

センター理論による日本語談話の省略解析

田村 浩二
富士通奥村 学
北陸先端科学技術大学院大学
情報科学研究科

[概要]

本稿では、センター理論を用いて複文を含む一般の談話の省略を扱う手法について述べる。センター理論は、代名詞・ゼロ代名詞の照応を解析する手法の1つであり、日本語の省略解析では現在もっとも一般的な手法である。しかし、単文など構造が簡単な文にしか適用できないなどの問題があった。

本研究では、複文は単文+接続助詞に分解して談話を単文の列に変換し、その単文の列に対してセンター理論を適用して省略の補完を行なう。その際、接続助詞の機能を考慮した制約条件を用いて先行詞候補の絞り込みを行なう。

[キーワード] 談話, 照応, 省略, センター理論

Japanese ellipsis resolution based on centering theory

TAMURA Kouji
FujitsuOKUMURA Manabu
School of Information Science,
Japan Advanced Institute of Science and Technology

Abstract

In this paper, we present a method of resolving complex Japanese ellipsis using centering theory. Centering theory is one of anaphora resolving methods, and a popular method of Japanese ellipsis resolution. However, it is impossible to deal with complex sentences.

We regard complex sentences as a series of simple sentences, and apply centering theory to each sentence. We restrict candidates of antecedent using constraint derived from conjunctive postpositional particle.

Key Words discourse, anaphora, ellipsis, centering

1 はじめに

日本語の文章では、主語、述語、目的語などの文の要素が省略されることが多いが、文の意味を理解するためには省略された要素を特定する必要がある。省略には、先行詞が文中に明示されていないものと、明示されているものがある。先行詞が明示されていない省略を補完するためには日本語に関する様々な知識や常識が必要であるが、先行詞が明示されている省略は、省略箇所的前後を調べれば省略された要素を特定することができる。

現在、省略補完に関する研究のほとんどは、先行詞が明示されている省略についてである。そのような研究の1つにセンター理論がある。センター理論とは、文中で話題の中心になっているものほど省略されやすい、という性質を利用して省略を補完する理論である。センター理論は、大規模な知識を必要としない、言語依存性が低い、計算機上で実現容易であるなどの利点を持っているが、従来は単文など構造が簡単な文にしか適用できなかった。

そこで、本研究ではセンター理論を用いて複文を含む一般の談話の省略を扱う枠組を示す。センター理論を使って複文の省略解析をするために、センター理論に対して以下の2つの拡張を行う。

- 述語を複数含むような文があった場合、その文を「単文」に分割して処理する。単文とは述語を1つだけ含む文のことで、元の複文では節に相当する。
- 接続助詞の情報を使って、先行詞候補の絞り込みと優先度の決定を行なう。接続助詞をその機能によって3つに分類し、各分類ごとに補完の手法、先行詞の優先度決定法などを定義する。

最後に、本研究で提案した手法を計算機上に実装し、実際にどの程度の省略が補完可能なかを検証する。

2 センター理論

センター理論 (Centering Theory) は英語の代名詞の照応関係を決定する手法として Grosz らによって提案された [4]。亀山は、英語の代名詞を日本語の省略に置き換えることでセンター理論を日本語の省略解析に適用し、更に、センター理論だけでは説明できない省略に対応するため、属性共有制約 (property sharing constraint) を導入した [1]。

2.1 センター理論の基本規則

センター理論では、各発話は Cf (Forward-looking Centers) と Cb (Backward-looking Center) という2つの構造を持つ。Cb と Cf の直感的な説明を以下に示す。

Cb 文中で一番注目されているもの。前の文から引き続き話題になっているもの。

Cf 文中の名詞句を先行詞になりやすい順番に並べたもの。先行詞になりやすい順番は日本語では以下のように定義されている。

トピック > 視点 > 主語 > 目的語 > その他

視点 (Ident) とは、授受の動詞など視点性がある動詞の必須格のうち、話者の視点があるものことである。久野の共感と同じ概念である [5]。

亀山の属性共有制約では、以下の2つの基本規則を用いて省略の補完を行う。

- 現在の発話の Cb が前の発話の Cb と同じならば、現在の発話の Cb は省略されているはずである
- そのような Cb の省略が連続する場合、それらのゼロ代名詞は文法属性 (主語、目的語など) を共有すべきである

2.2 省略補完処理のアルゴリズム

以下に属性共有制約を用いた省略補完処理の概略を示す。

1. 一文読み込む。
2. 文中に省略があれば補完する。
 - 先行詞の候補は直前文の Cf である
 - 先行詞は Cb や文法属性が変わらないものを優先する
3. Cf, Cb を決定する

2.3 センター理論の問題点

センター理論+属性共有制約では以下のような場合にはうまく処理することができない。

(1) 同じ文法属性の要素が複数ある場合

Cf中の名詞句の優先順位は、トピック>視点>主語>目的語>その他、であると決められているが、同じ文法属性の要素(例えば主語)が2つ以上出てきた場合の順番は決められていない。そのため、このような場合にはCfを適切に設定することができない。

(例)

太郎が次郎と遊んでいると 三郎がやってきた。

(2) 文法属性が異なる先行詞

連続する省略で文法属性が異なる場合、補完に失敗することが多い

(例)

(a) 太郎は一生懸命勉強していた。

(b1) そのため、次郎が(φに)声をかけても

(b2) (φが)気がつかなかった

属性共有制約では、文法属性の変わらない先行詞を優先するため、(c)で次郎を補完してしまう。

3 センター理論の拡張

2.3節で述べた問題点を解決するために、センター理論に以下のような拡張を行なう。

● 複文の分割処理

複文には複数の述語が含まれているが、その述語の直後で複文を切って複数の文に分割し、分割後の文にセンター理論を適用する。

● 接続助詞による候補の絞り込み

述語に付随する接続助詞の情報を用いて先行詞候補の絞り込みおよび優先度の決定を行なう。

3.1 複文の分割処理

2.3節で述べたように、センター理論では複文がうまく扱えない。そこで、複文が入力された場合には、複文を述語の直後で切って複数の文に分割する。本稿では、分割後の文のことを単文と呼ぶことにする。それぞれの単文を独立した文とみなしてセンター理論で処理することにより、単文も複文も同じように処理できるようになる。

(分割前)

(a) 太郎が次郎と遊んでいると三郎がやってきた。

(分割後)

(a1) 太郎が次郎と遊んでいると

(a2) 三郎がやってきた

このように処理すると、1つの文に主語が複数存在するということがなくなるため、2.3節で指摘したような問題は起こらない。さらに、これにより同一文内の照応も扱えるようになる。

3.1.1 先行詞候補の拡張

センター理論では先行詞の候補となるのは直前文のCfである。この直前文とは、複文を処理の単位としていた場合のことであるので、複文を単文に分割して処理するようにすると、直前の単文のCfだけでは候補が不十分になってしまう。

そこで、1文当りの単文数や先行詞が実際に何文前にあるかなどを調査した結果から、4文前のCfまでを先行詞候補にすることにした。

3.2 接続助詞

3.2.1 接続助詞の分類

3.1節で複文が入力された場合には単文に切って、それぞれ独立した文とみなして処理することにした。しかし、もともと1つの文であったものを分割して処理しているため、うまく補完できないことも多い。

そこで、元々単文である文と複文を分割した文を比較したところ、複文の方には述語の後に接続助詞がついていることが多いということがわかった。接続助詞は、その前後の単文の関係を表しているため、接続助詞の持つ意味を省略の補完処理に利用することで解析精度の向上が期待できる。接続助詞の情報を省略の補完に利用する研究は、吉本によって行われている[6]。

接続助詞を、その前後の単文での主語の一致・不一致という点に注目して3つに分類した(表1)。分類や命名に関しては吉本の研究を参考にしている。

A類 接続助詞の前後で主語が必ず一致するものをA類の接続助詞と呼ぶ。

B類 接続助詞の前後でどちらか一方の主語が省略されている場合、明示されている主語と省略されている主語が一致しないことが多い接続助詞をB類の

表 1: 接続助詞の分類

		性質	例
A 類	前/後略	接続助詞の前後で	ながら, て たり, つつ
	両方省略	主語が一致する	
B 類	前/後略	接続助詞の前後で	ても, ので けれど, ば
	両方省略	文脈による	
C 類	前/後略	文脈による	が
	両方省略		

接続助詞と呼ぶ。両方省略されている場合は前後の文脈により一致したり一致しなかったりする。

C類 接続助詞の前後での主語の一致・不一致が文脈によって決まるものをC類の接続助詞と呼ぶ。

3.2.2 接続助詞による制約

A類の接続助詞による制約

A類の接続助詞の前後では、主語が必ず一致するという強い統語的制約が働いている。そのため、接続助詞の前の主語が決まると、接続助詞の後の主語も自動的に決まってしまう。これは、センター理論や属性共有制約の文脈的なヒューリスティクスよりも強い、統語的な制約である。そこで、A類の接続助詞の直後の補完では属性共有制約などのヒューリスティクスは使わないことにし、主語は前文と同じものを補完する。

B類の接続助詞による制約

B類の接続助詞の前後では、どちらかの主語が省略されている場合、省略されている主語と明示されている主語が一致しないことが多い。つまり、B類の接続助詞を含む節の名詞句は、その直後の節の省略の先行詞になりにくいということである。そこで、B類の接続助詞を含む節の名詞句は、省略補完の際に先行詞候補として優先しないことにする。

C類の接続助詞による制約

C類の接続助詞の前後では、主語の一致・不一致は文脈によって決まるので、特別な制約は設けずに通常の文と同じように処理する。

接続助詞による制約を使った例を以下に示す。

(例)

1. 太郎は遅刻をして(A)
2. (ϕ が) 授業に遅れたが(C)
3. 先生も遅れたので(B)
4. (ϕ が) 怒られなかった。

2行目の主語が省略されているが、ここは直前文の接続助詞がA類の接続助詞「て」であるため、A類の接続助詞の制約により1行目の主語=2行目の主語となり、 ϕ =太郎と補完される。2行目にはC類の接続助詞「が」があるが、C類には制約がないため ϕ =太郎を補完しても矛盾するようなことはない。

4行目の主語の省略は、通常ならば直前文の先生が優先されるが、直前文にはB類の接続助詞「ので」があるため先生は優先されず、1行目の太郎が補完される。

4 実験

3節で述べたセンター理論の拡張の有効性を確かめるため、属性共有制約と本研究での提案手法を計算機上へ実装し、例文を処理させて正答率を比較する実験を行った。実験の条件は以下の通りである。

- 入力される文は複文を単文に分割したものである
- 入力文は正しく構文解析されている
- 補完の対象は各述語の必須格のものである
- 連体修飾節内の省略、先行詞が文中に明示されていない省略は補完の対象としない(全省略必須格の約30%)

実験の結果を表2に示す。

5 考察

表2の1行目は、入力文の複文を単文に切ってそのまま属性共有制約で処理した結果である。そのままでも約70%が正解になるが、この手法では不正解のうち候補無しのものが多いのがわかる。候補無しとは、正解の先行詞が先行詞候補にない場合、すなわち先行詞が直前文にない場合である。この状態で優先度などを工夫しても正答率は上がらないため、正答率向上のためには先行詞候補を増やす必要がある。

表 2: 実験結果

手法	正解	不正解		正答率
		候補有	候補無	
属性共有制約	224	10	80	71.3 %
属性共有制約 + 4 文前	214	70	30	68.2 %
属性共有制約 + 4 文前 + 接続助詞	251	32	31	79.9 %

表 2 の 2 行目は、入力文の複文を単文に切り、先行詞の候補を 4 文前から取るようにしたものである。候補を多く取ることにより候補なしの不正解は減少したが、正解数も減少している。正解数が減少したのは、先行詞候補の優先度をトピック > 主語としているためである。つまり、直前文にトピックではない正解の先行詞があり、それより前の文にトピックがある場合、トピックを優先してしまふため正答率が下がったのである。このように、ただ単に候補の数を増やしても正答率はあまり上がらない。増えた候補から正解のものを的確に取り出す必要がある。

表 2 の 3 行目は、先行詞を 4 文前から取るようにし、さらに接続助詞による優先・制約を加えたものである。接続助詞の制約を加える前に比べて正答率が約 10 % 上昇した。これは、主に A 類の接続助詞の前後での解析誤りが減少したためである。B 類の接続助詞は 1 文章中に 2 ~ 3 個しか使われていないので、この制約による顕著な効果は見られなかった。

5.1 複文への対応

本研究では複文を処理する手法として、複文を単文に切って処理するという手法を用いた。表 2 の結果からわかるように、最高で約 80 % の正答率をあげている。この結果から、複文を単文に分割し個別にセンタリングで処理するという手法でも省略解析が可能である事がわかる。

よって 2.3 節で示した問題点 (1) はこの手法によって解消される。

5.2 接続助詞による制約

接続助詞による制約を加えた事により、ただ単に文法属性が同じものや近い文のものを優先するのではなく、

節の意味を考慮した補完が行えるようになった。そのため、2.3 節の問題点 (2) で示したような、隣接する単文の省略で文法属性が異なる場合でも正しく解析できるようになった。

5.3 エラー連鎖

センター理論を使って省略の補完を行なうと、エラー連鎖という連続した誤りがしばしば起こる。これは、ある補完誤りにより間違った Cf, Cb、すなわち間違った先行詞候補が作成され、その影響でそれに続く文が誤りになってしまうというものである。この逆に、ある補完誤りの影響でそれに続く文が正解になってしまうということもある。

今回の実験では、エラー連鎖が原因の不正解は全体の 30.2 % (不正解 63 個のうち 19 個)、連鎖が原因の正解は全体の 1.2 % (正解 251 個のうち 3 個) であった。

5.4 その他の不正解箇所

エラー連鎖以外の原因で不正解になったのは以下のような場合である。

- 先行詞が 5 文以上前にある場合
先行詞が少し離れた場所にある場合は、たいてい先行詞と省略の間に短いエピソードが挿入されている。センター理論は局所的な話題の連続性を前提に処理しているため、このような省略を処理するためにはセンター理論以外のグローバルな談話情報を利用するような枠組みが必要である。
- 形式名詞や連体修飾節中の名詞が先行詞の場合
形式名詞や先行詞になることが少ないので、文中に現れても先行詞候補にしていない。そのため、形式名詞が先行詞になると必ず補完に失敗する。ただし、このような照応はまれである。
連体修飾節中の名詞句は優先度が低いので、先行詞になると失敗することが多い。このような省略を正しく処理するためには、連体修飾節と被修飾名詞の意味関係などを処理する枠組みが必要である。

6 今後の課題

- 意味的情報の利用
今回は実装を簡単にするために、語の意味的情報はほとんど使わなかった。しかし、不正解になった

ものの中には、語の意味を考えることにより正解になるものが少なくない。

● 他の方式との比較

本研究で提案した方式の有効性を客観的に評価するためには、既存のセンター理論や他の省略補完アルゴリズムとの比較を行うことが有効である。しかし、センター理論を拡張した他の方式で、複文をどのように扱っているかが明らかでなかったため、今回は他の方式との比較は行わなかった。

7 おわりに

本論文では、センター理論を用いて複文を含む一般の談話の省略を扱う枠組を示した。述語を複数含む文は、述語をちょうど1つだけ含む単文に分割し、センター理論を使うことにより解析可能であった。

先行詞候補の名詞句は直前文だけから取るのではなく4文前まで遡って候補にし、さらに、接続助詞をその前後の主語の一致/不一致という視点から3つに分類し、その情報を補完処理に使うことにより、精度を大幅に改善することができた。

本論文で提案した手法を計算機上に実装し、例文を処理させた結果、約80%の省略が処理可能であった。

参考文献

- [1] Megumi Kameyama, "A Property-Sharing Constraint in Centering", 24th Annual Meeting of the Association for Computational Linguistics, pp.200-206, 1986.
- [2] Marilyn Walker, Masayo Iida, Sharon Cote, "Japanese Discourse and the Process of Centering", Computational Linguistics, Volume20, Number 2, pp.193-232, 1994.
- [3] Shingo Takada, Norihisa Doi, "Centering in Japanese: A Step Towards Better Interpretation of Pronouns and Zero-Pronouns", COLING 94, 1994.
- [4] Barbara J. Grosz, Aravind K. Joshi, Scott Weinstein, "Providing a Unified Account of Definite Noun Phrases in Discourse", 21st Annual Meeting of the Association for Computational Linguistics, pp.44-50, 1983.
- [5] 久野章, "談話の文法", 大修館書店, 1978.
- [6] 吉本啓, "談話処理における日本語ゼロ代名詞の扱いについて", 自然言語処理 56-4 (1986.7.25), 1986.