自動獲得した名詞関係辞書に基づく共参照解析の高度化

笹野　遼平[†,††]・黒橋　禎夫[†††]

本稿では，自動獲得した知識を用いた日本語共参照解析システムを提案する．日本語における共参照の多くを占める名詞句間の共参照の解析では，語彙的知識が重要となり，中でも同義表現知識が非常に有効となる．そこでまず，大規模なコーパスおよび国語辞典の定義文から同義表現の自動獲得を行い，自動獲得した同義表現を用いた共参照解析システムを構築する．さらに，より精度の高い共参照解析システムの構築のため，自動構築した名詞格フレームを用いた名詞句の関係解析を行い，その結果を共参照解析の手掛りとして使用する．新聞記事およびウェブテキストを用いた実験の結果，同義表現，および，名詞句の関係解析結果を用いることにより，共参照解析の精度は向上し，手法の有効性が確認できた．

キーワード：共参照解析，同義表現抽出，名詞句の関係解析

Improving Coreference Resolution Using Automatically Acquired Knowledge of Nominal Relations

Ryohei Sasano[†,††] and Sadao Kurohashi[†††]

We present a knowledge-rich approach to Japanese coreference resolution. In Japanese, noun phrase coreference occupies a central position in coreference relations. To improve coreference resolution for such language, wide-coverage knowledge of synonyms is required. We first acquire knowledge of synonyms from large raw corpus and dictionary definition sentences, and then resolve coreference relations based on the knowledge. Furthermore, to boost the performance of coreference resolution, we integrate bridging reference resolution system that uses automatically constructed nominal case frames into coreference resolver. We evaluated our approach on news paper article and WEB corpus and confirmed that the performance of coreference resolution is improved by using automatically acquired synonyms and bridging reference resolution.

Key Words: *Coreference, Synonym Extraction, Bridging Reference*

1　はじめに

共参照解析とは，ある表現が他の表現と同一の対象を指していることを同定する解析のことであり，計算機による自然言語の意味理解を目指す上で重要な技術である．本研究では，日本

† 東京大学大学院情報理工学系研究科, Graduate School of Information Science and Technology, University of Tokyo
†† 日本学術振興会特別研究員 DC, Research Fellow of the Japan Society for the Promotion of Science
††† 京都大学情報学研究科, Graduate School of Infomatics, Kyoto University

語文における，同一文章内の表現間の共参照である文章内共参照を解析の対象とする．文章内共参照では，ある表現（照応詞）が文章中の先行する表現（先行詞）と同一の対象を指している場合にそれを認識することが目的となる．

共参照における照応詞としては，普通名詞，固有名詞，代名詞の3つが考えられる．英語などの言語では照応詞として代名詞が頻繁に使用されるが，日本語では代名詞の多くはゼロ代名詞として省略されるため，照応詞の多くは普通名詞，固有名詞が占めている．ゼロ代名詞の検出・解析（ゼロ照応解析）も，意味理解を目指すためには欠かすことのできない解析であり，多くの研究が行われている (Seki, Fujii, and Ishikawa 2002; Kawahara and Kurohashi 2004; Iida, Inui, and Matsumoto 2006)．ゼロ照応解析は，先行する文中から先行詞を同定するという点では共参照解析と同じであるが，ゼロ代名詞の認識が必要である点，省略されているため照応詞自体に関する情報がない点で異なっており，より応用的なタスクであると言える．本研究では，高精度な照応解析システムを実現するためには，まず基礎的な照応解析である共参照解析の精度向上が重要であると考え，共参照解析の精度向上を目指す．

共参照解析の手法としては大きく分けて，人手で作成した規則に基づく手法と，タグ付きコーパスを用いた機械学習に基づく手法がある．英語を対象とした共参照解析では，これらの2手法によりほぼ同程度の精度が得られている (Soon, Ng, and Lim 2001; Ng and Cardie 2002; Guodong and Jian 2004)．一方，日本語の場合は規則に基づく手法で高い精度が得られる傾向がある (Iida, Inui, Takamura, and Matsumoto 2003; 村田，長尾 1996)[1]．日本語において規則に基づく手法で高い精度が得られるのは，普通名詞，固有名詞間の共参照関係が大部分であり，語彙的情報が非常に大きな役割を占めるため，機械学習によって得られる性向が，人手で作成した規則でも十分に反映できているためであると考えられる．そこで本研究では基本的に，人手で設定した規則に基づく共参照解析システムを構築する．

照応詞が普通名詞，固有名詞となる場合，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 その他（クラスとインスタンス，上位語と下位語など）：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．ただし，末尾が一致する場合すべてが共参照関係にあるわけではなく，精度の高い解析のためには照応詞，先行詞が指すものを解析する必要がある．例えば次のような2文があった場合，いずれの文にも「結果」という語が複数回出現するが，a

[1] これらの研究では使用しているコーパスが異なるため単純には比較できないものの，新聞記事を対象とした予備実験の結果，規則に基づく手法でより高い精度が得られた．

ではそれらが同一の内容を指しているのに対し，bでは異なる内容を指している．

(1) a. 2006 FIFA ワールドカップ優勝国予想アンケートを行った．結果はブラジルがトップだった．アンケート結果の詳細はWebで見られる．
b. 先月行なわれた韓国との親善試合の結果を受けアンケートを行った．アンケート結果から以下のようなことが判明した．

これらの違いを正しく解析するためには，a中の「結果」はともに「アンケートの結果」を意味しているのに対し，b中の「結果」は順に「試合の結果」，「アンケートの結果」を意味していることを認識する必要がある．そこで本研究では，係り受け解析，および，自動構築した名詞格フレームに基づく橋渡し指示 (bridging reference) 解析により名詞句の関係を解析し，その結果を共参照解析の手掛りとして用いる．

2は「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となっている場合である．同義表現を用いた言い換えとなっている場合，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．そこで，同義表現に関する知識を事前にコーパスや国語辞典から自動的に獲得し，獲得した同義表現知識を共参照解析に使用する．

3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なわず，今後の課題とする．

2 同義表現の自動獲得

2.1 獲得に用いるリソース

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や，テキストの局所的な文脈依存性を利用する手法 (山本 2002)，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 (酒井，増山 2003)，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 (上野，森，木戸，中川 2004)，複数の著者の表記の違いを利用した手法 (村上，那須川 2004) 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している括弧表現を用いた手法を用いる．

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，新語や未知語への対応力は強いものの，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は抽出できない点が挙げられる．

(2) 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，極めて常識的な言い換え表現を獲得するため国語辞典からも同義表現の抽出を行う．国語辞典から獲得できる同義表現ペアには，新語などは含まれないものの，括弧を用いて表記されないような極めて一般的な同義表現ペアが含まれていると考えられ，括弧表現と国語辞典の2つのリソースを用いて同義表現を抽出することで，多くの同義表現を獲得できると考えられる．

2.2 括弧表現からの同義表現の抽出

括弧の解析に関する先行研究としては久光らの研究 (久光，丹羽 1997) や，Okazaki らの研究 (Okazaki and Ishizuka 2008) がある．久光らは統計量とルール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．久光らの手法は，小規模なコーパスからも大量に同義表現を抽出できるという特徴がある．実験には日経新聞 1992 年 1 年分を使用しており，もっとも高い精度となる Yate 補正した χ^2 とルールを組合せた場合の言い換え抽出精度は，χ^2 値上位 500 位に含まれる 437 個の言い換え表現の獲得に対しては約 99.3%，χ^2 値上位 501 位～6366 位に含まれる約 3400 個の言い換え表現の獲得に対しては約 96.5%である[2]．

Okazaki らは，新たに以下の 2 つの条件を同時に満たす文書は「A → B」の語彙的言い換えであると認定することで計算される言い換え発生率を指標として導入し，精度 95.7%，適合率 90.0%，再現率 87.6%を得ている．

1 「A(B)」のパターンが出てくる前の文において，表現 B が出現しない．

2 「A(B)」のパターンが出てきた後の文において，表現 A よりも表現 B の出現頻度が高い．

実験には 1998–1999 年の毎日新聞・読売新聞に含まれる括弧表現のうち共起頻度が 8 よりも大きい語彙対を使用しており，その中に含まれる言い換え可能な表現は 1,430 事例である．再現率が 87.6%であることから約 1,250 個の言い換え表現を獲得していることになる．

本研究では，共参照解析において有用となる同義表現を獲得することを目的とし，できるだけ出現頻度の高い同義表現を精度良く獲得することを目指す．同義表現であるならば，「A(B)」のパターンに加えて，「B(A)」のパターンも出現する（双方向性がある）可能性が高いと考え，双方向性に注目することにより高精度に同義表現を抽出する手法を提案する．抽出する同義表現を「A(B)」のパターンに加えて，「B(A)」のパターンも出現するものに限定することにより，コーパスサイズに対する獲得同義表現数は少なくなるものの，高い精度で抽出できると考えられる．提案する括弧表現からの同義表現の自動獲得の手順は以下の通りである．

1. 長い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

[2] 久光らは，適切な文字列の削除／追加により正解となるものを半正解としているが，ここでは本研究の基準と同様にそれらを不正解として計算している．

パスから取り出し，AとBを同義表現の候補とする．例えば，(3) 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「金融システムの危機について焦点となっている長銀」を取り出す．

(3) 現在のところ，金融システムの危機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，….

2. 短い同義表現候補の抽出

Bの形態素解析を行い末尾の名詞句B'がBと異なる場合はB'を取り出し，AとB'のペアも同義表現候補とする[3]．例えば (3) 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．

3. 同義表現の決定

A(B) とB(A) の両方が出現しているものに対し，表1に示すような同義表現候補のタイプごとに設定した閾値を満足する同義表現候補を同義表現として抽出する．

表1に示した閾値は事前に100個程度の同義表現候補とその出現頻度を参考に決定した．また，実験には毎日新聞12年分と読売新聞14年分，計26年分，約2,600万文を使用した．約2,600万文中に文頭に出現するものを除いて括弧は約1,000万回出現し，短い同義表現候補の異なり数は約110,000個，双方向性のある語彙対は5,800個であった．獲得された固有表現の種類と数，正しいと判断されたものの割合（正解の割合）を表2に示す．正解の割合はタイプ1，タイプ2に関してはランダムに抽出した200個を，タイプ3，タイプ4に関しては抽出されたすべての同義表現対を人手で評価し算出している[4]．

表2の結果から約2,600個の同義表現を精度約99%という高い精度で同義表現を獲得できており，双方向性に注目した絞り込みが有効であったことが確認できる．また，先行研究と比較した場合，大規模なコーパスを使うことにより，抽出精度を落とさずに多くの同義表現の獲得

表 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	同義表現とみなす条件
1　一方が英字，もう一方が英字以外	頻度の積 $>$ 2 [0]
2　一方がカタカナ，もう一方がカタカナ以外	頻度の積 $>$ 5 [2]
3　共に漢字で，一方が他方の部分集合	頻度の積 $>$ 1 [0] & 文字長の差 $>$ 2
4　その他	頻度の積 $>$ 200 [20]& それぞれの頻度 $>$ 8
	[ ] 内の数字は比較実験に用いた緩い閾値

[3] 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AとB'のいずれかが，人名のみ，または地名のみで構成されている場合は候補としていない．

[4] 抽出された同義表現対が正解であると判断する基準は，それらが同一文章中に出現した場合に同じ対象を指していることが多いと考えられるかどうかである．

表 2 括弧表現を用いた同義表現抽出の結果

タイプ	数	正解の割合	抽出された同義語の例
1	1,572	99.5%	国連平和維持活動=PKO, 北大西洋条約機構=NATO
2	727	98.5%	関税貿易一般協定=ガット，金融派生商品=デリバティブ
3	239	98.7%	住宅金融専門会社=住専，動力炉・核燃料開発事業団=動燃
4	110	96.4%	朝鮮民主主義人民共和国=北朝鮮，二酸化炭素=CO2
合計	2,648	99.0%	

表 3 緩い閾値を用いて新たに抽出された同義表現

タイプ	数	正解の割合	誤って抽出された同義語の例
1	554	95.0%	タレント養成所=NSC, 現行=NHK
2	151	96.0%	店頭市場=ジャスダック，労働=ヒト
3	48	93.8%	社会教育=社会，副会長理事=理事
4	18	66.7%	大分地検検事正=最高検検事，申請本=検定前
合計	771	94.8%	

に成功していると言える．

次に，使用した閾値の妥当性を確認するため同義表現であると判断する閾値を表 1 に "[ ]" を用いて記した閾値に緩めて同義表現の抽出を行った．新たに同義表現と判断された語の数と評価を表 3 に示す．新たに約 730 個の正しい同義表現が獲得され，その精度は約 95%であった．このことから，使用するコーパスから出来るだけ多くの同義表現の獲得を目的とする場合，閾値を緩めた方が良いと考えられる．しかしながら，同義表現の獲得数はより大規模なコーパスを使用することで増やすことができると考えられること，精度の高い同義表現データの方が汎用性が高いと考えられることから本研究では緩い閾値は採用せず，表 2 に示した同義表現を知識として使用する．

2.3 国語辞典からの同義表現の抽出

括弧表現を用いるだけでは抽出できないと考えられる 2.1 節の (2) 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については多くの研究が行なわれており (鶴丸，竹下，伊丹，柳川，吉田 1991; Tokunaga, Syotu, Tanaka, and Shirai 2001; Nichols, Bond, and Flickinger 2005)，それらの多くは国語辞典から出来る限り多くの情報を抽出することを目的としている．本研究では，共参照解析に有用な同義表現の獲得を目的としており，また，コーパス中に出現する共参照関係にある同義表現のうち括弧表現から抽出できないものの多くは常識的な地名の言い換えであることから，これらの地名を含む常識的な言い換えが抽出できれば十分であると考える．

そこで，これらが抽出できるような簡単な規則を設定し国語辞典からの同義表現抽出を行う．同義表現抽出のために用いた規則を以下に示す．

1　対象の語の見出し語Aを取り出す．
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4に示すような見出し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，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．「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．

実験に用いた国語辞典は「例解小学国語辞典 (田近 1997)」と「岩波国語辞典 (西尾，岩淵，水谷 2000)」である．「例解小学国語辞典」は小学生向けの辞書で基本的な語が比較的平易な定義文により記載されており，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．

表5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

表4　例解小学国語辞典の定義文の例

ソビエトれんぽう
　表記：ソビエト連邦
　品詞：地名
　・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.

ふけい
　表記：婦警
　・「婦人警察官」の略.

表 5 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（見出し語）	定義文中の語
「～の略.」	婦警	婦人警官
	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
	加算	足し算
「～.」	ソビエト連邦	ソ連
	アメリカ合衆国	米国
	アルミ	アルミニウム

彙数に対して少ないと言えるが，目的とした常識的な国名の言い換え表現は獲得できており，誤った同義表現のペアは含まれていなかった．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．

3 名詞句の関係解析

本研究では，名詞句間の関係を解析するため，構文解析，および，橋渡し指示 (bridging reference) 解析を行う．

構文解析は，KNP(黒橋，河原 2007) を用いて行う．構文解析の結果，文節ごとの係り受け関係，および，連体修飾であるなど係り受け関係にある 2 文節がどのような関係にあるかが分かる．例えば，以下のような文があった場合，「立てこもる」が「事件」を連体修飾していることなどが分かる．

(4) 女性を人質に立てこもる事件があった．

橋渡し指示とは，(5) 中の「チケット」と「値段」の関係である．これらは直接係り受け関係にはないが，「チケットの値段」という意味となっている．橋渡し指示解析は自動構築した名詞格フレームを用いて行う (Sasano, Kawahara, and Kurohashi 2004)．橋渡し指示解析の結果，(5) 中の「値段」は「チケット」の値段という意味であることなどが分かる．

(5) 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．

4 共参照解析

4.1 文字列のマッチングを用いた共参照解析

本研究では，基本的な共参照解析システムとして，文字列のマッチングを用いた共参照解析システムを用いる．本節では，文字列のマッチングを用いた共参照解析システムについて説明する．

4.1.1 照応詞，先行詞として考える単位

共参照解析システムを構築するにあたり問題となるのが，照応詞，先行詞として扱う単位をどのようにするかである．特に，複合名詞句があった場合，その構成素のうちどの部分を照応詞，先行詞として考えるかが問題となる．まず，考えられるのが Iida ら (Iida et al. 2003) の基準である．Iida らは共参照解析を行うにあたり，照応詞を文節の主辞（最右の名詞自立語）に限定している．

(6) 携帯電話／PHS の利用に関するウェブ・アンケート調査を実施し，207 名から回答を得ました．調査内容は…

しかしながら，(6) のような文があった場合，「ウェブ・アンケート調査」と，「調査内容」の「調査」は同じ対象を指しており，カバレッジの大きな共参照解析システムの構築を目指す場合，主辞となっていない形態素が照応詞となる共参照関係も認識できることが望ましいと考えられる．そこで，本研究では，複合名詞の構成素すべてを照応詞の候補として考える．ただし，固有表現については結び付きが強いと考えられることから例外として扱い，固有表現の部分構成素は照応詞，先行詞として考慮しないものとする．

京都テキストコーパス (河原，黒橋，橋田 2002) では，複合名詞句の構成素も含むすべての自立語を照応詞，先行詞として扱っている．例えば，次のような文があった場合，後続する「ロシア軍」に含まれる「ロシア」および「軍」にはそれぞれ別々に，先行する「ロシア」，「軍」と共参照関係にあるというタグが付与されている．

(7) グロズヌイからの報道では，ロシア 軍は…．首都防衛はうまくいっており，ロシア 軍の戦車五十両を破壊したと発表．

しかしながら，複合名詞のある構成素が，先行する同表記の複合名詞の構成素と共参照関係にある場合，その複合名詞の他の構成素も対応していることは自明であると考えられる．例え

ば，(7) のような文があった場合，「軍」が同一の対象を指しているならば，「ロシア」が同じ対象を指していることは自明である．そこで本研究では，1つの文節に対してはより右側に出現した照応詞1つのみを解析の対象とする．

以上より，本研究における先行詞，照応詞として扱う基準は以下のとおりである．

- 文章内に出現したすべての名詞句，複合名詞句の構成素を先行詞の候補とする．
- 文章内に出現したすべての名詞句，複合名詞句の構成素を照応詞の候補とする．ただし，1つの文節に対しては，より右側に出現した照応詞1つのみを対象とする．
- 固有表現については例外として扱い，その部分構成素は先行詞，照応詞として考えない．

4.1.2 基本的な方針

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を伴なって出現す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・対象を指す表現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，同一，または，より簡潔な表現で表される場合が多いと考えられる．また，同一文章中に先行する文章中で出現した表現が同一，または，より簡潔な形で出現した場合は，それらは同一の内容・対象を指す可能性が高いと考えられる．

そこで本研究では基本的に，先行する文章中に出現した表現が同一，または，より簡潔な形で出現した場合に，それらが同一の内容・対象を指すと考える．ただし，指示詞や「同」に修飾されている表現については，先行する表現を照応していると考えた方が自然であるので，これらの語を伴なっていた場合も先行する表現を照応していると考える．また，固有表現は，修飾語によって限定されることはないと考えられるので，修飾語を伴なっていた場合も先行する同表記の固有表現を照応していると考える．以下では，同一，または，より簡潔な形で出現したと判断し，照応詞候補と先行詞候補が共参照関係にあると判断する基準を，共参照関係認定基準と呼ぶ．

4.1.3 共参照関係認定基準

文章中に出現した表現が，共参照関係にあると判断する基準，すわなち，照応詞候補が先行詞候補と，同一，または，より簡潔な形であると判断する基準として以下の2つの基準を用いる．ただし，いずれの場合も指示詞，および，「同」は考慮しない．

共参照関係認定基準 1: （文節内のみ考慮）

照応詞，先行詞候補を含む文節を比較し，照応詞候補を含む文節の照応詞候補以前の部分が，先行詞候補を含む文節に含まれている場合，同一，または，より簡潔であるとする．

共参照関係認定基準 2: （文節間の係り受けも考慮）

1の条件に加え，照応詞候補が他の文節から修飾されていない場合のみ，同一，または，

より簡潔であるとする．

例として，(8) のような文を考える．共参照関係認定基準 1 を使用した場合は，a，b，c の場合に，後続する「出場者」は先行する「出場者」と同一，または，より簡潔な表現だと判断し，共参照関係認定基準 2 を使用した場合は，a，b の場合のみ，同一，または，より簡潔な表現だと判断する．

(8) a. 会場に集まった出場者が…．出場者たちは…．
b. 会場に集まった出場者が…．同出場者たちは…．
c. 会場に集まった出場者が…．決勝に残った出場者たちは…．
d. 会場に集まった出場者が…．決勝戦出場者たちは…．

4.1.4 文字列のマッチングを用いた共参照解析のアルゴリズム

以上の方針に基づく，文字列のマッチングを用いた共参照解析のアルゴリズムを以下に示す．

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 文頭の文節から順に，すべての名詞句，および，複合名詞の構成素を照応詞候補とする．だだし，固有表現と解析された名詞句については，それ以上分割しない．
3 各照応詞候補について，以下の基準で先行詞を探し，先行詞が見つかった場合は，それらの照応詞候補，先行詞は共参照関係にあると判断する．ただし，1 文節中に複数の照応詞がある場合は，より主辞の近く（右側）に出現したものを優先する．
   (a) 先行する文章中から同一の表現を探す．
   (b) 照応詞候補が固有表現である場合は，より簡潔な表現であるかどうかを考慮せず，同一の固有表現があれば先行詞と判断する．
   (c) それ以外の場合は，照応詞候補がその表現と同一，または，より簡潔な形である場合，その表現を先行詞とする．先行詞の条件を満たす表現が複数あった場合は，照応詞候補の近くに出現したものを優先する．

4.2 名詞句の関係解析の利用

構文解析の結果，連体修飾関係にある 2 つの文節と，それらに含まれる自立語を含む複合名詞句があった場合，連体修飾されている名詞句と複合名詞句は同じ対象を指している可能性が高いと考えられる．例えば，(9) 中の「北海道北部」に連体修飾された「占領」と「北海道北部占領」や，(10) 中の「立てこもる」に連体修飾された「事件」と「立てこもり事件」は同一の対象を指していると考えられる．

そこで，照応詞候補を含む文節の照応詞候補以前の部分が，先行詞候補を含む文節に含まれていない場合であっても，含まれていない部分を原形に直したものが，先行詞候補を連体修飾

している文節の原形に含まれている場合，これらは共参照関係にあると考える．

(9) …，ソ連の当時の最高指導者スターリンが，日本の**北海道北部**の占領とともに，……
ソ連の北海道北部占領計画は既に知られているが…

(10) …女性を人質に**立てこもる**事件があった．今回の立てこもり事件について…

同様に，橋渡し指示解析の結果，先行詞候補と関係があると解析された表現の原形を補うことにより，照応詞候補が先行詞候補に含まれるようになった場合，これらは共参照関係にあると考える．例えば(11)のような文があった場合，橋渡し指示解析の結果，2文目に出現する「結果」がアンケートの結果であると認識され，3文目の「アンケート結果」と同一の対象を指していると解析できるようになる．

(11) 2006 FIFA ワールドカップ優勝国予想アンケートが行った．結果はブラジルがトップだった．アンケート結果の詳細は Web で見られる．

4.3 同義表現の利用

4.1.4節で説明したアルゴリズムでは，同義表現を用いた言い換えに対応できない．そこで，4.1.4節の3(a)において，先行する文章中から同一の表現を探す際に，自動獲得した同義表現も対象とする．この結果，以下のような文があった場合，「北大西洋条約機構」と「NATO」の間の共参照解析を認識できるようになる．

(12) 米国は北大西洋条約機構加盟国に対し，タリバンとの衝突が激化した南部地域への増派を求めており，7日からのNATO国防省理事会で主要議題になる見通し．

5 実験と考察

京都テキストコーパス，および，ウェブから集めた文章に京都テキストコーパスと同様の基準 (河原，笹野，黒橋，橋田 2005) で共参照タグを付与したウェブコーパスを用いて，共参照解析実験を行なった．京都テキストコーパスは，毎日新聞 322 記事 2098 文から成り，2870 個の共参照タグが付与されている．ウェブコーパスは，186 記事 979 文から成り，717 個の共参照タグが付与されている．

実験は，共参照関係認定基準 1，および，共参照関係認定基準 2 それぞれに対し，文字列のマッチングのみを用いた手法，それに名詞句の関係解析，自動獲得した同義表現，および，その両方を追加した計 4 手法を行った．また，共参照関係認定基準の妥当性を確かめるため，よ

り簡潔であるかどうかに関わらず，照応詞候補があった場合，先行する直近の同一の表現を先行詞と判断するという手法も用いた．すなわち，「大統領官邸」という表現の前に「首相官邸」という表現がある場合，「官邸」が同一の対象を指していると判断する．ただし，より長い表現間のマッチングを優先し，「首相官邸」より前に「大統領官邸」という表現があった場合は，「大統領官邸」を先行詞とする．結果を表6に示す．表6中のF値は，適合率と再現率の調和平均である．

先行する直近の同一の表現を先行詞と判断する手法と文字列のマッチングのみを用いた手法を比較すると，いずれの共参照関係認定基準を用いた場合も，僅かな再現率の減少で，適合率は大幅に上昇しており，照応詞を先行詞と同一，または，より簡潔な表現とするという制約が有効であることが確認できる．

共参照関係認定基準1を用いた場合と共参照関係認定基準2を用いた場合とを比較すると，共参照関係認定基準2の方が厳しい制約であるため，再現率が低下するかわりに，適合率が上昇している．F値に関しては，京都テキストコーパスを用いた実験では共参照関係認定基準1を用いた場合の方が，ウェブコーパスを用いた実験では共参照関係認定基準2を用いた場合の方が高くなっている．これは，新聞記事では比較的長い名詞句が多いため，同一の複合名詞句であれば同じものを指している場合が多く文節内のみを考慮すれば十分であるのに対し，ウェ

表6　共参照解析結果

	京都テキストコーパス			ウェブコーパス		
共参照関係認定基準	適合率 (%)	再現率 (%)	F値	適合率 (%)	再現率 (%)	F値
共参照関係認定基準1: (文節内のみ考慮)	72.2 (2191/3033)	76.3 (2191/2870)	74.2	68.6 (583/850)	81.3 (583/717)	74.4
＋名詞句の関係解析	72.0 (2209/3068)	77.0 (2209/2870)	74.4	68.1 (586/861)	81.7 (586/717)	74.3
＋同義表現知識使用	72.6 (2239/3086)	78.0 (2239/2870)	75.2	68.7 (586/853)	81.7 (586/717)	74.6
＋関係解析＋同義表現	72.3 (2257/3121)	**78.6** (2257/2870)	**75.3**	68.2 (589/864)	**82.1** (589/717)	74.5
共参照関係認定基準2: (文節間の係り受けも考慮)	78.1 (1946/2492)	67.8 (1946/2870)	72.6	82.5 (515/624)	71.8 (515/717)	76.8
＋名詞句の関係解析	77.6 (2004/2583)	69.8 (2004/2870)	73.5	80.5 (532/661)	74.2 (532/717)	77.2
＋同義表現知識使用	**78.4** (1995/2544)	69.5 (1995/2870)	73.7	**82.6** (518/627)	72.2 (518/717)	77.1
＋関係解析＋同義表現	77.9 (2052/2634)	71.5 (2052/2870)	74.6	80.6 (535/664)	74.6 (535/717)	**77.5**
先行する直近の同一の表現を先行詞と判断	57.4 (2251/3925)	78.4 (2251/2870)	66.3	56.2 (585/1041)	82.1 (585/717)	66.6

ブコーパスでは短い名詞句が多いため文節間の修飾関係も考慮する必要があるためだと考えられる．

同義表現を用いない場合と用いる場合を比較すると，同義表現を用いることにより適合率，再現率はともに上昇しており，自動獲得した同義表現を共参照解析に用いることは有効であると言える．表7に同義表現の利用により新たに共参照関係にあると解析された例を示す．共参照関係認定基準1を用いた場合，同義表現を用いることにより，京都テキストコーパスとウェブコーパス合わせて新たに56個がシステムにより共参照関係にあると解析されるようになった．そのうち51個が正しい解析となっており，新たに誤って解析されるようになったものは表7に示した「衛星」と「BS」など5個のみであった．また，51個中21個が国語辞典から抽出された同義表現であり，国語辞典から抽出された同義表現は，数は少ないものの共参照解析の性能向上に貢献していることが分かる．

一方，名詞句の関係解析を用いた場合，再現率は上昇したものの，適合率は減少しており，ウェブコーパスに対し共参照関係認定基準1を用いた実験ではF値も低下している．しかし，ウェブコーパスに対し，より高い精度となる共参照関係認定基準2を用いた場合は再現率は大幅に上昇しており，また，いずれのコーパスに対しても，もっとも高い精度が得られたのは同義表現，名詞句の関係解析の両方を用いた場合であることから，名詞句の関係解析を用いることも共参照解析にある程度有効であると考えられる．共参照関係認定基準1を用いた場合に，名詞句の関係解析の利用により新たに共参照関係にあると解析された例を表8に示す．

表8において，名詞句の関係解析を用いることにより新たに正しく認識できるようになったものは，それぞれ2回出現する「所感」，「結果」，「漁民」が「首相の所感」，「アンケートの結果」，「ベトナム系の漁民」と解析されたことにより，これらの間の共参照関係を認識できるようになった．一方，新たに誤って認識するようになったものは，それぞれ2回出現する「候補」，「燃料」が「連絡協議会の候補」，「核の燃料」と解析されることから，これらの表現が共参照関係にあると判断したものの，この場合は，それぞれ「有力」，「独自の」，また，「初回分の」，「代替」という異なる修飾語で限定されていることから，これらの表現は同一のものを指している

表7 同義表現の利用により新たに共参照関係にあると解析された例

新たに正しく認識できるようになったもの
- 少なくとも日本に生まれ育った在日韓国・朝鮮人を地方公務員として排斥する理由はない．
- …連合を支持基盤とする民社党や民主改革連合，…．民改連は夏の参院選で…
- 一方，朝鮮民主主義人民共和国は，…．今回，どういう形で行うかが注目されるが，北朝鮮の…
- 韓・チリの自由貿易協定批准が1年以上漂流しているなか，韓国の対中南米貿易が…

新たに誤って解析されるようになったもの
- 衛星通信，ケーブルテレビは将来どのようなシナリオで…現在五チャンネルのBSが…

とは言えず，誤った解析となっている．

続いて，1章で分類した照応詞と先行詞の関係ごとの傾向を調べるため，京都テキストコーパスから無作為に抽出した共参照タグ250個について，照応詞と先行詞の関係，および，システムが正しく認識できているか否かを調べた．結果を表9に示す．照応詞の表記が先行詞の表記に含まれている場合は高い再現率が実現できていることが確認できる．また，同義表現による言い換えとなっている場合は，出現数が少ないものの，ある程度高い再現率が実現されていると考えられる．その他に分類されたものは本システムでは原理的に解析できないため，22個すべてが解析できていない．その他に分類された例を(13)に示す．このような共参照関係は全体の約9%程度を占めており，より高い再現率をもつ共参照解析システムを構築するためには，これらの認識を行う必要があると考えられる．

表8　名詞句の関係解析の利用により新たに共参照関係にあると解析された例

新たに正しく認識できるようになったもの

- 村山富市首相は年頭の記者会見で，「創造とやさしさの国造りのビジョン」と題する所感を発表した．…村山富市首相が発表した「年頭所感」の要旨は次の通り．
- 昨年十二月実施した全衆院議員アンケートで，…．結果は現首相の「村山富市氏」を挙げた議員は全体の二八%．…アンケート結果を集計すると，…
- カンボジアの漁業は，イスラム系住民のチャム族とベトナム系の漁民が担ってきた．が，一昨年の総選挙を前に，ポル・ポト派がベトナム系漁民を襲撃．

新たに誤って解析されるようになったもの

- …独自候補擁立へ向け，都議と合同で「連絡協議会」を新年に発足させる．同協議会メンバーで，有力候補とささやかれる鳩山氏は…
- …初回分の燃料用重油五万トンを予定通り今月中に供与する姿勢を示した．米国は，北朝鮮が核兵器開発を凍結する見返りに，軽水炉建設と代替燃料の供給を約束．

表9　照応詞と先行詞の関係ごとの再現率

照応詞と先行詞の関係	再現率
1. 照応詞の表記が先行詞の表記に含まれている	86.9 (192/221)
2. 同義表現による言い換え	71.4 (5/7)
3. その他（クラスとインスタンス，上位語と下位語など）	0.0 (0/22)
合計	76.1 (197/250)

表 10 先行研究との比較

	適合率	再現率	F 値
村田ら (村田, 長尾 1996) (名詞のみの評価)	78.7 (89/113)	77.3 (89/115)	78.1
Iida ら (Iida et al. 2003) (主辞のみを対象)	76.7 (582/759)	65.9 (582/883)	70.9
本手法 (京都テキストコーパス)	72.3 (2257/3121)	78.6 (2257/2870)	75.3
本手法 (ウェブコーパス)	80.6 (535/664)	74.6 (535/717)	77.5

(13) a. 小選挙区に立候補するには現金三百万円か同額の国債証書を…

b. …ロシア軍は一日までの激戦で，首都グロズヌイを事実上制圧した模様だが，…．しかし，市内を完全に制圧するまでには，…

最後に，先行研究との比較結果を表 10 に示す．対象とする共参照の定義，および，使用しているコーパスが異なるため単純には比較できないものの，提案システムは，ある程度高い精度を実現していると考えられる．

6 関連研究

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．

6.1 規則ベースの手法

Zhou ら (Guodong and Jian 2004) 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．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．

村田ら (村田，長尾 1996) 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，複合名詞の構成素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．

6.2 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．これらの手法の多くは，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 (Soon et al. 2001) 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，決定木を用いて学習を行ない，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．

Ngら (Ng and Cardie 2002) はSoonらの手法を2つの点において改良している．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．

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 Iidaら (Iida et al. 2003) の研究がある．Iida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．Iida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．

7 おわりに

本稿では，まず，コーパスおよび国語辞典の定義文から同義表現の自動獲得を行った．続いて，獲得した同義表現，および，名詞句の関係解析結果を用いた日本語共参照解析システムの構築を行った．京都テキストコーパス，および，ウェブコーパスを使った実験の結果，同義表

現，および，名詞句の関係解析結果を用いることにより，共参照解析の精度は向上し，手法の有効性が確認できた．今後の課題としては，文字列のマッチングや同義表現による言い換えでは解析できないような共参照関係の認識が挙げられる．

参考文献

Guodong, Z. and Jian, S. (2004). "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.

Iida, R., Inui, K., and Matsumoto, Y. (2006). "Exploiting Syntactic Patterns as Clues in Zero-Anaphora Resolution." In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 625–632.

Iida, R., Inui, K., Takamura, H., and Matsumoto, Y. (2003). "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.

Kawahara, D. and Kurohashi, S. (2004). "Zero Pronoun Resolution based on Automatically Constructed Case Frames and Structural Preference of Antecedents." In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP-04)*, pp. 334–341.

Ng, V. and Cardie, C. (2002). "Improving Machine Learning Approaches to Coreference Resolution." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 104–111.

Nichols, E., Bond, F., and Flickinger, D. (2005). "Robust ontology acquisition from machine-readable dictionaries." In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-2005*, pp. 1111–1116.

Okazaki, N. and Ishizuka, M. (2008). "A Discriminative Approach to Japanese Abbreviation Extraction." In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*, pp. 889–894.

Sasano, R., Kawahara, D., and Kurohashi, S. (2004). "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.

Seki, K., Fujii, A., and Ishikawa, T. (2002). "A Probabilistic Method for Analyzing Japanese Anaphora Integrating Zero Pronoun Detection and Resolution." In *Proceedings of the 19th*

International Conference on Computational Linguistics, pp. 911–917.

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

Tokunaga, T., Syotu, Y., Tanaka, H., and Shirai, K. (2001). "Integration of heterogeneous language resources: A monolingual dictionary and a thesaurus." In *the 6th Natural Language Processing Pacific Rim Symposium*, pp. 135–142.

鶴丸弘昭，竹下克典，伊丹克企，柳川俊英，吉田将 (1991). "国語辞典情報を用いたシソーラスの作成について." 情報処理学会自然言語処理研究会 1991-NL-083, pp. 121–128.

酒井浩之，増山繁 (2003). "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.

西尾実，岩淵悦太，水谷静夫（編） (2000). 岩波国語辞典. 岩波書店.

村田真樹，長尾眞 (1996). "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, **3** (1), pp. 67–81.

河原大輔，黒橋禎夫，橋田浩一 (2002). "「関係」タグ付きコーパスの作成." 言語処理学会第8回年次大会発表論文集, pp. 495–498.

河原大輔，笹野遼平，黒橋禎夫，橋田浩一 (2005). 格・省略・共参照タグ付けの基準.

黒橋禎夫，河原大輔 (2007). "日本語構文解析システム KNP version 3.0 使用説明書." 京都大学大学院情報学研究科.

久光徹，丹羽芳樹 (1997). "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.

村上明子，那須川哲哉 (2004). "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.

上野友司，森辰則，木戸冬子，中川裕志 (2004). "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.

山本和英 (2002). "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.

田近洵一（編） (1997). 例解小学国語辞典. 三省堂.

略歴

笹野　遼平：2004年東京大学工学部電子情報工学科卒業．2006年同大学院情報理工学系研究科修士課程修了．現在，同大学院博士課程在学中．省略解析，照応解析の研究に従事．

黒橋　禎夫：1989年京都大学工学部電気工学第二学科卒業．1994年同大学院博士課程修了．京都大学工学部助手，京都大学大学院情報学研究科講師，東京

大学大学院情報理工学系研究科助教授を経て，2006年京都大学大学院情報学研究科教授，現在に至る．自然言語処理，知識情報処理の研究に従事．

(2008年2月6日　受付)
(2008年5月14日　再受付)
(2008年7月1日　採録)