主題・焦点を考慮した照応解析システムの解析精度向上

伊澤 友輔*　韓 東力**　原田 実**

青山学院大学　理工学研究科　経営工学専攻*　　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

原田研究室はこれまで、文章中の指示代名詞とゼロ代名詞の照応先を決定する照応解析システムAnasys[1]を開発してきた。従来の Anasys は，ＥＤＲ電子化辞書による照応詞と先行詞の語意の類似性の評価および語間距離と、照応詞の特性の３つの得点を線形的に足し合わせて先行詞を決定しており、近い語意の先行詞候補が複数ある場合には誤った先行詞候補を検出していた。これらを調査の結果、主題・焦点情報を加味することで正解を導けることが分かった。そこで主題・焦点を新たな得点として導入する。また正解先行詞となりうる確率を、学習データをもとに判別分析を行い、統計的に根拠を持つ得点の統合化を行った。

2.　Anasys における照応解析

Anasys は指示代名詞とゼロ代名詞を解析対象とし、大きく照応詞検出部と先行詞解析部の２つに分かれる。

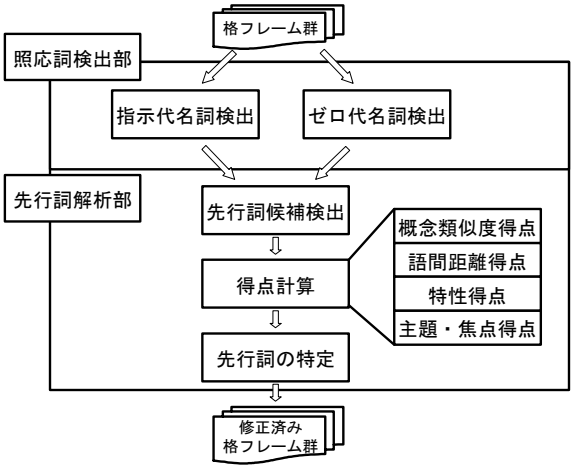

図 1　Anasys の基本的処理フロー

2.1.　照応詞検出

まず、SAGE[2]の出力である格フレーム群を入力とし、各用言の深層格の EDR の共起辞書内における出現割合を計算することで、統計的に各用言の必須格を決定し、文章中にない格をゼロ代名詞として検出する。指示代名詞は SAGE の品詞解析情報を基に検出する。

2.2.　先行詞解析

照応詞 300 文例を調査した結果を図 2に示す。それにより、照応詞を含む文とその前２文後１文を先行詞解析の範囲とする。その範囲から抽出された先行詞候補ごとに「概念類似度得点」、「語間距離得点」、「特性得点」、「主題・焦点得点」の４つの得点を計算し、学習データを基に正解先行詞確率を計算する。そして先行候補の中で最も正解先行詞確率の高い先行詞候補を正解先行詞と特定し、格フレーム群を補完する。以下に４つの得点について述べる。

	それ以前	前2文	前1文	同文	後1文	それ以降
割合(%)	0.3	13.0	50.0	32.7	4.0	0.0

図 2　先行詞の出現場所頻度

2.2.1.　概念類似度得点

概念類似度計算は、照応詞の係り元用言の共起事例中の係り先語と先行詞候補の語意類似度を計算し、その平均値を求めるものである。

$$\text{概念類似度} = \frac{\sum \text{語意類似度}}{\text{全ての共起事例}\ \text{共起事例の総数}}$$

図 3　概念類似度算出式

語意類似度の算出方法を述べる。 語ａと語ｍの共通上位語をｃとすると、最上位概念である「概念」から共通上位語ｃまでの EDR 電子化辞書の概念階層状の距離を dc とする。同様に語ａから語ｃまでの距離をar、語ｍから語ｃまでの距離を mr として図 4に示すような計算式で語意類似度を求める。

$$\text{語意類似度} = \left(\frac{0.6dc}{0.6dc + (1-0.7^{ar}) + (1-0.7^{mr})}\right) \times 100$$

図 4　語意類似度得点算出式

2.2.2.　語間距離得点

先行詞は前方照応の可能性が高く、後方照応の可能性は低い。また、先行詞は指示代名詞やゼロ代名詞の近くに存在する可能性が高い。以上の２点の性質から図 5に示す計算式を用いて語間距離得点を求める。

Improvement of accuracy of the correspondence analysis in consideration of the theme and the focus

Yusuke Izawa*, Dongli Han ** and　Minoru Harada **
*Graduate School of Industrial and Systems Engineering, Aoyama Gakuin University.
**Department of Integrated Information Technology, Faculty of Science and Engineering, Aoyama Gakuin University.

$$\text{語間距離得点（前方照応）} = \frac{\text{照応詞から数えて何番目の単語か}}{\text{照応詞の前方にある単語の総数}} \times 100$$

$$\text{語間距離得点（後方照応）} = \frac{\text{照応詞から数えて何番目の単語か}}{\text{照応詞の前方にある単語の総数}} \times 50$$

図 5　語間距離得点算出式

2.2.3.　特性得点

指示代名詞とゼロ代名詞の特性（例:指示代名詞「ここ」、「そこ」、「あそこ」の場合には、場所を表す語を指示対象としやすい）を、EDR 電子化辞書のコーパス辞書等の事例検索から指示代名詞がどのような事物を指示対象としやすいかを判断し、得点化する。

2.2.4.　主題・焦点得点

主題・焦点の特徴として、ゼロ代名詞や指示代名詞の指示対象になり易いというものがある。しかしながら主題・焦点の抽出方法はまだまだ曖昧で、決定的なアルゴリズムは構築されていない。そこで本研究では文章ごとに主題・焦点を一意に抽出するのではなく、主題・焦点候補の指示対象へのなり易さの度合いを、村田ら[3]の手法を基に独自の考察を加えて図 6のように得点化した。

主題の得点

表層表現	例	得点
名詞　は／には	太郎はした	100

焦点の得点

表層表現（「は」がつかないもので）	例	得点
名詞　が／も／だ／なら／こそ	太郎もした	75
名詞　を／に／、／。	太郎にした	70
名詞　へ／で／から／より	学校へ行く	65

図 6　主題・焦点得点

3.　得点の統合化

まず、x_1：概念距離得点、x_2：語間距離得点、x_3：特性得点、x_4：主題・焦点得点とし、学習データとしてゼロ代名詞 300 文例、指示代名詞 300 文例の先行詞候補をあらかじめ人手にて正解、不正解に判別しておく。その上で新たな先行詞候補を解析する際、それが正解、不正解どちらの群に近いかをマハラノビスの汎距離を用いて判別し、正解先行詞確率という形で算出する。以下に正解先行詞確率を算出するまでの手順を示す。

$V_{(1)}^{-1}$　：正解群の分散・共分散行列の逆行列

$u_{(1)}$　：正解群変数の平均ベクトル

$$u_{(1)} = (\bar{x}_{1(1)} \cdots \bar{x}_{4(1)})$$

X　：未知の先行詞候補の得点ベクトル

$$X = (x_1 \cdots x_4)$$

とすると、正解群からのマハラノビス距離 $D_{(1)}^{\ 2}$ は

$$D_{(1)}^{\ 2} = (X - u_{(1)}) V_{(1)}^{-1} (X - u_{(1)})$$

となる。

同様に不正解群からのマハラノビス距離 $D_{(2)}^{\ 2}$ を求め、

$$f_1 = e^{-\frac{1}{2}D_{(1)}^{\ 2}}, \quad f_2 = e^{-\frac{1}{2}D_{(2)}^{\ 2}}$$

とすれば、未知の先行詞候補が正解に属する確率は

$$P_1 = 100 \times \frac{f_1}{f_1 + f_2}$$

となる。

これを正解先行詞確率とし、先行詞候補の中で最大となったものを出力する。

4.　実験・評価

本研究の有効性を確かめるため、A:概念類似度得点、B:語間距離得点、C:特性得点、D:主題・焦点得点、の各得点を単純和したもの、更にE:マハラノビスの汎距離により各得点の統合化を行ったものの、5 つの手法に対する精度評価実験を行った。

組み合わせ	A	A+B	A+B+C	A+B+C+D	A+B+C+D+E
正答率(%)	31.3	59.4	62.5	84.4	87.5

図 7　組み合わせと解析精度

図 7は毎日新聞記事の指示代名詞を含む 50 文例、ゼロ代名詞を含む 50 文例による実験結果である。なお、これらはマハラノビスの汎距離に用いた学習データとは、異なる文例のものである。

これにより、4 つの得点をマハラノビスの汎距離を用いて統合化した場合に最も精度が良くなることが分かった。

謝辞

本研究の一部は文部省科学研究費基盤研究 C『日本語文章の常識を用いた意味理解・文脈理解システムの開発研究』の補助金を用いて行われました。

参考文献

[1] 南 旭瑞, 原田 実 : ”語意の類似性を用いた照応解析システムの開発Anasys”, 情報処理学会第64回全国大会論文集, 3M-06 第 2 分冊 pp. 53-54 (2002. 3).

[2] 原田実, 田淵和幸, 大野博之:”日本語意味解析システム SAGE の高速化・高精度化とコーパスによる精度評価”, 情報処理学会論文誌, Vol. 43, No. 9, pp. 2894-2902, (2002. 9).

[3] 村田真樹, 長尾真 : 用例や表層表現を用いた日本語文章中の指示詞・代名詞・ゼロ代名詞の指示対象の推定, 自然言語処理, Vol. 2, No. 3, pp. 3-26 (1995. 7).