ゼロ連体格代名詞の自動検出システム

磯江　健史
広島市立大学
情報科学部
isoe@nlp.its.hiroshima-cu.ac.jp

竹井　光子
広島市立大学
情報科学研究科
yamuram@nlp.its.hiroshima-cu.ac.jp

相沢　輝昭
広島市立大学
情報科学部
aizawa@its.hiroshima-cu.ac.jp

1　はじめに

日本語の代表的な結束表現（文と文のつながりを保証する言語手段）であるゼロ代名詞は，日本語談話理解の鍵となる要素の一つである。これまで，機械翻訳などのための照応解析を中心に重ねられてきたゼロ代名詞研究では，その対象は主として「ゼロ連用格代名詞」（ガ格，ヲ格などの省略）であった。しかし，日本語談話において結束関係を形成するゼロ代名詞には，以下の例[1]が示すように，これに加えて「ノ格の省略」がある。

> ［（女）ガ］［（男）ヲ］待っていると，男が帰ってきた。［（男）ガ］紙コップを持っている。［（男）ノ］表情に変化はなかった。

本研究では，この「ノ格の省略」に焦点を当て，その存在メカニズムを解明するとともに検出手法を整備し，システムを実装，その評価を行った。

2　ゼロ連体格代名詞と ATN

名詞には，その意味を完成させるために「～の」という情報を必要とする種のものがある。例えば，前述の談話例の「表情」は，「誰の」という情報を本質的に要求する性質を持っている。そして，その情報は文脈から類推可能な場合しばしば省略される。そこで，この元々「A の B」の形で完結する表現のうち省略された「A の」の部分，すなわち「ノ格の省略」を「ゼロ連体格代名詞」と呼ぶことにする。そして，「A の」という情報を要求するか否かは名詞 B の属性によって決定されると仮定し，「ノ格」の情報を必要とする名詞を Argument-Taking Noun (ATN) と呼ぶことにする。さらに，「ゼロ連体格代名詞」を検出することは，談話中に単独で現れた ATN を検出することと同義ととらえ，ATN の検出をシステムの目的をする。

ATN という概念は，言語学の先行研究において「関係名詞」，「相対名詞」，「一項名詞」などの名称で論じられてきた。これらの研究を踏まえ，より網羅的にコーパス調査した Yamura-Takei (2003) の分析結果（以後，予備調査と呼ぶ）および，それを利用した提案アルゴリズム（以後，基本システムと呼ぶ）を本研究の基盤とする。

予備調査では，コーパス中に現れた「（省略された A）の B」の名詞 B が ATN であると仮定し，その「（人手により補完された A）の B」の表現を収集し，島津ら (1986) の「A の B」の意味関係分析にしたがって分類した。さらに，各分類グループの B の名詞の特徴を示すものとして「日本語語彙体系（NTT/岩波書店）」にある名詞の属性を割り当てた。この予備調査によって得られた名詞 B の特徴を，以後「ATN 属性」と呼ぶことにする。

3　ゼロ連体格検出システム

検出手法としては，形態素情報および構文情報を利用した言語学的知見に基づく規則から成るアルゴリズムを採用する。「ゼロ連用格代名詞」は，節中の動詞の結合価情報と顕在格要素との照合を見ることによって検出を実現させた (Yamura-Takei et al. 2002)。

「ゼロ連体格代名詞」の検出では，候補となる名詞の属性を前述の ATN 属性と照合することをシステムの中核とする。システム開発は，まず予備調査に基づく「基本システム」を実装し，「実験用コーパス[2]」において評価実験，

---

[1]　星新一『車内の事件』「エヌ氏の遊園地」講談社文庫 (1971) より

[2]　日本語学習用教科書 5 冊分中の読解教材，全 66 テキスト分，人手によって検出されたゼロ連体格代名詞 366 個 を含む。

誤り分析を繰り返すことにより，精度向上に必要な改良を加える方法をとった。

最終的に作成した「改良システム」は，(1) ATN 候補の抽出準備，(2) ATN 候補の絞り込み，(3) ATN 選別，(4) 出力，の４つのモジュールから成る。各モジュールの詳細を説明する。

3.1 ATN 候補の抽出準備モジュール

入力テキストは，形態素解析および係り受け解析[3]を経て，係り受け情報を持った形態素列として，次のモジュールに渡される。

3.2 ATN 候補の絞込みモジュール

ATN の候補は裸名詞（修飾部分を持たない名詞）とする。候補を裸名詞に限定したのは，予備調査において人手で ATN と判断された名詞のほとんどが修飾部分を持たない裸の名詞として現れていたからである。

さらに，裸名詞でありながら ATN 候補として扱わない方がよい表現を振るい落とすために，次の 2 つのフィルタを設定した。

3.2.1 熟語フィルタ

「目が覚める」の「目」など，ATN と判定すべき意味属性の名詞でも慣用的表現として使われている場合には，ゼロ連体格代名詞表現と考えるより熟語として認識する方が自然であると考えた。それらの表現内の名詞は，熟語フィルタによって，裸名詞検出の段階で振るい落とすことにした。「目」は，後述の ATN 意味属性によって ATN と判定されるはずであるが，熟語表現内に現れた場合のみ，フィルタによって振るい落とされることになる。

熟語フィルタには，熟語辞書などを参考に収集した 1003 項目を登録した。

3.2.2 不定名詞フィルタ

ATN はゼロ連体格代名詞の存在によって，先行文脈とのつながりを持つ。例えば，「（男の）表情」は，先行文脈の要素（男）と照応関係にある。したがって，その名詞表現は不特定の対象を示すものではなく，ある特定の対象を指示しているはずである。つまり，ATN は「定性」という特性を持つことになり，「不定性」の表現は検出対象から排除しなければならいないことにある。（定・不定）冠詞というシステムを持たない日本語は，「定性・不定性」を示すマーカーは名詞句に付加されず，両者とも裸名詞として現われる。そこで，村田ら (1996) を参考に，「不定性」を示す表層的な情報を用いて，本来なら ATN 候補の裸名詞として検出される名詞から「不定名詞」を削除するフィルタを追加し過剰検出を防いだ。

不定名詞と判定する規則としては，「名詞＋と＋は（例：留学とは）」など７つを採用した。この「留学」という名詞は，後述の ATN 文法属性によると ATN となるが，この表現内では，フィルタによって振るい落とされる。

3.3 ATN 選別モジュール

裸名詞であることを前提に，２つのフィルタによって絞り込まれた ATN 候補は，ここでデータベース（日本語語彙体系）にアクセスし，ATN 属性との照合によって， ATN か NON_ATN かの選別が行われる。

3.3.1 ATN 属性

予備調査を基に規定した ATN 属性には，文法属性と意味属性の２種がある。文法属性，意味属性の順に照合を行い，どちらかで照合を見た場合に ATN と判定する。

ATN と規定する名詞の文法属性（品詞情報）は以下の通りである。

表 1：ATN 文法属性

	品詞（群）	例
1	サ変動詞型名詞	計画
2	転成名詞	長さ
3	形式名詞	他
4	数詞＋接尾	一割
5	形容動詞語幹＋接尾	危険性

次に意味属性のチェックを行う。当初，日本語語彙体系の意味属性木構造（2,715 ノード，最大 12 段）のうち，「人間<親族関係>」，「動物<部分>」，「数量」など比較的木構造上位の 8 つのノードを ATN 意味属性として規定していたが，改良実験の結果，見直しの必要性

[3] 茶筅および南瓜を使用した。

が明らかとなった。　主として，過剰検出の改善のため，親ノードから子ノード，孫ノードへの規定属性の細分化を，過少検出の改善のために<部分>を表す属性などの追加を行った。この修正の結果，ATN と規定する意味属性は，38 ノードとなった。

「日本語語彙体系」の単語体系は，元々，本研究で意図する ATN 属性，すなわち名詞の統語的な特性（自立性：必須の項をとるかどうか）や特殊な意味属性（相対性，関係性），による分類を意図して作成されたものではない。よって，ATN 検出のためにいくら注意深く属性を取捨選択しても，過剰・過少検出は避けられず，属性による選別には限界がある。検出誤りを減らすためには他の視点からの選り分けの規則が必要だとして，以下の２つの補助機能を付加した。

3.3.2　特定表現辞書

名詞には，ほとんどの使われ方において NON_ATN であるが，ある特定の表現の時のみ ATN となるものがある。例えば，「国」は ATN 属性の名詞と規定していないが，「国へ（に）帰る」という表現中に現われた場合にのみ，ATN と判定することにした。登録した特定表現の総数は 9 項目である。

3.3.3　補完辞書

ATN 属性は，その中に含まれる名詞が ATN である確率が高いものを選定しているが，選定されていない属性の中にも ATN となる名詞がわずかに含まれている場合がある。評価実験の過程で，ATN 属性とすれば過剰検出の可能性が高いが，その中に頻度の高い ATN 名詞が含まれている場合には，一つ一つ補完辞書に登録し，過少検出を防いだ。補完辞書に登録されている名詞は，評価実験の中で抽出された「住所」，「教え子」など 49 項目である。

3.4　出力モジュール

検出された ATN の前には「ゼロ連体格代名詞」が存在するはずである。よって，出力は，省略された「ノ格」を明示する形式になっている。出力例を図 1 に示す。

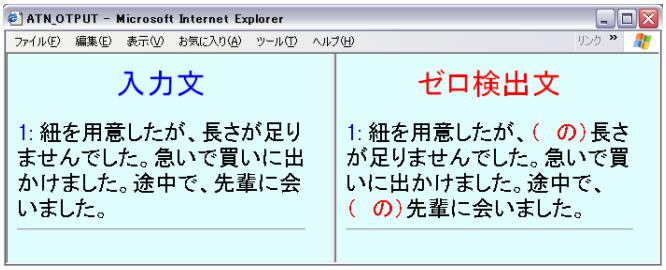

図 1：出力例

3.5　システムの概要

「改良システム」の概要を図式化すると図 2 のようになる。網掛けの部分が評価実験を経て改良を加えた部分である。

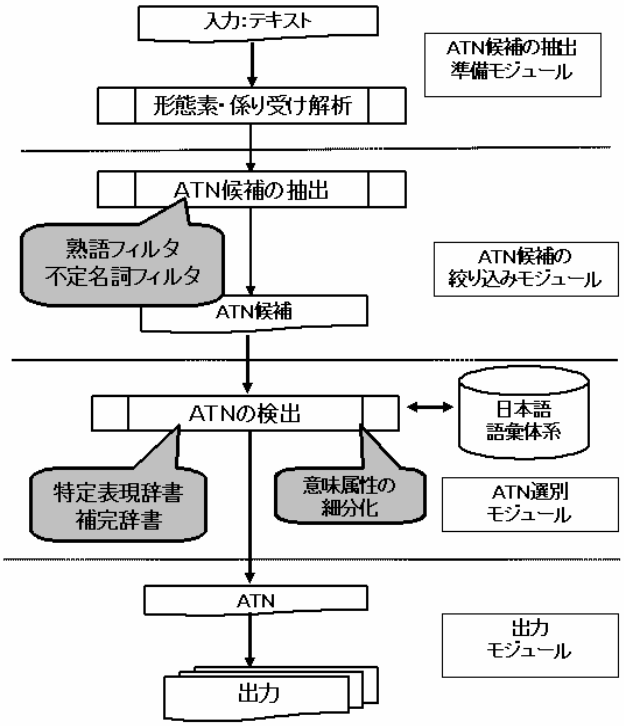

図 2：改良システムの概要

4　システムの評価

4.1　検出精度

評価は，予め人手によって検出された ATN との比較によって行った。まず，改良実験に用いた「実験用コーパス」による検出精度の変化を表 2 に示す。

表 2：基本・改良システムの比較[4]

	再現率	精度	F 値
基本システム	58.73 %	58.24%	0.58
改良システム	86.43%	84.10%	0.85

再現率，精度ともに 25% 以上の上昇がみられ，改良実験によって適用した規則や属性の見直しが有効に働いていることがわかる。

次に，「改良システム」の評価として，「評価用コーパス[5]」上でシステムを実行した結果は，表 3 に示す通りである。

表 3：改良システムの評価

	再現率	精度	F 値
改良システム	73.74%	70.19%	0.72

4.2 考察

改良システムは，再現率，精度ともに約 7 割であった。3 割の誤り例を分析した結果，その 9 割が文脈情報を必要とするエラーであった[6]。よって，ATN 属性を規定し，形態素，係り受けなどの表層情報からの経験則を利用した手法としては，これが限界と考えられる。ATN 属性そのものが，動詞の結合価と同様，文脈によってゆれの生じやすい概念であることからも，文脈情報を組み入れることが必須であろう。

5 まとめと今後の課題

本研究では，予備調査で示された ATN 検出手法の実装を行い（基本システム），評価実験を繰り返すことにより ATN 属性の見直しや補助機能の追加などを行った（改良システム）。

改良システムの評価を行った結果，再現率・精度ともに約 7 割の結果を得ることが出来た。語彙レベルの ATN 属性による選別を核として，統語レベルの情報による補助機能を追加した提案手法は，ある程度有効であると言えるであろう。しかしながら，更なる精度向上のためには，誤り分析の結果からも分かるように，文脈情報の利用が不可欠である。

本システムは，応用先の一つとして日本語読解支援システムを想定している。そのため，先行研究の「ゼロ連用格代名詞検出システム（ゼロ・ディテクター）」(Yamura-Takei et al. 2002) と統合を行った。その結果，2 種類のゼロ代名詞を検出できるようになり，1 章で示した談話例のような両タイプのゼロ代名詞により形成される結束関係を視覚的にとらえやすくなった。

［　が］［　を］待っていると，男が帰ってきた。［　が］紙コップを持っている。［　の］表情に変化はなかった。

このゼロ代名詞の明示機能によって，結束関係の理解を目的とする教育支援システムとしての効果が期待できる（詳細は，竹井ら，2004）。

[4] F 値は，再現率と適合率の相対的重みを 1 として計算してある。

[5] 実験用コーパスとは別の日本語学習用教科書内の読解教材 15 テキスト分。人手によって検出されたゼロ連体格代名詞（正解）を 99 個含む。

[6] 例えば，「先生」という名詞は，先行文脈の「前田先生」を指示している場合と(NON_ATN)，先行文脈の「茂君」を受けて「（茂君の）先生」と照応している場合(ATN)とに，文脈によって分かれる。

参考文献

島津明，内藤昭三，野村浩郷．1986．助詞「の」が結ぶ名詞の意味関係の解析．計量国語学会　第 15 巻第 7 号．pp. 250-260.

竹井光子，磯江健史，相沢輝昭．2004．第二言語習得理論におけるインプット強化と自然言語処理技術．言語処理学会第 10 回年次大会発表論文集.

村田真樹，黒橋禎夫，長尾真．1996. 表層表現を手がかりとした日本語名詞句の指示性と数の推定．自然言語処理　３巻４号 pp. 31-48.

Yamura-Takei, M., Fujiwara, M., Yoshie, M. and Aizawa, T. 2002. Automatic linguistic analysis for language teachers: the case of zeros. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-02)*, 1114-1120.

Yamura-Takei, M. 2003. Approaches to zero adnominal recognition. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03), Student Research Workshop*, 87-94.