ゼロ代名詞の分類と先行詞同定の新しいベースラインについて

鈴木　久美
Microsoft Research
One Microsoft Way, Redmond WA 98052 USA
hisamis@microsoft.com

概要

本論文では、構文的な特徴に基づいた日本語ゼロ代名詞の分類と、それに基づいた、照応解析の新しいベースラインを提案する。まず、新聞記事と百科事典記事からなる文書集合に現れるゼロ代名詞を5つの構文タイプに分け、それぞれについて、頻度と先行詞同定の難易度を比較する。次に、この分類をもとに、言語学的な知見に基づいたベースライン・アルゴリズムを提案する。これは、「ゼロ代名詞という形で省略が可能なのは、「は」でマークされうる主題であり、格助詞でマークされた要素ではない」という久野[3]の分析に基づいている。この手法を構文解析と併用することにより、新聞記事データで67%の精度という、先行研究のベースラインを改善する照応解析結果が得られた。

1　はじめに

格要素の省略、いわゆるゼロ代名詞の使用は日本語では頻繁に観察される現象である。文の格要素を正しく認識することは、機械翻訳や自動要約などさまざまなアプリケーションにとって重要なことから、省略現象の解析については言語処理の立場からも、これまでにさまざまな提案がなされてきた。しかし、ゼロ代名詞の定義の違いや訓練・評価データの違いなどから、これまでの諸研究を比較し、現段階で得られている研究結果を包括的に評価することは難しい。また、ゼロ代名詞には、先行詞の同定が非常に容易なものと難しいものがあるが、これらはどうちがっているのか、それぞれどのような特徴があり、その分布はどうなっているのかなど、照応解析が対象としている現象自体はあまり研究の対象とされてこなかった。

このような現状をふまえ、本論文はまずゼロ代名詞をできるだけ広義にとらえ、その構文的な特徴に基づいた分類を試みる。この分類によって過去に提案されたアプローチを整理し、またカテゴリ別に最適の解析方法を検討するのが目的である。つぎに、この分類に基づいたベースライン・アルゴリズムを提案する。近年、機械学習などの洗練された手法を用いた照応解析のモデルも多く提案されているが、これらのモデルによる貢献度を正しく評価するためにも、単純で現実的なベースラインを設定することが望ましい。本論文で提案するベースラインは非常に単純で実装も容易であることから、現段階での現実的なベースラインになりうると考えられる。

2　ゼロ代名詞の分類と分布

先行研究においてゼロ代名詞と呼ばれているのは、用言の格要素が省略されているものである。格要素の格はほとんどの場合表層格としてみられており、深層格としてみられているものは少ない。以下に言及する先行研究も表層格要素を対象としている。表層格要素のうち省略されている格を同定する作業は、省略があるか否かの判断が、用言の語義に左右されるものであり、また語義が一意に決まりにくいため、人手でも難しい作業である。その中で、ガ格は例外的に省略の有無が一意に決まりやすく、またもっとも省略が起こりやすい格要素でもあるので、本論文はこの格のみを対象とする。なお、格要素の省略を用言（動詞、形容詞、名詞＋判定詞）に対するものだけでなく、サ変名詞の名詞用法に広げている研究もあるが[2]、ここでは用言の格要素にしぼって議論する。

2.1　ゼロ代名詞の分類

「ゼロ代名詞」(以下例文中ではφで表す)を「用言の格要素が表層で省略されているもの」と広く定義した場合、その先行詞との関係において、次のように分類できる。この分類は統語的な関係に基づいてはいるものの、純粋に言語学的なものではなく、言語処理での有益性を考慮に入れて行ったものである。

A. 主題化にともなう省略

先行詞は助詞「は」によって主題化された格要素であり、主題と用言の間に独立した別の節が存在しない。たとえば(a1)の例では、「したがい」の省略されたガ格がこれにあたる（以下の例では先行詞を下線で示す。また当該カテゴリに属する省略のみを明示する）。

(a1) クレオパトラ$_i$は[φ$_i$が]もうひとりの弟と結婚した。

B. 関係節化にともなう省略

関係節の中に省略があり、先行詞は関係節の係り先の要素である場合。次の例のφで明示されているゼロ代名詞はすべてこのカテゴリに属する。

(b1) [φ$_i$が]地方に追われたクレオパトラ$_i$は、[φ$_j$が]ポンペイウスを追って[φ$_j$が]アレクサンドリアに遠征してきたカエサル$_j$の支持をえて内紛に勝利をおさめた。

C. 節の並列化・従属化にともなう省略

Aoneら[5]で、「擬似ゼロ代名詞」(quasi-zero)と呼ばれているカテゴリに相当する。並列節あるいは従属節を含む文で、第一節以外の節のゼロ代名詞の先行詞が文頭の節の主語になっているもの、と定義されている。ここでは、文頭の節の主語が「は」あるいは「が」でマークされているもののみを含めた。次の例では「結婚する」の省略されたが格がこれにあたる。

(c1) 彼ら$_i$は王家の習慣にしたがい[ϕ_iが]結婚するはずだった。

D. その他の省略

いわゆる狭義のゼロ代名詞がこのカテゴリの中心である。これは、英語など、ゼロ代名詞を容認しない言語では代名詞が使用されるところに、日本語ではゼロ代名詞が使われているものと考えられ、上述のカテゴリと区別される。(d1)のように先行詞が同一文の中にないゼロ代名詞が、このタイプのゼロ代名詞の代表的な例である。ただし、本論文では言語処理の観点から見て、このタイプと区別しにくい省略現象もこのカテゴリに加えた。これらには、ゼロ代名詞と同一文にある先行詞が「は」「が」以外の助詞によりマークされているもの(d2)と、後方照応(d3)がある。

(d1) [ϕ_iが]オクタウィアとは離婚していた。

(d2) しかし、アントニウスの敗北がさけられなくなると、クレオパトラは艦隊をひきあげ、 [ϕ_iが]アントニウス$_i$もアレクサンドリアに逃げかえった。

(d3) [ϕ_iが]エジプトでしばらく彼女とすごしたのち、アントニウス$_i$は、カエサルの後継者オクタウィアヌス(のちのローマ皇帝アウグストゥス)の姉オクタウィアと政略結婚するためローマにもどった。

E. 名詞句先行詞のない省略

外界照応しているもの(たとえば、例(e1)のひとつめの省略)や、先行詞が名詞句以外のもの(同ふたつめの省略)をこれに含めた。

(e1) 18世紀後半には[ϕ_iが]ひとりでさす雨傘はまだめずらしく、1778年にイギリスの商人ハンウェーが雨傘をさしてロンドンの街をあるいて$_j$[ϕ_jが]話題をよんだ。

表1は日本語ゼロ代名詞の照応解析の先行研究のうちいくつかをこの分類をもとにまとめたものである。研究の対象とされているカテゴリ(研究に明示されているもののみならず、例文などから演繹したものもふまえている)に✓をつけて示してある。表から、先行研究のゼロ代名詞の定義にはばらつきがあることがわかる。

2.2 ゼロ代名詞の分布

次に、上述の分類をもとに、ゼロ代名詞の分布をコーパスを使って調べてみた。コーパスは、人手でゼロ代名詞とその先行詞(先行詞がない場合あるいは名詞句以外の場合はその旨)を付与した毎日新聞記事とエンカルタ百科事典の記事各10記事、計20記事(365文)を使用した。表層ガ格のゼロ代名詞は634個ふくまれていた。図1はその分布を示したものである。

	Aoneら [5]	河原ら [2]	関ら [4]	飯田ら [1]
A.主題化		✓	✓	✓
B.関係節化				
C.並列・従属化	✓	✓	✓	✓
D.狭義のゼロ	✓	✓	✓	✓
E.先行詞なし		✓		

表1: ゼロ代名詞の分類と先行研究[1]

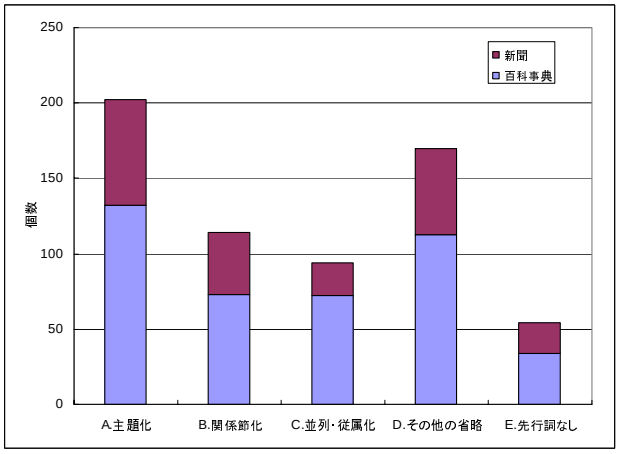

図1: ゼロ代名詞の分布

図1をみると、Aの主題化による省略が全体の約3分の1を占め、Dのゼロ代名詞が約4分の1、これ以外のカテゴリが約10%から20%となっている。とりわけ注目に値するのは、Aの「主題化による省略」とCの「並列・従属節化による省略」という、構文的に先行詞を同定しやすいカテゴリで分布の約半分を占めていることである。次節で提案するベースライン・アルゴリズムはこの分布の特徴を考慮している。

3 ベースライン

3.1 アルゴリズム

本論文で提案するのは、複雑なルールや学習機を使わない、ゼロ代名詞の照応解析のベースライン・アルゴリズムである。このベースラインの設計に、上述の分類から得られる知見を取り入れた。すなわち、同一文中に先行詞があり、統語構造の同定が比較的容易であると考えられるカテゴリ、Aの「主題化による省略」とCの「並列・従属節化による省略」は、構文解析タスクの一部とみなし、

[1] 飯田ら[1]は後方照応を考慮にいれていない。また河原ら[2]と関ら[4]はニ格とヲ格も対象に含めている。

特別な解析は行わない。Bの「関係節化による省略」は、関係節中の省略された要素が実際に関係節化によるものなのか、あるいはDのゼロ代名詞にあたるのかは処理が難しいが、これも構文解析の一部とみなし、特別な処理は行わないこととした。これに対し、Dのゼロ代名詞については、このカテゴリの先行詞同定には明確に定義できる統語構造が存在しないため、図2に述べるようなこのタスクに特化したごく単純なヒューリスティック・ルールを導入した。このようにゼロ代名詞の照応解析というタスクを構文解析に依存するタスクとそれ以外のタスクに分け、それぞれを別に扱っていることが、本ベースラインの特徴である。

ここで提案するベースラインの解析のプロセスを図2に示す。行番号2から6は上位のルールの条件が満たされなかったときにのみ下位に移るようになっている。行番号1から4までは、Aの主題化、Cの並列・従属節化、Bの関係節化による省略の処理であり、ここで使用した構文解析システム[10]の既存のルールをそのまま使用している。

1. 入力文を構文解析し、ゼロ代名詞を特定する
2. Aの「主題化による省略」の構文構造がみたされたとき、主題を先行詞に同定する
3. Cの「並列・従属節化による省略」の構文構造がみたされたとき、文頭節の主題を先行詞に同定する
4. ゼロ代名詞が関係節中にあるとき[2]、先行詞を関係節の係り先の名詞句に同定する
5. 同一文中に「は」による主題があるとき、これを先行詞に同定する
6. 同一文中に「は」による主題がないとき、図3の方法で設定された主題を先行詞に同定する

図2: ベースライン・アルゴリズム

行番号5と6がDのゼロ代名詞の先行詞同定のベースライン・アルゴリズムにあたるものである。このベースラインの特徴は、従来のように、ゼロ代名詞の先行詞が同一文にはなく、先行文にあった場合、その先行名詞句を直接ゼロ代名詞の先行詞とみなすのではなく、実は省略されているのは主題であり、その主題をゼロ代名詞の先行詞とみなす、という点にある。以下の例でNPaをゼロ代名詞が参照している名詞句とすると、従来のアプローチでは、(a)のように省略要素と先行詞のペアを直接見つける方法がとられてきた。これに対し、ここでは(b)のように、まずゼロ代名詞が存在している文の主題を想定し、その主題をゼロ代名詞の先行詞とみなすのである[3]。

(a) … NPa …。… [φ が] …

(b) … NPa …。[φ は] … [φ が] …

この見方は、ゼロ代名詞という形で省略が可能なのは、「は」でマークされうる主題であり、格助詞でマークされた要素ではない、という久野[3]の分析に基づいている。このように各文に主題を想定することによって、ゼロ代名詞の先行詞を探すという問題は、すべて同一文章の中から先行詞を探す、という問題に還元できるのである。

省略された主題の復元は、現在は図3のような単純なヒューリスティックによっている。

1. 当該文に「NPは」があれば、それを主題に設定する（複数個の「NPは」が存在するときは最初のものを使用する）
2. 当該文に「NPは」が存在しないとき、第1番目の文では、「NPが」を主題に設定する
3. それ以外の文では、ひとつ前の文の主題を、当該文の主題に設定する

図3:主題の設定

各文に主題のような談話的な構造を想定するのは、センタリング理論の中心的な考え方でもあるが、現段階のベースラインは、順序つきリストやセンタリング推移(centering transition)などの考え方は考慮にいれていないことから、あくまでベースラインとみなすのが適当である[4]。

3.2 評価

図4はベースライン・アルゴリズムの精度を2節で述べたコーパス上で評価したものである。まず、Aの「主題化による省略」は、現段階でも90%以上の高い精度を示している。これはこのタイプの省略が構文解析の結果からほぼ確実に同定できるということである。Cの「並列・従属化による省略」については、現段階での精度はAよりも多少低いが、基本的に同じことが言える。このカテゴリの精度がAよりも低くとどまっているのは、複数の節からなる文の係り受け解析の精度が落ちることが主要な理由として考えられる。同じ理由でBの「関係節化にともなう省略」の精度も改善の余地がある。これらのカテゴリの精度は、構文解析の精度をあげることによって、先行詞同定の精度も自動的にあがることが期待できる。なお、日本語の関係節の研究については、[6]など言語処理の観点から見たものも多くあるが、関係節とその係り先の名詞句の関係を

[2] 使用した構文解析システムでは、実際の条件はこれよりも多少複雑である。

[3] 例(b)では省略された主題を「φは」で表しているが、厳密には主題は談話構造の概念であり、「NPは」と同義ではない。ただしこのベースライン・アルゴリズムは、図3に述べるように「NPは」を使用して主題を設定している。

[4] 英語の代名詞の照応解析においても、センタリング理論を単純化したモデルの有効性が報告されている[9]。

対象にしており、関係節内の省略を対象にしたものではない。また関係節化による省略は表1からもわかるように通常、照応解析の研究の対象外となることが多いが、関係節内の省略はそれが関係節化によるものなのか、あるいは狭義のゼロ代名詞なのか、処理上峻別が困難な場合も多いため、関係節内の省略は照応解析に含めて考えるほうが現実的と思われる。

これらの構文的なカテゴリに対し、Dのゼロ代名詞におけるベースラインの精度は、新聞記事で33%、百科事典記事で53%となっている。解析が失敗した例には、主題の特定の失敗に起因するものが数多く見られる。これは「NPは」がいつも主題を表しているとは限らないのに、現時点ではこれを主題としたためである。Eの「先行詞なし」は、現在のベースライン・アルゴリズムに組み込まれていないため、精度は0%である。

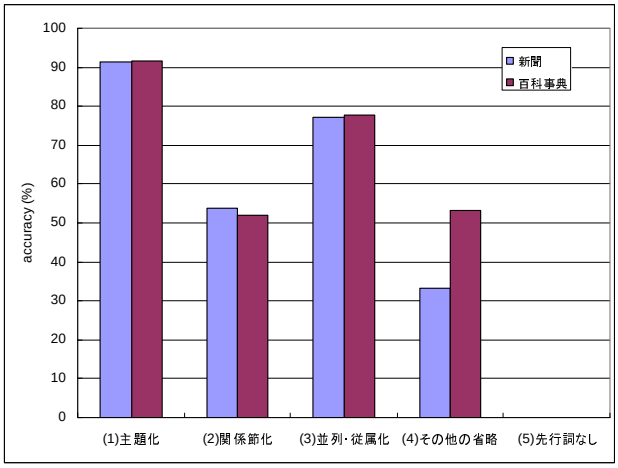

図4:ベースラインの評価

このような、構文解析と、ゼロ化した主題の回復による単純なヒューリスティックを組み合わせたベースラインの精度を、これまでに提案されている手法の精度と比較することは可能だろうか。直接的な比較は難しいが、あえて飯田ら[1]に報告されている結果と比較してみた。飯田らは、機械学習を使った最新のモデルに加え、センタリング理論を応用したNariyama[7]のルールに基づいたモデルと、Ngら[8]の手法を日本語に応用・実装したモデルの精度を、新聞データを使って比較している。対象としているゼロ代名詞のカテゴリは、A、C、Dである。飯田らの評価の条件に近づけるため、本ベースラインの精度は、解析の対象をA、C、Dのカテゴリにしぼり、新聞データ10記事のみを使って評価した[5]。この結果、本ベースラインの精度は67.11%であった。これはNariyamaのモデル(~46%)を上回り、Ngらの決定木によるモデル(~69%)と同程度である。飯田らのモデルの精度は75.1%と報告されている。

現在提案されている日本語のゼロ代名詞の照応解析には、ルールの条件や機械学習に用いる素性などに構文解析の結果が幅広く利用されていることから、本論文で提案した、構文解析に単純なヒューリスティックを組み合わせたアルゴリズムは、現段階での現実的なベースラインとして考えることができると思われる。

4 おわりに

本論文ではゼロ代名詞の分類と、それにもとづいたベースライン・アルゴリズムを提案した。今後の照応解析の精度の向上のためには、構文解析の質の向上と共に、現段階では精度の低い、構文構造に依存しないゼロ代名詞の先行詞同定に絞って精度向上をめざすことが有益と考えられる。後者に関しては、主題の省略を想定して解析を行うことの有益性も示唆した。これらの結果が今後、機械学習などの手法においても、学習問題や素性を設定する際に利用されていくことを期待している。

参考文献

[1] 飯田龍、乾健太郎、松本裕治. 文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定. 情報処理学会論文誌. Vol.25-3: 906-918. 2004.

[2] 河原大輔、黒橋禎夫. 自動構築された格フレーム辞書に基づく省略解析. 言語処理学会7回年次大会発表論文集. 2001.

[3] 久野暲. 日本語文法研究. 大修館書店. 1973.

[4] 関和広、藤井敦、石川徹也. 確率モデルを用いた日本語ゼロ代名詞の照応解析. 自然言語処理. Vol.9-3: 63-85. 2002.

[5] Aone, C. and Bennett, S.W. Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies. In *Proceeding of ACL*, 1995.

[6] Baldwin, T. Making Sense of Japanese Relative Clause Constructions. In *Proceedings of 2nd Workshop on Text Meaning and Interpretation*, ACL 2004.

[7] Nariyama, S. Grammar for Ellipsis Resolution in Japanese. In *Proceedings of TMI*, 2002.

[8] Ng, V. and Cardie, C. Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of ACL*, pp.104-111. 2002.

[9] Strube, M. Never Look Back: An Alternative to Centering. In *Proceedings of ACL*, pp.1251-1257. 1998.

[10] Suzuki, H. Phrase-Based Dependency Evaluation of a Japanese Parser. In Proceedings of LREC, pp.863-866. 2004.

[5] ただしこの評価に使ったコーパスに含まれるゼロ代名詞の数は149であり、飯田らの使用した約500(2781の5分割の交差検定)に比べて小さい。