E-050

動詞の意味情報を用いた代名詞の指示対象の推定法

Determining the Referent of a Pronoun Using Semantic Information of Verb

上條 敦史† 石川 勉†
Atsushi Kamijo Tsutomu Ishikawa

1 はじめに

我々は、自然言語インターフェースのコンサルティングシステムやWeb情報等の電子化文書を利用した質問応答システムについて研究している。ここでは、自然言語文を一定の知識表現に変換し、それを用いて推論することが不可欠となる。この場合、一文ごとの知識変換だけでなく、文中の代名詞の指示対象の特定（照応解析）が重要な課題となる。このような照応解析に関してはこれまで多くの研究がなされ、例えば、村田らは名詞の指示性、修飾語、所有者に関する特性に着目した手法[1]を、また、杉村らは意味解析の情報を利用した手法[2]を提案している。

本報告では、従来から用いられていた重要な規則の他に、新たに動詞の意味情報を利用した指示対象推定法について提案する。

2 利用辞書およびツール

本手法は、動詞に関するいくつかの意味情報を使用し指示対象の推定を行う。この意味情報としては、EDR電子化辞書[3]から得た動詞の深層格情報および独自に開発した単語の類似性判別ツール[4]から得た動詞間の関連性情報を使用する。以下に、これらツールについて説明する。

1) EDR電子化辞書

主要な動詞について表層および深層の格に関連する情報を記述した辞書である動詞共起副辞書（動詞総数は15,225）、各概念とその概念識別子で構成される単語辞書および各概念の上位下位の関係を示した概念体系辞書を用いる。

2) 類似性判別ツール

概念ベース[5]（以下、GB）および日本語語彙体系[6]（以下、NGT）のシソーラス情報を利用した、概念（単語）間の意味の類似度（0〜1）を算出するツールである。

GBは国語辞書中の見出し語を概念とし、各概念についての語義文中の独立語を属性、その出現頻度を属性値とし、これらをベクトル表現したもので、基本的にはtf･idfの考え方に基づいて構築されている。この類似性判別能力は、NGTやEDR等のシソーラスよりも平均的には優れていることが確認されているが[7]、機械的に作られているため、人間の直感と合わない類似度が算出されることもある。このため、表記ツールでは類似度をGBとNGTのシソーラスの両方で算出し、その差が一定値以上の場合には、それらを統合して最終的な類似度としている。

このツールに対して、各概念が”よく似ている”から”まったく関連がない”までの5段階で評価したところ、”ある程度関連がある”の類似度の平均値は0.173となった。したがって、ここでは類似度が0.1以上となった場合、各概念に関連性があるとする。

3 指示対象特定法

前方照応のみを考慮するものとし、対象とする代名詞以前の文を形態素・構文解析（それぞれ茶筌[8]、南瓜[9]を利用）し、その結果から指示対象の候補となる名詞を獲得する。獲得した候補名詞に対して複数の規則により評価値を与える。これら規則には、従来から指示対象推定に用いられている代名詞と候補名詞との距離、性別、数の関係性および文間の主題の連続性等の情報の他に、後述する動詞の意味情報を用いる。最後に規則ごとの評価値を統合し、指示対象を特定する。以下、それぞれの処理について述べる。

3.1 指示対象の候補となる名詞の獲得

出現した代名詞の前文N個から全ての名詞を取得し、指示対象の候補とする。ただし非自立名詞や代名詞の型と一致しない名詞は候補から除外する。具体的には、代名詞が指示代名詞だった場合、人称名詞は候補から除外し、逆の場合は人称名詞だけを候補とする。例えば、「林檎がテレビの上にある。太郎はそれを食べた。」の場合、「林檎」「テレビ」「上」「太郎」が取得されるが、「上」と「太郎」が除外され「林檎」と「テレビ」が候補となる。

3.2 指示対象の特定規則

i) 直前名詞のスコア付け

代名詞と距離が近い名詞程、高い点（評価値=$\frac{1}{\text{距離}}$）を与える。また、距離については、文節を単位とすることも考えられるが、ここでは単純に各候補名詞と代名詞の近さの順を距離とする。

ii) 性別の判定

代名詞や候補名詞の性別の判定を行い、その一致をチェックする。これには、NGTの一般名詞や固有名詞のカテゴリを用いる。

iii) 数の一致

代名詞、候補名詞がそれぞれ単数、複数のどちらを表すかを調べ、一致した候補名詞に評価値を与える。また、単複の判断は「等、ら、方」などの接尾語が付加していた場合に複数と、それ以外は単数とみなす。

iv) 主題のチェック

主題は、その文が何について述べているのかを示すものである。主題となる候補名詞は話の中心であることから、代名詞等の指示対象になることが多いと考えられる。

主題となりうる名詞に係る助詞には「は」「が」があり、「は」は話し手の事柄に対する係わりのありようを表すもので、話の中心になることが多い。一方、「が」は事柄のありようのみが述べられることが多い[10]。従って、候補名詞の係助詞が「は」「が」の場合には主題の連続性を考慮し評価値を付

† 拓殖大学工学部情報工学科

与するが、その重みは「は」>「が」と設定する。

v) 格関係のチェック

代名詞が出現した文の動詞とその表層格を用い、動詞共起副辞書から該当する共起パターンを調べる。表層格部分にある概念識別子を取得し、候補名詞の概念識別子と一致、または下位概念に位置しているかを概念体系辞書を用いて調べる。

例えば、3.1 節の例文「…それを食べた」の場合、動詞共起副辞書では動詞「食べる」にかかる目的語は、飲食物 (3f9639) や動物 (30f6bf) 等の下位概念に位置する。例文から獲得された林檎 (3bd8db)、テレビ (3bd32c) の二つの候補名詞では林檎が飲食物の下位概念であり、「食べる」の深層格となりうる。よって、これに評価値を与える。

vi) 動詞の関連性のチェック

文書には、「A は … α した。B は … β した。C は … γ した。」といった表現が散見される。ここで、A, B は候補名詞、C は代名詞、α, β, γ は動詞とする。このような表現では、規則 ii)〜v) が全て同じ条件だった場合、規則 i) によって距離が優先され、常に B が指示対象となってしまう。しかし、一般には A が指示対象となる場合も多い。例えば、「太郎は勉強した。次郎は遊んだ。彼は合格した。」の場合、「彼」の指示対象は「太郎」である。従って、こういった表現で有効な特定規則の設定が望まれる。

ここでは、γ と α, β 間の動詞の関連性をチェックし、関連が高い方に対応する候補名詞が指示対象である可能性が高いとする規則を設定する。具体的には、前述した類似性判別ツールを用い、γ-α 間、γ-β 間の類似度を算出し、それが 0.1 以上となった場合、それらの間に関連性があるとし、それを含む文中の候補名詞に評価値を付与する。実際、先の例文の場合、「合格」と「勉強」および「合格」と「遊ぶ」の類似度はそれぞれ 0.49、0.00 となり、「太郎」を指示対象と推定できる。

3.3 評価値の統合

各規則の評価値は以下のように統合する。

$$F_m = \sum_{i=1}^{n} a_i \cdot f_i$$

ここで、F_m は候補名詞 m の統合評価値、a_i は規則 i の重み、f_i は規則 i での評価値、n は規則の個数である。また、それぞれの重みは $0 < a \leq 1$ とする。

4 評価

代名詞が出現する文をインターネット等から取得し、一部を修正した文 100 件に対し評価した。評価結果を表 1 に示す。

表 1. 実験結果

規則	成功率 (%)
直前名詞の規則のみ	26.0
性別の規則追加	32.0
数の規則追加	37.0
主題の規則追加	41.0
格関係の規則追加	47.0
動詞の関連性の規則追加	58.0
各規則の重みの調整	61.0

同表から、動詞の意味情報を用いることで指示対象特定の成功率が約 10 %弱向上していることが分かる。

以下に、成功例を示す。

(a)「太郎達は次郎と旅行に行った。旅行先で、次郎のパスポートが盗まれた。彼らは先に帰国した。」

(b)「明智光秀は謀反人として戦った。このため、豊臣秀吉は多くの人達を味方にした。その結果、彼は山崎で敗れた。」

(a) は規則 i)〜iv) のみで特定に成功した例である。(b) は、規則 vi) が効果的に働いて成功した例である。

また、3.2 節の全規則を適用しても指示対象の特定ができなかった例文の一部を以下に示す。

(a)「塩素はナトリウムと結合すると、食塩となる。人体には欠かせない物質である。しかし、それに酸素が結合すると劇薬となる。」

(b)「花子は子供と動物園に行こうとした。しかし、仕事の都合で行けなくなったので、花子は友人の友子に頼んだ。友子は快く了承し、彼女の息子と動物園へ行った。」

(a) を正しく特定するためには、代名詞の後に出現する名詞や動詞を利用した規則が考えられる。すなわち酸素が結合できるものを候補名詞として限定していくような規則が必要となる。これについては前述した EDR 辞書の利用で可能と考えられ、今後作成していく予定である。また、(b) の指示対象は「花子」であるが、単純な意味解析では特定が難しいと考えられる。このような文では、常識知識等を利用した深いレベルでの意味理解が必要といえる。

5 まとめ

本報告では、従来の規則の他に動詞の意味情報を利用した代名詞の指示対象の推定法を提案した。動詞の意味情報としては、代名詞が出現する文の動詞の深層格情報および代名詞が出現する文の主動詞とその前文の主動詞との関連性情報を用いた。評価の結果、これら情報の利用により特定精度の向上が可能との見通しを得た。

参考文献

[1] 村田真樹，長尾真：“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)

[2] 杉村和徳，松田源立，原田実：“意味解析に基づく照応解析の研究”，情報処理学会第 69 回全国大会，1C-5(2007)

[3] EDR 電子化辞書：http://www2.nict.go.jp/r/r312/EDR/J_index.html

[4] 野口洋平, 石川勉：“あらゆる概念表記への対応・精度向上を目指した意味的類似度算出ツール”，情報処理学会第 69 回全国大会，4ZB-2(2007)

[5] グエン・ベト・ハー, 帆苅譲，石川勉，笠原要：“単語の意味に関する大規模概念ベースの構成と評価”，情報処理学会論文誌，43 巻 10 号,pp.3127-3136(2002)

[6] 池原悟ほか：“日本語語彙体系”，岩波書店 (1997)

[7] 川島貴広，石川勉：“言葉の意味の類似性判別に関するシソーラスと概念ベースの性能評価”，人工知能学会誌論文誌，20 巻 5 号 B,pp.326-336(2005)

[8] 形態素解析システム茶筌 (Chasen)：http://chasen-legacy.sourceforge.jp

[9] 日本語係り受け解析器南瓜 (CaboCha)：http://chasen.org/~taku/software/cabocha/

[10] 森田良行：“日本語文法の発想”，ひつじ書房，pp.59-83