意味解析に基づく照応解析システムＡＮＡＳＹＳの精度向上と大規模テキストコーパスによる評価実験

村上春佳　笠間千秋　松田源立　原田 実

青山学院大学　理工学部　情報テクノロジー学科

1.　背景・研究目的

原田研究室で研究を続けている意味解析システム SAGE[1]内には照応解析システム ANASYS[2]が組み込まれている。意味解析の精度も向上し、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[3]では照応解析部分の精度の向上が必要とされてきている。本研究の照応解析システムは多くの他の既存研究[4]とは違い意味解析に基づいておこなわれている。これにより、解析中の語に EDR 辞書中での語意が付与されているので、この概念を概念体系木中の他の概念との類似度を計算でき、人が照応解析を判断する時に考える「この動詞にはこの様な概念の名詞を先行詞としてとりやすい」という判断を、計算を用いて行うことができる。ANASYS の照応解析には「指示代名詞の解析」と「ゼロ代名詞の解析」が存在している。本研究では、特に「ゼロ主語の解析」における精度の向上をおこなった。また、本研究では[4]と異なり、文内照応だけでなく、前文や後文に先行詞がある場合や、外界照応を扱えるようにした。なお、客観的な学習・評価を行うために、NAIST テキストコーパス[5]を利用し、学習・評価実験をおこなった。

2.　基本的な考え方

本手法で扱う照応解析は、代名詞の検出から先行詞の特定まで一連の処理を、図 1に示すように意味解析結果の情報を用いながら行う。意味解析システム SAGE は、単一文内における各語の語意と、係り受け関係にあるすべての 2 文節間の深層格を与えるが、照応解析機能を組み込むことで、複数の文にわたる語間の照応関係の解析もおこなえる。

3.　システム概要

ANASYS の処理の流れを図 1に示す。各工程ごとに解説する。

3.1.　照応詞検出部

主語を表わす深層格（主語格）を持たない動詞節、動名詞節、断定節の３種類の述語節に対しては、その主語格を補完する必要がある。よってそれらをゼロ代名詞の照応解析が必要な照応詞とみなす。

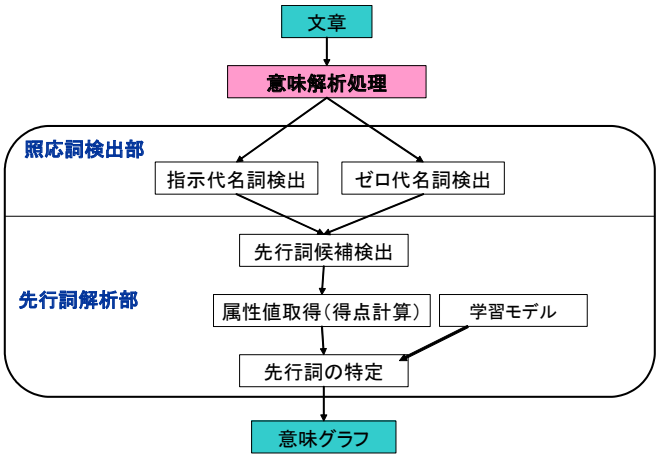

図 1　ANASYS の処理の流れ

3.1.1.　主語格の決定方法

ゼロ代名詞の場合、主語を表わす深層格（主語格とする）となりうる格は agent 格（有意志動作の主体）、a-object 格（属性の主体）、o-agent 格（無意志動作の主体）の３種類である。この主語格の決定は、照応詞の文節品詞に基づいて以下のルールに従って決定する。

１）断定節の場合

照応詞とガ格の共起レコードを調べ、１つでも a-object 格を持つレコードが存在すれば、主語格を a-object とする。存在しない場合は、agent 格と object 格の出現数を比べ、agent 格が多ければ agent 格とする。object 格が多ければ、o-agent 格とする。

２）動詞節と動名詞節の場合。

照応詞が scene 格や place 格など別の深層格での係り受けが存在する場合には、共起レコード中に同じ係り受けが存在するかを調べる。存在するならば、その係り受けの例文を参照し、例文中でどの深層格で使われているかを見る。agent 格で使われているなら agent 格、a-object 格で使われているなら a-object 格、object 格で使われているなら o-agent 格を主語格とする。例を以下に示す。また、係り受けが存在しない場合は、１）と同様に agent 格と object 格の出現回数で主語格を決定する。

例）

私は港に入るのを見た。

①照応詞が「入る」で、この「入る」は「港に」と goal 格で係り受け関係がある。

②「入る」と共起関係子「に」で検索する。
③共起レコード中の係り側に「港」と同じ概念が存在したら、その例文「船が港に入る。」などを取り出す。
④例文では、「入る」と「船」が object 格で存在している。
⑤照応詞「入る」の主語格を o-agent 格とする。

3.2. 先行詞解析部

3.2.1. 先行詞候補群の検出

本文中からは、図 2に示すように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節と断定節を先行詞候補とする。本来は、先行詞として照応詞に直接かかっていないものが選択されるべきだが、place 格などの別の深層格で主語となる節が係っている場合も存在する。そこで、先行詞候補の中には直接係っている文節も入れ、最終的にその候補が先行詞として最有力となった場合には、その照応詞を照応解析の対象外とすることにした。他にもタイトルは常に先行詞候補とすることとし、前６文までの主題と考えられる文節も候補とした。また、先行詞が本文中に存在する名詞ではなく、筆者や読者などである場合を考え、外界として先行詞候補に加えた。コーパスでは外界の定義を一人称、二人称、一般の３種類としていたが、本研究では概念を重視するため以下の５種類とした。

１）一人称
筆者が自分の考えを述べている場合などである。例えば、「本を読みたいと思う。」の「読む」と「思う」のは筆者の動作である。

２）二人称
読者に提案を挙げている場合などである。例えば、「もう帰って寝たらどうですか。」の「帰る」と「寝る」のは読者の動作である。

３）事
一般的な事象において、人ではなく、出来事が起こしている事象の場合である。例えば、「そろそろ円安に転じる。」の「転じる」動作を行う対象が事象である。

４）人
一般的な事象において、誰かが人的に行った事象の場合である。例えば、「税金を納めるのは当然だ。」の「納める」の動作の主語となるのは世間一般の人々である。

５）物
一般的な事象において、物など意識を持たない物体が主語となる場合である。例えば、「故障したら修理に出しましょう。」の「故障した」の主語が物である。

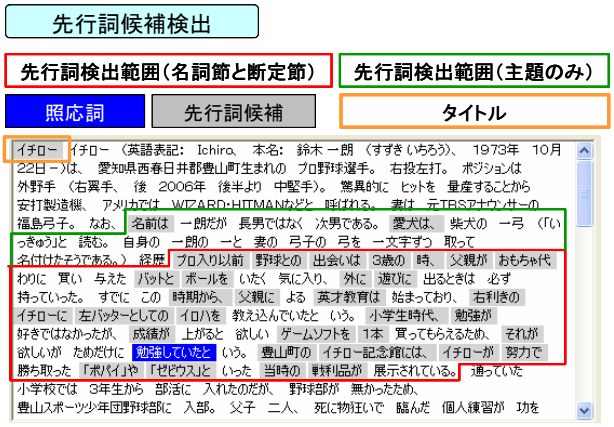

図 2 先行詞候補検出範囲の一例

3.2.2. 属性値の取得

図 3に示すように、各先行詞候補に対し、概念距離得点、語間距離得点、特性得点、主題得点、固有名詞得点の５つの属性値を取得する。

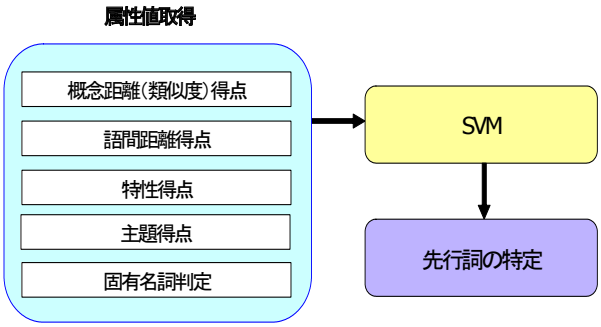

図 3 属性値の概要

１）概念距離得点
照応詞と先行詞候補の関係の成り立ちやすさを概念を用いて計算した得点。共起辞書を調べ、照応詞がどのような概念の単語を主語としているかを調べ、その例文の概念と先行詞の概念の距離を計算して求めている。例えば、「太郎は京都生まれである。しかし、東京で育ったらしい。」という例文では、照応詞は「育った」の部分である。まず共起関係詞「が」と照応詞「育つ」を持つ共起レコードを検索する。次に「私が育った」などの共起レコードの係り側「私」と先行詞候補「太郎」との概念距離を計算する。これを繰り返し、全共起レコードで行い上位５つの平均値を概念距離得点とする。

２）語間距離得点
先行詞と照応詞の文節間の距離を計算する。主語を省略している文章はその主語となる言葉が近くに存在しているので表記しないという場合が多い。そこで先行詞と照応詞の距離が近いものほど

先行詞となりやすいと考えられる。表記上、照応詞と先行詞が近いほど、点数が高くなる。

３）特性得点

先行詞候補が agent 格になりやすいかを得点化する。agent 格は有意志動作を引き起こす時につけられる深層格であるので、有意志の先行詞候補が先行詞となりやすいと考えられる。初期値を 0 とし、先行詞候補の上位概念に「人間または人間と似た振る舞いをする主体」を持っていた場合、得点を 1 とする。

４）主題得点

先行詞候補が主題となりうる場合に得点を与える。主題や焦点となる言葉はその話題の中心である為に、先行詞となりやすいと考え得点化する。たとえば、ハ格を持つ場合は 1.0、ガ格を持つ場合は 0.8 を与える。また照応詞から 1 つ遠くなる毎に点数を減らす。

５）固有名詞得点

先行詞候補が固有表現かどうかを判断し、得点を与える。固有表現を持つ言葉も話題の中心となっている可能性が高く、先行詞となりやすいからである。

3.2.3. 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。決定には、SVM の線形カーネルを利用し、識別関数値が一番大きくなるものを先行詞として決定した。また、全ての照応詞に対して、必ず先行詞は存在するものと考え、先行詞候補群の中から 1 つ選択する。

3.2.4. 先行詞特定の事例

本研究では、先行詞決定に使われる属性値に概念類似度を用いている。概念類似度は人が先行詞を決定する時に用いる判断を機械が行なえる手段である。例えば、「その物音は近いのか遠いのか分らないほどかすかであって、この広い屋敷の壁の中から響くのか、･･･」という文章は文法的に「物音」が「響く」に係っていることを見つけるのは困難であるが、「響く」の共起事例の係り側と「物音」の概念類似度を考えることで、「物音」が先行詞として抽出しやすくなる。属性値取得の結果例を図 4に示す。

	響くのか	格	o-agent				
先行詞候補	正1負2	概念類似度	語間距離	特性得点	主題得点	固有名詞
曲物	2	1.65	−1.36	0	0	0
レイモンド	2	2.98	−1.26	0	0.49	0
物音は、	2	3.24	−0.44	0	0.7	0
闇の	2	2.86	−0.02	0	0	0
物音は	1	3.24	0.8	0	1	0
屋敷の	2	1.62	1.62	0	0	0
壁の	2	1.54	1.73	0	0	0
庭の	2	1.63	1.63	0	0	0
木立の	2	1.58	1.39	0	0	0
彼女は	2	1.58	−0.6	0	1	0

図 4 先行詞取得例

4. 先行詞特定の学習データ作成

SVM の学習のためにゼロ代名詞用、指示代名詞用で学習データを作成し、特にゼロ代名詞には解析文章の分野別に物語文、新聞記事、辞典文、クレーム文の 4 種類の学習データを作成した。学習器で用いた学習データは、新聞記事の文章には NAIST テキストコーパスを用いて学習データ作成支援ツールから自動的に作成している。また、物語文、辞典文、クレーム文は筆者らのグループにより人手により解析したデータを学習データとして利用した。学習データは、図 5に示した形式であり、3.2.2で述べた属性値と先行詞(1)か非先行詞(-1)の情報で成っている。

学習データ(ゼロ代名詞)
形式
先行詞(1)･非先行詞(-1) 1:概念類似度得点 2:語間距離得点 3:特性得点 4:主題得点 5:固有名詞属性
実例
-1 1:0.43 2:-0.67 3:1.00 4:0.20 5:0
-1 1:-0.40 2:1.17 3:0.00 4:0.20 5:0
1 1:0.61 2:1.17 3:1.00 4:0.20 5:0
-1 1:-0.33 2:-1.02 3:0.00 4:0.00 5:0
-1 1:-0.35 2:-0.70 3:1.00 4:0.49 5:1
-1 1:-0.91 2:-0.54 3:0.00 4:0.00 5:0
-1 1:-0.73 2:0.10 3:0.00 4:0.70 5:0
-1 1:0.01 2:0.26 3:1.00 4:0.00 5:0
-1 1:2.17 2:0.42 3:1.00 4:0.00 5:1
-1 1:-0.73 2:0.90 3:0.00 4:0.00 5:0
-1 1:-0.35 2:1.07 3:1.00 4:0.00 5:1
1 1:2.17 2:1.23 3:1.00 4:1.00 5:0
（以下続く）

図 5 学習データ例

学習データを作成する処理の流れを図 6に示す。処理の流れは、１）コーパスのテキストデータを SAGE で解析を行い、意味グラフとして出力する。２）ANASYS MODEL FORM に意味グラフと、コーパスの照応関係が書かれている xml データを入力する。３）先行詞候補に対する属性値とコーパスによる正負判定が書かれている学習支援データが TinySVM[6]に適した形で出力される。４）そのデータを TinySVM に入力することで、学習モデルデータが作成される。

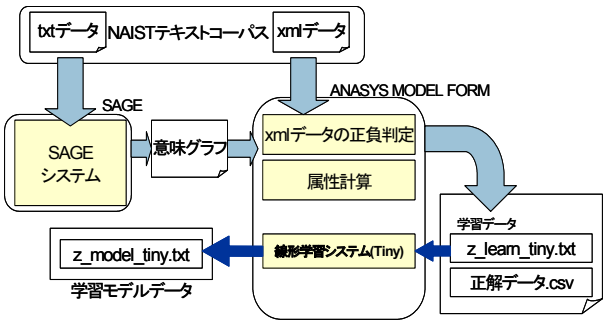

図 6　学習データ作成の流れ

5.　評価実験

具体的な例を用いて評価実験を行った。本研究では利用データにコーパスを用いているため、客観的に照応詞の判定を行えるようになった。そこで、照応詞判定の精度と先行詞判定の精度に分けて評価実験を行った。

5.1.　照応詞判定における精度評価

コーパスにおける照応詞判定の精度を新聞記事3512 事例を用いて評価した。適合率と再現率の算出式を以下に示し、結果を表 1に示す。昨年のシステムでは解析が行われる照応詞の数が少なかった。本年は解析が行なわれる照応詞の抽出ルールの改善を行い、90%近く抽出できるようになった。

$$\text{適合率} = \frac{\text{システムが正しく照応詞とした総数}}{\text{システムが出力した照応詞の総数}}$$

$$\text{再現率} = \frac{\text{システムが正しく照応詞とした総数}}{\text{コーパスによる照応詞の総数}}$$

表 1　照応詞判定における精度評価

	適合率	再現率	F値
新聞	89.56% （3122/3486）	88.90% （3122/3512）	89.23

5.2.　先行詞判定における精度評価

本研究では、先行詞特定の計算を分野別で行っているため、各分野ごとで精度評価を行うこととする。新聞記事の場合には5.1で利用したデータと同様の物を利用している。他分野では筆者らのグループが人手により作成した正解データを用いる。物語文では物語文章 64 事例、辞典文では wikipedia 文章 51 事例、クレーム文ではミドリカワ電気 59 事例を利用した。適合率と再現率の算出式を以下に示し、結果を表 2に示す。新聞記事の精度が低い。同じ新聞のコーパスを対象とする飯田ら[4]の研究では精度 75%が報告されているが、これは文内ゼロ主語である。新聞では文章が長く、また先行詞候補の数が他分野に比べても多く、文章中に組織名や人物名など先行詞となりやすい語が多く登場し、また外界照応を考慮している。新聞記事の特徴を探し出し、更なる属性値取得などに工夫が必要であることがわかった。一方、物語文・辞典・クレーム文の精度は新聞記事より高く、また３つの分野に共通して 50%近い数値を出せたことは良い結果といえる。

$$\text{適合率} = \frac{\text{システムが正しく先行詞を選んだ総数}}{\text{システムが出力した照応詞の総数}}$$

$$\text{再現率} = \frac{\text{システムが正しく先行詞を選んだ総数}}{\text{コーパス・人手の正解先行詞がANASYSの先行詞出力条件を満たす照応詞の総数}}$$

表 2　先行詞判定における精度評価

	適合率	再現率	F値
物語文	50.00%（34/68）	53.94%（34/63）	51.9
新聞	25.27% （881/3486）	29.07% （881/3031）	27.0
辞典	50.94%（27/53）	56.25%（27/48）	53.5
クレーム	40.68%（24/59）	44.44%（24/54）	42.5

6.　結論

照応詞判定の精度は改善され、90%近く照応詞を抽出できるようになり、先行詞判定では物語文・辞典・クレーム文での精度が高くなった。また、属性値調整の結果、概念類似度を利用して先行詞決定を行なえるようになった。多様な文での先行詞特定は困難な問題であり、３つの分野に共通して 50%近い数値は良い結果を出せているということは、本研究のアプローチの有効性を示している。

7.　参考文献

[1] 川口純一,青木洋,松田源立,原田実:"意味解析システム SAGE の精度向上"情報処理学会第 69 回全国大会論文集,1C-04,第 2 分冊 pp. 77-78. (2007.3).
[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 回全国大会論文集,1C-05,第 2 分冊 pp. 79-80. (2007.3).
[3] Minoru Harada, Yuhei Kato, Kazuaki Takehara, Masatsuna Kawamata, Kazunori Sugimura, and Junichi Kawaguchi: “QA System Metis Based on Semantic Graph Matching “,Proc. of the 6th International Conference on NII Test Collection for IR Systems(NTCIR6), Tokyo, Japan, pp.448-459, (2007.5).
[4] 飯田龍, 乾健太郎, 松本裕治: 文の構造を利用した文内ゼロ照応解析, 言語処理学会第 12 回年次大会 発表論文集, pp.488-491(2006).
[5] NAIST Text Corpus：http://cl.naist.jp/nldata/corpus/
[6] TinySVM:http://chasen.org/~taku/software/TinySVM/