# 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

笠間　千秋（15804022）、村上　春佳（15804071）
原田研究室

## 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]内には照応解析システム ANASYS[2]が組み込まれている。本年度は、照応解析部分の中でも「ゼロ代名詞の解析」における精度の向上をおこなった。また、本研究では既存研究とは異なり、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[3]を利用し、学習・評価実験をおこなった。

## 2.　手法

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。各工程について説明する。

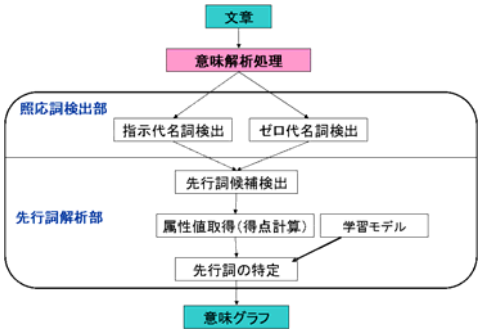


図 1　ANASYS の処理の流れ

### 2.1.　照応詞検出部

主語を表わす深層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。

### 2.2.　先行詞解析部

#### 2.2.1.　先行詞候補検出

本文中からは、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。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入れた。外界は、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。

#### 2.2.2.　属性値取得（得点計算）

各先行詞候補に対し、以下の５つの属性値を取得する。

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。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。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。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。

#### 2.2.3.　先行詞特定

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。先行詞特定で利用するモデルファイルは、物語文、新聞記事、辞典文、クレーム文の４種類に対して TinySVM[4]を学習器として作成した。この時に使われる学習データは、学習データ作成支援ツールを用いて、新聞記事の場合にはコーパスを用いて自動的に作成し、物語文、辞典、クレーム文は筆者らのグループにより人手により作成した。

## 3.　評価実験結果

### 3.1.　照応詞判定の精度評価

本研究では、コーパスを利用していることから客観的な照応詞の数を計測できるようになった。そこで、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。結果を表 1 に示す。

**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 |

### 3.2.　先行詞判定の精度評価

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。新聞記事では 3.1 と同様のデータを利用している。他分野では人手により作成したデータを用いる。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。結果を表 2 に示す。

**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 |

## 4.　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った。また、先行詞判定では新聞記事の精度が低い。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、また外界照応が多い。そのため、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・クレーム文での精度が高くなった。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。

## 5.　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]NAIST Text Corpus：http://cl.naist.jp/nldata/corpus/

[4]TinySVM:http://chasen.org/~taku/software/TinySVM/