particular case slot, but also the knowledge of the relationships among a predicate and its multiple case slots. To obtain such knowledge, very large corpora seem to be necessary; however it is still unknown how much corpora would be required to obtain good coverage.

For example, Kawahara and Kurohashi proposed a method for constructing case frames from large corpora [22] in order to acquire wide-coverage case frames, and a model for syntactic

and case structure analysis of Japanese that based upon case frames [11]. However, they did not demonstrate whether the coverage of case frames was wide enough for these tasks and how dependent the performance of the model was on the corpus size for case frame construction.

This chapter aims to address these questions. A very large Japanese corpus consisting of approximately 100 billion words, or 1.6 billion unique sentences, was collected from the Web. Subsets of the corpus were then randomly selected to obtain corpora of different sizes ranging from 1.6 million to 1.6 billion sentences. Case frames were constructed from each corpus and applied to syntactic and case structure analysis, and zero anaphora resolution. The relationships between the corpus size and the performance for these analyses confirmed the effectiveness of the larger corpora.

6.2 Related Work

Many NLP tasks have successfully utilized very large corpora, most of which were acquired from the Web [74]. Volk [69] proposed a method for resolving PP attachment ambiguities based upon Web data. Modjeska et al. [70] used the Web for resolving nominal anaphora. Lapata and Keller [71] investigated the performance of web-based models for a wide range of NLP tasks, such as MT candidate selection, article generation, and countability detection. Nakov and Hearst [75] solved relational similarity problems using the Web as a corpus.

With respect to the effect of corpus size on NLP tasks, Banko and Brill [76] showed that for content sensitive spelling correction, increasing the training data size increased the accuracy. Atterer and Schütze [72] investigated the effect of corpus size when combining supervised and unsupervised training for disambiguation; they found that the combined system only improved the performance of the parse for small training sets. Brants et al. [73] varied the amount of language model training data from 13 million to 2 trillion tokens and applied these models to machine translation systems. They reported that translation quality continued to improve with increasing corpus size for training language models at even size of 2 trillion tokens. Suzuki and Isozaki [77] provided evidence that the use of more unlabeled data in semi-supervised learning could improve the performance of NLP tasks, such as part-of-speech tagging, syntactic chunking, and named entities recognition.

There are several methods to obtain the information extracted from very large corpora. Search engines, such as Google and Altavista, are often used to obtain Web counts (e.g. [78,79]). However, search engines are not designed for NLP research and the reported hit counts are

subject to uncontrolled variations and approximations. Therefore, several researchers have collected corpora by themselves for their work from the Web. For English, Banko and Brill [80] collected a corpus with 1 billion words from variety of English texts. Liu and Curran [81] created a Web corpus for English that contained 10 billion words and showed that for content-sensitive spelling correction the Web corpus results were better than using a search engine. Halacsy et al. [82] created a corpus with 1 billion words for Hungarian from the Web by downloading 18 million pages. Others utilize publicly available corpus such as the North American News Corpus (NANC) and the Gigaword Corpus [83]. For instance, McClosky et al. [84] proposed a simple method of self-training a two phase parser-reranker system using NANC.

As for Japanese, Kawahara and Kurohashi [22] collected 23 million pages and created a corpus with approximately 20 billion words. Google released Japanese n-gram constructed from 20 billion Japanese sentences [85]. Several news wires are publicly available consisting of tens of million sentences. Kotonoha project is now constructing a balanced corpus of the present-day written Japanese consisting of 50 million words [86].

6.3 Discourse Analysis Based on Case Frames

6.3.1 Model for Syntactic and Case Structure Analysis

In order to investigate the effect of corpus size on complex NLP tasks, the constructed cases frames were applied to an integrated probabilistic model for Japanese syntactic and case structure analysis proposed by Kawahara and Kurohashi [11] and a probabilistic model for Japanese zero anaphora resolution. In this section, we briefly describe the integrated probabilistic model for Japanese syntactic and case structure analysis.

Kawahara and Kurohashi [11] proposed an integrated probabilistic model for Japanese syntactic and case structure analysis based upon case frames. Case structure analysis recognizes predicate argument structures. Their model gives a probability to each possible syntactic structure T and case structure L of the input sentence S, and outputs the syntactic and case structure that have the highest probability. That is to say, the system selects the syntactic structure T_{best} and the case structure L_{best} that maximize the probability $P(T, L|S)$:

$$\begin{aligned}(T_{best}, L_{best}) &= \mathrm{argmax}\,(T, L) P(T, L|S) \\ &= \mathrm{argmax}\,(T, L) P(T, L, S) \end{aligned} \quad \text{(i)}$$

The last equation is derived because $P(S)$ is constant. $P(T, L, S)$ is defined as the product of a

probability for generating a clause C_i as follows:

$$P(T, L, S) = \prod_{i=1..n} P(C_i|b_{h_i}) \quad \text{(ii)}$$

where n is the number of clauses in S, and b_{h_i} is C_i's modifying *bunsetsu*[1]. $P(C_i|b_{h_i})$ is approximately decomposed into the product of several generative probabilities such as $P(A(s_j)=1|CF_l, s_j)$ and $P(n_j|CF_l, s_j, A(s_j)=1)$, where the function $A(s_j)$ returns 1 if a case slot s_j is filled with an input case component; otherwise 0. $P(A(s_j)=1|CF_l, s_j)$ denotes the probability that the case slot s_j is filled with an input case component, and is estimated from resultant case structure analysis of a large raw corpus. $P(n_j|CF_l, s_j, A(s_j)=1)$ denotes the probability of generating a content part n_j from a filled case slot s_j in a case frame CF_l, and is calculated by using case frames. For details see [11].

6.3.2 Model for Zero Anaphora Resolution

Anaphora resolution is one of the most important techniques for discourse analysis. We apply the probabilistic model for Japanese zero anaphora resolution described in Chapter 5. In this section, we summarize the model.

Our model first resolves coreference and identifies discourse entities; then gives a probability to each possible case frame CF and case assignment CA when target predicate v, input case components ICC and existing discourse entities ENT are given, and outputs the case frame and case assignment that have the highest probability. That is to say, their model selects the case frame CF_{best} and the case assignment CA_{best} that maximize the probability $P(CF, CA|v, ICC, ENT)$:

$$\begin{aligned}&(CF_{best}, CA_{best})\\ &\quad = \mathrm{argmax}\,(CF, CA) P(CF, CA|v, ICC, ENT)\end{aligned} \quad \text{(iii)}$$

$P(CF, CA|v, ICC, ENT)$ is approximately decomposed into the product of several probabilities. Case frames are used for calculating $P(n_j|CF_l, s_j, A(s_j)=1)$, the probability of generating a content part n_j from a case slot s_j in a case frame CF_l, and $P(n_j|CF_l, s_j, A'(s_j)=1)$, the probability of generating a content part n_j of a zero pronoun, where the function $A'(s_j)$ returns 1 if a case slot s_j is filled with an antecedent of a zero pronoun; otherwise 0.

[1] In Japanese, *bunsetsu* is a basic unit of dependency, consisting of one or more content words and the following zero or more function words. It corresponds to a base phrase in English.

Table 6.1: Corpus Sizes and Thresholds.

corpus size for case frame construction (sentences)	1.6M	6.3M	25M	100M	400M	1.6G
threshold α introduced in Sec. 2.3.2	2	3	4	5	7	10
corpus size to estimate generative probability (sentences)	1.6M	3.2M	6.3M	13M	25M	50M

$P(n_j|CF_l, s_j, A'(s_j) = 1)$ is similar to $P(n_j|CF_l,\ s_j, A(s_j) = 1)$ and estimated from the frequencies of case slot examples in case frames. However, while $A'(s_j) = 1$ means s_j is not filled with an overt argument but filled with an antecedent of zero pronoun, case frames are constructed from overt predicate argument pairs. Therefore, the content part n_j is often not included in the case slot examples. To cope with this problem, their model also utilizes generalized examples to estimate $P(n_j|CF_l,\ s_j, A(s_j){=}1)$.

6.4 Experiments

6.4.1 Construction of Case Frames

In order to investigate the effect of corpus size, case frames were constructed from corpora of different sizes. Japanese sentences were first collected from the Web using the method proposed by Kawahara and Kurohashi [22]. Approximately 6 billion Japanese sentences consisting of approximately 100 billion words were acquired from 100 million Japanese web pages. After discarding duplicate sentences, which may have been extracted from mirror sites, a corpus was acquired comprising of 1.6 billion (1.6G) unique Japanese sentences consisting of approximately 25 billion words. The average number of characters and words in each sentence was 28.3, 15.6, respectively. Subsets of the corpus were then randomly selected for five different sizes; 1.6M, 6.3M, 25M, 100M, and 400M sentences to obtain corpora of different sizes.

Case frames were constructed from each corpus. JUMAN and the rule-based syntactic parser, KNP [12] were employed to parse each corpus. The threshold α introduced in Section 2.3.2 was changed depending upon the size of the corpus as shown in Table 6.1. Completing the case frame construction took about two weeks using 600 CPUs. Table 6.2 shows the statistics for the constructed case frames. The number of predicates, the average number of examples

Table 6.2: Statistics of the Constructed Case Frames.

Corpus size (sentences)		1.6M	6.3M	25M	100M	400M	1.6G
# of predicate		2460	6134	13532	27226	42739	65679
(type)	verb	2039	4895	10183	19191	28523	41732
	adjective	154	326	617	1120	1641	2318
	noun with copula	267	913	2732	6915	12575	21629
average # of case frames for a predicate		15.9	12.2	13.3	16.1	20.5	25.3
average # of case slots for a case frame		2.95	3.44	3.88	4.21	4.69	5.08
average # of examples for a case slot		4.89	10.2	19.5	34.0	67.2	137.6
average # of unique examples for a case slot		1.19	1.85	3.06	4.42	6.81	9.64
average # of generalized examples for a case slot		0.14	0.24	0.37	0.49	0.67	0.84
file size(Byte)		8.9M	20M	56M	147M	369M	928M

and unique examples for a case slot, and whole file size were confirmed to be heavily dependent upon the corpus size. However, the average number of case frames for a predicate, and case slots for a case frame did not.

6.4.2 Coverage of Constructed Case Frames

Setting

In order to investigate the coverage of the resultant case frames, we created a syntactic relation, case structure, and anaphoric relation annotated corpus consisting of 186 web documents (979 sentences) as mentioned in Section 4.3.1. This corpus was manually annotated using the same criteria as relevance-tagged corpora [87]. There were 2,390 annotated relationships between predicates and their direct (not omitted) case components and 837 zero anaphoric relations in the corpus.

Two evaluation metrics were used depending upon whether the target case component was omitted or not. When evaluating the overt case component of a predicate, the target component was judged to be covered by case frames if the target component itself was included in the examples for one of the corresponding case slots of the case frame. When evaluating the omitted case component of a predicate, not only the target component but also all mentions that refer to the same entity were checked.

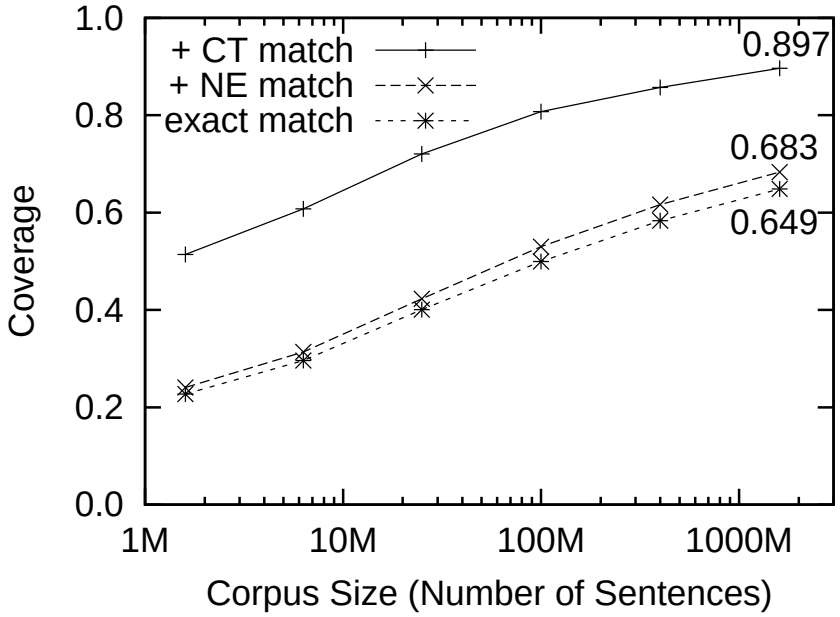


Figure 6.1: Coverage of CF (overt argument).

Coverage of Case Frames

Figure 6.1 shows the coverage of case frames for the overt argument, which would have tight relations with case structure analysis. The lower line shows the coverage without considering generalized examples, the middle line shows the coverage considering generalized NE examples, and the upper line shows the coverage considering all generalized examples.

Figure 6.2 shows the coverage of case frames for the omitted argument, which would have tight relations with zero anaphora resolution. The upper line shows the coverage considering all generalized examples, which considered to be the upper bound of performance for the zero anaphora resolution system described in Chapter 5.

When compared with Figure 6.1, two characteristics were confirmed. First, the lower and middle lines of Figure 6.2 were located lower than the corresponding lines in Figure 6.1. This may reflect that some frequently omitted case components are not described in the case frames because the case frames were constructed from only overt predicate argument pairs. Secondly, the effect of generalized NE examples was confirmed to be more evident for the omitted argument reflecting the important rule of NEs in zero anaphora resolution.

Both figures confirm that the coverage was improved by using larger corpora and there was no saturation even when the largest corpus of 1.6 billion sentences was used. When using the

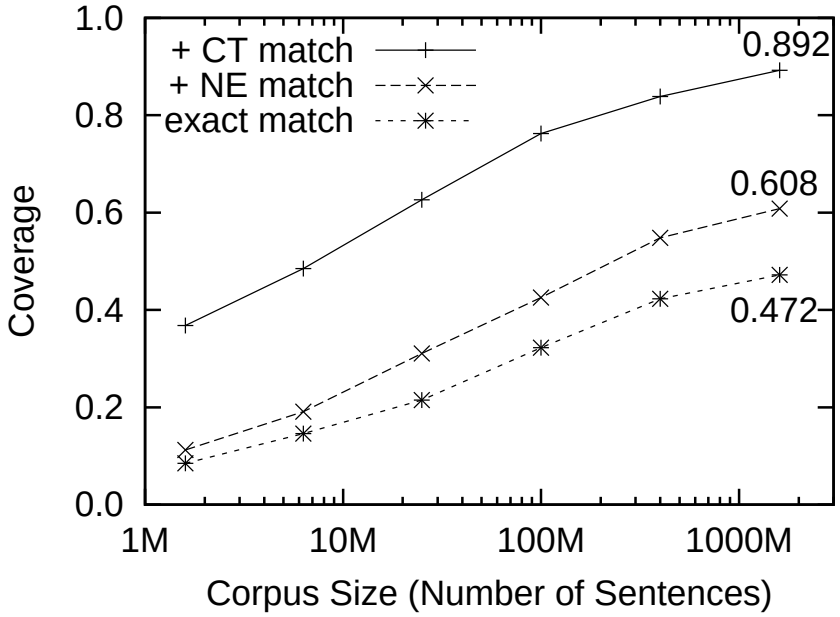


Figure 6.2: Coverage of CF (omitted argument).

largest corpus and all generalized examples, the case frames achieved a coverage of almost 90% for both the overt and omitted argument.

Figure 6.3 shows the coverage of case frames for each predicate type, which is calculated for both overt and omitted argument considering all generalized examples. The case frames for verbs achieved the coverage of 93.0%. There are 189 predicate-argument pairs that are not included case frames; 11 pairs of them are due to lack of a case frame of target predicate itself. For adjective, the coverage was 78.8%. The main cause of the lower coverage than the coverage for verb may be that several adjectives are mainly used in restrictive manner and the predicate argument relations concerning these adjectives are not used for case frame construction. For noun with copula, the coverage was only 54.5%. However, most predicate argument relations concerning nouns with copula are easily inferred, and thus the low coverage would not quite affect the performance of discourse analysis.

6.4.3 Syntactic and Case Structure Analysis

Accuracy of Syntactic Analysis

We evaluated syntactic structures analyzed by the method described in Section 6.3.1. Our experiments were run on hand-annotated 759 web sentences. The resultant syntactic structures

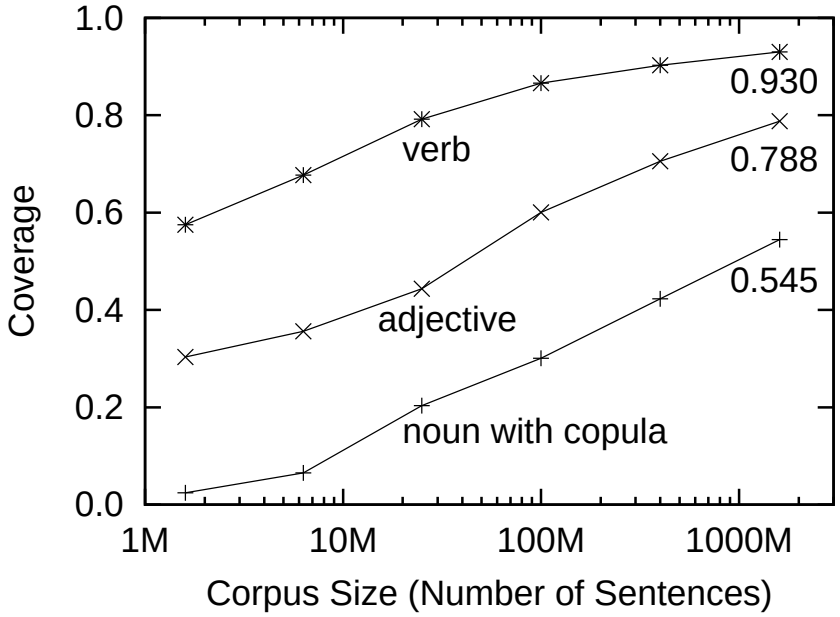


Figure 6.3: Coverage of CF for Each Predicate Type.

were evaluated with regard to dependency accuracy, the proportion of correct dependencies out of all dependencies[2].

Figure 6.4 shows the accuracy of syntactic structures. These experiments were conducted with case frames constructed from corpora of different sizes. The corpus size to estimate generative probability of a case slot in Chapter 5 was also changed depending upon the size of the corpus for case frame construction as shown in Table 6.1.

In Figure 6.4, "w/o case frames" shows the accuracy of the rule-based syntactic parser KNP that does not use case frames. Since the model described in Chapter 5 assumes the existence of reasonable case frames, when we used case frames constructed from very small corpus, such as 1.6M and 6.3M sentences, the accuracy was lower than that of the rule-based syntactic parser. Moreover, when we tested the model described in Chapter 5 without any case frames, the accuracy was 88.5%.

We confirmed that better performance was obtained when using case frames constructed from larger corpora, and the accuracy of 89.4%[3] was achieved when using the case frames constructed from 1.6G sentences. However the effect was limited. This may be because there

[2]Note that Kawahara and Kurohashi [11] exclude the dependency between the last two *bunsetsu*, since Japanese is head-final and thus the second last *bunsetsu* unambiguously depends on the last *bunsetsu*.

[3]It corresponds to 87.7% in Kawahara and Kurohashi's [11] evaluation metrics.

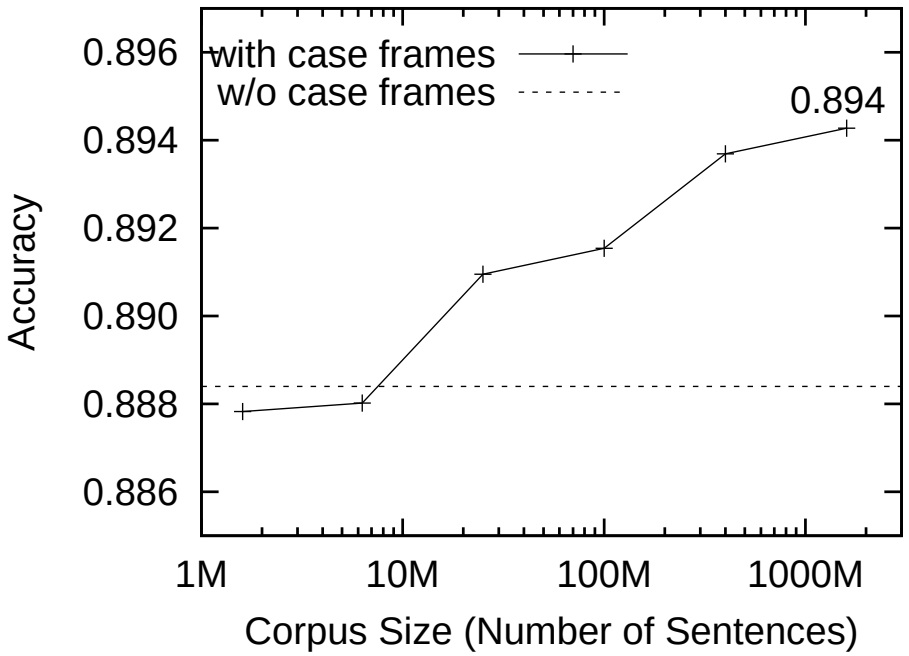


Figure 6.4: Accuracy of Syntactic Analysis.

are various causes of dependency error and the case frame sparseness problem occupies at most 10% of them.

Although the model proposed by Kawahara and Kurohashi [11] does not utilize generalized examples of case frames, we considered that these examples can benefit for the accuracy of syntactic analysis. Therefore, we tried several models that utilize generalized examples. However, we cannot confirm any improvement.

Accuracy of Case Structure Analysis

Case structure analysis was conducted on 215 web sentences in order to investigate the accuracy. The case markers of topic marking phrases and clausal modifiers are evaluated by comparing them with the gold standard in the corpus. The experimental results are shown in Figure 6.5. We confirmed that the accuracy of case structure analysis strongly depends on corpus size for case frame construction.

Analysis Speed

Table 6.3 shows the time for analyzing syntactic and case structure of 759 web sentences. Although the time for analysis became longer when using case frames constructed from a larger

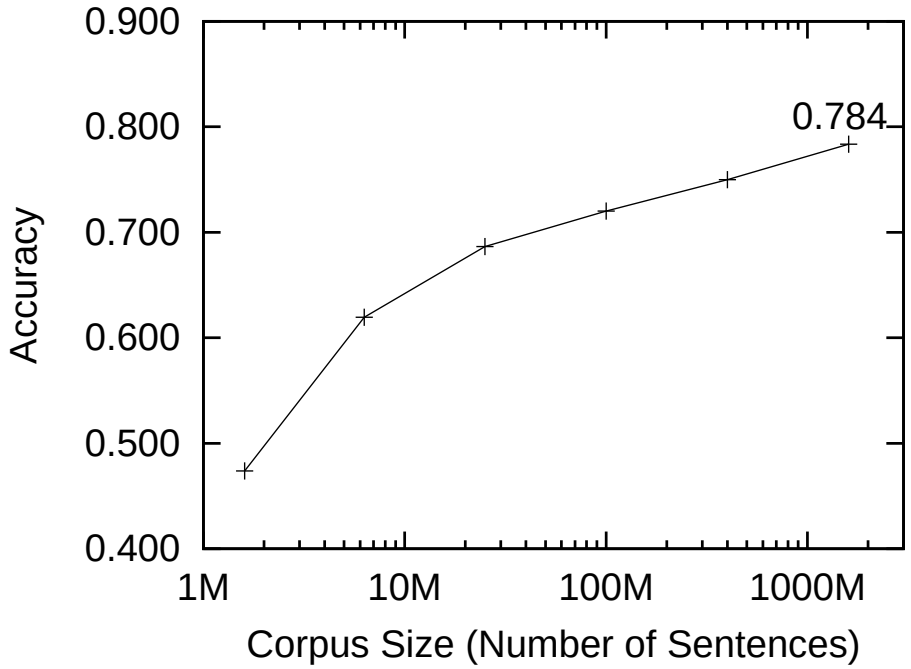


Figure 6.5: Accuracy of Case Structure Analysis.

Table 6.3: Corpus Size and Time for Syntactic and Case Structure Analysis.

corpus size	1.6M	6.3M	25M	100M	400M	1.6G
time (sec.)	850	1244	1833	2696	3783	5553

corpus, the growth rate was smaller than the growth rate of the size for case frames described in Table 6.2.

The increase of analysis time cannot be ignored comparing the increase in accuracy of syntactic analysis. Therefore, we cannot absolutely say that case frames constructed from larger corpora are desirable for syntactic analysis. However, since there is enough increase in accuracy of case structure analysis, we can say that case frames constructed larger corpora are desirable for case structure analysis.

6.4.4 Zero Anaphora Resolution

Accuracy of Zero Anaphora Resolution

We used an anaphoric relation annotated corpus consisting of 186 web documents (979 sentences) to evaluate zero anaphora resolution. As well as mentioned in Section 5.3.1, we used

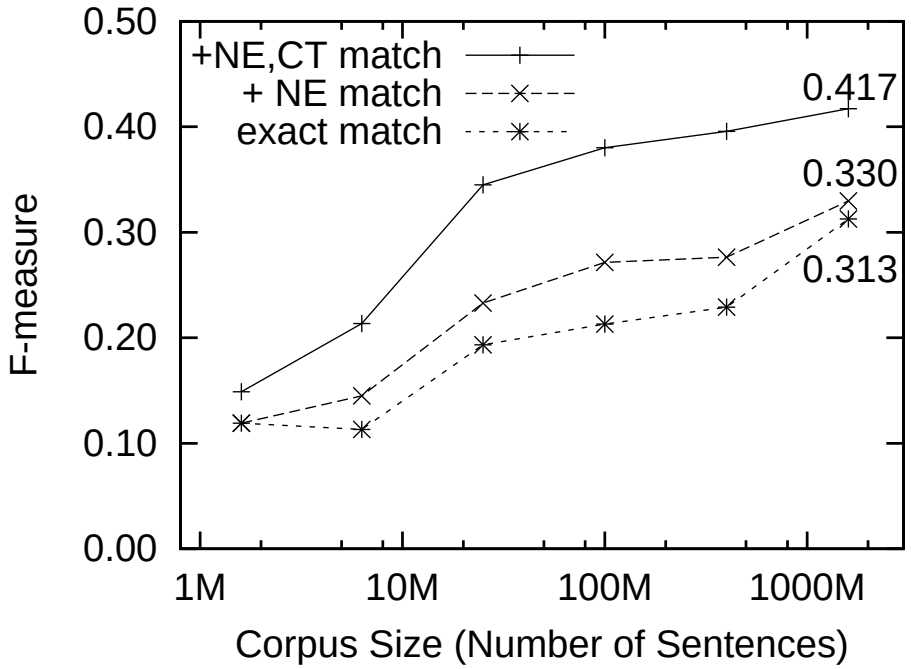


Figure 6.6: F-measure of Zero Anaphora Resolution.

first 51 documents for test and used the other 135 documents for calculating several probabilities. In the 51 test documents, 233 zero anaphoric relations were annotated between one of the mentions of the antecedent and corresponding predicate that had zero pronoun.

In order to concentrate on evaluation for zero anaphora resolution, we used the correct morphemes, named entities, syntactic structures and coreference relations that were manually annotated. Since correct coreference relations were given, the number of created entities was the same between the gold standard and the system output because zero anaphora resolution did not create new entities. The experimental results are shown in Figure 6.6 - 6.8.

Figure 6.6 shows the F-measure. The upper line shows the performance using all generalized examples, the middle line shows the performance using only generalized NEs, and the lower line shows the performance without using any generalized examples. While generalized categories much improved the F-measure, generalized NEs contributed little. This tendency is similar to that of coverage of case frames for omitted argument shown in Figure 6.2. These experimental results also show the effectiveness of the corpus size on zero anaphora resolution.

Figure 6.7 and 6.8 show the recall and precision, respectively. From Figure 6.7, we can confirm that the larger corpus size benefits for the recall of zero anaphora resolution. On the

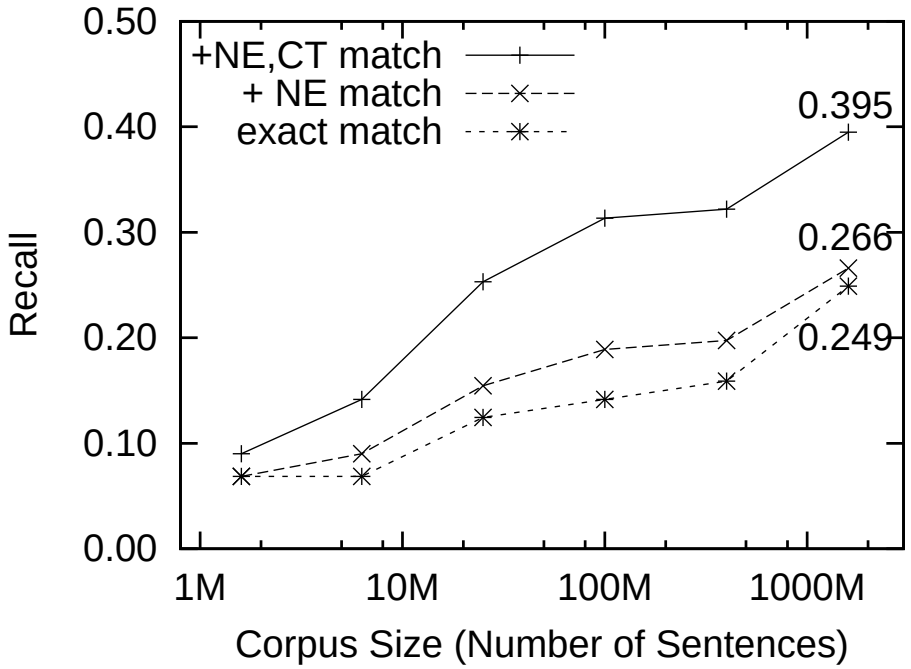


Figure 6.7: Recall of Zero Anaphora Resolution.

other hand, there was little correlation between the precision and the corpus size for case frame construction.

Analysis Speed

Table 6.4 shows the time for resolving zero anaphora in 51 web documents consisting of 278 sentences. The time for analysis became longer when using case frames constructed from larger corpora, which tendency is similar to the growth of the time for analyzing syntactic and case structure.

However, unlike syntactic and case structure analysis, the performance for the zero anaphora resolution is quite low when using case frames constructed from small corpora, and we can say that case frames constructed from larger corpora are essential for zero anaphora resolution even though the time for analysis become long.

6.4.5 Discussion

Experimental results of both syntactic and case structure analysis, and zero anaphora resolution show the effectiveness of a larger corpus in case frame acquisition for Japanese discourse analysis. Up to the corpus size of 1.6 billion sentences, or approximately 100 billion words, these

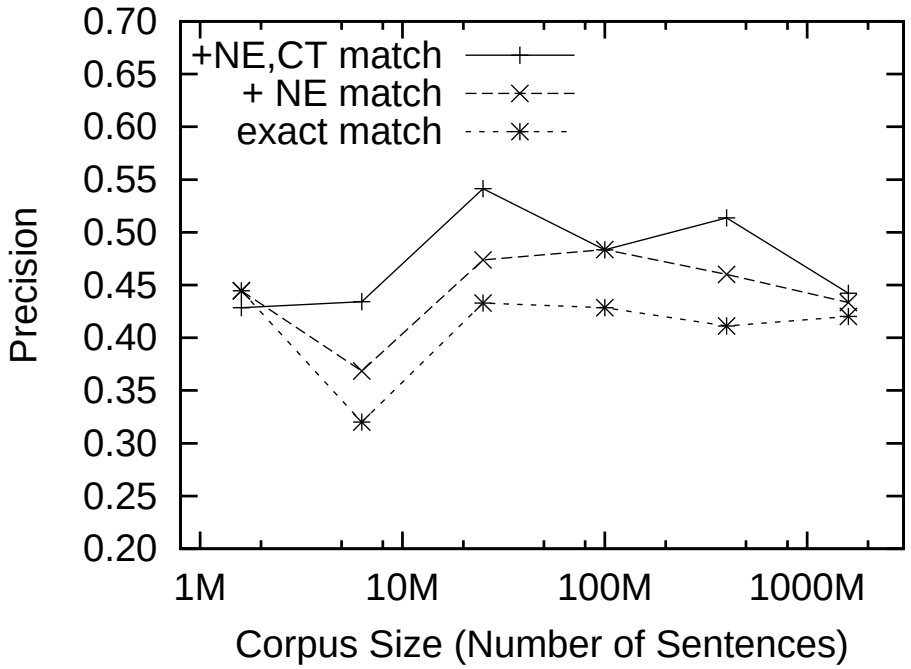


Figure 6.8: Precision of Zero Anaphora Resolution.

Table 6.4: Corpus Size and Time for Zero Anaphora Resolution.

corpus size	1.6M	6.3M	25M	100M	400M	1.6G
time (sec.)	538	545	835	1040	1646	2219

experimental results still show a steady increase in performance.

However, since the coverage of case frames approaches to 1.0 as shown in Figure 6.1 and Figure 6.2, much improvement of case frame coverage cannot be expected. Thus, the effect of corpus size for discourse analysis is considered to be saturated with a few orders of magnitude larger corpus size.

6.5 Summary of this Chapter

This chapter has reported the effect of corpus size on case frame acquisition for syntactic and case structure analysis, and zero anaphora resolution in Japanese. Case frames were constructed from corpora of six different sizes ranging from 1.6 million to 1.6 billion sentences; these case frames were then applied to Japanese syntactic and case structure analysis, and zero anaphora

resolution. Experimental results showed the better results were obtained using case frames constructed from larger corpora, and the performance showed no saturation even when the corpus size was 1.6 billion sentences.

Chapter 7

Conclusion

What is represented in natural language text has originally a network structure, in which several mentions refer to the same entities, and several entities have tight relations with each other. However, due to the linear constraints of text, most of them are not obviously expressed in the normal form of text; thus automatic recognition of such relations is considered to be an essential step in natural language understanding. Anaphora resolution is the task to recognize anaphoric relations in text, which include anaphoric relations between coreferential mentions, zero anaphoric relations, and bridging relations. In this thesis, we focused on Japanese text; proposed an NE resolver using non-local information, and integrated model for anaphora resolution using case frames constructed from very large corpora. Our NE resolver achieved state-of-the-art performance; the integrated model for anaphora resolution, which is the first model to resolve anaphora resolution including coreference resolution, zero anaphora resolution and bridging reference resolution, achieved reasonable performance. We also reported that better anaphora resolution results were obtained by using case frames constructed from larger corpora; the performance was not saturated even with a corpus size of approximately 100 billion words.

7.1 Summary

Chapter 2 described how to acquire world knowledge automatically from very large corpora. We first acquired knowledge of synonyms, which was utilized for coreference resolution, from large raw corpus and dictionary definition sentences. As a result, we acquired 2,798 synonym pairs with an accuracy of 99%. Then, we constructed case frames from modifier-head examples in the resulting parses. We took a gradual approach that began to acquire basic case frames and gradually acquired richer ones by doing both case frame acquisition and text understanding one after another. Furthermore, in order to deal with data sparseness problem, we generalized

the examples of case slots. By using 1.6 billion Japanese sentences, we acquired about 1.6 million case frames for 65,679 unique predicates. Finally, we constructed nominal case frames. The point of our method was the integrated use of a dictionary and example phrases from large corpora. By using 1.6 billion Japanese sentences, we acquired about 566 thousand nominal case frames for about 564 thousand nouns.

Chapter 3 described the NE resolution (NER) system that used non-local information. While conventional Japanese NER system has been often performed immediately after morphological analysis and rely only on local context, our system performed after structural analyses, and used four types of non-local information: cache features, coreference relations, syntactic features and caseframe features, which were obtained from structural analyses. We evaluated our approach on CRL NE data and obtained an F-measure of 89.43%, which is higher than existing approaches that do not use structural information. We also conducted experiments on IREX NE data and an NE-annotated web corpus and confirmed that structural information improves the performance of NER.

Chapter 4 presented a knowledge-rich approach to Japanese coreference resolution. In Japanese, since pronouns are often omitted, called zero pronouns, proper noun coreference and common noun coreference occupy a central position in coreference relations. To improve coreference resolution for such language, wide-coverage knowledge of synonyms is useful. We first introduced knowledge of synonyms into coreference resolver. Furthermore, to boost the performance of coreference resolution, we integrated primitive bridging reference resolution system into coreference resolver. As a result of experiments on Japanese newspaper articles and web text, we confirmed that the use of automatically acquired synonyms and the result of bridging reference resolution boosted the performance of coreference resolution and the effectiveness of our integrated method.

Chapter 5 presented a probabilistic model for Japanese zero anaphora and bridging reference resolution. First, this model recognized discourse entities and linked all mentions to them. Zero pronouns were then detected by case structure analysis based on automatically constructed case frames. Their appropriate antecedents were selected from the entities with high salience scores, based on the case frames and several preferences on the relation between a zero pronoun and an antecedent. Case structure and zero anaphoric relations were simultaneously determined based on probabilistic evaluation metrics. As a result of experiments on Japanese web text, our system achieved an F-measure of 41.2% for zero anaphora resolution and an F-measure of 32.4% for bridging reference resolution.

Chapter 6 reported the effect of corpus size on case frame acquisition for discourse analysis. For this study, a Japanese corpus consisting of up to approximately 100 billion words was collected from the Web, and case frames were constructed from corpora of six different sizes: 1.6M, 6.3M, 25M, 100M, and 400M sentences, respectively. These case frames were then applied to syntactic and case structure analysis, and zero anaphora resolution. Better results were obtained by using case frames constructed from larger corpora; the performance was not saturated even with a corpus size of approximately 100 billion words.

7.2 Future Directions

This thesis described Japanese anaphora resolution model for Japanese text, which included named entity recognition, coreference resolution, bridging reference, and zero anaphora resolution. Although our named entity recognizer and coreference resolver achieved desirable performance, the performance of zero anaphora and bridging reference resolution was not satisfactory and there are still several problems:

Consideration of entities that are not written in the text obviously.

In some texts, such as essays and monologues, most of zero pronouns refer to the author of the sentences or speaker of the utterance, which are not written in the text obviously. However, since our model does not consider such entities, our model often outputs erroneous antecedents for such zero pronouns.

To consider the text type and detect such unwritten discourse entities would benefit the performance of anaphora resolution.

Consideration of semantic attributes for unknown words.

We generalized case frame examples and used semantic attribute of discourse entities by considering the categories of common nouns and the types of named entities. However, our model cannot consider semantic attributes of unknown words that do not refer to named entities.

(7.1) *Meiku-san-ga ○○○-san-no hada-ni "dôran"-wo nurikomi · · ·*
makeup artist Mr. ○○○ skin greasepaint rub
(The makeup artist rubbed greasepaint into Mr. ○○○'s skin.)

For example, both "*meiku-san*" and "∘∘∘-*san*" have semantic attribute "person." However, since "∘∘∘" is an unknown word and "*meiku*" usually means makeup itself, the system cannot recognize that they have semantic attribute "person." To recognize such semantic attribute is considered to be useful for anaphora resolution.

Macroscopic point of view for discourse structure.

By introducing salience score, our zero anaphora model considers global discourse to some extent. However, there is much more information concerning discourse structure, such as paragraph boundaries.

Recognition of such information and its application would benefit the performance of anaphora resolution.

Acquisition of knowledge about relations between predicates and its application.

Our model does not use the relations between predicates. However, the components of several predicates have tight relationships.

(7.2) *Tachiiri-ga kinshi-sareru-beki chiiki-daga hontouni kisei-sareta baai*. . . .
entry should be prohibited area indeed be restricted case
(Though the entry to the area should be prohibited, in case ϕ is restricted indeed . . . .)

For example, both the accusative cases of "*kinshi*" (prohibit) and that of "*kisei*" (restrict) in (7.2) are the same entity "*tachiiri*" (entry), and the accusative case of "*kisei*" (restrict) is considered to be easily resolved if there is knowledge that the accusative case of "*kisei*" (restrict) is often identical with the accusative case of "*kinshi*." To acquire such knowledge about relations between predicates and apply the knowledge to zero anaphora resolution system is our future work.

Expansion of case types for zero anaphora resolution.

In proposed model, we concentrated upon three case slots for zero anaphora resolution: "*ga* (nominative)," "*wo* (accusative)" and "*ni* (dative)." Since these cases cover about 90% of zero anaphora, this restriction is reasonable for primitive stage of anaphora resolution

system construction. However, there is another important case, *second nominative case*, which denotes topic of the sentence.

(7.3) *Zou-wa hana-ga nagai.*
elephant-**TM** trunk-non long
([(literally)] As for elephant, trunk is long. = Elephant's trunk is long.)

(7.3) is a double nominative sentence, and "*zou*" in (7.3) is the second nominative case. The second nominative cases also often omitted and to resolve such omissions is also our future work.

Bibliography

[1] Takehito Utsuro, Manabu Sassano, and Kiyotaka Uchimoto. Combing outputs of multiple named entity chunkers by stacking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.

[2] Hideki Isozaki and Hideto Kazawa. Speeding up support vector machines for named entity recognition (in Japanese). *Trans. of Information Processing Society of Japan*, 44(3):970–979, 2003.

[3] Masayuki Asahara and Yuji Matsumoto. Japanese named entity extraction with redundant morphological analysis. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 8–15, 2003.

[4] Kenta Fukuoka. Named entity extraction with semi-Markov conditional random fields (in Japanese). Master's thesis, Nara Institute of Science and Technology, 2006.

[5] Hiroyasu Yamada. Shift reduce chunking for Japanese named entity extraction (in Japanese). In *IPSJ SIG Notes NL-179-3*, pages 13–18, 2007.

[6] Masaki Murata and Makoto Nagao. An estimate of referent of noun phrases in Japanese sentences. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 912–916, 1998.

[7] Ryu Iida, Kentaro Inui, Yuji Matsumoto, and Satoshi Sekine. Noun phrase coreference resolution in Japanese most likely candidate antecedents (in Japanese). *Journal of Information Processing Society of Japan*, 46(3):831–844, 2005.

[8] Hideki Isozaki and Tsutomu Hirao. Japanese zero pronoun resolution based on ranking rules and machine learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 184–191, 2003.

[9] Daisuke Kawahara and Sadao Kurohashi. Zero pronoun resolution based on automatically constructed case frames and structural preference of antecedents. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP-04)*, pages 334–341, 2004.

[10] Ryu Iida, Kentaro Inui, and Yuji 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, 2006.

[11] Daisuke Kawahara and Sadao 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, 2006.

[12] Sadao Kurohashi and Makoto Nagao. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4):507–534, 1994.

[13] Nobuhiro Kaji Daisuke Kawahara and Sadao Kurohashi. Japanese case structure analysis by unsupervised construction of a case frame dictionary. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING2000)*, pages 432–438, 2000.

[14] Naoaki Okazaki and Mitsuru Ishizuka. A discriminative approach to Japanese abbreviation extraction. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*, pages 889–894, 2008.

[15] Takenobu Tokunaga, Yasuhiro Syotu, Hozumi Tanaka, and Kiyoaki Shirai. Integration of heterogeneous language resources: A monolingual dictionary and a thesaurus. In *the 6th Natural Language Processing Pacific Rim Symposium*, pages 135–142, 2001.

[16] Eric Nichols, Francis Bond, and Daniel Flickinger. Robust ontology acquisition from machine-readable dictionaries. In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-2005*, pages 1111–1116, 2005.

[17] Jun-ichi Tajika, editor. *Reikai Syogaku Kokugojiten*. Sanseido, 1997.

[18] Minoru Nishio, Etsutaro Iwabuchi, and Mizutani Shizuo, editors. *Iwanami kokugo jiten (Iwanami Japanese Dictionary)*. Iwanami Shoten, 2000.

[19] Information-Technology Promotion Agency, Japan. Japanese verbs : A guide to the IPA lexicon of basic Japanese verbs, 1987.

[20] Japan Electronic Dictionary Research Institute, Ltd. EDR electronic dictionary technical guide, 1995. (In Japanese).

[21] Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentarou Ogura, Yoshifumi Oyama, and Yoshihiko Hayashi. Japanese lexicon, 1997.

[22] Daisuke Kawahara and Sadao Kurohashi. Case frame compilation from the web using high-performance computing. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1344–1347, 2006.

[23] The National Language Institute for Japanese Language. *Bunruigoihyo*. Dainippon Tosho, (In Japanese), 2004.

[24] Daisuke Kawahara and Sadao Kurohashi. Fertilization of case frame dictionary for robust Japanese case analysis. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 425–431, 2002.

[25] Ted Briscoe and John Carroll. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 356–363, 1997.

[26] Udo Hahn, Michael Strube, and Katja Markert. Bridging textual ellipses. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 496–501, 1996.

[27] Renata Vieira and Massimo Poesio. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–592, 2000.

[28] Michael Strube and Udo Hahn. Functional centering – grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344, 1999.

[29] Massimo Poesio, Tomonori Ishikawa, Sabine Schulte im Walde, and Renata Vieira. Acquiring lexical knowledge for anaphora resolution. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1220–1224, 2002.

[30] Masaki Murata, Hitoshi Isahara, and Makoto Nagao. Resolution of indirect anaphora in Japanese sentences using examples "X no Y"(Y of X). In *Proceedings of ACL'99 Workshop on 'Coreference and Its Applications'*, 1999.

[31] Sadao Kurohashi and Yasuyuki Sakai. Semantic analysis of Japanese noun phrases: A new approach to dictionary-based understanding. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 481–488, 1999.

[32] Lisa Ferro Nancy Chinchor, Erica Brown and Patty Robinson. Named entity recognition task definition, 1999.

[33] IREX Committee, editor. *Proceedings of the IREX Workshop*, 1999.

[34] MUC-6, editor. *Proceedings of the Sixth Message Understanding Conference*. Morgan Kaufmann Publishers, INC., 1995.

[35] Keigo Nakano and Yuzo Hirai. Japanese named entity extraction with bunsetsu features (in Japanese). *Trans. of Information Processing Society of Japan*, 45(3):934–941, 2004.

[36] R. Malouf. Markov models for language-independent named entity recognition. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*, pages 187–190, 2002.

[37] Hai Leong Chieu and Hwee Tou Ng. Named entity recognition: A maximum entropy approach using global information. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7, 2002.

[38] Heng Ji and Ralph Grishman. Improving name tagging by reference resolution and relation detection. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 411–418, 2005.

[39] Behrang Mohit and Rebecca Hwa. Syntax-based semi-supervised named entity tagging. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 57–60, 2005.

[40] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, 2005.

[41] Vijay Krishnan and Christopher D. Manning. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1121–1128, 2006.

[42] Taku Kudo and Yuji Matsumoto. Chunking with support vector machines. In *Proceedings of the second meeting of North American chapter of association for computational linguistics (NAACL)*, pages 192–199, 2001.

[43] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, and Masayuki Asahara. Morphological analysis system ChaSen 2.3.3 users manual, 2003.

[44] Sadao Kurohashi and Daisuke Kawahara. Japanese morphological analysis system JUMAN version 6.0 manual, 2007.

[45] Masayuki Asahara and Yuji Matsumoto. *IPADIC User Manual*. Nara Institute of Science and Technology, Japan, 2002.

[46] John C. Platt, Nello Cristiani, and John ShaweTaylor. Large margin DAGs for multiclass classification. In *Advances in Neural Information Processing System 12*, 2000.

[47] The National Language Institute of Japanese Language, NLRI, editor. *Bunrui Goi Hyo (in Japanese)*. Shuuei Publishing, 1993.

[48] Jing Jiang and ChengXang Zhai. Exploiting domain structure for named entity recognition. In *Proceedings of the Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting*, 2006.

[49] Jun'ichi Kazama and Kentaro Torisawa. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 08)*, pages 407–415, 2008.

[50] Ken'ichi Fukushima, Nobuhiro Kaji, and Masaru Kitsuregawa. Use of massive amounts of web text in Japanese named entity recognition. In *Proceedings of Data Engineering Workshop (DEWS2008)*, pages A3–3, 2008.

[51] Xiaofeng Yang, Guodong Zhou, Jiau Su, and Chew Lim Tan. Improving noun phrase coreference resolution by matching strings. In *Proceedings of 1st International Joint Conference of Natural Language Processing (IJCNLP04)*, pages 326–333, 2004.

[52] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.

[53] Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, pages 135–142, 2004.

[54] Vincent Ng. Machine learning for coreference resolution: From local classification to global ranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 157–164, 2005.

[55] Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. Construction of a Japanese relevance-tagged corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 2008–2013, 2002.

[56] Zhou Guodong and Su Jian. A high-performance coreference resolution system using a constraint-based multi-agent strategy. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 522–528, 2004.

[57] Vincent Ng and Claire Cardie. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 730–736, 2002.

[58] Pascal Denis and Jason Baldridge. Joint determination of anaphoricity and coreference resolution using integer programming. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 236–243, 2007.

[59] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, 2002.

[60] Aron Culotta, Michael Wick, and Andrew McCallum. First-order probabilistic models for coreference resolution. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 81–88, 2007.

[61] Xiaofeng Yang, Jian Su, Jun Lang, Ghew Lim Tan, Ting Liu, and Sheng Li. An entity-mention model for coreference resolution with inductive logic programming. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 08)*, pages 843–851, 2008.

[62] Aria Haghighi and Dan Klein. Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 848–855, 2007.

[63] Hoifung Poon and Pedro Domingos. Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 650–659, 2008.

[64] Ruslan Mitkov, Richard Evans, and Constantin Orăsan. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, 2002.

[65] Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 911–917, 2002.

[66] Xiaoqiang Luo. Coreference or not: A twin model for coreference resolution. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 73–80, 2007.

[67] Shalom Lappin and Herbert J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–562, 1994.

[68] Ryu Iida, Kentaro Inui, and Yuji Matsumoto. Zero-anaphora resolution from the perspective of cohesion and coherence (in Japanese). In *IPSJ SIG Notes NL-178-7*, pages 45–52, 2008.

[69] Martin Volk. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proceedings of the Corpus Linguistics Conference*, pages 601–606, 2001.

[70] Natalia N. Modjeska, Katja Markert, and Malvina Nissim. Using the web in machine learning for other-anaphora resolution. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 176–183, 2003.

[71] Mirella Lapata and Frank Keller. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*, 2:1:1–31, 2005.

[72] Michaela Atterer and Hinrich Schütze. The effect of corpus size in combining supervised and unsupervised training for disambiguation. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 25–32, 2006.

[73] Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, 2007.

[74] Adam Kilgarriff and Gregory Grefenstette. Introduction to the special issue on the web as corpus. *Computational Linguistic*, 29(3):333–347, 2003.

[75] Preslav Nakov and Marti A. Hearst. Solving relational similarity problems using the web as a corpus. In *Proceedings of ACL-08: HLT*, pages 452–460, 2008.

[76] Michele Banko and Eric Brill. Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing. In *Proceedings of the Conference on Human Language Technology*, 2001.

[77] Jun Suzuki and Hideki Isozaki. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT 08)*, pages 665–673, 2008.

[78] Preslav Nakov and Marti Hearst. A study of using search engine page hits as a proxy for n-gram frequencies. In *Proceedings of RANLP'05*, 2005.

[79] Ann Gledson and John Keane. Using web-search results to measure word-group similarity. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 281–288, 2008.

[80] Michele Banko and Eric Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, 2001.

[81] Vinci Liu and James R. Curran. Web text corpus for natural language processing. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 233–240, 2006.

[82] Peter Halacsy, Andras Kornai, Laszlo Nemeth, Andras Rung, Istvan Szakadat, and Vikto Tron. Creating open language resources for Hungarian. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 203–210, 2004.

[83] David Graff. English Gigaword. Technical Report LDC2003T05, Linguistic Data Consortium, Philadelphia, PA USA., 2003.

[84] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, 2006.

[85] Taku Kudo and Hideto Kazawa. Web Japanese N-gram version 1, published by Gengo Shigen Kyokai, 2007.

[86] Kikuo Maekawa. Kotonoha, the corpus development project of the National Institute for Japanese language. In *Proceedings of the 13th NIJL International Symposium*, pages 55–62, 2006.

[87] Daisuke Kawahara, Ryohei Sasano, and Sadao Kurohashi. Toward text understanding: Integrating relevance-tagged corpora and automatically constructed case frames. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1833–1836, 2004.

List of Publications by the Author

Major Publications

[1] Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. Automatic construction of nominal case frames and its application to indirect anaphora resolution. In *Proceedings of the 20th International Conference on Computational Linguistics (Coling 2004)*, pages 1201–1207, 2004.

[2] Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. Automatic construction of nominal case frames and its application to indirect anaphora resolution (in Japanese). *Journal of Natural Language Processing*, 12(3):129–144, 2005.

[3] Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. Improving coreference resolution using bridging reference resolution and automatically acquired synonyms. In *Anaphora: Analysis, Algorithms and Applications, 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2007)*, pages 125-136, 2007.

[4] Ryohei Sasano and Sadao Kurohashi. Japanese named entity recognition using structural natural language processing. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*, pages 607–612, 2008.

[5] Ryohei Sasano, Daisuke Kawahara, and Sadao 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, 2008.

[6] Ryohei Sasano and Sadao Kurohashi. Improving coreference resolution using automatically acquired knowledge of nominal relations (in Japanese). *Journal of Natural Language Processing*, 15(5):99–118, 2008.

[7] Ryohei Sasano and Sadao Kurohashi. Japanese named entity recognition using non-local

information (in Japanese). *Journal of Information Processing Society of Japan*, 49(11):3765–3776, 2008.

[8] Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. The effect of corpus size on case frame acquisition for discourse analysis. In *Proceedings of the Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting (NAACL/HLT 2009)*, (to appear).

Other Publications

[1] Daisuke Kawahara, Ryohei Sasano, and Sadao Kurohashi. Text understanding for conversational agent. In *Proceedings of the 1st International Workshop on Intelligent Media Technology for Communicative Intelligence (IMTfCI2004)*, pages 68–71, 2004.

[2] Daisuke Kawahara, Ryohei Sasano, and Sadao Kurohashi. Toward text understanding: Integrating relevance-tagged corpora and automatically constructed case frames. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1833–1836, 2004.

[3] Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. Automatic construction of nominal case frames and its application to indirect anaphora resolution (in Japanese). In *Proceedings of the 10th Annual Meeting of the Association for Natural Language Processing*, pages 472–475, 2004.

[4] Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. Integrated anaphora resolution based on automatically acquired knowledge (in Japanese). In *Proceedings of the 11th Annual Meeting of the Association for Natural Language Processing*, pages 480–483, 2006.

[5] Ryohei Sasano and Sadao Kurohashi. Automatic recognition of rendaku and onomatopoeia in morphological analysis (in Japanese). In *Proceedings of the 12th Annual Meeting of the Association for Natural Language Processing*, pages 819–822, 2007.

[6] Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. Improving the coverage of case frames using larger corpus and generalized examples (in Japanese). In *Proceedings of the 13th Annual Meeting of the Association for Natural Language Processing*, pages 528–531, 2008.

[7] Kei Hamada, Ryohei Sasano, Tomohide Shibata, Daisuke Kawahara, and Sadao Kurohashi. Case frame construction based on distributional similarity (in Japanese). In *Proceedings of the 13th Annual Meeting of the Association for Natural Language Processing*, pages 532–535, 2008.