Vol.2009-NL-191 No.5
Vol.2009-SLP-76 No.5
2009/5/21

意味解析に基づく照応解析システムＡＮＡＳＹＳ
-EM アルゴリズムによる先行詞同定の学習の導入-

西尾公秀[†]　村上春佳[††]　松田源立[†††] 原田実[†††]

我々は意味解析に基づく照応解析システム ANASYS を開発した．ANASYS は文内・文間照応を問わず，また文中に存在しない単語への外界照応も含めた照応解析を行う．本研究では先行詞候補に意味解析によって付与された EDR 辞書中での語意から計算した概念類似度，先行詞候補に係る深層格，文構造などを得点化して計 7 個の素性を用意した．これらの素性を持つ学習データから EM アルゴリズムを用いて混合正規分布のパラメータを推定することで各先行詞候補が正解先行詞になる確率モデルを同定した．この確率モデルを用いて NAIST テキストコーパス全 609 記事(外界照応も含む照応詞 13967 個を含む)について精度実験を行った結果，適合率 33%，再現率 41%という結果になった．

Anaphoric analysis system ANASYS based on semantic analysis
- Antecedent identification by EM algorithm -

Masahide Nishio[†], Haruka Murakami[††],
Yoshitatsu Matsuda[†††] and Minoru Harada[†††]

We developed anaphoric analysis system ANASYS based on the semantic analysis. ANASYS resolves the anaphoric correspondence including the exophoric correspondence to the word which does not exist in the sentence including the anaphora. In this research, to each antecedent candidate, seven features such as the concept similarity calculated from the meaning of a word given by the semantic analysis, the deep case representing the function of the antecedent candidate and the sentence structure, are given. By using the EM algorithm for the learning data having these seven features, we estimated the parameters of the contaminated normal distribution according to which each antecedent candidate becomes a correct antecedent. Experimentation using NAIST text corpus of 609 articles (including 13967 anaphora) resulted in precision 33% and recall 41%.

† 青山学院大学大学院理工学研究科
†† ソニー株式会社
††† 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][2]内には照応解析システム ANASYS[3][4]が組み込まれている．意味解析の精度も向上し(語意精度 95%，深層格精度 90%程度)，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．

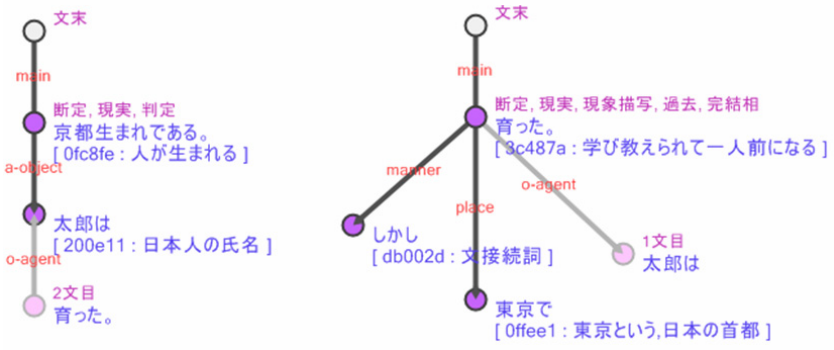

図 1　意味解析システム SAGE における照応解析

特に質問応答の応用研究[5]では照応解析の精度の向上が特に必要とされてきている．照応解析の既存研究として，規則を作成する手法と照応関係タグ付きコーパスを用いた学習手法がある．田村ら[6]はセンタリング理論に基づいた規則によって文中の名詞句を先行詞になりやすい順番に順位付けして先行詞同定をおこない，さらに複文についても解析がおこなえるよう拡張した手法を提案している．飯田ら[7][8][9]はセンタリング理論による規則を素性とし，学習をおこなっている．また先行詞同定にはトーナメントモデルというモデルを用いている．トーナメントモデルとは，照応詞に一番近い先行詞候補から順に 1 対 1 で勝ち抜き戦のような形で先行詞を選ぶ. つまり，文章の一番前方にある先行詞候補は，それまでを勝ち抜いてきた先行詞候補とどちらが先行詞らしいかを比較することになる．

本研究の照応解析システムはこれらの既存研究と違い，意味解析に基づいておこな

Vol.2009-NL-191 No.5
Vol.2009-SLP-76 No.5
2009/5/21

う．これにより，解析中の語に EDR 辞書[10]中での語意が付与されているので，任意の二つの語の語意間の概念類似度を計算し素性に用いることができ，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，先行詞候補と照応詞の共起レコード中の係り側間の概念類似度計算を用いて行うことができる．また，本研究では文内照応だけでなく，前文や後文に先行詞がある場合や，外界照応も扱える．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「ゼロ代名詞の解析」における精度の向上として，並列文などの構文に依存した素性の導入をおこなった．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[11]を利用し，先行詞になる確率を EM アルゴリズムを用いて混合正規分布として推定し，最も確率の高い先行詞を選ぶという方式を取った．

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 2 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．

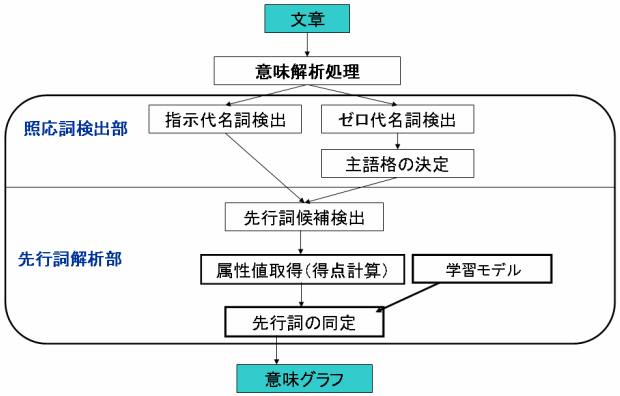

図 2　　ANASYS の処理の流れ

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．ただし，断定節でもサ変名詞のものや，動詞節でも一部の機能表現はゼロ代名詞としては扱わない．

2.1.1 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)となり得る格は 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．この主語格を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

照応詞の共起レコードで表層格がガ格でありかつ深層格が a-object 格であるレコードが少なくとも 1 つ存在すれば，主語格を a-object とする．そのようなレコードが存在しない場合は，共起レコード中の agent 格と o-agent 格の出現数を比べ，agent 格が多ければ agent 格とする．o-agent 格が多ければ，o-agent 格とする．

２）動詞節と動名詞節の場合

照応詞が原文中に scene 格や place 格など別の深層格での係り受けを持つ場合には，照応詞の共起レコード中に同じ格での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」の object 格として「船」が存在している．

⑤照応詞「入る」の主語格を o-agent 格とする．

2.2 先行詞検出部

2.2.1 先行詞候補の検出

本文中からは，先行詞は照応詞の近くに存在することが多いという理由から，図３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(名詞と断定の助動詞からなる)断定節を先行詞候補とする．先行詞候補は，照応詞に直接係っている文節は含まない．また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．図３において照応詞は青く塗られている「勉強していたと」であり，赤の枠で示されている範囲が先行詞候補の検出範囲，オレンジの枠で示されている「イチロー」がタイトル，緑の枠で示された範囲が主題の検出範囲である．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る．コーパスでは外

Vol.2009-NL-191 No.5
Vol.2009-SLP-76 No.5
2009/5/21

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種類とする．

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，でき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，物など意志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．

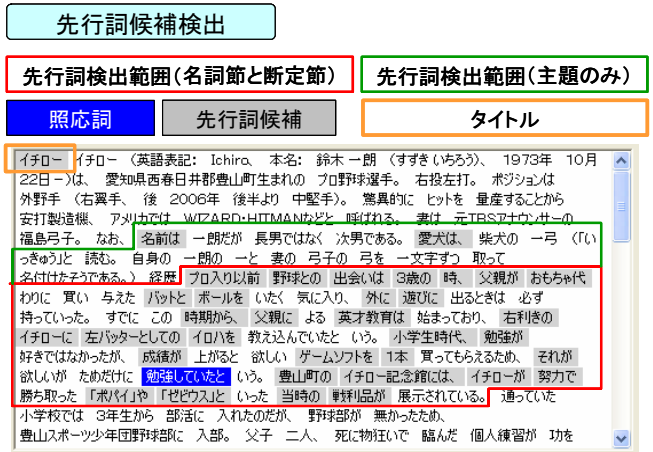

図 3　先行詞候補検出範囲の例

2.3　素性取得(得点計算)

図 4 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，主題得点，固有名詞得点，同一主語得点，主語格得点，係り受け距離得点の合計 7 個の素性を用意した．以下でそれぞれの得点について説明する．

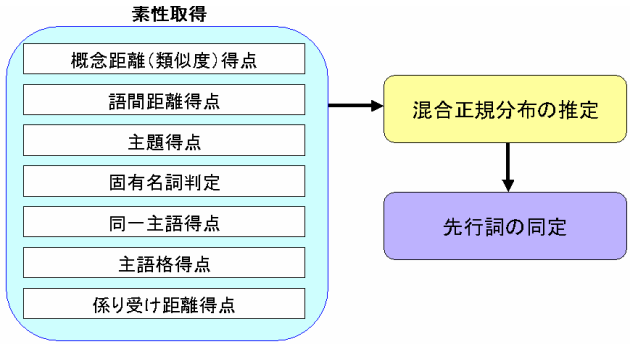

図 4　7 個の素性を元にした混合正規分布確率による先行詞の同定

1）概念距離得点
先行詞候補と照応詞との主語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る．具体的には照応詞の共起レコード中の係り側の概念と先行詞の概念の概念類似度を求める．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．まず共起関係子「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．外界については表 1 に示す概念を代わりに用いる．

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．外界は意味解析によって得られるモダリティに基づいて距離を経験的に設定した．

3）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．例えば先行詞が持つ場合は 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．外界は一人称・人が主題になりやすいと考え，0.7 を与えている．

4）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．固有表現を持つ語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．

Vol.2009-NL-191 No.5
Vol.2009-SLP-76 No.5
2009/5/21

5）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 格や 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，Ａの同一主語得点に１を与える．
6）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は Vs の先行詞になりやすいため，主語格得点に１を与える．
7）係り受け距離得点
文節の数で距離を測る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木において照応詞 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．外界は語間距離得点同様モダリティを用いて距離を経験的に設定した．

表 1 外界の概念

クレーム	
外界(一人称)	2dc301: c#I
外界(二人称)	3ce735: 営利事業を目的として設立した社団法人
外界(人)	103c4d:人間
外界(物)	1e850b:人力以外の動力によって作動し,一定の仕事をする仕組みの装置
外界(事)	3aa963:状態
小説、新聞、wiki	
外界(一人称)	2dc301: c#I
外界(二人称)	2dc303:c#you
外界(人)	103c4d:人間
外界(物)	1e850b:人力以外の動力によって作動し,一定の仕事をする仕組みの装置
外界(事)	3aa963:状態

2.3.1 先行詞特定

全ての先行詞候補に対して前出の 7 種類の素性の値を取得後，それらの素性ベクトル $\vec{x}$ を用いて先行詞を 1 つに決定する．この為に，確率モデルとして混合正規分布を仮定し，正例，負例それぞれに分けた学習データより,Weka[12]のＥＭアルゴリズムを用いて，そのパラメータ($a_i, \mu_j^i, \sigma_{jj}^i$)を推定する．ここで得られたパラメータを用いて各先行詞候補について正例を元に学習した確率分布 p positive($\vec{x}$)と負例を元に学習した確率分布 p negative($\vec{x}$)を計算する．その値より，以下の式で最終的なスコア(Score($\vec{x}$))を計算し，スコアが最も大きくなるものを正解として１つ選択する．計算式を以下に示す． x_j は j 番目の素性を表す．

$$p(\vec{x}) = \sum a_i \cdot p_i(\vec{x})$$

$$p_i(\vec{x}) = \prod_j \frac{1}{\sqrt{2\pi}\sigma_{jj}^i} \exp\left\{ -\frac{(x_j - \mu_j^i)^2}{2(\sigma_{jj}^i)^2} \right\}$$

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

得られたスコアは，前に定めておいた閾値によって評価できる．最大スコアをもつ先行詞であっても閾値を下回るスコアしか持たない場合は照応の対象外とし，高いスコアを得た信頼性の高いもののみを抽出することもできる．

3. 先行詞特定の学習データ作成

EM アルゴリズムでの学習のために，正例・負例それぞれについて分野別に，物語文，新聞記事，辞典文，クレーム文の合計 8 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,新聞記事については NAIST テキストコーパスから 280 記事 7077 文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．このコーパスには述語と表層格（ガ格，ヲ格，ニ格）の関係，事態性名詞と表層格（ガ格，ヲ格，ニ格）の関係，名詞句間の共参照の情報が付与されている．この中で，述語と表層格（ガ格）の関係を本研究に用いている．また，新聞以外の分野については物語文 106 文，辞典文 94 文，クレーム文 41 文について，筆者らが人手によって解析したデータを学習データとして利用した．学習データの一例を表 2 に示す．学習データは 2.2.2 で述べた素性の値から成り，各値はそれを正規化したものである．

表 2　学習データ例

GainenKyo	GokanKyo	Koyu	Syudai	Doitsu	SyugoKaku	Kakariuke
0.710041	−0.13933	−0.38868	0.725567	−0.06763	−0.27651	−0.49487
−0.34564	0.633414	−0.38868	1.086521	−0.06763	−0.27651	0.657598
0.808512	0.633414	−0.38868	1.086521	−0.06763	−0.27651	0.657598
0.632745	0.633414	−0.38868	1.086521	−0.06763	−0.27651	0.657598
−0.4048	0.633414	−0.38868	1.086521	−0.06763	−0.27651	0.657598
−0.71472	0.633414	−0.38868	1.086521	−0.06763	−0.27651	0.657598
0.957789	1.406293	−0.38868	0.31305	−0.06763	−0.27651	3.538772
−1.09179	0.633414	−0.38868	1.086521	−0.06763	−0.27651	0.657598
−0.21687	0.633414	−0.38868	1.086521	−0.06763	−0.27651	0.657598
−0.3093	0.066506	−0.38868	1.086521	−0.06763	−0.27651	0.427105

新聞記事に対する学習データを作成する処理の流れを図 5 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）意味グラフと，コーパスの照応関係が書かれている XML データに対して ANASYS と

Vol.2009-NL-191 No.5
Vol.2009-SLP-76 No.5
2009/5/21

同様の処理を行って各先行詞候補の各素性を求め，コーパスの正解データに合致するものは正例とし，そうでないものを負例として，学習データを作成する．この学習データの形式は Weka に適した形で出力する．3）このデータを Weka に入力することで，混合正規分布のパラメータを推定する．

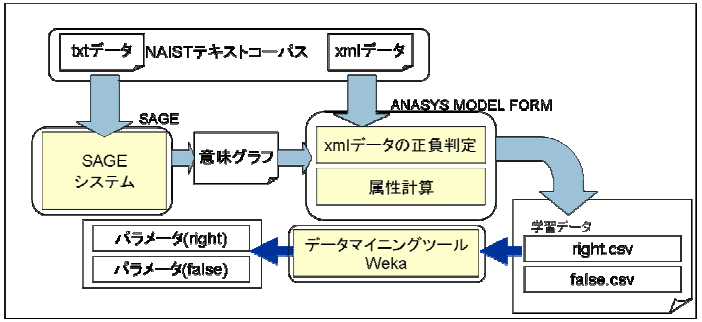

図 5　学習データ作成の流れ

4. 評価実験

4.1 先行詞特定の事例

具体的な先行詞特定の事例を挙げる．

図 6 と表 3 で示す事例は新聞記事の 1 記事を解析した結果である．「(前略)…だが，日本政府の対応に疑義がなかったわけではない．なおきちんと説明されていない問題を，少なからず残している．…(後略)」という文章中において，照応詞である動詞節「残している」について 20 以上の先行詞候補から正しい先行詞「日本政府の」を補完することができている．これは，ハ格を取らない事例だが，素性の中で概念類似度得点が高く，効果的に影響した事例である．

図 7 と表 4 で示す事例は新聞記事の 1 記事を解析した結果である．「(前略)…美容サロンを経営するレイラさんは健康維持のためボクシングを始め，1 年前にプロボクサー転向を決意した．178 センチ，76 キロ．…(後略)」という文章中において，照応詞である断定節「76 キロ．」について 20 以上の文中の先行詞候補から正しい先行詞「レイラさんは」を補完することができている．この事例では，断定節「76 キロ．」について正解の「レイラさんは」の概念類似度得点は低いが主題得点，係り受け距離得点などによって正解を得ることができている．

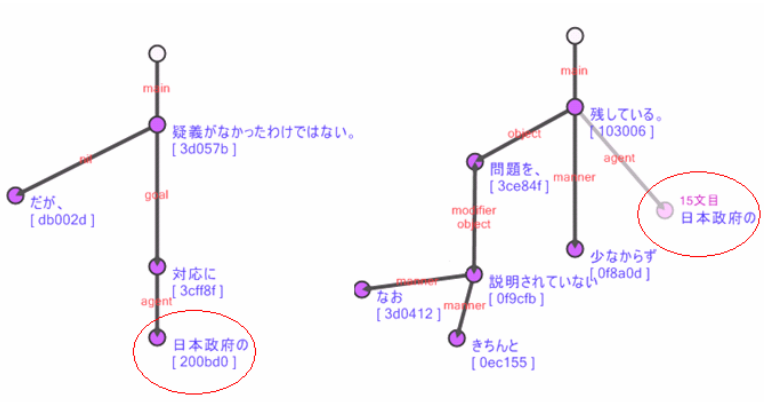

図 6　先行詞特定の事例 1(ViVi での出力)

表 3　先行詞特定の事例 1(得点)

	残している	格	agent						
先行詞候補	スコア	概念類	語間距	主題得	固有名	同一主	主語格	係り受	判定
事態は	0.538864	−0.72	−0.3	0.92	2.573	−0.07	−0.28	−0.7	FALSE
北朝鮮は、	0.979469	1.21	−1.48	1.63	2.573	−0.07	3.616	−0.49	FALSE
期限までに	8.34E−07	−1.09	−1.33	−0.72	−0.39	−0.07	−0.28	−0.49	FALSE
合意に	5.36E−05	0.38	−1.18	−0.72	−0.39	−0.07	−0.28	−0.61	FALSE
実験炉の	0.000164	0.79	−0.88	−0.72	−0.39	−0.07	−0.28	−0.61	FALSE
凍結解除・	0.000404	−1.09	−0.72	0	−0.39	−0.07	−0.28	−0.49	FALSE
KEDOの	0.403936	1.21	−0.42	−0.72	2.573	−0.07	−0.28	−0.35	FALSE
発足は、	0.057852	−1.09	−0.27	2.63	−0.39	−0.07	−0.28	−0.17	FALSE
米朝合意の	0.466762	0.38	−0.12	−0.72	2.573	−0.07	−0.28	−0.49	FALSE
前進を	0.077438	−1.09	0.03	0.31	−0.39	−0.07	−0.28	−0.35	FALSE
日本政府の	0.99071	0.91	0.64	−0.72	−0.39	−0.07	3.616	0.08	TRUE
対応に	0.023632	−1.09	0.79	−0.72	−0.39	−0.07	−0.28	0.43	FALSE
軽水炉転換	0.021049	−1.09	0.28	−0.72	−0.39	−0.07	−0.28	0.08	FALSE
資金は、	0.107084	0.35	−0.68	1.09	−0.39	−0.07	−0.28	0.43	FALSE
韓国と	0.818863	1.21	−1.27	−0.72	2.573	−0.07	−0.28	0.43	FALSE
日本が	0.87697	1.21	−1.53	0.83	2.573	−0.07	−0.28	0.43	FALSE
外界(一人称	0.506791	0.86	0.07	1.34	−0.39	−0.07	−0.28	0.95	FALSE
外界(二人称	0.178278	0.86	−0.5	0.83	−0.39	−0.07	−0.28	0.43	FALSE
外界(事)	0.25171	−0.68	0.63	0.83	−0.39	−0.07	−0.28	0.43	FALSE
外界(人)	0.648526	0.88	0.63	1.34	−0.39	−0.07	−0.28	0.95	FALSE
外界(物)	0.411103	0.79	0.63	0.83	−0.39	−0.07	−0.28	0.43	FALSE

Vol.2009-NL-191 No.5
Vol.2009-SLP-76 No.5
2009/5/21

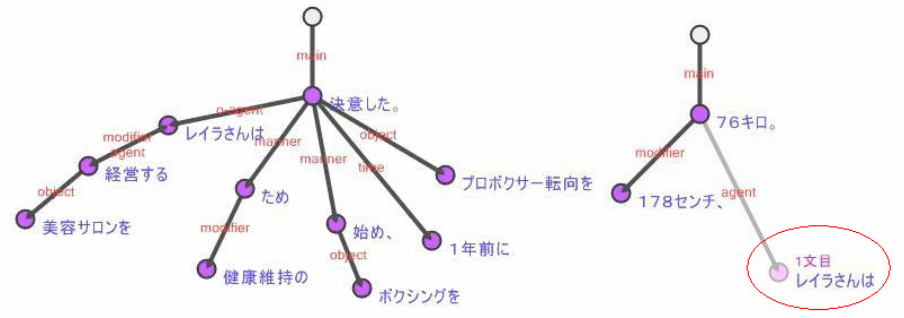

図 7　先行詞特定の事例 2(ViVi での出力)

表 4 先行詞特定の事例 2(得点)

	76キロ。格		agent						
先行詞候補	スコア	概念類	語間距	主題得	固有名	同一主	主語格	係り受け	判定
ボクシング　元世界ヘビー	0.0102	−1.09	−1.52	−0.72	−0.389	−0.068	−0.277	−0.7	FALSE
モハメド・アリ氏の	0.0059	−1.09	−1.4	−0.72	2.5728	−0.068	−0.277	−0.61	FALSE
娘が	0.4981	0.86	−1.28	0.73	−0.389	−0.068	−0.277	−0.49	FALSE
プロデビュー。	0.0085	−0.14	−1.16	−0.72	−0.389	−0.068	−0.277	−0.35	FALSE
元世界ヘビー級チャンピ	0.0011	−1.09	−0.92	−0.72	−0.389	−0.068	3.6165	−0.49	FALSE
モハメド・アリ氏	0.0648	−1.09	−0.8	−0.72	2.5728	−0.068	−0.277	−0.49	FALSE
末娘レイラさん	0.6443	−1.09	−0.56	1.34	−0.389	−0.068	3.6165	−0.35	FALSE
ニューヨークで	0.4995	−1.09	−0.09	−0.72	2.5728	−0.068	−0.277	−0.35	FALSE
プロボクサーとして	0.0373	−1.09	0.03	−0.72	−0.389	−0.068	−0.277	−0.35	FALSE
美容サロンを	0.4213	−1.09	0.51	−0.21	−0.389	−0.068	−0.277	−0.17	FALSE
レイラさんは	1	−1.09	0.75	2.63	−0.389	−0.068	−0.277	0.43	TRUE
健康維持の	0.2024	−0.12	0.87	−0.72	−0.389	−0.068	−0.277	0.08	FALSE
ため	0.5272	−1.09	0.99	−0.72	−0.389	−0.068	−0.277	0.43	FALSE
ボクシングを	0.8808	−0.14	1.11	0	−0.389	−0.068	−0.277	0.08	FALSE
プロボクサー転向を	0.9937	−0.14	1.46	0.31	−0.389	−0.068	−0.277	0.43	FALSE
練習は	0.9916	−0.14	0.28	1.09	−0.389	−0.068	−0.277	0.43	FALSE
週	0.0451	−1.09	−1.27	−0.72	−0.389	−0.068	−0.277	0.43	FALSE
外界(一人称)	0.9904	−1.09	0.07	1.09	−0.389	−0.068	−0.277	0.43	FALSE
外界(二人称)	0.9561	−1.09	−0.5	0.83	−0.389	−0.068	−0.277	0.43	FALSE
外界(事)	0.9813	1.15	0.63	0.83	−0.389	−0.068	−0.277	0.43	FALSE
外界(人)	0.9987	−1.09	0.63	1.09	−0.389	−0.068	−0.277	0.66	FALSE
外界(物)	0.9927	−1.09	0.63	0.83	−0.389	−0.068	−0.277	0.43	FALSE

図 8 と表 5 に示す事例も新聞記事の例である．「自衛隊は現在，那覇を含め全国五カ所にＡＳＷＯＣを持ち，日本の周辺海域を行動する極東ロシア軍などの潜水艦をＰ３Ｃで捜索し，その艦名や動向などを調べている．」という文章中において，照応詞は「持ち」と「捜索し」である．本研究では意味解析結果を用いて解析を行うため，これら二つの照応詞が動詞節「調べている」と並列関係にあるという情報が与えられている．同一文中の並列関係にある動詞の主語は同じになりやすく，これを表す同一主語得点が効果を発揮して，正しく先行詞「自衛隊は」を補完することができている．

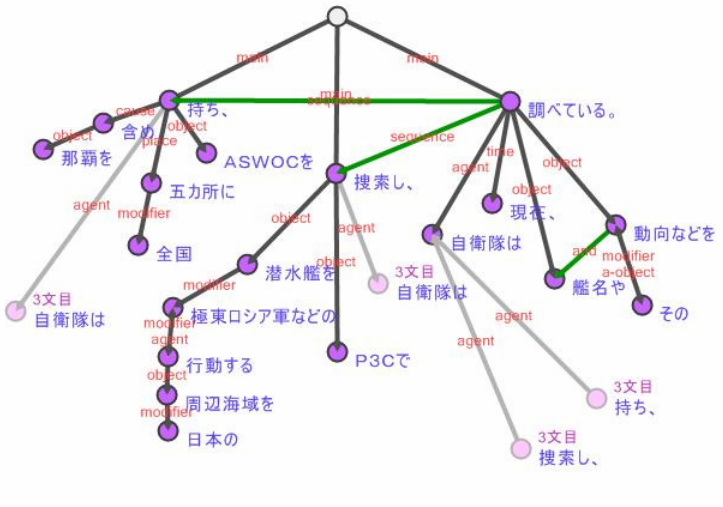

図 8 先行詞特定の事例 3(ViVi での出力)

表 5 先行詞特定の事例 3(得点)

	捜索し、	格	agent						
先行詞候	スコア	概念類	語間距	主題得	固有名	同一主	主語格	係り受け	判定
自衛隊は	1	0.36	−1.4	2.63	−0.389	14.785	3.6165	1.81	TRUE
現在、	0.159283	−1.09	−1.16	−0.72	−0.389	−0.068	−0.277	1.81	FALSE
那覇を	0.721622	0.27	−0.92	−0.21	2.5728	−0.068	−0.277	0.95	FALSE
全国	0.481488	−1.09	−0.44	−0.72	−0.389	−0.068	−0.277	0.95	FALSE
ASWOCを	0.938948	−1.09	0.03	0	−0.389	−0.068	−0.277	1.81	FALSE
日本の	0.831998	0.3	0.51	−0.72	2.5728	−0.068	−0.277	0.43	FALSE
周辺海域を	0.997287	−1.09	0.75	0.31	−0.389	−0.068	−0.277	0.95	FALSE
極東ロシア	0.996185	−0.12	1.22	−0.72	2.5728	−0.068	−0.277	3.54	FALSE
艦名や	0.32385	−1.09	1.27	−0.72	−0.389	−0.068	−0.277	1.81	FALSE
動向などを	0.960765	−1.09	0.81	0.06	−0.389	−0.068	−0.277	1.81	FALSE
那覇防衛	0.60486	0.18	−0.23	1.09	2.5728	−0.068	−0.277	0.43	FALSE
建築基準	0.019238	−1.09	−1.03	−0.72	−0.389	−0.068	−0.277	0.08	FALSE
同センター	0.01615	−1.09	−1.43	−0.72	−0.389	−0.068	−0.277	0.08	FALSE
建築工事	0.26844	−1.09	−1.53	−0.18	−0.389	−0.068	−0.277	0.43	FALSE
那覇市に	0.292123	0.36	−1.58	−0.72	2.5728	−0.068	−0.277	0.43	FALSE
外界(一人	0.973843	−0.48	−0.63	1.09	−0.389	−0.068	−0.277	0.43	FALSE
外界(二人	0.927969	−0.48	−0.97	0.83	−0.389	−0.068	−0.277	0.43	FALSE
外界(事)	0.965021	−1.09	−0.3	0.83	−0.389	−0.068	−0.277	0.43	FALSE
外界(人)	0.992907	−0.44	−0.3	1.09	−0.389	−0.068	−0.277	0.66	FALSE
外界(物)	0.965021	−1.09	−0.3	0.83	−0.389	−0.068	−0.277	0.43	FALSE

Vol.2009-NL-191 No.5
Vol.2009-SLP-76 No.5
2009/5/21

図 9 と表 6 に示す事例は外界照応が成功している事例である.「…(前略)製品による事故で損害を受けた場合，従来ならばメーカー側の責任追及では「過失」を立証しなければならなかった．…(後略)」という文章中において，照応詞「受けた場合」と「立証しなければならなかった.」はそれぞれ文章中に書かれていない一般的な人が主語となる．ANASYS では外界を先行詞候補にでき，概念類似度・主題得点などによって先行詞「外界(人)」を同定することができている．

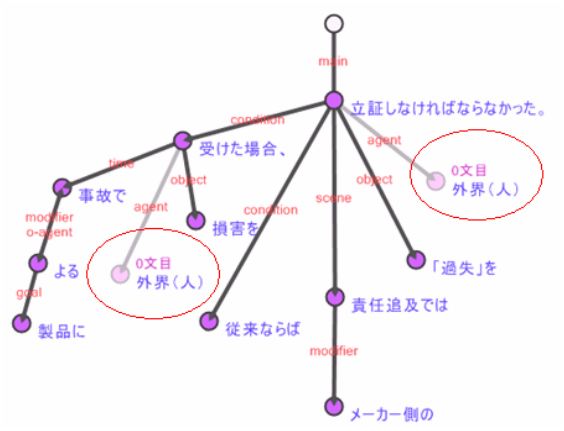

図 9　先行詞特定の事例 4 外界の事例(ViVi)

表 6　先行詞特定の事例 4 外界の事例(得点)

	受けた場合格		agent						
先行詞候補	スコア	概念類	語間距	主題得	固有名	同一主	主語格	係り受け	判定
製造物責任	0.300939	0.53	−1.38	1.34	2.5728	−0.068	−0.277	−0.35	FALSE
消費者の	0.00625	1.37	−0.61	−0.72	−0.389	−0.068	−0.277	−0.61	FALSE
立場に	0.005491	1.76	−0.35	−0.72	−0.389	−0.068	−0.277	−0.49	FALSE
製品に	0.437937	1.27	0.93	−0.72	−0.389	−0.068	−0.277	1.81	FALSE
メーカー側	0.550463	1.78	1.1	−0.72	−0.389	−0.068	−0.277	1.81	FALSE
責任追及で	0.737819	−0.24	0.49	−0.72	−0.389	−0.068	−0.277	3.54	FALSE
「過失」を	0.52725	−0.34	−0.14	0.06	−0.389	−0.068	−0.277	3.54	FALSE
ハイテク時	0.01859	−1.09	−1.08	−0.72	−0.389	−0.068	−0.277	0.08	FALSE
消費者が	0.595115	1.37	−1.35	0.83	−0.389	−0.068	−0.277	0.08	FALSE
立証するこ	0.013096	−0.24	−1.58	−0.72	−0.389	−0.068	−0.277	0.08	FALSE
外界(一人	0.968172	1.5	0.07	1.09	−0.389	−0.068	−0.277	0.43	FALSE
外界(二人	0.881303	1.32	−0.5	0.83	−0.389	−0.068	−0.277	0.43	FALSE
外界(事)	0.982543	1.06	0.63	0.83	−0.389	−0.068	−0.277	0.43	FALSE
外界(人)	0.98289	1.98	0.63	1.09	−0.389	−0.068	−0.277	0.43	TRUE
外界(物)	0.981067	1.17	0.63	0.83	−0.389	−0.068	−0.277	0.43	FALSE

4.2　照応詞判定の精度評価

本研究では，新聞記事については NAIST テキスト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 609 記事 18721 文を用いて評価した．結果を表 7 に示す.

表 7　照応詞判定における精度評価

	適合率	再現率	F値
新聞	80.53% (13823/17166)	98.97% (13823/13967)	88.8

4.3　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，分野ごとに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，TinySVM[13]，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．閾値設定は手作業で最適な値と定めたものを使用し，2.2.3 先行詞特定において得られるスコアが閾値以上のもののみを照応解析対象とし，閾値以下のものは解析対象外とすることとした．それぞれ結果を表 8 と表 9 に示す．表 8 中の ANASYS2007 は 2007 年度の論文における実験結果[3]で，概念類似度得点，語間距離得点，主題得点，固有名詞得点からなる 4 つの素性と SVM によって学習している．　ANASYS2008S は本論文で述べた 7 個の素性を用いて SVM で学習した場合で，ANASYS2008E は同様の 7 個の素性を用いて EM アルゴリズムで学習した場合である．ANASYS2008ET はスコアが経験的に設定した閾値を上回った場合のみを照応の対象とした場合である．なお表 8 において外界無というのは正解先行詞が外界になる照応詞を評価データから外した場合である．有の場合はすべてのデータを評価対象とした場合である．なお両者において推定された統計モデルは同じものを使った．表 9 は同様に[]内が ANASYS2007 による実験結果である.

Vol.2009-NL-191 No.5
Vol.2009-SLP-76 No.5
2009/5/21

表 8　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
ANASYS2007	有	21.51% (1372/6374)	24.55% (1372/5589)	22.9
ANASYS2008S	有	30.72% (5273/17167)	37.76% (5273/13965)	33.9
	無	32.53% (4165/12802)	43.37% (4165/9603)	37.2
ANASYS2008E	有	33.32% (5720/17167)	40.95% (5720/13965)	36.7
	無	33.80% (4327/12802)	45.06% (4327/9603)	38.6
ANASYS2008ET	有	49.15% (1035/2106)	7.4% (1035/13965)	12.8
	無	54.18% (1035/1591)	10.7% (1035/9603)	17.8

表 9　先行詞判定における精度評価(新聞以外)

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]

5. 結論

先行詞判定における精度では，新聞において本研究の 7 個の素性を用いたところ，ＳＶＭによる学習，ＥＭアルゴリズムによる学習それぞれについてＦ値が 30 を上回る精度を得た．これは，同一主語得点など文構造を用いた素性を導入したことにより，文内照応の精度が大きく向上したため ANASYS2007 に比べて 10%以上の精度向上に繋がったと思われる．SVM に比べ EM アルゴリズムを用いた場合の方が 3%ほど精度が高かったことは，照応解析における EM アルゴリズムによる学習の有効性を示しているといえる．また，ＥＭアルゴリズムによって先行詞になりやすい確率を直接求めることで，閾値を設定して信頼性の高いものを選ぶことができるようになった．その結果 50%近い精度を出すことができた．しかし適合率を上げるために事例を絞り込むと，再現率は下がるためＦ値が低くなってしまった．得点の調整などにより信頼性の高い事例の数を増やすことでこの問題は解決できる．本研究では外界照応を含んだ照応解析を行うため，外界の素性をどう設定するかが結果に大きく影響を与える．語間距離得点や主題得点など外界の素性の得点を 0.1 変えるだけでも解析結果に大きな影響を与え，外界を正解として取りすぎてしまったり，逆に殆ど外界が選ばれなくなってしまうこともある．よって外界無しの解析に比べて高い精度を得ることが困難であるといえる．このためには，外界については，コーパス中で外界を先行詞にとる動詞を共通上位概念などでまとめなれないかなどを調査すべきである．

参考文献

1) 原田実, 尾見孝一郎, 岩田隆志, 水野高宏: 日本語文章からの意味フレーム自動生成システム SAGE(Semantic frame Automatic GEnerator)の開発研究, 人工知能学会第 13 回全国大会論文集, pp. 213-216 (1999).
2) 原田実，水野高宏:　EDR を用いた日本語意味解析システム SAGE，人工知能学会論文誌, Vol.16, No.1, pp.85-93 (2001.1).
3) 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, 言語処理学会第 14 回年次大会発表論文集, E3-2 , pp. 552-555 (2008.3).
4) 西尾公秀, 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 EM アルゴリズムによる学習の導入, 卒業論文, 青山学院大学(2008).
5) 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).
6) 情報処理学会研究会報告, 95-NL-107, pp. 91−96 (1995)
7) 飯田龍, 乾健太郎, 松本裕治. 文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定, 情報処理学会論文誌, Vol .45, No. 3, (2004).
8) 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, pp.488-491.(2006).
9) 飯田龍, 乾健太郎, 松本裕治. 結束性と首尾一貫性から見たゼロ照応解析. 情報処理学会自然言語処理研究会予稿集, NL-178-7. pp.45-52. (2008).
10) (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(2002).
11) NAIST Text Corpus：http://cl.naist.jp/nldata/corpus/
12) Weka: http://www.cs.waikato.ac.nz/ml/weka/
13) TinySVM: http://chasen.org/~taku/software/TinySVM/