最尤先行詞候補を用いた日本語名詞句同一指示解析

飯田 龍[†] 乾 健太郎[†]
松本 裕治[†] 関根 聡[††]

日本語における名詞句の同一指示関係を同定するための新しい解析手法を提案する．従来の同一指示解析方法は定名詞か否かを判定する際に局所文脈の情報しか参照していないという問題や，非照応詞と先行詞候補との関係を訓練事例として抽出していないため，非照応詞が棄却されることが保証されないという問題が存在する．また，日本語の場合は，英語などの言語と比較して冠詞の情報がないため同一指示関係の解析の問題が難しくなると考えられる．そこで，本稿では，照応詞の候補に対して先行詞となる可能性のある候補を提示することで，より広い文脈の情報を参照して照応詞か否かの分類問題を解くための手法を提案する．この手法では，非照応詞に対しても最も先行詞らしい候補を見せることで極端に負例が多くならないという利点もある．この提案手法を用いて日本語名詞句同一指示関係の同定実験を行い，先行研究の機械学習を用いた手法より精度良く同一指示関係の同定ができたことを報告する．

Noun Phrase Coreference Resolution in Japanese Based on Most Likely Antecedent Candidates

Ryu Iida,[†] Kentaro Inui,[†] Yuji Matsumoto[†]
and Satoshi Sekine[††]

We propose a new approach to coreference resolution in Japanese. In conventional approaches, noun phrases are classified as definite or not by referring only to information in the local context, and these approaches might not be able to reject non-anaphoric entities because the instances of the non-anaphoric entities and antecedent candidates are not extracted. Furthermore, resolving coreferential relations in Japanese is more difficult than in English because of the absence of articles. This paper proposes a model to resolve the classification problem of whether a candidate is truly an anaphor or not by referring to information from a much larger context. This model has an advantage of being able to reduce negative training instances by using the most likely antecedent candidate selected for a given non-anaphoric entity. Application of the proposed model to noun phrase coreference resolution showed that the model outperformed earlier machine learning-based models.

1. はじめに

ある言語表現が，文脈中の表現もしくは文脈外の要素と同じ内容や対象をさすとき，これらの表現は**照応関係**☆にあるという．このとき，指示先の表現を**先行詞**，指示元の表現を**照応詞**という．一般に，先行詞が照応詞よりも前にある照応関係は**前方照応**といい，逆に，先行詞が照応詞より後にくる照応関係を**後方照応**という．また，前方照応や後方照応のように，先行詞が文脈の中に存在する関係を**文脈照応**といい，逆に先行詞が文脈中に認められない場合，この関係を**外界照応**という．このような照応関係を特定する処理を照応解析といい，照応解析は照応詞を認定する処理（照応詞認定）と，認定した照応詞の先行詞を同定する処理（先行詞同定）からなる．照応解析は，対話モデルや高品質な機械翻訳システムを実現するために必要とされ，情報抽出や自然言語での質問応答タスクなどの応用分野で特に重要である．これまでの照応解析の手法は大きく理論指向の規則作成に基づく手法とコーパスを用いた学習手法に分類できる．

規則作成に基づく解析手法では，さまざまな言語的な手がかりを人手で規則に取り入れる試みが行われている[2),6),16),17)]．この手法では，対象となる名詞句の

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
†† ニューヨーク大学コンピュータサイエンス学科
Computer Science Department, New York University

☆ 照応と同一指示の関係の詳細については付録 A.1 でまとめる．

意味役割や先行詞候補の出現順序，照応詞と先行詞の間の意味的な互換性などの手がかりに加え，センタリング理論[4]のような言語学的な知見をもとに規則を記述する．MUC-7☆における照応解析のタスクでは，約70%の精度と約60%の再現率が報告されているが[2]，機械翻訳などの現実的な応用を考えた場合，満足できる精度とはいえない．さらに，規則が特定のドメインに特化している場合は，他のドメインで同様の精度を得ることが難しい．このような事実を考慮すると，人手による規則の洗練は難しく，コストも大きいと考えられる．

これに対し，照応タグ付きコーパスを用いた統計的な手法[3),7),10),12)]は，コストが低いという利点を持ちながらも，MUC-6やMUC-7の照応解析の評価セットを用いた実験で規則ベースの手法と同程度の精度を得ている．しかし，日本語を対象とした名詞句照応解析では，冠詞の情報がないため，名詞句の指示性の推定が英語などの言語に比べてさらに困難になると考えられる．

本稿では，日本語における名詞句の照応関係（かつ，同一指示関係）同定のための新しい解析モデルを提案する． 2章では先行研究の照応解析手法について述べ， 3章では新しい名詞句照応解析モデルを提案する．このモデルは，我々が以前提案した先行詞候補間の先行詞らしさをとらえるモデル（トーナメントモデル）[20]を拡張したモデルであり，照応詞候補に対して最も先行詞らしい候補（最尤先行詞候補）を決定した後，最尤先行詞候補と照応詞候補の対を用いて照応詞か否かを分類する．次に4章では，日本語の名詞句照応解析の実験を行い，実験結果について報告する．5章で関連研究との比較を述べ，最後に6章でまとめる．

2. 先行研究

同一指示解析の手法は大きく規則作成に基づく解析手法と同一指示関係タグ付きコーパスを用いた学習手法に分類できる．この章では規則ベースの手法として，村田ら[16),17)]の日本語名詞句同一指示解析手法を示し，また，機械学習を用いた解析手法として，Soonら[10)]とNgら[7)]の手法を説明する．

2.1 村田らの日本語名詞句同一指示解析手法

村田ら[16),17)]は，名詞句の指示性を総称名詞，定名詞，不定名詞の3種類に分類し（**図 1**），人手で作成した86個の規則を適用し文章中の名詞句を3種類のうちいずれかに分類している．作成された規則には，(1) 名詞自身，(2) 助詞や動詞の格関係，(3) 修飾句（節），(4) 既出の名詞句が再び出現する，(5) 規則が適用されない場合に不定名詞の可能性を高くする，という大きく5種類の情報を用いている．それぞれの規則には，名詞句がどの指示性となるかの可能性に関する評価値と，可能性の評価値が等しい場合にどの指示性となるかの優先度の値が人手で設定されている．実際に指示性を推定する際は，適用可能な規則をすべて適用し，名詞句をある指示性に唯一に決定する．次に，定名詞と分類した名詞句のみを対象に，2つの名詞句の文字列の一致度や修飾要素の情報に基づき，名詞句間の同一指示関係を決定する．

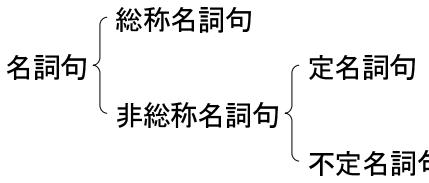

図 1 村田らの導入した名詞句の指示性
Fig. 1 The referential property of NP introduced by Murata, et al.

村田らが導入した規則では，既出の名詞句が再度出現した場合，再度出現した名詞句は定名詞となる可能性を高くするよう可能性と優先度を設定してあるが，どのような種類の名詞句がどのような状況で出現した場合に定名詞となるかは前文脈に出現している名詞句と対象としている名詞句の関係を考慮する必要があり，それらを人手で規則と書き尽くすことは困難である．

2.2 機械学習を用いた英語の名詞句同一指示解析手法

機械学習を用いた同一指示解析はすでにいくつかの手法が提案されており，たとえばSoonら[10)]やNgら[7)]のモデルは，MUCの照応解析のタスクにおいて規則ベースの手法と同程度の精度を得ている．

Soonらのモデルでは，同一指示解析の問題を，与えられた照応詞に対して，先行詞の候補となる名詞句の各々が先行詞となるか否かを判別する2値分類問題に分解する．**図 2**を用いて説明しよう．図2では，照応詞 ANP に対して，7つの名詞句 $NP_1, \ldots, NP_7$ が先行文脈に出現している状況を仮定している．NP_2 と NP_4，NP_3 と NP_5，NP_6 と NP_7 はそれぞれ同一指示関係にあり，ANP の先行詞は NP_5（NP_3）とする．この状況で，分類器は名詞句 NP_i （$i \in \{1, \ldots, 7\}$）が先行詞かどうかという2値分類問題を解く．

訓練時には，図2(a)のように照応詞と（照応詞から

☆ The 7th Message Understanding Conference（1998）：http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

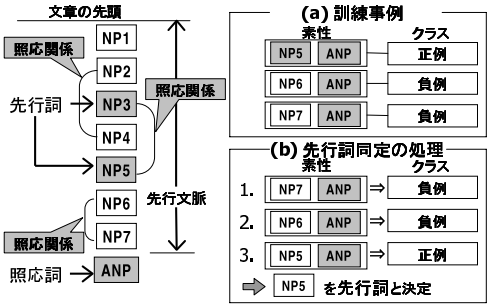

図 2 Soon ら（Ng ら）のモデル
Fig. 2 Soon's (Ng's) model.

最も近い）先行詞の対 ANP-NP_5 を正例，照応詞と他の（先行詞と照応詞の間の）各名詞句の対 ANP-NP_6 および ANP-NP_7 を負例として学習する．新しい同一指示問題を解く際には，訓練時と同様に，照応詞から先行文脈に向かって，先行詞候補となる名詞句の各々について，それが先行詞かどうか分類していく．そして，分類器がいずれかの名詞句を先行詞として決定した時点で解析を終了する．分類器が，先行する名詞句をすべて先行詞ではないと分類した場合は，対象としている照応詞は先行詞を持たないと判断する．図 2 (b) の例では，照応詞 ANP と ANP に最も近い先行詞候補 NP_7 との対，ANP と次に近い NP_6 との対と順に分類してゆき，分類器がはじめて正例に分類した ANP と NP_5 との対をそれぞれ照応詞と先行詞として決定する．Soon らの実験では，**表 1** に示した 12 個の限られた素性を用い，C5.0[9] を使用して決定木学習を行っている．

Ng ら[7] は Soon らの手法を 2 つの点において改良した．1 つは素性集合を拡張し，語彙的な素性や意味的な素性など，53 個に素性を増やした．もう 1 つは先行詞同定の探索アルゴリズムの変更である．Soon らが照応詞に近い名詞句から順に先行詞かどうかを決定的に決めるのに対し，Ng らはすべての先行する名詞句を分類器にかけ，分類器が先行詞と決定した名詞句の中で，最も先行詞らしいと判定した名詞句を先行詞とする．ここでも，すべての名詞句が先行詞でないと分類された場合には，照応詞は先行詞を持たないと判断する．

ただし，Soon らの手法や Ng らの手法では，図 2 (a) の訓練事例の抽出例のように，照応詞に対してどの先行詞候補が先行詞となりやすいかを判断するための訓練事例しか抽出しておらず，先行文脈のどの名詞句とも同一指示関係にない名詞句（非照応詞）についてはどの候補との対も抽出していない．そのため，非照応詞が適切に棄却されるかは不明である．非照応詞についても同様に先行詞候補との関係を抽出することが考えられるが，単純に先行詞候補と非照応詞の組合せを抽出した場合，負例の数が正例に比べて極端に多くなるという問題が起こる．

3. 提 案 手 法

2 章で示した先行研究のように，最初に名詞句の指示性を分類し，次に推定した定名詞のみを対象に同一指示関係を推定する，というアプローチは一見適切であるように思われる．しかし，日本語では冠詞などの情報がないため，名詞句の指示性の推定はそれほど容易でない．指示性の推定誤りはそのまま同一指示解析の失敗につながるので，この問題は慎重に扱う必要がある．指示性の認定には，大きく (1) 名詞句の意味カテゴリの粒度や名詞句の係り先，係り元などの語彙・統語的な情報と，(2) 先行詞候補群から得られる情報の 2 種類の情報が必要となると考えられるが，本研究では，特に (2) の情報をもとに同一指示関係の同定を試みる．

先行詞候補群から得られる情報を効果的に利用するために，提案手法では，以下の手続きで同一指示関係を解析する．

(1) 文章の先頭から順に照応詞☆の候補となりうる候補を検出する．
(2) 対象とする照応詞候補に対して先行文脈から先行詞候補をすべて抽出する．
(3) 照応詞候補に対して最も先行詞らしい候補（最尤先行詞候補）を選択する．
(4) (3) で選択した照応詞候補と最尤先行詞候補の対が真に照応詞とそれに対応する先行詞か否かの 2 値分類問題を解く．
(5) 解析の対象となる照応詞候補がなくなるまで (1)〜(4) を繰り返す．

(3) の最尤先行詞候補の同定については 3.1 節に，また，(4) の照応詞の認定については 3.2 節に詳細をまとめる．

3.1 最尤先行詞候補の同定

照応詞候補に対して最尤先行詞候補を決めるモデルとして，我々がが提案したトーナメントモデル[20] を用いる．このトーナメントモデルでは照応詞候補に対して，先行するすべての名詞句のうちでどれが最も先行詞らしいかを決定するために先行詞候補間の勝ち抜き戦を行い，最尤先行詞候補を決定する．

☆ 本研究では，文脈内照応の関係にある名詞句のみを照応詞とする．外界照応を対象としない理由については 4.1 節で後述する．

表 1 Soon の実験で用いられる素性
Table 1 Feature set used in Soon's experiments.

素性の種類	素性名	詳細
Lexical	SOON_STR	冠詞を除いた名詞句 NP_i と NP_j が一致するならば C．それ以外は I.
Grammatical	PRONOUN_1	NP_i が代名詞ならば Y．それ以外は N.
	PRONOUN_2	NP_j が代名詞ならば Y．それ以外は N.
	DEFINITE_2	NP_j が "the" で始まるならば Y．それ以外は N.
	DEMONSTRATIVE_2	NP_j が "this"，"that"，"these"，"those" で始まるならば Y．それ以外は N.
	NUMBER	NP_i と NP_j の数が一致するならば C．一致しない場合は I．NP の数の情報が決定できない場合は NA.
	GENDER	NP_i と NP_j の性が一致するならば C．一致しない場合は I．NP の性の情報が決定できない場合は NA.
	BOTH_PROPER_NOUN	NP_i と NP_j が共に固有名詞ならば C．片方が固有名詞の場合は NA．それ以外は I.
	APPOSITIVE	NP_i と NP_j が同格の関係にあるならば C．それ以外は I.
Semantic	WNCLASS	NP_i と NP_j が同じ WordNet の意味クラスに属するならば C．属さない場合は I.
	ALIAS	NP の片方が他方の別名である場合は C．それ以外は I.
Positional	SENTNUM	NP 間の文の距離.

NP_i が先行する名詞句を表し，NP_j は照応詞を表す．素性は個々の要素についての素性と要素間の関係についての素性を含んでおり，個々の要素についての素性は，対象となっている NP_i に対してその性質を満たすか（YES）満たさないか（NO）の 2 値をとる．要素間の関係を表す素性は対象としている NP_i-NP_j の対に対して，その性質が矛盾しない（COMPATIBLE），矛盾する（INCOMPATIBLE）の 2 値をとり，その性質が適用できない場合は NOT APPLICABLE の値をとる．

トーナメントモデルの概要について **図 3** を用いて説明しよう．図 3 では，照応詞 ANP に対して，7 つの名詞句 $NP_1, \ldots, NP_7$ が先行文脈に出現している状況を仮定している．NP_2 と NP_4，NP_3 と NP_5，NP_6 と NP_7 はそれぞれ同一指示関係にあり，ANP の先行詞は NP_5（NP_3）とする．ここでは，すでに解析された照応関係を考慮し，ANP に対して 4 つの名詞句 NP_1，NP_4，NP_5，NP_7 を先行詞候補とする．すなわち，NP_4 と照応関係にある NP_2 は候補としない．NP_3，NP_6 も同様である．さて，トーナメントモデルでは，正しい先行詞である NP_5 は他の先行詞候補に対して勝ち残る必要がある．そのため，この関係を学習するために，図 3 (a) に示す 3 つの訓練事例を抽出する．クラス right（left）は与えられた先行詞の候補のうち，右（左）側の候補が勝ちである（より先行詞らしい）ことを示している．

トーナメントモデルで解析を行う際には，照応詞候補に対して先行詞候補となる名詞句の間で勝ち抜き戦を行う．勝ち抜き戦は照応詞候補から文章の先頭に向かって処理する．最初の比較では，図 3 (b) 1. に示すように，最も照応詞候補に近い 2 つの候補 NP_7 と NP_5 を比較し，分類器はより先行詞らしい名詞句を選択する．以降の比較では，1 つ前の比較において勝ち残った（より先行詞らしいと判定された）候補と新たな先行詞候補との比較を行う．たとえば，NP_7 と NP_5 の比較で NP_5 が勝ったとすると，次は NP_5 と新たな候補 NP_4 を比較する（図 3 (b) 2.）．この処理を繰り返し，最後の比較では，文章の先頭に最も近い先行詞候補との比較を行い，勝ち残った候補を与えられた照応詞に対する最尤先行詞候補と決定する．この最尤先行詞候補と照応詞候補を 3.2 節の照応詞認定の入力とする．

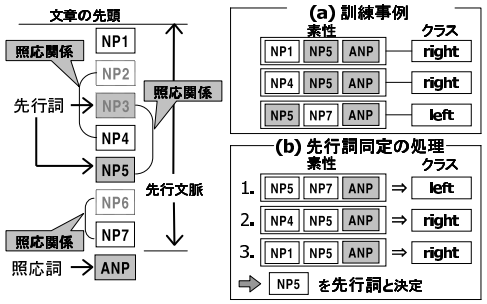

図 3 トーナメントモデル
Fig. 3 The tournament model.

3.2 照応詞の認定

トーナメントモデルを用いて照応詞候補と対となる最尤先行詞候補を決定したのち，次にその対の情報を参照しながら照応詞候補が照応詞か非照応詞かの分類問題を解くことで照応詞を認定する．そのため，照応詞候補が真に照応詞であり，最尤先行詞候補がその先行詞である場合は正例として分類し，照応詞候補と最尤先行詞候補の対が同一指示関係にない（照応詞候補が非照応詞である）場合は棄却する分類モデルを作成する必要がある．訓練事例の抽出を **図 4** を用いて説明しよう．図 4 では，照応詞 ANP に対して 4 つの名詞句（$NP_1, \ldots, NP_4$）が先行文脈に出現している状況を仮定している．ANP の先行詞は NP_2 とする．この状況で，分類器は照応詞候補が照応詞か否かの 2 値分類問題を解く．訓練時には，照応詞とその先

表 2 Soon らの手法と提案手法の比較
Table 2 Comparison between Soon's method and the proposed method.

	訓練事例に非照応詞を利用	解析速度	解析手順
Soon らの手法	利用しない	$O(n)$	照応詞と先行詞を同時に決定
提案手法	利用する	$O(n)$	最尤先行詞候補同定後に照応詞認定

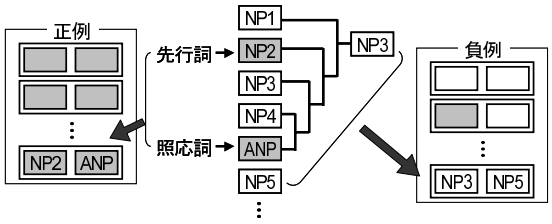

図 4 訓練事例の抽出
Fig. 4 Extracting training examples.

行詞（先行詞が複数ある場合は最も近い先行詞）の対（NP_2-ANP）を正例とする．また，非照応詞である照応詞候補とその照応詞候補の最尤先行詞候補の対を負例とする．図 4 の例で NP_5 が非照応詞だったとすると，NP_5 に対してトーナメントモデルを用いて最尤先行詞候補 NP_3 を決定し，その対（NP_5-NP_3）を負例として訓練事例に追加する．このように訓練事例を作成することで，3.1 節の段階で照応詞に対して先行詞を正しく決定できた場合はその同一指示関係を認定し，また非照応詞の場合には，非照応詞と最尤先行詞候補の対を適切に棄却できると考えられる．

3.3 先行研究（Soon らのモデル）との比較

ここでは，機械学習を用いた同一指示解析の代表例である Soon らの手法と提案手法を比較する．それぞれの手法とその特徴をまとめたものを **表 2** に示す．まず，照応詞を認定する際の非照応詞の利用についてだが，2.2 節に示したように Soon らの手法では非照応詞と先行詞候補の関係について訓練事例を抽出していない．そのため，非照応詞を適切に棄却することは保証されない．それに対し，提案手法では，非照応詞についても最尤先行詞候補との対を負例とすることで，非照応詞が適切に棄却されるように学習を行う．また，Soon らの手法では照応詞の 1 つ前の候補から照応詞に最も近い先行詞までしか訓練事例を作成していないため，先行詞より前方文脈の先行詞候補に対する振舞いが保証されないのに対し，提案手法では前方文脈の候補集合全体から最尤先行詞候補を決定するため，照応詞認定モデルの負例作成の際には先行詞候補全体を考慮して事例作成を行えるという利点がある．また，解析速度についても，前方文脈中の先行詞候補の数を n とした場合，Soon らの手法が照応詞候補と各先行詞候補について同一指示関係となるか否かの分類問題をそれぞれ解くため，計算量は $O(n)$ となるのに対して，提案手法では照応詞候補について最尤先行詞候補を求めるために n 回の勝ち抜き戦を行い，次に 1 回だけ照応詞認定の分類問題を解くため，こちらも計算量は $O(n)$ となり，2 つのモデルの計算量は等しい．

4. 評価実験

4.1 同一指示タグ付与の問題点

作成した同一指示解析モデルを評価するために，名詞句同一指示関係タグ付きのコーパスが必要となるが，同一指示関係のタグを付与する際，以下に示すような問題を考慮する必要がある．

4.1.1 総称名詞や不定名詞の同一指示関係

総称名詞や不定名詞に関しても同一指示関係と認定可能な場合がある．たとえば，次の例の総称名詞「**図書館** $_1$」と総称名詞「**図書館** $_2$」は同一の概念を指しているために同一指示関係として認定すべきかもしれない．

> **図書館** $_1$ には**本** $_1$ というものが置いてある。
> **図書館** $_2$ の**本** $_2$ は借りることができる。

これに対し，総称名詞「**本** $_1$」と総称名詞「**本** $_2$」の場合は，「**本** $_1$（というもの）」が「本を意味する類に属するすべての要素」を指すのに対し，「**本** $_2$」は「図書館の本（図書館に置いてある本）」を指し，“**本** $_1$ ⊃ **本** $_2$” という包含関係が成立つため，「**本** $_1$」と「**本** $_2$」を同一指示の関係として認定するか否かの判断が困難となる．また不定名詞においても同様の現象が起こる．

4.1.2 外界照応

指示代名詞（「それ」など）や人称代名詞（「私」など），指示連体詞（「その」など）は文章外の要素と同一指示関係にある場合がある．たとえば，次の例文において「**その角**」は文章外のある場所を指示しているため外界照応の関係にある．

> 銀行は**その角**を曲がったところにあります。

この外界照応の関係を名詞句同一指示解析の対象に含める，つまり，ある照応詞が文脈外の要素と同一指示関係にある場合，その関係をタグ付与するか否かが問題となるが，指示連体詞などの明示的な手がかり語がない名詞句，たとえば定名詞句「村山首相」のよう

な場合でも，外界のある人物と外界照応の関係にあると見なすことができるため，ある文章内の名詞句がどのような場合に外界照応の関係となっているかを揺れなくタグ付与することは困難であると考えられる．

4.1.3 複合名詞句の構成素

照応詞もしくは先行詞の候補となる名詞句が複合名詞句である場合，その名詞句の構成素を解析の対象とするか否かが問題となる．たとえば，「[八重洲 東] [駐車 場]」という 4 つの形態素を含む複合名詞句を考えた場合，「八重洲」，「[八重洲 東]」，「[駐車 場]」などの構成素も解析の対象となるが，たとえば，「[八重洲 東][駐車 場]」から切り取った構成素「[駐車 場]」はそれ単体で総称の名詞句として解釈するか，それとも「[八重洲 東]」から修飾された「[駐車 場]」として解釈するのかを決定する必要がある．前者の場合は，「[駐車 場]」と「パーキングエリア」などの概念間の同義性を厳密に定義できなければタグを付与することができず，また後者の場合は，「[八重洲 東][駐車 場]」と「[駐車 場]」に同じタグを付与することになり冗長である．

4.2 同一指示タグ付与の基準

4.1 節にあげたような関係を同一指示関係と解釈するか否かは議論の余地があるが，同一指示関係であると考えたとしても，人が揺れなく判断することさえ容易ではない．そこで，今回は 4.1 節に示した関係を便宜的に同一指示関係にないと見なすことにし，以下に示すような 3 つの名詞句同一指示関係のタグ付与の基準を設定した☆1．

- 総称名詞と不定名詞は照応詞，先行詞として考えない．
- 談話内に出現した名詞句のみを先行詞とする．
- 照応詞と先行詞☆2は文節の主辞（最右の名詞自立語）のみを対象とする．

ただし，便宜的とはいえ，今回の同一指示解析の処理は，[1] 対象名詞句が次の (a) または (b) である場合を棄却して，[2] 対象名詞句が前方文脈中に先行詞を持つ定名詞句である場合だけを掬いとるタスクということができ，言語学的な意味は失っていない☆3．

また，上の 3 つの基準に従うと同一指示関係のタグが付与されていない名詞句（非照応詞）は，(a) 総称名詞もしくは不定名詞である名詞句，もしくは，(b) 文脈の前方に先行詞がない定名詞である名詞句のいずれかとなる．(a) の場合は，照応詞認定の処理（3.2 節）で照応詞候補とその最尤先行詞候補の対の両方の情報を参照することにより，総称名詞（もしくは不定名詞）であることを推定し棄却できる見込みがある．(b) の場合も，定名詞と最尤先行詞候補のそれぞれの意味属性が明らかに異なる場合は棄却可能であると考えられる．

4.3 訓練・評価データ

4.2 節のタグ付けの基準に従い，京大コーパス[14] の報道 90 記事に対して名詞句同一指示関係のタグを付与した．今回の実験では，883 の名詞句間の同一指示関係を抽出し 10 分割交差検定を行った．

実験では，対象とする文章に対して茶筌[21] と CaboCha[19] を用い形態素解析，固有表現タグ付与，係り受け解析を行った．また，学習器として Support Vector Machine（SVM）[11] を用い，カーネルには線形カーネルを使用した．

4.4 素　　性

実験では**表 3** に示す 4 種の素性を導入した．表に示す素性の多くは 2 つの学習で共通に使用する．ただし，‘*’ で記した素性は 3.2 節で示した照応詞認定でのみ用いるのに対し，‘**’ で記した素性は 3.1 節で示した最尤先行詞候補同定でのみ用いる．すべての素性の値は 2 値で表現し，文字列長や距離の情報のように表現するべき内容が変動するものは，それぞれを異なる素性として用いた．たとえば，文間の距離の場合，同一文内，1 文，2 文，3 文以上をそれぞれ異なる素性とし，また，一致文字列長の場合は，1 文字一致，2 文字一致，3 文字以上一致を異なる素性とした．たとえば，照応詞候補が「富士山」，先行詞候補「山」である場合，文字列の一致に関して抽出する素性は，表 3 の＜ Last_Match(1 文字一致) ＞と＜ Part_Match(1 文字一致) ＞，＜ Head_Match(1 文字一致) ＞となる☆4．

以下で 4 種類の素性についてまとめる．

4.4.1 語彙的な情報を用いた素性（Lexical）

照応詞候補と最尤先行詞候補の 2 つの文字列の “完全一致”，“前方一致”，“後方一致”，“主辞（最右の内容語）一致”，“部分一致”，“構成文字列の一致” などの情報を素性として導入した．これらの素性を用いることにより，たとえば，最尤先行詞候補「村山富一首

☆1 今回導入した同一指示関係の定義については付録 A.1 を参照．

☆2 主辞の品詞の再分類が “非自立”，“形容動詞語幹”，“数”，“接尾–形容動詞語幹” である名詞句は経験的に先行詞となりにくいことが分かっているので，それらの品詞となる名詞句を除いたものを最終的な先行詞の候補として抽出して実験した．

☆3 総称名詞や不定名詞間の同一指示関係を扱う課題や，初出の定名詞の外界照応関係を扱う課題は，今回の課題設定と矛盾することなく後から追加できる．

☆4 つまり，＜ Part_Match(1 文字一致) ＞と＜ Part_Match(2 文字一致) ＞，＜ Part_Match(3 文字以上一致) ＞はそれぞれ別の素性として区別して扱われる．

表 3 実験に用いた素性
Table 3 The feature set used in our experiments.

素性の種類	素性名	詳細
Lexical	BF_COMB*	ANP-NP_i の対の主辞（文節の最右の内容語）の組合せ.
	DOU_MATCH	ANP が “同” を含む場合，“同” を除いた文字列が NP_i に部分一致する場合は Y. それ以外は N.
	FIRST_PERSON_MATCH	ANP，NP_i の最初の形態素が人名であり，かつそれらが完全一致する場合は一致した文字列長.
	COMP_MATCH	ANP，NP_i が完全一致する場合は一致した文字列長.
	LAST_MATCH	ANP が NP_i に対して後方一致する場合は一致した文字列長.
	FIRST_MATCH	ANP が NP_i に対して前方一致する場合は一致した文字列長.
	PART_MATCH	ANP が NP_i に対して部分一致する場合は一致した文字列長.
	LAST_REVERSE_MATCH	NP_i が ANP に対して後方一致する場合は一致した文字列長.
	FIRST_REVERSE_MATCH	NP_i が ANP に対して前方一致する場合は一致した文字列長.
	STRING_MATCH	NP_i が文字の並び順に ANP のすべての文字を含む場合は一致した文字列長.
	HEAD_MATCH	NP_i と ANP の主辞（ただし，“さん”，“氏” などの場合は主辞を移動）が一致する場合は一致した文字列長.
Grammatical	POS	“名詞–固有名詞”，“名詞–サ変接続” のような NP_i（ANP）の品詞.
	DEFINITE	NP_i（ANP）がソ系の代名詞（“それ”，“その”，“そんな” など）である場合は Y. それ以外は N.
	DEMONSTRATIVE	NP_i（ANP）がコ系もしくはア系の代名詞（“これ”，“ここ”，“あの”，“あそこ” など）である場合は Y. それ以外は N.
	PARTICLE	“は”，“が”，“を” のような NP_i（ANP）に続く助詞.
	DOU	NP_i（ANP）が文字列 “同” を含む場合は Y. それ以外は N.
	DEP_PAST**	NP_i（ANP）に対して述語が “た”（過去形）で係る場合は Y. それ以外は N.
	DEP_PRED**	NP_i（ANP）に対して述語が “た”（過去形）以外で係る場合は Y. それ以外は N.
Semantic	NE	NP_i（ANP）の固有表現の種類：PERSON，ORGANIZATION，LOCATION，ARTIFACT，DATE，TIME，MONEY，PERCENT もしくは N/A.
	EDR_HUMAN	NP_i（ANP）が EDR 概念辞書の中の “人間”，“人間の属性” に含まれる語である場合は Y. それ以外は N.
	EDR_PSEDOU_ORG	NP_i（ANP）が EDR 概念辞書の中の “人間または人間と似た振舞いをする主体” に含まれる語である場合は Y. それ以外は N.
	PRONOUN_TYPE	NP_i（ANP）が代名詞である場合，代名詞に応じて 3 つのクラスに分類（e.g. “彼” → PERSON，“そこ” → LOCATION，“それ” → OTHERS）
	SEM_COMB*	ANP と NP_i のそれぞれの主辞の文字列から辞書引きした意味属性の組合せ.
	SEM_MATCH**	ANP と NP_i のそれぞれの主辞の文字列から辞書引きした意味属性が一致する場合は Y. それ以外は N.
	DEP_NE**	NP_i（ANP）に対して “Named Entity +の” が係る場合は Y. それ以外は N.
	DEP_NO**	NP_i（ANP）に対して “NP_j の” が係る場合は Y. それ以外は N.
Positional	SENTNUM_ANP	NP_i と ANP の文間の距離.
	SENTNUM_NPS**	NP_1 と NP_2 の文間の距離.
	DEP_MAIN	NP_i（ANP）が主節に係る場合は Y. それ以外は N.
	BEGINNING	NP_i（ANP）が文頭にある場合は Y. それ以外は N.
	END	NP_i（ANP）が文末にある場合は Y. それ以外は N.

ANP は照応詞を表し，$NP_{i\in\{1,2\}}$ は先行詞候補を表す．素性は個々の要素についての素性と要素間の関係についての素性を含んでおり，個々の要素についての素性は，対象となっている NP_i に対してその性質を満たすか（YES）満たさないか（NO）の 2 値をとる．要素間の関係を表す素性は対象としている NP_1-NP_2 もしくは NP_i-ANP の対に対して，その性質が矛盾しない（COMPATIBLE），矛盾する（INCOMPATIBLE）の 2 値をとり，その性質が適用できない場合は NOT APPLICABLE の値をとる．‘*’ で記された素性は照応詞認定モデルの学習・分類に用いられ，‘**’ で記された素性は最尤先行詞候補同定モデルの学習・分類に用いられる．

相」と照応詞候補「村山首相」の場合は，“主辞の一致” と “構成文字列の一致” の 2 つの一致情報を同一指示関係の根拠として学習・分類することができる．

4.4.2 形態・統語的な情報を用いた素性（Grammatical）

照応詞候補と最尤先行詞候補それぞれの品詞，指示詞，助詞，対象とする名詞句の連体修飾要素の時制を素性として導入した．これらは，たとえば指示連体詞「その」が名詞句に係る場合は，その名詞句は定名詞である可能性が高い，また「昨日摘みとった**果物**は味がいいです。」のように，連体修飾要素が過去形で対象とする名詞句「果物」に係る場合は，定名詞である可能性が高いなどの村田らの規則に基づく．

4.4.3 意味的な情報を用いた素性（Semantic）

照応詞候補と最尤先行詞候補の意味属性が異なる場合は同一指示関係とならない可能性が高い．今回の実

表 4 名詞句同一指示解析の結果
Table 4 Result of resolving NP coreference.

	主辞一致のモデル	Ng らのモデル	提案手法のモデル
精度	31.5% (630/2000)	40.3% (742/1842)	76.7% (582/759)
再現率	71.3% (630/883)	84.0% (742/883)	65.9% (582/883)
F 値	43.7	54.5	70.9%

験では，CaboCha が出力した固有表現のタグや分類語彙表[13] の分類項目を素性として用いた．具体的には，照応詞候補の主辞（最右の内容語）の文字列を含む分類項目と先行詞候補の主辞を含む分類項目の組合せを明示的に素性とした．たとえば，照応詞候補の主辞の文字列が「人」，先行詞候補の主辞の文字列が「者」である場合，それぞれ分類項目として，{ 1.1960-単位，1.2010-自他，1.2020-人間 } と { 1.1000-こそあど，1.2020-人間 } を得る．これらの多義性を解消することは困難なので，今回は可能な組合せをすべて素性とした．つまり，上の 2 つの文字列から抽出される素性は，＜ 1.1960-単位—1.1000-こそあど＞，＜ 1.1960-単位—1.2020-人間＞，＜ 1.2010-自他—1.1000-こそあど＞，＜ 1.2010-自他—1.2020-人間＞，＜ 1.2020-人間—1.1000-こそあど＞，＜ 1.2020-人間—1.2020-人間＞の 6 つとなる．

4.4.4 名詞句間の距離情報を用いた素性（Positional）

照応詞候補と最尤先行詞候補の距離が離れるほど同一指示関係とならない可能性が高い．そこで，照応詞候補と最尤先行詞候補の文間の距離を素性として導入した．

4.5 ベースラインモデル

本研究で提案する解析モデルの有効性を示すため，以下に示す 2 つのベースラインモデルと提案手法のモデルを比較評価した．

- 主辞一致情報を用いた解析モデル
- Ng らの解析モデル

4.5.1 主辞一致情報を用いた解析モデル

規則ベースの最も単純な解析モデルの一例として，照応詞候補と先行詞候補の主辞（文節内の最右の内容語）が一致するか否かの情報を用いたベースラインモデルを作成した．このモデルでは，照応詞候補と先行詞候補のそれぞれの主辞が一致する場合でかつそのときに限り同一指示関係と判断する．ただし，照応詞候補に対して複数の候補の主辞が一致する場合は，照応詞候補から最も近い候補を 1 つだけ選択する．

4.5.2 Ng らの解析モデル

機械学習を用いたベースラインモデルとして，2.2 節に示した Ng らの同一指示解析モデルを使用した．このモデルでは，表 3 に示した素性のうち提案手法の照応詞認定で用いる素性と同じ素性を用い，また学習器には決定木ではなく SVM を使用した．Ng らのモデルでは，照応詞と先行詞候補の対についての関係のみを学習し，非照応詞についての振舞いを学習していない．そのため，このベースラインモデルと提案手法のモデルを比較することで，非照応詞と先行詞候補の関係について学習することの有用性を調査することができる．

4.6 実験結果

評価実験の結果を **表 4** に示す．この評価実験における再現率は照応詞と先行詞の正しい対を認定できた割合であり，精度はモデルが出力した対のうち，照応詞と先行詞の正しい対となっていた割合となるため，表 4 の精度，再現率，F 値はそれぞれ以下の式を用いて求められる．

$$\text{精度} = \frac{\text{正しく同一指示解析できた個数}}{\text{システムが照応詞と判断した候補数}},$$

$$\text{再現率} = \frac{\text{正しく同一指示解析できた個数}}{\text{照応詞の総数}},$$

$$\text{F 値} = \frac{2 \times \text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}}.$$

結果より，主辞一致のモデルや Ng らのモデルは提案手法のモデルより再現率が良いことが分かる．しかし，ベースラインとした 2 つのモデルはともに過剰に照応詞と先行詞の対を認定しているために精度が悪く，結果として F 値が低くなっていることが分かる．これに対し，提案手法のモデルでは 3 章に示した事例作成の工夫により，非照応詞を精度良く棄却できていることが分かる．

提案手法では最初に照応詞候補に対して最尤先行詞候補を同定したのちに照応詞を認定するが，最初の処理，つまり最尤先行詞候補の同定において，照応詞に対して正しく先行詞を同定する精度は 86.6% (765/883) であった．この結果から，提案手法の方が他の 2 つのモデルより精度良く先行詞が同定できているが，照応詞認定の処理で正解となる対も含めて過剰に棄却しているため，最終的な再現率は他の 2 つのモデルより悪くなっている．

また，訓練事例の量を変えたときの再現率，精度，

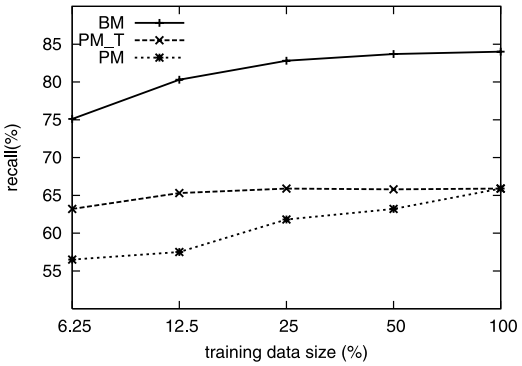

図 5 学習曲線（再現率）
Fig. 5 Learning curve (Recall).
BM：ベースラインモデルの訓練事例を変動,
PM_T：最尤先行詞候補同定の訓練事例を変動,
PM：照応詞認定の訓練事例を変動

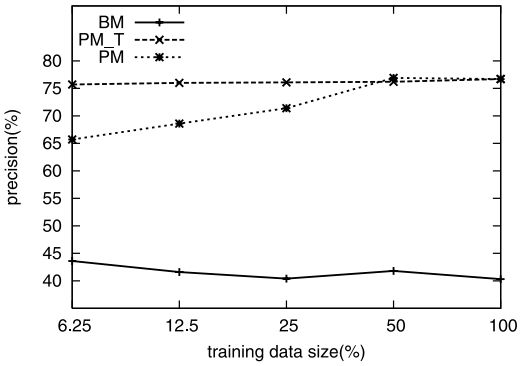

図 6 学習曲線（精度）
Fig. 6 Learning curve (Precision).
BM：ベースラインモデルの訓練事例を変動,
PM_T：最尤先行詞候補同定の訓練事例を変動,
PM：照応詞認定の訓練事例を変動

F 値の学習曲線をそれぞれ **図 5**，**図 6**，**図 7** に示す．提案手法では最尤先行詞候補同定モデルの訓練事例と照応詞認定モデルの訓練事例の両方を変動させることができるので，前者を変動させたものを図中 PM_T で示し，後者を変動させたものを PM で示す．また Ng のモデルの訓練事例を変動させたものを BM で示す．3 つの図を通して PM_T は事例を増減しても変化がないことが分かる．逆に PM は事例が増加するに従いモデルの質が向上していることが分かる．PM と PM_T が異なる原因の 1 つとして，モデルの学習に用いる訓練事例の総数の違いがあげられる．最尤先行詞候補同定モデルの事例作成は，3.1 節に示したように，ある照応詞に対して先行詞と他の先行詞候補の組合せすべてを訓練事例とする．それに対して，照応詞認定モデルの事例作成は，3.2 節で述べたように，1 つの照応詞候補に対して作成できる訓練事例は 1 つだけである．

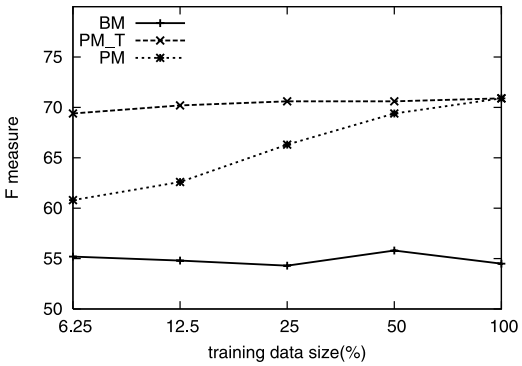

図 7 学習曲線（F 値）
Fig. 7 Learning curve (F-measure).
BM：ベースラインモデルの訓練事例を変動,
PM_T：最尤先行詞候補同定の訓練事例を変動,
PM：照応詞認定の訓練事例を変動

表 5 照応詞の種類と再現率の関係
Table 5 Recall of each anaphor type.

種類	(a) 最尤先行詞候補同定	(a) + 照応詞認定
固有表現	94.8% (368/388)	84.3% (327/388)
普通名詞	81.5% (392/481)	52.8% (254/481)
代名詞	35.7% (5/14)	7.1% (1/14)

そのため，今回の実験で用いた訓練事例は，最尤先行詞候補同定の訓練事例が約 4 万であるのに対して，照応詞認定の訓練事例は約 6 千程度であった．そのため，照応詞認定の問題については，訓練事例を増やすことで再現率や精度が向上するかを確かめる必要がある．

次に，表 4 のうち，提案手法のモデルの解析結果を照応詞の種類で分類した（**表 5**）．ただし，固有表現には CaboCha の出力する IREX の 8 種の固有表現を，また代名詞には茶筌の解析結果を用いた．固有表現，代名詞に分類しなかった名詞句を普通名詞とした．表 5 より，照応詞が固有表現である場合に解析精度が最も良いことが分かる．固有表現の場合は照応詞と先行詞の文字列が一致する場合が多いため，その特徴をうまく学習できたと考えられる．代名詞に関しては，訓練事例が少なく，またセンタリング理論[4] などの言語学的な手がかりを直接的に用いていないために，代名詞がどのような場合にどの先行詞候補と同一指示関係になりやすいかの傾向が学習できなかったと考えられる．

4.7 誤り分析

実験の結果より，先行詞同定の解析誤りと照応詞認定の解析誤りそれぞれについて人手で分析した結果を報告する．

表 6 先行詞同定における誤りの割合
Table 6 Error distribution in the antecedent detection.

誤りの原因	割合
(1) 名詞意味属性の適用	35.6% (42/118)
(2) 特徴的な語の過剰な重み	16.9% (20/118)
(3) 文字列素性が過剰に働く	18.6% (22/118)
(4) 文章内外の情報が必要	15.3% (18/118)
(5) 定名詞の推定誤り	9.3% (11/118)
(6) その他	22.9% (27/118)

4.7.1 先行詞同定の解析誤り

最尤先行詞候補同定の処理において，照応詞に対して先行詞同定を誤った 118 事例☆1を分析した結果を**表 6** に示す☆2．表 6 のうち，典型的な誤りについて以下にまとめる．

(1) の名詞意味属性の適用の問題については，照応詞と先行詞は同じ意味属性を持つことが望ましいが，素性として導入する意味属性の粒度やある意味属性に含まれる語の集合は用いる言語資源に依存する．質の良い意味属性を用いれば，照応詞「今年」に対して名詞句「事業」のような明らかに異なる意味属性に分類されている名詞句を先行詞候補から除外できる可能性がある．しかし，現在入手可能な既存のシソーラスを使う限り,「兄」と「妹」のように，明らかに同一指示関係にならない対が同一の意味属性を持つケースも少なくない．今後は，同一指示解析に必要な意味属性の分け方や，それらがどの程度文脈に依存するかを明らかにする必要がある．

次に表 6 (2) の問題を説明しよう．SVM で学習して得られたモデルそのものを分析したところ，先行詞候補が指示代名詞や人称代名詞，接頭辞「同」である素性は他の素性に比べ過剰に重みが付与されていることが分かった．このため，これらの素性を含む先行詞候補が正解の先行詞よりもより先行詞らしいと判断されてしまい解析を誤ることになる．具体例を下の文章を用いて説明する．この文章では，下線部「現行制度」が照応詞であり，下線部「受信量制度」がその先行詞である．また太字の「それ」が誤って先行詞と解析した候補である.「受信料制度」と「それ」はそれぞれ先行詞候補となるが，今回の手法では「それ」自体が何と同一指示関係にあるか，もしくはどの候補とも関係がないかが分からない状態で学習・分類をしている☆3．

☆1 全事例 883 事例から正解の 765 事例を引いた値.
☆2 これらの問題は同時に起こる場合もあるため，その場合は 1 つの事例を複数のカテゴリに分類した.
☆3 同一指示解析の結果から得られる情報を用いてもよいが，すべての照応詞候補が正しく解析できるとは限らないので，そのまま扱うとノイズになる可能性がある.

さらに，今回対象とした訓練・評価データでは，先行詞に「それ」のような語を選ぶ事例が選ばない事例と比較して極端に多かったため，それらの文字列が素性として抽出された場合，その素性に過剰な重みが付与されてしまい，その結果，先行詞候補が特定の文字列を含む場合には誤って最尤先行詞候補として同定されてしまった．このような事例については，Ng らのように指示詞などの場合は候補から除くという選択肢もあるため，先行詞候補の取捨選択という問題として再度考えたい．

「パソコンで NHK 放送を見ても受信料は必要。しかし、今後、受信料制度 の抜本的な見直しが必要」。マルチメディアパソコン時代の NHK 受信料について、郵政省放送行政局がこんな見解をまとめた。五日までに同局が作成した「パソコン画面上のテレビ映像について」で明らかになった。**それ**によると、現行法令上は受信契約が必要としている。ただし「受信の意思がなくパソコンを買った人に受信契約義務が生じる 現行制度 の再検討が必要」と、直ちに受信料徴収に乗り出すわけではないことも明記している。

さらに，今回導入した文字列一致の素性が過剰に働くために解析を誤る場合も見られた（表 6 (3)）．文字列の素性は固有表現の特徴をとらえるためには有効であると考えられるが，逆に照応詞候補「キリスト教会」に対して「キリスト教会色」のように明らかに同一指示関係にない先行詞候補を最尤先行詞候補として決定してしまう．また，表層文字列が同じ名詞句（たとえば,「2 人」など）が複数回出現して，かつそれらが異なる談話実体を指す場合も誤って最尤先行詞候補とする傾向が見られた．

4.7.2 照応詞認定の解析誤り

照応詞認定の解析誤りを分析する際には解析の信頼度☆4を導入し，その信頼度が高く，かつ解析を誤った 50 事例を分析した．分析結果を**表 7** に示す．照応詞認定の誤りの多くは照応詞候補（もしくは最尤先行詞候補）が定名詞か否かの特徴をとらえることができず，総称名詞（もしくは不定名詞）である場合でも同一指示関係を認定するという誤りであった．誤りの具体例を下の文章を用いて説明しよう．この文章で太字の「女性」が照応詞候補であり，下線部の「女性」が最尤先行詞候補として同定された候補である．この 2 つの「女性」は同一指示関係にはないが，今回導入し

☆4 解析の信頼度には照応詞認定のモデルが出力する分離平面からの距離を用いた.

表 7 照応詞認定における誤りの割合
Table 7 Error distribution in the anaphor identification.

誤りの原因	割合
(1) 定名詞の推定誤り	50.0% (25/50)
(2) 文字列素性が過剰に働く	14.0% (7/50)
(3) 文章内外の情報が必要	12.0% (6/50)
(4) その他	22.0% (11/50)

た素性だけでは照応詞候補が非照応詞であることの特徴をとらえることができず，誤って同一指示関係として解析している．このような誤りが多数存在するため，名詞句の指示性をとらえるための手がかりを調査することこそ今後我々が取り組むべき課題の 1 つであるといえる．

戦後五十年間で女性が一生の間に産む子供の数は三分の一に減り、働く場所は家の中から外へ――。総理府男女共同参画室が三日付で発表した「女性の歩み五十年」で、戦後における**女性**の地位や生活の変化が改めて浮き彫りになった。

5. 関連研究との比較

英語の規則に基づく手法として，Baldwin [2] の手法がある．この手法では，名詞句同一指示の先行詞同定規則を人手で記述し，高い精度で同一指示関係を同定している．Baldwin は，実際に物語文の照応の現象を観察し，統語・意味役割のレベルで 8 つの規則を導入している．精度 92%，再現率 64%という結果が報告されているが，規則適用の制約が厳しすぎるために再現率が低くなるという問題がある．一方，日本語の規則に基づく手法には，村田ら[16),17)] の手法がある．この手法では，最初に名詞句の指示性を推定し，次に定名詞と推定した名詞句間の同一指示関係を同定する．物語文などを対象に同一指示解析の実験を行い，精度 79%，再現率 77%という結果が報告されているが，2 章でも述べたように先行詞候補群から得られる情報を有効に利用できていない．これに対し，本提案手法では，最初に最尤先行詞候補を特定し，最尤先行詞候補と照応詞候補の対を用いることで，より多くの文脈情報を考慮したうえで照応詞か否かを分類することが可能である．

英語の機械学習ベースの手法としては，2.2 節で述べた Soon らや Ng らの解析手法がある．この手法では，先行詞候補それぞれに対して照応詞候補が真に照応詞であるか否かの分類問題を考える．Ng らは MUC-7 の照応解析タスクにおいて，精度 78.0%，再現率 64.2%という結果を得たが，非照応詞に対して学習事例を抽出していないため，非照応詞が適切に棄却されることが保証されていない．そのため，4.6 節の結果からも分かるように過剰に同一指示関係を同定してしまう．これに対し，本手法では，非照応詞に関しても最尤先行詞候補との対を負例として抽出するため，先行研究の手法と比較して精度良く非照応詞を棄却できる．一方，日本語のための機械学習ベースの手法として，Aone ら[1)] の手法がある．彼女らは，ゼロ代名詞，固有名詞，限定詞を対象に，決定木を用いた同一指示解析手法を提案した．新聞記事コーパスを用いて評価実験を行い，約 90%の精度を得ている．しかし，この実験では，組織を照応するものに限定されている．これにより，照応詞候補や先行詞候補を会社名など組織に関係する名詞句に限定すればよく，この制約により候補を削減することができる．

6. おわりに

本稿では，最尤先行詞候補を同定したうえで照応詞を認定する名詞句同一指示解析手法を提案した．新聞報道記事に対して同一指示解析の実験を行い，再現率 65.9%，精度 78.4%を得た．照応詞が固有表現である場合に解析精度 84.3% を得たが，普通名詞の指示性については表層文字列から得られる単純な素性を用いただけでその特徴をとらえることが困難なことが結果より分かる．今後は，同一指示解析においてもセンタリング理論で導入されている局所文脈の情報をより直接的に用い，かつ同一指示の関係を同定するためにより適切な意味の粒度を調査する必要もある．さらに名詞句の指示性と新情報と旧情報の観点からの名詞句の分類[8)] の関係を明らかにし，解析対象の候補選択の問題にも取り組みたい．

参考文献

1) Aone, C. and Bennett, S.W.: Evaluating automated and manual acquisition of anaphora resolution strategies, *Proc. 33rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.122–129 (1995).
2) Baldwin, B.: CogNIAC: A Discourse Processing Engine, Ph.D. Thesis, Department of Computer and Information Sciences, University of Pennsylvania (1995).
3) Ge, N., Hale, J. and Charniak, E.: A Statistical Approach to Anaphora Resolution, *Proc. 6th Workshop on Very Large Corpora*, pp.161–170 (1998).
4) Grosz, B.J., Joshi, A.K. and Weinstein, S.:

Centering: A framework for modeling the local coherence of discource, *Computational Linguistics*, Vol.21, No.2, pp.203–226 (1995).
5) Karttunen, L.: Discourse referents, *Syntax and Semantics 7: Note from the Linguistic Underground*, McCawley, J.(Ed.), pp.363–386, Academic Press, New York (1971).
6) Mitkov, R.: Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches, *Proc. ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution* (1997).
7) Ng, V. and Cardie, C.: Improving Machine Learning Approaches to Coreference Resolution, *Proc. 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.104–111 (2002).
8) Prince, E.F.: Toward a taxonomy of given-new information, *Radical Pragmatics*, Cole, P.(Ed.), pp.223–255, Academic Press, New York (1981).
9) Quinlan, J.R.: *C4.5: Programs for Machine Learning*, The Morgan Kaufmann Series in Machine Learning, Morgan Kaufmann (1993).
10) Soon, W.M., Ng, H.T. and Lim, D.C.Y.: A Machine Learning Approach to Coreference Resolution of Noun Phrases, *Computational Linguistics*, Vol.27, No.4, pp.521–544 (2001).
11) Vapnik, V.N.: Statistical Learning Theory, *Adaptive and Learning Systems for Signal Processing Communications* and control, John Wiley & Sons (1998).
12) Yang, X., Zhou, G., Su, J. and Tan, C.L.: Coreference Resolution Using Competition Learning Approach, *Proc. 41st Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.176–183 (2003).
13) 国立国語研究所：分類語彙表，Vol. 国立国語研究所資料集 6，秀英出版 (1993).
14) 黒橋禎夫，長尾　眞：京都大学テキストコーパス・プロジェクト，言語処理学会第 3 回年次大会発表論文集，pp.115–118 (1997).
15) 山梨正明：推論と照応，くろしお出版 (1992).
16) 村田真樹，黒橋禎夫，長尾　真：表層表現を手がかりとした日本語名詞句の指示性と数の推定，自然言語処理，Vol.3, No.4, pp.31–48 (1996).
17) 村田真樹，長尾　真：名詞の指示性を利用した日本語文章における名詞の指示対象の推定，自然言語処理，Vol.3, No.1, pp.67–81 (1996).
18) 大塚高信，中島文雄：新英語学辞典，研究社 (1992).
19) 工藤　拓，松本裕治：Support Vector Machine を用いた Chunk 同定，自然言語処理，Vol.9, No.5, pp.3–21 (2002).
20) 飯田　龍，乾健太郎，松本裕治：文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定，情報処理学会論文誌，Vol.45, No.3, pp.906–918 (2004).
21) 松本裕治，北内　啓，平野善隆，松田　寛，高岡一馬，浅原正幸：形態素解析システム『茶筌』version 2.2.9 使用説明書，奈良先端科学技術大学院大学 (2002).

付　　録

A.1 同一指示関係の定義

同一指示（coreference）関係は言語哲学や意味論などの分野において，さまざまな異なった定義が導入されており，統一した解釈が定められていない．そこで以下では，本研究で対象とする同一指示関係を取り決め，その範囲内で問題を考える．同一指示の関係を定義するため，まず指示（reference）と談話要素（referent）について，新英語学辞典[18]の解釈を下に示す．

語の意味として外延（denotation）と内包（connotation）とを区別するとき，前者すなわち，外的世界にあって指示対象となるものをその語の referent と呼び，それを指示する機能あるいは作用を reference と呼ぶ．

この解釈に従った談話要素とは，Karttunen[5]の discourse referent や Prince[8]の discourse entity と等価なものと考える．つまり，外的世界（もしくはある仮想世界）へ写像する際の機能が ‘指示’ であり，写像したある世界での要素が ‘談話要素’ となる☆．本研究では，この解釈を採用し，次のように同一指示の関係を定義する．

2 つの名詞句が同一指示の関係にあるとは，それらの談話要素が同一であることである．

これに対して，照応（anaphora）☆☆とは，文章内に出現したある要素が前文脈に出現した要素を指す現象のことをいう．たとえば，文 (1) において，‘*she*’ は前文脈の ‘*the Queen*’ を指している．

(1) *The Queen* is not here yet but *she* is expected to arrive in the next half an hour.

次に，同一指示と照応の差分を示すことで 2 つの概念の関係をまとめる．同一指示の関係にあるとは，文章を越えてある 2 つの名詞句をある空間に写像したう

☆ この写像が現実世界への写像であるかの議論には立ち入らない．つまり，‘The present king of France is bold.’ という文において ‘present king of France’ が実世界にすでに存在しないために写像することが不可能であるという問題は起こらない．

☆☆ 山梨[15]は reference の訳語として “照応” という用語を用いているが，ここでは anaphora のみを指すこととする．

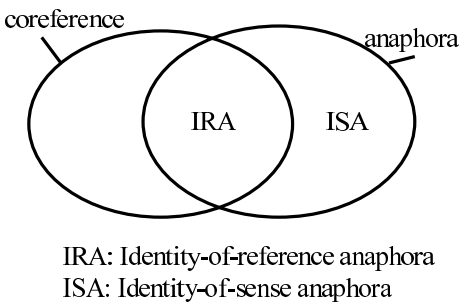

図 8 IRA と ISA を用いた coreference と anaphora の関係
Fig. 8 The relation between coreference and anaphora using the concepts of IRA and ISA.

えでそれらの談話要素が同一であるかを定めることが可能な関係である．また，同一指示関係の場合，2 つの談話要素間に方向性は存在しない．これに対して，照応関係にあるとは，ある閉じた文章内で対象となる名詞句が照応詞である場合，その照応詞は前方文脈に対して先行詞を考えるといった方向性のある関係となる．照応関係は同一指示の観点から大きく identity of reference anaphora （IRA）と identity of sense anaphora（ISA）に分けることができる．

IRA は，同一指示関係かつ照応関係にある関係である．文 (2) では，‘**it**$_i$’ が前方文脈に出現している ‘**their cottage**$_i$’ を指しており，かつ，同一指示関係にある．

(2) In Barcombe, East Sussex, a family had to flee **their cottage**$_i$ when **it**$_i$ was hit by lightning.

これに対して，ISA は語義レベルでの等価性のみしか問題としない．つまり，照応関係にはあるが，同一指示の関係にはない関係である．文 (3) の ‘**the one**’ と ‘**it**’ はそれぞれ ‘the man’ と ‘his hair’ を指しているが，それぞれの名詞句をある空間に写像した際，同一の関係になっていないため，同一指示関係にはない．

(3) The man who has his hair cut at the barber’s is more sensible than **the one** who **it** done at the hairdresser’s.

この IRA と ISA を用いて同一指示，照応の概念の包含関係をまとめると**図 8** のようになる．

近年の機械学習を用いた同一指示解析[7),10),12)] では，同一指示の関係を IRA の関係ととらえ，照応詞と先行詞の存在を仮定して解析を行っている．本研究でも同様に，同一指示関係となる 2 つの談話要素について，先に出現した要素を先行詞，後に出現した要素を照応詞として，前方照応解析の枠組で同一指示解析の問題を考える．たとえば，以下の (4)，(5)，(6) の文章において，照応詞はそれぞれ，固有表現，普通名詞，代名詞である．本研究では，このような名詞句すべてを解析の対象とする．

(4) 村山富市首相$_i$ は三十一日夕、日航機で羽田空港に到着した。
　村山氏$_i$ は記者団に対し、「...」と感想を述べた。

(5) 村山富市首相$_i$ は三十一日夕、日航機で羽田空港に到着した。
　首相$_i$ は記者団に対し、「...」と感想を述べた。

(6) 村山富市首相$_i$ は三十一日夕、日航機で羽田空港に到着した。
　彼$_i$ は記者団に対し、「...」と感想を述べた。

(平成 16 年 5 月 26 日受付)
(平成 17 年 1 月 7 日採録)

飯田　龍（学生会員）
1980 年生．2002 年九州工業大学情報工学部知能情報工学科卒業．現在，奈良先端科学技術大学院大学情報科学研究科博士後期課程在学中．自然言語処理の研究に従事．

乾　健太郎（正会員）
1967 年生．1995 年東京工業大学大学院情報理工学研究科博士課程修了．同年より同研究科助手．1998 年より九州工業大学情報工学部助教授．1998〜2001 年科学技術振興事業団さきがけ研究 21 研究員を兼任．2001 年より奈良先端科学技術大学院大学情報科学研究科助教授．現在にいたる．博士（工学）．自然言語処理の研究に従事．言語処理学会，人工知能学会，電子情報通信学会，ソフトウェア科学会各会員．

松本　裕治（正会員）

1955 年生．1977 年京都大学工学部情報工学科卒業．1979 年同大学大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．1984〜1985 年英国インペリアルカレッジ客員研究員．1985〜1987 年（財）新世代コンピュータ技術開発機構に出向．京都大学助教授を経て，1993 年より奈良先端科学技術大学院大学教授．現在にいたる．工学博士．専門は自然言語処理．人工知能学会，日本ソフトウェア科学会，言語処理学会，認知科学会，AAAI，ACL，ACM 各会員．

関根　　聡（正会員）

Assistant Research Professor, New York University．1987 年東京工業大学応用物理学科卒業．同年松下電器東京研究所に入社．1990〜1992 年 UMIST 客員研究員．1992 年 UMIST 計算言語学科修士．1994 年から NYU, Computer Science Department，Assitant Research Scientist．1998 年 Ph.D. 同年から現職．自然言語処理の研究に従事．コーパスベース，パーザ，分野依存性，情報抽出，情報検索等に興味を持つ．言語処理学会，人工知能学会，ACL 各会員．