自動獲得した知識に基づく統合的な照応解析

笹野 遼平　　河原 大輔　　黒橋 禎夫
東京大学大学院情報理工学系研究科
{ryohei, kawahara, kuro}@kc.t.u-tokyo.ac.jp

1　はじめに

ある表現が他の表現と同一の内容を指していることを同定する照応解析は計算機による意味理解において重要である。もっとも基本的な照応現象は、照応詞とその照応先である先行詞がともに文章中に明示的に出現し直接同一の内容を表す直接照応であるが、それ以外にも照応詞が省略されるゼロ照応、照応詞が間接的に先行詞を照応する橋渡し指示 (bridging) など様々な種類が存在する。例えば (1) の例では「値段」が「チケット」を橋渡し指示している。

(1)　チケットを買った。値段 (ノ:チケット) は 2000 円だった。

(2)　対露関係悪化を承知で「ロシア (=:露) 非難」に踏み切ったものとみられる。

このような照応解析を行う場合、様々な知識が必要となる。例えば、(1) の例において「チケットの値段」という関係を認識するには「値段」という語が何らかの「商品」の「値段」であるという知識が必要となる。(2) における「露」と「ロシア」は直接照応の関係にあるが、この関係を認識することは「露」と「ロシア」が同義表現であるという知識がなければ困難である。このように、高度な照応解析を行うためには、事前に有用な知識を獲得しておくことが必要となる。

また、これらの照応現象は相互に関連しており、統合的に解析を行うことで、より高精度な照応解析システムを構築できると考えられる。本稿では、まず自動獲得した同義表現を用いた直接照応解析を行い、橋渡し指示解析解析との統合を行うことにより、照応解析の高精度化を目指す。

2　同義表現の獲得

2.1　括弧表現を用いた同義表現獲得

まず、コーパス中に出現する括弧表現を用いた同義表現獲得を行った。以下に手順を示す。

表 1: 括弧表現を用いた同義表現獲得のための閾値

タイプ	双方向	同義表現とみなす条件
英字と英字	yes	頻度の相乗平均 > 3
以外のペア	no	頻度 > 50
カタカナとカタ	yes	頻度の相乗平均 > 4
カナ以外のペア	no	頻度 > 300
漢字で構成される語とその省略[1]	yes	頻度の相乗平均 > 1
	no	−

1. 括弧の中の表現 A と、その直前の同一文字種からなる文字列 B のペアの出現頻度を数える。
2. (a) 表 1 に示す 3 つのタイプに該当し、かつその出現回数が設定した閾値を越えたペアを同義表現とする。この際、B(A) に対する A(B) のように互いを入れ替えた表現が存在する場合の閾値は緩く設定する。

   (b) 表 1 に示す 3 つのタイプに該当しなかった場合でも、互いを入れ替えた表現が存在し、かつ出現回数の相乗平均が 30 を越える場合は同義表現とする。

毎日新聞 12 年分と読売新聞 14 年分、計 26 年分、約 2,600 万文中に出現した約 800 万個の括弧表現から同義表現の自動獲得を行った。結果を表 2 に示す。獲得できた同義表現は 1,588 ペアで、誤ったペアを獲得しないように閾値を設定したため、すべて正しい同義表現ペアであった。

2.2　辞書の定義文からの同義表現獲得

括弧表現を用いるだけでは抽出できないと考えられる「米国」と「アメリカ」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、国語辞典の定義文を用いた同義表現獲得も行った。国語辞典の見出し語ごとに以下の処理を行う。

[1] 一方の構成漢字が他方にすべて含まれている場合のみ可。

表 2: 括弧表現を用いた同義表現獲得の結果

タイプ	数	主な例
英字と英字以外のペア	1052	国連平和維持活動=PKO 北大西洋条約機構=NATO
カタカナとそれ以外	220	関税貿易一般協定=ガット 金融派生商品=デリバティブ
漢字とその省略	241	住宅金融専門会社=住専 動力炉・核燃料開発事業団=動燃
その他	75	朝鮮民主主義人民共和国=北朝鮮 二酸化炭素=CO2
合計	1588	

表 3: 国語辞典を用いた同義表現獲得

	主な例	
定義文のタイプ	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
	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
「〜。」	ソビエト連邦	ソ連
	アメリカ合衆国	アメリカ
	アルミ	アルミニウム

1. 定義文が「の略。」、「のこと。」で終わっている場合はその前の部分すべてを、それ以外の定義文については句点より前の部分すべてを取り出し B とする。
2. 取り出した B が「」で囲まれている場合、または、第一義でかつ B が辞書の見出しに含まれる場合、A と B を同義表現とする。

「例解小学国語辞典」と「岩波国語辞典」を用いて抽出を行った結果、 表 3 に示すような同義表現が 150 ペア抽出された。個数は多くないものの、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」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が抽出できた。

3 同義表現を用いた直接照応の解析

直接照応における照応詞と先行詞の関係は表 4 のように分類できる。このうち 1〜3 は比較的簡単な規則と、同義表現や固有表現認識技術を用いることにより認識を行うことが可能であると考えられる。

一方、4 については意味素性など様々な要素を考慮に入れなければならず、人手で規則を作成して解析を行うことは困難である。しかし、実際にはこのような

表 4: 直接照応の分類

1. 表層的な情報から認識が可能なもの (ex. 大統領官邸=同官邸)
2. 同義表現による言い換え (ex. 露=ロシア)
3. 照応詞が代名詞となっているもの (ex. 松井=彼)
4. その他 (文脈の理解が必要なもの) (ex.1995 年= 前年)

直接照応はあまり出現しないため、4 にあたる直接照応が解析できなくても十分に高精度な照応解析を行うことができると考えられる。そこで、本研究では 1〜3 の場合の解析を行うことを目指し、人手による規則に基づく直接照応解析を行う。

まず、複合名詞中のどのような語が照応詞、先行詞になり得るかを考える必要があるが、多くの場合は先頭形態素を含む部分列のみを考えれば十分であると言える。例えば「調査内容」という表現があった場合も「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。そこで次のような方針を立てる。

方針 1 複合名詞の先頭を含まない部分形態素列は照応詞、先行詞として考えない

同一の文章中に先行する表現と同じ表現や同義表現が出現した場合、特に修飾語などによって限定されていない場合は同一の内容を指していると考えるのが自然である。逆に何らかの修飾語によって限定されている場合は先行する表現と別の内容を指している場合が多いと考えられる。そこで、次のような方針を立てる。

方針 2 照応詞は修飾されないと考える

これら 2 つの方針、および同義表現知識に基づく照応解析のアルゴリズムを表 5 に示す。また、その際用いる修飾に関する条件を表 6 に示す。

4 橋渡し指示解析

橋渡し指示解析は自動構築した名詞格フレーム辞書を用いて行う。ここで名詞格フレーム辞書とは、「値段」おける「商品」のように、各名詞に必須的な要素を記述した辞書であり、コーパス中に出現する名詞句「A の B」の用例などを用いて自動構築を行った [2]。名詞格フレーム辞書や、固有表現認識結果、直接照応解析結果を用いた橋渡し指示解析のアルゴリズムを表 7 に示す。

[2]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
[3]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: 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 6 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、先行詞と認定する。(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[2](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[3](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)

表 6: 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。

表 7: 橋渡し指示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. 固有表現に含まれず、かつ、直接照応していない普通名詞を照応詞候補とする。
3. 文頭から順に各照応詞候補について以下の処理を行う。
   (a) 照応詞候補の名詞格フレームが存在しない場合は、照応詞でないとして処理を終了する。
   (b) 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、格フレームの用例との類似度を計算する。
   (c) 閾値を満足し、かつ最もスコアの良いものを先行詞とする。ただし、直接係り受け関係にある場合や、先行詞候補が主題である場合などは優遇する。

5 統合的な解析

表 5 のアルゴリズムは基本的に照応詞は修飾されないものとしている。このため、前方に出現した表現が修飾されて再び出現した場合には同一の内容を表していてもそのことを認識できない。

(3) 村山首相は年頭の記者会見で**所感**を発表した。首相が発表した 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、構文解析、直接照応解析、橋渡し指示解析を行った結果は図 1 のようになり、2 文目に出現する「所感」は「発表した」により修飾されていることから照応詞候補とならないため、1 文目の「所感」と同一の内容を指していることは認識できない。

しかしこの場合、橋渡し指示解析の結果、同一の「首相」を橋渡し指示していると解析されることから、いずれも「首相の所感」であることが分かる。このように、修飾されている要素であっても、橋渡し指示の結果、意味内容が補完されることにより同一性の認識を行うことが可能となる場合も存在する。

そこで、表記が同一で、橋渡し指示解析の結果、同一の対象を指示していると判断された場合は、直接照応の関係とするという規則を加えて解析を行う。

6 実験と考察

95 年の毎日新聞の記事に、直接照応に関するタグが付与された京都テキストコーパス Version 4.0[1] を用いて直接照応解析の実験を行った。使用した記事は 30 記事、合計 181 文で、これらの記事には解析対象

表 8: 修飾に関する条件ごとの直接照応解析の精度

修飾に関する条件	適合率	再現率	F 値
考慮しない	75.5 (231/306)	**80.5** (231/287)	77.9
主辞と同様に扱う	**84.1** (212/252)	73.9 (212/287)	78.7
主辞以外は修飾されないと考える	82.6 (223/270)	77.7 (223/287)	80.1
名詞格フレーム辞書を用いる	83.2 (223/268)	77.7 (223/287)	**80.4**

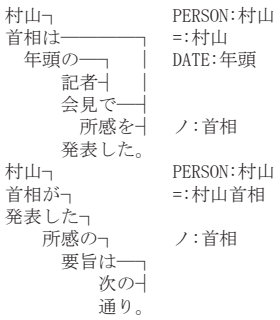

図 1: 直接照応解析、橋渡し指示解析の結果

とする直接照応タグ[4]は 287 個付与されていた。また、解析対象とする橋渡し指示を表すタグは 59 個付与されていた。

まず、表 8 に橋渡し指示解析の結果を用いなかった場合の直接照応解析結果を示す。修飾を考慮しなかった場合は再現率こそ高くなるものの、F 値は最も低いことから、修飾を考慮することの有効性が確認できる。また、修飾に関する条件としては名詞格フレームを用いた場合がもっとも良い精度を得られた。この名詞格フレーム辞書を用いた場合の照応の種類ごとの解析精度を表 9 に示す。表層的な情報から認識が可能なものに加えて、同義表現が必要なものについても高精度で解析できている。

表 10 に橋渡し指示解析の結果を示す[5]。自動構築した名詞格フレーム辞書は単独の名詞ごとに必須要素を記述しているため、単独で出現する名詞に対してのみ評価を行った。再現率については高い精度を得ることができたが、適合率は高い精度を得られなかった。これは、橋渡し指示の解析では、照応性の判断を名詞格フレーム辞書にのみ頼っているためだと考えられる。

続いて、修飾に関する条件として名詞格フレームを用い、さらに橋渡し指示の解析結果を用いた場合の精度を先行研究の精度とともに表 11 に示す。橋渡し指示解析の結果を使用することにより再現率、および F 値が向上していることが確認できる。また、使用しているコーパスや対象が異なるため単純な比較はできないものの、先行研究と比べても十分に高い精度で解析できていると言える。

先行研究において、村田らは名詞句の指示性を 86 個の規則を用いることにより推定し、さらに名詞の指示性を考慮した 9 個の規則を用いて名詞の同一性の解析を行っている [3]。飯田らの研究は機械学習を用いた手法で、分類器として SVM、学習の用いる素性としては語彙的な情報や、形態・統語的な情報、名詞句間の距離など計 30 あまりの素性を用いて解析を行っている [4]。

[4]京都テキストコーパスでは固有表現の一部となるような表現を先行詞、または照応詞とするようなタグも付与されているが本実験では評価から除く。また、「小泉首相」などの表現が出現した場合、「小泉」と「首相」の間にも=タグが付与されているが本研究では評価から除く。

[5]適合率と再現率の分子が異なるのは、再現率を評価するときは正解に含めないものの、システムが出力した場合は適合率の評価では正解として扱うタグが存在するため。

表 9: 照応の種類ごとの解析精度

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)
照応詞が代名詞	50.0	(4/8)
その他 (文脈理解が必要)	0.0	(0/19)
合計	77.7	(223/287)

表 10: 橋渡し指示解析の精度

適合率	再現率	F 値
31.0(45/145)	69.5(41/59)	42.2

7 おわりに

本稿では、まずコーパスおよび国語辞典から同義表現の自動獲得を行い、続いて同義表現を用いた直接照応解析や、名詞格フレーム辞書を用いた橋渡し指示の解析を行い、最後に橋渡し指示解析の結果を用いた直接照応解析を行った。新聞記事を用いた実験の結果、先行研究と比べても十分に高い精度を得ることができ、同義表現および橋渡し指示の解析結果を直接照応解析に用いる手法の有効性が示された。

参考文献

[1] 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, 2005.

[2] 笹野遼平, 河原大輔, 黒橋禎夫. 名詞格フレーム辞書の自動構築とそれを用いた名詞句の関係解析. 自然言語処理, Vol. 12, No. 3, pp. 129–144, 2005.

[3] 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.

[4] 飯田龍, 乾健太郎, 松本祐治, 関根聡. 最尤先行詞候補を用いた日本語名詞句同一性指示解析. 情報処理学会論文誌, Vol. 46, No. 3, pp. 831–844, 2005.

表 11: 橋渡し指示の結果を用いた直接照応解析の精度

	適合率	再現率	F 値
村田ら (参考)	78.7 (89/113)	77.3 (89/115)	78.1
飯田ら (参考)	76.7(582/759)	65.9(582/883)	70.9
橋渡し指示なし	83.2(223/268)	77.7(223/287)	80.4
橋渡し指示使用	83.2(227/273)	79.1(227/287)	81.1