修 士 論 文

# 新聞記事の固有表現を対象とした参照関係の解析

指導教官　白井 清昭

北陸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

佐竹 正臣

2002 年 2 月 15 日

## 要 旨

本稿では，新聞記事の固有表現抽出の精度を改善するために，固有表現の照応解析を行う手法を提案する．現在の固有表現抽出技術では，固有名詞が同一の対象を指しているとき，同じ固有表現タグを付与するようにタグの整合性を取ることが試みられていない．このため，同一の対象を表す固有表現が抽出されない場合や同一の対象に対して同じタグが付与されない場合があった．このような問題を解決するために，固有表現を対象とした新しい照応解析手法を提案し，同一の対象を表わす固有表現を特定する．さらに，同一の対象を表わす固有表現に付与する固有表現タグの整合性を取ることにより，固有表現抽出の精度を改善させることを目的としている．

# 目 次

# 図 目 次

# 表 目 次

# 第 1 章

# はじめに

## 1.1 研究の背景と目的

固有名詞に組織名や人名などの属性タグを付与する固有表現抽出は，テキスト処理における基礎的な技術として重要である．特に新聞記事には，時間表現や数値表現などの固有表現が多く含まれているため，新聞記事を対象に固有表現抽出を行なう研究が数多く行なわれている．固有表現抽出の先行研究の多くは固有名詞の周辺にある単語の情報を手がかりに，固有表現タグを付与する規則を自動的に学習している [9][4][15]．また，固有表現タグの付与は，同一文書にある他の固有表現に対するタグの付与とは独立に行なわれる場合が多い．そのため，以下に挙げる 2 つの問題点がある．

- 同一の対象を表す固有表現が抽出されない

  例えば，同一文章中に「公正取引委員会」と「公取委」という 2 つの固有名詞があるとする．このとき，前者には “組織名” というタグを付与するが，後者には固有表現タグを付与しない，つまり固有名詞として抽出されない場合がある．しかし，これらは共に固有表現として抽出し，同じ固有表現タグを付与するべきである．

- 同一の対象に対して同じタグが付与されない

  例えば，同一文章中に「山岸章」と「山岸」という固有名詞があり，両者は同一の対象を表しているとする．しかし，従来の固有表現抽出技術では，独立に固有表現タグを付与するため，前者に “人名”，後者に “組織名” といったように，異なる固有表現タグを付与する可能性がある．両者は同一の対象を表しているので，これらには同じ固有表現タグを付与するべきである．

このような問題に対処するために，本研究では新聞記事を対象に記事内の固有名詞の照応解析を行ない，その結果を利用して固有表現抽出の精度を向上させることを目的とする．具体的には，照応解析によって同一の対象を表す固有表現を特定し，それらに同一の固有表現タグを付与するように初期の固有表現抽出結果を修正する．また，固有表現を対象とした新しい照応解析アルゴリズムも提案する．

## 1.2 本論文の構成

2 章では，固有表現抽出と照応解析のそれぞれの関連研究について説明し，3 章では，固有表現抽出システムの実装と各モジュールについて説明する．4 章では，照応解析手法について述べる．5 章では，システムの評価実験結果を示し，提案手法と従来法との比較や考察を述べる．6 章では結論と今後の課題を述べる．

# 第 2 章

# 関連研究

## 2.1 固有表現抽出

固有名詞に組織名や人名などの属性タグを付与する固有表現抽出は，テキスト処理における基礎的な技術として重要である．また，固有表現抽出は情報抽出における基礎的な技術として認識されているだけではなく，形態素，構文解析の精度向上にもつながる技術でもある．

固有表現抽出は近年活発に研究されている研究分野である．特に MUC(Message Understanding Conferences) の Named Entity Task では様々な団体が参加して行なわれた．MUC(Message Understanding Conferences) は米国国防総省をスポンサーとする Tipster プログラムの一環として行なわれている情報抽出国際会議であり，いくつかのタスクの 1 つが Named Entity Task である．MUC は主に英語を対象としていたが，それから派生した MET(Multilingual Entity Task) は日本語をはじめとして，いくつかの言語を対象とした固有表現抽出のコンテストで，MUC6，7 と並行して MET1，2 が開催された．それから日本で独自のコンテストを開催する気運が高まって，IREX(Information Retrieval and Extraction Exercise) が開催され，その一課題として固有表現抽出課題が設定され，様々な方法で精度を競った．

このタスクで固有表現として抽出するのは,「共和党」のように組織の名称を表すもの,「小泉純一郎」のように人名を表すもの,「東京銀座」のように地名を表すもの,「ノーベル賞」のように固有物の名称を表すもの,「５月１３日」のように日付を表すもの,「午前７時」のように時間を表すもの,「５００億円」のように金額を表すものや「０．５%」のように割合を表すものである．固有表現抽出は，新聞記事などのテキストに対して，以下に挙げる例のような固有表現タグを付与することによって，固有表現を抽出するとともに，

固有表現の内容も分類する技術である.

- <ORGANIZATION> **組織名** </ORGANIZATION>

  複数の人間で構成され，共通の目的を持った組織等の名称

  例えば,「日本銀行」や「ＪＲ東日本」

- <PERSON> **人名** </PERSON>

  固有の人を指す名前

  例えば,「森喜朗」や「クリントン」

- <LOCATION> **地名** </LOCATION>

  固有の場所を指す名称

  例えば,「岐阜県大垣市」や「国道１号」

- <ARTIFACT> **固有物名** </ARTIFACT>

  人間の活動によって作られた固有の物の名称

  例えば,「ペンティアムプロセッサ」や「サンフランシスコ平和条約」

- <DATE> **日付表現** </DATE>

  単位が２４時間以上のもの

  例えば,「２００１年5月１３日」や「秋」

- <TIME> **時間表現** </TIME>

  単位が２４時間以下のもの

  例えば,「午前７時」や「正午」

- <MONEY> **金額表現** </MONEY>

  金額を表す表現

  例えば,「４０ドル」や「１１４円」

- <PERCENT> **割合表現** </PERCENT>

  割合を表す表現

  例えば,「２０%」や「2倍」

図 2.1: 固有表現抽出の例

<DOC>
<DOCNO>940413099</DOCNO>
<SECTION>経済</SECTION>
<AE>無</AE>
<WORDS>147</WORDS>
<HEADLINE><ORGANIZATION>タカラ</ORGANIZATION>社長に副社長の<PERSON>佐藤博久</PERSON>氏が昇格</HEADLINE>
<TEXT>
　おもちゃ大手の<ORGANIZATION>タカラ</ORGANIZATION>は<DATE>十二日</DATE>、<PERSON>佐藤博久</PERSON>副社長が社長に昇格する人事を内定した。創業者の<PERSON>佐藤安太</PERSON>社長は会長に就任する見通し。<PERSON>博久</PERSON>氏は、<PERSON>安太</PERSON>氏の長男。正式決定は<DATE>六月下旬</DATE>。
　◇<PERSON>佐藤博久</PERSON>氏（さとう・ひろひさ）<DATE>１９７９年</DATE>慶大法卒、<DATE>８０年</DATE><ORGANIZATION>タカラ</ORGANIZATION>入社。常務などを経て<DATE>９２年４月</DATE>から副社長。<LOCATION>東京都</LOCATION>出身、３８歳。
</TEXT>
</DOC>

固有表現抽出の例を図 2.1に示す.

固有表現は多様性に富み，また次々と新たに生み出されるためにそのすべてを辞書に登録することは不可能である．そのため，辞書だけを手がかりに固有表現を同定することは不可能である.

固有表現を抽出する方法は大きく分けて２つに分類される.

### 2.1.1　パターン型

パターン型とは人手で明示的なパターンを抽出し，それらを用いたものである．初期の MUC では構文解析等の技術を用いるのが主流であったが，パターンマッチングによる方法の方が性能的に優れていたために，現在ではあまり研究されていない．パターンマッチングは文や文の一部にマッチするパターンを用意しておいて，それを決まった順序で適用

しながら決定的に抽出する方法である．また，パターンマッチングは難しい技術を使用せず，深層的な理解を試みることなく固有表現抽出が簡単に実現できるという利点がある．

しかし，パターン型の最大の問題点はパターンをドメインごとに用意しなければならないことである．例えば,「政治」というドメインと「経済」というドメインとではパターンが異なる場合がある．ドメインによっては大量のパターンを必要とするため，その度にシステムをつくり直さなければならないため，移植性に欠ける．そこで，パターンを自動的に作成する方法である学習型が考えられた．

### 2.1.2 学習型

学習型とは大量の文章を元に品詞や固有表現の出現傾向などの情報を用いて，自動的にパターンを学習するものである．現在労力の少なさから学習型が主流である．まず，学習型の従来研究の 1 つである関根ら [9] の方法について説明する．

関根らは決定木を用いて自動的にパターンを学習する方法をとった．従来の学習システムの問題点は部分的に (1) 手作業のルールを使用すること，(2) 手動で調節しなければならないパラメータを持つこと，(3) 自動的な手段では性能が良くないことが挙げられている．また，(4) 決定木は決定的であるため，固有表現の種類や範囲に矛盾が生じる問題も挙げられている．しかし，関根のシステム (rika と呼ばれる) は JUMAN が出力する品詞，文字型 (漢字，ひらがな，アルファベット，数，記号，それらの組み合わせ)，辞書，固有表現の始め (opening)，続き (continuation)，終り (closing) を表すタグ (例えば，“金沢市片町” は “金沢” が open-cont，“市” が cont-cont，“ 片町” が cont-close となる) を決定木 (C4.5[8]) の入力に与えることで (1) から (3) を解決し，テキスト内で最も確率の高い首尾一貫したタグを選びだすことで，(4) を解決した．また，少量のトレーニングデータで十分パフォーマンスが得られ，移植性も高いことが記されている．

学習型ではこの他，内元ら [4] の最大エントロピーと書き換え規則を使用する方法や，山田ら [15] の Support Vector Machine を使用する方法などがある．

内元ら [4] は最大エントロピーと書き換え規則を用いている．固有表現には一つあるいは複数の形態素からなるものと，形態素単位より短い部分文字列の 2 種類ある．前者の固有表現は固有表現の始まり，中間，終りなどを表すラベルを４０個用意し，そのラベルを最大エントロピーモデルによって推定することによって抽出する．最大エントロピーモデルはデータスパースネスに強いため，大量の学習データがなくても高い精度が得られる．また，後者の固有表現は，学習コーパスに対するシステムの解析結果と正解データとの差異から自動獲得した書き換え規則によって抽出する．実験により，着目している形態素

の前後２形態素の関する見出し語と品詞情報が素性として有効であるとしている．また，固有名詞辞書を利用して，IREX-NE の本試験データに対して，F 値で 80.17(一般ドメイン) の精度を得ている．

また，山田ら [15] らは Support Vector Machine をいている．Support Vector Machine は自然言語処理における様々な問題に対して適用され，他の学習アルゴリズムに比べて，良い成績を収めている．また，過学習に頑健なアルゴリズムとして知られている．Support Vector Machines は二値分類器であるので，これを多値分類に対応できるように拡張した．様々な素性を組み合わせて比較実験したところ，素性として，語彙，品詞細分類，文字種を用いて学習した結果が良いことがわかった．また，固有表現抽出を文頭から文末にかけて行なう右向き解析と，文末から文頭にかけて行なう左向き解析を比較したところ，左向き解析の方が精度が良かった．同一データで比較したわけではないが，同等以上 (F 値：83.01) の精度を得られたと報告している．

### 2.1.3 固有表現抽出の問題

先行研究では，固有表現タグの付与は，同一文書にある他の固有表現に対するタグの付与とは独立に行なわれる場合が多いため，以下の２つの問題が見受けられた．

1. 同一の対象を表す固有表現が抽出されない

   例えば，同一文章中に「山岸章」と「山岸」という固有名詞があり，両者は同一の対象を表しているとする．しかし，従来の固有表現抽出技術では，独立に固有表現タグを付与するため，前者に “人名”，後者に “組織名” といったように，異なる固有表現タグを付与する可能性がある．両者は同一の対象を表しているので，これらには同じ固有表現タグを付与するべきである．

2. 同一の対象に対して同じタグが付与されない

   例えば，同一文章中に「公正取引委員会」と「公取委」という２つの固有名詞があるとする．このとき，前者には “組織名” というタグを付与するが，後者には固有表現タグを付与しない，つまり固有名詞として抽出されない場合がある．両者は同一の対象を表しているので，これらは共に固有表現として抽出し，同じ固有表現タグを付与するべきである．

これらの問題点が生じる理由として，タグの整合性を取らなかったことが上げられる．タグの整合性とは，ここでは同一の対象が表す固有名詞に同一の固有表現タグを付与する

ことを指す．そこでタグの整合性を取るため，本研究では固有表現の照応関係を解析することを考える．

2.2 節では，照応関係の解析についての関連研究を述べる．

## 2.2 照応解析

照応解析の手法は様々な方法があるが，自然言語処理における一般的な照応解析手法であるセンタリング理論について 2.2.1 項で述べる．また，本研究では,「同社」「同日」など,「同」を含む表現について，新しい照応解析アルゴリズムを提案する．2.2.2 項では「同社」を含む表現を対象とした過去の照応解析手法について述べる．

### 2.2.1 センタリング理論

センタリング理論 [1] は英語の代名詞の照応関係を決定する手法として提案された．センタリング理論では，文中で話題の中心になっているものをセンターと呼び，談話中でセンターが連続している場合，すなわち話題が連続している場合には代名詞が使われているはずである，という基本規則を利用して照応の解析を行なっている．

また亀山 [2] は，英語の代名詞を日本語の省略に置き換えることでセンタリング理論を日本語の省略解析に適用した．更に，センタリング理論だけでは説明できない省略に対応するために，属性共有制約 (property sharing constraint) を導入した．属性共有制約とは，隣接する文でセンターが継続する場合，その 2 文中のゼロ代名詞は文法属性を共有すべきである，というものである．

Walker ら [13][12] は，日本語の省略解析に Transition を適用し，Transition を使うことで属性共有制約と同等以上の解析ができることを示した．Transition とは，省略の補完に複数の解釈が可能な場合，なるべくセンターが変わらない解釈を優先するという手法のことである．

また，田村 [5]，高田ら [10] はセンタリング理論を応用して，文間・文内照応解析を行っている．

#### センターの定義

談話単位中の各発話 $U$ には，**前向き中心** $(foward\text{-}looking\text{-}center)Cf(U)$ と**後向き中心** $(backward\text{-}looking\text{-}center)Cb(U)$ が結び付いている．$Cf$ は発話 $U$ において実現されてい

る対象のリストで，$Cf$ のうち，現在の話の中心になっている特別な要素が $Cb$ である．$Cf$ の要素は日本語の場合，次のランキングで順序付けられる．

$$(\text{文法・ゼロ})\ \text{主題} > \text{視点} > \text{ガ格} > \text{ニ格} > \text{ヲ格} > \text{その他} \tag{2.1}$$

「主題」は固有名詞が主題化されているとき，ガ格，ニ格，ヲ格はそれぞれの表層格の格要素になっているときを表す.「視点」は授与動詞「(〜して) やる」のガ格や「(〜して) くれる」のニ格など，話し手の共感がおかれる対象を指す．また，$Cf$ の中で最も序列の高い要素を**優先中心** $(preferred\text{-}center)Cp$ と呼ぶ．

**センターの制約と規則**

1. 発話列 $U_1, ..., U_m$ からなる談話単位中の各発話 $U_i$ について，以下の制約が成り立つ．

   (a) ただ１つの後ろ向き中心 $Cb(U_i)$ が存在する．

   (b) 前向き中心のリスト $Cf(U_i)$ のあらゆる要素は，$U_i$ で実現されている．

   (c) $Cb(U_i)$ は，$Cf(U_{i-1})$ の要素のうち $U_i$ で実現されているものの中で，$Cf(U_{i-1})$ での序列が最も高かったものである．

2. 発話列 $U_1, ..., U_m$ からなる談話単位中の各発話 $U_i$ について，以下の規則が適用される．

   (a) $Cf(U_{i-1})$ のある要素が $U_i$ で代名詞として実現されるなら，$Cb(U_i)$ もまた代名詞として実現される．

   (b) $Cb$ の遷移には以下の優先順序がある．

$$\text{CONTINUE} > \text{RETAIN} > \text{SMOOTH-SHIFT} > \text{ROUGH-SHIFT} \tag{2.2}$$

CONTINUE とは先行文脈から引き継いだ $Cb$ がその後も続くと予測される場合で，RETAIN はそれまで続いてきた $Cb$ が次には移動すると予測される場合である．また，SMOOTH-SHIFT と ROUGH-SHIFT は先行文脈から $Cb$ が移動した場合である．センターの遷移パターンを表 2.1に示す．

**センタリング理論の適用例**

本節ではセンタリング理論の適用例を以下に示す．(出典：[13])

表 2.1: センターの遷移パターン

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

a 太郎$_i$ は 花子$_j$ を映画に誘いました.

b i $\phi_{i\ ガ}$ 一日中何も手につきませんでした.

ii $\phi_{j\ ガ}$ 一日中何も手につきませんでした.

表 2.2: センターの遷移

| | $Cb$ | $Cf$ | 遷移パターン |
|---|---|---|---|
| a | 太郎 | [太郎<主格>，映画<二格>，花子<ヲ格>] | |
| b.i | 太郎 | [太郎<ガ格>] | CONTINUE |
| b.ii | 花子 | [花子<ガ格>] | SMOOTH-SHIFT |

表 2.2の b.i のようにゼロ代名詞の先行詞を「太郎」と仮定すると，$Cb$ の遷移は CONTINUE に，b.ii のようにゼロ代名詞の先行詞を「花子」と仮定すると，$Cb$ の遷移は SMOOTH-SHIFT になる．式 (2.2) により，CONTINUE が SMOOTH-SHIFT より優先されることから先行詞は優先的に「太郎」と解釈される.

### 田村らによるセンタリング理論の拡張

田村ら [5] は従来のセンタリング理論では以下の問題があったと指摘している.

- 複文のように，同じ文法属性の要素が複数存在する場合 (例えば，主語が２つ) の順序は決められていないため，解析できない.
- 省略すべき箇所が複数ある場合，先行詞候補の優先度が決められていない.
- 同一文内の照応に対応していない.

- 目的語の補完に失敗する例が多い.

そこで，以下を行なうことで解決した.

- 複文を単文に分割

  複文を単文に分割することで，センタリングを用いることができる.

- 接続助詞を省略の補完に利用

  接続助詞を 3 種類に分類し，それに対応した戦略を提案している.

- 候補の優先度とスコア付け

  要素を全順序で決めることは難しく，先行詞候補に対して「どちらかどれだけ良いか」はわからない．そのため，先行詞の候補をスコアリングしている.

上記の方法を用いることで，複文を含む談話の省略を扱うことができるといっている.

### 2.2.2 同社を対象とした照応解析

同社を対象とした照応解析の先行研究として，若尾 [11]，木谷 [3] がある．若尾は「同社」を対象とした照応解析を行なった．この論文では，人手によって同社とそれらの位置を決定し，もっとも距離が近い社名をとる単純法と若尾が提案する 2 つのヒューリスティックを評価している．以下にその手法について説明する.

1. 単純法 (Simple Closest Method(SCM))

   もっとも距離が近い社名をとる.

2. 「同社＋が」のために SCM を改良したヒューリスティック

   - もし同社の前に同じ文の中に 1 つ以上の社名があるならば，照応としてもっとも距離が近い社名をとる.
   - もし同社の前のどこかに「は」,「が」,「では」,「によると」のすぐ近くに社名があるならば，照応としてそのもっとも距離が近い社名をとる.
   - もし前の文が社名で終っている場合のように社名が強調されているならば，それを照応とする.

- もし前の文で「社名＋の＋人間の名前」のパターンであれば，そのとき，照応として社名を使う.「人間の名前」として典型的に現われるのは，社長や会長である．

3. 「同社＋は」のためにSCMを改良したヒューリスティック

- もし社の前のどこかに「は」,「が」,「では」,「によると」のすぐ近くに社名があるならば，照応としてもっとも距離が近い社名をとる．
- もし前の文が社名で終っている場合のように社名が強調されているならば，それを照応とする．
- もし前の文で「社名＋の＋人間の名前」のパターンであれば，そのとき，照応として社名を使う．

これらの方法で選んだ先行詞が正確であるかをテストしたところ,「同社＋が」では，SCMが67％（27／42）の正解率，2つ目の方法が90％（38／42）の正解率であり,「同社＋は」では，SCMが52％（34／66）の正解，3つ目の方法が94％（62／66）の正解であった．この結果，2つのヒューリスティックは，照応の発見に良い性能を示したと言っている．

また，木谷は同社に加えて省略表現（例：松下（松下電器産業の略））や言い換え表現（例：日本移動通信（IDO）),「両社」「自社」を対象としている．これらの表現を解析する際に，文のトピックとLCS（longest common subsequence），さらにヒューリスティックを用いることで照応解析を行なっている．評価の対象はベンチャー企業やマイクロエレクトロニクスの合弁企業について書かれた新聞記事である．以下にそのヒューリスティックを示す.

- 「両社」は合弁企業の記事で説明されている2つのタイアップしている会社と関係している．
- 同じ文の中で「同社」の前に２つ以上の会社があるとき，照応はその記事のトピックの会社よりもむしろ一番近い会社を指す傾向がある．
- 「同社」はトピックの会社と関係することが少ないが,「自社」はしばしばトピックの会社と関係する．
- 「自社」は１つ以上の会社を指すことができる．

- 「自社」は時々「会社」か「a company」のような会社についての一般的な表現を指す.

実験の結果，318 の省略のうち，再現率と精度はおのおの 67%と 89%であった．また，単純な文字列のマッチングはよく似た社名（系列会社など）を誤まる可能性があると指摘している.

これらの論文では，解析対象の表現が「同社，両社，自社」と省略に限られている．しかし，他の表現も照応解析の対象として考えるべきである.

## 2.3 本研究との違い

本研究では，2.1.3項で述べたように，固有表現抽出の 2 つの問題

- ある固有表現と同一の対象を表す固有表現が抽出されない場合
- 同一の対象に対して同じタグが付与されない場合

を解決するために，照応解析を行なう．照応解析により，同一の対象を表している固有表現を特定し，同一の対象に対して同じタグを付与するように，タグの整合性を取る．また，固有名詞を指す普通名詞については，新たに固有表現として抽出し，参照する固有名詞に付けられたタグをそのまま付与する.

また，2.2.2項で述べた研究では,「同社，両社，自社，省略」など，限られた表現を対象としていた．しかし,「同県」「同日」「同氏」など,「同」を使った表現は多様に存在する．そこで本研究ではこれらの表現も取り扱えるような方法を考えて，より幅の広い表現に対応した照応解析の実現を目指す.

# 第 3 章

# 固有表現抽出システムの実装

## 3.1 処理の流れ

システムの処理の流れを図 3.1に示す.

最初に固有表現抽出を行ない，初期の固有表現タグを付与する．次に，照応解析を行なう．まず，照応解析の解析対象となる表現を参照表現と呼び，これを抽出する．次に，参照表現の先行詞を決定する．最後に，照応解析の結果をもとに，同一の対象を指す表現には同じ固有表現タグを付与するようにタグを統一する.

図 3.1における各モジュールの説明を次節以降に述べる.

## 3.2 固有表現抽出モジュール

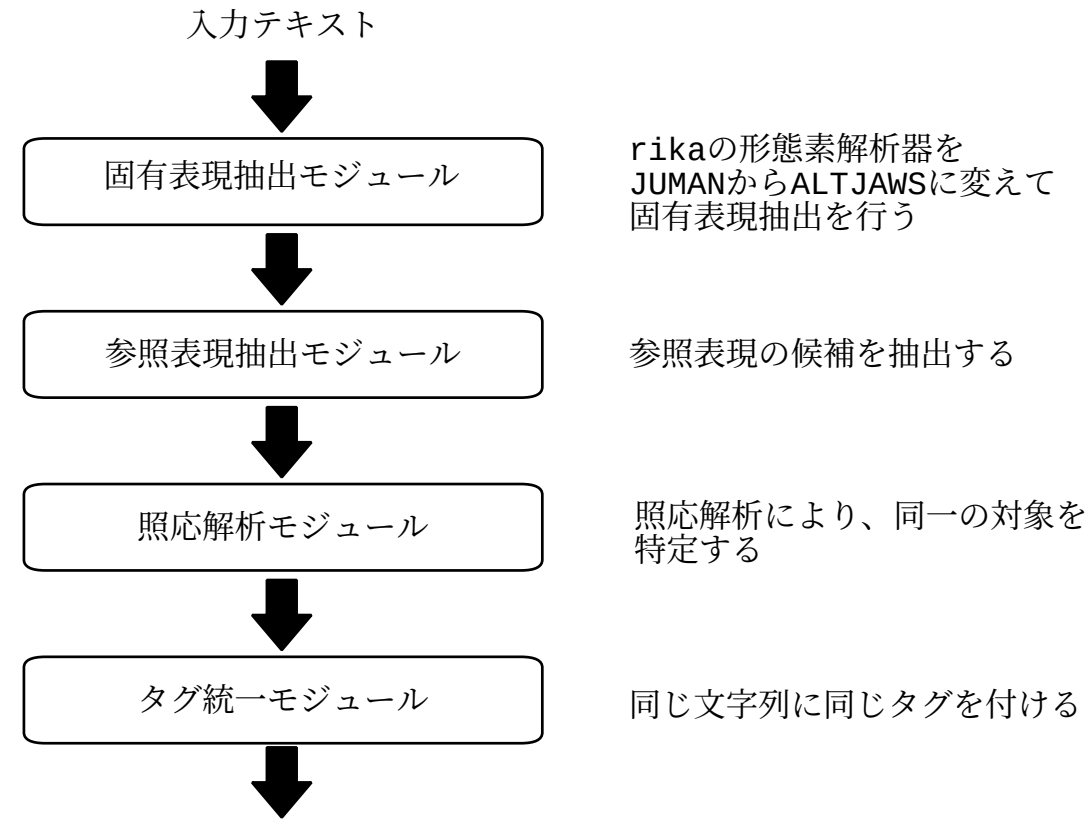
このモジュールでは初期の固有表現タグを付与する.

### 3.2.1 rika

2.1節で述べたように，rika は関根らによって提案された固有表現抽出システムである．本研究は，この rika を用いて固有表現抽出を行なう．ここでは，その使用方法を簡単に説明する.

始めにトレーニングデータを入力し，決定木学習器 C4.5 に渡すデータを作成する．決定木学習器はプログラムパッケージである C4.5[8] を用いている．C4.5 に渡すデータは JUMAN が出力する品詞，文字型 (漢字，ひらがな，アルファベット，数，記号，それらの組み合わせ)，辞書，固有表現の始め (opening)，続き (continuation)，終り (closing) を

図 3.1: システム図




表すタグなどのデータである[1]．次にそのデータを C4.5 に渡し，決定木を自動的に学習する．最後に，その決定木を用いて，タグ付けされていないテキストデータに固有表現タグを付与する．

### 3.2.2 rika の改良

rika で用いられている形態素解析器を本研究では ALTJAWS[14] に置き換える．その理由は以下の通りである．後述する照応解析モジュールでは，単語の意味素の情報を必要とする．また，意味素として日本語語彙大系 [7] の意味属性を必要とする．ALTJAWS は形態素解析結果と同時に日本語語彙大系の意味属性番号を出力することができるので形態素解析器を JUMAN から ALTJAWS に変更する．

[1]opening，continuation，closing の例は文献 [9] の付録に記載されている．

## 3.3　参照表現抽出モジュール

このモジュールでは参照表現を抽出する．

### 3.3.1　参照表現の定義

まず，文書から照応解析の対象となる名詞を抽出する．本研究では「固有名詞」または「固有名詞を参照している表現」を参照表現と呼び，照応解析の対象とする．参照表現としては，以下の 3 種類を考える．

1. 省略表現

   例：「松下電器産業」を単に「松下」と表す

2. 固有表現を指す普通名詞

   例：「東京大学」を単に「大学」と表す

3. 「同」を用いた表現

   例：「同社」「同県」

   本研究では，他に「両社」や「自社」等も取り扱う．これらも「同」を用いた表現と同じように取り扱う．

### 3.3.2　参照表現の抽出方法

本研究では，初期の固有表現抽出で抽出されなかった固有名詞や固有名詞を参照する普通名詞に対しても固有表現タグを付与するべきであると考える．そこで上記の 2，3 のような固有名詞以外の普通名詞も参照表現として抽出し，照応解析を行なう．

参照表現の抽出は，初期の固有表現抽出の際に得られる ALTJAWS[14] の形態素解析結果を元に行なう．まず，ALTJAWS によって固有名詞を表す品詞が与えられた形態素を参照表現として抽出する．また，普通名詞を表す品詞が与えられた形態素についても，ALTJAWS が出力する意味属性が組織名や人名などのように固有名詞に近い場合には，参照表現として抽出する．さらに,「同」を用いた表現については「同」「両」「自」といった文字を手がかりにして抽出する．

以上のように，形態素結果をもとに抽出される参照表現の他に，初期の固有表現抽出によって抽出された固有表現も参照表現として加える．

ALTJAWS による形態素区切りは固有名詞を表す単位としては細かすぎる場合がある．そこで以下の場合には形態素を統合し，一つの参照表現として取り扱う．

1. 似ている意味属性を持つ名詞の連続

   例) 東京都＋杉並区，村山＋富市

2. 名詞 (句) ＋接辞

   例) 関西＋国際空港，社会＋党

3. 括弧で括られた名詞 (句)

   例) 「村山政権を支え社民リベラル政治をすすめる会」

### 3.3.3 参照表現抽出モジュールの実装

参照表現抽出モジュールの実装について説明する．

1. 固有表現抽出モジュールの形態素解析結果と出力結果から，形態素情報と意味素情報を得る．

   形態素情報と意味素情報は ALTJAWS[14] の出力である．

2. 得られた情報から参照表現の候補を抜き出す．また，参照表現をいくつかの種類に分類する．

   形態素情報と意味素情報を元に参照表現の種類 (組織名，人名等) を振り分ける．振り分ける種類は，以下の９つである．

   - ORGANIZATION(組織名)
   - PERSON(人名)
   - LOCATION(地名)
   - ARTIFACT(固有物名)
   - DATE(日付表現)
   - TIME(時間表現)
   - MONEY(金額表現)
   - PERCENT(割合表現)

- OPTIONAL

  OPTIONAL とは，どの種類に属するかは現段階ではわからないが，参照表現の可能性のある表現である．

これら参照表現の種類は固有表現のタグの種類とほぼ等しい．しかし，ここでの参照表現の分類では，ALTJAWS の形態素解析結果のみを用いて簡潔に行ない，初期の固有表現抽出とは全く独立に行なわれる．

以下に具体的な方法を説明する．

(a) はじめに，ALTJAWS が出力する品詞が「名詞」又は「接辞」であるなら形態素を取り出す．また，括弧で括られた名詞句は，それ全体で１つの参照表現として抽出する．

(b) 次に，(a) で抽出されたもののうち，形態素が連続するものに対して，以下のパターンを適用する．

  以下のパターンで使用されている記号を説明する．“ (A，B，C) ” は，パターンマッチするべき形態素の，品詞，意味素情報，表記を表す．ただし，A，B，C が “＊” のときは，その条件についてはいかなる形態素についてもマッチすることを表す．例えば，(接辞，＊,「同」「両」) は,「同」または「両」という接辞にマッチすることを表す．

  i. (接辞，＊,「同」「両」) ＋ (名詞，組織名，＊) → ORGANIZATION
  ii. (接辞，＊,「同」「両」) ＋ (名詞，人名，＊) → PERSON
  iii. (接辞，＊,「同」「両」) ＋ (名詞，地名，＊) → LOCATION
  iv. (接辞，＊,「同」「両」) ＋ (名詞，日付表現，＊) → DATE
  v. (接辞，＊,「同」「両」) ＋ (名詞，時間表現，＊) → TIME
  vi. (固有名詞：人名，＊，＊) ＋ (固有名詞：人名，＊，＊) → PERSON
  vii. (未定義名詞，＊，＊) ＋ (固有名詞：人名，＊，＊) → PERSON
  viii. (固有名詞：地名，＊，＊) ＋ (接辞，地名，＊) → LOCATION
  ix. (固有名詞：地名，＊，＊) ＋ (固有名詞：地名，＊，＊) → LOCATION
  x. (固有名詞：地名，＊，＊) ＋ (接辞，組織名，＊) → ORGANIZATION
  xi. (数詞，＊，＊) ＋ (接辞，金額表現，＊) → MONEY
  xii. (数詞，＊，＊) ＋ (接辞，日付表現，＊) → DATE
  xiii. (数詞，＊，＊) ＋ (接辞，時間表現，＊) → TIME

xiv. (時詞，＊，＊) + (接辞，時間表現，＊) → TIME

xv. (記号，＊,「(」「「」等) +（＊，＊，＊）+ ··· + (名詞，＊，＊) + (記号，＊,「)」「」」等) → OPTIONAL

(c) 上記のパターンにマッチしなかった形態素は，単一の形態素で構成される参照表現の可能性があると考え，以下のパターンを適用して，固有表現の種類を特定する．

i. (名詞，組織名，同) → ORGANIZATION

ii. (名詞，人名，同) → PRESON

iii. (名詞，地名，同) → LOCATION

iv. (名詞，日付表現，同) → DATE

v. (名詞，組織名，＊) → ORGANIZATION

vi. (名詞，人名，＊) → PRESON

vii. (名詞，地名，＊) → LOCATION

viii. (名詞，固有物名，＊) → ARTIFACT

ix. (名詞，日付表現，＊) → DATE

x. (名詞，時間表現，＊) → TIME

xi. (名詞，金額表現，＊) → MONEY

xii. (名詞，割合表現，＊) → PERCENT

xiii. (未定義名詞，＊，＊) → OPTIONAL

(d) 上記までのパターンがマッチしなかった形態素で，初期の固有表現抽出で抽出されたものは，参照表現として抽出する．また，その種類は，初期の固有表現抽出タグによって決める．

(e) 上記までのパターンがマッチしなかった形態素は，たとえその品詞が「名詞」であっても参照表現とて抽出しない．

### 3.3.4 参照表現抽出モジュールの問題点

本モジュールには，問題点がある．ALTJAWS は一般名詞意味属性番号を最大５つ出力するが，どれの番号が尤らしいかはわからない．そのため，固有表現の種類を分けるときに，最初にマッチした一般名詞意味属性番号で固有表現の種類を特定している．よって，固有表現に曖昧性があるとき (例えば，組織名か地名か) は曖昧性のあるまま処理せずに一意に決めている．

## 3.4 照応解析モジュール

このモジュールでは照応解析を行なう．すなわち，3.3 節で述べた方法で抽出したすべての参照表現について，その先行詞を求める．詳しくは，4章で述べる．

## 3.5 タグ統一モジュール

タグの整合性を取るために，照応解析の結果から同一と判定された表現に対して，同一の固有表現タグが付与されるように初期の結果を修正する．

1. 照応解析の結果，同一と判定された表現のどれか 1 つが固有名詞であるか，rika で出力された固有表現かのどちらかであれば，その表現に同じタグを付与する．そうでなければ，同じタグを付与しない．なぜなら，参照表現には多くの普通名詞が混在しており，同一と判定された参照表現の中にもすべてが普通名詞である可能性があるからである．

2. また，同一の対象に対して，タグの種類が複数存在するとき，rika が出力する固有表現タグの信頼度を調べる．

   (a) どれかの信頼度が高ければ，それを正解タグとしてタグを統一する．

   (b) 信頼度が同じであれば，タグの多数決を取り，多い方でタグを統一する．

3. 同一記事内で照応解析によって参照表現として抽出された表現と全く同じ表現が同一記事内の他の場所で現われても，形態素解析結果の相違などによって，参照表現として抽出されないときがある．このような表現は照応解析でも同一の対象を指しているとは判定されない．そのため，本モジュールでは，同一記事内の形態素で，表現が全く同じであるものすべてに同一の固有表現タグを付与する操作を行なう．

# 第 4 章

# 照応解析

本章では照応解析手法について述べる．

はじめに照応解析の素性に用いる素性について述べる．照応解析に用いる素性は，先行研究で提案された素性と，本稿で提案する新たな素性の両方を用いる．次に，それらの素性を用いた照応解析手法について述べる．

## 4.1 照応解析の素性

若尾 [11] や木谷 [3] のヒューリスティックを観察すると，(1) 先行詞と照応詞の距離，(2) 先行詞の格情報やトピック，という 2 つの素性を元にしてヒューリスティックを作成していることが分かる．つまり，それら 2 つの素性は照応解析の手がかりとなると考えられる．また,「同社」「両社」「自社」といった特定のパターンに特化したヒューリスティックを数多く導入している．しかし，これらのヒューリスティックは，他の「同」を用いた表現に用いることは難しく，移植性に欠ける．より一般性を持たせるためには，先程の (1) 先行詞と照応詞の距離，(2) 先行詞の格情報やトピックやその他の素性を用いて，特定のパターンに依存しない方法を構築することが望ましい．

そこで本研究では，照応解析の手がかりとなる素性として，以下の 3 つを考える．これらの素性は固有名詞の先行詞のなりやすさの指標として用いる．

- **センタリング理論に基づく文法属性**

  センタリング理論 [1][13][6] では，名詞の格などの文法的な属性に着目し，式 (2.1) のような順序で名詞が先行詞になりやすいとしている．式 (2.1) を式 (4.1) として再掲する．

  主題 > 視点 > ガ格 > ニ格 > ヲ格 > その他 (4.1)

「主題」は固有名詞が主題化されているとき，ガ格，ニ格，ヲ格はそれぞれの表層格の格要素になっているときを表す．本研究では,「視点」を除き，式 (4.2) の順序で固有名詞が先行詞になりやすいとする．

$$\text{主題} > \text{ガ格} > \text{ニ格} > \text{ヲ格} > \text{その他} \tag{4.2}$$

また,「AのB」のように，AもBも共に固有表現であった場合，どちらが先行詞になりやすいかを調査した．調査対象は CRL 固有表現データの一部である．CRL(郵政省通信総合研究所) 固有表現データは，毎日新聞 95 年 1 月 1 日から 10 日までの全記事，約 1 万文に対して，固有表現をタグ付けしたデータである．表 4.1 に調査結果を示す．

表 4.1: A の B の調査結果

| | | |
|---|---|---|
| | 「A の B」の総数 | 1225 |
| (1) | A のみが固有表現 | 105 |
| (2) | B のみが固有表現 | 18 |
| (3) | どちらも固有表現 | 27 |
| (4) | 「A の B」の全体が固有表現 | 8 |
| (5) | (1) で A が先行詞 | 53 |
| (6) | (2) で B が先行詞 | 7 |
| (7) | (3) で A が先行詞 | 2 |
| (8) | (3) で B が先行詞 | 5 |
| (9) | (3) でどちらも先行詞 | 7 |
| (10) | (4) 自体が先行詞 | 1 |
| (11) | (5)+(7)+(9) | 62 |
| (12) | (6)+(8)+(9) | 19 |
| | (1)+(2)+(3)+(4) | 158 |

表 4.1 の結果，事例は少ないが,「A の B」の A，B ともに固有表現であるとき (27 事例)，A が 2 回，B が 5 回先行詞になっていることから，$A < B$ とすることがよい ($(7) : (8) = 2 : 5$) と思われる．そこで,「A の B」のとき，B の方を優先する．

- **距離**

  距離とは，固有名詞とその固有名詞の先行詞との間に存在する単語数であると定義する．ここでは距離が小さいほど，先行詞になりやすいとする．

  若尾 [11] の実験では「同社＋が」に対して，SCM（最も近くに存在する社名を先行詞とする方法）の精度が 67 %（27 ／ 42）で,「同社＋は」に対して，SCM の精度が 52 %（34 ／ 66）であったことが報告されている．この結果は「同社」のみの結果なので，単純に「同社」以外の「同」を用いた表現に適応できるかは不明であるが，有効な素性であると考えられる．

- **言及クラス**

  本研究で新しく提案する素性である．この素性の定義と，素性を提案した理由については，4.2 節で説明する．

## 4.2 言及クラス

### 4.2.1 照応解析に有効な素性の調査

本研究では，照応解析に有効な素性を調べるために，CRL 固有表現データの一部 (６５記事) に対して，図 4.1のような照応タグを付与したデータを作成した．

作成したデータに付与した<COREF>タグと</COREF>タグは，同一記事内に複数回言及された固有表現のみに付与されている．また，<COREF>タグの中の属性は以下の通りである．

- **ID**

  ID 属性はそのエレメントのユニークな識別子である．

- **REF**

  REF 属性は，その表現の先行詞の ID 番号である．先行詞が複数存在する場合は，すべて最初に出現した先行詞の番号が付与されている．

照応タグを付与した新聞記事を調べたところ，ある同一の対象が異なる表現で出現した場合，それ以降もそれまでとは異なる別の表記で出現する傾向がみられた．始めに例を図 4.2に示す．

図 4.2の 1 行目から 2 行目にある

<COREF ID="2">ポレワノフ</COREF>

図 4.1: 照応タグを付与したデータの例

```
<DOC>
<DOCID>950101044</DOCID>
<TEXT>
サッカー界のスーパースター、<COREF ID="1">ディエゴ・マラドーナ</COREF>
選手とその家族が<COREF ID="2">昨年１２月３０日</COREF>、<COREF ID="3">
キューバ</COREF>の<COREF ID="4">カストロ</COREF><COREF ID="5">国家評議
会</COREF>議長と会い、ニッコリ記念撮影。<COREF ID="6" REF="4">カストロ
</COREF>議長と<COREF ID="7" REF="1">マラドーナ</COREF>選手は旧知の仲。
<COREF ID="8" REF="3">同国</COREF>で<COREF ID="9">２日</COREF>まで家族
水入らずの休暇を楽しむ予定。
</TEXT>
</DOC>
```

と同一の対象が 11 行目に

```
<COREF ID="10" REF="2">同副首相</COREF>
```

として現われ，異なる表現で言及されている．さらに 17 行目になると，同一の対象が

```
<COREF ID="15" REF="2">同氏</COREF>
```

として現われ，先ほどとは別の異なる表現で言及されている．

これとは逆に，同一の対象が同じ表現で現われるとどうなるだろうか．そこで先ほどの図 4.2の

```
<COREF ID="4">チュバイス</COREF>
```

に注目する．“チュバイス” は図 4.2を通してすべて同じ表現で現われている．何度も同じ表現で出現した表現は，その後も同じ表現で言及していることが分かる．“同氏” の先行詞を特定するときに，“ポレワノフ” か “チュバイス” かを考えた場合，この言及の方法によって “同氏” の先行詞が特定出来そうである．

したがって,「同」という表現が出現したとき，それ以前に様々な表記で出現した対象を指しやすいと考える．また，逆に一定の表記で出現した対象は指しにくいと考える．そこで，この言及の方法という新たな素性を加えることにし，その言及の方法に関するクラス（**言及クラス**）を定義する．

図 4.2: 同一の対象が異なる表現で出現する例

<COREF ID="1">ロシア</COREF>の民営化政策を担当する<COREF ID="2">ポレワノフ</COREF>副首相兼<COREF ID="3">国家資産管理委員会</COREF>議長が「非民営化、再国営化」の基本姿勢を打ち出した。前任の<COREF ID="4">チュバイス</COREF>氏の急進的な民営化政策を大幅に修正するものであり、<COREF ID="5">エリツィン</COREF>政権の<COREF ID="6">一九九五年</COREF>の経済運営を占う意味でも注目される。<COREF ID="7">三十日</COREF>の<COREF ID="8">セボードニャ紙</COREF>によると、<COREF ID="9" REF="2">ポレワノフ</COREF>副首相は「政府の経済路線を変更し、企業に対する国家の指導を強化することが必要」と強調するとともに「これまでに誤って民営化された企業を再国営化させる法案の採択を目指す」と語った。

国営に戻すべき分野として<COREF ID="10" REF="2">同副首相</COREF>はアルミニウム、エネルギー、軍需産業を挙げ、「外国企業が<COREF ID="11">一五%</COREF>の株式を取得し役員会への代表派遣を可能にしたことは、<COREF ID="12" REF="1">ロシア</COREF>国家の安全保障に直接脅威を与える」と述べた。

さらに<COREF ID="13">イズベスチヤ紙</COREF>も「民営化の足下に爆弾」という見出しで「<COREF ID="14" REF="2">ポレワノフ</COREF>副首相の基本姿勢は非民営化にある」と伝えた。その中で<COREF ID="15" REF="2">同氏</COREF>は「これまでの民営化は一方向だけへの動きだった。昼は夜を持ち、生には死があるように民営化にも国営化がある」と述べ、民営化が行き過ぎたとの認識を明らかにした。<COREF ID="16" REf="1">ロシア</COREF>の民営化政策は<COREF ID="17" REF="4">チュバイス</COREF>氏の指導で<COREF ID="18">九二年十月</COREF>から始まった。

民営化証券を使って株式を取得するという第一段階は<COREF ID="19">九四年六月</COREF>に終了し、現在は現金で株を購入できる第二段階に入っている。

<COREF ID="20">九四年十一月</COREF>段階で中小企業の<COREF ID="21">七五%</COREF>が民営化されたほか、大企業の民営化もかなり進んでいる。

またこれまでに自治体所有の<COREF ID="22">七一%</COREF>の資産が民間に売却された。

### 4.2.2 言及クラスの定義

言及クラスとは，ある同一の対象が次にどう表現されていたかを表すクラスであり，同一の対象ごとに与えられる．言及クラスには以下の３つのクラスがある．

- **相違言及クラス**

  ある対象と同一の対象が異なる表現で言及されたもの．同一の対象が記事中に異なる表現で出現した回数が２回以上の時に “相違言及クラス” になる．例えば「九条署」が後に「同署」となった場合，相違言及となる．

- **未言及クラス**

  ある対象が文章内で始めて出現し，それ以前にはまだ言及されていないもの．

- **同一言及クラス**

  ある対象と同一の対象が同じ表現で言及されたもの．同一の対象が記事中に同じ表現で出現した回数が３回以上の時に “同一言及クラス” になる．例えば「九条署」が後も「九条署」となり，それが２回続く場合，同一言及となる．

本研究では，相違言及，未言及，同一言及の順に固有名詞が先行詞になりやすいとする．また，先行詞の候補となる複数の固有名詞が同じ言及クラスに属する場合，言及回数の多い固有名詞ほど先行詞になりやすいとする．言及回数とは，同一の対象を指す固有名詞が記事中に出現した回数である．

## 4.3 提案する照応解析手法

センタリング理論の文法属性と，先行詞と照応詞の距離，さらに言及クラスを用いた照応解析手法を提案する．

### 4.3.1 アルゴリズム

提案手法の詳細なアルゴリズムを示す．まず，3.3節で述べた手法で抽出された参照表現の集合を $r_1, \cdots, r_n$ とする．ただし，$r_i$ は文書中の出現順序にしたがって並べられているとする．また，先行詞候補のリスト $AL$ を用意する．$AL$ は，先行詞の候補となりうる既出の参照表現を，先行詞のなりやすさの順序で並べたリストである．$AL$ には，以下の９種類のリストがある．これらのリストは，3.3.3 項で述べた参照表現のそれぞれの種

類に対応している．参照表現は，その種類に応じて，これら 9 つのリストの 1 つの中から先行詞が選ばれる．ただし，OPTIONAL については，他の８つのリストで先行詞が選ばれなかったときに，先行詞の探索が行なわれるリストである.「同」を用いた表現の先行詞が OPTIONAL の要素であるとき，その要素は「同」を用いた表現が属するリストに挿入される．例えば,「同社」の先行詞が ORGANIZATION リストの要素でないとき, OPTIONAL リストの要素を探索する．もし,「同社」の先行詞が OPTIONAL リストの要素の１つである「デル」であったとき,「デル」は ORGANIZATION リストに挿入され, OPTIONAL リストから削除される.

- ORGANIZATION リスト
- PERSON リスト
- LOCATION リスト
- ARTIFACT リスト
- DATE リスト
- TIME リスト
- MONEY リスト
- PERCENT リスト
- OPTIONAL リスト

  どの種類に属するかは現段階ではわからないが，参照表現の可能性のあるリスト．このリストには，未知語や括弧で括られた名詞句などが挿入される.

参照表現 $r_i$ の先行詞を決定する手続きは以下の通りである.

1. $AL$ を空にする．$i = 1$ とする.
2. $AL = \{a_1, \cdots, a_m\}$ であるとする．要素 $a_j$ について，リストの先頭から順に，$r_i$ の先行詞が $a_j$ となるかどうかを調べる．$a_j$ が $r_i$ の先行詞となる条件は以下の通りである.

- $r_i$ が「同」を用いた表現以外のとき

  $a_j$ と $r_i$ の $LCS$ が 2 以上のとき，$r_i$ の先行詞は $a_j$ であるとみなす．$LCS$ とは，2 つの文字列のうち，同じ順序で現れる同じ文字の数のことである．例えば,「松下電器」と「松下」の $LCS$ は 2 である．

- $r_i$ が「同」を用いた表現のとき

  「同」を用いた表現のときは,「同社」などの表現は長さが短く，表記から得られる情報が少ないため，$LCS$ によって先行詞を特定することは難しい．そこで，$a_j$ の ALTJAWS による品詞タグが「固有名詞」であるなら，$r_i$ の先行詞は $a_j$ であるとみなす．

3. $AL$ の中から $r_i$ の先行詞が見つからなければ，$r_i$ は文書中の最初に出現した固有名詞であるとみなす．また，$r_i$ は後続する固有名詞の先行詞となる可能性があるので，これを $AL$ に追加する．

4. $AL$ の要素の順序を並び替える．順序は 4.1 節で述べた素性をもとに決定する．5 章で述べる評価実験では，以下の 3 つの基準で先行詞の候補の並び替えを行った．

   **基準 1** 言及クラス+言及回数 > 文法属性 > 距離

   **基準 2** 文法属性 > 言及クラス+言及回数 > 距離

   **基準 3** 距離

   “基準 1” では，まず「言及クラス」「言及回数」によって $AL$ の要素の順序を決める．これが同じときには「文法属性」によって $AL$ の要素の順序を決め,「文法属性」も同じときには「距離」によって決定する．“基準 2” と “基準 3” も同様である．

5. $i = n$ なら終了．それ以外は 2. へ戻る．

### 4.3.2 解析例

提案アルゴリズムによる照応解析の例を図 4.3に示す．この例では主にPERSON の先行詞候補のリストに注目する．ID 番号 23 の “同氏” の前のPERSON の先行詞の候補は以下の 5 つである．

<COREF ID="1">ベルルスコーニ</COREF>
<COREF ID="3">スカルファロ</COREF>

<COREF ID="15">スコニャミリオ</COREF>

<COREF ID="16">ピベッティ</COREF>

<COREF ID="22">ディニ</COREF>

参照表現の先行詞を順に決めていく過程と，先行詞候補のリストの変化を表 4.2に示す.

表 4.2: リストの並び替え

| ID | 参照表現 | REF | リスト |
|---|---|---|---|
| 1 | ベルルスコーニ | - | ベルルスコーニ (以下，ベル) |
| 3 | スカルファロ | - | スカルファロ (以下，スカ), ベル |
| 6 | ベルルスコーニ | 1 | ベル, スカ |
| 9 | スカルファロ | 3 | スカ, ベル |
| 10 | ベルルスコーニ | 1 | ベル, スカ |
| 11 | スカルファロ | 3 | スカ, ベル |
| 15 | スコニャミリオ | - | スコニャミリオ (以下，スコ), スカ, ベル |
| 16 | ピベッティ | - | ピベッティ, スコ, スカ, ベル |
| 17 | スカルファロ | 3 | ピベッティ, スコ, スカ, ベル |
| 18 | ディニ | - | ディニ, ピベッティ, スコ, スカ, ベル |
| 19 | ディニ | 18 | ディニ, ピベッティ, スコ, スカ, ベル |
| 20 | ベルルスコーニ | 1 | ディニ, ピベッティ, スコ, ベル, スカ |
| 22 | ベルルスコーニ | 1 | ディニ, ピベッティ, スコ, ベル, スカ |
| 23 | 同氏 | 18 | (ディニ), ピベッティ, スコ, ベル, スカ |

表 4.2の最終行に示したリストから，“同氏” の先行詞は “ ディニ” となる.

図 4.3: 照応解析の例

<COREF ID="1">ベルルスコーニ</COREF>・<COREF ID="2">イタリア</COREF>首相の辞任を機に表面化した政治危機で、調停役の<COREF ID="3">スカルファロ</COREF>大統領は<COREF ID="4">三日</COREF>から、再び関係者との協議に乗り出し、<COREF ID="5">今週末</COREF>には暫定政権発足の見通しだ。
しかし、<COREF ID="6" REF="1">ベルルスコーニ</COREF>与党内には依然<COREF ID="7">議会</COREF>解散・総選挙を主張する声が強く、<COREF ID="8">四日</COREF>に行われる<COREF ID="9" REF="3">スカルファロ</COREF>・<COREF ID="10" REF="1">ベルルスコーニ</COREF>会談の結果次第では、暫定政権発足に失敗する可能性も残されている。
<COREF ID="11" REF="3">スカルファロ</COREF>大統領は新年のメッセージの中で、<COREF ID="12" REF="7">議会</COREF>解散・総選挙の道は<COREF ID="13" REF="2">イタリア</COREF>政治の国内外の信用にさらにダメージを与えかねないことを強調し、暫定政権発足を念頭に各政党間、議会関係者の調停を進めていることをより明確にした。
<COREF ID="14" REF="4">三日</COREF>から始まった第二ラウンドの調停では、大統領はまず<COREF ID="15">スコニャミリオ</COREF>、<COREF ID="16">ピベッティ</COREF>上下院議長と会談し、暫定政権発足の協力を強く要請した。
<COREF ID="17" REF="3">スカルファロ</COREF>大統領は、暫定政権の首相第一候補には前イタリア中央銀行副総裁でベルルスコーニ内閣の国庫相を務めた<COREF ID="18">ディニ</COREF>氏を推す意向といわれる。
<COREF ID="19" REF="18">ディニ</COREF>氏は無党派の人間ながら、<COREF ID="20" REF="1">ベルルスコーニ</COREF>氏の側近の一人で、経済通として内外に知られている。
さらに、野党の受けもいいため、大統領は<COREF ID="21" REF="8">四日</COREF>の<COREF ID="22" REF="1">ベルルスコーニ</COREF>氏との会談で、<COREF ID="23" REF="18">同氏</COREF>の説得に成功すれば、<COREF ID="24" REF="5">週末</COREF>には<COREF ID="25" REF="18">ディニ</COREF>氏を首相に指名し直ちに<COREF ID="26" REF="18">ディニ</COREF>暫定政権を発足させたい考えだ。

# 第 5 章

# 評価実験

提案手法を評価する実験を行った．実験に用いるデータは IREX の NE_DRYRUN データで，そのうち 34 記事を使用している．元のデータは 36 記事であるが，ALTJAWS の形態素解析のエラーが出る記事は除いている．これらの記事に対して，4.3.1 項で述べた手法により固有表現抽出を行なった．使用した固有表現タグは以下の 8 種類である．

- 組織名 (ORGANIZATION)
- 人名 (PERSON)
- 地名 (LOCATION)
- 固有物名 (法律名，商品名など)(ARTIFACT)
- 日付表現 (DATE)
- 時間表現 (TIME)
- 金額 (MONEY)
- 割合表現 (PERCENT)

IREX のデータには，正解となる固有表現タグが付与されている．この正解とシステムの出力とを比較し，再現率，精度，F 値を調べて評価を行なった．再現率，精度，F 値は以下の式で計算される．

$$\text{再現率} = R = \frac{\text{システムが出力した個数}}{\text{正解データの個数}} \tag{5.1}$$

$$\text{精度} = P = \frac{\text{システムの正解の個数}}{\text{正解データの個数}} \tag{5.2}$$

$$F\text{ 値} = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R} \tag{5.3}$$

ただし，式 (5.3) の係数 $\beta$ は 1 とする．

さらに，照応解析結果についても，再現率，精度，F 値による評価を行った．正解となる参照表現とその先行詞は人手で与えた．

## 5.1 固有表現抽出結果

固有表現抽出の結果を表 5.1から表 5.4に示す．また，表 5.1から表 5.4までに使われている用語を説明する．

**GLD** 正解データに含まれる固有表現の個数

**SYS** システムが作成したデータに含まれる固有表現の個数

**COR** システムが作成したデータのうち正解データにも含まれる固有表現の個数

**MIS** システムが抽出し損なった固有表現の個数 (GLD-COR)

**OVG** システムが過抽出した固有表現の個数 (SYS-COR)

**REC** 再現率 (COR/GLD)

**PRE** 精度 (COR/SYS)

**F-MEASURES** F 値

表 5.1は，rika の固有表現抽出結果を表している．rika の結果は [9] で報告されている値よりも劣っている．特に，ARTIFACT の抽出精度が極端に悪い．これは special dictionary[9] を使用していないことが原因であると考えられる．本研究では，既存の固有表現抽出の精度を照応解析により，向上させることを目的としていたので，この結果がどのように変化するのかをみていく．

表 5.2は，照応解析の結果，同一の対象とみなされた参照表現のうち，初期の固有表現抽出で抽出されなかった固有名詞に対して，新たにタグを付与したときの結果である．ここでは，同一の対象を指す参照表現に付与された固有表現タグが矛盾していても，それらを統一する処理は行なっていない．

初期の固有表現抽出によって抽出できなかった表現が抽出されたため，再現率は良くなっている．しかし，表 5.2と表 5.1を比較すると，PERSON の精度が悪い．この原因のひとつは，ALTJAWS の形態素解析の誤りである．特に,「てる」や「しょう」などの形態素を人名として誤認識していた．このために，これらの形態素が参照表現として抽出されて，PERSON タグが付与された．表 5.2の OVG(過抽出) の数が多くなったのはこのためである．

表 5.3は，表 5.2の結果から，タグを統一した結果であり，本稿で述べた手法の結果である．しかし，表 5.2と比べて，F 値は変化がない．結果を観察すると，LOCATION の過抽出が目立つ．この原因は１文字の固有表現（例えば，日，米）も記事内のすべての表現に LOCATION タグを付与することにより，固有表現とは関係のない文字も抽出してしまっていたためである．

以上のことから，PERSON については照応解析の結果を用いないようにし，さらに LOCATION についてはタグを統一しないようにシステムを改良した．その結果が表 5.4 である．改良したことによって，F 値がわずかに上がった．しかし，このシステムは純粋なオープンテストではなく，あくまで参考データである．

rika による固有表現抽出結果と，本研究で提案した手法による固有表現抽出の結果を改めて表 5.5 に示す．表 5.5 の右側は表 5.3の再掲である．

提案手法は，rika と比べて，F 値がわずかに向上した．次節で示すように，照応解析の再現率や精度が十分に高くないのも関わらず，固有表現抽出の F 値が向上したことから，照応解析の結果を利用して固有表現抽出のタグ付け結果を修正することは有効であると言える．照応解析の性能が向上すれば，固有表現抽出の精度や再現率もさらに改善されると期待できる．

精度と再現率を見ると，照応解析結果を用いることにより，精度は低下したが再現率は向上している．再現率が向上したのは，rika で固有表現として抽出されなかった参照表現も，照応解析結果に基づくタグの修正によって，新たに固有名詞タグを付与することができたためと考えらえる．

表 5.1: rika(ALTJAWS) のみの結果

| | GLD | SYS | COR | MIS | OVG | REC | PRE |
|---|---|---|---|---|---|---|---|
| ORGANIZATION | 188 | 120 | 80 | 108 | 40 | 42.55 | 66.67 |
| PERSON | 134 | 128 | 100 | 34 | 28 | 74.63 | 78.12 |
| LOCATION | 190 | 172 | 117 | 73 | 55 | 61.58 | 68.02 |
| ARTIFACT | 42 | 1 | 0 | 42 | 1 | 0.00 | 0.00 |
| DATE | 112 | 120 | 85 | 27 | 35 | 75.89 | 70.83 |
| TIME | 22 | 19 | 14 | 8 | 5 | 63.64 | 73.68 |
| MONEY | 33 | 33 | 33 | 0 | 0 | 100.00 | 100.00 |
| PERCENT | 6 | 5 | 5 | 1 | 0 | 83.33 | 100.00 |
| ALL SLOTS | 727 | 598 | 434 | 293 | 164 | 59.70 | 72.58 |
| F-MEASURES | | | | | | | 65.51 |

表 5.2: 照応解析の結果から新たに固有表現タグを付与した結果

| | GLD | SYS | COR | MIS | OVG | REC | PRE |
|---|---|---|---|---|---|---|---|
| ORGANIZATION | 188 | 155 | 98 | 90 | 57 | 52.13 | 63.23 |
| PERSON | 134 | 147 | 101 | 33 | 46 | 75.37 | 68.71 |
| LOCATION | 190 | 214 | 133 | 57 | 81 | 70.00 | 62.15 |
| ARTIFACT | 42 | 1 | 0 | 42 | 1 | 0.00 | 0.00 |
| DATE | 112 | 125 | 91 | 21 | 34 | 81.25 | 72.80 |
| TIME | 22 | 19 | 14 | 8 | 5 | 63.64 | 73.68 |
| MONEY | 33 | 33 | 33 | 0 | 0 | 100.00 | 100.00 |
| PERCENT | 6 | 5 | 5 | 1 | 0 | 83.33 | 100.00 |
| ALL SLOTS | 727 | 699 | 475 | 252 | 224 | 65.34 | 67.95 |
| F-MEASURES | | | | | | | 66.62 |

表 5.3: タグの統一処理をした結果

| | GLD | SYS | COR | MIS | OVG | REC | PRE |
|---|---|---|---|---|---|---|---|
| ORGANIZATION | 188 | 162 | 102 | 86 | 60 | 54.26 | 62.96 |
| PERSON | 134 | 151 | 103 | 31 | 48 | 76.87 | 68.21 |
| LOCATION | 190 | 226 | 135 | 55 | 91 | 71.05 | 59.73 |
| ARTIFACT | 42 | 1 | 0 | 42 | 1 | 0.00 | 0.00 |
| DATE | 112 | 125 | 91 | 21 | 34 | 81.25 | 72.80 |
| TIME | 22 | 20 | 14 | 8 | 6 | 63.64 | 70.00 |
| MONEY | 33 | 33 | 33 | 0 | 0 | 100.00 | 100.00 |
| PERCENT | 6 | 5 | 5 | 1 | 0 | 83.33 | 100.00 |
| ALL SLOTS | 727 | 723 | 483 | 244 | 240 | 66.44 | 66.80 |
| F-MEASURES | | | | | | | 66.62 |

表 5.4: 改良した結果

| | GLD | SYS | COR | MIS | OVG | REC | PRE |
|---|---|---|---|---|---|---|---|
| ORGANIZATION | 188 | 162 | 102 | 86 | 60 | 54.26 | 62.96 |
| PERSON | 134 | 152 | 106 | 28 | 46 | 79.10 | 69.74 |
| LOCATION | 190 | 213 | 133 | 57 | 80 | 70.00 | 62.44 |
| ARTIFACT | 42 | 1 | 0 | 42 | 1 | 0.00 | 0.00 |
| DATE | 112 | 125 | 91 | 21 | 34 | 81.25 | 72.80 |
| TIME | 22 | 20 | 14 | 8 | 6 | 63.64 | 70.00 |
| MONEY | 33 | 33 | 33 | 0 | 0 | 100.00 | 100.00 |
| PERCENT | 6 | 5 | 5 | 1 | 0 | 83.33 | 100.00 |
| ALL SLOTS | 727 | 711 | 484 | 243 | 227 | 66.57 | 68.07 |
| F-MEASURES | | | | | | | 67.32 |

表 5.5: 固有名詞抽出の実験結果

| | rika | | 提案手法 | |
|---|---|---|---|---|
| | 再現率 | 精度 | 再現率 | 精度 |
| ORG | 42.55 | 66.67 | 54.26 | 62.96 |
| PERSON | 74.63 | 78.12 | 76.87 | 68.21 |
| LOCATION | 61.58 | 68.02 | 71.05 | 59.73 |
| ARTIFACT | 0.00 | 0.00 | 0.00 | 0.00 |
| DATE | 75.89 | 70.83 | 81.25 | 72.80 |
| TIME | 63.64 | 73.68 | 63.64 | 70.00 |
| MONEY | 100.00 | 100.00 | 100.00 | 100.00 |
| PERCENT | 83.33 | 100.00 | 83.33 | 100.00 |
| ALL SLOTS | 59.70 | 72.58 | 66.44 | 66.80 |
| F 値 | 65.51 | | 66.62 | |

## 5.2 照応解析結果

まず，照応解析の結果を表 5.6 に示す．表 5.6 における “基準 1”, “基準 2”, “基準 3” は，4.3.1 項で述べた先行詞候補のリスト $AL$ の順序を決定する 3 つの基準を表わす．この表からわかるように，照応解析の再現率，精度ともに十分高いとは言えない．この原因のひとつは，ALTJAWS の形態素解析の誤りである．例えば,「見てる」を ALTJAWS で形態素解析した場合,「てる」が人名になり,「同氏」などの先行詞として誤認識される．このとき，その文書に現われる「同氏」はすべて不正解となり，精度を著しく下げる要因となっている．また，再現率が低い理由として，ALTJAWS の形態素区切りが参照表現の区切りと一致せず，参照表現として取り出すべき名詞が取り出されていないことが挙げられる．これについては，3.3.2 項で説明したように，複数の形態素を連結して参照表現を抽出するように工夫はしているが，十分な成果が得られていないことがわかった．

本研究では,「同」を用いた表現の先行詞を決定するために,「言及クラス」「言及回数」といった素性を新たに導入した．しかし,「言及クラス+言及回数」を最も重視した基準 1 と,「文法属性」を最も重視した基準 2 とでは，F 値にほとんど差が見られない．これは，実験に用いた新聞記事の中に,「同」を含む表現があまり存在しなかったためである．評価した記事の中に参照表現は 460 個あったが，そのうち「同」を含む表現はわずかに 13 であった．したがって，本研究で提案する「言及クラス」や「言及回数」が，照応解析に用いる素性として有効であることを確かめることができなかった．今後,「同」を含む表

表 5.6: 照応解析の結果

| | 基準 1 | 基準 2 | 基準 3 |
|---|---|---|---|
| 再現率 | 34.78 | 34.78 | 31.52 |
| 精度 | 54.42 | 54.23 | 52.16 |
| F 値 | 42.44 | 42.38 | 39.29 |

現の事例を増やして，評価実験を行う予定である．

### 5.2.1 センタリング理論との比較

2.2 節で述べたセンタリング理論と本手法の比較を行なった．CRL 固有表現データ中で「同」を用いた表現が存在し，先行詞候補が２つ以上の記事 (１８記事で,「同」を用いた表現が２２事例ある）に対して，人手でセンタリング理論と提案手法を適用して実験を行なった．実験を行なうにあたり，以下の項目を設定して実験を行なった．

- すべての参照表現は正しい範囲で認識されている
- 固有表現タグの種類は正しく認識されている
- センタリング理論において，直前の文に先行詞がない場合は，文を遡って先行詞を特定する
- センタリング理論において，文法属性で $Cf$ の順序が決定されない場合は，不正解とする

  例 :「○○会社と□□会社は … 」という文の場合,「○○会社」と「□□会社」のどちらを優先すべきかについては不明である．

- センタリング理論は２つの方法で解析する

  **センタリング理論 (1)** 複文のまま解析 (Walker ら [13] の方法)

  **センタリング理論 (2)** 複文を述語で区切ることで，複文を単文に分割し，文内照応 (先行詞が照応詞と同じ文に存在する場合の照応) も解析 (田村 [5] の方法)

表 5.7: センタリング理論と提案手法の比較

| | 精度 |
|---|---|
| センタリング理論 (1) | 59.09 (13/22) |
| センタリング理論 (2) | 77.27 (17/22) |
| 提案手法 | 86.36 (19/22) |

実験結果を表 5.7に記す.

センタリング理論 (1) と提案手法を比較すると，精度が顕著に向上した．提案手法は参照表現が正しく認識されれば，新聞記事のような複雑な文書に対して，センタリング理論よりも良い精度があることがわかった．センタリング理論は複雑な文書には弱いことが知られている．これは，文法属性で $Cf$ の順序が決定されない場合，どちらを優先させるかを判断できないこと（例えば，上記の例では，○○会社と□□会社のどちらが優先させるかわからない）が主な原因である．また，センタリング理論は直前の文の要素しか先行詞の候補としない．今回の実験では，先行詞が見つかるまで遡って特定しているが，直前の文のみを先行詞の候補の対象にすると，精度はさらに悪くなる．このため，文書内の複文を単文にすることにより，この問題を解決した方法がセンタリング理論 (2) である．センタリング理論 (2) と提案手法を比較した．今回の実験のデータ量が少量であるので，センタリング理論 (2) より良いと断定できないが，少なくとも提案手法は，センタリング理論 (2) と遜色のない精度であると言える．

今回は小規模な実験しか行なえなかったが，さらにデータを増やして実験して，提案手法の有効性を確かめたい.

# 第 6 章

# おわりに

本稿では固有表現抽出の精度を向上させるために照応解析の結果を利用する手法を提案し，その有効性を実験により確認した．また，固有名詞を対象とした照応解析の新たな素性として「言及クラス」を提案し，他の素性と併用する照応解析手法を提案した．

最後に，今後の課題を 3 つ挙げる．

1. 言い換え表現について

   提案手法は言い換え表現には対応していない．言い換え表現とは，例えば「第二電電」と「DDI」などのように，日本語表記が英語表記の省略形になった表現や,「朝鮮民主主義人民共和国」と「北朝鮮」などのように，省略とは違う形で言い換えられる表現である．言い換え表現は $LCS$ によって関連性を見つけることが難しく，固有名詞の言い換え表現を収集した辞書が必要となる．

2. 照応解析の再現率の改善

   今回の実験では，形態素解析による単語区切りと参照表現の境界が一致しなかったとき，いくつかのパターンを用意して単語境界を修正し，参照表現の抽出を行なった．しかし，このような参照表現の結合パターンは数多く存在すると考えられるので，参照表現の結合パターンについても自動的に学習すれば，再現率は向上するだろう．

3. 固有表現を指す普通名詞と省略表現について

   3.3 節で挙げた 3 つの参照表現のタイプのうち，固有表現を指す普通名詞と省略表現は $LCS$ を用いて先行詞を特定していた．しかし，例えば「北陸先端大」の後に単に「大学」とあれば，$LCS$ では同じ順序で文字が現われていないため，同一の対象

とは見なされない．$LCS$ によるスコアが低くても，ALTJAWS が出力する意味属性が等しければ，先行詞と見なす手法も試みたが，精度が悪化した．$LCS$ に代わる判定基準を検討する必要がある．

# 謝辞

本研究を進めるにあたり，終始熱心な御指導を賜わりました白井清昭助教授に心から感謝致します．また，本研究に関して当初ご指導を賜りました東京工業大学 精密工学研究所 奥村学助教授に心から感謝致します．さらに数多くの御教授を頂きました島津明教授に厚く御礼申し上げます．望月源助手ならびに自然言語処理学講座の皆様には，貴重な御意見、討論をして頂きました事を感謝致します．

本研究において固有表現抽出システムを使用させて頂きましたニューヨーク大学 コンピュータサイエンス学科 関根聡 助教授に感謝致します．

最後に，多くの方々の御援助によって本研究を行なうことができましたことを厚く御礼申し上げます．

# 参考文献


[1] Barbara J.Groze, Arvind K.Jpshi, and Scott Weinstein. Providing a unified account of definite noun phrases in discourse. *21st Annual Meeting of the Association for Computational Linguistic*, pp. 44–50, 1983.

[2] Megumi Kameya. A property-sharing constraint in centering. *24th Annual Meeting of the Association for Computational Linguistic*, pp. 200–206, 1986.

[3] Tsuyoshi Kitani. Merging information by discourse processing for information extraction. *tenth IEEE Conference on Aritficial Intelligence for Applications*, pp. 412–418, 1994.

[4] 内元清貴, 馬青, 村田真樹, 小作浩美, 内山将夫, 井佐原均. 最大エントロピーモデルと書き換え規則に基づく固有表現抽出. 自然言語処理，Vol.7 No.2, pp. 63–90, 2000.

[5] 田村浩二. センター理論による日本語談話の省略解析. Master's thesis, 北陸先端大学院大学情報科学研究科, 1995.

[6] 石崎雅人, 伝康晴. 談話と対話. 東京大学出版会, 2001.

[7] NTT コミュニケーション科学研究所. 日本語語彙大系. 岩波書店, 1997.

[8] J. Ross Quinnan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, 1995. 邦訳：AI によるデータ解析，古川 康一監訳，トッパン，1995.

[9] Satoshi Sekine, Ralph Grishman, and Hiroyuki Shinnou. A decision tree method for finding and classifying names in japanese texts. *COLING-ACL'98,Proceedings of the Sixth Workshop on Very LargeCorpora*, pp. 171–178, 1998.

[10] 高田眞吾, 土居範久. 省略された代名詞の解釈-工学系-. 日本語学，vol.14, 1995.

[11] Takahiro Wakao. Reference resolution using semantic patterns in japanese newspaper articles. *In Proceedings of COLING 94*, pp. 1133–1137, 1994.

[12] Marilyn Walker, Masayo Iida, and Sharon Cote. Centering in japanese discourse. *In Proceedings, 13th International Conference on Computational Linguistics(COLING-90)*, 1990.

[13] Marilyn Walker, Masayo Iida, and Sharon Cote. Japanese discourse and the process of centering. *Computational Linguistics*, Vol. 20, No. 2, pp. 193–232, 1994.

[14] ＮＴＴコミュニケーション科学基礎研究所知能情報研究部 翻訳コミュニケーション研究グループ. ＮＴＴ日本語形態素解析ライブラリ ＡＬＴＪＡＷＳ Version 2.0, 1999.

[15] 山田寛康, 工藤拓, 松本祐治. Support vector machines を用いた日本語固有表現抽出. 情報処理学会研究会報告,No.NL-142-18, pp. 121–128, 2001.