修士論文

自動獲得した知識に基づく

統合的な照応解析

指導教員　　黒橋　禎夫　助教授

東京大学大学院 情報理工学系研究科　電子情報学専攻

46413　　笹野　遼平

平成18年2月3日作成

内 容 梗 概

ある言語表現が、これに先行する言語表現と同一の内容ないしは同じ対象を指す場合、これらの表現は照応関係にあるという。照応関係を認識する照応解析は、計算機による意味理解において重要な技術である。これまで照応解析に関する研究は、照応現象は照応詞のタイプごとにその性質が異なるため、直接照応、ゼロ照応など照応の種類ごとに独立して行われることが多かった。本論文では、文章中の要素間の様々な関連性を認識することを目指し、直接照応や、間接的な照応現象の一種である橋渡し指示などの照応の解析を自動獲得した知識や、固有表現認識技術を用いて統合的に行う。

目 次

第 1章： 序論

書き手が読み手に伝えたい内容はさまざまな概念や事象が相互に関連し、ネットワーク構造をなしている。例えば次のような文章があった場合、「その店」というのは「近所の洋菓子屋」であり、「値段」とは「ケーキの値段」のことであり、また、「良く買っている」のは「太郎」である。

(1) 太郎は近所の洋菓子屋でケーキを買った。値段は高いがその店のケーキはおいしいので良く買っている。

ところが、文章は一次元の文字列で表現されるため、これらの関係の多くは明示されない。したがって、読み手が書き手の伝えたいことを理解するためには、文章に現れる文と文、語と語、語と概念、あるいは概念同士などの関連性を、背景の状況やそれまでの話の過程などと結びつけながら認識することが必要となる。

計算機を用いてこれらの関連性を認識するためには、形態素解析 (morphological analysis) や固有表現認識、構文解析 (parsing)、格解析 (case analysis)、照応解析 (anaphora resolution) などの処理を行う必要がある。形態素解析とは文を構成する最小の意味単位である形態素に分解する処理、固有表現認識は組織名や人名を認識する処理、構文解析とは文中の修飾-非修飾関係 (係り受け) を認識する処理、格解析とは対象の文の格フレーム構造を決定する処理のことである。照応 (anaphora) とは、ある言語表現が他の言語表現と同じ内容や対象をさす現象のことであり、照応解析とは照応関係を同定する処理である。このとき先に出現した表現を先行詞 (antecedent)、これに対応する後続の表現を照応詞 (anaphor) と呼ぶ。

これらの解析のうち、形態素解析については 99%程度、構文解析については 90%程度の高精度な解析がすでに実現されている。一方、照応解析に関しては、近年様々な研究が行われているが、50〜70%程度の精度の解析しか実現できていない。高精度な照応解析システムの構築は文章中の要素間の関連性理解に向けて重要であり、また、自動要約、機械翻訳、質問応答などの言語処理アプリケーションの高度化に向けても大きな役割を果たすと考えられる。

照応に関する研究としては、直接的な照応関係の一種である共参照 (coreference) や、ゼロ照応 (zero anaphora)、橋渡し指示 (bridging reference) に関する研究が行われている。ここで、ゼロ照応とは、日本語などの言語では多く見られる現象で、(2) の例文 b のように用言の格要素にあたる照応詞が省略される場合の照応のことを指す、

(2) a. お菓子を買った。それをおいしく食べた。
b. お菓子を買った。(ϕを) おいしく食べた。

また、次の例において「値段」とは「りんごの値段」のことを指すが、橋渡し指示とはこの例における「値段」と「りんご」の間にあるような間接的な指示関係のことを言う。

(3) りんごを買った。値段は 100 円だった。

これまでの照応解析に関する研究では、直接照応やゼロ照応、橋渡し指示などが独立して研究されてきた。また、例えば直接的な照応現象については照応詞が現実世界の実体を指示する表現の場合である共参照についてのみを扱ったり、照応詞として文節の主辞のみを考えるなど、様々な制限が設けられていた。しかし、直接照応やゼロ照応、橋渡し指示などといった照応関係、さらには形態素解析、固有表現認識、構文解析、格解析などといった処理は相互に関連性を持っており、これらの解析を統合的に行うことによりそれぞれの解析を高精度化することが可能だと考えられる。

本研究では、文章中の要素間の関連性認識を目指し、従来個別に扱われてきた直接照応やゼロ照応、橋渡し指示などといった照応現象を、固有表現認識技術や事前に獲得した知識を用いて統合的に解析することを目的とする。

本論文の構成は以下のようになっている。まず、2 章で扱うべき照応現象について述べたあと、3 章で直接照応の解析を行う際に有用となる同義表現や、橋渡し指示の解析を行う際に有用となる名詞格フレームの自動獲得について説明する。続く 4 章で照応解析において重要となる固有表現認識、5 章、6 章では、直接照応解析、橋渡し指示の解析について説明する。7 章では固有表現認識なども含む統合的な照応解析システムの構築について述べ、最後に 8 章で結論を述べる。

第 2章 :　照応現象について

ある言語表現が、これに先行する言語表現と同一の内容ないしは同じ対象を指す場合、これらの表現は「照応関係」(anaphoric relation) にあるとされる。この場合、前者の表現は「先行詞」(antecedent)、これに対する後続の表現は「照応詞」(anaphora) と呼ばれる。本章では、まず照応現象の分類を行い、続いて本研究で扱う照応現象を明らかにする。

2.1　照応現象の分類

2.1.1　先行詞の出現位置による照応の分類

先行詞と照応詞の順序関係からみた場合、照応は「前方照応」(anaphora) と「後方照応」(cataphora) に分けられる。

(4) **赤外自由電子レーザー**と それ を用いる光科学

(5) いつも そう なんだが、**彼は食べるのが早い。**

(4) のような文章があった場合、照応詞「それ」は、先行する「赤外自由電子レーザー」を指している。一般に、このように先行詞が照応詞よりも前にある照応は「前方照応」と呼ばれる。一方、(5) の場合、照応詞「そう」の先行詞にあたる表現「彼は食べるのが早い」は照応詞よりも後に来ている。このように先行詞が照応詞より後に来る照応は「後方照応」と呼ばれる。

これらの例の場合、先行詞は照応詞の前に出現するか、後に出現するかの違いはあるものの、いずれの場合も文章中に認められる。照応詞の指示する対象が文章中に出現する照応はまとめて「文脈照応」(endophora) と呼ばれる。一般に照応現象という場合には、このタイプの照応を問題とする。しかし、照応を広い意味に理解するならば、次のような現象もその一部とみなすこともできる。

(6) ここ に移り住んで一年になる。

(7) ちょっと そのコップ 取って。

この場合には、照応詞の照応先としての先行詞は、文章中には認められず、問題の発話における言語外の場面の中に認められる。この種の照応を、「外界照応 (exophora)」として、広い意味での照応現象の下位類とみなすこともできる。照応現象を広く解釈するならば、照応は図 2.1 のように分類することができる。

照応 $\begin{cases} \text{文脈照応 (Endophora)} - \begin{cases} \text{前方照応 (Anaphora)} \\ \text{後方照応 (Cataphora)} \end{cases} \\ \text{外界照応 (Exophora)} \end{cases}$

図 2.1: 照応の分類

2.1.2 照応詞の種類による照応の分類

(8) a. **自民党**の政調会長は幹事長・総務会長と共に、自民党 の 3 役の 1 つだ。
b. **自民党**の政調会長は幹事長・総務会長と共に、党 3 役の 1 つだ。

文脈照応は照応詞の現れ方によって分類できる。(8) のような文章があった場合、a では後続する「自民党」が先行する「自民党」を照応しており、b では「党」が「自民党」を照応している。いずれの場合も照応詞は明示的に存在しており、照応詞は名詞句となっている。このように、名詞句が照応詞となる照応現象を「名詞句照応」と呼ぶ。

(9) 4 番は**松井**だと考えていた。彼 はそれだけの存在だ。

この場合も (8) の場合と同様に、「松井」を先行詞とする照応詞「彼」は文章中に明示的に存在しているが、この場合の照応詞は代名詞である。このように、人称代名詞や、指示代名詞なども照応詞となる。このような照応現象を「代名詞照応」と呼ぶ。

(10) **太郎**が学校から帰ってきた。しかし、すぐに ($\underline{\phi}$ガ) 出かけてしまった。

一方このような文章があった場合、「出かけてしまった」のガ格にあたる要素は、「太郎」を指示しているが、照応詞は明示的には出現していない。このような照応現象をゼロ照応 (zero anaphora) と呼び、省略された照応詞をゼロ代名詞 (zero pronoun) と呼ぶ。「名詞句照応」、「代名詞照応」、「ゼロ照応」の関係を図 2.2 に示す。

$\begin{cases} \text{照応詞が明示的に存在} - \begin{cases} \text{照応詞が名詞句} - \text{名詞句照応} \\ \text{照応詞が代名詞} - \text{代名詞照応} \end{cases} \\ \text{照応詞が省略されている} - - - - - - - - - - \text{ゼロ照応} \end{cases}$

図 2.2: 照応詞の種類による照応の分類

2.1.3 先行詞の種類による照応の分類

照応の中には、先行詞に相当する要素が先行文脈中に明示的には出現せず、推論のプロセスを介することによって初めて理解される照応現象が存在する。このため、先行詞が明示的かどうかによっても以下のように照応現象を分類することができる。

表 2.1: 直接照応、ゼロ照応、間接照応の分類

	照応詞が明示的	照応詞が省略
先行詞が明示的される	直接照応	ゼロ照応
先行詞が明示されない	間接照応	

- 問題の照応詞に対応する先行詞が、前後関係を規定する文脈や状況、場面に直接的に認められる場合
- 問題の照応詞に対応する先行詞が、前後関係を規定する文脈や状況、場面には明示されず間接的に推定される場合

(4)、(5)、(8)、(9)、(10) などの例が前者に該当する。このように先行詞が直接的に認められる場合は直接照応と呼ばれる。ただし、通常、直接照応という言葉は先行詞が直接的に認められるのに加え、照応詞が省略されていない場合、すなわち照応詞および先行詞がともに明示的に認められる場合に使われる。本論文でも直接照応といった場合、ゼロ照応を含まないものとする。

これに対し、後者のような場合の照応現象を間接照応と呼ぶ[1]。表 2.1 に本研究における「直接照応」、「間接照応」、「ゼロ照応」の関係を示す。山梨 [25] は間接照応の例として以下のような例を挙げている。

(11) その村では、今年もまたひとりの少年が家出した。彼ら は都会へのあこがれを捨てることができないらしい。

この場合、「彼ら」は複数の対象を意味する代名詞であり、「その村の少年達」のことを指しているが、これらの代名詞に直接対応する先行詞は先行文脈には存在していない。

(12) 行きつけの…アナ場にやって来た旦那は、女房を待ちながら、それ までにでっかいのを釣りあげて彼女を驚かしてやろうと糸を投げた。

この場合の代名詞「それ」は、「女房がアナ場にやって来る」を意味するが、このような表現は先行文脈には存在しない。しかし、先行文脈の「女房を待ちながら」の部分からの推論のプロセスを介して、問題の先行詞の理解が可能となる。

(13) チンパンジーは、ありずかの小さな穴を見つけると、近くの木の皮をはいできて、歯と手をうまく使って長さ 20 センチメートルぐらいの細い棒を作ります。チンパンジーはこの棒を作るのにしなやかで強い木の皮を探します。そう でないと、穴にさし込む時、途中で折れたり曲がったりして、シロアリの所まで届かないからです。

[1]英語の場合、direct anaphora を主辞が同一の表現である場合の照応現象に、indirect anaphora を主辞が異なる語である場合の照応現象に用いられることが多い。

この例の場合も、複合的な推論のプロセスが照応の理解に関わってくる。この場合の「そう」の先行詞は、一見すると、その直前の文「棒を作るのにしなやかで強い木の皮を探す」に相当するように見える。しかし、しなやかで強い木の皮を探すこと自体が問題になっているわけではない。この場合、推論のプロセスを経ることにより、「しなやかで強い木の皮でできた棒を使う」が、問題の代名詞の先行詞として理解されることになる。

2.1.4 橋渡し指示について

次のような文章があった場合、それぞれ「チケット」と「値段」、「小屋」と「屋根」は関連性を持っている。

(14) **チケット**を買った。値段は 2000 円だった。

(15) 池のほとりに**小屋**があった。屋根は白かった。

(14) における「値段」とは「チケットの値段」であり、「チケット」と「値段」は強い関連性を持っている。「何かの値段」であることは明示的に表現されてはいないが、「値段」とは「何かの値段」であり、この場合はその「何か」にあたるものが「チケット」である。同様に (15) の文における「屋根」とは「小屋の屋根」である。

これらの例は、それぞれ「チケットの値段」、「小屋の屋根」が文章中に明示的に出現していないので間接照応の一種であると言える。しかし、2.1.3 節で紹介した例と比べると、「チケット」と「値段」、「小屋」と「屋根」という語同士の関係が強いと言え、それぞれ「チケット」「小屋」という語を、「A の B」という形で結ばれるような関係で、間接的に照応しているとも考えられる。このように、照応詞にあたる語がある要素を必要としており、その要素にあたる語を間接的に照応しているような場合を特に橋渡し指示 (bridging reference) と呼ぶ。

橋渡し指示は間接照応に一種であり基本的には推論が必要となるが、照応詞にあたる語がどのような知識を必要とするかさえ分かれば、解決が可能となる指示関係であると言える。以降、橋渡し指示については、照応詞と直接、同一の対象・内容を指示するわけではないが、(14) における「チケット」や、(15) における「小屋」にあたる語を先行詞と呼ぶことにする。

2.2 解析対象とする照応現象

2.2.1 照応現象であるとみなす範囲

照応と近い意味を持つ語として共参照という語がある。ある 2 つの表現が共参照 (coreference) 関係にあるという場合は、一方の表現ともう一方の表現が同じ対象を指示する場合を意味する。この場合の「同一性」はせまい意味での「指示的な同一性」(referential

表 2.2: 照応、共参照の捉え方

出現しない	文脈中に先行詞に相当する要素が出現	
外界照応	文脈照応	
	共参照	
指示的 (referential)		
現実世界の実体を指示する		できない

identity) が基準となっていることが多く、他の表現と共参照関係となるような表現は指示的な名詞句 (現実世界の実体を指示する表現) に限られる。

照応における「同一性」の基準を、せまい意味での「指示的な同一性」とし、照応詞を現実世界の実体に指示する表現に限定することも考えられる。しかしながら、文脈中に現れる照応現象の照応詞は必ずしもこのような表現であるとは限らない。

(16) a. **犬**は鼻がいつも濡れています。
これは、犬は人間のように汗をかいたりできないためです。
b. **犬**は鼻がいつも濡れています。
これは、人間のように (ϕが) 汗をかいたりできないためです。

例えばこのような文があった場合、通常、2 文目の「汗をかく」のガ格にあたる照応詞は、総称名詞である 1 文目の「犬」を照応していると考えられる。

本研究では、できるだけ多くの照応現象の認識を行うため、照応表現の同一性を問題にする場合には、問題の言語表現の概念的な「意味の同一性」(conceptual identity) も広い意味で含めることとする[2]。表 2.2 に本研究における「照応」や「共参照」などの語の指す範囲を示す。

また、言語学の分野などでは、次の例のように照応詞と先行詞にあたる表現が表層的に同一である場合を照応とは呼ばないことがあるが、本研究では、これらの表現も照応関係にあると考えることとする。すなわち、次のような文があった場合、2 文目に出現する「自民党」は、1 文目に出現した「自民党」を照応しているとみなす。

(17) 政党支持率は**自民党**がトップだった。しかし、政権選択については民主党中心が 自民党 中心を上回った。

このように考えることにより、すべての名詞句は次の 2 つに分けることが出来るようになる。

- 先行文脈中に直接的または間接的に同一の内容ないし同じ対象を指す表現が存在する
- 先行文脈中に直接的にも間接的にも同一の内容ないし同じ対象を指す表現が存在しない

[2]現実世界の実体を指示する表現であるかどうかの判断は別の課題であると考える。

本研究では、文章中の要素間の関連性認識を行うことを目的とする。このため基本的に外界照応については扱わない。文脈照応については、先行詞が直接的に認められる場合 (直接照応) と、橋渡し指示に限り先行詞が間接的に認められる場合 (間接照応) の解析を行う。ゼロ照応については既に格フレームを用いて解析する枠組みが既に構築されているので解析は行うものの、本論文ではその手法の説明、評価は行わない。

次節では、直接照応と橋渡し指示に分けてどのような現象を解析対象とするかについて詳述する。

2.2.2 解析対象とする照応現象

直接照応

直接照応の照応詞としては、まず、普通名詞、固有名詞、代名詞などが考えられる。以下の (18)、(19)、(20) の例では下線によって照応詞を表わしているが、それぞれ普通名詞、固有名詞、代名詞となっている。本研究ではいずれの場合も解析の対象とする。ただし、(21) における「これ」のように、先行詞にあたる要素が全文の意味内容である場合など、文章中の単語で表せない場合は解析対象としない。

(18) **ブッシュ大統領**は再選に向けた準備を進めている。大統領はこれまでに 2 億ドル近くの選挙資金を集めた。

(19) **ブッシュ大統領**は再選に向けた準備を進めている。ブッシュ氏はこれまでに 2 億ドル近くの選挙資金を集めた。

(20) **ブッシュ大統領**は再選に向けた準備を進めている。彼はこれまでに 2 億ドル近くの選挙資金を集めた。

(21) **伊達はバックサイドにボールを集め、相手のミスを誘う作戦に出た。**これが功を奏し、このセットをものにした。

また、単独の名詞以外にも複合名詞や、その部分形態素列が照応詞となる場合も考えられる。

(22) 金融派生商品取引の全容を解明するため世界統一の調査を実施することを明らかにした。調査内容は、金融派生商品の対象となる為替、金利、エクイティ、コモディティの四つの大きなグループに分け、…

例えば、このような文章があった場合は 1 文目「金融派生商品」と 2 文目の「金融派生商品」、1 文目の「調査」と 2 文目の「調査」が同一の対象を指していると言うことができる。このため、基本的に、複合名詞全体や、複合名詞中の部分形態素列も照応詞、先行詞として考える。ただし、固有表現中の部分形態素列については、本研究では照応の要素としては扱わないこととする。これは固有表現は全体で意味をなす場合が多いためである。

表 2.3: MUC-6 における markable の定義

1. 名詞 (nouns)、名詞句 (noun phrases)、代名詞 (pronoun) に分類される要素 (elements) を markable とする。
2. 固有表現 (named entities) については、定義については NE タスクに準拠し、すべて markable とする。ただし、部分文字列は markable とはしない。
3. 動名詞はすべて markable としない。
4. 代名詞については、所有格 (possessive forms) の代名詞、および人称代名詞 (personal pronouns) を markable とする。
5. 裸名詞 (bare nouns) については名詞を前から修飾する場合 (prenominal occurences) は markable とする。主辞だけ取り出して markable とすることはしない。
6. ゼロ代名詞 (implicit pronouns) は英語にはないとして扱わない。
7. 複数の主辞を持つ名詞句 (conjoined noun phrase) は markable としない。

(23) 日本最大の生命保険会社である 日本 生命は、… (日本生命=“組織名”)

例えば、(23) における「日本生命」中の「日本」は、「日本」だけを取り出して意味を考えれば「日本」という国を意味しているが、通常は「日本生命」という組織名の一部として捉えられ、「日本生命」中の「日本」が前方の「日本」を照応しているとは考えない。

日本語に関しては、複合名詞や固有表現を構成する語をどこまで照応詞、先行詞として考えるべきかの一般的な基準は定まっていないが、英語に関しては、MUC-6 において照応現象に関わることが可能な単位を markable と呼び、照応解析を行う対象が定義されている。もちろん、MUC-6 で markable であると定義された語以外も、照応詞、先行詞となる場合を考えることはできるが、ある程度の一般性を持った基準を定義することは妥当なことだと考えられる。表 2.3 に MUC-6 における markable の定義を示す。

基本的には、名詞句全体を markable とし、固有表現以外の名詞句については、名詞を前から修飾する名詞については markable としている。例えば次の例文のような場合、“aluninum siding” は固有表現ではないので、“aluninum” は markable と判断され、後続する”aluninum” と共参照解析の対象とされる。

(24) The price of aluninum siding has steadily incresed, as the market for aluninum reats to the strike in Chile

一方、次のような文章があった場合、“Iowa Cos.” は組織名であるので、後続する “Iowa” と “Iowa Cos.” 中の “Iowa” の関係は解析の対象とはされない。

(25) Equitable of Iowa Cos. ... located in Iowa. (Iowa Cos.=“組織名”)

表 2.4: 本研究における照応詞、先行詞になりうる語の定義

1. 名詞句全体
2. 複合名詞中の部分形態素列 (ただし、固有表現はそれ以上分割しない)
3. 代名詞

MUC-6 では先行詞、照応詞がともに markable となるようなペアのみを共参照解析の対象としている。この定義は、すべての照応現象を網羅できるわけではないが、基本的な照応関係の多くを含んでいると考えられる。

表 2.4 に本研究における照応詞、先行詞になりうる表現の定義を示す。照応詞、先行詞がともにこの条件を満たすような照応関係のみを、直接照応の解析対象として考える。

例えば、(22) のような文章があった場合は、2 文目に出現する「金融派生商品」が、この「金融派生商品取引」中の「金融派生商品」を照応していると考える。また、固有表現中の語は、照応現象に関連する語として扱わないため、次の例における固有表現「神戸製鋼」中に含まれる「神戸」と後続する「神戸」は照応しているとは考えない。

(26) **神戸**製鋼は第 11 節で、地元**神戸**でワールドと対戦する。(神戸製鋼=“組織名”)

橋渡し指示

橋渡し指示とは直接照応していないものの、間接的に先行詞を照応する現象として捉える。基本的に、文章中に出現する名詞句の全体、またはその部分列が必須要素を持っており、その必須要素が文章中に出現している場合、それらの間には橋渡し指示の関係があるとする。ただし、その名詞句の必須要素にあたる表現が先行文脈中に出現している場合でも、その名詞句が直接照応する表現が先行文中に出現しているならば、橋渡し指示をしているとは考えない。

従って、橋渡し指示の照応詞となりうる要素としては、前方の表現を直接照応していない複合名詞、および複合名詞の部分形態素列、先行詞としてはすべての複合名詞、および複合名詞の部分形態素列を考える。

第 3章 :　知識の自動獲得

本章では、照応解析に必要な知識について考察し、直接照応の解析を行う際に必要となる同義表現、および、橋渡し指示の解析を行う際に必要となる名詞格フレーム辞書の自動獲得を行う。

3.1　同義表現

3.1.1　同義表現情報の必要性

直接照応における照応詞と先行詞の関係としては、互いがまったく同一の表現である場合や、照応詞が先行詞の主辞となっている場合、照応詞が先行詞の言い換え表現になっている場合、照応詞が代名詞となっている場合などが考えられる。このうち言い換え表現となっている場合、表層的に異なる 2 つの表現が同じものを指すという知識がないと対象の 2 つの言語表現が同一の内容ないし対象を指していることを認識するのは極めて困難である。

このことは、人間が文章を理解する場合にもあてはまり、そのため、一般的でない言い換えを用いる場合、新聞記事などでは、最初に次の例のように異なる表記が括弧などを用いて併記されるなど、同義表現についての情報が与えられる場合が多い。

(27) 鈴木勝也日朝国交正常化交渉担当兼**朝鮮半島エネルギー開発機構**(KEDO) 大使の勇退を認めた。
(28) 現在のところ、金融システムの機器について焦点となっている**長銀**(**日本長期信用銀行**) に関しては、… 。

しかし、文章中に出現する同義表現がすべて、説明されて使用されるわけではなく、「ロシア」と「露」のように極めて日常的に使われる表現の場合はこのような情報なしに異なる表現が使用されることが多い。また、括弧等を用いて文章中にそれらの表現が同義表現であるという情報が記されている場合も、計算機にとっては、それらの表現が同義表現であることを認識するのは容易ではない。このため、高精度な直接照応解析の実現のためには、このような同義表現に関する知識を事前に獲得しておく必要がある。

3.1.2　同義表現抽出に用いるリソース

同義表現を獲得するためのリソースとしては、コーパスや辞書が考えられる。

コーパスを用いた同義表現に関する研究としては、括弧表現を用いる手法、テキストの局所的な文脈依存性を利用する手法 [24] や、コーパスから名詞と略語をその出現頻度

に関するルールを用いて獲得する手法 [26]、係り受けおよび共起関係を利用し同義表現を抽出する手法 [27]、複数の著者の表記の違いを利用した手法 [30] などが提案されている。本研究ではこのうち比較的高い精度を実現できると考えられる括弧表現を用いた手法を用いる。

括弧の解析に関する先行研究としては久光ら [20] の研究がある。久光らは統計量とルールをを組み合わせて括弧表現を、同義表現や、読みを表している場合、補足している場合などに分類し、同義表現などの有用な情報の抽出を行っている。しかし、この研究で用いられたルールは同義表現の抽出に特化してはいないためその精度は十分に高いとは言えず、また、抽出に用いたコーパスの規模が新聞 1 年分と大きくないため抽出された同義表現の数は多くない。しかし、この問題はタスクを同義表現抽出に特化し、使用するコーパスの規模を大きくすることにより改善できると考えられる。

括弧表現から獲得できる同義表現の特徴としては、常識となっていない事柄、すなわち新語や聞き慣れない語も含むため新語、未知語への対応力頻度などは強いと考えられる一方、次の例文における「日」と「日本」のように極めて常識的な言い換えについては、括弧表現を用いるだけでは抽出できないと考えられる。

(29) 在**日**外国人への所得課税を優遇する要件を厳しくし、主に 日本 で働く外国人には国内外のすべての所得に課税できるようにする。

そこで、基本的な地名や略語に関する同義表現を獲得するため、辞書を用いた同義表現抽出も合わせて行う。同義表現の知識源として辞書を用いた場合、新語に対応できないという問題点があるが、極めて常識的な言い換え表現を獲得することができると考えられる。

3.1.3 コーパス中の括弧を用いた同義表現抽出

まず、括弧表現を用いた同義表現抽出について説明する。本研究では大規模コーパスを用いることにより、高精度で大規模な同義表現知識の獲得を目指す。まず、同義表現の分類を行い、続いて抽出の手順を示す。

同義表現の分類

抽出する同義表現のタイプを以下の 4 種類に分類する。

type1 英字と英字以外のペア

ex. 朝鮮半島エネルギー開発機構 (KEDO)
ASEAN(東南アジア諸国連合)

type2 カタカナとカタカナ以外のペア

ex. 主要国首脳会議 (サミット)

アパルトヘイト (人種隔離政策)

type3 共に漢字表記で一方の構成漢字がもう一方に全て含まれているペア

ex. 日本長期信用銀行 (長銀)
住専 (住宅金融専門会社)

type4 type1〜type3 以外

ex. 朝鮮民主主義人民共和国 (北朝鮮)
少額貯蓄非課税制度 (マル優)

同義表現の自動獲得の手順

コーパス中の括弧で記された情報を用いた同義表現の自動獲得の手順は以下の通りである。

1. 同義表現候補の抽出

括弧の中の表現 A と、その前に出現した句読点から括弧の前までの表現 B のペアをコーパスから取り出す。例えば、(27) のような文があった場合は A として「KEDO」、B として「鈴木勝也日朝国交正常化交渉担当兼朝鮮半島エネルギー開発機構」が、(28) のような文があった場合は A として「日本長期信用銀行」が B として「金融システムの機器について焦点となっている長銀」が取り出される。

2.type1〜type3 に該当する同義表現を抽出

1. B の形態素解析を行い末尾の名詞句 B' を取り出し、A と B' のペアが type1〜type3 に該当する場合、同義表現候補として、その出現頻度を数える。ただし、A と B' のいずれかが、人名のみ、または地名のみ[1]で構成されている場合は候補としない。これは、人名とその所属組織、地名とそこに位置する組織名などの組み合わせを除くためである。例えば (28) のような文があった場合は「日本長期信用銀行」と「長銀」のペアが抽出される。
2. コーパス全体における A と B' のペアの出現頻度に対して設定した閾値を越えるペアを同義表現とする。B'(A)、A(B') というよう互いを入れ替えた表現が存在する (以下では双方向性があると呼ぶ) ものについては、それぞれの出現回数の相乗平均に対して閾値を設定する。双方向性があるものは同義表現である可能性が高いと考えられることから、双方向性のあるペアに対する閾値は, 一方向性しかないペアに設定した閾値と比べ緩く設定した。実際に用いた閾値を表 3.1 に示す。

[1]形態素解析器 JUMAN の出力として与えられる人名、地名を用いる。

表 3.1: 括弧表現を用いた同義表現抽出のために設定した閾値

タイプ	双方向性	同義表現とみなす条件
type1	あり	頻度の相乗平均 > 3
	なし	頻度 > 50
type2	あり	頻度の相乗平均 > 4
	なし	頻度 > 300
type3	あり	頻度の相乗平均 > 1 & 文字長の差 > 2
	なし	–
type4	あり	頻度の相乗平均 > 30[2]
	なし	–

表 3.2: 括弧表現を用いた同義表現抽出の結果

タイプ	数	主な例
type1	1052	国連平和維持活動=PKO
		北大西洋条約機構=NATO
type2	220	関税貿易一般協定=ガット
		金融派生商品=デリバティブ
type3	241	住宅金融専門会社=住専
		動力炉・核燃料開発事業団=動燃
type4	75	朝鮮民主主義人民共和国=北朝鮮
		二酸化炭素=CO2
合計	1588	

3.type4 に該当するペアを抽出

type1〜type3 に該当しない同義表現ペア (type4) を抽出するため、A と B のペアの出現頻度を数え、設定した閾値を満足したペアを同義表現とする。文字種などによる制限をかけない場合、「エネルギー (1 人前) 」(新聞 26 年分に 2521 回出現) や「一等 (１０００万円)」(新聞 26 年分に 1157 回出現) など出現頻度は高いものの同義表現でないものが多く存在するため、双方向性のあるペアに限定し、またその閾値も type1〜type3 に比べてきつく設定する。

毎日新聞 12 年分と読売新聞 14 年分、計 26 年分、約 2,600 万文を用いて同義表現の自動獲得を行った。獲得された固有表現の種類と数を表 3.2 に示す。約 2,600 万文中に文頭に出現するものを除いて括弧は約 800 万回出現した。

[2]実際には文字長によって変動させている。

本実験では、誤ったペアを抽出しないように閾値を設定したため、誤った同義表現ペアは抽出されていない。また、大規模なコーパスを用いたためその数も十分なものであると言える。

3.1.4 辞書の定義文を用いた同義表現抽出

本研究では、括弧表現を用いるだけでは抽出できないと考えられる、(29) の例における「日」と「日本」のような極めて常識的な言い換え表現も含めた同義表現辞書を構築するために、国語辞典からの同義表現抽出も行う。

国語辞典からの同義表現抽出は各見出し語に対して以下のような規則を用いることにより抽出した。抽出に用いた規則の概要は以下の通り。

1. 対象の語の見出し語 A を取り出す[3]。
2. 対象の語の定義文を順に見ていき、「の略。」、「のこと。」で終わっている定義文である場合はその前の部分を、それ以外の定義文については句点より前の部分を取り出し B とする。
3. 取り出した B が「」で囲まれているか、または B が国語辞典に見出し語として載っている場合のみ次の処理に進む。
4. その定義文が対象の語の第一義である場合、または B が地名として国語辞典にに登録されているならば、A と B を同義表現とする。

例えば、表 3.3 に示すような見出語と定義文があった場合の処理は次のようになる。

「ソビエトれんぽう」に対しては、まず、表記として「ソビエト連邦」が取り出される。続いて定義文を順に取り出していき、条件 3 を満足するかどうかを調べると、最後の定義文「ソ連」のみが辞書の表記として含まれているので、最後の 4 の処理に進む。この場合、「ソ連」は辞書に地名として載っているので、「ソビエト連邦」と「ソ連」は同義表現であると判断される。

「ちゅうごく」に対しては、まず、表記として「中国」が取り出され、続いて定義文から「中華人民共和国」と「日本の中国地方」が取り出される。このうち辞書に載っているのは「中華人民共和国」で、第一義であるので、「中国」と「中華人民共和国」は同義表現として抽出される。

「ふけい」に対しては、まず、表記として「婦警」が取り出され、続いて、定義文から「婦人警察官」が取り出される。「婦人警察官」は辞書には登録されていないが、「」で囲まれた表現なので 4 の処理に進み、第一義であるので「婦警」と「婦人警察官」は同義表現として抽出される。

本研究では、国語辞典として「例解小学国語辞典 [34]」と「岩波国語辞典 [28]」を用いている。「例解小学国語辞典 [34]」は小学生向けの辞書で基本的な語が比較的平易な定義

[3] “表記” という項目が別にある場合はそちらを用いる。

表 3.3: 例解小学国語辞典の定義文の例

見出：ソビエトれんぽう
　表記：ソビエト連邦
　品詞：地名
　定義文:
　・一九一七年に、ロシア帝国を革命によってたおして、新しくつくられた国。
　・はじめての共産主義による政治がおこなわれたが、一九九一年に解体した。
　・ソ連。

見出：ちゅうごく
　表記：中国
　定義文:
　・中華人民共和国のこと。
　・日本の中国地方のこと。

見出：ふけい
　表記：婦警
　定義文:
　・「婦人警察官」の略。

文により記載されている。約 3 万語が記載されている。一方、「岩波国語辞典」は一般向けの辞書であり、語彙数は約 6 万語である。複数の辞書を用いているのは同じ語であっても、辞書により定義文が異なり、より多くの定義文を用いた方が設定したルールに該当する用例が増え、より多くの同義表現を抽出できるようになるためである。ただし、辞書に表記として含まれているかどうかの判断には「例解小学国語辞典」のみを用いている。

表 3.4 に自動抽出された同義表現の例を示す。抽出された同義表現は 150 個であった。掲載語彙数に対して少ないと言えるが、これは曖昧性のない同義表現の抽出を目的とし、厳しい絞り込みを行っているためである。このため、誤った同義表現のペアは含まれておらず、括弧表現から抽出した同義表現ペアと重複しているのは「国連」と「国際連合」、「北朝鮮」と「朝鮮民主主義人民共和国」など 6 つのみであり、「高校」と「高等学校」、「米国」と「アメリカ」など括弧表現から抽出することができない極めて常識的な同義表現の抽出に成功していると言える。

表 3.4: 国語辞典を用いた同義語抽出

定義文のタイプ	主な例 表記 (見出語)	定義文中の語
「〜の略。」	婦警	婦人警官
	高校	高等学校
	原爆	原子爆弾
	日	日本
「〜のこと。」	中国	中華人民共和国
	アメリカ	アメリカ合衆国
	都	東京都
	加算	足し算
「〜。」	ソビエト連邦	ソ連
	アメリカ合衆国	アメリカ
	アメリカ合衆国	米国
	アルミ	アルミニウム

3.2 名詞格フレーム辞書

3.2.1 名詞格フレーム辞書とは

橋渡し指示における関係、すなわち (14) における「チケット」と「値段」や、(15) における「小屋」と「屋根」の間の関係は、主に後者の名詞の性質に起因していると考えられる。「値段」とは「ある商品の値段」であり、「値段」には「何という商品の値段であるか」という情報が必要となる。同様に「屋根」には、「何という建物の屋根であるか」という情報が必要となる。このように多くの名詞にはその名詞にとって必須的な要素が存在しており、橋渡し指示の解析には、他の表現を橋渡し的に指示している名詞がどのような要素を必須としているかに関する知識が必要となる。本研究ではこれらの情報を記述した計算機用の辞書を名詞格フレーム辞書と呼ぶことにする。

3.2.2 名詞格フレーム辞書に関する先行研究

名詞格フレーム辞書は数万語あるいは数十万語という規模が必要となることから人手で作成することはほとんど不可能である。このためコーパスや既存の辞書などから自動構築する必要がある。

動詞・形容詞など用言の格フレーム辞書については、自動構築する手法がすでに提案されている [1, 5]。これらの手法ではコーパス中に表層的に明示された格情報とその用例を用いて格フレームを自動構築している。名詞でこれらの手法を応用する場合、ノ格として対象の名詞に係る表現の用例を用いることが考えられる。しかし、用言の場合はガ

格やヲ格といった格の種類によって対象の語との関係がある程度制限されるのに対し、名詞の場合、ノ格により 2 つの名詞が関係を持つことが明示されていても、それらの持つ関係は様々である。このため、名詞格フレーム辞書を構築する場合には名詞句の関係の解析を行うことが必要となる。

英文における橋渡し指示に関連した研究としては、人手で作成したルールを使用したものや、WordNet のような人手で作成した辞書的知識を用いたものがある [3, 16, 14]。Poesio らは"B of A" という名詞句から辞書的知識を獲得する手法を提案している。対象としているのは (15) の例における「小屋」と「屋根」のような全体-部分関係のみであり、(14) における「チケット」と「値段」の関係は解析できない [9]。

日本語に関しては、村田ら [7] が名詞句「A の B」の用例を名詞格フレーム辞書として用いて、橋渡し指示の解析を行う手法を提案している。しかし、基本的に全ての「A の B」の用例を名詞格フレーム辞書として使用しており、名詞句「A の B」の解析を高精度に行うことにより解析の精度が上がる可能性に言及している。

3.2.3 名詞句「A の B」の意味解析

本研究では、コーパスから名詞句「A の B」、複合名詞「AB」を収集し、名詞句の意味解析を国語辞典の定義文を利用した手法 [6] を用いて行い、その結果から名詞格フレーム辞書を自動構築した。本節では、国語辞典の定義文を利用した名詞句「A の B」の意味解析の概要を示す

名詞句「A の B」における表層的に同じ接続助詞「の」で結ばれる二つの名詞は、所有、目的、道具などの様々な意味的関係を持ち得る。黒橋らは国語辞典に出現する必須的な要素に注目し、さらに一部の語についてはシソーラスを用いることで、「A の B」の意味解析が高精度に行えることを示した。以下ではその手法を簡単に示す。

国語辞典の定義文には、その語に必須的な要素が含まれていることが多い。例えば、『例解小学国語辞典』[34] における「値段」の定義文は次のようになっており、「品物」という必須要素が含まれている。

【値段】品物を売り買いするときの金額。

このような定義文に含まれる必須要素との対応付けを行うことにより、例えば「チケットの値段」という名詞句は「チケットを売り買いするときの金額」という意味であると解析できる。

【コピー】① 書類など、もとのものと同じものを写しとること
② 広告の文案。

また、「コピー」のように多義性のある名詞の場合、「資料のコピー」という名詞句では「資料」と定義文中の「書類」を、「CMのコピー」という名詞句では「CM」と定義文中の「広告」を対応付けることにより多義性を解消することが可能となる。

解析手順を以下に示す。

表 3.5: 意味属性に関して設定したルールの例

A:《人》, B:《人間＜親族関係＞》 → <必須格 (親族)>	e.g. 私の姉
A:《人》, B:《人間＜対人関係＞》 → <必須格 (対人)>	e.g. 父の上司
A:《＊》, B:《度量衡》 → <必須格 (属性)>	e.g. 子供の年齢
A:《＊》, B:《場》 → <必須格 (位置)>	e.g. 家の中
A:《サ変》or《性質》or《状態》or《形状》, B:《＊》 → <修飾>	e.g. 公式の会談
A:《資材》, B:《具体》 → <修飾>	e.g. 木の机
A:《人》, B:《人》 → <修飾>	e.g. 男性の客
A:《人》, B:《動作》 → <主体>	e.g. 医師の薦め
A:《時間》, B:《＊》 → <時間>	e.g. 夏の大会
A:《組織》, B:《主体》 → <所属>	e.g. 高校の教師
A:《場所》or《家屋 (部分＜場＞)》, B:《具体》 → <場所>	e.g. 庭のベンチ
A:《主体》, B:《＊》 → <所有>	e.g. 自分の能力

1. Bの定義文から、Aと対応付ける候補語としてBの上位語以外の名詞Wをとりだす。基本的に定義文の末尾の語を上位語として扱う。

2. シソーラス [17] を用いてWとAとの類似度を計算し、その値をWとAとの対応付けのスコアとする[4] 。

3. 以上の処理により最大のスコアをとるWにAを関連付け、そのスコアを国語辞典を用いた意味解析のスコアとする。

国語辞典中に適当な定義文のない語や、国語辞典に載っていない語が存在する。例えば、『例解小学国語辞典』における「姉」の定義文は次のようになっており、この定義文だけを用いて「彼の姉」における「彼」と「姉」の関係を解析するのは困難である。

【姉】年上の、女のきょうだい。

このような語に対しても高精度な解析を行えるようにするため、国語辞典を用いた解析に加えて、シソーラスの意味属性に関してルールを設定しルールに基づいた意味解析を行う。意味解析の結果として与える関係として、< 必須格 (親族)>、<必須格 (対人)>、<必須格 (位置)>、<必須格 (属性)>、<所属>、<所有>、<主体>、<場所>、<時間>、<修飾>を設定する。このようなルールの例を表 3.5 に示す。このルールに基づいた解析によって、例えば「彼の姉」には<必須格（親族）>、「赤色の帽子」には<修飾>の関係を与えることができる。

多くの場合、国語辞典を用いた解析も意味属性を用いた解析も解析結果を出力する。このため、国語辞典を用いた解析と意味属性を用いた解析、両方の解析で解析結果が得

[4] 2 語の類似度は基本的にシソーラスの木構造での 2 語の近さから算出する。XとYの類似度を計算する場合、木構造におけるXの深さ d_X、Yの深さ d_Y、XとYの共通の親の深さ d_S を用いて、$(d_S \times 2) / (d_X + d_Y)$ を類似度として用いる。

られる場合がある。このような場合、<必須格（親族）>という関係が得られたものや、国語辞典を用いた解析で高いスコアが得られたものなど、より信頼性が高いと考えられる解析結果を使用するまた、いずれの解析においても適当な解析結果が得られない場合は、名詞 A と名詞 B の間に関係がないものと判断し、解析結果なしとする。

3.2.4 名詞格フレーム辞書の自動構築

名詞格フレーム辞書とは、名詞がどのような必須格をとるのか、また各必須格の用例にはどのようなものがあるのかを記した計算機用の辞書である。このため、名詞格フレーム辞書を作るために使用できる情報源としては、国語辞典の定義文とコーパス中に出現する名詞句「A の B」などが考えられる。しかし、国語辞典の定義文中にはその語の必須要素が含まれていることが多いものの、国語辞典を使用するだけでは定義文中のどの名詞が必須要素であるか判断することは困難であり、また、全ての必須要素が定義文中に含まれているとも限らない。一方、名詞の必須格の多くは対象の名詞にノ格でかかることから名詞句「A の B」の用例を使用することが考えられるが、名詞句「A の B」を使用するだけでは、どのような要素が必須要素であるか判断することが困難であり、多義性に対応することもできない。そこで、これらの 2 つの情報源を統合的に使用することにより名詞格フレーム辞書の自動構築を行った [12]。

また、本研究では名詞句「A の B」に加えて補助的に複合名詞「AB」も使用する。これは、必須要素が多くが複合名詞の形で出現するような名詞も存在すると考えるためである。以下に名詞格フレーム辞書の自動構築の概要を示す。

名詞句「A の B」、複合名詞「AB」の収集・解析

コーパスから名詞 B にノ格で係る名詞 A の用例、および複合名詞「AB」の形で B に係る名詞 A の用例を収集する。名詞 B の格フレームの用例となる名詞 A を集めるのが目的である。したがって、「A の B」の形をしていても、「A の B の C」や「A の BC」のように A が B に係っているのが明らかでないものは収集せず、「A の B が」や「A の B を」のように B がガ格やヲ格であるものなど、A が B に係っている可能性が高いもののみ収集する[5]。また、複合名詞「AB」の場合も同様に収集する。

次に、収集された「A の B」に含まれる各 B について集まった A の用例を先に述べた方法で解析し (国語辞典として『例解小学国語辞典』を用いる)、解析結果ごとにまとめる。解析結果のまとめ方は次のとおりである。複合名詞「AB」の場合は「A の B」という形で出現しているもののみを使用し、「A の B」と記載されているものとして解析する。ただし、「A の B」よりも必須性を表している度合いは小さいと考え、用例の出現回数としては $\alpha(< 1)$ 倍して扱っている。

- 定義文中の語に関係付けられた用例は定義文中の語ごとにまとめる。ただし、定義

[5]「A の B を」という形であっても、「監督の優勝を決めた瞬間の表情」のように A が B に係らない場合もある。しかし、このような用例は稀である。

表 3.6: 「Aのひさし」の用例の収集・解析結果

国語辞典を用いた解析の結果	
1. 家の出入り口や窓などの上にさし出た小さな屋根。	
〔家〕11	玄関 2, ビル 1, 家屋 1, 建物 1,…
〔窓〕8	屋根 2, 窓 2, 入り口 1, 天井 1,…
2. ぼうしの前についているつば。	
〔ぼうし〕13	帽子 8, 兜 1, フード 1, よけ 1,…
意味属性を用いた解析の結果	
<場所>18	駐車場 3, 店舗 3, コーナー 2, 店 2,…
<修飾>10	コンクリート 1, 下 1, 金属製 1, 銀色 1,…
意味解析結果のないもの	
<その他>14	部分 3, 式 1, 信号機 1, 電話 1,…

コーパス中の「ひさし」の出現回数:292 回

文中の並列語に集まった用例はひとつにまとめる。

- 意味属性による解析で関係が得られたものはその関係でまとめる。
- 意味解析結果の得られなかったものは<その他>としてまとめる。

例えばBが「ひさし」である場合、Aの用例は表 3.6 のようにまとめられる (表中の格スロットに続く数字はその格スロットに分類される用例の出現回数の合計、用例に続く数字はその用例の出現回数を表している)。以下ではこのようなまとまりのそれぞれを格スロットと呼ぶことにし、国語辞典に関連付けられた格スロットを “〔...〕” で、それ以外の格スロットを “<...>” で表記する。

格スロットのクラスタリング

同じ必須要素を表す格スロットが重複して作成される場合がある。例えば表 3.6 の「ひさし」の場合、〔家〕と〔窓〕、<場所>はほぼ同じ内容を表している。このような格スロットをまとめるため、類似度が 0.5 以上である格スロットを統合する。格スロットの類似度は各用例間の類似度の上位 1/4 の値の平均値として計算する。ただし、異なる定義文に関係付けられた格スロット同士は統合の対象としない。また、統合する格スロットの中に定義文に関連付けられた格スロットが含まれている場合は、以降では統合された格スロットを定義文に関連付けられた格スロットとして扱う。

「ひさし」の場合、まず〔家〕と〔窓〕の類似度が 0.80 と計算されるので、これらの格スロットは統合され〔家・窓〕となる。次に、〔家・窓〕と<場所>の類似度も 0.67 と計算されるので、これらの格スロットも統合される。〔ぼうし〕と〔家・窓〕の類似度も 0.58 と計算されるが、これらは異なる定義文に関連付けられていることから統合されない。

格スロットの選択

用例の解析結果をまとめた格スロットには必須的でない要素も含まれている。このため、どの格スロットが必須的か判断し、格フレーム辞書に含める格スロットを絞りこむことが必要となる。

必須的な格スロットであればその用例も多いと考えられる。また、定義文中に含まれる語に関連付けれられた格スロットなどは必須要素である可能性が高く、意味属性を用いた解析で<修飾>と解析された格スロットは必須要素にまずならないなど、格スロットの種類によって必須要素である可能性が異なると考えられる。そこで、格スロットの種類ごとに必須格とみなす閾値を設定し、閾値を満足した格スロットのみを残す。閾値は 50 個の名詞の用例の収集結果を参考に設定した。設定した閾値を表 3.7 に示す。

「ひさし」の場合、「ひさし」の出現回数は 292 回であるので、閾値を満足する格スロットは用例の出現回数が 19 回の〔家・窓〕と用例の出現回数が 13 回の〔ぼうし〕の 2 つだけであり、これら 2 つの格スロットのみが残される。

表 3.7: 必須要素と判断する閾値

格スロットの種類	用例の出現頻度
定義文に関連付け	1/240
<必須格 (親族・対人・属性・位置)>	1/36
<所有>・<所属>	1/48
<主体>・<場所>	1/24
<その他>	1/12
<修飾>・<時間>	(使用しない)

出現頻度:対象の名詞の出現回数に対する用例の出現回数の割合

格フレームの構築

格フレームとは、ある語のとり得る格の制約を記述したものであり語義ごとに必要となる。このため、同一の語義に対する格スロットはひとつの格フレームの異なる格スロットとして扱い、異なる語義に対応する格スロットは別々の格フレームとして扱う。ここで問題となるのは、どの格スロットがどの語義に対応していると考えるかである。定義文に関連付けられた格スロットは語義との対応関係が明らかなので、意味属性による解析で得られた格スロットと<その他>としてまとめられた格スロットをどう扱うかが問題となる。

まず、意味属性による解析で<必須格>としてまとめられた格スロットについて考える。このような格スロットは国語辞典に適切な定義文がないために構築されたものと考えられることから、定義文に関連付けられた格スロットとは異なる語義に対応していると考える。また、<その他> としてまとめられた格スロットについても、<その他>としてまとめられた格スロットが残っている場合はその出現頻度は 1/12 以上と高頻度であることか

ら (表 3.7 参照)、<必須格>としてまとめられた格スロットと同様に、定義文に関連付けられた格スロットとは異なる語義に対応していると考える[6]。

一方、意味属性を用いた解析結果によって<所有>、<場所>、<所属>、<主体>としてまとめられた格スロットに対しては、これらの格スロットが残るのはこれらの要素が国語辞典の定義文中にあまり記述されない要素でためであると捉え、他の必須要素とも共存できる要素、すなわち、他の格スロットと同じ語義に関連付けられる格スロットであると考える。

このような考えに基づく格フレームの構築アルゴリズムを以下に示す。

- 定義文に関連付けられた格スロットがある場合は定義文ごとに 1 つの格フレームを構築する。
- 意味属性による解析によって<必須格>としてまとめられた格スロットおよび<その他>としてまとめられた格スロットが統合されず残っている場合は、それぞれに対して新たな格フレームを構築する。
- 意味属性を用いた解析結果によって<所有>や<場所>、<所属>、<主体>としてまとめられた格スロットが残っている場合は、既に存在している格フレームに付け加える。ただし他の格スロットが存在しない場合は新しい格フレームを作成する。

この方法で自動構築された名詞格フレームの例を表 3.8 に示す。「ひさし」の場合、〔家・窓〕と〔ぼうし〕の 2 つの格スロットは異なる定義文に関連付けられた格スロットがあるので各定義文ごとに異なる格フレームが構築される。一方「表情」の場合は、同一の定義文に関連付けられた複数の格スロットであるのでこれらの格スロットは同一の格フレームの別の格スロットとして扱われる。「引き出し」の場合、定義文に関連付けられた格スロットの他に、<その他>としてまとめられた格スロットが存在しているので、これらは異なる語義に対する格スロットであると判断され 2 つの格フレームが構築される。「コーチ」の場合は定義文に関連付けられた格スロットの他に存在するのが<所属>という格スロットであるので多義性はないと判断され 1 つの格フレームにまとめられる。

3.2.5 自動構築された名詞格フレーム辞書の規模

毎日新聞 12 年分および読売新聞 14 年分の約 2,600 万文を用いて名詞格フレーム辞書の自動構築を行ったところ、約 1 千万個の「A の B」から、約 18,000 語の名詞について格フレームが構築された。「A の B」の形で出現する名詞 B は約 2 万 9 千種類であった。1 語あたりの格フレーム数の平均は 1.07 個、1 つの格フレームに含まれる格スロット数の平均は 1.21 個であった。

[6] もし、定義文に関連付けられた格スロットと同じ語義に対応しているとすると、これだけの高頻度で出現する極めて必須的な要素が定義文に記述されていないことになるが、このような可能性は低いと考えられる。

表 3.8: 名詞格フレームの例

	格スロット	用例
ひさし (1)	(家の出入り口や窓などの上にさし出た小さな屋根。)	
	〔家・窓〕	駐車場 3, 店舗 3, 玄関 2, 屋根 2, 窓 2,…
ひさし (2)	(ぼうしの前についているつば。)	
	〔ぼうし〕	帽子 8, 兜 1, フード 1, よけ 1,…
表情 (1)	(自分の気持ちを顔や身ぶりにあらわすこと。)	
	〔自分〕	<人>人々137, 人 80, 市民 61, 相手 49, 首相 45,…
	〔気持ち〕	<動作>安ど 599, 余裕 397, 困惑 247, 苦渋 245,…
引き出し (1)	(机・たんすなどにある、引いて出し入れができる箱。)	
	〔机・たんす〕	机 216, たんす 36, タンス 33, 鏡台 6, デスク 4…
引き出し (2)	<その他>	預金 250, 資金 31, 配当金 21, 貯金 13, 資産 13,…
コーチ (1)	(スポーツなどでそのやり方などを教えること。)	
	〔スポーツ〕	野球 12, サッカー 9, 水泳 6, スケート 6, 体操 4,…
	<所属>	<組織>チーム 97, 部 36, クラブ 10, 母校 10,…
株式 (1)	(株式会社の資本構成単位。)	
	〔会社〕	<組織>企業 1078, 会社 628, 銀行 169, 子会社 142,…

WEB から収集した文章、約 5 億文を用いて名詞格フレーム辞書の自動構築を行ったところ、約 1 億 8 千万個の「A の B」、および約 4 億 5 千万個の複合名詞「AB」から、約 41,500 語の名詞について格フレームが構築された。「A の B」の形で出現する名詞 B は約 47 万種類であった。1 語あたりの格フレーム数の平均は 1.04 個、1 つの格フレームに含まれる格スロット数の平均は 1.10 個であった。

第 4章 :　固有表現認識

固有表現 (Named Entity) 認識とは、情報検索、情報抽出の基礎技術として、テキスト中から組織名、人名、地名等の自動的な認識を行う技術であり、高精度な照応解析を行う際に必要となる技術である。本章では固有表現認識について述べる。

4.1　固有表現認識とは

固有表現認識とは、与えられた文書から人名・地名・組織名といった予め決められたタイプの表現を認識する問題である。固有表現認識は、次章で述べる直接照応において照応詞であるかどうかの判定に使用できたり、ある表現が例えば “組織” を表していることを認識することにより、橋渡し指示の対象であると判断することが可能となったりするなど、高精度な照応解析を行う上で非常に重要な技術であると言える。

認識する表現のタイプとしては人名、地名、組織名などの他に時間や割合などを対象とする場合もあり、また入れ子構造をどのように扱うかなど様々な設定を考えることができる。同一の文章に対して、どのようなタイプの表現の認識を行うかという問題設定より求めるべき固有表現は異なってくる。このためワークショップなどで定義された基準に従って認識を行い、評価される場合が多い。ヨーロッパ系言語を対象としたワークショップとしては、Message Understanding Conference(MUC-6) や、CoNLL-2002(Coreference on Natural Language Learning 2002)、CoNLL02003 などが、日本語を対象としたものとしては IREX(Information Retrieval and Extraction Exercise) などが行われている。MUC-6 および、IREX において定義された固有表現を表 4.1、表 4.2 にそれぞれ示す。

表 4.1:　MUC6 で定義された固有表現

固有表現の種類
ORGANIZATION
PERSON
LOCATION
DATE
TIME
MONEY
PERCENT

表 4.2: IREX で定義された固有表現

	固有表現の種類	
固有名詞的表現	組織名	ORGANIZATION
	人名	PERSON
	地名	LOCATION
	人工物名	ARTIFACT
時間表現	日付	DATE
	時刻	TIME
数値表現	金額	MONEY
	割合	PERCENT

4.2 固有表現認識に関する先行研究

4.2.1 先行研究の手法

日本語における固有表現認識は入力文を適当な解析単位 (トークン) に分割し、各トークンが固有表現中であるか、固有表現中である場合はどの種類の固有表現のどの位置にあたるのかを判定し、これらをまとめあげる手法が一般的である。トークンの単位としては、形態素を単位とする手法や、文字を単位とする手法が提案されている。形態素を単位とする手法は、文字を単位とする手法と比べて、トークンの単位が大きいため、広範囲の情報を用いられるという利点がある。しかし、形態素中に固有表現の区切りが来るような固有表現をそのまま扱うことができないという問題がある。これらの手法は、表 4.3 に示すように現在のところほぼ同等の精度が得られている。

ある表現が固有表現であるかの判定に用いられる手法としては大きく分けて、人手により規則を作成しそれらのルールに基づいて固有表現を認識する手法と、固有表現がタグ付けされたテキストから、機械学習を用いて認識規則を学習する手法の 2 つがある。

人手により規則を作成しそれらのルールに基づいて固有表現を認識する手法ではある程度の精度は比較的簡単に実現できることが期待できる反面、高精度化のためにはどのようなルールを加えれば良いのかなどについて高度な考察が必要となり、多大なコストを要するという欠点がある。IREX の NE タスクにおいて最も良いスコアを達成したシステムは、人手による規則ベースの手法であったが、89,000 個の固有名詞が登録され、また規則を作成するのに 5,000 文が用いられている。このシステムは自由トピックの記事 72 記事に対して、83.86 の F 値を達成した。

これに対して、固有表現がタグ付けされたテキストから、機械学習を用いて抽出規則を学習する手法では、固有表現がタグ付けされたテキストが必要となるものの、固有表現がタグ付けされたテキストが増えるに従って、精度が上昇していくと考えられ、認識規則の生成に関する複雑な考察を必要としない利点がある。近年は、毎日新聞 95 年度版 1,174 記事、約 11,000 文に対して IREX で定義された固有表現のタグが付与された CRL

表 4.3: 機械学習を用いた固有表現認識の先行研究

	CRL データ交叉検定	IREX テストデータ	学習手法	形態素解析器	解析単位	シソーラス
内元 2000[36]		80.17	ME	JUMAN	形態素	無
颯々野 2000[37]		82.8	ME	BREAKFAST	形態素	無
山田 2002[23]	83,2		SVM	ChaSen	形態素	無
磯崎 2003[18]	86.77	85.11	SVM	ChaSen	形態素	無
浅原 2004[29]	86.35		SVM	ChaSen	文字	無
	87.23		SVM	ChaSen	文字	有
中野 2003[33]	88.72		SVM	ChaSen	文字	無
	89.03		SVM	ChaSen	文字	有

固有表現データが公開されていることもあり、主に機械学習を用いた手法が研究されている。表 4.3 に機械学習を用いた固有表現認識の先行研究の精度を示す。精度はすべて F 値で表示してあり、CRL データ交叉検定とは CRL データの 5 分割交叉検定の精度、IREX テストデータとは IREX のテストデータのうち自由トピックの記事 72 記事を解析した場合の精度である。

4.2.2 機械学習を用いた固有表現認識

表 4.3 が示すように、機械学習を用いた手法としては近年、サポートベクタマシン (SVM[15]) を用いた手法が提案されており、高い精度を実現している。SVM を用いた固有表現認識システムでは、固有表現認識を形態素 (または文字) の分類問題として捉える。ひとつの固有表現は 2 つ以上の形態素からなることがあるので、対象の形態素が固有表現のどの位置にあたるかに応じて分類する必要がある。

磯崎 [18] らは各形態素を IREX 固有表現の 8 種類に対して先頭、途中、末尾、単独の 4 種類に分け、固有表現以外を加えた合計 $4 \times 8+1=33$ 種類に分類している。「組織名の先頭」の次には「組織名の途中」か「組織名の末尾」であるなどという制約が存在するので、これらの制約を満たすべての組合せの中で、Viterbi アルゴリズムを用いて、文全体でベストになる組合わせを選んでいる。Viterbi アルゴリズムを使用する際に必要となるコストとしては、SVM の出力に次のようなシグモイド関数を適用することにより、確率類似量を計算し用いている。

$$s(x) = \frac{1}{1 + \exp(-\beta x)} \tag{4.1}$$

ここで、固有表現認識では 33 クラスへの分類が必要である。SVM は 2 クラスの分類しかできないため、SVM を複数組み合わせて使用する必要があるが、磯崎らは、あるクラスに属すか、属さないかの 2 値分類を行う one-versus-rest を採用している。また、分類

の際の素性としては、分類対象の形態素とその前後 2 つずつの計 5 形態素に対し、出現形、ChaSen[35] による品詞、形態素を構成する文字種という 3 種類の素性を用いている。

山田ら [23] は固有表現の位置として、Chunk 同定問題で使用される IOB1 や IOB2、IOE1、IOE2 などといった表記を利用している。ここで、IOB1 は固有表現である単語に対して I というタグを付与し、同種類で別の固有表現が連続した場合は、後続する固有表現の開始単語に B というタグを付与する。固有表現以外の単語には O というタグを付与する。IOB2 は開始位置に付与するタグが IOB1 と異なり、固有表現の開始単語に必ず B というタグを付与する。また、IOE1 と IOE2 では終了位置に注目し、E というタグを付与する。これらのタグを IREX 固有表現の 8 種類に対し適用し、合計 $2 \times 8+1=17$ 種類に分類している。SVM の解析結果は文頭、または文末から決定的に適用している。形態素解析器としては ChaSen を利用している。

浅原ら [29] は解析単位として文字を採用した上で、やはり IOB などのチャンクタグを用いて固有表現認識を行なっている。また、シソーラスの使用によって精度が向上することも示している。中野ら [33] はさらに文節区切りを行い、文節内外の素性を文節素性として用いることにより、F 値約 89 という結果を得ている。

4.3 固有表現認識システムの構築

4.3.1 固有表現認識の方針

本研究では、解析単位として形態素を用いて固有表現認識システムの構築を行う。解析単位として文字ではなく形態素を用いたのは、解析単位が異なる以外はほぼ同様の手法を用いている磯崎らの手法と、シソーラスを用いなかった場合の浅原らの手法では、僅かではあるが解析単位として形態素を採用した磯崎らの方が良く、また、本研究で用いる形態素解析器 JUMAN[22] では、形態素の単位としてより細かい単位をとるようになっており、実際に形態素中に固有表現区切りがくるような固有表現はあまり多くなく、また特定のものが多数を占めているためである。

実際に、CRL 固有表現データで付与されている固有表現 20,258 個のうち、その境界が形態素中であるものは 242 個のみである。このうち、現在ひとつの形態素となっている「自社」、「日中」、「米朝」を切り、「華」、「半」、「倍」、「英」、「憲」、「京」、「邦」、「仏」、「長」、「員」を独立した形態素と考えることにより解決するものが 2/3 近い 160 個を占める。また、この残りの 82 個についても、意味のある最小の区切りとしての形態素解析の単位を変更してまで扱うメリットのある固有表現であるとは限らない。

(30) 代表質問では 政権準備委 員長の海部氏が「施政方針演説」を行うほか、「財政演説」、「外交演説」と位置付ける質問も行う方針。

例えば、この例文において「政策準備委員長」は形態素解析の結果「政策」、「準備」、「委員」、「長」と分けられるのに対し、「政策準備委」に組織であるというタグがついているが、「委」と「員」の間に強い意味の区切りがあるとは考えにくい。

以上の理由から本研究では、解析単位として形態素を採用し、形態素を解析単位とする先行研究の中で最も良い精度を実現している磯崎らの手法に基づいてシステムの構築を行う。また、文節素性という素性を用いることによって先行研究で最も良い精度を実現している中野らの手法に倣い、文節区切りの解析も行い、そこから得られる情報も使用する。

まず、磯崎らと同様に形態素を 33 種類に分類し Viterbi アルゴリズムを用いて最終的なタグを決定する。形態素を分類するための素性としては、磯崎らが用いている前後 2 形態素までの、形態素表記、品詞 (細分類も含む)、文字種に加え、本研究独自の素性として、固有表現の末尾に関する情報、対象の形態素の以前の文での解析結果を用いた。

固有表現の末尾に関する情報とは、形態素解析器 JUMAN によって与えられる固有表現の末尾に関する情報である。JUMAN による解析の結果は組織名の末尾となる場合の多い「会」や「機構」、「支店」などといった 70 余りの形態素には組織名末尾になりやすい、という情報が与えられる。また、人名の直後に来ることの多い「会長」、「監督」、「候補」などといった 200 余りの形態素にも人名の直後に来やすい、という情報が与えられている。本研究では、これらの情報がその形態素に与えられているかどうか、さらに構文解析器によって区切られた文節内にこれらの情報が与えられた形態素が存在しているかどうかを SVM で用いる素性として用いている。以下、これらの素性のことを文節素性と呼ぶ。

対象の形態素の以前の文での解析結果を素性として用いるのは、文章中に同じ形態素が出現したのならば、それらは同じ固有表現や、その一部になる可能性が高く、また、初出の個所の方が固有表現であることを認識する手がかりが多い可能性があると考えられるためである。以下、この素性をキャシュと呼ぶ。キャッシュは以前の文で固有表現の一部と判断されたときのみ使用する。

(31) 島根県出雲市で行われた第六十五期将棋棋聖戦五番勝負第三局は、六日午後七時五十二分、六十一手で**羽生**善治棋聖が挑戦者の島朗八段に勝ち、三連勝で四連覇を達成した。十二日から滋賀県彦根市の彦根プリンスホテルで行われる第四十四期王将戦七番勝負で、羽生 は谷川浩司王将に挑戦、史上初の七冠を目指す。

例えばこのような文があった場合、1 文目の羽生が「人名の先頭」と解析されたという情報を用いないと 2 文目の文頭の羽生は「場所 (単独)」と判断されてしまうが、1 文目で羽生が「人名の先頭」と解析されたという情報を用いることによって、正しく「人名 (単独)」と判断される。

(31) の文中の「、羽生は谷川浩司王将に挑戦」の形態素解析結果を表 4.4 に示す。また、このうち「は」、「谷川」を解析する際に用いる素性を表 4.5 に示す。

4.3.2 固有表現認識実験

学習には磯崎らと同じく IREX で作成された CRL 固有表現データ 1,174 記事を、解析対象として IREX のテストデータを用い、形態素解析器としては JUMAN ををを用いて固

表 4.4: 形態素解析結果の例

	··· 、	羽生	は	谷川	浩司	王将	に	挑戦 ···
品詞	特殊	名詞	助詞	名詞	名詞	名詞	助詞	名詞
細分類	読点	地名	副助詞	普通名詞 or 人名	人名	普通名詞	格助詞	サ変名詞
その他 (特徴)						人名直後 に多い		

有表現認識実験を行った。さらに、先行研究との比較を行うために CRL 固有表現データを用いて 5 分割交叉検定による実験も行った。結果を表 4.6 に示す。固有表現区切りが形態素中にある固有表現は評価から除いている。

また、どの素性がどのくらい効いているのかを調べるために、本研究で新たに追加した追加した素性であるキャッシュや文節素性を用いない場合の実験も行なった。実験には IREX テストデータを用い、磯崎らの用いた素性のみの場合、文節素性を用いない場合、キャッシュを用いない場合の実験を行い、これら 3 つの場合の固有表現認識結果と、キャッシュや文節素性などすべての素性を用いた場合の固有表現認識結果の比較を行った。結果を表 4.7 に示す。この結果から、キャッシュ、文節素性共に良く効いているのが確認できた。また、両方を同時に用いることにより、単独で用いた場合の精度向上よりも大きく向上することが判った。

ここで、磯崎らの用いている形態素列、品詞、文字種のみを用いた場合の精度が磯崎らの実験よりも悪いのは使用している形態素解析器が異なるためである。まず、JUMAN では辞書は必要最低限の大きさに留めるという考えのもと、それ以上分割できないような語しか辞書に登録されていない。このことは形態素中に固有表現の区切りが来ることが少ないという点では有利に働くが、固有表現がそのまま辞書に登録されていることが少なくなるため、少なくとも過去の記事を解析する限りにおいては不利に働いていると考えられる。JUMAN の辞書に登録されている全語彙数は約 6 万語、そのうち固有名詞として登録されているのは約 3 万語であるのに対し、ChaSen の辞書に登録されている全語彙数は約 24 万語、そのうち固有名詞として登録されているのは約 14 万語である。

さらに、JUMAN の辞書と ChaSen の辞書の違いとして細分類の細かさが挙げられる。JUMAN では固有表現に関する細分類は基本的に「組織名」、「人名」、「地名」があるだけ[1]であるが、ChaSen の辞書では「人名」がさらに「姓」と「名」と「一般」に、「地名」がさらに「国」と「一般」に分かれていたり、「固有名詞」という別の分類があるなど JUMAN より細かく分かれている。このことも JUMAN を用いた固有表現認識の方が ChaSen を用いたものよりも精度が低くなる原因であると考えられる。

構築した固有表現認識システムは、同じく形態素解析器として JUMAN を用いている内元らよりは高い精度を、より大きな辞書を用いた磯崎らとほぼ同じ精度を実現しており、また、新たに加えた素性による精度向上が確認できたことから、用いた手法はある

[1]JUMAN でも「固有名詞」という細分類も存在するが僅か 51 語しか登録されていない。

表 4.5: 「は」、「谷川」の解析に用いる素性

	「は」	「谷川」
正解	その他	人名先頭
2つ前の形態素	「、」	「羽生」
2つ前の品詞	特殊	名詞
2つ前の細分類	読点	地名
2つ前の文節素性	–	–
2つ前のキャッシュ	–	人名先頭
1つ前の形態素	「羽生」	「は」
1つ前の品詞	名詞	助詞
1つ前の細分類	地名	副助詞
1つ前の文節素性	–	–
1つ前のキャッシュ	人名先頭	–
対象の形態素	「は」	「谷川」
対象の品詞	助詞	名詞
対象の細分類	副助詞	普通名詞、人名
対象の文節素性	–	後方に人名直後に多い語
対象のキャッシュ	–	–
1つ後の形態素	「谷川」	「浩司」
1つ後の品詞	名詞	名詞
1つ後の細分類	普通名詞、人名	人名
1つ後の文節素性	後方に人名直後に多い語	後方に人名直後に多い語
1つ後のキャッシュ	–	–
2つ後の形態素	「浩司」	「王将」
2つ後の品詞	名詞	名詞
2つ後の細分類	人名	普通名詞
2つ後の文節素性	後方に人名直後に多い語	人名直後に多い
2つ後のキャッシュ	–	–

表 4.6: 固有表現認識の精度

	CRL データ交叉検定			IREX テストデータ		
固有表現の種類	再現率	適合率	F 値	再現率	適合率	F 値
組織名	75.67	85.87	80.45	69.55	83.84	76.03
人名	85.71	90.56	88.07	86.65	89.30	87.96
地名	87.22	88.73	87.97	84.83	87.44	86.12
人工物名	37.41	65.17	47.54	31.25	40.54	35.29
日付	92.22	94.78	93.48	95.38	95.38	95.38
時刻	88.78	94.06	91.34	92.59	94.34	93.46
金額	91.49	96.47	93.92	100.00	100.00	100.00
割合	91.03	96.97	93.91	100.00	100.00	100.00
総合	83.86	89.74	86.70	82.29	87.89	85.00

表 4.7: 用いる素性ごとの固有表現認識精度 (F 値)

固有表現の種類	磯崎らの素性のみ	キャッシュを使用	文節素性を使用	すべて使用
組織名	73.95	73.30	75.38	76.03
人名	82.85	84.05	86.28	87.96
地名	85.25	85.33	84.57	86.12
人工物名	32.61	35.29	31.82	35.29
日付	94.43	95.38	94.23	95.38
時刻	95.33	93.46	95.33	93.46
金額	100.00	100.00	100.00	100.00
割合	100.00	100.00	100.00	100.00
総合	82.88	83.36	83.74	85.00

表 4.8: 検出できなかった主な固有表現

認識できなかった回数	固有表現	正解	主な原因
14	長銀	組織名	接頭辞 + 普通名詞 として解析
13	インターポール	組織名	「ポール」を人名と判断
4	東大寺	組織名	場所と判断
4	商法	人工物名	全体でひとつの形態素と判断
3	共同	組織名	前後から推測できず
3	ティフアナ	場所	前後から推測できず
2	朴セリ	人名	漢字とカタカナからなる人名が稀
2	村上・亀井派	組織名	「村上」を単独で人名と判断
2	額賀福志郎	人名	「賀福志郎」を人名と判断
2	小渕	人名	接頭辞 + 普通名詞として解析

程度有効な方法であると言える。Chasen を用いた先行研究のうち、文節素性を用いた中野らのスコアを実現することはできなかったが、小さい辞書を用い、細かい単位で語を扱うことは、まったく未知の固有表現が多く出現するテキストでは有利に働くこともあると考えられる。

4.3.3 主な解析誤り

IREX テストデータを用いた実験における主な誤りを表 4.8 に示す。表中に含まれるのは、正しく認識できなかった回数 3 回以上の固有表現すべてと、2 回の固有表現の一部である。

これらのうち、上位 2 つの語が全解析誤りの 10%以上を占める。これらは ChaSen の辞書にはそれぞれの固有表現として登録されているため、ChaSen を用いた場合は解析できると考えられる。

第 5章 :　直接照応解析

本章では、自動獲得した同義表現や、固有表現認識技術を用いた直接照応解析について述べる。

5.1　直接照応解析に関係する先行研究

直接照応解析に関係する先行研究で用いられた手法としては大きく、人手で作成した規則に基づく解析手法と、タグ付きコーパスを用いた学習手法に分けられる。本節では、英語と日本語それぞれについて、規則ベースの先行研究と機械学習を用いた先行研究を簡単に紹介する。

5.1.1　規則ベースの手法

Zhou ら [2] は、英文に対して、coreference を 7 種類に分類し、照応の種類ごとに規則を作成し直接照応の解析を行っている。Zhou らの手法の概要を以下に示す。

1. 照応詞と先行詞間の性や数が一致しなければならないことや、意味的つながりから生じる一般的な制約を用いて、先行詞候補の絞り込みを行う。
2. 照応のタイプごとに作成した制約を用いて先行詞候補をさらに絞り込む。ここで照応のタイプとしては、Name alias coreference、Apposition coreference など 7 種類のタイプを定義している。
3. 残った先行詞候補から、照応詞に近いものほど先行詞になりやすいという単純なルールに基づいて先行詞を決定する。

照応のタイプとしては以下のような 7 種類に分類している。

- 先行詞と照応詞が同義表現である場合 (Name alias coreference)
- 先行詞と照応詞が同格関係にある場合 (Apposition coreference)
- 先行詞、照応詞が述語-項関係にある場合 (Predicate nominal coreference)
- 照応詞が代名詞である場合 (Pronominal coreference)
- 照応詞が定名詞句である場合 (Definite noun phrase coreference)

表 5.1: MUC-6, MUC-7 の coreference 課題の解析精度

	MUC-6			MUC-7		
	再現率	適合率	F 値	再現率	適合率	F 値
Soon ら	58.6	67.3	62.6	56.1	65.5	60.4
Ng ら	64.2	78.0	70.4	57.4	70.8	63.4
Zhou ら	65.8	84.7	73.9	55.7	82.8	66.5

- 照応詞が指示詞に修飾されている場合 (Demonstrative noun phrase coreference)
- 照応詞が裸名詞である場合 (Bare noun phrase coreference)

各段階で必要となる制約は基本的にデータから人手で作成している。Zhou らはこの手法により、MUC-6 に対して 73.9%、MUC-7 に対して 66.5%の F 値という解析結果を得ている (表 5.1)。

村田ら [31] は、日本語を対象として、名詞の指示性を考慮した 9 個のルールを用いて名詞の同一性の解析を行っている。名詞句の指示性に関しては、人手で作成した 86 個の規則を適用することにより、すべての名詞を総称名詞、定名詞、不定名詞の 3 種類に分類している。童話や新聞記事を用いた実験を行い、結果として適合率 79%、再現率 77%を得ている。童話、新聞記事それぞれの精度、および、複合名詞の部分形態素列が関係する照応をどこまで扱っているかなどは不明である。

5.1.2 機械学習を用いた手法

機械学習を用いた同一指示解析手法はいくつかの手法が提案されている。ここでは Soon ら手法 [13]、およびその改良である Ng ら [8] の手法を紹介する。

これらの手法では、共参照解析の問題を、照応詞候補に対して、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 値分類問題として扱っている。分類器は対象の名詞句が先行詞かどうかという 2 値分類問題を解く。

Soon らの手法では、訓練時には、先行詞と照応詞の対を正例、先行詞と照応詞の間の各名詞句と照応詞の対を負例として学習した。照応問題を解く際には、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を分類していく。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。

Soon らの実験では、12 個の素性を用い、C5.0[11] を使用して決定木学習を行ない、MUC-6 に対して 62.6%の F 値、MUC-7 に対して 60.4%の F 値と、規則ベースの手法と同程度の精度を得ている (表 5.1)。

Ng らの手法では Soon らの手法が 2 つの点において改良された。一つは素性集合を拡張し、語彙的な素性や意味的素性など、53 個の素性に増やした。もう一つは先行詞同定

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。Ng らのモデルは Soon らのモデルよりも先行詞同定の精度がよく、MUC-6 に対して 70.4%、MUC-7 に対して 63.4%の F 値を得ている (表 5.1)。

日本語における機械学習を用いた同一指示性解析に関する研究としては飯田ら [4] の研究がある。飯田らは日本語では冠詞などの情報が無く、名詞句の指示性の推定がそれほど容易でないことから、まず名詞の指示性の判断を行った後に先行詞の同定を行うのではなく、まずある表現に対する最尤先行詞候補を決定した後先行詞候補の情報も用いて名詞の指示性の判断を行っている。飯田らの手法の概要を以下に示す。

1. 文章の先頭から順に照応詞の候補となり得る候補を検出する。
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尤先行詞候補を決定する。
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無くなるまで 1〜4 を繰り返す。

飯田らは分類器として SVM を用い、語彙的な情報を用いた素性や統語的な情報を用いた素性、意味的な情報を用いた素性、名詞句間の距離情報を用いた素性計 30 あまりの素性を用いている。京大コーパスの報道 90 記事に対して名詞句同一指示関係のタグを付与し、10 分割交叉検定を行った結果、F 値として 70.9% を得ている。

5.2 直接照応解析

5.2.1 解析の方針

先行研究において規則を用いた手法の方が良い精度を実現していることから、直接照応という現象は、比較的単純な規則によって解決できる場合が多いと考えられる。そこで本研究では機械学習を用いずに、直接照応の性質に関して設定した規則を用いた直接照応解析を行うことにする。

照応詞と先行詞の関係は大きく以下のように分類できる。

1. 照応詞の表記が先行詞の表記に含まれているもの (ex. 大統領官邸=官邸)
2. 同義表現による言い換え (ex. 北大西洋条約機構= NATO)
3. 照応詞が代名詞となっているもの (ex. 松井=彼)
4. その他 (文脈の理解が必要となるもの) (ex.1995 年=前年)

このうち、1 は基本的に照応詞が先行詞と一致する場合や、末尾に含まれている場合で、特別な知識がなくても認識が可能である。また、人名である場合は先行詞がフルネームで照応詞が名字である場合なども考えられる。この場合は固有表現を正しく認識できれば認識が可能となる。

一方、「北大西洋条約機構」と「NATO」のように、同義表現を用いた言い換えは人間が同一性を理解する場合も、事前の知識がないと困難な場合も多いと考えられる。本研究では 3 章で自動獲得した知識を用いることによりある程度の同一性の認識を目指す。3 については固有表現認識により人名の認識ができれば解析が可能である。4 については、シソーラスを用いたり、文脈的な手がかりを用いることによって解決できる場合があると考えられるが本研究では解析を行わない。このため、本研究で認識できる言い換え表現は、基本的に 2 つの表現の間に常にイコールの関係が成立するような同義表現を用いた言い換えのみで、「犬」と「その動物」などといったような上位下位関係などを用いた言い換えは認識できない。

次に、複合名詞中のどのような部分形態素列が照応現象に関係するかを考える。例えば「調査内容」とあった場合に、その部分形態素列「調査内容」、「調査」、「内容」のうちどれが照応詞、先行詞になり得るかと考えると、先行文中に「内容」にあたる語が出現していたり、後方から「内容」が照応される可能性はあるものの、それはあくまで「調査内容」であり、「調査内容」と「調査」を照応詞・先行詞として考えれば十分であると考えられる。そこで本章では、直接照応解析を行うにあたり、まず次のような方針を用いる。

方針 1 複合名詞の先頭を含む部分形態素列を照応可能要素とする

ここで、照応可能要素とは、照応詞、先行詞になりうる要素のことである。例えば、「金融派生商品」という語があった場合は「金融派生商品」、「金融派生」、「金融」などの語が照応可能要素となり、「派生商品」、「派生」、「商品」などの語は照応可能要素にはならない。

また、一般的に、文章中に新しい概念が登場する際は、その性質や内容を表す節によって修飾されている場合が多いと考えられる。これに対して、既に文章中に出現している内容・対象を指す語の場合はすでに行われた説明を繰り返すと冗長になるため、多くの場合修飾されずに出現する。そこで、どのような表現を照応詞の候補とするかについては次のような方針を用いる。

方針 2 修飾されていない照応可能要素のみを照応詞候補とする

ただし、指示詞や「同」に修飾されている表現については、むしろ前方を照応していると考えた方が自然であり、これらの語による修飾は例外とする。また、固有表現は、修飾語によってさらに限定されることはないと考えられるので、固有表現については修飾されている場合も照応詞として考える。

5.2.2 直接照応解析のアルゴリズム

前節で述べたような方針に基づく、直接照応解析のアルゴリズムを表 5.2 に示す。ただし、表 5.2 中の 3 における修飾に関する条件は表 5.3 に示す。

表 5.2: 直接照応解析のアルゴリズム

1. 対象とする文章について、形態素解析、固有表現認識、構文解析を行う。
2. 文章中に出現するすべての複合名詞、代名詞、固有表現を照応可能要素とする。また、複合名詞の先頭の形態素から始まる、複合名詞中のすべての部分形態素列を照応可能要素とする。
3. 文頭の文節から順に、文末まで以下の処理を行う。
   (a) 次の条件のいずれかを満足する照応可能要素を照応詞候補とする。ただし、同一の複合名詞中の照応可能要素については長いものを優先する。
      - 固有表現である
      - 指示詞、または「同」に修飾されている
      - 修飾に関する条件 (表 5.3 参照)
   (b) 各照応詞候補について、以前に出現した照応可能要素を近いものから順に先行詞候補とし、以下の条件のいずれかを満足した場合、照応詞候補と先行詞候補は直接照応関係にあると認定する。条件を満たす先行詞候補が見つからなかった場合は、その照応詞候補は照応詞ではなかったと判断する。(以下の ex. は先行詞、照応詞の順)
      i. 照応詞候補が先行詞候補の末尾に含まれる (ex. 大統領官邸=官邸)
      ii. 同義表現に置換することにより照応詞候補が先行詞候補の末尾に含まれる (ex. 露=ロシア)
      iii. 先行詞候補、照応詞候補がともに “人名” で、照応詞候補が先行詞候補の先頭に含まれる[1](ex. 小泉純一郎=小泉)
      iv. 先行詞候補、照応詞候補がともに “地名” で先行詞候補の末尾 1 文字を除いたものと照応詞候補が一致する[2](ex. 神戸市=神戸)
      v. 照応詞候補が人称代名詞で照応詞が “人名” である (ex. 松井=彼)

(32) アジア・太平洋地域の 情報化 推進の司令塔の役割を果す。

[1]前方でフルネーム、後方で名字だけが出現する場合があるため
[2]地名の場合、末尾の「市」などが省略される場合があるため

表 5.3: 修飾に関する条件

以下の 4 つの条件でそれぞれ実験する。

1. **修飾を考慮しない** 修飾されている場合も含めて、すべての照応可能要素を照応詞候補とする。

(以下では修飾されていない要素のみ照応詞候補とする)

2. **主辞と同様に扱う** 文節中の主辞以外の要素については、主辞が修飾されている場合は修飾されているとみなす。

3. **主辞以外は修飾されていない** 文節の主辞以外は修飾されていないと考える。

4. **名詞格フレームを用いる** 主辞以外の語については、直前の文節の主辞が自分の格フレーム辞書の用例に含まれている場合のみ修飾されていると考える。

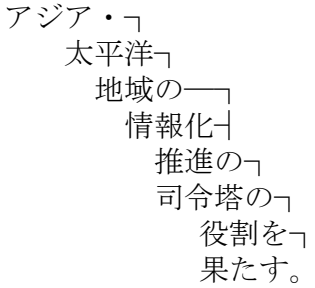


図 5.1: 構文解析結果

例えば (32) のような文章があった場合、構文解析結果は図 5.1 のようになるが、「情報化」という表現は、1(修飾を考慮しない)、3(文節の主辞以外は修飾されていないと考える) の方法を用いた場合は照応詞候補と判断される。2(主辞と同様に扱う) の条件を用いた場合は、主辞「推進」は「地域の」によって修飾されているので、「情報化」は修飾されていると判断し照応詞候補としない。また、4 の方法を用いた場合は「情報化」の格フレーム辞書を用いるが、この場合はその用例に「地域」が含まれているので「情報化」は修飾されていると判断され、照応詞候補とはしない。

このアルゴリズムは前節で述べたような 2 つの方針に従っており、また、解析できる直接照応現象を、知識を必要としないようなもの、および自動獲得した同義表現を用いて解析できるものに限定している。このため以下の 5 つの場合は本手法で原理的に解析できない。

照応詞が照応詞候補とならないもの

1. 照応詞が複合名詞の先頭形態素を含まない部分形態素列である

2. 照応詞が固有表現でなく、かつ指示詞、「同」以外の表現に修飾されている

照応詞と先行詞が同一のものを指していることが認識できないもの

3. 照応詞の方が先行詞よりも簡潔な表現である
4. 同義表現の関係にあるのに、自動獲得した同義表現に含まれていない
5. インスタンスとクラスまたは上位下位の関係にあり、主辞が異なる

5.3 新聞記事を用いた実験

5.3.1 使用するコーパスについて

直接照応解析に関する実験には、京都テキストコーパス Version 4.0[19] から 30 記事、合計 181 文を使用する。京都テキストコーパスは、95 年の毎日新聞の記事に、格関係に関するタグ (省略も含む)、名詞間の関係を表すタグ (橋渡し指示に相当)、共参照タグ (本研究では直接照応に相当)、および固有表現タグが付与されている。これらの記事は CRL 固有表現データで固有表現が付与されているので、これらの固有表現タグを付与して用いる。

このコーパスでは意味関係を考える単位をタグ単位と呼び、タグ単位は、基本的に自立語 1 語を核として、その前後に存在する接頭辞、助詞、助動詞などの付属語をまとめたものになっている。ただし、人名の姓名の 2 つ以上の連続は 1 つのタグとして、「前」「後」などの接頭辞、「率」「者」などの接尾辞は独立したタグ単位として扱うという例外がある。本研究でもタグ単位で照応の評価を行う。

このコーパスに付与された共参照タグは、ある表現が既に出現した表現に対して、同じ内容・対象を指していると考えられる場合に「=」というタグが付与されている。また、次の例のように「人名+役職」のような表現にも「=」タグが付与されている。

(33) クリントン 大統領 (=:クリントン)

(34) 金・前商工資源 相 (=:金)

固有表現については特別な扱いをしておらず、固有表現の 1 部分が照応詞、先行詞となるようなタグも付与されている。また、照応詞、先行詞が用言となるようなものや、先行詞が前文となるようなタグも付与されている。また、複合名詞については各タグごとに照応に関するタグが付与されている。

本研究では、名詞句以外を照応詞、先行詞とする照応や、固有表現の 1 部分が照応詞、先行詞となるような照応を対象から除いているので、これらのタグは評価の対象から除く。「人名+役職」のような表現に付与された「=」タグについては、表 5.4 に示すような固有表現解析結果を用いた簡単なルールにより比較的容易にタグを付与することができるので、解析を行い、評価の対象にも入れる。

解析対象とする 30 記事のコーパス中に解析対象とする直接照応タグは、合計 327 個が付与されている。その内訳を表 5.5 に示す。

表 5.4: 人名+役職の認識ルール

次の条件を満たす B を含むタグに直接照応タグを付与する

- 同一文節中に人名と判断された形態素 (末尾を A とする) と、役職を表す形態素 (B とする) がある
- A と B の間に、固有表現、「副」、「元」などの一部の接頭辞、「・」以外の形態素が存在しない

ex. 小泉首相、ブッシュ・米大統領、小沢一郎・自民党元幹事長

表 5.5: コーパスに付与された直接照応タグの内訳

先行詞と照応詞の関係	個数	割合
表層的な表記が類似	242	74%
同義表現	19	6%
普通名詞と代名詞	8	2%
人名と役職	40	12%
その他 (文脈理解が必要)	18	6%
合計	327	100%

5.3.2 評価方法

次のような式により再現率、適合率、F 値の計算を行う。

$$recall(\text{再現率}) = \frac{\text{(分母) のうちシステムも付与したもの}}{\text{コーパスに付与された直接照応タグ}} \tag{5.1}$$

$$precision(\text{適合率}) = \frac{\text{(分母) のうちコーパスにも付与されているもの}}{\text{システムが付与した直接照応タグ}} \tag{5.2}$$

$$F\text{ 値} = \frac{2}{1/recall + 1/precision} \tag{5.3}$$

この際、京都テキストコーパスでは主辞同士にしかタグが付与されていないが、自動解析で付与されたものと主辞が一致していれば正解とする。また、人手で付与された直接照応タグと自動解析で付与されている直接照応タグの照応先が異なる場合も、直接照応タグを辿ることによって、結果的に同一の照応関係を意味している場合は正解とする。

コーパスと自動解析結果でタグの切れ目がずれている場合はコーパスを自動解析結果に合わせた。

表 5.6: 直接照応の解析実験の結果

修飾判定	適合率	再現率	F 値
考慮しない	78.3	**82.6**	80.4
	(270/345)	(270/327)	
主辞が修飾されていれば	**86.3**	76.8	81.2
修飾されていると考える	(251/291)	(251/327)	
主辞以外は修飾	84.5	79.8	82.4
されないと考える	(262/309)	(262/327)	
名詞格フレーム	85.3	80.1	**82.6**
辞書を用いる	(262/307)	(262/327)	

表 5.7: 照応の種類ごとの解析結果

先行詞と照応詞の関係	再現率	
表層的な表記が類似	85.1	(205/241)
同義表現	73.7	(14/19)
普通名詞と代名詞	50.0	(4/8)
人名と役職	97.5	(39/40)
その他 (文脈理解が必要)	0.0	(0/19)
合計	80.1	(262/327)

5.3.3 実験と考察

形態素解析には JUMAN、構文解析には KNP[21] を用いて、上記のアルゴリズムで直接照応の解析を行った。固有表現認識は学習に用いたクローズドな条件で自動解析を行なっているため、ほぼ正しい解析結果が使っている。

解析結果を表 5.6 に示す。修飾の判定方法については、格フレームを用いた場合に最も良い精度が得られた。また、主辞以外は修飾されないとした場合もほぼ同様の精度を得ることができた。

また、もっとも精度の良かった、修飾判定に関する条件として名詞格フレームを用いる手法を採用した場合の解析結果の詳細を表 5.7 に示す。「人名」と「役職」の組合わせが最も良い精度で認識できていることが判る。唯一の誤りは次の例である。複数の役職が「兼」で結ばれているため、後者の役職について正しいタグを付与できなかった。

(35) **ポレワノフ**副首相兼 国家資産管理委員会議長 が … (ポレワノフ=PERSON)

先行研究との比較のため、「人名と役職」のタグを除いた精度を、村田らと飯田らの先行研究の精度と一緒に表 5.8 に示す。また、他の条件は同一にして固有表現認識の結果を用いなかった場合の精度も一緒に示す。

表 5.8: 先行研究との比較

	適合率	再現率	F 値
村田ら (名詞のみの評価)	78.7 (89/113)	77.3 (89/115)	78.1
飯田ら (主辞のみを対象)	76.7 (582/759)	65.9 (582/883)	70.9
本手法 (固有表現不使用)	80.5 (207/257)	72.1 (207/287)	76.1
本手法 (固有表現使用)	83.5 (223/267)	77.7 (223/287)	80.5

いずれの先行研究も使用しているコーパスが異なり、また対象とする照応現象もずれがあるため一概には比較できないが、本手法で十分な精度が得られていると考えられる。また、直接照応解析に先立ち固有表現認識を行うことは効果的であることが判る。

正しい直接照応関係が認識できない、すなわち再現率を低くする主な要因としては、まず文脈の理解が必要となる場合を扱っていないことが挙げられる。

(36) 高知県の一般事務職の採用は**日本**国籍が要件。国は「公権力の行使や公の意思の構成に携わる … 。

例えば、このような文章があった場合、京都テキストコーパスでは「国」の「日本」の間には共参照タグが付与されているが、本手法ではこれを認識することはできない。このような場合は「日本」が「国」であるというシソーラス的知識を用いることが考えられるが、その結果、適合率が低下する可能性がある。

再現率を下げる別の主な要因としては、照応詞は修飾されないという仮定を置いていることが挙げられる。表 5.6 の結果からも、照応詞が修飾されにくいことは確認できるが、例外も存在する。以下の例では 2 つの「国際貢献」は直接照応の関係にあるとのタグが京都テキストコーパスでは付与されているが、2 つ目の「国際貢献」も修飾されているため、この関係は認識されない。

(37)「わが国にふさわしい**国際貢献**による世界平和の創造」と銘打った非軍事分野の国際貢献など「四つの創造」を打ち出している。

第 6章： 橋渡し指示

本章では、まず橋渡し指示に関する先行研究を紹介した後、自動構築した名詞格フレーム辞書を用いた橋渡し指示解析について述べる。

6.1 橋渡し指示に関する先行研究

6.1.1 Google、WordNet を用いた間接照応解析

Poesio らは英文を対象とし、先行詞と照応詞の間の条件に関する知識として、Google および WordNet を用いた手法 [10] を提案した。Poesio らは間接照応解析に用いる情報として辞書的知識と局所的な主題情報を用いている。このうち辞書的知識としては Google distance を用いたシステムと WN distance を用いたシステムの 2 つの手法を用いている。ここで、Google distance は照応詞候補語を NBD、先行詞候補を NPA としたときの “the NBD of the NPA” というクエリに対する Google のヒット数 N に対して次のように定義される。

$$\text{Google distance} = \begin{cases} 1 & \text{if } N = 0 \\ 1/N & \text{otherwise} \end{cases}$$

また、WN distance は、NBD と NPA の共通の上位語をまでの各々の WordNet 上での最短距離の和を ShtstWNDist、WordNet 上で最も離れた 2 語間の距離を MaxWNDist としたとき、次のように定義される。

$$\text{WN distance} = \begin{cases} 1 & \text{if no path} \\ \dfrac{\text{ShtstWNDist}}{\text{MaxWNDist}} & \text{otherwise} \end{cases}$$

また、局所的な主題情報としては発話距離、直前 5 文での初出位置、文章全体における初出位置などを用いており、学習器としては Multi-layer perceptrons を用いている。

153 個の正解を含む Gnome コーパスを用いて行った実験の結果を Table. 6.1 に示す。対象とする間接照応が全体-部分関係のみという制約があるものの、高精度な解析が実現できていると言える。

6.1.2 名詞句「A の B」を用いた手法

村田ら [32] は日本語を対象とし、名詞格フレーム辞書の代わりに、名詞句「A の B」の用例を用いて、橋渡し指示の解析を行っている。具体的には、名詞句「A の B」の用例と、用言の格フレーム辞書を名詞格フレーム辞書の代わりに用い、さらに指示性などに

表 6.1: Poesio らによる間接照応解析に実験結果

	再現率	適合率	F 値
WordNet	84.5	75.4	79.6
Google	86.2	70.6	77.6

関して、直接照応解析のための規則を 9 つ、橋渡し指示解析のための規則を 4 つ用いて解析している。

童話や新聞の社説など合わせて 171 文を用いた実験の結果、63%の再現率、68%の適合率を得ている。

6.2 直接照応解析の必要性

橋渡し指示の解析を行うためには、まず直接照応解析を行う必要がある。

(38) **チェチェン共和国**の**首都**グロズヌイに侵攻したロシア軍は … 。**ロシア** 側は首都 制圧の最終段階に入ったとみられる。

この例のように、先行する表現と同一の内容を表しうる表現が出現した場合は、修飾要素によって限定されていたりしない限りは、橋渡し指示をしていると考えるよりも、直接照応していると考える方が自然であり、橋渡し指示をしているかどうかの判断には、その表現が直接照応しているかどうかの判断が用いることが可能である。

(39) **日本**は 人口 が減少し始めたが、**インド** では、人口 が …

当然、この例のように 2 回出現する修飾されていない「人口」がそれぞれ「日本の人口」と「インドの人口」といったように直接照応せず、別々の要素を橋渡し指示により指している場合も考えられるが、あまり数は多くないと考え、このような例の解析は今後の課題とする。

6.3 名詞格フレームを用いた橋渡し指示解析

6.3.1 解析の方針

基本的なルールとして基本的に橋渡し指示は、先行する表現を間接的に照応している現象なので、直接照応している場合は、別の要素を橋渡し指示している可能性はほとんどない。このため、直接照応している表現は橋渡し指示の対象から除く。

また、固有表現はそれ自体が固有のものなので、仮に何らかの表現に修飾されていたとしても、その情報がなくても同一のものを指すと考えられることから、それは必須的な要素とは言えない。このため、固有表現も橋渡し指示の対象から除く。

表 6.2: 橋渡し解析のアルゴリズム

1. 対象とする文章について、形態素解析、固有表現認識、構文解析、直接照応解析を行う。
2. 文章中に出現する普通名詞のうち、固有表現に含まれず、かつ、直接照応していない語を照応詞候補とし、先頭から順に各照応詞候補について以下の処理を行う
3. 照応詞候補の名詞格フレームが存在するかを調べ、存在しなかった場合は次の照応詞候補に移る。
4. 格フレームが存在した場合は、直接係り受け関係にある語が、必須要素に該当するかを調べる。具体的には、格スロットの用例との類似度が α 以上だった場合は、対応する格スロットを埋める。まだ、対応付けられていない格スロットが残っている場合のみ次の処理に進む。
5. 照応詞候補を含む文、およびその前 2 文に含まれるすべての名詞を先行詞候補とし、格スロットの用例との類似度を計算する。
6. 閾値 β を満足し、かつ最もスコアの良いものをその格スロットに割り当てる。ただし、格スロットに記された、<人>、<組織>などに該当する先行詞の場合は β より低い閾値 γ を適用する。スコアが同点だった場合は以下の条件を考慮して先行詞候補を決定する。
   (a) 先行詞候補が照応詞候補よりも前に出現する
   (b) 先行詞候補に “主題” が与えられている
   (c) マッチした用例の頻度が多い
   (d) 先行詞候補と照応詞候補の表層的な距離が近い

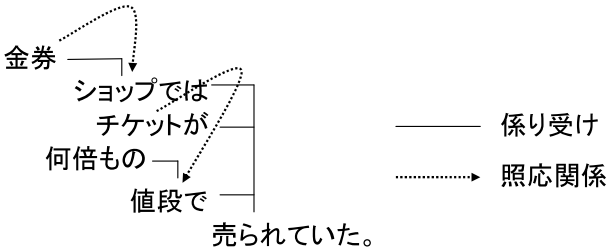
6.3.2 橋渡し指示解析のアルゴリズム

構築した橋渡し指示解析のアルゴリズムを表 6.2 に示す。橋渡し指示解析の例として、次のような文章の解析を考える (図 6.1 参照)。

(40) 金券ショップではチケットが何倍もの値段で売られていた。

形態素・構文解析などの後、まず「金券」の解析が行われ、「金券」に対する格フレームがないことから「金券」の必須要素はないと判断され、照応詞ではないとする。「ショップ」については、直接係っている「金券」が閾値 α を満足するので格スロット〔商品〕に「金券」が対応付けられる。

次に、「チケット」の解析が行われる。直接係っている名詞がないことから、先行詞の候補となる「値段」、「ショップ」、「金券」に対して格スロットに含まれる用例との類似



	格スロット	用例	解析結果
ショップ	〔商品〕	商品 4, 品 3, 雑貨 3, 工芸品 2,…	金券
チケット	〔乗り物・劇場〕	戦 54, コンサート 45, 公演 44,…	なし
値段	〔品物〕	土地 112, 商品 82, チケット 23,…	チケット

【ショップ】商品をならべて売る所。
【チケット】乗り物や劇場などで、料金をはらった印にくれる紙の札。
【値段】品物を売り買いするときの金額。

図 6.1: 橋渡し指示解析

度が計算される。しかし、閾値 α を満足するものがないため必須要素はないと判断される[1]。

最後に、「値段」の解析が行われる。「値段」には「何倍」が直接係っているが閾値 α を満足しないことから他の要素から先行詞が探される。この場合、「チケット」と格スロット〔品物〕の用例との類似度が計算され。用例中に「チケット」があることから最も高いスコアとなり、また閾値を満たすので格スロット〔品物〕にチケットが対応付けられる。

6.4 新聞記事を用いた実験

6.4.1 使用するコーパスについて

橋渡し指示の解析実験にも、直接照応で用いたのと同じく京都テキストコーパス Version 4.0[19] から 30 記事、合計 181 文を使用する。

このコーパスには橋渡し指示に関するタグとしてノ格タグとノ？格タグが付与されている。これは橋渡し指示の場合、3.2.1 節の例 (14) における「チケット」と「値段」のように、誰が見ても「値段」の意味を理解するために必要とであると考えられる関係から、次の例における「日本語」と「ひらがな」の関係のように必須要素であるとは考えにくいような例まで、連続的に存在しており、橋渡し指示であるか否かの判断が困難となる場合があるためである。

[1]必須要素なしと判断されるのは、あくまで入力文中に必須要素が含まれていないためである。本来、チケットとは「<試合・公演>のチケット」であり、この文の前にコンサートなどの単語が出現している場合は、その単語が必須要素であると判断される。

(41) 日本語では、**ひらがな**に加えて、漢字とカタカナが用いられる。

ノ格タグは 3.2.1 節の例 (14) における「チケット」と「値段」のように誰が見ても「値段」の意味を理解するために必要であると考えられるような関係があった場合に、ノ？格タグは「日本語」と「ひらがな」の関係のように、単なる修飾以上の関係ではあるが必須要素であるかどうか微妙だと判断された場合に付与されている。また、「茶色の鞄」のように単なる修飾と判断された場合には修飾タグが付与されている。

ノ格は名詞の必須要素を記述しているので、必ずしも橋渡し指示でない場合も付与されている。具体的には直接、必須要素に修飾されている場合や、前方の要素を直接照応していて、前方に出現した際に必須要素が明示されている場合も付与されている。本研究では橋渡し指示の解析を行うので、直接照応しているタグ、および必須要素に直接修飾されている場合は評価から除く。これらのタグを除いた結果、解析対象とする橋渡し指示に関するタグは 96 個付与されていた。

6.4.2 評価方法

次のような式により再現率、適合率、F 値の計算を行う。直接照応解析の評価と同様に、人手で付与された橋渡し指示の先行詞と自動解析で付与されている橋渡し指示の先行詞が異なる場合も、直接照応タグを辿ることによって、結果的に同一の内容・対象であることが判る場合は正解とする。

$$recall(\text{再現率}) = \frac{(A)\text{のうちシステムも付与したもの}}{\text{コーパスに付与されたノ格タグ } (A)} \tag{6.1}$$

$$precision(\text{適合率}) = \frac{(B)\text{のうちコーパスにノ格またはノ？格タグが付与されているもの}}{\text{システムが付与した橋渡し指示タグ } (B)} \tag{6.2}$$

$$F\text{ 値} = \frac{2}{1/recall + 1/precision} \tag{6.3}$$

再現率を求める際は、京都テキストコーパスで付与されているノ格タグのみを、適合率を求める際は京都テキストコーパス付与されているノ格タグに加えて、ノ？格タグも使用するため、再現率と適合率を求める式の分子は必ずしも一致しない。

6.4.3 実験と考察

橋渡し指示解析の結果を表 6.3 に示す。再現率についてはある程度、高い精度が実現できているものの、適合率は、特に複合名詞中に出現した名詞について、極めて低い精度となっている。

まず適合率が低いのは、名詞の照応性の判定を名詞格フレーム辞書の有無にのみ頼っているためだと考えられる。

表 6.3: 橋渡し指示解析の精度

	適合率	再現率	F 値
単独で出現した名詞	30.3(44/145)	71.4(40/56)	43.0
複合名詞中に出現した名詞	17.0(25/147)	57.5(23/40)	26.2
合計	23.6(69/292)	65.6(63/96)	34.7

(42) 世界でも類を見ない新しい**芸術**を創造した。

例えばこのような文章があった場合、名詞「世界」は自動構築された格フレームを持ち、その用例には「芸術」が含まれる。当然「芸術の世界」という意味で「世界」という名詞が使用されることも考えられ、この格フレーム自体が間違っているわけではない。しかし、この場合の「世界」は一般的な意味での「世界」であり、「芸術の世界」という意味は持たないにも関わらず、「世界」が名詞格フレームを持っているためこの「世界」は照応的と捉えられ、「芸術」を橋渡し指示していると認識されてしまう。

また、次のような文章があった場合も上手く解析できない。

(43) **社会党**は今年 … 。… 新党準備会を旗揚げする方針を崩しておらず、久保亘書記長ら執行部は … 。

この場合、「執行部」は「社会党」を橋渡し指示しており、「社会党の執行部」であると考えられる。ところが、名詞格フレームは単独の名詞ごとに構築しているため、「執行部」は形態素解析の結果「執行 (=普通名詞) ＋部 (=普通名詞)」と判断されるため、「執行部」の名詞格フレームは構築されず、「部」の名詞格フレームを用いて解析を行うことになる。しかし、「社会党の執行部」の用例は多いと考えられるのに対し、「社会党の部」という用例は存在せず「部」の名詞格フレームの用例に「社会党」は含まれないため、「部」の必須要素は、「部」の格フレームの用例に含まれる「会」であると認識されてしまう。

このような橋渡し指示にも対応するためには、一部の複合名詞については、複合名詞の名詞格フレームを構築する必要があると考えられる。

第 7章 :　統合的な解析システムの構築

本章では、固有表現認識や、直接照応、橋渡し指示を統合的に行う利点、および実際にどのようにして統合的に行うかについて説明する。

7.1　統合的な解析について

既に述べたように、本研究で用いる固有表現認識には形態素解析結果が必要となる。また、直接照応解析を行うには固有表現認識や構文解析結果が必要となり、橋渡し指示の解析には直接照応の結果が必要である。さらに、 次節で挙げる例のように、構文解析には固有表現認識結果を利用できる場合がある。

このため、これらの処理は基本的には、まず形態素解析を行い、続いて固有表現認識、構文・格解析、さらに直接照応解析、橋渡し指示解析、ゼロ照応解析を行うといったような順序となる。しかし、例えば固有表現認識の結果を用いて形態素解析結果を修正したり、直接照応解析の結果を用いて固有表現認識結果を修正したりすることが考えられ、後に行う処理の結果をすでに行った解析にフィードバックすることが考えられる。

以下では、固有表現認識が構文解析に役立つ例、および、後に行った処理をフィードバックできる例について紹介する。

固有表現認識結果の利用

固有表現認識結果が構文解析に役立つ例として次のような文章が考えられる。

(44) 韓国の金泳三大統領は新年の辞を発表した。(金泳三=“人”)

(45) 年金未納三閣僚も花押を押していた。

これらの文章における「三」はいずれも形態素解析の結果、数詞と判断されるため、固有表現認識結果を用いずに構文解析を行うと図 7.1 に示すように、「三」の直前で文節が切れるような結果が得られる。しかし、「金泳三」が人名を表しているという固有表現認識結果を用いることにより、正しく構文解析を行うことが可能となる。

また、固有表現認識結果を用いることにより品詞を特定できる場合がある。例えば形態素解析器 JUMAN と構文解析器 KNP を用いて (46) の文を解析すると、「さきがけ」という形態素は “動詞” の連用形であるが、直後に「の」という格助詞が来ているため名詞であると判断される。

(46) 新党さきがけの武村代表は … 。

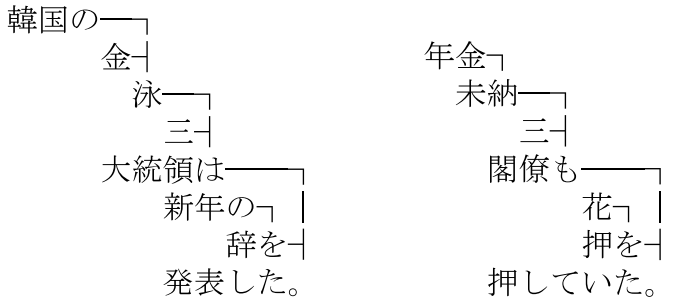


図 7.1: 構文解析例

(47) 武村代表が さきがけ 主導の新党結成に改めて意欲を示す。

しかし、(47) のような文章があった場合、直後に格助詞を伴っていないため、この「さきがけ」は名詞であるとは判断されず、動詞であるとされてしまう。しかし、固有表現認識の結果「さきがけ」は組織名であることが判るので、「さきがけ」を名詞であることが認識できるようになる。

直接照応解析結果の利用

4 章で述べたように、固有表現認識においてキャッシュを導入することにより、同表記の固有表現については、以前の文での固有表現認識結果を用いて固有表現の認識を行えるようになったが、同義表現辞書を用いている場合は必ずしも同じタグと認識されるわけではない。

(48) **日本長期信用銀行**も 1766 億円の資本増強が決定し、株価は 373 円まで値上りした。しかし、バブル崩壊で深手を負った 長銀 にとって、この金額は焼け石に水だった。

例えばこのような文章があった場合、SVM を用いた固有表現認識を行った結果、「日本長期信用銀行」は “組織名” であると認識されるが、「長銀」は「長」が接尾辞と「銀」が普通名詞と形態素解析されるため組織名と認識されない。しかしながら、「長銀」と「日本長期信用銀行」は、自動構築した同義表現辞書には含まれており、直接照応解析により同一のものであると解析される。この解析結果と「日本長期信用銀行」が組織名であることから、「長銀」が組織名であることを認識することができる。

この例の場合、事前に同義表現辞書に登録されている語が、固有表現であるかどうかを判断しておくことでも解決することができるが、例えば「ヨーロッパ共同体」という同義表現と「電子商取引」という同義表現をもつ「ＥＣ」や、「イングランドサッカー協会」という同義表現と「フリーエージェント」という同義表現を持つ「ＦＡ」のような例もあることから、事前に登録するだけでは上手くいかない場合も考えられる。

橋渡し指示解析結果の利用

5.2.2 節で述べた方法で照応解析を行った場合、照応詞は修飾されないものとしているた

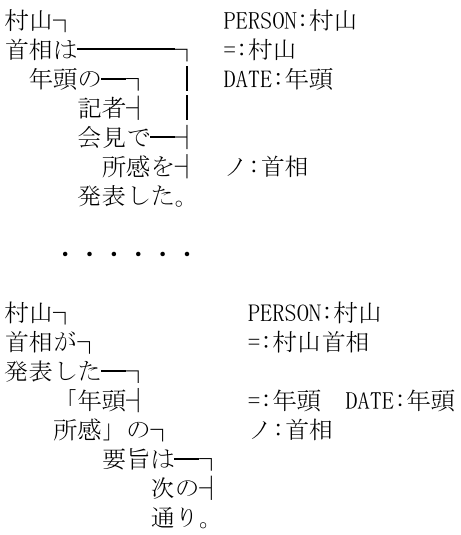


図 7.2: 直接照応解析、橋渡し指示解析の結果

め、前方に出現した表現が再び修飾されて出現した場合には同一の対象・内容を指示している場合もそのことを認識できない。しかしながら、直接照応の照応詞となることは少ないものの、後から説明が加わったりする場合など、修飾された表現が前方を照応していることも考えられる。

(49) 村山首相は年頭の記者会見で**所感**を発表した。… 。村山富市首相が発表した「年頭 所感」の要旨は次の通り。

例えば、このような文章があった場合、2 度目に出現する「所感」は前方の出現した「所感」と同一の内容を指している。しかし、修飾されていると判断されるため、5 章の表 5.2 に示したようなアルゴリズムではこれらの同一性を認識できない。

しかし、橋渡し指示についてはその語が修飾されている場合も解析を行っており、この場合は、1 文目の「所感」と同様に橋渡し指示先として同一の対象を指す「首相」を指示していると解析される (図 7.2 参照)。いずれも同一の「首相の所感」であることが判り、2 度目に出現する「所感」が 1 文目の所感と同一のものを指していることは容易に推測される。このように、修飾されている要素であっても、橋渡し指示による意味内容が補完を、直接照応解析に用いることができる。

7.2 統合的な解析システム

7.2.1 システムの概要

基本的には以下のような順序で解析を行っていく。

1. 形態素解析

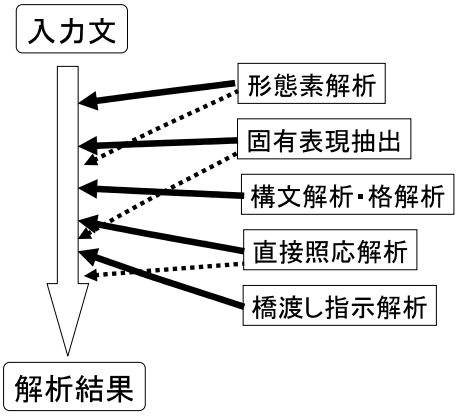


図 7.3: 統合的な解析

2. 固有表現認識

3. 構文・格解析

4. 直接照応解析

5. ゼロ照応・橋渡し指示解析

ただし、2 の固有表現認識が終了した後に、その結果を用いた形態素解析の修正を、4 の直接照応解析が終わった後に、その結果を用いた固有表現の新たな認識を、さらに 5 の橋渡し指示解析が終わった後に、その結果を用いた直接照応解析をそれぞれ行うものとする (図 7.3 参照)。

7.2.2 直接照応解析を用いた固有表現認識

直接照応結果を用いた固有表現認識として、先行する組織名を直接照応している表現に組織名を与えた。ただし、インスタンスである組織名に対するクラスにあたる表現であることも考えられるので、表現が完全に一致するか、または新聞中の括弧表現から自動構築された同義表現辞書に登録されている表現である場合のみ、固有表現タグを引き継ぐこととする。

使用したコーパスは、CRL 固有表現データと IREX テストデータで、前者については 5 分割交叉検定により評価を行った。その結果と、それぞれのデータを用いた最も精度の良い先行研究 (シソーラスを用いなかった場合) の精度を表 7.1 に示す。

いずれのデータを用いた場合も精度の向上が見られ、CRL 固有表現を用いた交叉検定ではあまり大きな変化がなかったが、IREX テストデータでは大きく精度が向上した。これは、IREX テストデータには CRL 固有表現データで削除されている括弧表現が残っており、同義表現の関係にある表現ペアが多く出現しているためだと考えられる。

表 7.1: 直接照応解析結果を用いた固有表現認識

	CRL データ	IREX テストデータ
先行研究 (順に中野ら、磯崎ら)	88.72	85.11
直接照応解析結果使用せず	86.70	85.00
直接照応解析結果を使用	86.72	85.84

表 7.2: 橋渡し指示解析結果を用いた直接照応解析

橋渡し指示	適合率	再現率	F 値
使用しない	85.3	80.1	82.6
	(262/307)	(262/327)	
使用する	85.3	81.3	**83.3**
	(266/312)	(266/327)	

改善された例と比べると数は少ないものの、いくつか誤った認識も増えた。この主な原因は前方文脈で誤って認識された固有表現をそのまま引き継いでしまったのが原因であり、直接照応の誤りによる誤認識はなかった。

7.2.3 橋渡し指示を用いた直接照解析

修飾されている照応可能要素であっても、橋渡し指示解析の結果、先行する同一表現と、同一の対象・内容を橋渡し指示していると判断されたものは、直接照応の関係を与えるというルールを追加して、直接照応解析を行った。解析結果を表 7.2 に示す。修飾の判定には名詞格フレームを用いた手法を使っている。

修飾されているために検出できなかった 2 つの直接照応関係を新たに認識することができた。大きく改善することはできないものの、誤った解析結果が新たに出力される可能性が低いことから、この手法は有効であると考えられる。

第 8章： 結論

本研究では、文章中の要素間の関連性認識を目指し、従来個別に扱われることの多かった直接照応解析や、橋渡し指示解析などといった照応解析を、固有表現認識技術や事前に獲得した知識を用いて統合的に行なった。

まず、照応解析の基礎となる固有表現認識については、局所的な情報に加えて以前の文での解析結果なども考慮に入れることにより、従来よりも少ない知識しか用いずに高い精度を実現することができた。また、直接照応解析の結果を用いることにより精度が向上することが確認できた。

直接照応解析については、修飾されていない表現は基本的に前方を照応しており、逆に新しい対象や内容を指示する表現は修飾されて出現することが多いという性質に着目し、さらに事前に自動獲得した同義表現に関する知識や、固有表現認識結果を用いることにより、適合率 85.3%、再現率 80.1%と、機械学習を用いた従来の手法よりも高い精度を実現することができた。

このうち、従来認識することが難しかった同義表現が関係する照応現象については、70%を越える再現率を実現することができた。このことから、コーパス中に出現する括弧表現、および、国語辞典からの同義表現の自動獲得が有効であったことが判った。さらに、修飾されているかどうかの判定については、名詞格フレーム辞書を用いた場合について、最も良い精度を実現することができ、橋渡し指示解析以外の用途についても自動構築した名詞格フレーム辞書を用いることができることが判った。

また、一部修飾されている表現についても、橋渡し指示解析結果により省略された必須要素が同じであることが解ったものについては同一性を認定したところ、ほとんど適合率を落とすことなく、再現率を上昇させることができ、照応解析と橋渡し指示解析を統合的に行うことの有効性が確認された。

橋渡し指示解析については、再現率についてはある程度高い精度を実現することができたが、適合率については十分な精度を実現することはできなかった。しかし、解析結果を直接照応解析にフィードバックすることにより、直接照応解析が向上したことから、使用方法によっては有効に用いることもできると考えられる。また、高い再現率が実現できたことから、名詞の必須要素の多くが、自動構築した名詞格フレーム辞書に記述されていることが確認できた。

今後の課題としては、直接照応解析については、シソーラスをコーパスや国語辞典などから自動構築し、それを用いて文脈理解が必要となるようなものについても解析できるものを増やしていくことが考えられる。また、修飾されている要素についても限定的な修飾でない場合は、解析の対象とすることが考えられる。

橋渡し指示については、本研究で構築した名詞格フレーム辞書では、複合名詞の橋渡し指示を正しく扱えないという問題がある。このため、複合名詞にも対応した名詞格フ

レームの構築することが必要となる。また、適合率を向上させるために、どういう場合に橋渡し指示の照応詞となるのかや、それぞれの格スロットには、どのような格要素として出現した語が入りやすいのかなど、文脈を考慮したシステムの構築を行っていくことが必要となる。

本研究では、ゼロ照応についてはあまり触れなかったが、既に格フレームを用いて解析する枠組みが構築されている。今後、ゼロ照応に関する既存の枠組みを、固有表現認識や、直接照応、橋渡し指示の解析結果を用いて高精度化し、また、逆にゼロ照応解析の結果を直接照応や、橋渡し指示にフィードバックしていき、今後、さらなる統合的な解析システムの構築していくことが考えられる。

謝 辞

本論文を執筆するにあたり、熱心にご指導下さいました黒橋禎夫助教授に心より感謝致します。

また、日ごろから研究室の皆さまにはお世話になりました。中でも学術研究支援員の河原大輔氏には研究の方針やプログラムに関することなど多くのことを教えていただきました。同じく学術研究支援員の岡本雅史氏には照応現象の言語学における扱いなどに関することなど多くのご助言をいただきました。博士 2 年の柴田知秀氏には計算機に関することを中心に多くのご助言をいただきました。また、秘書の町居信子氏には研究室環境を整備していただくなど、いろいろとお世話になりました。大変感謝致します。

参 考 文 献

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

[2] ZHOU GuoDong and SU Jian. A high-performance coreference resolution system using a constraint-based multi-agent strategy. In *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 522–528, 2004.

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

[4] Ryu Iida, Kentaro Inui, Hiroya Takamura, and Yuji Matsumoto. Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics Workshop on The Computational Treatment of Anaphora*, pp. 23–30, 2003.

[5] Daisuke Kawahara and Sadao Kurohashi. Fertilization of Case Frame Dictionary for Robust Japanese Case Analysis. In *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 425–431, 2002.

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

[7] Masaki Murata, Hitoshi Isahara, and Makoto Nagao. Resolution of Indirect Anaphora in Japanese Sentences Using Examples "X no Y" (Y of X). In *Proceedings of ACL'99 Workshop on 'Coreference and Its Applications'*, 1999.

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

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

[10] Massimo Poesio, Pahul Mehta, Axel Maroudas, and Janet Hitzeman. Learning to Resolove Bridging References. In *Proceedings of the 42th Meeting of the Association for Computational Linguistics (ACL'04)*, 2004.

[11] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1993.

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

[13] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, Vol. 27, No. 4, pp. 521–544, 2001.

[14] Michael Strube and Udo Hahn. Functional centering – grounding referential coherence in information structure. *Computational Linguistics*, Vol. 25, No. 3, pp. 309–344, 1999.

[15] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[16] Renata Vieira and Massimo Poesio. An empirically based system for processing definite descriptions. *Computational Linguistics*, Vol. 26, No. 4, pp. 539–592, 2000.

[17] 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.

[18] 磯崎秀樹, 加沢秀人. 固有表現抽出のための SVM の高速化. 情報処理学会論文誌, Vol. 44, No. 3, pp. 970–979, 2003.

[19] 河原大輔, 笹野遼平, 黒橋禎夫, 橋田浩一. 格・省略・共参照タグ付けの基準, 2004.

[20] 久光徹, 丹羽芳樹. 統計量とルールを組み合わせて有用な括弧表現を抽出する手法. 情報処理学会 自然言語処理研究会 1997-NL-122, pp. 113–118, 1997.

[21] 黒橋禎夫. 日本語構文解析システム KNP version 2.0 b6 使用説明書. 京都大学大学院 情報学研究科, 6 1998.

[22] 黒橋禎夫, 河原大輔. 日本語形態素解析システム JUMAN version 4.0 使用説明書. 京都大学大学院 情報学研究科, 7 2003.

[23] 山田寛康, 工藤拓, 松本裕治. Support vector machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol. 43, No. 1, pp. 44–53, 2002.

[24] 山本和英. テキストからの語彙的換言知識の獲得. 言語処理学会 第 8 回年次大会発表論文集, 2002.

[25] 山梨正明. 推論と照応. くろしお出版, 1992.

[26] 酒井浩之, 増山繁. コーパスからの名詞と略語の対応関係の自動獲得. 言語処理学会第 9 回年次大会発表論文集, 2003.

[27] 上野友司, 森辰則, 木戸冬子, 中川裕志. 係り受けの 2 部グラフと共起関係を利用した同義語抽出. 言語処理学会 第 10 回年次大会発表論文集, 2004.

[28] 西尾実, 岩淵悦太, 水谷静夫（編）. 岩波国語辞典. 三省堂, 2000.

[29] 浅原正幸, 松本裕治. 日本語固有表現抽出におけるわかち書きの問題の解決. 情報処理学会論文誌, Vol. 45, No. 5, pp. 1442–1450, 2004.

[30] 村上明子, 那須川哲哉. 複数の著者の表記の違いを利用した同義表現抽出. 情報処理学会 自然言語処理研究会 2004-NL-162, pp. 117–124, 2004.

[31] 村田真樹, 長尾眞. 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.

[32] 村田真樹, 長尾眞. 意味的制約を用いた日本語名詞における間接照応解析. 自然言語処理, Vol. 4, No. 2, 1997.

[33] 中野圭吾, 平井有三. 日本語固有表現抽出における文節情報の利用. 情報処理学会 自然言語処理研究会 2003-NL-156-2, pp. 7–14, 2003.

[34] 田近洵一（編）. 例解小学国語辞典. 三省堂, 1997.

[35] 奈良先端科学技術大学院大学自然言語処理学講座松本研究室. 日本語形態素解析システム『茶筌』, 2003.

[36] 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.

[37] 颯々野学, 宇津呂武仁. 統計的日本語固有表現抽出における固有表現まとめ上げ手法とその評価. 情報処理学会 自然言語処理研究会 2000-NL-139-1, pp. 1–8, 2000.