意味解析に基づく照応解析システムＡＮＡＳＹＳの精度向上と EM アルゴリズムによる学習の導入

西尾公秀（15805055）
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」の精度の向上を行った.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,学習には EM アルゴリズムを用いた.

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.

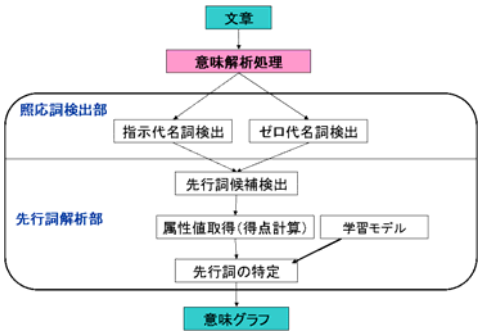

図 1　ANASYS の処理の流れ

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,一般的な人,事柄などである場合を考え,外界として先行詞候補に入れる.

2.2.2.　属性値取得（得点計算）

各先行詞候補に対し,昨年度までは概念距離得点,語間距離得点,特性得点,主題得点,固有名詞得点の５つの属性値を取得していたが,本年度はさらに,以下の５つの属性値を追加した.

１）同一主語得点：文中の係り受け関係において,照応詞 Vs と manner 格,sequence 格など,連接・並列などの関係にある動詞 V1 の主語になっている先行詞候補Ａは,Vs の先行詞になりやすいため,該当するＡの同一主語得点に１を与える.

２）主語格得点：文中の係り受け関係において,先行詞候補Ａが,照応詞Vsの主語格と同じ深層格で他の文節に係っているとき,Ａの主語格得点に１を与える.

３）分裂文得点：「Vs したのはＡである」といった分裂文構造のとき,Vs の先行詞リストに断定節「Ａである」を加え,分裂文得点に１を与える.

４）一人称得点：照応詞 Vs が「思う」や「聞こえる」などを上位概念に持つと,新聞記事において外界（一人称）を取りやすい.よって該当する Vs の先行詞リスト中の外界(一人称)の一人称得点に１を与える.

５）係り受け距離得点：文節の数で距離を測っていた語間距離得点とは異なり,係り受け解析によって得られた係り受け関係の木構造から,照応詞Vsから先行詞候補 A までの枝の数を測り,それをＡの係り受け距離得点とする.Ａが Vs と異なる文に存在する場合は,Ａから文末までの距離＋文番号の差を得点とする.

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,正例,

負例それぞれの学習データについて,Weka[5]のＥＭアルゴリズムによって下記の混合正規分布を導出し,そこで得られたパラメータより確率分布 Pi を計算する.

$$p_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{(x-\mu_i)^2}{2\sigma_i^{\,2}}\right\}$$

$$p(x) = \sum a_i \cdot p_i(x)$$

$$Score(x) = \frac{p_{positive}(x)}{p_{positive}(x) + p_{negative}(x)}$$

得られたスコアは,閾値によって判別して信頼性の高いもののみを抽出することも出来る.なお,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,コーパスより自動的に作成した.

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.そこで,コーパスにおける照応詞判定の精度を,新聞記事 280 記事を用いて評価した.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	82.99% (4109/4951)	98.42% (4109/4175)	90.0

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.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.新聞記事では 3.1 で求めた照応詞を基に,SVM,EM アルゴリズム,EM アルゴリズムで閾値設定の各条件においての解析精度を評価した.他それぞれ結果を表 2,表 3 に示す.

表 2　先行詞判定における精度評価

[]内は昨年の精度を示す

分野	適合率	再現率	F値
物語文	41.22%(47/114) [49.25%]	51.64%(47/91) [51.56%]	45.8 [50.4]
辞典	39.31%(46/117) [29.03%]	42.20%(46/109) [30.51%]	40.7 [29.7]
クレーム	41.60%(57/137) [26.83%]	46.34%(57/123) [28.45%]	43.8 [27.6]

表 3　先行詞判定における精度評価(新聞)

新聞	外界	適合率	再現率	F値
昨年度	有	21.51% (1372/6374)	24.55% (1372/5589)	22.9
新得点 [SVM]	有	30.22% (1496/4951)	35.83% (1496/4175)	33.0
	無	28.47% (1077/3783)	35.82% (1077/3007)	32.1
新得点 [EMアルゴリズム]	有	28.78% (1425/4952)	34.13% (1425/4175)	31.5
	無	24.45% (925/3784)	30.76% (925/3007)	27.6
新得点 [EMアルゴリズム] (閾値)	有	49.53% (423/854)	49.47% (423/855)	49.5
	無	54.36% (368/677)	54.28% (368/678)	54.3

4.　結論

昨年度特に低かった新聞のゼロ代名詞の先行詞判定精度において,昨年度 20%弱であった適合率が 30%ほどにまで上がった.主な要因として,新得点の同一主語得点が 300 例近い数の正解を導き出していることがいえる.また,EM アルゴリズムの導入によって先行詞になる確率を相対的な数値として処理し,閾値を用いて信頼性の高いものを選ぶことが出来るようになった.その結果,応用研究で必要とされる外界以外の照応の部分で 50%を上回る精度を出すことが出来た.一方で辞典において 10%、クレームにおいて 15%ほどの精度向上が見られた.これらの分野に特化した得点を追加したわけではないので,EM アルゴリズムによる学習の効果であるといえる.一方で物語文には精度向上が見られなかった.本年度に追加した得点は新聞記事など解説・論説の解析に効果的な得点であり,物語などの文学的な文書の解析には適していないため,学習時に正解の妨げになってしまった可能性があると考えられる.

5.　参考文献

[1]梅澤俊之, 西尾華織, 松田源立, 原田実:" 意味解析システム SAGE の精度向上とモダリティの付与と辞書更新支援系の開発",卒業論文,青山学院大学(2007).

[2]村上春佳, 笠間千秋, 松田源立, 原田実:" 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",卒業論文,青山学院大学(2007).

[3]NAIST Text Corpus：http://cl.naist.jp/nldata/corpus/

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

[5]Weka: http://www.weka-jp.info/