1C-5

意味解析に基づく照応解析の研究

杉村 和徳† 松田 源立‡ 原田 実‡
†青山学院大学大学院 理工学研究科 理工学専攻 ‡青山学院大学 理工学部 情報テクノロジー学科

1.はじめに

原田研究室で研究を続けている意味解析システム SAGE[1]の精度も向上し、意味解析の対象となる文章も多様化してきている。そうした中で、文章要約や質問応答等の応用研究も盛んに行われてきたが、特に質問応答の応用研究では照応解析の実装が必要とされてきている。以下に質問応答研究からの要求の一例を示す。

北里柴三郎は細菌学者だ。
彼はペスト菌を発見した。

上記のような文章があった場合、人が見れば、2 文目の「彼」は 1 文目の「北里柴三郎」の事であると判断できる。しかし、従来の 1 文毎に解析を行う意味解析では文内で解析が完結しているため、2 文目の「彼」が、1 文目の「北里柴三郎」のことを指示していると解析することができない。この結果、質問応答システムが回答として「彼」を返し、正しい答「北里柴三郎」を返すことが出来ない事になる。ここで、意味解析で「彼」が「北里柴三郎」を指示していると解析出来るようになると、質問応答側では「北里柴三郎はペスト菌を発見した。」といった知識文とのマッチングが取れるようになり、正しい答えを返すことが出来るようになる。

原田研究室では、以前にも照応解析の研究が行われていたが、その時点では意味解析システム SAGE の精度も含め、長い文章を対象にできる状態ではなかった。そこで、本研究では SAGE の精度向上に合わせて、以前の照応解析を大幅に改良し、新しい照応解析システム Anasys を研究・開発した。

2.先行研究

照応解析の先行研究においては、名詞の指示対象の推定を村田ら[2]が行っている。村田らの手法では、人手による 100 に及ぶ規則を用い、先行詞の指示対象としての可能性を得点付けすることで、指示対象の推定を行う。解析精度は適合率 79%、再現率 77%、F 値 78.0 となっている。

また、飯田ら[3]は、機械学習による名詞句照応解析を行っている。飯田らの手法は、先行詞同定のモデルと、最尤先行詞候補と照応詞の対による照応性判定モデルの 2 つを用意して解析を行う。先行詞同定モデルにおいて必ず 1 つの最尤先行詞候補を返し、照応性判定モデルにおいて照応詞との対が照応関係にあるかどうかを判定し、解析を行う。解析精度は適合率 78.4%、再現率 65.9%、F 値 71.6 となっている。

3.照応解析システム Anasys

本研究では、意味解析システム SAGE による意味解析の情報を用いることで精度の高い照応解析を実現することを目標としている。本手法で扱う照応解析は、ゼロ代名詞解析の中でもゼロ主語の解析と、指示代名詞の解析を対象とし、代名詞の検出から先行詞の特定まで一連の処理を意味解析システムの情報を用いながら行う。

Research of Anaphoric analysis based on Semantic analysis
Kazunori Sugimura†, Yoshitatsu Matsuda‡ and Minoru Harada‡
†Graduate School of Science and Engineering, Aoyama Gakuin University.
‡Department of Integrated Information Technology, Faculty of Science and Engineering, Aoyama Gakuin University.

4.Anasys のシステム構成

照応解析システムである Anasys は， 図 1 に示すように、大きく分けて照応詞検出部と先行詞解析部の 2 つの処理によって構成され、意味解析システム SAGE に、1 つの処理として組み込まれる。意味解析システム SAGE とは、係り受け関係にあるすべての 2 文節の主辞間の深層格を決定するシステムであり、従来の処理では単一の文の解析を主としていたが、この照応解析を実装することで、複数の文にわたる語間の照応関係の解析機能を有することになる。

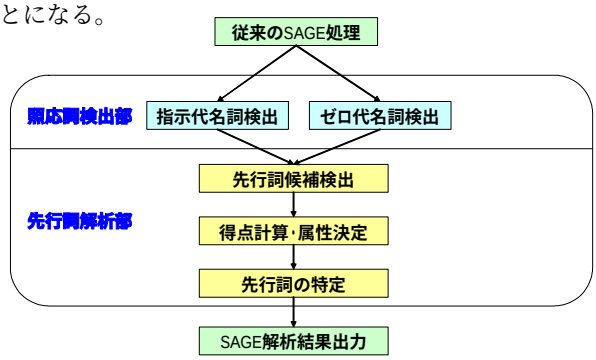

図 1.Anasys の処理の流れ

5.システム概要

図 1 に示すように、照応詞検出部は指示代名詞検出とゼロ代名詞検出の 2 つに分かれ、代名詞が検出された場合に先行詞解析部へと移行する。先行詞解析部は、指示代名詞候補検出と得点計算・属性決定と先行詞特定という 3 つの処理によって構成される。処理の流れは、1)先行詞候補群の検出、2)先行詞候補それぞれに対しての得点付けと属性の取得、3)それらのデータを用いた機械学習による先行詞の特定となる。最終的には意味解析情報に照応解析の結果を付加したデータが出力される。ここでは、それぞれの処理を解説する。

5.1.照応詞検出部

5.1.1.指示代名詞検出

意味解析システム SAGE によって、詳細品詞が名詞形態指示詞または連体詞形態指示詞とされたものを、基本的には指示代名詞と判定する。なお、代名詞である「彼」「彼女」も今回は補完対象として扱うこととする。

5.1.2.ゼロ代名詞検出

意味解析によって係り受け間の主語を表す深層格である agent 格と a-object 格を持たない述語節があるとその必須格情報を参照してそれらを照応詞として検出する。必須格情報は、EDR 電子化辞書[4]を用いて作成した情報で、ある用言に対してどの深層格による係り受けが統計的に多いかを保持している。本研究では、照応詞候補の文節が、agent 格や a-object 格を必須格として持っているが、それを含む文中にこられの格が欠けている場合に補完対象と判定する。

5.2.先行詞解析部

5.2.1.先行詞候補検出

照応詞を含む文とその前３文と後１文を対象として探索し、その範囲にある名詞を先行詞候補とする。この時、形式名詞や同一文内で照応詞への係り受け関係を持つ文節は、対象外とする。

5.2.2.属性決定

先行詞候補として検出されたそれぞれに対して、概念距離得点、前方語間距離得点、後方語間距離得点、特性得点の 4 つの得点と、文内の文節位置、共起関係詞種別、固有名詞判定の３つの属性値を取得する。

1-a)[概念距離得点(指示代名詞)]例えば、「太郎は次郎にパンを渡した。次郎はそれを食べた。」という文章があった場合、指示詞「それ」の係り受け関係「それを食べた。」に注目する。そして、共起関係詞「を」と受け側文節「食べた。」で共起辞書を検索し、共起レコード群を取得する。レコード群は「○○を食べた。」という共起関係のレコード群であり、各レコードの係り側文節と先行詞候補(例:太郎、次郎、パン等)の語意を用いて概念類似度を計算し、最も高い類似度を示した値をその先行詞の概念類似度得点として用いる。

1-b)[概念距離得点(ゼロ代名詞)]ゼロ代名詞の場合も、基本的な処理手順は指示代名詞の場合と同じである。但し、共起辞書引きに用いる要素は「先行詞の共起関係詞」と「照応詞」とする。

2)[前方語間距離得点(ゼロ代名詞/指示代名詞)]文章の中で先行詞が照応詞より前に存在する場合、つまり前方照応の場合に、前方語間距離得点を計算する。得点は、照応詞と先行詞の表記上の距離が近いほど高得点になり、離れているほど得点が低くなる。計算式を以下に示す。

$$\text{語間距離得点} = \left(1 - \frac{\text{照応詞から先行詞までの文節数}}{\text{照応詞から前方(後方)の文節総数}}\right)$$

3)[後方語間距離得点(ゼロ代名詞/指示代名詞)]前方語間距離得点と対になる得点で、文章の中で先行詞が照応詞より後に存在する後方照応の場合に計算する。計算式は、2)で示した通りである。

4)[特性得点(ゼロ代名詞/指示代名詞)]それぞれの照応関係においてどのような事物を指示対象としやすいかを先行詞候補の語意を元にルール化し、得点化を行う。計算方法は、100 点をベースとしてルールによって得点を増減する。例えばゼロ代名詞解析において agent 格を補完する場合、先行詞候補文節の語意から上位概念を取得し、そこに「人間または人間と似た振る舞いをする主体」“3aa911”が含まれていると、先行詞となる可能性が高いと考えられるため、加点対象となる。逆に「時」“30f776”や「場所」“3aa938”が含まれる場合には、減点対象となる。

5)[共起関係詞種別(ゼロ代名詞/指示代名詞)]先行詞候補が持つ共起関係詞が「は、には、が、も、なら、こそ、を、に、にも、へ、で、から、より、その他」のどれかである時にそれぞれ１をとる 14 個の属性。

6)[文内の文節位置(ゼロ代名詞/指示代名詞)]文の先頭文節を文節番号 1 とし、以下に示す計算により、文節位置を計算する。

$$\text{文節位置} = \left(1 - \frac{\text{対象文節の文節番号}}{\text{文に含まれる文節の総数}}\right) \times 100$$

7)[固有名詞判定(ゼロ代名詞/指示代名詞)]その文節に固有名詞が含まれているかどうかを 2 値属性で表す。

5.2.3.先行詞の特定

全ての先行詞候補に対して属性を取得後、それらのデータを用いて先行詞を 1 つに決定する。学習器には原田研究室の成果である Flexible Data Miner(FDM)[5]を用い、SVM 非線形カーネルでの学習を行っている。ここでは、必ず 1 つの先行詞を決定するのではなく、場合によっては補完しない場合もある。それは、照応詞と判定された文節であっても、先行詞が著者や読者であるような時はその前後に必ず先行詞が存在するとは限らない為である。

6.先行詞特定の学習データ

先に述べた学習器で用いる学習データは、本研究室で開発したデータ作成補助ツールを利用して人手で作成した。対象となる文章はインターネット記事・社説等を利用し、ゼロ代名詞事例 78 と指示代名詞事例 112 のそれぞれに対応する先行詞候補群を学習データとして用いた。

なお、データ作成補助ツールを用いて、テストデータ作成や精度評価等も行う事が出来る。

7.評価実験

表１に実験結果を示す。精度評価には、インターネット記事・社説等を中心に、データ作成補助ツールを利用して人手によって作成した事例を用いた。再現率と適合率は以下に示す式によって求めた結果である。

$$\text{適合率} = \frac{\text{システムが正しく解析した総数}}{\text{システムが出力した照応関係の総数}}$$

$$\text{再現率} = \frac{\text{システムが正しく解析した総数}}{\text{人手による正解照応関係の総数}}$$

表 1:照応解析実験結果

	適合率	再現率	F値
ゼロ代名詞	74.4%(32/43)	55.2%(32/58)	63.4
指示代名詞	71.7%(43/60)	58.9%(43/73)	64.7
総合	72.8%(75/103)	57.3%(75/131)	64.1

8.おわりに

本稿では、意味解析の情報を用いた照応解析手法を提案した。今回の手法では、先行研究と比較して精度は少し劣るが、照応詞検出部に意味解析を用い、先行詞特定部にも照応性判定の能力を持たせることで、無駄な補完を極力省けるという利点も持つ。また、今回の意味解析を用いた照応解析で、70%を超える精度を示すこと成功したため、今後は属性得点等の調整・追加や、意味解析情報の有効活用を追加検討することで、さらなる精度の向上が期待できると考えている。

参考文献

[1]山本哲哉, 小林寛之, 米澤太一:意味解析システム SAGE の精度向上と利便性の向上,卒業論文,青山学院大学(2005).
[2]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).
[3]飯田龍, 乾健太郎, 松本裕治:先行文脈と局所文脈を併用した照応性判定モデルの学習, 言語処理学会第 11 回年次大会 発表論文集, pp.1048-1051(2005).
[4] (株) 日本語電子辞書研究所: EDR 電子化辞書仕様説明書 (第２版) ,(株) 日本語電子辞書研究所(1995).
[5]櫻沢研一, 越前陽祐:操作性・汎用性の高いデータマイニングツール FlexibleDataMiner(FDM)の開発, 卒業論文,青山学院大学(2005).