日英対訳コーパスからのゼロ代名詞とその指示対象の自動抽出[1]

中岩浩巳　山田節夫
ＮＴＴコミュニケーション科学研究所

１．はじめに

自然言語では通常，相手（読み手もしくは聞き手）に容易に判断できる要素は，文章上表現しない場合が多い．この現象は，機械翻訳システムや対話処理システム等の自然言語処理システムにおいて，大きな問題となる．例えば，機械翻訳システムにおいては，原言語では陽に示されていない要素が，目的言語で必須要素になる場合，陽に示されていない要素の同定が必要となる．特に日英機械翻訳システムにおいては，日本語の格要素が省略される傾向が強いのに対し，英語では訳出上必須要素となるため，この省略された格要素（ゼロ代名詞と呼ばれる）の照応解析技術は重要となる．

日本語ゼロ代名詞の照応解析に関しては，従来から様々な手法が提案されてきているが[1][2][3][4]，翻訳対象分野を限定しない機械翻訳システムに応用することを考えると，解析精度の点や対象とする言語現象が限られる点，また，必要となる知識量が膨大となる点で問題があり，実現は困難である．これらの問題に対しては，照応解析条件として，用言の意味属性[5]，様相表現，接続表現を用い，これらを表現の持つ意味に応じて分類し，その代表的属性値に応じて照応要素を決定することによりこれらの問題を考慮にいれた，機械翻訳に適した照応解析手法が提案されている[6][7][8]．

しかし，これら従来から提案されている手法では，基本的に人間が照応解析のための規則を作成する必要がある．よって，網羅的な照応解析ルールを作成するためには，かなりの専門知識と労力が必要となる．さらに，解析対象となる分野に応じて，異なった照応要素を認定する必要があるゼロ代名詞が存在するので，分野に依存した照応解析ルールを作成する必要がでてくる．しかし，分野毎にこのルールを作成することは，このための労力や時間を考慮すると，実際的には実現不可能である．よって，このゼロ代名詞の照応解析ルールを効率的に獲得する手法の実現が望まれている．

自然言語処理システムにおける解析ルールの効果的な獲得のためには，従来から，既存のコーパスを用いて，コーパス中に現れる言語現象を分析し，分析した結果を基に解析ルールを抽出する手法が提案されている．ゼロ代名詞照応解析ルールの自動抽出に関してもいくつかの手法の提案がされてきており，それらは多くは基本的に解析対象となる言語のコーパスのみを用いている[9],[10]．しかし，解析対象言語のコーパスのみを使用する場合，その言語ではほぼ常にゼロ化される要素を補完するための規則を抽出することは困難である．また，解析対象となるタイプのゼロ代名詞の指示対象を決定するための情報を含む言語現象が，その解析対象文以外の文中に現れなければ，有効な解析規則を抽出できない．

このような問題を考慮にいれると，ゼロ代名詞の解析ルールの抽出に用いるコーパスとしては，解析対象の言語のみからなるコーパスではなく，解析対象の言語と他の言語の対訳コーパスを利用することが有望であると期待できる．特に，日本語と英語のように言語族が異なる場合には，省略現象が現れる傾向が異なるため，ある言語の文ではゼロ化されている要素が，その文と対訳関係にある別の言語の文では明記される場合が多々有り，その利用が有望である．

対訳関係にある文の集合である対訳コーパスから各種解析ルールを抽出するためには，それら集合から，対訳関係にある文対を抽出する技術，対訳関係にある文対から対訳関係にある単語・表現対を抽出する技術が重要となり，従来から様々な手法が提案されてきている．ゼロ代名詞照応解析ルールの抽出

[1] Automatic Identification of Zero Pronouns and their Antecedents within Aligned Sentence Pairs

という観点で考えると，対訳関係にある一方の文からゼロ代名詞を抽出し，他方の文からそのゼロ代名詞の指示対象を抽出する技術が必要となる．これに関連した技術としては，最近，対訳コーパス中の段落や図，表など文章単位での省略箇所を抽出する手法が提案されているが[11]，文中でゼロ化された箇所とそこに補うべき要素の抽出に関する手法の提案はされていない．

本稿では，このような目的を達成するための第１段として，１文対１文で対訳関係にある日英の対訳文からなる日英対訳コーパスから，日本語文中のゼロ化された箇所と，英語文中のそこに補うべき要素を抽出する手法を提案する．

２．対訳文中の省略とその指示対象

本節では，対訳関係にある日本語文と英語文を用いて，日本語文中の省略された要素と，英語文中の省略された要素に相当する要素の傾向について，日本語ゼロ代名詞と所有格的な表現のゼロ化にしぼって考察する．

２．１　日本語ゼロ代名詞の傾向

よく知られているように，日本語では，聞き手もしくは読み手が文脈や常識から容易に推測できる場合には，主語や目的語等の格要素は省略される場合がほとんどである．それに対し，英語では，明示される場合がほとんどである．例えば，

(1)　(φ-が) 本を読みたい

“I want to read a book.”

という表現では，希望の様相表現「たい」を伴うため，読み手もしくは聞き手は，特別な文脈がなければ，「本を読みたい」のは，話し手か書き手であることが容易に推測できるため，動詞「読む」のガ格である「私」が省略されている．しかし，このような場合でも，英語では，動詞”read”の主語として，”I”が明示される．このような，日本語では省略される格要素が，英語では明示される表現としては，下記の様な２種類が考えられる．

- 話し手又は書き手や聞き手又は読み手等のような直示的な(deictic)要素が格要素である場合．
- 同一文内中の要素や文章中の他の文の中の要素を指示するような照応的な(anaphoric)要素が格要素である場合．

この内で，前者の直示的な要素に関しては，英語では”I”や”you”のような人称代名詞が代名詞が使われる場合が多い．それに対し，後者の照応的な要素に関しては，英語では人称代名詞に加えて，”that”等の指示詞，”the company”等の定冠詞を伴う定表現，”each other”等の相互代名詞，”one”等が使われることが知られている[12]．よって，英語文中のこのような表現に着目することによってゼロ代名詞の指示対象候補を効果的に抽出できることが期待される．

２．２　所有格的表現のゼロ化傾向

日本語と英語の省略傾向の差が顕著なものに，日本語における「ＡのＢ」表現の中の「Ａの」に相当する，「の」を介した修飾を含む表現である所有格的表現の取り扱いがある．日本語では，例えば，「ＡのＢ」の日本語表現で，「Ａ」が直示的な(deictic)要素である場合に，省略される傾向が強い．例えば[13]，

(2)　(φ-の)鼻が痒い．

“My nose itches”

という表現では，その「痒い鼻」の持ち主が誰であるかの情報が日本語文では明示されていないが，それと対訳関係にある英語文では，所有代名詞”my”で”nose”を修飾することによって，読み手又は書き手の鼻であることを明示している．この日本語表現を分析すると，「痒い」という動詞が示す現象は，この「痒み」を経験をしている人しか「痒い」かどうかを認識できないので，「痒い」のガ格としては，話し手か書き手，もしくは話し手か書き手の身体の一部分であるという制約から，このゼロ化された所有格的表現の指示対象が話し手又は書き手と決められる．

このような，直示的な要素以外にも，２．１で示した照応的な要素による所有格的表現のゼロ化も行われる．例えば，

(3) 犬が(φ-の)尾を振る．

“A dog wags its tail.”

では，「振られる」尾の持ち主が「振る」のガ格である「犬」となるが，日本語文ではそれが省略され，英語では”A dog”を指示する所有代名詞”its”として明示されている．このようなゼロ化された照応的な要素の所有格的表現が英語で明示される表現としては，”-’s”を伴う所有格表現や，”of ..”や”in ...”等の前置詞を伴い名詞を修飾する表現等が考えられる．

日本語で明示されないこのような所有格的表現の

細かな分析によると[13]，<人間>と<鼻>の場合のように，一方の単語が他方の単語と"has a"関係などの直接的な関係を持つ場合が多い．よって，このような他の語と明らかな関係を持つ語であるかを判断することによって，英語で所有代名詞を生成すべきかをどうかが決められる傾向にある．

本節で議論した，日本語における所有格的表現は，日本語のみを考慮にいれると省略現象とは言えない場合も多い．しかし，日本語から英語のように他の言語へ機械翻訳する場合や，文中に暗に含まれる情報を意味理解する解析系を構築する場合は，ここで述べた所有格的要素の情報を補う技術は大変重要となり，自然言語解析システムの立場から見ると，これらは一種の省略現象であると言える．

3．自動抽出手法

2節での議論をもとに，本節では，1対1で対訳関係にある文対からなる日英対訳コーパスから，日本語文中のゼロ代名詞等の省略箇所と，その省略箇所に補完すべき英語文中の補完要素を抽出する手法について提案する．

3．1　省略補完要素抽出の基本規則

日本語文中の省略箇所および対訳英語文中の補完要素を抽出するために本手法で利用した基本的な規則について以下にまとめる．

日本語ゼロ代名詞候補の抽出規則1

日本語文の解析の結果，省略格要素と認定された要素を，日本語ゼロ代名詞の候補として抽出．

これは，例えば，日本語自動詞におけるガ格や，日本語他動詞におけるガ格もしくはヲ格に相当する格要素が省略された場合に，ゼロ代名詞候補として抽出する．

日本語ゼロ代名詞候補の抽出規則2

日英文中に対訳関係の認定された述語があり，その英語述語と格関係のある英語格要素中に，対訳関係のある日本語要素を持たない英語格要素があり，その英語格要素と同じ格関係をその日本語述語との間で持つ日本語格要素が存在しないばあい，その日本語格要素を日本語ゼロ代名詞の候補として抽出．

これは，例えば，日本語では自由格的な格要素であるが，英語では同じ格関係を持つ格要素として訳出されている場合に，ゼロ代名詞候補として抽出する．

ゼロ代名詞の英語指示対象候補の抽出規則1

英語文中の要素で，人称代名詞，指示詞，定冠詞を伴う定表現，相互代名詞，"one"表現を英語指示対象の候補として抽出．

これは，2．1節で議論した現象を取り扱うための規則である．

ゼロ代名詞の英語指示対象候補の抽出規則2

日本語ゼロ代名詞候補の抽出規則2で検出された英語格要素を英語指示対象の候補として抽出．

ゼロ代名詞の英語指示対象候補の抽出規則3

英語指示対象候補のうち，日本語意味解析の結果，日本語ゼロ代名詞に課せられた意味的制約を満たさないものがある場合，その要素を英語指示対象候補から除外．

ゼロ代名詞の英語指示対象候補の抽出規則4

日本語文中で英語指示対象候補が決まっていない日本語ゼロ代名詞が1箇所残っており，それと対訳関係にある英語文中に補完先となる日本語ゼロ代名詞が決まっていない英語指示対象候補が1箇所残っている場合，その残り同士を照応関係にある要素と認定．

日本語ゼロ化所有格的表現候補と英語補完候補の抽出規則1

英語文中に存在する単語を修飾する要素の中で，所有代名詞，"-'s"等の所有格表現，"of .."や"in ..."等の前置詞を伴い名詞を修飾する表現を日本語のゼロ化所有格的表現への英語補完要素として抽出．

これは，2．2節で議論した現象を取り扱うための規則である．

日本語ゼロ化所有格的表現候補と英語補完候補の抽出規則2

日英で対訳関係が認定された単語対が存在し，日本語側の単語にはその単語を「の」で修飾する要素は存在しないが，英語側の単語にはその単語を修飾する要素が存在する場合，その日本語単語に修飾する要素を日本語ゼロ化所有格的表現候補と認定し，その英単語を修飾する要素を英語補完候補と認定する．

これは，(2)の「鼻」に対する"my nose"や，(3)の「尾」に対する"its tail"に相当する現象を取り扱うためである．

3．2　システム構成

3．1節の基本規則に基づいて，日英対訳コーパスから日本語文中の省略要素と英語文中の補完要素

を抽出し，その結果をもとに，日本語の省略要素を補完するルールを自動生成するシステムの構成図を図１に示す．図の通り，入力された日英対訳コーパス中の対訳関係にある日本語文と英語文を解析し，その文対から対訳関係にある表現対を抽出する．次に，３．１節の規則を用いて，日本語省略箇所およびそれに補う英語補完要素を抽出する．そして，英語補完要素をもとに，日英対訳辞書等を用いて，日本語補完要素を抽出する．以上の結果をもとに，日本語解析結果を参考に，省略要素補完ルールを作成する．その後は，その新規作成された省略要素補完ルールを用いて，同じ対訳コーパスを対象に，新規作成された省略要素補完規則の有効性を検証しつつ，再度この学習過程を繰り返す．

この省略要素補完ルール自動抽出処理系は，日英機械翻訳システム ALT-J/E 中でその日本語解析系を利用して実装中である．図１の構成では，結果として，日本語省略箇所に対する英語補完要素も抽出できるので，これをそのまま流用することで，日英機械翻訳システムでの日本語省略要素の翻訳規則も抽出できる．

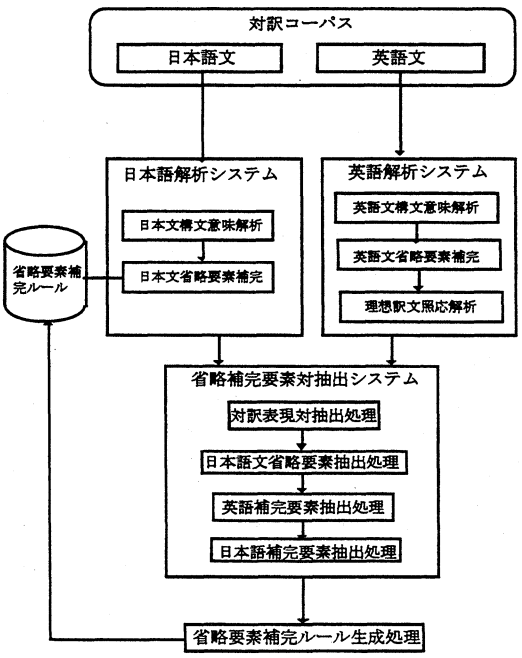

図１　省略要素補完ルール自動抽出処理の構成図

４．まとめ

本稿では，日英対訳コーパスを用いた，日本語文内のゼロ代名詞などの省略箇所と，英語文内に出現するそれに対する補完要素を自動的に抽出する手法の基本的な方式を提案した．本方式の有効性に関しては，現在，評価データの集計中であり，今後その結果をまとめて報告したい．また，省略要素補完ルール抽出方法の詳細に関しても実システムを用いてより詳細な検討を行っていきたい．

謝辞

1995 年から 1996 年までのマンチェスタ理工科大学(UMIST)滞在中，本技術に関して貴重な議論をしていただいた辻井潤一教授に感謝致します．

参考文献

[1]Kameyama,M.: A Property-sharing Constraint in Centering, Proc of ACL (1986).

[2] Walker, M. et al.: Centering in Japanese Discourse, Proc of COLING'90 (1990).

[3] Yoshimoto, K. : Identifying Zero Pronouns in Japanese Dialogue, Proc of COLING'88 (1988).

[4] 堂坂：語用論的条件の解釈に基づく日本語ゼロ代名詞の指示対象同定，情報処理学会論文誌, Vol.35 No.5 (1994).

[5] 中岩,池原：日英の構文的対応関係に着目した日本語用言意味属性の分類，情報処理学会論文誌, Vol.38 No.2 (1997).

[6] 中岩,池原：日英翻訳システムにおける用言意味属性を用いたゼロ代名詞照応解析，情報処理学会論文誌,Vol.34 No.8 (1993).

[7] 中岩,池原：語用論的意味論的制約を用いた日本語ゼロ代名詞の文内照応解析，自然言語処理, Vol 3 No.4 (1996).

[8] Nakaiwa, H. et al : Resolution of Japanese Zero Pronouns with Deictic Reference, Proc of COLING'96 (1996).

[9] Nasukawa, T.: Full-text processing: improving a practical NLP system based on surface information within the context, Proc of COLING-96 (1996).

[10] 村田，長尾: 用例や表層表現を用いた日本語文章中の指示詞・代名詞・ゼロ代名詞の指示対象の推定，自然言語処理, Vol.4, No.1 (1997).

[11] Melamed, I. D.: Automatic Detection of Omissions in Translations, Proc of COLING-96 (1996).

[12]今井，浅野: 照応と削除，大修館書店 (1990).

[13] Bond, F., et.al. : Possessive pronouns as determiners in Japanese-to-English machine translation, Proc of PACLING-95 (1995).