NAIST-IS-DD0461004

Doctoral Dissertation

Combining Linguistic Knowledge and Machine Learning for Anaphora Resolution

Ryu Iida

March 23, 2007

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Ryu Iida

Thesis Committee:

Professor Yuji Matsumoto	(Supervisor)
Professor Kiyohiro Shikano	(Co-supervisor)
Professor Shin Ishii	(Co-supervisor)
Associate Professor Kentaro Inui	(Co-supervisor)

Combining Linguistic Knowledge and Machine Learning for Anaphora Resolution*

Ryu Iida

Abstract

This thesis focuses on attempting to incorporate linguistically contextual clues into machine learning-based approaches for *anaphora resolution*, which is the process of identifying whether or not an expression refers to another expression. As the state-of-the-art of morpho-syntactic analysis and named entity recognition grows more increasingly sophisticated, research focus in natural language processing (NLP) has shifted to more semantically motivated tasks such as anaphora resolution. These tasks are particularly important as they often provide a critical bridge between basic NLP techniques and end-level applications.

Conventional approaches to anaphora resolution have been roughly evolving in two different but complementary directions. One is theory-oriented rule-based approaches and the other is empirical corpus-based approaches. In rule-based approaches, efforts have been directed toward the manual encoding of various linguistic cues into a set of rules, however it is extremely difficult to manually encode linguistic findings into rules while considering widely ranging aspects from lexical to discourse factors. In contrast, empirical corpus-based approaches have been mainly developed with shallow morpho-syntactic information such as part-of-speech and gender/number information as features, while having achieved a performance comparable to the best-performing rule-based system. Given this background, this thesis deals with effectively combining machine learning and linguistic knowledge mainly used in discourse theory for anaphora resolution.

First of all, we discuss how to annotate predicate-argument relations including zero-anaphoric relations and coreference relations. In order to develop a trainable

*Doctoral Dissertation, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0461004, February 23, 2007.

model for anaphora or coreference resolution, a large-scale corpus annotated with information about predicate-argument structures and coreference is needed. To our best knowledge, however, there is no large-scale corpus including such tags in Japanese. So, we develop new criteria for our annotation processes by examining previous work on annotationg tasks. Chapter 3 explains our annotating specification cultivated through actual annotating processes for the texts in Kyoto Text Corpus version 3.0 and discusses the future directions.

In Chapter 4 we describe how in the Centering Theory the preference of antecedents between candidate antecedents is generally formalized by comparison between candidates. In that spirit, we propose the *tournament model*, a machine learning-based model that can directly compare two candidates in series of matches. This new model dramatically outperforms conventional pairwise classification models in experiments on Japanese zero-anaphora resolution.

Secondly, in Chapter 5 we proposed the selection-then-classification model, a process that reverses the order of the steps in the classification-then-search model proposed by Ng and Cardie (2002b), inheriting all the advantages of that model. We conducted experiments on resolving noun phrase anaphora in Japanese. The results show that with the selection-then-classification based modifications, our model outperforms earlier learning-based approaches.

Finally, we approach the zero-anaphora resolution problem by decomposing it into intra-sentential and inter-sentential zero-anaphora resolution. For the former problem, syntactic patterns of the appearance of zero-pronouns and their antecedents are useful clues. Taking Japanese as a target language, we empirically demonstrate that incorporating rich syntactic pattern features in a state-of-the-art learning-based anaphora resolution model dramatically improves the accuracy of intra-sentential zero-anaphora, which consequently improves the overall performance of zero-anaphora resolution.

Keywords:

anaphora, zero-pronoun, corpus, machine learning, Centering Theory

照応解析のための言語学的知識と機械学習手法の融合*

飯田 龍

内容梗概

本研究では機械学習に基づく照応解析の処理に解析に役立つ文脈的な手がかりを導入する試みについて報告する．照応解析とは，文章内に出現する表現のうち，一方の表現が他方の表現を指す関係を自動的に同定する処理をいう．形態素解析や構文解析，固有表現抽出などの自然言語処理の基盤技術が成熟し，情報抽出などの応用処理が表層レベルから一歩意味に踏み込んだ結果を求める現状において，基盤技術と応用処理の中間に存在する照応解析の処理を実用的なレベルまで向上させることは現在の自然言語処理の枠組みにおいて非常に重要であると考えられる．

従来の照応解析の処理は，人手で作成された規則に基づく解析手法と照応関係がタグ付与されたコーパスを利用した機械学習に基づく手法に分けて考えることができ，それぞれ相補的に研究が進められてきた．規則に基づく手法では，センタリング理論など談話研究に基づいた手がかりを人手で規則に導入する試みがなされており，一方，機械学習に基づく手法では，品詞や文字列の情報など主に表層的な手がかりを学習に利用する素性に導入し規則ベースの手法と同程度の解析精度を得ている．本研究では規則ベースの手法で主に導入されてきた言語学的な知見を機械学習ベースの手法で効果的に利用できるようモデル化する．

まず最初に，今回解析対象とする照応関係の問題設定について議論する．先行研究の照応解析の問題は共参照関係との関係でさまざまな問題設定がなされている．また，日本語を対象とした照応解析についても厳密な問題設定の議論がないまま，いくつもの解析手法が提案されている．そこで，3章で日本語を対象とした照応関係・共参照関係の仕様を提案し，実際に作業者にタグ付与作業を行ってもらい，分析/評価/学習用のタグ付与コーパスを作成する．また，タグ付与作業の際に起こった問題点についても議論する．

*奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 博士論文, NAIST-IS-DD0461004, 2007年2月23日.

次に，4章で，先行詞の序列を理論的に説明したセンタリング理論の考えを採用し，この序列を学習ベースのモデルで利用する二つの方法について述べる．一つは，先行詞候補が前文脈から得られた*forward-looking center*（先行詞候補集合を先行詞らしさの序列に基づき並べたリスト）のどの箇所に含まれているかを示す素性（**センタリング素性**）を学習手法で直接利用するやり方である．もう一つは，先行詞候補全体の間の序列を二つの候補の間の関係に分解して考え，二つの候補の間の選好を学習し，候補間で勝ち抜き戦を行うことにより，最終的に最尤の先行詞を決定することで先行詞同定を行うモデル（**トーナメントモデル**）を提案した．日本語ゼロ代名詞の先行詞同定の評価実験を通じて，特に後者が既存手法に比べ大幅に精度向上したことを報告する．

5章では，与えられた照応詞の候補が真に照応詞となるか否かを判定する照応性判定の問題を解く手法を説明する．この研究では，既存の手法で個別に利用されてきた先行文脈の情報と照応詞の局所的な文脈情報を効果的に組み合せることがどのていど精度に影響を及ぼすかを調査する．日本語名詞句照応解析を対象に，二つの情報を組み合わせることによって先行詞同定と照応性判定のそれぞれにおいて精度向上に貢献することを示す．

最後に，6章で，文内ゼロ照応に統語的なパタンを素性として導入する一手法を提案する．文内ゼロ照応では，ゼロ代名詞と先行詞がどのような構造的位置関係で出現しているかが解析のための大きな手がかりとなる．そこで，ゼロ代名詞と先行詞の間の統語的なパタンから有効な素性をマイニングし，ゼロ照応解析に利用する手法を提案する．

キーワード

照応, ゼロ代名詞, コーパス, 機械学習, センタリング理論

Acknowledgements

指導教官の松本裕治教授には，研究を行う上での真摯な態度を日頃から見せていただき，また私が取り組んだ各研究テーマについて非常に的確なコメントをいただきました．3章のトーナメントモデルや4章の探索先行分類型モデルの基本的なアイデアは松本先生がふとした瞬間にもらしたコメントやメールでの発言に基づいています．本当にありがとうございました．石井信教授には学内で発表を聞いていただく度に研究室のスタッフからは絶対出ないであろう興味深い指摘をいただきました．ありがとうございます．乾健太郎助教授には，自然言語処理という問題に取り組むことへのおもしろさを教えていただきました．学部3年のころから面倒を見ていただき，今この奈良先端大にいて研究を続けていられるもの一重に九州工業大学で乾さんに出会えたからだと思います．これからもよろしくお願いします．研究室の研究会ではいつも意義のあるコメントをしていただいた新保仁助手，筑波大学のEdson T. Miyamoto講師，東京工業大学の高村大也助手に心より感謝いたします．京都大学の高梨克也さん，大阪府立大学の橋本喜代太さん，通信総合研究所の黒田航さんには，工学の立場にいただけでは気付くことができない照応・共参照の問題の複雑さやその問題に関わる分野へのつながりを教えていただきました．ありがとうございます．ニューヨーク大学で好き勝手に研究をやらせてくれた関根聡研究助教授に感謝します．また，各研究テーマの評価実験では，すべて工藤拓さんの作成されたツールを利用しています．ありがとうございます．乾研のころから研究をやることへの意欲を示していただいた高橋哲朗さん，藤田篤さんに感謝します．同期の伊藤敬彦さん，小林のぞみさん，山崎貴宏さん，徳永泰浩さん，奥村研究室の藤木稔明さんに感謝します．英語の細かいチェックをしてくれたEric Nicholsさん，日々の生活でさまざまな問題への議論をしてくれた小町守さんに感謝します．さらに，研究活動や学校生活を暖かく支えて下さいました大川基子秘書，コーパスへのタグ付与作業を行い，その経験を通じて議論していただいた技術補佐員の津崎真由美さんと中川由樹子さんに感謝いたします．今回取り組んだ研究に対して有意義なコメントをしていただいたすべての方へ感謝します．最後に，5年間奈良での学校生活を精神面で支えてくれた野口睦子さんに心から感謝します．

Contents

List of Figures

List of Tables

Abbreviations

ABL	Ablative case
ACC	Accusative case
ADNOM	Adnominal
Cb	Backward-looking center
Cf	Forward-looking center
CONJ	Conjunctive
Cp	Preferred center
DAT	Dative case
D_OBJ	Direct Object
I_OBJ	Indirect Object
NOM	Nominative case
NP	Noun phrase
PUNC	Punctuation
SUBJ	Subject case
TOP	Topic
U	Utterance

Chapter 1

Introduction

In recent years, information retrieval techniques for widespread web data have been polished and refined. As the reader is likely aware, the search engines such as Google, for instance, provide a highly ranked list of web pages as the results of search queries by using algorithms such as $tf \cdot idf$ or *PageRank* (Page et al., 1998). This brings significant benefit to users seeking to effectively retrieve information from large amounts of web data. However, if the retrieved pages reach hundreds of thousands in numbers, it is difficult for the users to manage them. Given this background, one of the key issues is the task of *information extraction*, where the goal is effectively aggregating the information that the users want to find in large data. Techniques of information extraction have been attracting attention since the early 1990's (see Pazienza (1997)), especially as in the task definitions given by conferences such as the Message Understanding Conference (MUC)[1] or Automatic Content Extraction (ACE)[2].

The problems in information extraction can be divided into two tasks: detecting an entity or detecting relations among entities such as anaphora resolution. In the former task, one of the most famous subtasks is named entity recognition; the extraction of names from a given text. State-of-the-art methods of named entity extraction have reached a level where it is practical to apply them to other NLP applications. Thus, researchers are now focusing on this latter task of anaphora resolution and semantic role labeling.

[1] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

[2] http://www.ldc.upenn.edu/Projects/ACE/

1.1. Anaphora Resolution

Anaphora is a linguistic phenomenon that an expression points back to another expression in the preceding context. The word or phrase pointing back is called an *anaphor* and the expression which is referred to by an anaphor is its *antecedent*. The process of identifying anaphoric relations is called *anaphora resolution*. For example, for the text given in (1), we must identify that the NP *US Air*$_i$ in the second sentence has an anaphoric relation with the NP *US Group Inc.*$_i$, whereas *share*$_j$ is judged as non-anaphoric because it has no antecedent in the preceding context.

(1) a. A federal judge in Pittsburgh issued a temporary restraining order preventing Trans World Airlines from buying additional shares of *USAir Group Inc.*$_i$

b. The order, requested in a suit filed by *USAir*$_i$, dealt another blow to TWA's bid to buy the company for $52 a *share*$_j$.

Anaphora resolution is an important process for various NLP applications. Striving for the realization of a practical solution, many researchers have worked on it from a variety of perspectives. As an example of their works, the rule-based approaches from theory-oriented perspective (Hobbs, 1978; Kameyama, 1986; Lappin and Leass, 1994; Baldwin, 1995; Okumura and Tamura, 1996; Mitkov, 1997; Walker et al., 1997, etc.) have been attempted for pronominal anaphora resolution (for more details see Mitkov (2002)). Baldwin (1995), for instance, has previously reported that his rule-based system achieves round 73% precision with about 75% recall for all pronouns in MUC-6 Coreference task, applying eight sophisticated rules containing syntactic and semantic information in the linguistic theories. The performance of the system is appealing, however it is extremely difficult to manually encode linguistic findings into rules while considering widely ranging aspects from lexical to discourse factors.

In contrast, empirical corpus-based approaches which apply machine learning algorithms, such as decision tree and Support Vector Machines to anaphora resolution have been attracting attention since the end of the 1990's (McCarthy and Lehnert, 1995; Ge et al., 1998; Soon et al., 2001; Ng and Cardie, 2002a; Yang et al., 2003; Iida et al., 2005, etc.). The approaches have been mainly developed with shallow morpho-syntactic information such as part-of-speech and gender/number information as features, while having achieved a performance comparable to the best-performing rule-based system for the coreference task test sets of MUC-6 and MUC-7. Combining linguistic findings and machine learning-based approaches has a potential to be an effective solution, however combining these advantages is not a trivial problem because most findings are abstract and hard to encode into computationally-accessible forms and even where it is possible, it is not clear that they are effective in anaphora resolution.

1.2. Aims

Given this background, this thesis focuses on developing anaphora resolution models by incorporating linguistic findings into machine learning-based approaches if the findings have beneficial effects on the resolution. As reported in Ng and Cardie (2002a), they introduce 53 features into a machine learning-based model, which is ineffective for the test set in MUC-6 and MUC-7. Ng and Cardie end up manually selecting a subset of 22-26 features that achieved the best-performing result. As can be seen from the work by Ng and Cardie (2002a), the previous learning-based models do not always effectively exploit these findings as features. So, one of the challenging issues we should explore next is investigating how to design a learning-based models that reflect the beneficial linguistic findings. Specifically, the thesis examines the following three topics.

- We propose an antecedent identification model that can capture the antecedenthood between candidate antecedents implicitly dealt with in Centering Theory (Grosz et al., 1995). This model, called the *tournament model*, conducts a tournament consisting of a series of matches in which candidates compete with each other and the that prevails through the final round is judged as an antecedent.

- As well as antecedent identification, *anaphoricity determination*, which is the task of judging whether an candidate anaphor is anaphoric or non-anaphoric, is another important problem. For this task, we propose the *selection-then-classification model* that identifies an antecedent followed by determining anaphoricity to inherit the advantages of the previous models such as Ng and Cardie (2002a) and our tournament model.

- We approach the zero-anaphora resolution problem by decomposing it into intra-sentential and inter-sentential zero-anaphora resolution. For the former problem, we use syntactic pattern features since syntactic patterns of the appearance of zero-pronouns and their antecedents are useful clues.

1.3. Contributions

The contributions of this study are as follows.

- In antecedent identification tasks in Japanese, the performance of the tournament model outperforms that estimates the absolute likelihood of each candidate independently of other candidates. It indicates that comparing between two candidate antecedent is more efficient for identifying antecedent than the candidate-wise models such as the work by Soon et al. (2001) and Ng and Cardie (2002a).
- The selection-then-classification approach improves the performance of the previous learning-based models by combining their advantages, while overcoming their drawbacks. Taking the task of NP anaphora resolution in Japanese, we demonstrate that even if the parameters for their models are optimally turned, the proposed model significantly outperforms them when it employs the tournament model for antecedent identification.
- The result of intra-sentential zero-anaphora resolution shows that the selection-then-classification model with syntactic pattern features is significant better than the original one, which consequently improves the overall performance of zero-anaphora resolution.

1.4. Thesis outline

This thesis is organized as follows. Chapter 2 describes two kinds of previous work: theory-oriented rule-based approaches and empirical corpus-based approaches to anaphora resolution. Chapter 3 discusses how to define the problems of predicate-argument analysis including zero-anaphora resolution and coreference resolution in Japanese written text. In Chapter 4, we present a method that incorporates contextual cues motivated by Centering Theory (Grosz et al., 1995) into a machine learning-based model for identifying antecedents in zero-anaphora resolution task. In Chapter 5, we present a machine learning-based approach to noun phrase anaphora resolution that combines the advantages of previous learning-based models while overcoming their drawbacks. We explain our model that uses syntactic patterns as features for intra-sentential zero-anaphora resolution in Chapter 6 and finally conclude this thesis with some remarks in Chapter 7.

Chapter 2

Previous Work on Anaphora Resolution

Computational approaches to anaphora resolution have been roughly evolving in two different but complementary directions. One is theory-oriented rule-based approaches, and the other is empirical corpus-based approaches. This chapter briefly reviews each approach and discusses the advantages and drawbacks of each.

2.1. Rule-based approaches

In rule-based approaches (Mitkov, 1997; Baldwin, 1995; Nakaiwa and Shirai, 1996; Okumura and Tamura, 1996, etc.), efforts have been directed to the manual encoding of various linguistic cues into a set of rules. Such cues include, for example, the syntactic role of each target noun phrase, the appearance order of antecedent candidates, and the semantic compatibility between an anaphor and a candidate. Most rule-based approaches are also influenced, to a greater or less extent, by theoretical linguistic work, such as Centering Theory (Grosz et al., 1995; Walker et al., 1994; Kameyama, 1986) and the Systemic Theory (Halliday and Hasan, 1976). The best-achieved performance in MUC-7 [1] was around 70% precision with 60% recall, which is still far from being satisfactory for practical application in many tasks. Worse still, a rule set

[1] The Seventh Message Understanding Conference (1998): http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

Table 2.1. Centering transition states

	$Cb(U_i) = Cb(U_{i-1})$ or $Cb(U_i) = [?]$	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i) = Cp(U_i)$	CONTINUE	SMOOTH-SHIFT
$Cb(U_i) \neq Cp(U_i)$	RETAIN	ROUGH-SHIFT

tuned for a particular domain is unlikely to work equally as well for another domain due to domain-dependent properties of coreference patterns. Given these facts, further manual refinements of rule-based models will be prohibitively costly.

2.1.1 Centering-based approaches

Centering Theory (Grosz et al., 1995) is a theory about discourse coherence and is based on the idea that each utterance features a topically most salient entity called the *center*. The main idea of Centering Theory is that certain entities mentioned in an utterance are more central in discourse than others and this imposes certain constraints on the use of referring expressions and in particular on the use of pronouns. The centering model is very simple. Discourses consist of constituent segments and each segment is represented as part of a discourse model. Centers are semantic entities that are part of the discourse model for each utterance in a discourse segment. The set of *forward-looking centers*, $Cf(U_i)$, represents discourse entities evoked by an utterance U_i in a discourse segment. The *backward-looking center* is a special member of the *Cf*, which represents the discourse entity that the utterance U_i most centrally concerns, similar to what is elsewhere called the 'topic.' The *Cb* entity links the current utterance to the previous discourse.

The set of forward-looking centers is ranked according to discourse salience. This ranking is a partial order according to their discourse salience[2]. The highest-ranked element in *Cf* is called the *preferred center*, *Cp*. The preferred center represents a prediction about the *Cb* of the following utterance. We can classify relation between *Cb* and *Cp* into four types of transition relations across pairs of utterances (see Table 2.1).

[2] As an example of *Cf* ranking, Brennan et al. (1987) rank the items in *Cf* by obliqueness of grammatical relation of the subcategorized functions of the main verb: that is, first the subject, object, and object2, followed by other subcategorized functions.

In addition to the structures for centers, *Cb* and *Cf*, the original Centering Theory includes two rules and three constraints.

> For each utterance U_i in a discourse segment D consisting of utterances $U_1, \ldots, U_m$:
> **Constraint 1** There is precisely one backward-looking center $Cb(U_i)$.
> **Constraint 2** Every element of the forward center list, $Cf(U_i)$, must be realized in U_i.
> **Constraint 3** The center, $Cb(U_i)$, is the highest-ranked element of $Cf(U_{i-1})$ that is realized in U_i.
>
> **Rule 1** If some element of $Cf(U_i)$ is realized as a pronoun in U_{i+1}, then $Cb(U_{i+1})$ must also be realized as a pronoun.
> **Rule 2** Transition states are ordered. The *continue* transition is preferred to the *retain* transition, which is preferred to the *smooth-shift* transition, which is preferred to the *rough-shift* transition.

To exemplify the theory, here are two very simple discourses differing in the second sentence from Discourse (2) and (3):

(2) a. John$_i$ went to his favorite music store$_j$ to buy a piano.

b. He$_i$ had frequented the store$_j$ for many years.

c. He$_i$ was excited to be going to the store$_j$ to actually buy a piano.

(3) a. John$_i$ went to his favorite music store$_j$ to buy a piano.

b. It$_j$ was a store John$_i$ had frequented for many years.

c. He$_i$ was excited to be going to the store$_j$ to actually buy a piano.

The backward-looking center of (2)b and (2)c and the forward-looking center of (2)a, (2)b and (2)c are listed as in Table 2.2. In discourse (2), according to the constraints, sentence (2)c exhibits CONTINUE transition. In discourse (3), on the other hand, sentence (3)b is interpreted as RETAIN transition and sentence (3)c ROUGH-SHIFT based

Table 2.2. Center transition in the text (2)

	Cb	*Cf*	transition
a.	-	[John, store, piano]	
b.	John	[John, store]	CONTINUE
c.	John	[John, store, piano]	CONTINUE

Table 2.3. Center transition in the text (3)

	Cb	*Cf*	transition
a.	-	[John, store, piano]	
b.	John	[store, store]	RETAIN
c.	store	[John, store, piano]	ROUGH-SHIFT

on Table 2.3. Rule 2 provides an underlying principle for coherence of discourse. Frequent shifts detract from local coherence, whereas continues contribute to coherence. According to Rule 2, centering accounts for the coherence: discourse (2) is more coherent than discourse (3).

Focusing on the Centering Theory, the several researchers have proposed variants of anaphora resolution models (Kameyama, 1986; Brennan et al., 1987; Walker et al., 1994; Poesio et al., 2000; Tetreault, 2001, etc.), however they have the limitations (see also Kehler (1997)). For instance, the original centering model only accounts for local coherence of discourse. In anaphora resolution context, when the candidates for the antecedent of an anaphor in the current utterance U_i have to be identified, the centering model proposes that the discourse entities in the immediately preceding utterance U_{i-1} be considered.

As an extension of this theory, Nariyama (2002) proposed a algorithm of zero-anaphora resolution, including:

- An extra forward-looking center list, named *salient referent list* (*SRL*) is defined. The SRL can deal with entities in all of the preceding utterances, whereas the original Centering Theory does only accounts for the entities in the immediately preceding utterance. Furthermore, if there are more than one zero-pronouns in the target sentence, her algorithm identifies an antecedent among each entity in the SRL for a given zero-pronoun according to the order of the SRL, which is

Topicalized Subject (Global $>$ Local $>$ Quotation)
$>$ Subject $>$ Indirect Object $>$ Object $>$ Others

Figure 2.1. Salient referent order list

(a) ’[ϕ_X ...,] *X-wa* ... .’
(b) ’*X-wa* [ϕ_X ...,] ... .’
(c) ’[ϕ_X ... SS conjunctive,] *X-ga* ... .’

Note that square brackets denote subordinate clauses and SS (Same Subject) conjunctive is a member of the set of the conjunctive markers: $\{$*-nagara*, *-te*, *-si*, *-tutu*, -ϕ and *-tameni*$\}$. ϕ_X and X stand for a zero-pronoun and an antecedent of ϕ_X respectively. For example, the pattern (c) indicates the situation that an antecedent X appears in the matrix clause involving with the nominative case marker *ga* and an zero-pronoun ϕ_X appears in the preceding subordinate clause with SS conjunctive.

Figure 2.2. Nariyama's heuristics for subject zero-anaphora

followed as Figure 2.1.

- For anaphora resolution in complex sentences with zero-pronouns, a series of heuristics (given in Figure 2.2) are given precedence to the SRL-based antecedent identification. The system falls back to the SRL-based antecedent identification whenever the heuristics are not satisfied.

2.1.2 Baldwin's high precision pronoun resolution engine

Another famous rule-based model is Baldwin (1995)'s pronominal anaphora resolution model. His system, named *CogNIAC*, is a pronoun resolution engine designed around the assumption that there is a sub-class of anaphora that does not require full world knowledge and achieves greater than 70% precision with 70% and better recall for the test set in MUC-6. In order to avoid problems which require general purpose reasoning, CogNIAC only resolves pronouns when very high confidence rules have been satisfied. The core rules of CogNIAC are given in Table 2.4.

The performance of the system is appealing, however it is extremely difficult to manually encode linguistic findings into rules by considering widely ranging aspects from lexical to discourse factors. So, adding a new rule into the system may have a

Table 2.4. Baldwin's high precision rules

1) **Unique in Discourse**: If there is a single possible antecedent PA_i in the read-in portion of the entire discourse, then pick PA_i as the antecedent.

2) **Reflexive**: Pick nearest possible antecedent in read-in portion of current sentence if the anaphor is a reflexive pronoun.

3) **Unique in Current + Prior**: If there is a single possible antecedent PA_i in the prior sentence and the read-in portion of the current sentence, then pick PA_i as the antecedent.

4) *Possessive Pro*: If the anaphor is a possessive pronoun and there is a single exact string match PA_i of the possessive in the prior sentence, then pick PA_i as the antecedent.

5) **Unique Current Sentence**: If there is a single possible antecedent PA_i in the read-in portion of the current sentence, then pick PA_i as the antecedent.

6) **Unique Subject / Subject Pronoun**: If the subject of the prior sentence contains a single possible antecedent PA_i, and the anaphor is the subject of its sentence, then pick PA_i as the antecedent.

7) **Cb-Picking**: If there is a *backward-looking center* C_b in the current finite clause that is also a candidate antecedent, then pick C_b as the antecedent.

8) **Pick Most Recent**: Pick the most recent potential antecedent in the text.

The term 'possible antecedents' refers to the set of entities from the discourse that are compatible with the anaphor for gender, number and coreference restrictions.

bad effect on the performance if the rule is not compatible with other rules. The same can be said of the rule-based approaches in general.

2.2. Machine learning-based approaches

Corpus-based empirical approaches, such as (Soon et al., 2001; Ng and Cardie, 2002a), on the other hand, are cost effective, while having achieved a performance comparable to the best-performing rule-based systems for the coreference task test sets of MUC-6 and MUC-7. However, they tend to lack an appropriate reference to theoretical linguistic work on coherence and coreference. Given this background, one of the challenging issues we should explore next is to make a good marriage between theoretical linguistic findings and corpus-based empirical methods.

Previous learning-based methods for anaphora resolution can be classified into two approaches: the *search-based approach* and the *classification-based approach*. We will discuss the advantages and disadvantages of each in Section 5.2.

2.2.1 Search-based model

The search-based approach determines the anaphoricity of a given NP indirectly as a by-product of searching the preceding context for its antecedent. If an appropriate candidate for the antecedent is found, the NP is classified as anaphoric; otherwise, non-anaphoric. Models proposed by Soon et al. (2001) and Ng and Cardie (2002a) fall into this class. In Soon et al.'s method (see Figure 2.3), for example, given a target NP (*Ana*) for resolution, the model processes each of its preceding NPs (i.e. candidate antecedents) in the right-to-left order, determining whether or not it is coreferent with the NP_i, until a positive answer (i.e. antecedent) comes up. If all the preceding NPs are classified negative, *Ana* is judged to be non-anaphoric. We call this approach the *search-based approach*. It has the advantage of using *broader context information* in the sense that the model determines the anaphoricity of an NP by examining whether the context preceding the NP in the discourse has a plausible candidate antecedent or not. Soon et al., in fact, defined the feature set including broad contextual information such as that shown in Table 2.5.

Following Soon et al.'s work, Ng and Cardie (2002a) improved upon the model by (a) expanding the feature set (see Table 2.6 and Table 2.7), and (b) introducing a new search algorithm that searches for the NP with the highest coreference likelihood value. According to Ng and Cardie (2002a), their model outperforms the Soon et al.'s model, which has also been supported by our experiment on Japanese zero-anaphora resolution reported in Chapter 4.

2.2.2 Classification-then-search model

The second approach is to introduce the process of anaphoricity determination separately from antecedent identification (Ng and Cardie, 2002b; Ng, 2004). We call this approach the *classification-based approach*. Unlike the search-based approach, it has the advantage that it uses labeled instances derived from non-anaphoric NPs as well as

```
Function Search-for-Antecedent ( $Ana$: candidate anaphor,
                                 $C$: set of candidate antecedents )
    $Max\_Ant := \phi$;  $Max\_Score := -\infty$;
    for $NP_i \in C$ do
        // judge whether or not Ana is anaphoric with $NP_i$
        $Score$ := classify-antecedenthood ($Ana$, $NP_i$);
        if $Score > Max\_Score$ then
            $Max\_Ant := NP_i$;  $Max\_Score := Score$;
        end
    end
    if $Max\_Score > \theta_{ant}$ then
        return $Max\_Ant$
    else
        return $NULL$
    end
end
```

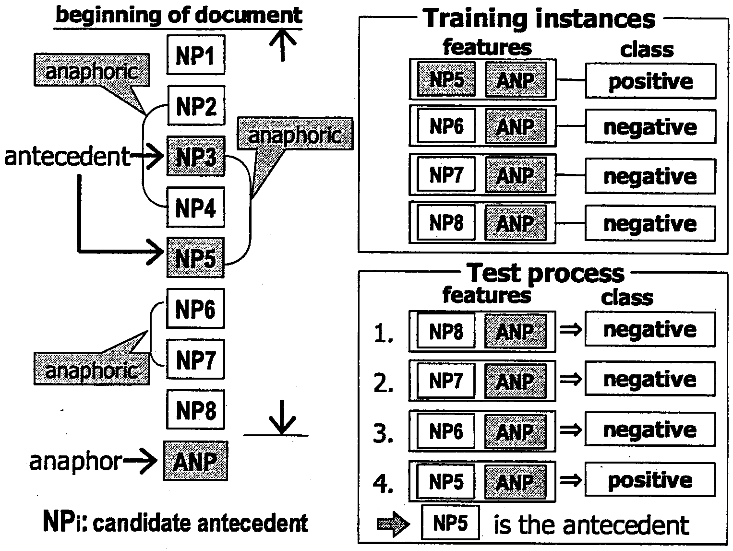
θ_{ant} is a global variable that indicates a global threshold parameter of antecedenthood.

Figure 2.3. The search-based model

those from anaphoric NPs to induce an anaphoricity classifier. For example, Ng (2004) proposed the following model (see Figure 2.5):

1. first carries out anaphoricity determination using a classification-based model to filter out a target NP (*Ana*) whose anaphoricity score *Ana_Score* is below threshold θ_{ana},
2. then searches for the antecedent for the remaining *Ana*, and
3. finally outputs the best-scored candidate antecedent *Max_Ant* if its score *Ant_Score* is above threshold θ_{ant}, or classifies the *Ana* as non-anaphoric otherwise.

Here we term this model the *classification-then-search model* because the model first determines the anaphoricity of a given candidate anaphor and then searches for the antecedent for the candidate anaphor.



The figure illustrates how model training and anaphora resolution are carried out, assuming that there are eight noun phrases, NP_1 through NP_8, which precede a noun phrase *ANP* in question. NP_2 and NP_4, NP_3 and NP_5, and NP_6 and NP_7 are coreferent respectively, and NP_5 (and its coreferent NP_3) is the antecedent of *ANP*. Under this situation, the model detects the antecedent by answering a sequence of candidate-wise boolean classification questions: whether or not NP_i is *ANP*'s antecedent for each $i \in \{1,\ldots,8\}$.

Figure 2.4. The search-based model proposed by Soon et al. and Ng and Cardie.

The classification-then-search model cautiously filters out non-anaphoric NPs according to the threshold parameter θ_{ana} at the first step. Second, the model also determines the anaphoricity of the remaining candidate anaphor according to the threshold parameter θ_{ant} as well as identifies an antecedent. This two-step anaphoricity determination model is designed because the anaphoricity determination component is not powerful enough to entirely free the antecedent identification component from the charge of anaphoricity determination. As Ng (2004) reports, optimizing the two threshold parameters could improve the performance for the overall task of anaphora.

```
Function Classify-Anaphor-and-Search-for-Antecedent(
                                        Ana: candidate anaphor,
                                        C: set of candidate antecedents )
    // judge whether or not Ana is anaphoric
    Ana_Score := classify-anaphoricity ( Ana );
    if Ana_Score > $\theta_{ana}$ then
        return Search-for-Antecedent ( Ana, C );
    else
        return NULL;
    end
end
```

θ_{ana} is a global variable that indicates a global threshold parameter of annaphoricity.

Figure 2.5. The classification-then-search model

Table 2.5. Feature set used in Soon et al.'s model.

Feature Type	Feature	Description
Lexical	SOON_STR	C if, after discarding determiners, the string denoting NP_i matches that of NP_j; else I.
Grammatical	PRONOUN_1	Y if NP_i is a pronoun; else N.
	PRONOUN_2	Y if NP_j is a pronoun; else N.
	DEFINITE_2	Y if NP_j starts with the word "the;" else N.
	DEMONSTRATIVE_2	Y if NP_j starts with a demonstrative such as "this," "that," "these," or "those;" else N.
	NUMBER	C if the NP pair agree in number; I if they disagree; NA if number information for one or both NPs cannot be determined.
	GENDER	C if the NP pair agree in gender; I if they disagree; NA if gender information for one or both NPs cannot be determined.
	BOTH_PROPER_NOUNS	C if both NPs are proper names; NA if exactly one NP is a propose name; else I.
	APPOSITIVE	C if the NPs are in an appositive relationship; else I.
Semantic	WNCLASS	C if the NPs have the same WordNet semantic class; I if they don't; NA if the semantic class information for one or both NPs cannot be determined.
	ALIAS	C if one NP is an alias of the other; else I.
Positional	SENTNUM	Distance between the NPs in terms of the number of sentences.

The feature set contains relational and non-relational features. Non-relational features test some property P of one of the NPs under consideration and take on a value of YES or NO depending on whether P holds. Relational features test whether some property P holds for the NP pair under consideration and indicate whether the NPs are COMPATIBLE or INCOMPATIBLE w.r.t. P; a value of NOT APPLICABLE is used when property P does not apply.

Table 2.6. Feature set used by Ng and Cardie (1/2)

Lexical		PRO_STR*	C if both NPs are pronominal and are the same string; else I.
		PN_STR*	C if both NPs are proper names and are the same string; else I.
		WORDS_STR	C if both NPs are non-pronominal and are the same string; else I.
		SOON_STR_NONPRO*	C if both NPs are non-pronominal and the string of NP matches that of NP; else I.
		WORD_OVERLAP	C if the intersection between the content words in NP and NP is not empty; else I.
		MODIFIER	C if the prenominal modifiers of one NP are a subset of the prenominal modifiers of the other; else I.
		PN_SUBSTR	C if both NPs are proper names and one NP is a proper substring (w.r.t. content words only) of the other; else I.
		WORDS_SUBSTR	C if both NPs are non-pronominal and one NP is a proper substring (w.r.t. content words only) of the other; else I.
Grammatical	NP type	BOTH_DEFINITES	C if both NPs start with "the;" I if neither start with "the;" else NA.
		BOTH_EMBEDDED	C if both NPs are prenominal modifiers ; I if neither are prenominal modifiers; else NA.
		BOTH_IN_QUOTES	C if both NPs are part of a quoted string; I if neither are part of a quoted string; else NA.
		BOTH_PRONOUNS*	C if both NPs are pronouns; I if neither are pronouns, else NA.
	heuristics	CONSTRAINTS*	C if the NPs agree in GENDER and NUMBER and do not have incompatible values for CONTRAINDICES, SPAN, ANIMACY, PRONOUN, and CONTAINS PN; I if the NPs have incompatible values for any of the above features; else NA.
		CONTAINS_PN	I if both NPs are not proper names but contain proper names that mismatch on every word; else C.
		DEFINITE_1	Y if NP starts with "the;" else N.
		EMBEDDED_1*	Y if NP is an embedded noun; else N.
		EMBEDDED_2	Y if NP is an embedded noun; else N.
		IN_QUOTE_1	Y if NP is part of a quoted string; else N.
		IN_QUOTE_2	Y if NP is part of a quoted string; else N.
		PROPER_NOUN	I if both NPs are proper names, but mismatch on every word; else C.
		TITLE*	I if one or both of the NPs is a title; else C.

*'d features are in the hand-selected feature set for at least one classifier/data set combination.

Table 2.7. Feature set used by Ng and Cardie (2/2)

Grammatical	role	BOTH_SUBJECTS	C if both NPs are grammatical subjects; I if neither are subjects; else NA.
		SUBJECT_1*	Y if NP is a subject; else N.
		SUBJECT_2	Y if NP is a subject; else N.
	linguistic con-straints	AGREEMENT*	C if the NPs agree in both gender and number; I if they disagree in both gender and number; else NA.
		ANIMACY*	C if the NPs match in animacy; else I.
		MAXIMALNP*	I if both NPs have the same maximal NP projection; else C.
		PREDNOM*	C if the NPs form a predicate nominal construction; else I.
		SPAN*	I if one NP spans the other; else C.
		BINDING*	I if the NPs violate conditions B or C of the Binding Theory; else C.
		CONTRAINDICES*	I if the NPs cannot be co-indexed based on simple heuristics; else C. For instance, two non-pronominal NPs separated by a preposition cannot be co-indexed.
		SYNTAX*	I if the NPs have incompatible values for the BINDING, CONTRAINDICES, SPAN or MAXIMALNP constraints; else C.
	ling. prefs	INDEFINITE*	I if NP is an indefinite and not appositive; else C.
		PRONOUN	I if NP is a pronoun and NP is not; else C.
Semantic		CLOSEST_COMP	C if NP is the closest NP preceding NP that has the same semantic class as NP and the two NPs do not violate any of the linguistic constraints; else I.
		SUBCLASS	C if the NPs have different head nouns but have an ancestor-descendent relationship in WordNet; else I.
		WNDIST	Distance between NP and NP in WordNet (using the first sense only) when they have an ancestor-descendent relationship but have different heads; else infinity.
		WNSENSE	Sense number in WordNet for which there exists an ancestor-descendent relationship between the two NPs when they have different heads; else infinity.
	Pos	PARANUM	Distance between the NPs in terms of the number of paragraphs.
	Other	PRO_RESOLVE*	C if NP is a pronoun and NP is its antecedent according to a naive pronoun resolution algorithm; else I.
		RULE_RESOLVE*	C if the NPs are coreferent according to a rule-based coreference resolution algorithm; else I.

Chapter 3

The NAIST Text Corpus: Annotating Predicate-Argument and Coreference Relations

In this chapter, we discuss how to annotate predicate-argument relations including zero-anaphoric relations and coreference relations in Japanese written text. Predicate argument analysis and coreference resolution are particularly important as they often provide a crucial bridge between basic NLP techniques such as morpho-syntactic analysis and end-level applications. They have been mainly developed with corpus-based empirical approaches. In order to train a classification model in such approaches, a large scale corpus annotated with information about predicate-argument and coreference is needed. To our best knowledge, however, there is no large-scale corpus including such tags in Japanese. In addition, we have difficulty adopting the existing specifications for annotating tags due to the problem setting of each task and differences in Japanese and English. So, we develop new criteria for our annotation processes by examining previous work on annotating tasks. This chapter explains our annotating specification cultivated through actual annotation of texts from Kyoto Text Corpus version 3.0[1], and discusses the future directions.

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

3.1. Introduction

Coreference resolution and predicate-argument analysis became extensive fields of research due to the demands of NLP tasks such as information extraction and machine translation which rely on their analysis. With research focus placed on these tasks, the specifications for annotating corpora, and the data sets themselves used in supervised techniques (Hirschman, 1997; Kingsbury and Palmer, 2002; Doddington et al., 2004) have also grown in sophistication.

In the task of coreference resolution, some annotation schemes have been already proposed and annotated corpora have developed according to these schemes (Hirschman, 1997; Kawahara et al., 2002; Hasida, 2002; Poesio et al., 2004; Doddington et al., 2004). For instance, in the Coreference (CO) task on Message Understanding Coreference (MUC) and Entity Detection and Tracking (EDT) task in Automatic Content Extraction (ACE) program, which is the successor of MUC, the details of specification for annotating coreference relations have been discussed over several years. The specification of predicate-argument analysis task, however, has been mainly discussed in the shared task[2] of the Conference on Computational Natural Language Learning (CoNLL) based on *PropBank* (Palmer et al., 2005).

In order to research in the areas of coreference and predicate-argument analysis in Japanese, a large annotated corpus is needed. However, the existing resources such as GDA-tagged corpus[3] and Kyoto Text Corpus version 4.0 (Kawahara et al., 2002) do not have enough annotated data to evaluate each task. Furthermore, we also have to consider the following two aspects:

- the problems caused by directly adopting the specifications of MUC and ACE, which are specific to the information extraction task, and
- the effects from the difference between English and Japanese on each task.

In this chapter, we investigate the previous work of annotating coreference and predicate-argument relations and discuss how to annotate each relations in Japanese written texts. In Section 3.2, we review the difference between "anaphora" and "coreference" and briefly introduce the previous work on annotating coreference and predicate-argument relations in Section 3.3. Next, Section 3.4 shows the guideline of our corpus

[2] http://www.lsi.upc.edu/~srlconll/

[3] The GDA (Global Document Annotation (Hasida, 2002))

based on the previous work. After that, we discuss the problems on the annotating process in Section 3.5 and conclude in Section 3.6. As the results of the current work of this chapter, we have released *NAIST Text Corpus*[4] version 1.2β.

3.2. Anaphora and coreference

Anaphora is a linguistic phenomenon where an expression points back to another expression in the preceding context. The word or phrase pointing back is called an *anaphor* and the expression which is referred to by an anaphor is its *antecedent*. In comparison, the relations between two or more mentions which refer to the same entity is called *coreference* relations. Note that some anaphoric relations (e.g. the relation between a person name and its pronoun) are also coreference relations. For example, in text (4), the pronoun $kare_i$ (he) points back to *Koizumi* $shusho_i$ (prime minister Koizumi) and these two mention refer to the same entity in the world and then we can regard them as both anaphoric and coreference relations.

(4)	*Koizumi*	$shusho_i$*-wa*	...
	Koizumi	prime minister$_i$-TOP	...
	$kare_i$*-wa*	...	
	he$_i$-TOP	...	

On the other hand, in text (5), we can also regard the relation between $iPod_i$ (iPod$_i$) and $sore_j$ (it$_j$) as anaphoric relation because $sore_j$ points back to iPod$_i$. However, these two mentions are not coreferential since they refer to the different entities in the world.

(5)	*Tom-wa*	$iPod_i$*-o*	*Kat-ta*	.
	Tom-TOP	iPod$_i$-ACC	buy-PAST	PUNC
	Tom bought an iPod.			
	Mary-mo	$sore_j$*-o*	*Kat-ta*	.
	Mary-TOP	it$_j$-ACC	buy-PAST	PUNC
	Mary also bought one.			

As above examples, anaphoric relations are classified into either two mentions refer to the same entity or not. The former case is called as *identity-of-reference anaphora*(*IRA*)

[4] http://cl.naist.jp/nldata/corpus/

and the latter *identity-of-sense anaphora(ISA)* in Mitkov (2002). Due to the crucial difference between IRA and ISA, there has been confusion on the treatments of these two relations in the previous work. As we show in Section 3.3, a variety of specifications for annotated corpora have been developed according to the different interpretations of anaphora and coreference.

3.3. Previous work

In this section, we briefly review the previous work on annotating coreference and predicate-argument relations in the corpora.

3.3.1 Annotating coreference relations

The task of coreference resolution has been mainly developed from an information extraction perspective. For instance, in the 6th and 7th Message Understanding Conferences , MUC being one of the more famous conferences in information extraction, coreference resolution is treated as a subtask of information extraction[5]. The annotated corpora built in MUC contains coreference relations between NPs, which are used as the gold standard data set for machine learning-based approaches to coreference resolution by researchers such as Soon et al. (2001) and Ng and Cardie (2002a). However, van Deemter and Kibble (1999) reported that the specification of the MUC coreference task guides us to annotate expressions which are not normally judged as coreferential as coreference relations, such as quantitative expressions (e.g. *every* and *most*) and appositive relations (e.g. *Julius Caesar*$_i$, *the/a well-known emperor*$_i$, ...).

In the task of Entity Detection and Tracking in the Automatic Content Extraction program (Doddington et al., 2004), which is the successor of MUC, the coreference relations is redefined by introducing the two concepts, *mentions* and *entities*, in order to avoid to redundant identification of such coreference relations. Mentions are the expressions appearing in the texts, including the proper nouns which is the extraction target in the information extraction task. On the other hand, entities stand for the conceptual entities consisting of all of the mentions in the texts. For example, in Figure 3.1, *John* and *He* are the mentions which refer to the same entity *entity*$_i$.

[5] http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html

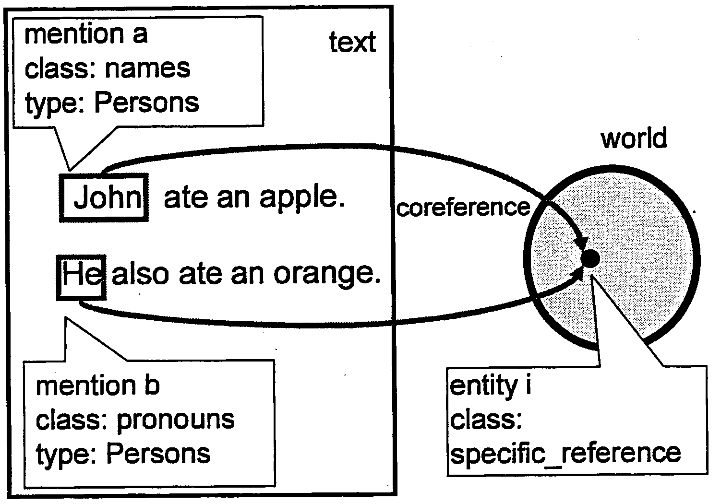


Figure 3.1. Relationship between mentions and entities

The target mentions in the annotation of EDT[6] are restricted to the expressions which are some sorts of named entity types such as PERSON and ORGANIZATION and refer to specific entities in the world. Therefore, ACE data set has the drawback since that not all coreference relations in the text are always exhaustively annotated, it is not enough to resolve all annotated coreference relations in order to properly analyze a text.

In Japanese, Kyoto Text Corpus version 4.0 (Kyoto Corpus) and GDA-tagged Corpus (GDA Corpus) contain the tags of coreference relations. For example, Kyoto Corpus includes as many as 114,729 coreference relations as well as dependency relations. Note that the relations between a mention referring to an entity and a mention referring to the corresponding attribute of that entity are regarded as coreferential as well as the relations where two mentions refers to the same entity. The GDA Corpus, on the other hand, contains the relations which refer to generic nouns as well as specific nouns, that is, the coreference relations in GDA Corpus are annotated as both IRA and ISA relations.

[6] http://projects.ldc.upenn.edu/ace/annotation/

3.3.2 Annotating predicate-argument relations

There are a variety of discussions over annotating task for a predicate-argument relations, which are annotated in terms of various annotating levels such as surface cases and thematic roles. *PropBank* (Palmer et al., 2005), which is one of the practical annotated corpus in English, contains 35 relations such as ARG0, ARG1, ..., ARG5, AA, AM, AM-ADV, which are conceptually related to thematic roles. In sentence (6), for instance, "the refiner" as ARG0, which relates to agent role, and "$66 million, or $1.19 a share" as ARG1, which relates to theme role, for a given target verb "earned".

(6) $[_{\text{ARGM-TMP}}$ *A year earlier*], $[_{\text{ARG0}}$ *the refiner*] $[_{\text{rel}}$ *earned*] $[_{\text{ARG1}}$ *\$66 million, or \$1.19 a share*].

Note that the range of annotating arguments on PropBank is limited to a given same sentence because arguments for a given predicate appear in the same sentence in languages such as English.

In contrast, since arguments are frequently omitted in Japanese due to zero-anaphoric phenomena, we have to search arguments beyond the sentence containing the target predicate. For this reason, Kyoto Corpus includes inter-sentential and exophoric zero-anaphora relations for each omitted argument. In text (7), for instance, the nominative argument of the predicate *kaeru (go back)* in the second sentence is omitted and refers to *Tom* in the first sentence. The dative of that predicate is also omitted, however the corresponding argument does not explicitly appear in text (7). In such case, this dative is annotated as "*exophoric use*" according to the specification of Kyoto Corpus.

(7) *Tom_i-wa kyo gakko-ni it-ta*
Tom_i-TOP today school-DAT go
Tom went to school today.

(ϕ_i-*ga*) ($\phi_{exophoric}$-*kara*) *kae-tte suguni asobi-ni dekake-ta*
ϕ_i-NOM $\phi_{exophoric}$-ABL go back immediately play-DAT go out-PAST
He went to play as soon as he came back from school.

In GDA Corpus, the predicate-argument relations are labeled as thematic role such as agent and theme, while ones in Kyoto Corpus are annotated as surface cases such as *ga* (nominative), *o* (accusative)[7]. To our best knowledge, GDA Corpus does not

[7] Strictly speaking, in Japanese the corresponding thematic roles for a given surface case differs depending on the appearing context of it.

contain intra-sentential zero-anaphoric relations as predicate-argument relations, so it has a serious drawback when we utilize that data set as training data set on machine learning techniques.

3.3.3 Annotating event-nouns and their arguments

In addition to annotating predicate-argument for verbs and adjectives, researchers have been focusing on annotating predicate-argument relations for the NPs representing an event in the context, which we call *event-nouns*.

As an example of creating event-noun resources, Meyers et al. (2004) built the *NomBank*, where predicate-argument relations of event-nouns are annotated based on the specification of PropBank (Palmer et al., 2005), taking Penn Treebank (Marcus et al., 1993) as the target corpus. For example, in phrase (8), the noun "*growth*" stands for some sorts of event meaning "*theme* grows *in some situations*" and "*in dividends*" and "*next year*" are annotated as ARG1, which is basically related to theme role, and ARGM-TMP, which is related to adjuncts, respectively.

(8) *12% growth in dividends next year* [REL=*growth*, ARG1=*in dividends*, ARGM-TMP=*next year*]

Note that arguments in NomBank are restricted as the NP appearing in the same sentence of a target event-noun, since NomBank complies with the PropBank specifications.

For Japanese, event-nouns and their arguments are also annotated in Kyoto Corpus. As shown in sentence (9), *akaji*$_i$ *(deficit)* is assigned to the nominative for the event-noun *eikyo (influence)*.

(9) *kono* *boueki* *akaji*$_i$*-wa* *waga* *kuni-no* *kyosoryoku*$_j$*-ni* ***eikyo**-o* *oyobosu*
this trade deficit-TOP our country-OF competitiveness-DAT **influence**-ACC affect
[REL=*eikyo* (influence), NOM=*akaji*$_i$ (deficit), DAT=*kyosoryoku*$_j$ *(competitiveness)*]
The trade deficit affects our competitiveness.

In some cases, the relation between an event-noun and its argument is compressed into a complex noun (or a complex noun phrase), such as "*kouho* (candidate) *senbatsu* (selection)", which means "candidate selection", and "*dassou (desertion) hei (soldier)*", which means "army deserter". So, we need to explicitly define when an NP in the context is treated as event-nouns.

3.4. Specification of NAIST Text Corpus

Taking on the previous work described in Section 3.3 into account, our annotated corpus called *NAIST Text Corpus* currently contains three relations: (a) predicate-argument relations, (b) event-noun and its arguments, and (c) coreference relations between two NPs which refer to the same entity, i.e. IRA relations.

3.4.1 Annotating predicate-argument relations

For recognizing predicates, annotators assign an expression to predicate by judging whether or not the expression is contained in the three parts-of-speech (verb, adjective or noun+copula) based on the lexical entries in *ipadic* (Asahara and Matsumoto, 2003).

For annotating arguments of a predicate, there are a variety of annotation layers: surface cases adopted by Kyoto Corpus, thematic roles used in GDA and the original specifications based on thematic roles in PropBank. In comparison to these previous work, what we want to extract from texts is a set of arguments for an active form of a given predicate as extracting pieces in information extraction perspective. So, if a predicate is used as a passive or causative form in the text, we interpret the predicate as a active form and annotate each argument of this active predicate. Note that it is unclear what kinds of information of predicates should be eliminated from surface cases, we currently annotate nominative, accusative and dative arguments of each predicate.

For example, in Kyoto Corpus *watashi$_i$ (I)*, *kare$_j$ (he)* and *ringo$_k$ (apple)* are annotated as the nominative, dative and accusative respectively for the causative verb *tabe-saseru (make one eat)*. In NAIST Text Corpus, on the other hand, *kare$_j$ (he)* and *ringo$_k$ (apple)* are annotated as the nominative and dative for the active verb "*taberu (eat)*". We also add an additional tag into the relationship between *eat* and *watashi*, because there is no information between them in which the predicate *eat* is treated as active voice.

(10) *watashi$_i$-wa* *kare$_j$-ni* *ringo$_k$-o* *tabe-saseru*
I$_i$-TOP he$_j$-DAT apple$_k$-ACC eat-CAUSATIVE
(I make him eat an apple.)
Kyoto Text Corpus: [REL=*tabe-saseru* (eat-CAUSATIVE), NOM=*watashi$_i$* (I$_i$), ACC=*ringo$_k$* (apple$_k$), DAT=*kare$_j$* (he$_j$)]
NAIST Text Corpus: [REL=*tabe-(ru)* (eat-ACTIVE), NOM=*kare$_j$* (he$_j$), ACC=*ringo$_k$* (apple$_k$), ADDITIONAL CASE=*watashi$_i$* (I$_i$)]

Table 3.1. Comparison of annotating predicate-argument relations

corpus	label	range
PropBank	pseudo thematic role	intra
GDA-tagged corpus	thematic role	inter, exo
Kyoto Text Corpus	surface case (including alternations)	intra, inter, exo
NAIST Text Corpus	surface case (not including alternations)	intra, inter, exo

intra: intra-sentential relations, inter: inter-sentential relations, exo: exophoric relations

A comparison of the specifications is summarized in Table 3.1.

3.4.2 Annotating event-nouns and their arguments

The relations between event-noun and its argument is also annotated based on obligatory surface cases such as *ga (nominative)*, *o (accusative)* and *ni (dative)* as well as predicate-argument relations for verbs and adjectives. For a given noun (or noun phrase), human annotators judge whether or not the noun represents an event in the context and if the noun is classified into event-noun, then search its arguments for that event-noun. In sentence (11), for instance, annotators have to judge *denwa$_i$ (phone)* as an event-noun and then annotate *kare$_a$ (he)* as nominative argument and *watashi$_b$* as dative for *denwa$_i$*, since it is interpreted as the core word in the event "*He called to me*". In contrast, *denwa$_j$* is not an event-noun because that word means "*my cell-phone*".

(11) *kare$_a$-karano* *denwa$_i$-niyoruto* *watashi$_b$-wa* *kare-no* *ie-ni* *wasure-tarasii*
he$_a$-ABL phone$_i$ according to I$_b$-NOM his-OF home-LOC leave-PAST
According to his phone call, I might leave my cell-phone in his home.

Compound nouns require special treatment. We apply the following steps to identify event-nouns for annotations.

1. The semantic compositionality test. If the meaning of the compound noun is clear from only the meaning of its composite words, it is considered compositional.

2. Evaluation of constituents as event-nouns. If a compound has been judged semantically compositional by the compositionality test, it is divided into con-

stituents, and any constituents that are event-nouns are annotated.

As an example, the compound "*itaku keiyaku (consignment contract)*" is found to pass the test and identified as event-nouns and both *itaku* and *keiyaku* will be considered for annotation. In contrast, "*furansu kakumei (French Revolution)*" does not pass the compositionality test.

3.4.3 Annotating coreference relations

The previous work for annotating coreference shows two choices, either the ISA relations are judged to be included in coreference relations in addition to the IRA relations or not. If the ISA relations are included in coreference relations, the annotators have to do complicated judges for annotation by considering *class inclusion*, whether or not the concept of NP_i includes the concept of NP_j for given all of the two $\langle NP_i, NP_j\rangle$ in the target text. For example, in text (12), the pair of generic noun $toshokan_a$ *(library)* and $toshokan_b$ may be judged as coreferential because the concept of $toshokan_a$ is equivalent to the concept of $toshokan_b$. However, the pair of the two nouns, hon_i *(book)* and hon_j, might not be judged as coreferential, since hon_i *(book)* refers to the concept "*a set of printed pages that are fastened together in a cover so that you can read them*", while because it is modified by "*toshokan no (library's)*", hon_j refers to the concept "books located in the library".

(12) *toshokan$_a$-niwa* *hon$_i$-ga* *oi-tearu*
library$_a$-LOC book$_i$-NOM place-ASPECT
There are books in the library.

toshokan$_b$-no *hon$_j$-wa* *kariru-kotogadekiru*
library$_b$-OF book$_j$-TOP borrow-CAN
We can borrow the books in the library.

As we can be seen in the above examples, whether a pair of two generic nouns is coreferential or not depends on the their contexts. It causes difficulty in judging coreference relations of generic nouns. For this reason, we deals with only the IRA relations as coreference in our specification, whereas the ISA relation is adopted in case of annotating predicate-argument relations and the relationship between event-nouns and their arguments.

As we described in Section 3.3.1, in EDT of ACE, mentions and entities are classified into some sorts of named entity types such as PERSON and ORGANIZATION, thus

Table 3.2. Difference of annotating coreference relations in previous work

corpus	annotating target
GDA-tagged corpus	IRA and ISA
ACE EDT	IRA (types and classes of entities are restricted)
Kyoto Text Corpus	IRA and ISA
NAIST Text Corpus	IRA

IRA: identity-of-reference anaphora, ISA: identify-of-sense anaphora

the noun which is not classified into any types can not be related to other nouns as a coreference relation even if it can be interpreted as coreferential. Therefore, in the current annotation process, named entity types of nouns are not restricted and coreference relations in texts are not restricted as are annotated according to the following three criteria:

1. An anaphor is annotated only when it appears in the syntactic head of the target NP.
2. An NP which explicitly appears in the discourse is regarded as an antecedent for a given anaphor.
3. A generic noun is not treated as both an anaphor and an antecedent.

A comparison between our specification and previous work is shown in Table 3.2.

3.4.4 Statistics

According to the specifications in Section 3.4.1, Section 3.4.2 and Section 3.4.3, two annotators worked on the task of annotating predicate-argument and coreference relations, taking all documents in Kyoto Text Corpus version 3.0 (containing 38,384 sentences in 2,929 texts) as a target corpus. The numbers of annotating predicate-argument relations are shown in Table 3.3. Each argument is categorized into five cases: (a) both a predicate and its argument appear in same phrases, (b) an argument depends on its predicate or a predicate depends on its argument, (c) a predicate and its argument has a intra-sentential zero-anaphora relation, (d) a predicate and its argument has a inter-sentential zero-anaphora relation and (e) an argument does not explicitly appear in the

Table 3.3. Statistics: annotating predicate-arguments relations

		ga (nominative)		o (accusative)		ni (dative)	
predicates 106,628	(a) same phrase	177	(0.002)	60	(0.001)	591	(0.027)
	(b) dependency relations	44,402	(0.419)	35,882	(0.835)	18,912	(0.879)
	(c) zero-anaphoric (intra-sentential)	32,270	(0.305)	5,625	(0.131)	1,417	(0.066)
	(d) zero-anaphoric (inter-sentential)	13,181	(0.124)	1,307	(0.030)	542	(0.025)
	(e) exophoric	15,885	(0.150)	96	(0.002)	45	(0.002)
	total	105,915	(1.000)	42,970	(1.000)	21,507	(1.000)
event-nouns 28,569	(a) same phrase	2,195	(0.077)	5,574	(0.506)	846	(0.436)
	(b) dependency relations	4,332	(0.152)	2,890	(0.263)	298	(0.154)
	(c) zero-anaphoric (intra-sentential)	9,222	(0.324)	1,645	(0.149)	586	(0.302)
	(d) zero-anaphoric (inter-sentential)	5,190	(0.183)	854	(0.078)	201	(0.104)
	(e) exophoric	7,525	(0.264)	42	(0.004)	10	(0.005)
	total	28,464	(1.000)	11,005	(1.000)	1,941	(1.000)

text (exophoric use). Table 3.3 shows that in annotation for predicates over 80% of both *o* (accusative) and *ni* (dative) arguments were annotated as dependency relations, while around 60% of *ga* (nominative) argument was annotated as zero-anaphoric relations. In comparison, in the case of event-nouns, *o* and *ni* arguments are likely to appear in same phrase of given event-nouns and about 80% of *ga* argument has a zero-anaphoric relations with event-nouns.

10,531 entities (25,357 anaphors) are annotated as annotated coreference relations. The number of coreference relations is quite smaller than that of Kyoto Corpus, because the IRA relations are only considered as coreferential in our specification, while Kyoto Corpus contains the ISA relations as coreference relations.

Next, to evaluate the agreement between two human annotators, randomly selected 30 articles were annotated by them. The annotation results are evaluated by calculating recall and precision in which one annotation result is regarded as correct examples and

Table 3.4. Agreement of annotating each relation

	recall		precision	
predicate	0.921	(806/875)	0.944	(806/854)
ga (nominative)	0.823	(683/830)	0.829	(683/824)
o (accusative)	0.899	(329/366)	0.954	(329/345)
ni (dative)	0.724	(105/145)	0.890	(105/118)
event-noun	0.965	(247/256)	0.792	(247/312)
ga (nominative)	0.735	(191/260)	0.743	(191/257)
o (accusative)	0.827	(86/104)	0.869	(86/99)
ni (dative)	0.389	(7/18)	0.583	(7/12)
coreference	0.813	(126/155)	0.813	(126/155)

the other as outputs of system. Note that arguments of predicates and event-nouns are considered for calculation of recall and precision only when predicates (event-nouns) are annotated by both annotators. For evaluation of coreference relations, we calculated recall and precision based on MUC score (Vilain et al., 1995). The results of each relation are shown in Table 3.4. According to Table 3.4, we can see that most annotating works were done with confidential quality except the minorities. However, each annotation still leaves the room for improvement. In Section 3.5, we will explain the problems of annotating each relation and discuss the future directions to solve them.

3.5. Difficulties in annotating task and future directions

In this section, we explain the difficulties on the annotating process of predicate-argument, coreference and event-nouns and its arguments in Japanese. After that, we discuss the future directions for them.

3.5.1 Difficulty in annotating predicates

As predicates sometimes has an ambiguity in the sense between a predicate and a compound functional expression which consists of more than one word including both content words and functional words, it causes to the inconsistency of judging whether such candidate predicate is a predicate or not. For this ambiguity, Tsuchiya et al. (2006) built the database of these kinds of compound functional expressions. They reported that the agreement ratio between annotators for annotating only functional expressions became higher in their experiments. On the other hand, in our experiments we achieved not good performance as shown in Table 3.4 compared with Tuchiya's evaluation. It may be caused by difficulty in judging a candidate expression as a predicate or not by considering its arguments. For instance, the compound functional expression "*toshite*" has two ambiguities, "*do*" as content usage and "*assignment of some meaning in one's perspective*" as functional usage, and the annotator judges its meaning depending on its appearing context, however it is difficult to exactly classify such meaning. In order to solve this problem, we are planing to predefine the preferred interpretation of each expressions. We believe that to present the definitions as clues will help annotator's work.

3.5.2 Difficulty in annotating event-nouns

In order to annotate event-nouns, we have to judge whether or not a complex noun can be compositionally decomposed into its constituents. However, the criteria for compositional decomposition between two annotators does not disagreed, then the agreement ratio shown in Table 3.4 has decreased. The ambiguity whether a given expression is an event-noun or not also causes to the annotation problem. The expressions such as *keiyaku (contract)*, *kisei (regulation)* and *toushi (investment)* are interpreted as the direct results of the event encoded in the noun as well as the event itself depending on the context. For example, in sentence (13), we can interpret *keiyaku (contract)* as either the event-noun or the result. Thus, such cases also make the agreement ratio decrease.

(13)	*sono*	*kaisha-wa*	***keiyaku**-o*	*kaijos-ite*	*liesus-areta*	*jettoki-o*	*henkyakus-ita*
	that	company-TOP	**contract**-THEME	dissolve	leased	jet-THEME	surrender-PAST

The company dissolved its contract and surrendered its leased jet.

3.5.3 Difficulty in annotating arguments of predicates/event-nouns

In annotating arguments of predicates and event-nouns, a variety of case patterns causes to the majority of annotation disagreements. For example, the predicate *jitsugen-suru (realize)* has two case patterns: "AGENT-*ga (nominative)* THEME-*o (accusative)* *jitsugen-suru*" and "THEME-*ga jitsugen-suru*". If all arguments of this predicate are omitted, we can annotate the nominative case of this predicate as either AGENT or THEME because of the two interpretations above. Similar to this problem, ambiguity of interpretation about *agentivity* also causes to the ambiguity of argument annotations. In sentence (14), for example, the predicate *shibaru (bind)* has two types of case patterns shown in (15) if we *kisoku (rule)* has a agentivity in this context. To avoid this problem, we will decide which case pattern is preferable among patterns for the convenience of annotators' works.

(14) *kisoku-ga* *hitobito-o* *sibaru*
rule-NOM people-ACC bind
The rule binds people.

(15) a. [REL = *sibaru* (bind), AGENT = *kisoku* (rule), THEME = *hitobito* (people)]

b. [REL = *sibaru* (bind), AGENT = ϕ (exophoric), THEME = *hitobito* (people), INSTRUMENT = *kisoku* (rule)]

In addition to the above problem, a problem occurs when the relationship between a predicate and its argument is omitted. Suppose the situation shown in Figure 3.2. In case (a), since *He* is annotated as the nominative argument and *John* and *He* are annotated as coreference relations, thus we can also regard *John* as the nominative argument. In case (b), on the other hand, two nouns, *children* and *kids*, are not coreferential in the case that *children* and *kids* are both generic nouns. Under the situation, we can not infer the relationship between *children* and its predicate even if *kids* is annotated as the nominative argument of the predicate.

3.5.4 Difficulties in annotating coreference

In the task of annotating coreference relations, we still had problems. As one of those problems, recognizing the IRA relation for given two NPs is the majority of annotating problems, since it is so difficult to judge whether or not two abstract nouns refer to the same entity. As we described in Section 3.4.3, it is undesirable to restrict the class of the NP, for creating incomplete training instances for the NP which is not assigned into any mention classes. However, as one of the future directions, we are planning to investigate how to improve the agreement by limiting the classes of abstract nouns when annotating the coreference relations.

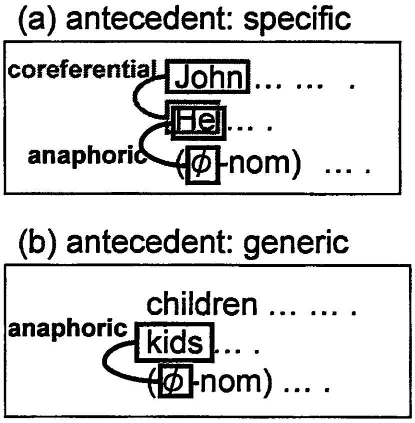


Figure 3.2. Difference of annotation between specific and generic antecedent

3.6. Summary

In this chapter, we reported the current specification of our annotated corpus for coreference resolution and predicate-argument analysis. According to the discussion in Section 3.4, we decided to annotate predicate-argument relations by ISA relation, whereas annotating coreference relations adopting IRA relation. Taking Kyoto Text Corpus version 3.0 as a target corpus, we built a large annotated corpus called *NAIST Text Corpus*. We also examined the revelation from the annotating process of our corpus, and discussed our future directions for refining the details of the specifications.

Chapter 4

Antecedent Identification Inspired from Centering Theory

4.1. Introduction

This chapter presents a method that incorporates contextual clues into machine learning-based approaches to anaphora resolution. As described in Chapter 2, in contrast with rule-based approaches, such as (Brennan et al., 1987; Lappin and Leass, 1994; Baldwin, 1995; Nakaiwa and Shirai, 1996; Okumura and Tamura, 1996; Mitkov, 1997), empirical, or corpus-based, approaches to this problem have shown to be a cost-efficient solution achieving performance that is comparable to the best performing rule-based systems (McCarthy and Lehnert, 1995; Ge et al., 1998; Soon et al., 2001; Ng and Cardie, 2002a; Strube and Müller, 2003; Iida et al., 2003; Yang et al., 2003). Given this background, one of the challenging issues we should explore next is to make a good marriage between theoretical linguistic findings and corpus-based empirical methods.

In this chapter, we report our attempt to enhance existing trainable coreference resolution models by incorporating such theoretical findings as the features utilized in Centering Theory. In Section 4.2, we discuss a significant drawback of Ng and Cardie's model and propose two solutions: (a) implementing the centering factors as what we call *centering features*, and (b) introducing a novel searching model, which we call a *tournament model*. We then report the results of our experiments on Japanese zero-anaphora resolution in Section 4.3 and conclude in Section 4.4.

4.2. Incorporating of contextual cues

As reviewed in Section 2.2.1, the search-based models such as (Soon et al., 2001; Ng and Cardie, 2002a) have achieved the performance comparable to the best-performing rule-based system. However, the search-based models have a serious drawback. Although Ng and Cardie (2002a) attempted to employ several types of features in their experiments (see Table 2.6 and Table 2.7), it should be pointed out that their model does not capture an important aspect of local context that has been proved useful for coreference interpretation in the literature of discourse analysis. We elaborate this flaw and propose two solutions.

4.2.1 A flaw of The baseline model

Consider the following two discourses:

(16) a. Mary went to see John$_i$.

b. He$_i$ was playing baseball.

(17) a. Tom$_i$ went to see John.

b. He$_i$ tried to explain what happened to him yesterday.

In (16), the subject of sentence (b), *He*, refers to the object of sentence (a), *John*. In (17), on the other hand, it is not the case although *He* and *John* fills the same syntactic role, respectively. An explanation for this difference derived from Centering Theory can be briefed as follows. In (17), *Tom* is chosen to be the preferred antecedent of *he* because:

(a) *Tom*, being the subject role filler, is the preferred center (i.e. the highest ranked entity of the forward looking centers) assigned in (a),

(b) *Tom* is thus most likely to be the backward looking center of (b), and

(c) if so is *Tom*, it must be realized as a pronoun.

In (16), on the other hand, *Mary*, the preferred center, violates the gender constraint imposed by *He*, and therefore the second ranked entity *John* is interpreted as the antecedent.

The essence of the above explanation is that it is derived from a model that takes into account the preference between candidates. Whether or not *John* is coreferent depends on the appearance of other entities, such as *Mary* and *Tom*, in its local context. This crucial property of local coherence is, however, not properly captured in Ng and Cardie's model because it views antecedent detection as a set of *candidate-wise* boolean classification problems.

4.2.2 Two solutions

Among various possibilities one may think of as a solution to the problem argued above, we have empirically examined two novel solutions.

Centering features

A straightforward solution is to augment the number of features that implement local contextual factors. For example, one may introduce a feature that indicates whether or not the antecedent candidate in question is the present preferred center. This feature can also be enhanced so that it can indicate whether or not the candidate is ranked the highest among the forward-looking centers while satisfying gender and number constraints. Such a feature would help the classification model to distinguish the two *Johns* in the previous examples. Note that the computation of such features requires the use of additional devices, such as a list for storing forward-looking centers, which has never been used in previous trainable models. We refer to such features as *centering features* for capturing centering state transitions. The centering features we used in our experiments will be presented in the next section.

The tournament model

Recall that what we wanted in *John*'s examples was a model that compares the first *John* with its opponent *Mary* and the second *John* with *Tom*. Our second solution is to implement a pairwise comparison between two candidates in reference to *ANP* as a binary classification problem (i.e. which candidate wins) and to conduct a tournament to check against the candidate. A tournament consists of a series of matches in which candidates compete with each other and the one that prevails through the final round is declared the winner, namely, identified as the antecedent. We call this new model the *tournament model*.

Observe the situation given in Figure 2.4 in Section 2.2.1 again, which we have reillustrated here as Figure 4.1. Now, due to the coreference chains, we have five candidates: NP_1, NP_4 (and its antecedent NP_2), NP_5 (NP_3), NP_7 (NP_6) and NP_8.

Let us first consider the training process. In the tournament, the correct antecedent NP_5 (NP_3) must prevail over any of the other four candidates. We thus extract four training examples from the present case as illustrated in the figure. The class *right* denotes that the succeeding one of a given pair of candidates prevails against (i.e. is more likely to be the antecedent than) the preceding one. Likewise, the class *left* denotes that the preceding candidate prevails over the succeeding one. Finally, we induce from a set of extracted training examples a pair-wise classifier that classifies a given feature vector into either *right* or *left*.

In the test phase, the model conducts a tournament for each given anaphor. In each tournament, it processes the antecedent candidates in the right-to-left order. In the first round, the model consults the trained classifier to judge which of the right-most (closest th *ANP*) two candidates is more likely to be the antecedent. Suppose anew that we are trying to resolve the problem illustrated in Figure 4.1. As shown in the "test process" part of the figure, the first match is arranged between the right-most two candidates NP_8 and NP_7. Here, we assume that NP_8 wins as shown in the figure. Then, each of the following matches is arranged in turn between the winner of the previous match and a right-most new challenger. In the case shown in the figure, the second match is arranged between the current winner NP_8 and the right-most new challenger NP_5. If NP_5 wins, it is next matched against all next challenger NP_4. This process is repeated until the left-most candidate participates. The model selects the candidate that prevails through the final round as the answer.

The introduction of the pairwise classification as above can incorporate the learning of centering factors, such as the expected center order; for example, the model may learn from *Tom* and *John*'s example that the subject role filler is preferred to the object role filler. The tournament model can also encode relational properties between candidates into features. One may, for example, add a feature that indicates the relative distance between a given candidate pair, expecting a tendency that the succeeding candidate is more likely to win when the relative distance between two candidates is longer.

4.3. Experiments

We conducted an empirical evaluation of Japanese zero-anaphora resolution. Japanese is characterized by an extensive use of zero-pronouns, which behave like pronouns in English texts. Zero-anaphora resolution has been receiving interest from an increasing number of researchers (Kameyama, 1986; Nariyama, 2002; Nakaiwa and Shirai, 1996; Seki et al., 2002; Yamamoto and Sumita, 1998).

4.3.1 Models

In the experiments, we compare the tournament model with the following two baseline models. For the first baseline model, we create a rule-based model based on Nariyama (2002)'s algorithm (see Section 2.1.1). Note that the algorithm includes some factors which can not be simulated computationally, so we implements the model according to the following way:

1. if there exists a candidate antecedent that satisfies the patterns shown in Figure 2.2, return the candidate as an antecedent.

2. if the current SRL is not empty, return the most likely candidate in the SRL as an antecedent.

3. otherwise; return NULL.

For the second baseline model, we employ Ng and Cardie's search-based model as a baseline model. By comparing these approaches with the tournament model, one can measure the effects of the comparison between two candidates.

4.3.2 Training and test sets

We extracted training and test data sets from a corpus with GDA-tagged[1] newspaper articles, which is annotated with anaphoric relation tags as well as various syntactic and semantic tags. The corpus contains over 25,000 sentences with roughly 20,000 anaphoric relation tags annotated. In the experiment, we preliminarily restricted our experiments for resolving subject zero-anaphors, 2,155 instances in total, and conducted five-fold cross-validation on that data set.

4.3.3 Feature set

We used five types of features as summarized in Table 4.1: (i) grammatical, (ii) semantic, (iii) positional, (iv) heuristic and (v) centering features. The features of types (i) to (iv) are defined so as to simulate Ng and Cardie's feature set, except the following three features:

- LOG_LIKE: indicates the largest value among the log-likelihood coefficients (Dunning, 1993) of the pairs of a noun in the coreference chain including the candidate and the predicate of the anaphor. Those coefficients are calculated with about ten millions of NOUN-VERB pairs extracted from other corpora (Shimbunsha, 1990 2000; Shimbunsha, 1991 1999).

- SELECT_REST: indicates whether or not a candidate satisfies selectional restrictions in Nihongo Goi Taikei (Japanese Lexicon) (Ikehara et al., 1997).

- CHAIN_LENGTH: indicates the number of all the preceding nouns in the coreference chain including the candidate.

We also introduce ANIMACY feature as in Ng's feature set, because an animate noun tends to be salient. ANIMACY indicates whether or not the candidate is an animate noun. A noun is regarded as animate if the noun is classified as PERSON or ORGANIZATION by a named entity tagger or the noun is included in PERSON or ORGANIZATION class of Nihongo Goi Taikei (Ikehara et al., 1997).

[1] The GDA (Global Document Annotation (Hasida, 2002)) tag set is designed to be a standard tag set which allows machines to automatically recognize the semantic and pragmatic structures of documents.

To define centering features, we adopted a Japanese anaphora resolution model proposed by Nariyama (2002) as the underlying theory. Nariyama's method is an expansion of Kameyama's work on the application of Centering Theory to Japanese anaphora (Kameyama, 1986). Nariyama expanded the original forward-looking center list into Salience Reference List (SRL) in order to take into account broader contextual information from preceding sentences. Analogous to common centering models, in SRL, discourse entities are stored in the salience order: TOPIC (marked by *wa*-particle) > SUBJ (*ga*) > I_OBJ (*ni*) > D_OBJ (*o*) > OTHERS. In the experiment, we introduced two features, SRL_ORDER and SRL_ORDER_COMP, to reflect the SRL-related contextual factors. The definition of them is given in Table 4.1. Nariyama's method is also devised to deal with state transitions in complex sentences, which was originally not handled in Kameyama's model on Japanese. We partially implemented this extension as another feature, GA_REF, expecting the strong tendency of coreference that some conjunctives convey.

In the experiment, all the features are automatically computed with the help of the following NLP systems: the Japanese morphological analyzer *ChaSen* (Matsumoto et al., 2000), the Japanese dependency structure analyzer *CaboCha* (Kudo and Matsumoto, 2002), and the named entity chunker *Yanee* (Yamada et al., 2002).

4.3.4 Results

While Ng and Cardie used the C4.5 decision tree induction system, we adopted Support Vector Machines (Vapnik, 1998) for classifier induction because of their state-of-the-art performance and considerable generalization ability, which had been proven for various NLP tasks.

The results are shown in Figure 4.2 and Table 4.2. Figure 4.2 shows the learning curves of each model by altering the training data size, and Table 4.2 shows the results with all of the training instances. Table 4.2 shows the performance of each machine learning-based model is significantly better than Nariyama's rule-based model. In our implementation of her algorithm, the candidates in the SRL are not always ranked exactly according to the salient referent order list shown in Figure 2.1, because it is difficult to computationally separate global and local information among the topicalized subjects. This leads to inaccurate antecedent selection.

In Figure 4.2, we can see the positive effects for introducing the centering features by comparing the learning curves of BM+CF with BM, and TM+CF with TM. Likewise, the differences between BM and TM show that the introduction of the tournament model significantly improved the performance regardless of the size of training data. It indicates that comparing between two candidate antecedent is more efficient for identifying antecedent than the candidate-wise comparison.

One can also introduce the notion of decision confidence into the tournament model. With a good confidence measure, one can effectively improve precision just by slightly sacrificing recall. In case of the tournament model, the likelihood (i.e. the degree of confidence) that the decision for a match is correct can be heuristically estimated by, for example, the absolute value of the SVM classifier's discrimination function for the corresponding classification problem. The likelihood that the winner of a tournament is correct is then given by the confidence value of the closest match the winner have played. Given such a confidence measure, one can obtain a recall-precision curve by moving the threshold of confidence values. Working of this is shown in Figure 4.3, which presents the recall-precision curve obtained by testing this heuristic measure.

4.4. Summary

In this chapter, we presented a trainable coreference resolution model that is designed to incorporate contextual cues by means of centering features and a tournament-based search algorithm. These two improvements worked effectively in our experiments on Japanese zero-anaphora resolution.

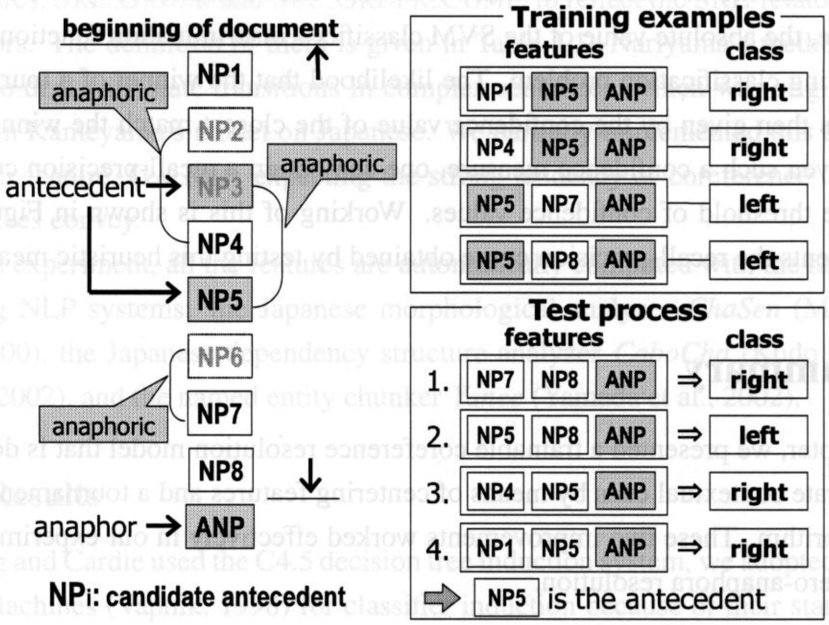


Figure 4.1. The tournament model

Table 4.1. Feature Set

Feature types	Feature names	Descriptions
Grammatical	POS	The part-of-speech of NP_i such as 'proper noun' and 'sahen noun'.
	DEFINITE	Y if NP_i is 'sore', 'soko', 'sono', 'sonna', etc; else N.
	DEMONSTRATIVE	Y if NP_i is 'kore', 'soko', 'ano', 'asoko', etc; else N.
	PARTICLE	The case marker attached to NP_i such as 'wa', 'ga' and 'o'.
Semantic	NE	Named entity class of NP_i: PERSON, ORGANIZATION, LOCATION, ARTIFACT, DATE, TIME, MONEY, PERCENT or N/A.
	EDR_HUMAN	Y if NP_i has the human attribute of EDR dictionary; else N.
	SELECT_REST	C if NP_i-ANP pair satisfies the selectional restriction; else I.
	LOG_LIKE	Five degree of the log-like coefficient of the NP_i-ANP pair.
	ANIMACY	Y if NP_i has the PERSON or ORGANIZATION class; else N.
	ANIMACY_COMP*	NP_1 if NP_1 has ANIMACY feature and NP_2 doesn't; else NP_2 if the opposite relation.
Positional	SENTNUM_ANP	Distance between NP_i and ANP in terms of sentences.
	SENTNUM_NPS*	Distance between NP_1 and NP_2 in terms of sentences.
	DEP_MAIN	Y if NP_i depends on the main clause; else N.
	EMBEDDED	Y if NP_i locates in an embedded clause; else N.
	BEGINNING	Y if NP_i locates in the beginning of the sentence; else N.
Heuristic	CHAIN_LENGTH	Length of a cohesive chain of NP_i
Centering	SRL_ORDER	The priority rank of NP_i in SRL.
	SRL_ORDER_COMP*	NP_1 if NP_1 is higher ranked than NP_2 in SRL; else NP_2
	GA_REF	Y if NP_i is the subject of a subordinate clause of a particuler conjunctive type and ANP is the subject of its matrix clause; else N.

ANP is an anaphor, and $NP_{i \in \{1,2\}}$ is an antecedent candidate. The feature set contains relational and non-relational features. Non-relational features test some property P of NP_i under consideration and take on a value of YES or NO depending on whether P holds. Relational features test whether some property P holds for the NP_1-NP_2 or NP_2-ANP pair under consideration and indicates whether the pair is COMPATIBLE or INCOMPATIBLE w.r.t. P; a value of NOT APPLICABLE is used when property P does not apply. Features with an asterisk are used only in the tournament model.

Table 4.2. The result with all of the training examples

Model	Accuracy
Nariyama (2002)'s rule-based model	45.6% (1269/2781)
Ng and Cardie (2002a)'s model (BM)	65.7% (1827/2781)
BM with centering features	69.0% (1918/2781)
The tournament model (TM)	74.3% (2065/2781)
TM with centering features	75.1% (2089/2781)

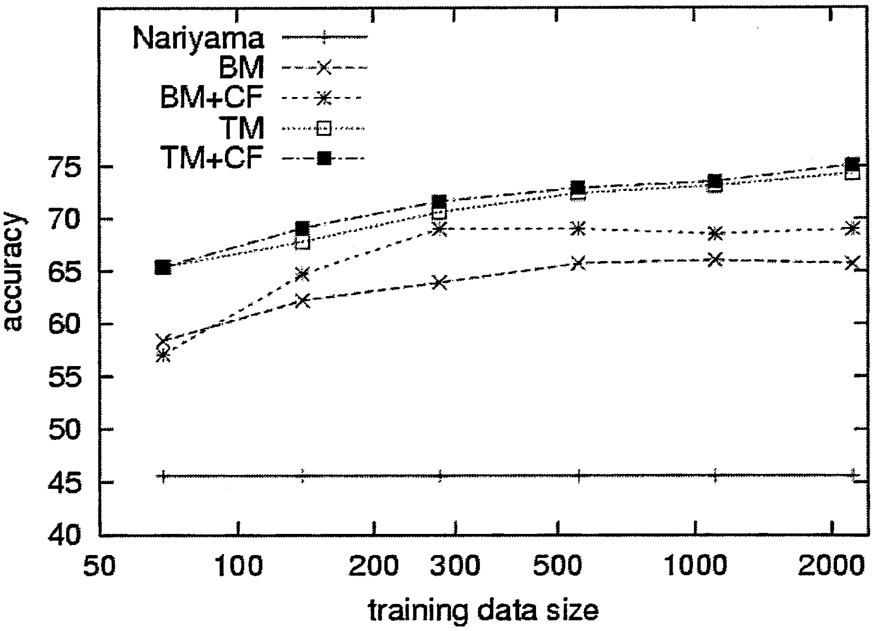


Figure 4.2. Learning curves

Nariyama: Nariyama's rule-based model, BM: Ng and Cardie's model, BM+CF: BM using centering features TM: Tournament model, and TM+CF: TM using centering features

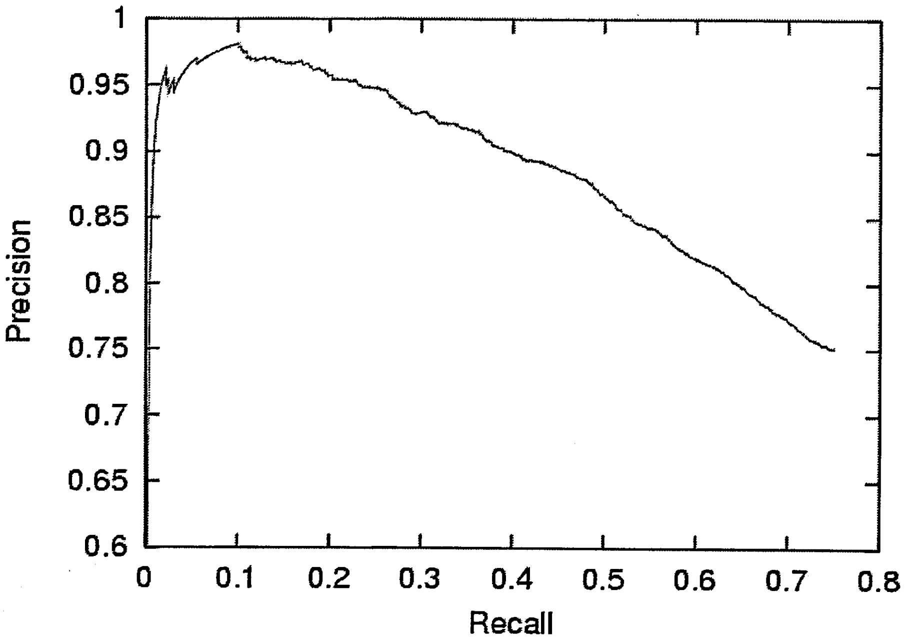


Figure 4.3. Precision-recall curve obtained with the tournament model

Chapter 5

Antecedent Identification Followed by Anaphoricity Determination

5.1. Introduction

Anaphora resolution can be decomposed into two subtasks: *anaphoricity determination* and *antecedent identification*. Anaphoricity determination is the task of classifying whether a given noun phrase (NP) is *anaphoric* or *non-anaphoric*. Here we say an NP is anaphoric if it has any antecedent (i.e. NP(s) that are anaphoric with it) in the context preceding it in the discourse, and non-anaphoric otherwise. The second task, antecedent identification, is identification of the antecedent(s) of a given anaphoric NP.

Early corpus-based work on anaphora resolution does not address anaphoricity determination; it assumes that the anaphora resolution system knows a priori all the anaphoric noun phrases. However, this problem has recently been given an increasing amount of attention (Bean and Riloff, 1999; Ng and Cardie, 2002b; Uryupina, 2003; Ng, 2004; Poesio et al., 2004) because:

- determining anaphoricity is not a trivial problem even in languages such as English and French, where definite articles can be used as clues (Ng and Cardie, 2002b), and

- the overall performance of anaphora resolution crucially depends on the accuracy of anaphoricity determination.

Obviously, the problem of anaphoricity determination is even more critical in the case of languages, such as Japanese, which do not have such clues as definite articles.

Previous efforts to tackle this problem have provided the following findings:

- One of the useful clues for determining the anaphoricity of a given NP can be obtained by searching for an antecedent. If an appropriate candidate for the antecedent is found in the preceding context of the discourse, the NP is likely to be anaphoric (Soon et al., 2001; Ng and Cardie, 2002a).
- Anaphoricity determination can be effectively carried out by a binary classifier that learns instances of non-anaphoric NPs as well as those of anaphoric NPs (Ng and Cardie, 2002b; Ng, 2004).

As we discuss in the next section, previous approaches to anaphora resolution (Ng and Cardie, 2002a; Ng and Cardie, 2002b; Iida et al., 2003) make use of a range of cues, but none of the previous models effectively combines from three previous approaches shown in Section 5.2. This leaves significant room for improvement in anaphora resolution.

In this chapter, we propose a machine learning-based model that effectively combines the sources of evidence used in existing models, while overcoming their drawbacks. We show the effectiveness of our approach through experiments on Japanese anaphora resolution comparing previous machine learning-based approaches including Ng and Cardie (2002a)'s search-based approach and Ng (2004)'s classification-then-search approach.

The rest of the chapter is organized as follows. In Section 5.2, we review previous machine learning-based approaches to anaphora resolution. Section 5.3 describes how the proposed model combines effectively advantages of each previous approach. We then report the results of our experiments on Japanese noun phrases anaphora resolution in Section 5.4 and conclude in Section 5.5.

5.2. Previous approaches

As reviewed in Section 2.2, previous learning-based methods for anaphora resolution can be classified into two approaches: the *search-based approach* and the *classification-*

Table 5.1. Advantages in each approach

	Search	Classification-then-search	Tournament
Use contextual clues?	√		√
Use non-anaphoric instances?		√	
Can determine anaphoricity?	√	√	
Balanced training instances?			√

based approach. We discuss their advantages and disadvantages below (see Table 5.1 for summary).

5.2.1 Search-based model

As described in Section 2.2.1, the search-based approaches have an advantage to deal with the broad contextual information. A flaw of this approach, on the other hand, is that models are not designed to learn non-anaphoric cases directly in the training phase. As an example, let us take a closer look at Soon et al.'s model (see Figure 2.4). For training, their model creates a positive instance from an anaphoric NP paired with its closest antecedent (NP_5-*ANP*) and a negative instance from each of the intervening NPs paired with the anaphor (NP_6-*ANP*, NP_7-*ANP* and NP_8-*ANP*). Note that no training instance is derived from non-anaphoric NPs. This drawback is shared also by other search-based models including (Ng and Cardie, 2002a; Yang et al., 2003). As we show in Section 5.4, this may well significantly degrade performance.

Another drawback of the approach is that it may suffer also from highly imbalanced distributions of positive and negative instances. The aforementioned method of generating training instances tends to generate much more negative instances than positive ones. For example, in the experiments described in Section 5.4, the ratio of the positive instances to the negative instances is 1 to 22. The model requires proper selection of training instances (Ng and Cardie, 2002c). However, it is not a trivial problem.

5.2.2 Classification-then-search model

As reported in Ng and Cardie (2002b) and also in Section 5.4 of this chapter, this model significantly outperforms the search-based model. However, it still has several

Table 5.2. Partial feature list relevant to the larger context information used in Ng and Cardie [2002b]'s model.

Feature Type	Feature	Description
Lexical	STR_MATCH	Y if there exists an NP NP_i preceding NP_j such that, after discarding determiners, NP_i and NP_j are the same string; else N.
	HEAD_MATCH	Y if there exists an NP NP_i preceding NP_j such that NP_i and NP_j have the same head; else N.
Semantic	ALIAS	Y i there exists an NP NP_i preceding NP_j such that NP_i and NP_j are aliases; else N.
	SUBCLASS	Y if there exists an NP NP_i preceding NP_j such that NP_i and NP_j have an ancestor-descendant relationship in WordNet; else N.

NP_i and NP_j indicate a candidate anaphor and a candidate antecedent respectively.

drawbacks and room for improvement.

First, Ng and Cardie (2002b) reports that the performance of the anaphoricity determination component is so low that applying it would not improve the performance of the overall task unless it incorporated features that effectively capture contextual information (see Table 5.2). This indicates that it is crucially important in anaphoricity determination to know whether or not the preceding context of the discourse contains NPs that are likely to be the antecedent of a current target NP. While such features as in Table 5.2 appear to be useful clues for this reason, they appear to be rather *ad hoc* and only provide an extremely rough summary of the context.

Second, in the classification-then-search model, not only the anaphoricity classifier but also the antecedent identification component takes charge of anaphoricity determination. This rather unclear way of division of labor constrains the range of algorithms that can be used for antecedent identification. The model cannot employ such a model as, for example, the tournament model, which we review below.

Third, as long as it employs such an algorithm as Ng and Cardie (2002a) for the antecedent identification subtask, the model inherits the drawbacks of the algorithm; in particular, it is important to note the problem of imbalanced distribution of positive and negative training instances.

5.2.3 Tournament model

For the task of antecedent identification alone, it is worth referring to a model called the tournament model proposed in Chapter 4. The model conducts a tournament consisting of a series of matches in which candidate antecedents compete with each other for a given anaphor. In the tournament, it processes the candidate antecedents in the right-to-left order. In the first round, the model consults a trained classifier to judge which of the right-most two candidates is more likely to be the antecedent for the anaphor. The winner then plays a match with the third right-most candidate. Likewise, each of the following matches is arranged in turn between the current winner and a right-most new challenger until the left-most candidate antecedent. The model selects the winner of tournament.

This model has several advantages over such previous antecedent identification models as reviewed in Section 2.2.1. First, it can incorporate the learning of some of centering factors, such as the expected center order, proposed in Centering Theory (Grosz et al., 1995). Second, unlike the previous models, the task of the classifier is to determine which of a pair of candidates is more likely to be the antecedent. This way of task decomposition inherently avoid the problem of imbalanced distributions of positive and negative instances which such a model as Soon et al. (2001) and Ng and Cardie (2002a, 2002b) would suffer from. Due to these advantages, Iida et al. (2003) report that the tournament model outperforms the Ng and Cardie (2002a)'s model in Japanese zero-anaphora resolution.

Despite these advantages, however, the tournament model has a strict limitation; namely, it is not capable of anaphoricity determination because it always select a candidate antecedent for a given NP whether the NP is anaphoric or not.

5.3. Selection-then-classification approach

This section discusses how to design an anaphora resolution model that inherits all the advantages of the previous models reviewed in the last section.

We explore an alternative way of incorporating contextual clues into anaphoricity determination. One way that has not yet been examined before is to implement an anaphora resolution process that reverses the steps of the classification-then-search model. Assuming that we have an antecedent identification model and an anaphoricity

```
Function Select-Antecedent-by-Tournament ( Ana: candidate anaphor,
                                           C: set of candidate antecedents )
    SC := sort_by_reverce_order_of_ appearance C;
    Max_Ant := $SC_1$; // the right-most candidate in SC
    SC := SC \ $SC_1$;
    for $i = 2, \ldots, n$ do
        // select which candidate is anaphoric with Ana
        Score := compare_antecedenthood ( Ana, $SC_i$, Max_Ant );
        if $Score > 0$ then
            Max_Ant := $SC_i$;
        end
    end
    return Max_Ant;
end
```

Figure 5.1. The tournament model

classification model, the new model processes each target noun phrase (*TNP*) in a given text in two steps (see Figure 5.7):

1. Select the *most likely candidate antecedent CA* (*NP_2* in Figure 5.7) for *TNP* using an antecedent identification model.

2. Classify *TNP* paired with *CA* as either *anaphoric* or *non-anaphoric* using an anaphoricity classification model. If pair *CA-TNP* is classified as *anaphoric*, *CA* is identified as the antecedent of *TNP*; otherwise, *TNP* is judged *non-anaphoric*.

To bring the contrast with the classification-then-search model, we call this model the *selection-then-classification model*.

To implement this new model, we extend a anaphoricity determination component designed in the classification-based approach so that the model determines whether a given NP paired with its most likely candidate antecedent is anaphoric or not. For training the classifier, we create a positive (anaphoric) and negative (non-anaphoric) training sets in the following way:

(i) For each NP appearing in the training corpus, we add the pair of the NP and its corresponding antecedent to the positive (anaphoric) training set if the NP is

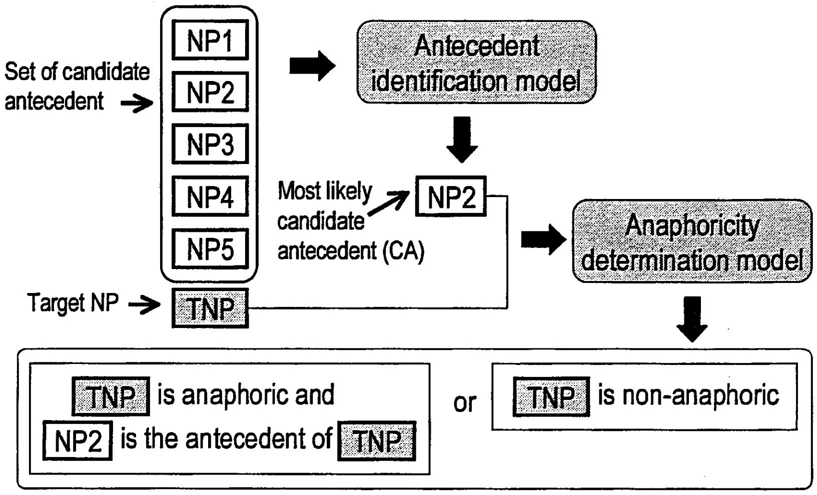

Set of candidate antecedent
NP1
NP2
NP3
NP4
NP5
Target NP
TNP
Antecedent identification model
Most likely candidate antecedent (CA)
NP2
Anaphoricity determination model
TNP is anaphoric and NP2 is the antecedent of TNP
or
TNP is non-anaphoric

Figure 5.2. Process of NP-anaphora resolution.

anaphoric. This is illustrated in the upper part of Figure 5.3, where an anaphoric noun phrase *ANP* paired with its antecedent NP_4 is added to the set of anaphoric instances.

(ii) If the NP is non-anaphoric, we first use the antecedent identification model that we employ in the antecedent identification step to select the most likely candidate antecedent for the NP, which we call the *pseudo-antecedent* of the NP. We then add the pair of the NP and its pseudo-antecedent to the negative training set. In the case of Figure 5.3 (the lower part of the figure), where we have a non-anaphoric noun phrase (*NANP*), we first select its most likely candidate antecedent NP_3 from candidate antecedents NP_1 through NP_5, and then add the pair NP_3-*NANP* to the non-anaphoric training set.

Provided an anaphoric and non-anaphoric training sets, we can use a wide range of classifier induction algorithms.

The new model might not look strongly different from such previous models as the classification-then-search model. However, the model in fact effectively combines the advantages of all the previous models we reviewed in Section 2.2.1.

First, the new model inherits the advantage of the search-based model. It determines the anaphoricity of a given NP taking into account the information of its most likely candidate antecedent. The candidate antecedent selected in the first step can

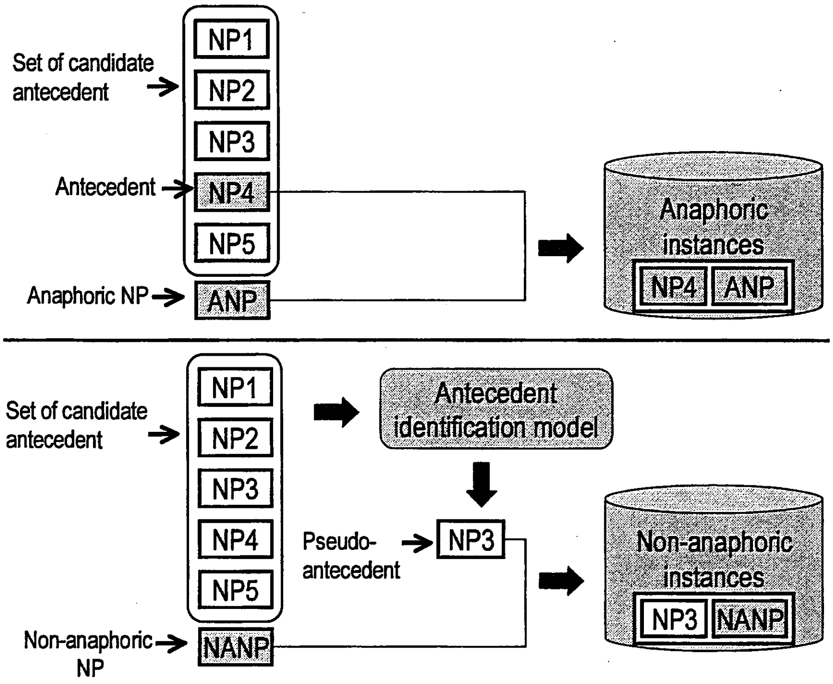


Figure 5.3. Training data collection for the anaphoricity determination model.

be expected to provide contextual information useful for anaphoricity determination; if the best candidate does not appear to be the real antecedent of the target NP, it is unlikely that the target NP has any antecedent in the discourse. In this respect, the proposed model makes better use of contextual clues than the classification-then-search model, which accesses to contextual information only through ad hoc string-based features.

Second, the proposed model uses non-anaphoric instances together with anaphoric instances to induce an anaphoricity classifier, which is an important advantage inherited from the classification-then-search model.

Third, in the proposed model, the division of labor between the two components is clearer than that in the selection-then-classification model. The antecedent identification component always selects a candidate antecedent for a given NP (i.e. candidate anaphor) whether the NP is anaphoric or not. This way of task decomposition allows us to employ the tournament model in antecedent identification (see Figure 5.4). Recall that in the classification-then-search model, the anaphoricity determination component

```
Function Select-Antecedent-and-Classify-Anaphor ( Ana: candidate anaphor,
                                                  C: set of candidate antecedents )
    Max_Ant := Select-Antecedent-by-Tournament ( Ana, C );
    // judge whether or not Ana is anaphoric with Max_Ant
    Score := classifier-anaphoricity ( Ana, Max_Ant );
    if ( $Score > \theta_{ana}$ ) then
        return Max_Ant;
    else
        return NULL;
    end
end
```

θ_{ana} is a global variable that indicates a global threshold parameter of annaphoricity.

Figure 5.4. The slection-then-classification model

is not reliable enough to entirely free the antecedent identification component from the charge of anaphoricity determination. This deficiency prohibits the model from incorporating the tournament model. As we report in Section 5.4.4, this gives a significant advantage to the new model.

5.4. Experiments on NP-anaphora resolution

We conducted an empirical evaluation of our method by applying it to Japanese newspaper articles. In the experiments, we compared three models: the search-based model, the classification-then-search model and the selection-then-classification model.

5.4.1 Models

For the search-based model, we created a model designed to simulate the model described in (Ng and Cardie, 2002a). Pseudocode describing the model is given in Figure 2.3 (see Section 2.2.1). We employed Support Vector Machines (Vapnik, 1998) for learning and used the distance between an input feature vector and the hyperplane as the score for classification.

For the classification-then-search model, we created a model based on the pseudocode given in Figure 2.5 (see Section 2.2.2). In these experiments, instead of preparing the development data for the estimation of two thresholds, we evaluated the performance by fine-tuning these thresholds by hand. In addition to the original classification-then-search model, we also implemented the model using the tournament model for the antecedent identification model instead of the search-based model. Thus, we can investigate whether or not the tournament model improves the classification-then-search model.

Regarding the selection-then-classification model, we implemented the model based on the process in Figure 5.4.

In addition to the original selection-then-classification model, we also implemented a model using the search-based model for the antecedent identification model instead of the tournament model. Thus, we can evaluate the effectiveness of the tournament model itself by comparing the two selection-then-classification models.

Like the search-based model, the classification-then-search model and the selection-then-classification model also used SVMs for both antecedent identification and anaphoricity classification.

5.4.2 Training and test instances

We created a coreference-tagged corpus consisting of 90 newspaper articles (1,104 sentences). The corpus contained 884 anaphoric NPs and 6,591 non-anaphoric NPs (7,475 NPs in total), each anaphoric NP being annotated with information indicating its antecedent. For each experiment, we conducted ten-fold cross-validation over 7,475 noun phrases so that the set of the noun phrases from a single text was not divided into the training and test sets.

5.4.3 Feature sets

We used the following five types of features:

- *ANA*: Features designed to capture the lexical, syntactic, semantic and positional information of a target noun phrase (i.e. a candidate anaphor)

- *ANT*: Features designed to capture the lexical, syntactic, semantic and positional information of a candidate antecedent

Table 5.3. Features used in each model

	SM	CSM	SCM	
			Antecedent identification	Anaphoricity determination
ANA	√	√	√	√
ANT	√		√	√
ANA-ANT	√		√	√
ANT_SET		√		
ANT-ANT			√	

SM: the search-based model, CSM: the classification-then-search model, and SCM: the selection-then-classification model.

- *ANA-ANT*: Features designed to capture the relation between the candidate antecedent and the target NP (e.g., the distance, semantic compatibility between the two)
- *ANT-ANT*: Features designed to capture the relation between two candidate antecedents (e.g. the distance between the two)
- *ANT_SET*: Features designed to capture the relation between the set of the candidate antecedents in the preceding context and the target NP (e.g., the binary feature that a target NP and an candidate antecedent in the preceding context contain the same string)

The features of the types *ANA*, *ANT* and *ANA-ANT* cover the feature set that Ng and Cardie (2002a) used in their search-based model. On the other hand, the *ANT-ANT* type of features were those that cannot be used in the search-based model but only in the tournament model because the search-based model refers only to a single candidate antecedent at the time of classification. The *ANT_SET* type of features is based on the feature set in Ng and Cardie's work (Ng and Cardie, 2002b). Table 5.3 summarizes which types of features were used for each model. Table 5.4, Table 5.5 and Table 5.6 present the details of the feature set.

In the experiment, all the features were automatically computed with the help of publicly available NLP tools, the Japanese morphological analyzer *ChaSen* (Mat-

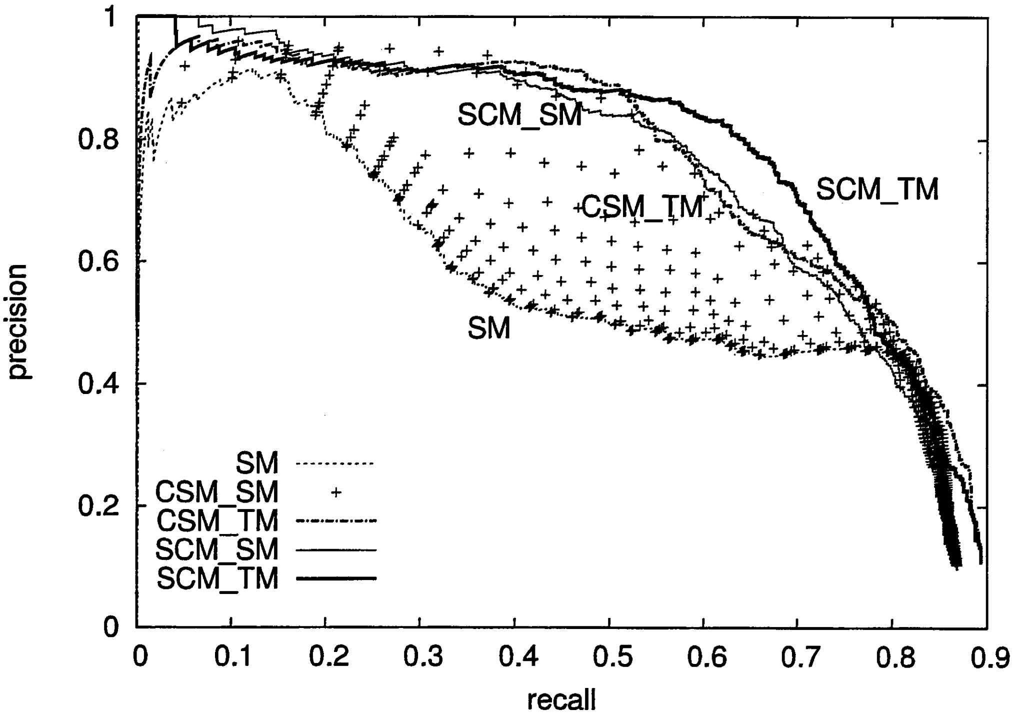


Figure 5.5. Recall-precision curves in NP-anaphora resolution
SM: the search-based model, CSM: the classification-then-search model, SCM_SM: the selection-then-classification model using the search-based model, and SCM_TM: the selection-then-classification model using the tournament model.

sumoto et al., 2000) and the Japanese dependency structure analyzer *CaboCha* (Kudo and Matsumoto, 2002), which also performed named-entity chunking.

5.4.4 Results

To compare the performance of the three models on the task of anaphora resolution, we plot a recall-precision curve for each model as shown in Figure 5.5 by altering threshold parameter θ_{ana} (and θ_{ant} in the case of the classification-then-search model using the search-based model (CSM_SM)), where recall R and precision P are calculated by:

$$R = \frac{\text{\# of detected anaphoric relations correctly}}{\text{\# of anaphoric NPs}},$$
$$P = \frac{\text{\# of detected anaphoric relations correctly}}{\text{\# of NPs classified as anaphoric}}.$$

Note that the curves of the classification-then-search model using the search-based model (CSM_SM) are plotted by altering two threshold parameters θ_{ana} and θ_{ant}. The curves indicate the upperbound of the performance of CSM_SM because in practical settings, these two parameters would have to be trained beforehand.

For the SCM algorithm, we implemented two models. One model employed SM for antecedent identification (SCM_SM) and the other employed the tournament model (SCM_TM).

The comparison between the search-based model and the classification-then-search model supports Ng and Cardie (2002b)'s claim that incorporating the anaphoricity classification process into the search-based model can improve the performance if the threshold parameters are appropriately selected.

By comparing the selection-then-classification model using the search-based model (SCM_SM) with the classification-then-search model using the search-based model (CSM_SM), one can measure the effects of using the most likely antecedent while preserving the advantage of referring to the non-anaphoric information. The performance of the SCM_SM approached the upper bound of the performance of the CSM_SM. Recall that the CSM_SM algorithm requires the two inter-dependent threshold parameters to be trained beforehand while the proposed model need to tune only one parameter. We consider it as an important advantage of the proposed model. This advantage comes from the design of the proposed model, where the model makes use of anaphoric/non-anaphoric training instances as well as contextual clues given by most likely candidate antecedents simultaneously in the anaphoricity determination phase.

The results also indicate that even if the parameters for CSM_SM are optimally tuned, the proposed model significantly outperforms it when it employs the tournament model for antecedent identification (i.e. SCM_TM). The performance of the search-based model (SM) and the tournament model (TM) for antecedent identification alone is compared in Table 5.7. The table shows that TM outperforms SM by 2.5 points

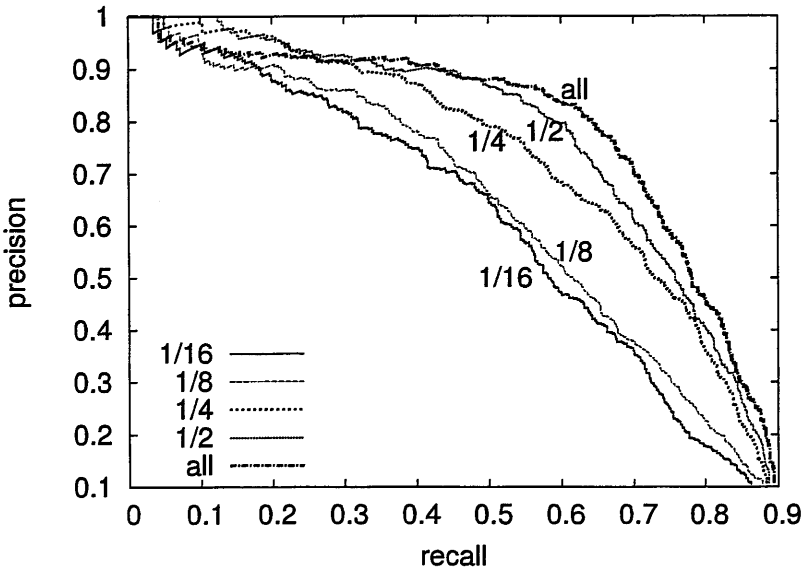


Figure 5.6. Change of recall-precision curves in NP-anaphora resolution

in accuracy. This difference is clearly reflected in the difference between SCM_TM and the SCM_SM. This is also an important advantage of the proposed model because previous model such as Ng (2004) cannot employ the tournament model as we noted in Section 2.2.2.

By comparing the selection-then-classification model using the tournament model (SCM_TM) with the classification-then-search model using the tournament model (CSM_TM), we can see whether or not the tournament model improves the CSM_TM. The results show that even if the tournament model is incorporated into the classification-then-search model, the SCM_TM still outperforms it.

Next, we evaluate the change in the performance for the different training data size on antecedent identification task, which is shown in Table 5.8. The results indicate that increasing training data size has little effect on antecedent identification. We also evaluate the transition of each recall-precision curve as shown in Figure 5.6 by altering training data size in NP-anaphora resolution including anaphoricity determination. Figure 5.6 shows that the performance is clearly improved by increasing data size and it may lead us to further improvement if the additional training instances are available for learning the anaphoricity determination model.

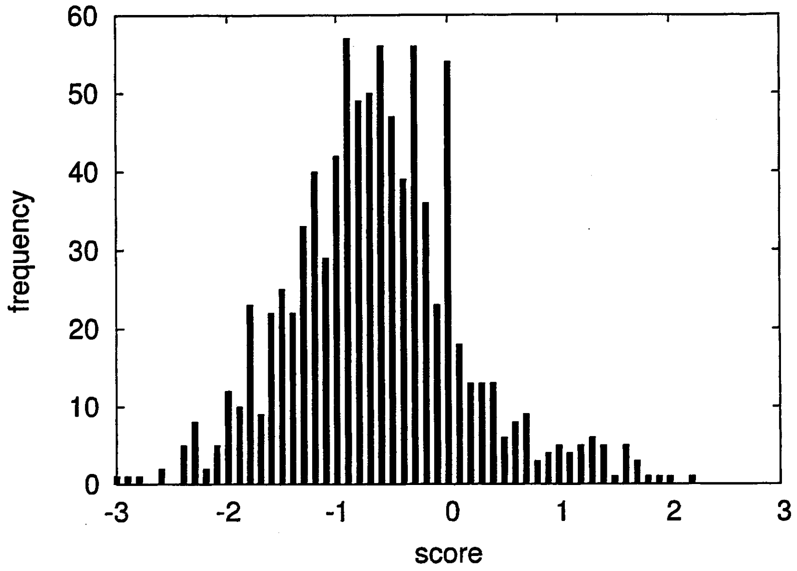


Figure 5.7. Distribution of anaphoricity determination score without correct antecedents.

Finally, we examine the behavior of our anaphoricity determination model. At the second step in the selection-then-classification model, in order to see which clue is more important, local context information such as contextual information around the target anaphor or preceding context information such as most likely antecedents, the model determines anaphoricity after eliminating correct antecedents for a given anaphor. In this experiment, if the model judges an input example as anaphoric, local context information is more important than preceding context information; otherwise preceding context information is preferable to local context information. Figure 5.7 illustrates the distribution of digitized values of the anaphoricity determination model. Note that the positive value supports that a candidate anaphor is anaphoric rather than non-anaphoric. Figure 5.7 shows that the peak of distribution is near -1, that is, it indicates that our model determines anaphoricity by utilizing preceding context information as important clues rather than local context information.

5.4.5 Discussion

According to our error analysis, a majority of errors are caused by the difficulty of judging the semantic compatibility between a candidate anaphor and candidate antecedent. For example, the lexical resources we employed in the experiments did not contain gender information; the model did not know that "*ani* (elder brother)" was semantically incompatible with "*kanojo* (she)" and thus could not be an antecedent of it. This raises an interesting issue, namely how to develop a lexical resource which includes a broad range of semantically compatible relations between nouns; for example, the model needs to know that *Russia* can be an antecedent of *Russian government*, but *president* is not compatible with *yesterday*. One of our future directions should aim at this issue.

There is also still room for improvement in the architecture of the proposed model. The model could make better use of the semantic information of candidate antecedents if it referred also to ancestors of coreference chains. For example, if a named-entity expression is referred to by such a word as "*dousha* (the/this company)" in the preceding context, we can enrich the coreference-chain information about by combining the relevant information from each noun phrase. This line of refinement will also lead us to explore methods to search for a globally optimal solution to a set of anaphora resolution problems for a given text, as discussed by McCallum and Wellner (McCallum and Wellner, 2003).

5.5. Summary

In this chapter, we reported that our selection-then-classification approach to anaphora resolution improves the performance of the previous learning-based models by combining their advantages, while overcoming their drawbacks. It does so in the following two respects:

(i) our model uses non-anaphoric instances together with anaphoric instances to induce an anaphoricity classifier, retaining the advantage inherited from the classification-based approach and

(ii) our model determines the anaphoricity of a given NP taking the information of its most likely candidate antecedent into account. Our argument has been supported by empirical evidence obtained from our experiment on Japanese NP-anaphora resolution.

Analogous to NP-anaphora resolution, zero-anaphora resolution also deals with the issue of anaphoricity determination. Motivated by this parallelism between NPs and zero-anaphora, in future work, we want to attempt anaphoricity determination for zero pronouns using the selection-then-classification approach proposed here.

Table 5.4. Feature set used in our experiments (1/3).

Feature Type	Feature	Description
Lexical	BF_COMB $_{AT}$	Combination of two characters of right-most morpheme in *ANP* and NP_i.
	DOU_MATCH $_{AT}$	1 if *ANP* contains the word "*dou* (i.e. same)" and the string of NP_i matches the *ANP* except for the word "*dou*"; otherwise 0.
	DOU_MATCH_SET $_{TS}$	1 if an NP_i preceding *ANP* exists such that *ANP* contains the word "*dou* (i.e. same)" and the string of NP_i matches the *ANP* except for the word "*dou*"; otherwise 0.
	FIRST_PERSON_MATCH $_{AT}$	1 if *ANP* and NP_i are classified as "Person" named entity class and *ANP* and NP_i share the same string; otherwise 0.
	FIRST_PERSON_MATCH_SET $_{TS}$	1 if an NP_i preceding *ANP* exists such that *ANP* and NP_i are classified as "Person" named entity class and *ANP* and NP_i share the same string; otherwise 0.
	FULL_MATCH $_{AT}$	1 if *ANP* and NP_i share the same string; otherwise 0.
	FULL_MATCH_SET $_{TS}$	1 if an NP_i preceding *ANP* exists such that *ANP* and NP_i share the same string; otherwise 0.
	FINAL_MATCH $_{AT}$	1 if *ANP* and NP_i share the same string-final morpheme; otherwise 0.
	FINAL_MATCH_SET $_{TS}$	1 if an NP_i preceding *ANP* exists such that *ANP* and NP_i share the same string-final morpheme; otherwise 0.
	FIRST_MATCH $_{AT}$	1 if *ANP* and NP_i share the same first morpheme; otherwise 0.
	FIRST_MATCH_SET $_{TS}$	1 if an NP_i preceding *ANP* exists such that *ANP* and NP_i share the same first morpheme; otherwise 0.
	PART_MATCH $_{AT}$	1 if *ANP* and NP_i share the same morpheme; otherwise 0.
	PART_MATCH_SET $_{TS}$	1 if an NP_i preceding *ANP* exists such that *ANP* and NP_i share the same morpheme; otherwise 0.

ANP indicates an anaphor, and $NP_{i\in\{1,2\}}$ indicates a candidate antecedent. '*'-ed features are used only in the experiments of antecedent identification. '*A*', '*T*', '*AT*', '*TS*' and '*TT*' indicate *ANA*, *ANT*, *ANA-ANT*, *ANT_SET*, and *ANT-ANT* features respectively.

Table 5.5. Feature set used in our experiments (2/3).

Feature Type	Feature	Description
Lexical	FINAL_INCUDED_MATCH $_{AT}$	1 if *NP*$_i$ and *ANP* share the same string-final morpheme and characters of *ANP* are included in *NP*$_i$; otherwise 0.
	FINAL_INCUDED_MATCH_SET $_{TS}$	1 if an *NP*$_i$ preceding *ANP* exists such that *NP*$_i$ and *ANP* share the same string-final morpheme and characters of *ANP* are included in *NP*$_i$; otherwise 0.
	FIRST_INCUDED_MATCH $_{AT}$	1 if *NP*$_i$ and *ANP* share the same first morpheme and characters of *ANP* are included in *NP*$_i$; otherwise 0.
	FIRST_INCUDED_MATCH_SET $_{TS}$	1 if an *NP*$_i$ preceding *ANP* exists such that *NP*$_i$ and *ANP* share the same first morpheme and characters of *ANP* are included in *NP*$_i$; otherwise 0.
	STRING_MATCH $_{AT}$	1 if morphemes in *ANP*$_i$ are included in *NP*$_i$ in the same order; otherwise 0.
	STRING_MATCH_SET $_{TS}$	1 if an *NP*$_i$ preceding *ANP* exists such that morphemes in *ANP*$_i$ are included in *NP*$_i$ in the same order; otherwise 0.

Table 5.6. Feature set used in our experiments (3/3).

Feature Type	Feature	Description
Grammatical	POS $_{A,\ T}$	Part-of-sppech of NP_i (*ANP*) followed by IPADIC [1].
	DEFINITE $_{A,\ T}$	1 if NP_i (*ANP*) contains the article corresponding to DEFINITE "the", such as "*sore*" or "*sono*"; otherwise 0.
	DEMONSTRATIVE $_{A,\ T}$	1 if NP_i (*ANP*) contains the article corresponeding to DEMONSTRATIVE "that" or "this", such as "*kono*", "*ano*"; otherwise 0.
	PARTICLE $_{A,\ T}$	Particle followed by NP_i (*ANP*), such as "*wa* (topic)", "*ga* (subject)", "*o* (object)".
	DOU $_{A,\ T}$	1 if NP_i (*ANP*) contains the word "*dou* (same)"; otherwise 0.
	DEP_PAST* $_{A,\ T}$	1 if some predicate (past form) depends on NP_i (*ANP*); otherwise 0.
	DEP_PRED* $_{A,\ T}$	1 if some predicate (not past form) depends on NP_i (*ANP*); otherwise 0.
Semantic	NE $_{A,\ T}$	Named entity of NP_i (*ANP*): PERSON, ORGANIZATION, LOCATION, ARTIFACT, DATE, TIME, MONEY, PERCENT or N/A.
	EDR_HUMAN $_{A,\ T}$	1 if NP_i (*ANP*) is included among the concept "a human being" or "atribute of a human being" in EDR dictionary; otherwise 0.
	EDR_AGENT $_{A,\ T}$	NP_i (*ANP*) is included among the concept "agent" in EDR dictionay; otherwise 0.
	PRONOUN_TYPE $_{A,\ T}$	Pronoun type of NP_i (*ANP*). (e.g. "*kare* (he)" → PERSON, "*koko* (here)" → LOCATION, "*sore* (this)" → OTHERS)
	SEM_PATH $_{AT}$	Depth of the lowest (most specific) common node between *ANP* and *NP* in Japanese thesaurus *Bunrui Goi Hyo* (Natural Language Research Institute, 1964) .
Positional	SENTNUM_ANP $_{AT}$	Distance between NP_i and *ANP*.
	SENTNUM_NPS* $_{TT}$	Distance between NP_1 and NP_2.
	BEGINNING $_{T,\ A}$	1 if NP_i (*ANP*) is located in the beggining of sentence; otherwise 0.
	END $_{A,\ T}$	1 if NP_i (*ANP*) is located in the end of sentence; otherwise 0.
	DEP_NE* $_{A,\ T}$	1 if NP_i (*ANP*) has the modifier "NAMED ENTITY+no (of)"; otherwise 0.
	DEP_NO* $_{A,\ T}$	1 if NP_i (*ANP*) has the modifier "*no* (of)"; otherwise 0.
	DEP_ANA $_{AT}$	1 if NP_i depends on *ANP*; otherwise 0.

Table 5.7. Result in the experiments of antecedent identification

	Search-based model	Tournament model
Accuracy	86.9% (768/884)	89.4% (790/884)

Table 5.8. Effects of altering training instances on antecedent identification in NP-anaphora resolution

training data	accuracy
1/16	0.862 (762/884)
1/8	0.878 (776/884)
1/4	0.883 (781/884)
1/2	0.887 (784/884)
1/1	0.894 (790/884)

Chapter 6

Exploitation of Syntactic Pattern Features

6.1. Introduction

Recent work on zero-anaphora resolution can be located in two different research contexts. First, zero-anaphora resolution is studied in the context of anaphora resolution (AR), in which zero-anaphora is regarded as a subclass of anaphora. In AR, the research trend has been shifting from rule-based approaches (Baldwin, 1995; Lappin and Leass, 1994; Mitkov, 1997, etc.) to empirical, or corpus-based, approaches (McCarthy and Lehnert, 1995; Ng and Cardie, 2002a; Soon et al., 2001; Strube and Müller, 2003; Yang et al., 2003) because the latter are shown to be a cost-efficient solution achieving a performance that is comparable to best performing rule-based systems (see the Coreference task in MUC[1] and the Entity Detection and Tracking task in the ACE program[2]). The same trend is observed also in Japanese zero-anaphora resolution, where the findings made in rule-based or theory-oriented work (Kameyama, 1986; Nakaiwa and Shirai, 1996; Okumura and Tamura, 1996, etc.) have been successfully incorporated in machine learning-based frameworks (Seki et al., 2002; Iida et al., 2003).

Second, the task of zero-anaphora resolution has some overlap with PropBank[3]-style semantic role labeling (SRL), which has been intensively studied, for example, in

[1] http://www-nlpir.nist.gov/related_projects/muc/

[2] http://projects.ldc.upenn.edu/ace/

[3] http://www.cis.upenn.edu/~mpalmer/project_pages/ACE.htm

the context of the CoNLL SRL task[4]. In this task, given a sentence "*To attract younger listeners, Radio Free Europe intersperses the latest in Western rock groups*", an SRL model is asked to identify the NP *Radio Free Europe* as the A0 (Agent) argument of the verb *attract*. This can be seen as the task of finding the zero-anaphoric relationship between a nominal gap (the A0 argument of *attract*) and its antecedent (*Radio Free Europe*) under the condition that the gap and its antecedent appear in the same sentence.

In spite of this overlap between AR and SRL, there are some important findings that are yet to be exchanged between them, partly because the two fields have been evolving somewhat independently. The AR community has recently made two important findings:

- A model that identifies the antecedent of an anaphor by a series of comparisons between candidate antecedents has a remarkable advantage over a model that estimates the absolute likelihood of each candidate independently of other candidates (Iida et al., 2003; Yang et al., 2003).
- An AR model that carries out antecedent identification *before* anaphoricity determination, the decision whether a given NP is anaphoric or not (i.e. discourse-new), significantly outperforms a model that executes those subtasks in the reverse order or simultaneously (Poesio et al., 2004; Iida et al., 2005).

To our best knowledge, however, existing SRL models do not exploit these advantages. In SRL, on the other hand, it is common to use syntactic features derived from the parse tree of a given input sentence for argument identification. A typical syntactic feature is the path on a parse tree from a target predicate to a noun phrase in question (Gildea and Jurafsky, 2002; Carreras and Marquez, 2005). However, existing AR models deal with intra- and inter-sentential anaphoric relations in a uniform manner; that is, they do not use as rich syntactic features as state-of-the-art SRL models do, even in finding intra-sentential anaphoric relations. We believe that the AR and SRL communities can learn more from each other.

Given this background, in this chapter, we show that combining the aforementioned techniques derived from each research trend makes significant impact on zero-

[4] http://www.lsi.upc.edu/~srlconll/

anaphora resolution, taking Japanese as a target language. More specifically, we demonstrate the following:

- Incorporating rich syntactic features in a state-of-the-art AR model dramatically improves the accuracy of intra-sentential zero-anaphora resolution, which consequently improves the overall performance of zero-anaphora resolution. This is to be considered as a contribution to AR research.
- Analogously to inter-sentential anaphora, decomposing the antecedent identification task into a series of comparisons between candidate antecedents works remarkably well also in intra-sentential zero-anaphora resolution. We hope this finding to be adopted in SRL.

The rest of the chapter is organized as follows. Section 6.2 describes the task definition of zero-anaphora resolution in Japanese. Section 6.3 described how the proposed model incorporates effectively syntactic features into the machine learning-based approach. We then report the results of our experiments on Japanese zero-anaphora resolution in Section 6.4 and conclude in Section 6.5.

6.2. Zero-anaphora resolution

In this chapter, we consider only zero-pronouns that function as an obligatory argument of a predicate for two reasons:

- Providing a clear definition of zero-pronouns appearing in adjunctive argument positions involves awkward problems, which we believe should be postponed until obligatory zero-anaphora is well studied.
- Resolving obligatory zero-anaphora tends to be more important than adjunctive zero-pronouns in actual applications.

A zero-pronoun may have its antecedent in the discourse; in this case, we say the zero-pronoun is *anaphoric*. On the other hand, a zero-pronoun whose referent does not explicitly appear in the discourse is called a *non-anaphoric* zero-pronoun. A zero-pronoun may be non-anaphoric typically when it refers to an extralinguistic entity (e.g. the first or second person) or its referent is unspecified in the context.

The following are Japanese examples. In sentence (18), zero-pronoun ϕ_i is anaphoric as its antecedent, '*shusho* (prime minister)', appears in the same sentence. In sentence (19), on the other hand, ϕ_j is considered non-anaphoric if its referent (i.e. the first person) does not appear in the discourse.

(18) *shusho$_i$-wa* *houbeisi-te* ,
prime minister$_i$-TOP visit-U.S.-CONJ PUNC
ryoukoku-no *gaikou-o*
both countries-BETWEEN diplomacy-ACC
(*ϕ_i-ga*) *suishinsuru*
(ϕ_i-NOM) promote-ADNOM
houshin-o *akirakanisi-ta* .
plan-OBJ unveil-PAST PUNC
The prime minister visited the united states and unveiled the plan to push diplomacy between the two countries.

(19) (*ϕ_j-ga*) *ie-ni* *kaeri-tai* .
(ϕ_j-NOM) home-DAT want to go back PUNC
(I) want to go home.

Given this distinction, we consider the task of zero-anaphora resolution as the combination of two sub-problems, antecedent identification and anaphoricity determination, which is analogous to NP-anaphora resolution:

> For each zero-pronoun in a given discourse, find its antecedent if it is anaphoric; otherwise, conclude it to be non-anaphoric.

6.3. Proposal

6.3.1 Task decomposition

We approach the zero-anaphora resolution problem by decomposing it into two subtasks: intra-sentential and inter-sentential zero-anaphora resolution. For the former problem, syntactic patterns in which zero-pronouns and their antecedents appear may well be useful clues, which, however, does not apply to the latter problem. We therefore build a separate component for each subtask, adopting Iida et al. (2005)'s selection-then-classification model for each component:

1. *Intra-sentential antecedent identification*: For a given zero-pronoun ZP in a given sentence S, select the most-likely candidate antecedent C_1^* from the candidates appearing in S by the intra-sentential tournament model

2. *Intra-sentential anaphoricity determination*: Estimate plausibility p_1 that C_1^* is the true antecedent, and return C_1^* if $p_1 \geq \theta_{intra}$ (θ_{intra} is a preselected threshold) or go to 3 otherwise

3. *Inter-sentential antecedent identification*: Select the most-likely candidate antecedent C_2^* from the candidates appearing outside of S by the inter-sentential tournament model.

4. *Inter-sentential anaphoricity determination*: Estimate plausibility p_2 that C_2^* is the true antecedent, and return C_2^* if $p_2 \geq \theta_{inter}$ (θ_{inter} is a preselected threshold) or return non-anaphoric otherwise.

6.3.2 Representation of syntactic patterns

In the first two of the above four steps, we use syntactic pattern features. Analogously to SRL, we extract the parse path between a zero-pronoun to its antecedent to capture the syntactic pattern of their occurrence. Among many alternative ways of representing a path, in the experiments reported in the next section, we adopted a method as we describe below, leaving the exploration of other alternatives as future work.

Given a sentence, we first use a standard dependency parser to obtain the dependency parse tree, in which words are structured according to the dependency relation between them. Figure 6.1(a), for example, shows the dependency tree of sentence (18) given in Section 6.2. We then extract the path between a zero-pronoun and its antecedent as in Figure 6.1(b). Finally, to encode the order of siblings and reduce data sparseness, we further transform the extracted path as in Figure 6.1(c):

- A path is represented by a subtree consisting of backbone nodes: ϕ (zero-pronoun), Ant (antecedent), Node (the lowest common ancestor), LeftNode (left-branch node) and RightNode.

- Each backbone node has daughter nodes, each corresponding to a function word associated with it.

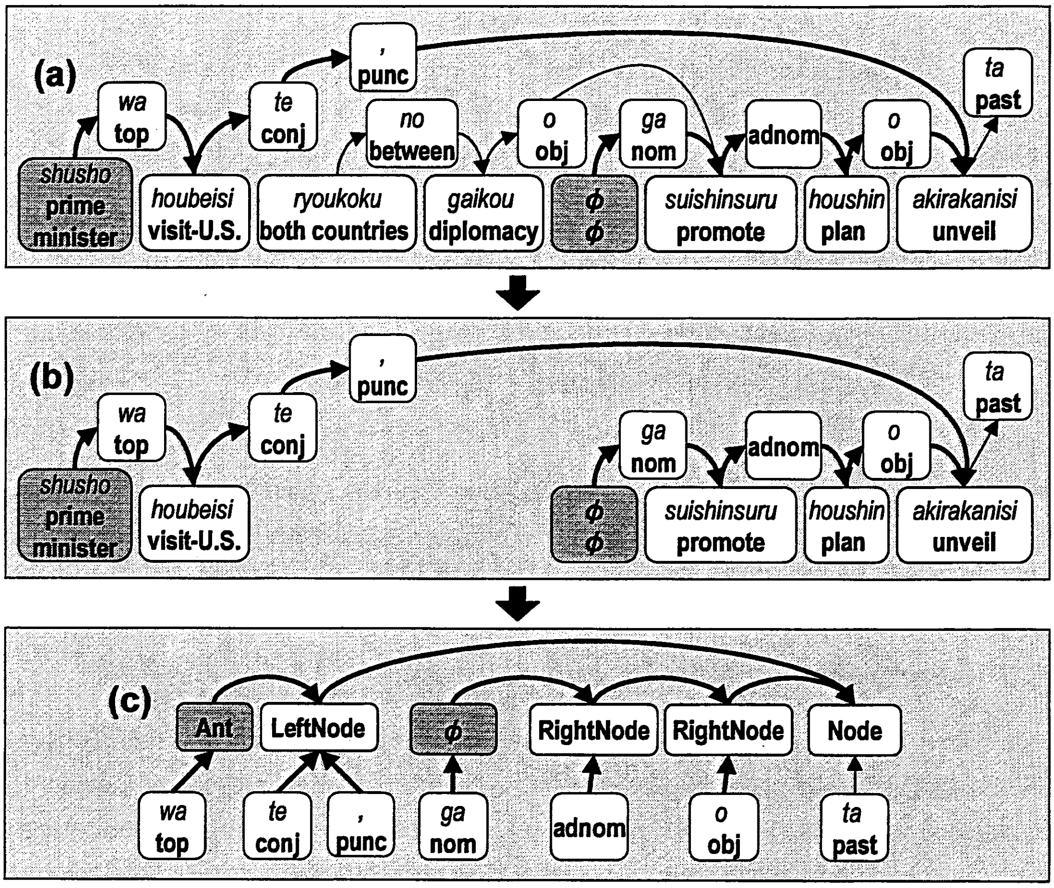


Figure 6.1. Representation of the path between a zero-pronoun to its antecedent

- Content words are deleted.

This way of encoding syntactic patterns is used in intra-sentential anaphoricity determination. In antecedent identification, on the other hand, the tournament model allows us to incorporate three paths, a path for each pair of a zero-pronoun and left and right candidate antecedents, as shown in Figure 6.2[5].

[5] To indicate which node belongs to which subtree, the label of each node is prefixed either with L, R or I.

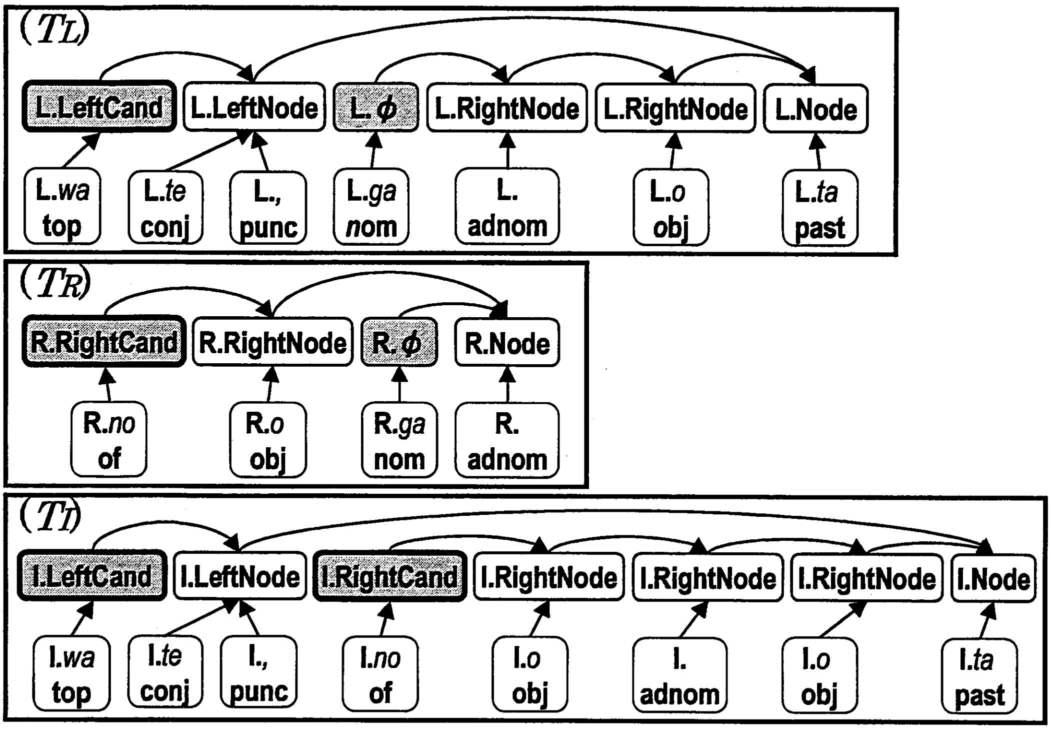


Figure 6.2. Paths used in the tournament model

6.3.3 Learning algorithm

As noted in Section 6.2, the use of zero-pronouns in Japanese is relatively less syntactically constrained compared, for example, with English. This forces the way of encoding path information given above to produce a staggering number of different paths, which inevitably leads to serious data sparseness problems.

This issue can be addressed in several ways. The SRL community has devised a range of variants of the standard path representation to reduce the complexity (Carreras and Marquez, 2005). Applying Kernel methods such as Tree kernels (Collins and Duffy, 2001) and Hierarchical DAG kernels (Suzuki et al., 2003) is another strong option. The Boosting-based algorithm proposed by Kudo and Matsumoto (Kudo and Matsumoto, 2004) is designed to learn subtrees useful for classification.

Leaving the question of selecting learning algorithms open, in our experiments, we have so far examined Kudo and Matsumoto (Kudo and Matsumoto, 2004)'s algorithm, which is implemented as the BACT system[6]. Given a set of training instances, each of which is represented as a tree labeled either positive or negative, the BACT system learns a list of weighted decision stumps with a Boosting algorithm.

The tree classification problem in BACT is defined to induce a mapping function $f(\mathrm{x})$: $\mathcal{X} \rightarrow \{\pm 1\}$, from given training instances $T = \{\langle \mathrm{x}_i, y_i \rangle\}_{i=1}^{L}$, where $\mathrm{x}_i \in \mathcal{X}$ is a labeled ordered tree and $y_i \in \{\pm 1\}$ is a class label associated with each training data. In each iteration of boosting, the decision stumps are trained to find a rule $\langle \hat{t}, \hat{y} \rangle$ that minimizes the error rate for the given training data $\{\langle \mathrm{x}_i, y_i \rangle\}_{i=1}^{L}$:

$$\langle \hat{t}, \hat{y} \rangle = \operatorname*{argmin}_{t \in \mathcal{F}, y \in \{\pm 1\}} \sum_{i=1}^{L} y_i d_i h_{\langle t,y \rangle}(\mathrm{x}_i), \tag{6.1}$$

where $\mathcal{F}$ is a set of candidate trees, $d_i (\sum_{i=1}^{L} d_i = 1, d_i \geq 0, \forall i = 1, \ldots, L)$ is a weight of each iteration, and $h_{\langle t,y \rangle}(\mathrm{x})$ is a decision stump classifier given by

$$h_{\langle t,y \rangle}(\mathrm{x}) \overset{\mathrm{def}}{=} \begin{cases} y & t \subseteq \mathrm{x} \\ -y & otherwise. \end{cases} \tag{6.2}$$

At the classification step, we use the following mapping function:

$$f(\mathrm{x}) = sgn(\sum_{k=1}^{K} \alpha_k h_{\langle t,y \rangle}(\mathrm{x}_k)) \tag{6.3}$$

where α_k is a weight of each decision stumps classifier $h_{\langle t,y \rangle}(\mathrm{x}_k)$. In this algorithm, α_k is calculated based on a variant of Boosting algorithm, *Arc-GV* (see Breiman:99).

The BACT algorithm has the important characteristic that the results of learning trees are more human-readable than those learned from algorithms such as Support Vector Machines, because the result of each iteration is given as a pair of weight α_k and decision stumps $h_{\langle t,y \rangle}$ in the training data set. So, we can easily interpret what kinds of sub-trees or features are useful for classification by viewing the results.

In antecedent identification, we train the tournament model by providing a set of labeled trees as a training set, where a label is either left or right. Each labeled tree has (i) path trees T_L, T_R and T_I (as given in Figure 6.2) and (ii) a set nodes corresponding to the binary features summarized in Table 6.3, each of which is linked to the root node

[6] http://chasen.org/˜taku/software/bact/

as illustrated in Figure 6.4. This way of organizing a labeled tree allows the model to learn, for example, the combinations of a subtree of T_L and some of the binary features. Analogously, for anaphoricity determination, we use trees $(T_C, f_1, \ldots, f_n)$, where T_C denotes a path subtree as in Figure 6.1(c).

6.4. Experiments

We conducted an evaluation of our method using Japanese newspaper articles. The following four models were compared:

1. BM: Ng and Cardie (2002a)'s search-based model.

2. BM_STR: BM with the syntactic features such as those in Figure 6.1(c).

3. SCM: The selection-then-classification model explained in Section 5.3.

4. SCM_STR: SCM with all types of syntactic features shown in Figure 6.2.

6.4.1 Setting

We created an anaphoric relation-tagged corpus consisting of 197 newspaper articles (1,803 sentences), 137 articles annotated by two annotators and 60 by one. The agreement ratio between two annotators on the 197 articles was 84.6%, which indicated that the annotation was sufficiently reliable.

In the experiments, we removed from the above data set the zero-pronouns to which the two annotators did not agree. Consequently, the data set contained 995 intra-sentential anaphoric zero-pronouns, 754 inter-sentential anaphoric zero-pronouns, and 603 non-anaphoric zero-pronouns (2,352 zero-pronouns in total), with each anaphoric zero-pronoun annotated to be linked to its antecedent. For each of the following experiments, we conducted five-fold cross-validation over 2,352 zero-pronouns so that the set of the zero-pronouns from a single text was not divided into the training and test sets.

In the experiments, all the features were automatically acquired with the help of the following NLP tools: the Japanese morphological analyzer *ChaSen*[7] and the Japanese

[7] http://chasen.naist.jp/hiki/ChaSen/

dependency structure analyzer *CaboCha*[8], which also carried out named-entity chunking.

6.4.2 Results on intra-sentential zero-anaphora resolution

In both intra-anaphoricity determination and antecedent identification, we investigated the effect of introducing the syntactic features for improving the performance. First, the results of antecedent identification are shown in Table 6.1. The comparison between BM (SCM) with BM_STR (SCM_STR) indicates that introducing the structural information effectively contributes to this task. In addition, the large improvement from BM_STR to SCM_STR indicates that the use of the tournament model has significant impact on intra-sentential antecedent identification. This finding may well contribute to semantic role labeling because these two tasks have a large overlap as discussed in Section 6.1.

Second, to evaluate the performance of intra-sentential zero-anaphora resolution, we plotted recall-precision curves altering threshold parameter and θ_{inter} for intra-anaphoricity determination as shown in Figure 6.5, where recall R and precision P were calculated by:

$$R = \frac{\#\text{ of detected antecedents correctly}}{\#\text{ of anaphoric zero-pronouns}},$$
$$P = \frac{\#\text{ of detected antecedents correctly}}{\#\text{ of zero-pronouns classified as anaphoric}}.$$

Note that we used the value of the BACT's discrimination function (i.e. (6.3)) as the score in intra-sentential zero-anaphora resolution. The curves indicate the upperbound of the performance of these models; in practical settings, the parameters have to be trained beforehand.

Figure 6.5 shows that BM_STR (SCM_STR) outperforms BM (SCM), which indicates that incorporating syntactic pattern features works remarkably well for intra-sentential zero-anaphora resolution. Furthermore, SCM_STR is significantly better than BM_STR. This result supports that the former has an advantage of learning non-anaphoric zero-pronouns (181 instances) as negative training instances in intra-sentential anaphoricity determination, which enables it to reject non-anaphoric zero-pronouns more accurately than the others.

[8] http://chasen.org/~taku/software/cabocha/

Table 6.1. Accuracy of antecedent identification.

BM	BM_STR	SCM	SCM_STR
48.0% (478/995)	63.5% (632/995)	65.1% (648/995)	**70.5%** (701/995)

Table 6.2. Effects of altering training instances on antecedent identification in intra-sentential zero-anaphora resolution

training data	accuracy
1/16	0.553 (550/995)
1/8	0.574 (571/995)
1/4	0.620 (617/995)
1/2	0.600 (597/995)
1/1	0.705 (701/995)

Next, we evaluate the transition of the performance for the different training data size on antecedent identification task. The empirical results shown in Table 6.2 indicates that increasing training data makes the performance increase excluding the trial where half of the training data was used. We also evaluate the transition of each recall-precision curve as shown in Figure 6.6 by altering training data size in intra-sentential zero-anaphora resolution. Figure 6.6 shows that the performance is improved by increasing data size.

6.4.3 Discussion

Our error analysis reveals that a majority of errors can be attributed to the current way of handling quoted phrases and sentences. Figure 6.7 shows the difference in resolution accuracy between zero-pronouns appearing in a quotation (262 zero-pronouns) and the rest (733 zero-pronouns), where "IN_Q" denotes the former (in-quote zero-pronouns) and "OUT_Q" the latter. The accuracy on the IN_Q problems is considerably lower than that on the OUT_Q cases, which indicates that we should deal with in-quote cases with a separate model so that it can take into account the nested structure of discourse segments introduced by quotations.

6.4.4 Impact on overall zero-anaphora resolution

We next evaluated the effects of introducing the proposed model on overall zero-anaphora resolution including inter-sentential cases.

As a baseline model, we implemented the original SCM, designed to resolve intra-sentential zero-anaphora and inter-sentential zero-anaphora simultaneously with no syntactic pattern features. Here, we adopted Support Vector Machines (Vapnik, 1998) to train the classifier on the baseline model and the inter-sentential zero-anaphora resolution in the SCM using structural information.

For the proposed model, we plotted several recall-precision curves by selecting different value for threshold parameters θ_{intra} and θ_{inter}. Note that we used the value of the BACT's discrimination function as the score for classification in intra-sentential zero-anaphora resolution, whereas we used the distance between an input feature vector and the hyperplane of SVM as the score for classification in the remaining problems. The results are shown in Figure 6.8, which indicates that the proposed model significantly outperforms the original SCM if θ_{intra} is appropriately chosen.

We then investigated the feasibility of parameter selection for θ_{intra} by plotting the AUC values for different θ_{intra} values. Here, each AUC value is the area under a recall-precision curve. The results are shown in Figure 6.9. Since the original SCM does not use θ_{intra}, the AUC value of it is constant, depicted by the SCM. As shown in the Figure 6.9, the AUC-value curve of the proposed model is not peaky, which indicates the selection of parameter θ_{intra} is not difficult.

6.5. Summary

In intra-sentential zero-anaphora resolution, syntactic patterns of the appearance of zero-pronouns and their antecedents are useful clues. Taking Japanese as a target language, we have empirically demonstrated that incorporating rich syntactic pattern features in a state-of-the-art learning-based anaphora resolution model dramatically improves the accuracy of intra-sentential zero-anaphora, which consequently improves the overall performance of zero-anaphora resolution.

In our next step, we are going to address the issue of how to find zero-pronouns, which requires us to design a broader framework that allows zero-anaphora resolution to interact with predicate-argument structure analysis. Another important issue is how to find a globally optimal solution to the set of zero-anaphora resolution problems in a given discourse, which leads us to explore methods as discussed by McCallum and Wellner (McCallum and Wellner, 2003).

Feature Type	Feature	Description
Lexical	HEAD_BF	characters of right-most morpheme in *NP* (*PRED*).
Grammatical	PRED_IN_MATRIX	1 if *PRED* exists in the matrix clause; otherwise 0.
	PRED_IN_EMBEDDED	1 if *PRED* exists in the relative clause; otherwise 0.
	PRED_VOICE	1 if *PRED* contains auxiliaries such as '*(ra)reru*'; otherwise 0.
	PRED_AUX	1 if *PRED* contains auxiliaries such as '*(sa)seru*', '*hosii*', '*morau*', '*itadaku*', '*kudasaru*', '*yaru*' and '*ageru*'.
	PRED_ALT	1 if PRED_VOICE is 1 or PRED_AUX is 1; otherwise 0.
	POS	Part-of-speech of *NP* followed by IPADIC (Asahara and Matsumoto, 2003).
	DEFINITE	1 if *NP* contains the article corresponding to DEFINITE 'the', such as '*sore*' or '*sono*'; otherwise 0.
	DEMONSTRATIVE	1 if *NP* contains the article corresponding to DEMONSTRATIVE 'that' or 'this', such as '*kono*', '*ano*'; otherwise 0.
	PARTICLE	Particle followed by *NP*, such as '*wa* (topic)', '*ga* (subject)', '*o* (object)'.
Semantic	NE	Named entity of *NP*: PERSON, ORGANIZATION, LOCATION, ARTIFACT, DATE, TIME, MONEY, PERCENT or N/A.
	EDR_HUMAN	1 if *NP* is included among the concept 'a human being' or 'atribute of a human being' in EDR dictionary (Jap, 1995); otherwise 0.
	PRONOUN_TYPE	Pronoun type of *NP*. (e.g. '*kare* (he)' → PERSON, '*koko* (here)' → LOCATION, '*sore* (this)' → OTHERS)
	SELECT_REST	1 if *NP* satisfies selectional restrictions in Nihongo Goi Taikei (Japanese Lexicon) (Ikehara et al., 1997); otherwise 0.
	COOC	the score of well-formedness model estimated from a large number of triplets ⟨*Noun, Case, Predicate*⟩ proposed by Fujita et al. (2004)
Positional	SENTNUM	Distance between *NP* and *PRED*.
	BEGINNING	1 if *NP* is located in the beggining of sentence; otherwise 0.
	END	1 if *NP* is located in the end of sentence; otherwise 0.
	PRED_NP	1 if *PRED* precedes *NP*; otherwise 0.
	NP_PRED	1 if *NP* precedes *PRED*; otherwise 0.
	DEP_PRED	1 if NP_i depends on *PRED*; otherwise 0.
	DEP_NP	1 if *PRED* depends on NP_i; otherwise 0.
	IN_QUOTE	1 if *NP* exists in the quoted text; otherwise 0.
Heuristic	CL_RANK	a rank of *NP* in forward looking-center list based on Centering Theory (Grosz et al., 1995)
	CL_ORDER	a order of *NP* in forward looking-center list based on Centering Theory (Grosz et al., 1995)

NP and *PRED* stand for a bunsetsu-chunk of a candidate antecedent and a bunsetsu-chunk of a predicate which has a target zero-pronoun respectively.

Figure 6.3. Feature set.

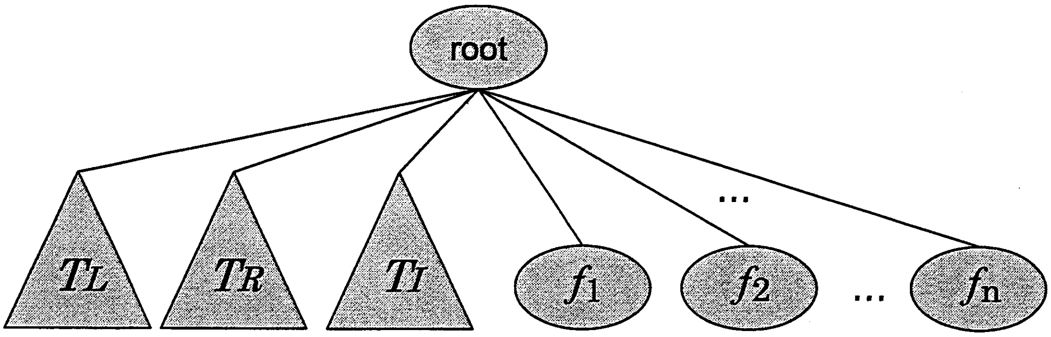


Figure 6.4. Tree representation of features for the tournament model.

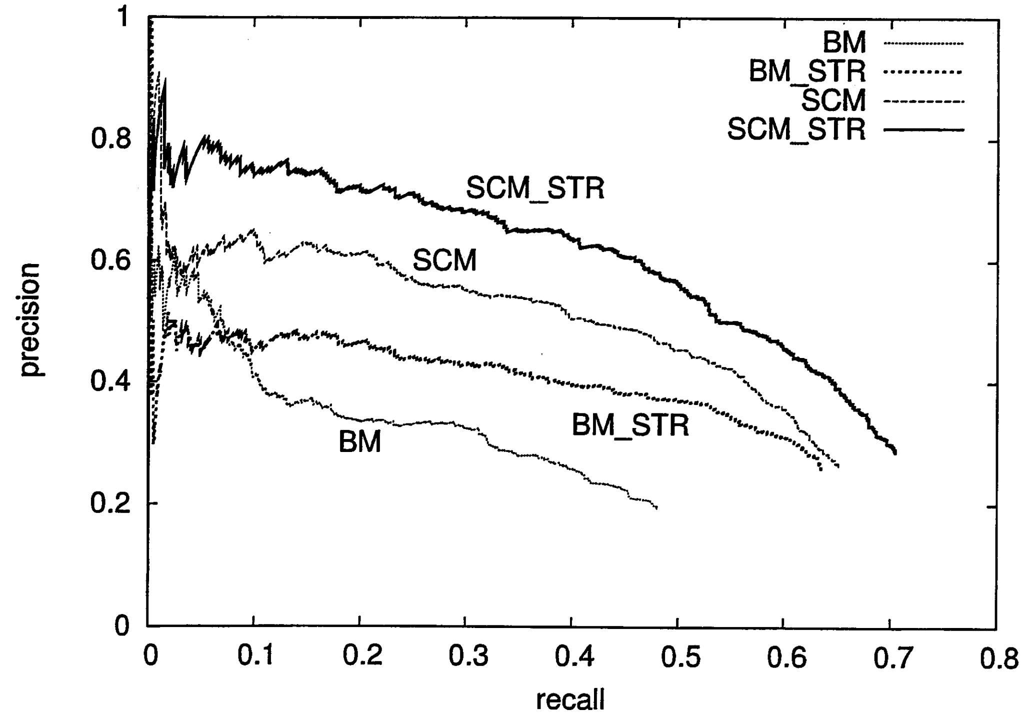


Figure 6.5. Recall-precision curves of intra-sentential zero-anaphora resolution.

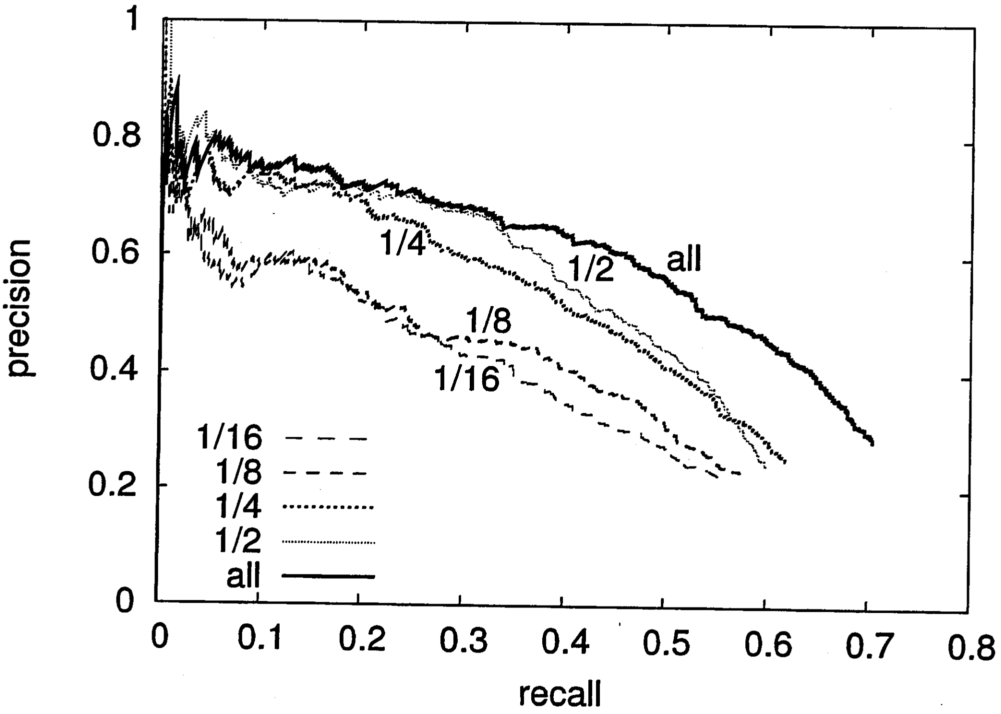


Figure 6.6. Transition of recall-precision curves in intra-sentential zero-anaphora resolution

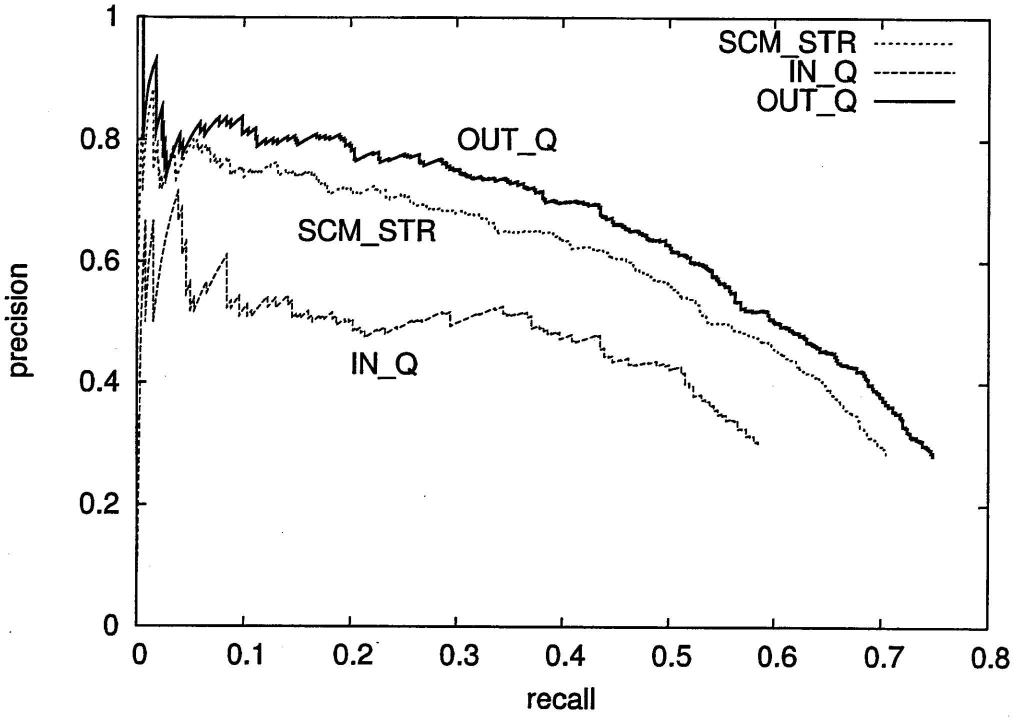


Figure 6.7. Recall-precision curves of resolving in-quote and out-quote zero-pronouns.

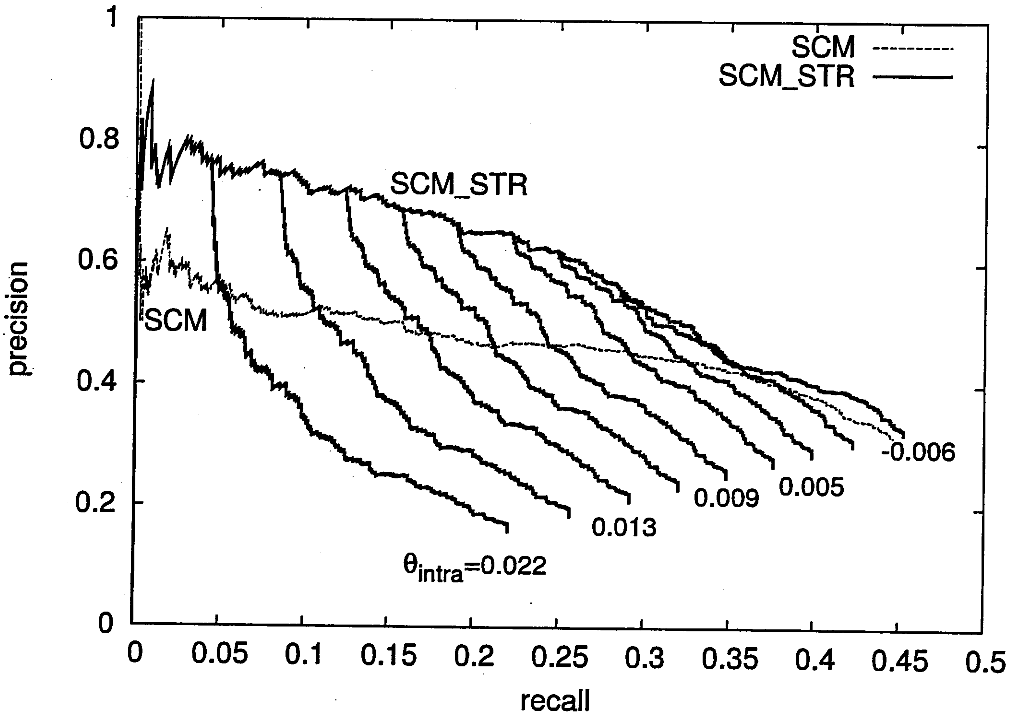


Figure 6.8. Recall-precision curves of overall zero-anaphora resolution.

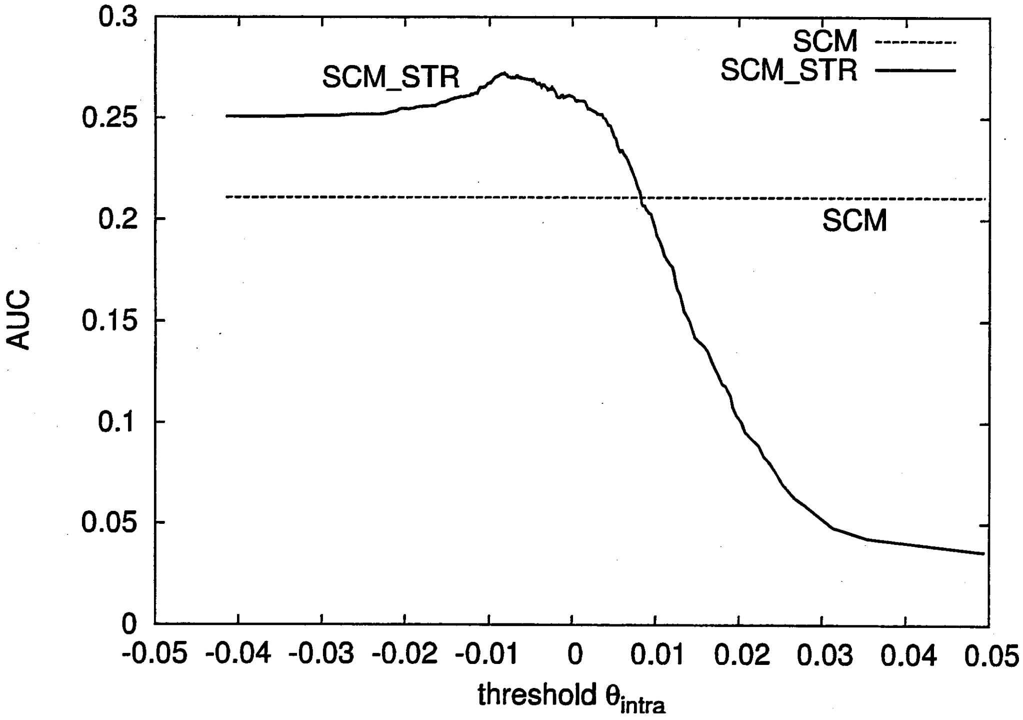


Figure 6.9. AUC curves plotted by altering θ_{intra}.

Chapter 7

Conclusion

In this thesis, we address three methods for incorporating contextual information into machine learning-based approaches for anaphora resolution. This chapter concludes the thesis with the following sections. Section 7.1 outlines the contribution of the thesis and Section 7.2 lists future work plans.

7.1. Summary of contributions

In Chapter 4, we explained the preference between candidate antecedents in forward-looking center introduced in Centering Theory is generally formalized by comparison between candidates. We introduced the tournament model, a machine learning-based model that can directly compare two candidates in series of matches, dramatically outperformed conventional pairwise classification models in the experiments of Japanese zero-anaphora resolution.

In addition to the tournament model, in Chapter 5 we proposed the selection-then-classification model that processes reverses the order of the steps in the classification-then-search model proposed by Ng and Cardie (2002b), inheriting all the advantages of that model. We conducted experiments on resolving noun phrase anaphora in Japanese. The results show that with the selection-then-classification based modifications, our model outperforms earlier learning-based approaches.

Finally, in Chapter 6 we investigated whether or not syntactic patterns between a given anaphor and a candidate antecedent are useful if an anaphor and a candidates appear in the same sentence. Taking Japanese as a target language, we empirically demonstrated that incorporating rich syntactic pattern features into a state-of-the-art learning-based anaphora resolution model dramatically improved the accuracy of intra-sentential zero-anaphora resolution, which consequently improves the overall performance of zero-anaphora resolution.

7.2. Future work

As described in Chapter 1, anaphora resolution is essential for various types of natural language applications. Finally, we would like to conclude the thesis with the future work.

Employment of the hierarchical structure between two candidate antecedents

In described in Chapter 4, *wa*-marked subtopics are often incorrectly selected as the most likely antecedent because the current model cannot capture topic-subtopic structures. Our next step will be to encode such hierarchical structures as a centering feature. Since a topic-subtopic relation holds between two NPs, it may be effective in the tournament model.

In addition, *wa*-marked NP in quoted sentence are incorrectly identified even if we employ a feature when indicates that a candidate antecedent appears in a quoted sentence. In such case, some predicates appearing in the quoted sentence often take the speaker as an argument. Thus, there is room for further improvement on the task of zero-anaphora resolution by identifying the speaker and incorporating such information into the current model.

Refinement of selectional preferences

We need to make selectional restrictions more effective for resolving anaphora. As described in Chapter 4 and Chapter 6, we introduce the probabilistic selectional restriction such as log-likelihood ratio based on triplet ⟨NOUN, CASE, VERB⟩. However, in case of employing such triplet, we cannot capture verb ambiguity and then the model identifies an incorrect candidate as the antecedent. In order to resolve the problem, verb frame information are available, but the task of constructing such resources involves technical problems such as verb sense disambiguation and classification of obligate and arbitral arguments.

Recently, automatic verb frame construction methods have been increasing attention (Resnik, 1993; Utsuro and Matsumoto, 1997; Kawahara et al., 2000; Gildea, 2002, etc.), which cluster verbs and arguments based on the similarity between instances in corpora. By adopting such a strategy, we obtain a scalable frame dictionary constructed automatically from large amounts of text data such as that available on the web. Of course, dictionaries that are automatically constructed are noisy and not always fine-grained, however, we are convinced that it is beneficial to introduce dictionaries into the process of anaphora resolution in light of these issues.

Semantic compatibility between two noun phrases

As we discussed in Chapter 5, in noun phrase anaphora resolution, the majority of errors are caused by the difficulty of judging the semantic compatibility between a candidate anaphor and candidate antecedent. To resolve this issue, we need resources that consist of equivalence relations for coreference resolution.

Interdependence between zero-pronouns

We should consider the interdependence between zero-pronouns appearing in the same discourse and the effect this relationship has on the syntax and semantics of its sentence, such that the nominative and dative slots in a clause must be occupied by distinct entities. This line of refinement will also lead us to explore methods to search for a globally optimal solution to a set of anaphora resolution problems for a given text, as discussed by McCallum and Wellner (2003).

References

M. Asahara and Y. Matsumoto, 2003. *IPADIC User Manual.* Nara Institute of Science and Technology, Japan.

B. Baldwin. 1995. *CogNIAC: A Discourse Processing Engine.* Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania.

D. L. Bean and E. Riloff. 1999. Corpus-based Identification of Non-Anaphoric Noun Phrases. In *Proceedings of the 37th ACL*, pages 373–380.

S. E. Brennan, M. W. Friedman, and C. Pollard. 1987. A Centering Approach to Pronouns. In *Proceedings of the 25th ACL*, pages 155–162.

X. Carreras and L. Marquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of the Ninth CoNll*, pages 152–164.

M. Collins and N.l Duffy. 2001. Convolution Kernels for Natural Language. In *Proceedings of the NIPS*, pages 625–632.

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. Automatic Content Extraction (ACE) program - task definitions and performance measures. In *Proceedings of the 4rd International Conference on Language Resources and Evaluation* (*LREC-2004*), pages 837–840.

T. Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

A. Fujita, K. Inui, and Y. Matsumoto. 2004. Detection of Incorrect Case Assignments in Automatically Generated Paraphrases of Japanese Sentences. In *Proceeding of the first IJCNLP*, pages 14–21.

N. Ge, J. Hale, and E. Charniak. 1998. A Statistical Approach to Anaphora Resolution. In *Proceedings of the 6th Workshop on Very Large Corpora*, pages 161–170.

D. Gildea and D. Jurafsky. 2002. Automatic Labeling of Semantic Roles. In *Computational Linguistics*, pages 245–288.

D. Gildea. 2002. Probabilistic Models of Verb-Argument Structure. In *Proceedings of the 19th International Conference on Computational Linguistics* (*COLING*, pages 308–314.

B. J. Grosz, A. K. Joshi, and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

M. A. K. Halliday and R. Hasan. 1976. *Cohesion in English.* English Language Series,Title No.9. Longman.

K. Hasida. 2002. Global Document Annotation (GDA). http://i-content.org/.

L. Hirschman. 1997. *MUC-7 coreference task definition.* Version 3.0.

J. Hobbs. 1978. Resolving Pronoun References. *Lingua*, 44:311–338.

R. Iida, K. Inui, H. Takamura, and Y. Matsumoto. 2003. Incorporating Contextual Cues in Trainable Models for Coreference Resolution. In *Proceedings of the 10th EACL Workshop on The Computational Treatment of Anaphora*, pages 23–30.

R. Iida, K. Inui, and Y. Matsumoto. 2005. Anaphora resolution by antecedent identification followed by anaphoricity determination. *ACM Transactions on Asian Language Information Processing* (*TALIP*), 4:417–434.

S. Ikehara, M. Miyazaki, S. Shirai A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. 1997. *Nihongo Goi Taikei (in Japanese).* Iwanami Shoten.

Japan Electronic Dictionary Research Institute, Ltd. Japan, 1995. *EDR Electronic Dictionary Technical Guide.*

M. Kameyama. 1986. A Property-Sharing Constraint in Centering. In *Proceedings of the 24th ACL*, pages 200–206.

D. Kawahara, N. Kaji, and S. Kurohashi. 2000. Japanese Case Structure Analysis by Unsupervised Construction of a Case Frame Dictionary. In *Proceedings of the 18th International Conference on Computational Linguistics* (*COLING*, pages 432–438.

D. Kawahara, T. Kurohashi, and K. Hasida. 2002. Construction of a Japanese Relevance-tagged Corpus (in Japanese). In *Proceedings of the 8th Annual Meeting of the Association for Natural Language Processing*, pages 495–498.

Andrew Kehler. 1997. Current Theories of Centering for Pronoun Interpretation: A Critical Evaluation. In *Computational Linguistics*, volume 23, pages 467–475.

P. Kingsbury and M. Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (*LREC-2002*), pages 1989–1993.

T. Kudo and Y. Matsumoto. 2002. Japanese Dependency Analysis using Cascaded Chunking. In *Proceedings of the 6th CoNLL*, pages 63–69.

T. Kudo and Y. Matsumoto. 2004. A Boosting Algorithm for Classification of Semi-Structured Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 301–308.

S. Lappin and H. J. Leass. 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535–561.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English:The Penn Treebank. In *Computational Linguistics*, pages 313–330.

Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara, 2000. *Morphological Analysis System ChaSen version 2.2.1 Manual*.

A. McCallum and B. Wellner. 2003. Object Consolidation by Graph Partitioning with a Conditionally Trained Distance Metric. In *Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pages 19–24.

J. F. McCarthy and W. G. Lehnert. 1995. Using Decision Trees for Coreference Resolution. In *Proceedings of the 14th IJCAI*, pages 1050–1055.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The NomBank Project: An InterimReport. In *Proceedings of the HLT-NAACL Workshop on Frontiers in Corpus Annotation*.

R. Mitkov. 1997. Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches. In *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Studies in Language and Linguistics. Pearson Education.

H. Nakaiwa and S. Shirai. 1996. Anaphora Resolution of Japanese Zero Pronouns with Deictic Reference. In *Proceedings of the 16th COLING*, pages 812–817.

S. Nariyama. 2002. Grammar for ellipsis resolution in Japanese. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 135–145.

Natural Language Research Institute. 1964. *Burui Goi Hyo (in Japanese)*. Shuuei Publishing.

V. Ng and C. Cardie. 2002a. Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 40th ACL*, pages 104–111.

V. Ng and C. Cardie. 2002b. Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution. In *Proceedings of the 19th International Conference on Computational Linguistics* (*COLING*), pages 730–736.

V. Ng and C. Cardie. 2002c. Combining Sample Selection and Error-Driven Pruning for Machine Learning of Coreference Rules. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 55–62.

V. Ng. 2004. Learning Noun Phrase Anaphoricity to Improve Coreference Resolution: Issues in Representation and Optimization. In *Proceedings of the 42nd ACL*, pages 152–159.

M. Okumura and K. Tamura. 1996. Zero Pronoun Resolution in Japanese Discourse Based on Centering Theory. In *Proceedings of the 16th COLING*, pages 871–876.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

M. T. Pazienza, (ed.). 1997. *Information Extraction *A Multidisciplinary Approach to an Emerging Information Technology*. Lecture Notes in Artificial Intelligence, Vol.1299. Springer-Verlag.

M. Poesio, H. Cheng, R. Henschel, J. Hitzeman, R. Kibble, and R. Stevenson. 2000. Specifying the Parameters of Centering Theory: a Corpus-Based Evaluation using Text from Application-Oriented Domains. In *Proceeding of the 38th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 400–407.

M. Poesio, O. Uryupina, R. Vieira, M. Alexandrov-Kabadjov, and R. Goulart. 2004. Discourse-New Detectors for Definite Description Resolution: A Survey and a Preliminary Proposal. In *Proceedings of the ACL 2004 Workshop on Reference Resolution and its Applications*, pages 47–54.

P. Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania.

K. Seki, A. Fujii, and T. Ishikawa. 2002. A Probabilistic Method for Analyzing Japanese Anaphora Integrating Zero Pronoun Detection and Resolution. In *Proceedings of the 19th COLING*, pages 911–917.

Nikkei Shimbunsha. 1990-2000. Nikkei Shimbun CD-ROM.

Mainichi Shimbunsha. 1991-1999. Mainichi Shimbun CD-ROM.

W. M. Soon, H. T. Ng, and D. C. Y. Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544.

M. Strube and C. Müller. 2003. A Machine Learning Approach to Pronoun Resolution in Spoken Dialogue. In *Proceedings of the 41st ACL*, pages 168–175.

J. Suzuki, T. Hirao, Y. Sasaki, and E. Maeda. 2003. Hierarchical Directed Acyclic Graph Kernel: Methods for Structured Natural Language Data. In *Proceeding of the 41st ACL*, pages 32–39.

J. R. Tetreault. 2001. A Corpus-Based Evalutation of Centering and Pronoun Resolution. In *Computational Linguistics*, volume 27, pages 507–520.

M. Tsuchiya, T. Utsuro, S. Matsuyoshi, S. Sato, and S. Nakagawa. 2006. Development and Analysis of An Example Database of Japanese Compound Functional Expressions. 47(6):1728–1741.

O. Uryupina. 2003. High-precision Identification of Discourse New and Unique Noun Phrases. In *Proceedings of the 41st ACL Student Research Workshop*, pages 80–86.

T. Utsuro and Y. Matsumoto. 1997. Learning Probabilistic Subcategorization Preference by Identifying Case Dependencies and Optimal Noun Class Generation Level. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 364–371.

K. van Deemter and R. Kibble. 1999. What is coreference, and what should coreference annotation be? In *Proceedings of the ACL '99 Workshop on Coreference and its applications*, pages 90–96.

V. N. Vapnik. 1998. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing Communications, and control. John Wiley & Sons.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A Model-Theoretic Coreference Scoring Scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pages 45–52.

M. Walker, M. Iida, and S. Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–233.

M. Walker, A. K. Joshi, and E. Prince (eds.). 1997. *Centering Theory in Discourse*. Oxford Univ. Press.

H. Yamada, T. Kudo, and Y. Matsumoto. 2002. Japanese Named Entity Extraction Using Support Vector Machine. *IPSJ Journal*, 43(1):44–53.

K. Yamamoto and E. Sumita. 1998. Feasibility Study for Ellipsis Resolution in Dialogues by Machine-Learning Technique. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (*COLING-ACL*), pages 1428–1435.

X. Yang, G. Zhou, J. Su, and C. L. Tan. 2003. Coreference Resolution Using Competition Learning Approach. In *Proceedings of the 41st ACL*, pages 176–183.

List of Publications

Journal Papers

1. R. Iida, K. Inui, and Y. Matsumoto. 2005. Anaphora Resolution by Antecedent Identification Followed by Anaphoricity Determination. *ACM Transactions on Asian Language Information Processing* (*TALIP*), 4:417–434.

2. R. Iida, K. Inui, Y. Matsumoto, and S. Sekine. 2005. Noun Phrase Coreference Resolution in Japanese Based on Most Likely Antecedent Candidates. *IPSJ Journal*, 46(3):831–844. (in Japanese).

3. R. Iida, K. Inui, and Y. Matsumoto. 2004. Identifying Antecedents of Japanese Zero-pronouns Using A Machine Learning Model with Contextual Cues. *IPSJ Journal*, 45(3):906–918.

International Conferences and Workshops

1. R. Iida, K. Inui, and Y. Matsumoto. 2006. 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* (*COLING/ACL*), pages 625–632.

2. R. Iida, K. Inui, and Y. Matsumoto. 2005. The Issue of Combining Anaphoricity Determination and Antecedent Identification in Anaphora Resolution. In *Proceeding of the IEEE International Conference on Natural Language Processing and Knowledge Engineering* (*IEEE NLP-KE*), pages 244–249.

3. R. Iida, K. Inui, H. Takamura, and Y. Matsumoto. 2003. Incorporating Contextual Cues in Trainable Models for Coreference Resolution. In *Proceedings of the 10th EACL Workshop on The Computational Treatment of Anaphora*, pages 23–30.

Other Publications

1. R. Iida, M. Komachi, K. Inui, and Y. Matsumoto. 2007. NAIST Text Corpus: Annotating Predicate-Argument and Coreference Relations. In *Information Processing Society of Japanese SIG Notes, NL-177-10*, pages 71–78. (in Japanese).

2. R. Iida, K. Inui, and Y. Matsumoto. 2006. Learning Syntactic Patterns for Zero-anaphora Resolution. In *Proceedings of the 2006 FIT*, pages 81–84. (in Japanese).

3. R. Iida, K. Inui, and Y. Matsumoto. 2006. Intra-sentential Zero-anaphora Resolution Using Dependency Structure Information. In *Proceedings of the 12th Annual Meeting of the Association for Natural Language Processing*, pages 488–491. (in Japanese).

4. R. Iida, K. Inui, and Y. Matsumoto. 2005. Noun Anaphora Resolution Combining Anaphoricity Determination and Antecedent Identification: Experiments and Analysis. In *Information Processing Society of Japanese SIG Notes, NL-169-15*, pages 93–100. (in Japanese).

5. R. Iida, K. Inui, and Y. Matsumoto. 2005. Learning An Anaphoricity Determination Model Combining Preceding and Local Contextual Information. In *Proceedings of the 11th Annual Meeting of the Association for Natural Language Processing*, pages 1048–1051. (in Japanese).

6. R. Iida, N. Kobayashi, K. Inui, and Y. Matsumoto. 2005. A Machine Learning-based Method to Extract Attribute-Value Pairs for Opinion Mining. In *Information Processing Society of Japanese SIG Notes, NL-165-4*, pages 21–28. (in Japanese).

7. R. Iida, K. Inui, Y. Matsumoto, and S. Satoshi. 2004. A Machine Learning-based Method for Resolving Japanese NP Coreference. In *Proceedings of the 10th Annual Meeting of the Association for Natural Language Processing*, pages 761–764. (in Japanese).

8. R. Iida, H. Takamura, K. Inui, and Y. Matsumoto. 2003. A Machine Learning-based Method for Zero-anaphora Resolution Using Contextual Cues. In *Proceedings of the 9th Annual Meeting of the Association for Natural Language Processing*, pages 585–588. (in Japanese).

9. R. Iida, H. Takamura, K. Inui, and Y. Matsumoto. 2003. One Method for Resolving Japanese Zero Pronouns with Machine Learning Model. In *Information Processing Society of Japanese SIG Notes, NL-154-23*, pages 161–168. (in Japanese).

10. R. Iida and E. T. Miyamoto. 2002. Corpus Counts of NP Sequences in Japanese. In *Technical report no. 2002015, November 2002. Nara Institute of Science and Technology* (*Ikoma, Japan*).

Awards

1. COLING/ACL 2006 Asian Federation of Natural Language Processing Best Asian NLP Paper Award. 2006. R. Iida, K. Inui, Y. Matsumoto. Exploiting Syntactic Patterns as Clues in Zero-Anaphora Resolution.

2. The 10th Annual Meeting of the Association for Natural Language Processing Best Presentation Award. 2004. R. Iida, K. Inui, Y. Matsumoto, and S. Satoshi. A Machine Learning-based Method for Resolving Japanese NP Coreference. (in Japanese).