文脈的手がかりを考慮した機械学習によるゼロ照応解析

飯田 龍 乾 健太郎 高村 大也 松本 裕治
奈良先端科学技術大学院大学
{ryu-i,inui,hiroya-t,matsu}@is.aist-nara.ac.jp

1 はじめに

自然言語では通常，読み手もしくは書き手に容易に判断できる要素は，文章上表現を簡略化する，あるいは省略する場合が多い．このような省略を文脈から補完するゼロ照応解析は，文脈解析において特に重要である．これまでの照応解析の手法はおおきく理論指向の規則作成に基づく手法とコーパスを用いた学習手法に分類できる．

規則作成に基づく解析手法では，さまざまな言語的な手がかりを人手で規則に取り入れる試みが行われている [8, 2, 18, 17, 21]．この手法では，対象となる名詞句の意味役割や先行詞候補の出現順序，照応詞と先行詞の間の意味的な互換性などの手がかりに加え，センタリング理論 [4, 13, 7] のような言語学的な知見をもとに規則を記述する．MUC-7 [1]における照応解析のタスクでは，約70%の適合率と約60%の再現率が報告されているが，機械翻訳などの現実的な応用を考えた場合，満足できる精度とは言えない．さらに，規則が特定のドメインに特化している場合は，他のドメインで同様の精度を得ることが難しい．このような事実を考慮すると，人手による規則の洗練は難しく，コストも大きいと考えられる．

これに対し，照応タグ付きコーパスを用いた統計的な手法 [1, 11, 10, 16] は，コストが低いという利点を持ちながらも，MUC-6 や MUC-7 の照応解析の評価セットを用いた実験で規則ベースの手法と同程度の精度を得ている．しかし，これまでの統計的手法は，照応に関して言語学で研究されてきた知見を考慮していないという問題がある．

そこで本稿では，統計的手法にセンタリング理論のような言語学的知見を取り入れた手法を提案する．2節では決定木学習を用いた Soon ら [11] の照応解析のモデルを示し，その後，このモデルを改良した Ng ら [10] のモデルについて述べる．3節では，Ng らのモデルの欠点を述べ，この欠点を補うために，センタリング理論の考えを考慮した素性 (センタリング素性) を導入するとともに，先行詞同定のための新たな探索モデル (トーナメントモデル) を提案する．次に4節では，日本語ゼロ照応を解消する実験を行い，先行研究と提案手法の比較を行う．最後に5節で現在のモデルのの問題と今後の方針について議論する．

2 先行研究

機械学習を用いた照応解析はすでにいくつかの手法が提案されており，例えば Soon ら [11] や Ng ら [10] のモデルは規則ベースの手法と同程度の精度を得ている．

Soon らのモデルは，照応解析の問題を，与えられた照応詞に対して，先行詞の候補となる名詞句の各々が先行詞となるかならないかを判別する2値分類問題に分解する．図1を用いて説明しよう．図1では，照応詞 *ANP* に対して，7つの名詞句 ($NP_1, \ldots, NP_7$) が先行文脈に出現している状況を仮定している．NP_2 と NP_4，NP_3 と NP_5，NP_6 と NP_7 はそれぞれ照応関係にあり，*ANP* の先行詞は NP_5 (NP_3) とする．この状況で，分類器は名詞句 NP_i ($i \in \{1, \ldots, 7\}$) が先行詞かどうかという2値分類問題を解く．

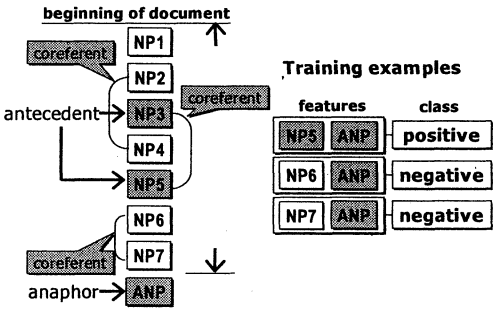

図 1: 訓練事例の作成 [11, 10]

訓練時には，照応詞から最も近い先行詞と照応詞の対 (*ANP-NP*$_5$) を正例，先行詞と照応詞の間の各名詞句と照応詞の対 (*ANP-NP*$_6$, *ANP-NP*$_7$) を負例として学習する．新しい照応問題を解く際には，訓練時と同様に，照応詞から先行文脈に向かって，先行詞候補となる名詞句の一つ一つについて，それが先行詞かどうか分類していく．そして，分類器がいずれかの名詞句を先行詞として決定した時点で解析を終了する．分類器が，先行する名詞句をすべて先行詞ではないと分類した場合は，対象としている照応詞は先行詞を持たないと判断する．Soon らの実験では，12個の限られた素性を用い，C5.0 を使用して決定木学習を行っている．

Ng ら [10] は Soon らの手法を2つの点において改良した．1つは素性集合を拡張し，語彙的な素性や意味的な素性など，53個に素性を増やした．もう1つは先行詞同定の探索アルゴリズムの変更である．Soon らが照応詞に近い名詞句から順に先行詞かどうかを決定的に決めるのに対し，Ng らはすべての先行する名詞句を分類器にかけ，分類器が先行詞と決定した名詞句の中で，最も先行詞らしいと判定した名詞句を先行詞とする．ここでも，すべての名詞句が先行詞でないと分類された場合には，照応詞は先行詞を持たないと判断する．Ng らのモデルは Soon らのモデルよりも先行詞同定の精度がよく，後述する日本語ゼロ照応解析の実験においても同様の結果と

[1] The Seventh Message Understanding Conference (1998): www.itl.nist.gov/iaui/894.02/related_projects/muc/

なった．そこで，Ngらのモデルを我々が提案する手法と比較する際の基準とする．

3 提案手法

3.1 先行研究の問題点

以下の2つ例文を用いて，SoonらやNgらのモデルの問題を考察する．

(1) a. メアリはジョン$_i$に会いに行った。
　　b. 彼$_i$は野球をしていた。

(2) a. トム$_i$はジョンに会いに行った。
　　b. 彼$_i$は昨日起こったことを説明しようとした。

(1)では，(b)の主題「彼」は(a)の目的格「ジョン」を指している．一方，(2)では，「彼」と「ジョン」がそれぞれ(1)と同じ意味役割であるにもかかわらず，「彼」が「ジョン」を指していない．この違いについてセンタリング理論では以下のように解釈する．(2)では，「トム」は前文の主題であるのでpreferred center ((a)のforward-looking centerの中で最も上位に位置する対象)となり，最も先行詞となりやすい．そのため，(b)では「トム」は代名詞で表現されなければならないが，「彼」=「トム」という解釈はこれと整合する．それに対して(1)では，「メアリ」がpreferred centerとなっているが，「彼」とgenderが一致しないため，2番目の候補である「ジョン」が先行詞として解釈される．

上述の解釈の重要な点は，センタリング理論のモデルが先行詞候補間の相対的な優先度を考慮している点である．上の例では，「ジョン」が照応関係にあるかどうかは「メアリ」や「トム」のような，局所的文脈中の他の要素の存在に依存している．このように，文脈中の他の要素との関係を考慮することが重要であると考えられる．しかし，Ngらのモデルでは，先行詞候補と照応詞だけを見て先行詞かどうかの2値分類を行っているために，周りの文脈の情報を扱えていない．

3.2 文脈の局所性を扱う2つの解決法

上述の問題に対してさまざまな解決策が考えられる．我々はこれまでに次の2つの解決策を試みた．

3.2.1 センタリング素性

最も直観的な解決法の一つは，素性集合に局所的な文脈情報を扱う素性を追加することである．このような素性をセンタリング素性と呼ぶことにし，以下にゼロ照応にセンタリング素性を導入する一例を示す．

まず，センタリング素性を定義するために，Nariyama [9]によって提案された日本語ゼロ代名詞解析の理論を示す．Nariyamaの理論は，センタリング理論を日本語の照応解析に適用したKameyama [7]の研究を拡張した理論である．前文のみしか扱えないセンタリング理論の一般的な考え方に対し，Nariyamaの導入したSalience Reference List (SRL)では，先行するすべての先行詞候補をゼロ代名詞の対象とすることができる．SRLでは，多くのセンタリング理論を扱ったモデルと同様に，**主題**(“は”，ゼロ)**>焦点**(ガ格)**>間接目的**(ニ格)**>直接目的**(ヲ格)**>その他**の順序で先行詞候補を保持する．SRLに先行詞候補を保持する際には，文章の先頭から順に出現した格要素を保持し，同じ格要素が出現した場合には，新しい要素を上書きする．

このSRLを用いることで，局所的な文脈を考慮できる例を示す．以下の例では，下線部のかけのガ格が省略されている．

> 兵庫県警は二日、サイコロとばくに客として加わったとして、同県高砂市緑丘二の同市教委スポーツ振興課副課長、清谷亨容疑者ら四人を常習とばくの疑いで逮捕した。調べでは、清谷容疑者ら四人は昨年三月二十三日夜から翌二十四日早朝にかけて、高砂市内のスナックで既にとばく開張図利容疑で逮捕されている山口組系暴力団幹部が開いたとばく場に参加、一回一万円から五十万円を(φガ)かけ、とばくをした疑い。

SRLでは最初に“は”で記された**兵庫県警**を主題として保持するが，途中で主題が遷移し，**四人**が新たな主題として保持される．最終的に省略の箇所まで計算されたSRLは，**四人>山口組系暴力団幹部>とばく場>五十万円>一万円**となり，最も優先度の高い「**四人**」がガ格の先行詞と決定される．このように，SRLに保持された情報を素性として扱うことで，局所的な文脈の情報を考慮できると考えられる．

また，Nariyamaのモデルでは複文における照応関係についても考慮されており，従属節の主語が“て”や“ながら”など特定のクラスの接続表現で主節に係る場合，主節も従属節と同じ主語になる強い傾向があるがあるため，同一文内でこのような関係となっているかどうかも素性として扱う．

3.2.2 トーナメントモデル

「ジョン」の例に戻って議論を進める．我々が考慮したい点は，「ジョン」に対して，「メアリ」か「トム」であるかの相対的な比較を行うことである．そのような比較を実現する方法の一つは，2つの先行詞候補間でどちらが先行詞らしいかの比較を行い，勝ち抜き方式で先行詞を決定する手法である．この手法をトーナメントモデルと呼ぶことにし，以下で詳細を述べる．

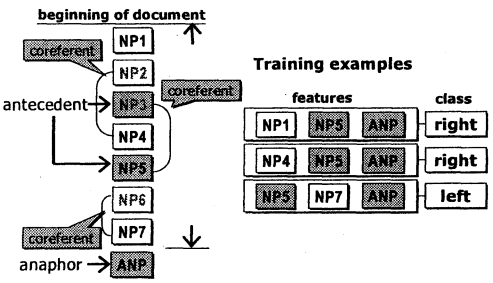

図2: トーナメントモデル

図1の状況を再び図2に描く．ここでは，すでに解析された照応関係を考慮し，*ANP*に対して4つの先行詞候補(*NP*$_1$，*NP*$_4$ (と照応関係にある*NP*$_2$)，*NP*$_5$ (*NP*$_3$)，*NP*$_7$ (*NP*$_6$))を扱う．勝ち抜き方式において，正しい先行詞である*NP*$_5$ (*NP*$_3$)は他の先行詞候補に対して勝ち残る必要がある．そのため，この関係を学習するために，図2に示してある3つの訓練事例を抽出した．クラスright (left)は与えられた先行詞の候補のうち，どちらの候補が勝ち抜けるか(先行詞らしいか)を示している．

勝ち抜き方式で解析を行う際には，先行詞候補となる名詞句の間で勝ち抜き戦を行う．勝ち抜き戦は照応詞から文章の先頭に向かって処理される．最初の比較では，最も照応詞に近い2つの候補(*NP*$_7$と*NP*$_5$)が比較され，分類器はより先行詞らしい名詞句を選択する．次の比較では，1つ前の比較において勝ち残った(より先行詞らしいと決定された)候補と新

たな先行詞候補との比較を行う．この処理を繰り返し，最後の比較では，文章の先頭に最も近い先行詞候補との比較を行い，勝ち残った候補を与えられた照応詞に対する先行詞と決定する．

この候補の比較を行うトーナメントモデルでは，センタリング理論の先行詞になるための順序を学習することが期待できる．例えば，(2)の「トム」と「ジョン」の例の場合，主格の要素が目的格の要素より先行詞になりやすいことを学習できる．またトーナメントモデルでは，2つの先行詞候補間の関係を素性として追加できるという利点がある．例えば，先行詞候補間の距離を素性として追加することができ，これによって候補間の距離が離れた場合，照応詞に近い要素が勝ち抜きやすいという性質を学習できる．

4 評価実験

この節では，日本語ゼロ照応解析の実験を行うことで，先行研究と提案手法のモデルを比較する．

4.1 訓練・評価データ

GDA[2]タグにはさまざまな統語・意味タグに加えて照応関係についてもタグが用意されており，評価実験では GDA タグでタグ付けされた新聞記事コーパスから訓練・評価のためのデータを抽出した．このコーパスは約 25,000 文を含み，約 20,000 箇所に照応関係のタグが付与されている．今回の実験では主題のゼロ代名詞に問題を限定して，2,155 事例を抽出しこのデータに対して 5 分割の交差検定を行った．

実験では，対象とする文章に対して茶筌 [20] と CaboCha [15] を用い形態構文解析を行い，また yanee [22] を用いて固有表現のタグを付与した．

4.2 素性

今回の実験で用いた 5 種類の素性 (grammatical, semantic, positional, heuristic, centering 素性) を表 1 に示す．grammatical, semantic, positional の 3 種類の素性は Ng らが用いた素性に概ね対応している．ただし，SELECT_REST, LOG_LIKE, CHAIN_LENGTH を新たに加えた．また，センタリング素性については 3 節で述べた素性を用いた (SRL_ORDER, SRL_ORDER_COMP, GA_REF).

4.3 実験結果

学習器として Ng らが決定木学習器 C5.0 を用いたのに対し，我々は汎化能力が高い Support Vector Machine (SVM) [12] を用いた．

実験の結果を図 3 に示す．結果より，Ng らの元のモデル (BM) に対してセンタリング素性を加えた Ng らのモデル (BM+CF) は，すべての学習事例を用いた場合，3%の精度の向上が見られた．また BM に対してトーナメントモデル (TM) では，訓練事例のサイズにかかわらず約 7%精度が向上した．精度が良くなった手法を組み合わせたモデル (TM+CF) は，少ないデータでは精度が悪い．しかし，今回の 4 つの実験の中で，訓練事例を増やした際の精度の上昇率が最も良いために，訓練事例を増やすことで精度向上が期待できる．

最も精度の良かったトーナメントモデルについて考察するために，解析の信頼度を導入する．まず，1 回の候補間の比較に関する信頼度として，1 つの候補に対してもう一つの候補がどのくらい先行詞らしいかを分類器が出力した値を用いる．その値を用いて，

[2]GDA (Global Document Annotation [19]) タグは計算機が文章の意味や語用について認識できるように作成されたタグセットである．

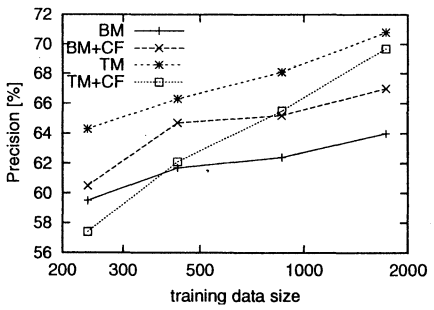

図 3: 学習曲線

BM: Ng らのモデル
BM+CF: センタリング素性を用いた Ng らのモデル
TM: トーナメントモデル
TM+CF: センタリング素性を用いたトーナメントモデル

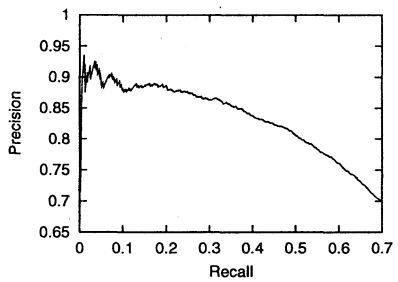

図 4: トーナメントモデルの Precision-recall 曲線

トーナメント全体の信頼度を，最後に勝ち残った候補が得た信頼度のうち，最も小さな信頼度の値とする．この全体の信頼度に基づき，評価事例をランキングすることにより Precision-Recall 曲線を描いた結果を図 4 に示す．照応解析の応用においては，誤って照応関係を解析するよりも少量の正しい解析結果のみを望む場合がある．そのような場合，この信頼度を用いて，再現率を犠牲にして適合率を効率良く向上させることができる．図 4 では，再現率を半分にすることで，適合率を 8 割まで上げることができることを示している．

5 おわりに

本稿では，言語学的な知見を考慮したセンタリング素性を用いて学習を行う手法と，先行詞候補間の関係を学習するトーナメントモデルの 2 つを提案し，この 2 つの改良が日本語ゼロ照応解析において効果的であることを示した．

今後の課題としては，まず，現在のモデルでは主題と副主題の構造を考慮していないために，“は”で記された副主題が誤って先行詞として選択されてしまうという問題がある．主題と副主題の関係は，今回提案したトーナメントモデルで効果的に扱うことができると考えられるので，我々は次の試みとして上記のような主題と副主題の関係を素性として組み込むことを考えている．別の課題として，ゼロ照応解析のためのより効果的な選択制限を考える必要がある．今回の実験で扱った log-likelihood 係数の計算には，コーパスに出現した名詞・動詞の文字列をそのまま用いたが，この名詞と動詞をどのように抽象化するかが選択制限を考える上で問題となる．また，

表 1: 実験に用いた素性

素性の種類	素性名	詳細
Grammatical	POS	"名詞-固有名詞", "名詞-サ変接続" のような NP_i の品詞.
	DEFINITE	NP_i がソ系の代名詞 ("それ", "その", "そんな" など) である場合は Y. それ以外は N.
	DEMONSTRATIVE	NP_i がコ系もしくはア系の代名詞 ("これ", "ここ", "あの", "あそこ" など) である場合は Y. それ以外は N.
	PARTICLE	"は", "が", "を" のような NP_i に続く助詞
Semantic	NE	NP_i の固有表現の種類: PERSON, ORGANIZATION, LOCATION, ARTIFACT, DATE, TIME, MONEY, PERCENT もしくは N/A.
	EDR_HUMAN	NP_i が EDR 概念辞書の中の "人間", "人間の属性" に含まれる語である場合は Y. それ以外は N.
	SELECT_REST	NP_i-*ANP* の対が日本語語彙体系で定義される選択制限を満たす場合は C. それ以外は I.
	LOG_LIKE	NP_i-*ANP* の対の log-likelihood 係数の値を 5 段階に分け, その値を付与.
	ANIMACY	NP_i が PERSON または ORGANIZATION である場合は Y. それ以外は N.
	ANIMACY_COMP*	NP_1 の ANIMACY が Y で NP_2 が N の場合は NP_1, 逆の場合は NP_2.
Positional	SENTNUM_ANP	NP_i と *ANP* の文間の距離. 同一文内の場合は 0.
	SENTNUM_NPS*	NP_1 と NP_2 の文間の距離. 同一文内の場合は 0.
	DEP_MAIN	NP_i が主節に係る場合は Y. それ以外は N.
	EMBEDDED	NP_i が連体句の中にある場合は Y. それ以外は N.
	BEGINNING	NP_i が文頭にある場合は Y. それ以外は N.
Heuristic	CHAIN_LENGTH	NP_i と照応関係にある名詞句の数.
Centering	SRL_ORDER	SRL の中での順位.
	SRL_ORDER_COMP*	NP_1 が NP_2 より高い優先度で順位付けされている場合は NP_1. 逆の関係の場合は NP_2.
	GA_REF	NP_i が従属節のガ格で, かつ特定の接続表現で主節に係っている場合は Y. それ以外は N.

ANP は照応詞を表し, $NP_{i\in\{1,2\}}$ は先行詞候補を表す. 素性は個々の要素についての素性と要素間の関係についての素性を含んでおり, 個々の要素についての素性は, 対象となっている NP_i に対してその性質を満たすか (YES) 満たさないか (NO) の 2 値をとる. 要素間の関係を表す素性は対象としている NP_1-NP_2 もしくは NP_i-*ANP* の対に対して, その性質が矛盾しない (COMPATIBLE), 矛盾する (INCOMPATIBLE) の 2 値をとり, その性質が適用できない場合は NOT APPLICABLE の値をとる. '*' で示された素性はトーナメントモデルでのみ使用できる素性である.

誤って解析された事例の中にはタグ付けの誤りも含まれており, 学習手法を頑健にすると同時にコーパスの質の向上も今後の課題としたい.

参考文献

[1] C. Aone and S. W. Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. *ACL*.

[2] B. Baldwin. 1995. CogNIAC: A Discourse Processing Engine. *Ph.D. Thesis, Department of Computer and Information Sciences, University of Pennsylvania*.

[3] N. Ge, J. Hale, and E. Charniak. 1998. A Statistical Approach to Anaphora Resolution. *WVLC*.

[4] B. J. Grosz, A. K. Joshi, and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discource. *Computational Linguistics*, 21(2).

[5] M. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman.

[6] T. Hofmann. 1999. Probabilistic Latent Semantic Indexing. *SIG-IR*.

[7] M. Kameyama. 1986. A Property-Sharing Constraint in Centering. *ACL*.

[8] R. Mitkov. 1997. Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches. *ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*.

[9] S. Nariyama. 2002. Grammar for ellipsis resolution in Japanese. *9th TMI*.

[10] V. Ng and C. Cardie. 2002. Improving Machine Learning Approaches to Coreference Resolution. *ACL*.

[11] W. M. Soon, H. T. Ng, and D. C. Y. Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4).

[12] V. Vapnik. 1998. *Statistical Learning Theory*. John Wiley.

[13] M. Walker, M. Iida, and S. Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2).

[14] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林. 1997. 日本語語彙大系. 岩波書店.

[15] 工藤 拓, 松本 裕治. 2002. Support Vector Machine を用いた Chunk 同定. 自然言語処理, 9-5.

[16] 関 和広, 藤井 敦, 石川 徹也. 2002. 確率モデルを用いた日本語ゼロ代名詞の照応解析. 自然言語処理, 9-3.

[17] 田村 浩二, 奥村 学. 1995. センター理論による日本語談話の省略解析. 情報処理学会報告 (自然言語処理研究会),107-12.

[18] 中岩 浩巳, 池原 悟. 1996. 語用論的・意味論的制約を用いた日本語ゼロ代名詞の文内照応解析. 自然言語処理, 3-4.

[19] 橋田 浩一. 2002. GDA 日本語タギングマニュアル 草稿 第 0.68 版. http://i-content.org/

[20] 松本 裕治, 北内啓, 平野 善隆, 松田 寛, 高岡 一馬, 浅原 正幸. 2002. 形態素解析システム『茶筌』 *version 2.2.9* 使用説明書. 奈良先端科学技術大学院大学.

[21] 村田 真樹, 長尾 真. 1997. 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. 情報処理学会研究会報告 (自然言語処理研究会),4-1.

[22] 山田 寛康, 工藤 拓, 松本 裕治. 2002. *Support Vector Machine* を用いた日本語固有表現抽出. 情報処理学会論文誌,44-53.

[23] 横井 俊夫. 1995. *EDR* 電子化辞書仕様説明書. 日本電子化辞書研究所.