自 然 言 語 処 理 154－23
（2003．3．7）

機械学習によるゼロ代名詞同定の一方法

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 情報科学研究科
〒 630-0192 奈良県生駒市高山町 8916-5
{ryu-i,inui,hiroya-t,matsu}@is.aist-nara.ac.jp

センタリング理論のような言語学的な知見を，機械学習を用いた照応解析に統合する一方法を提案する．先行研究である Soon ら (2001) の解析手法に対して，我々は (i) 局所的な要素を考慮した素性 (センタリング素性) の追加と，(ii) 先行詞候補間を比較するモデル (トーナメントモデル) の 2 点を改良した．この提案手法を用いて日本語ゼロ代名詞の同定を行い，先行研究の手法より精度よく先行詞の同定ができたことを報告する．

One Method for Resolving Japanese Zero Pronouns with Machine Learning Model

Ryu Iida, Kentaro Inui, Hiroya Takamura and Yuji Matsumoto

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192, JAPAN
{ryu-i,inui,hiroya-t,matsu}@is.aist-nara.ac.jp

We propose to enhance a machine learning model for coreference resolution by incorporating linguistically motivated contextual clues, such as the centering theory. In comparison to Soon et al. (2001), our model shows improvements arising from two sources: (i) the feature of local contextual factors and (ii) an augmentation of the learning model to take into account comparison between candidates. This model is applied to resolve Japanese zero-anaphors and outperforms earlier machine learning approaches.

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 (代名詞，指示表現など)，あるいは省略する (ゼロ代名詞など) 場合が多い．このような省略を適切に補完することは，文脈解析において特に重要である．この補完の解析は省略格要素の照応解析と呼ばれる．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 (Mitkov, 1997; Baldwin, 1995; 中岩ら, 1996; 奥村ら, 1995; 村田ら, 1997)．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を用いる．規則ベースの手法の多くはセンタリング理論 (Grosz et al., 1995; Walker et al., 1994; Kameyama, 1986) のような言語学的な研究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アプリケーションでの使用を考えた場合，これらの数値は満足できる値であるとは言えない．さらに特定のドメインに特化した規則は，他のドメインで同じような精度を求めることが難しい．このような事実を考慮すると，人手での規則の洗練は難しく，また規則作成のコストは大きいと考えられる．

それに対して，照応タグ付きコーパスを用いた統計的な手法 (Aone and Bennett, 1995; Soon et al., 2001; Ng and Cardie, 2002; 関ら, 2002) は人手での規則作成に対してコストが低いという利点がある．それにもかかわらず，MUC-6 や MUC-7 の照応解析の評価セットを用いた実験では，規則ベースの手法と同程度の精度を得ている．しかし，これらの統計的な手法は，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．

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な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 (2001) 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 (2002) 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**) を追加し，また，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，日本語ゼロ代名詞を同定する実験を行い，先行研究と提案手法の比較を行う．最後に 5 節でゼロ代名詞の同定を行う際の問題と今後の方針について議論する．

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

2 先行研究

機械学習を用いた照応解析はさまざまな手法が提案されているが，Soon ら (2001) や Ng ら (2002) の先行研究では，機械学習の手法を用いて，規則ベースの手法と同程度の精度を得ている．そこで，この Soon らのモデルを基盤とし，その欠点を改善することで精度向上を目指す．

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，名詞句が先行詞となるかならないかという 2 値分類問題に分解する．この手法を図 1 を用いて説明する．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する．それぞれ NP_2 と NP_4，NP_3 と NP_5，NP_6 と NP_7 は照応関係にあり，*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．

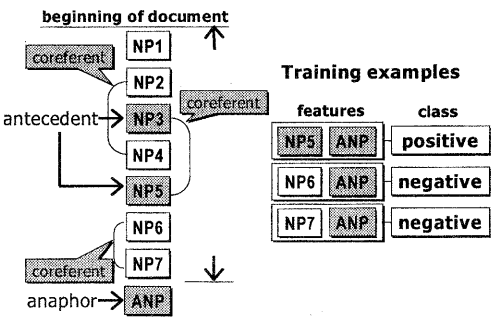

図 1: 訓練事例の作成 (Soon et al., 2001; Ng and Cardie, 2002)

訓練時には，Soon ら (2001) のモデルは，照応詞から最も近い先行詞と照応詞の対 ($ANP\text{-}NP_5$) を正例，先行詞と照応詞の間の名詞句それぞれと照応詞の対 ($ANP\text{-}NP_6$, $ANP\text{-}NP_7$) を負例として学習を

行う．同様に，先行詞を決定する際には，照応詞から先行文脈に向かって，対象となる名詞句を先行詞かどうか分類しながら処理を行う．そして，分類器が名詞句を先行詞として決定した時点で解析を終了する．この分類器が，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される．この実験には，12 個の限られた素性を用い，学習器は C4.5 (Quinlan, 1993) の拡張版である C5.0 を使用し，決定木学習を行っている．

Ng ら (2002) は Soon らの手法を 2 つの点において改良した．1 つは素性の拡張を行い，語彙的な素性や意味的な素性など学習に 53 個の素性を用いた．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先行詞かどうかの分類を行い，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判定された場合には，照応詞は先行詞を持たないと判断される．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代名詞同定のタスクにおいても同様の結果となった．そのため，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．

Soon らのモデルや Ng らのモデルは MUC のデータと評価方法に対しては，人手で記述した規則を用いた手法と同程度の精度を得ているが，情報検索や機械翻訳などへの照応解析の応用を考える上では，十分な精度であるとは言えない．この精度の低下を引き起こしている大きな原因は，この 2 つのモデルが，照応詞に対して 1 つの先行詞候補のみで学習を行っており，かつセンタリング理論のような文章の局所的な情報を無視しているためである．我々はこの問題について調べ，その結果考えられる 2 つの解決策を 3 節で示す．

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 らの問題について考察する．

(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) の主語“彼”は (a) の目的格である“ジョン”を指している．それに対して，(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 (i) 前文の主格であるので prefered center ((a) の forward-looking center の中で最も上位に位置する対象) となり，(ii)“トム”は (b) の backward-looking center の中で最も先行詞となりやすい．(iii) 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．それに対して (1) では，“メアリ”が prefered 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の間での優先度を考慮している点である．この例では“ジョン”が照応関係にあるかどうかは“メアリ”や“トム”のような局所的な文脈の中の他の要素に依存している．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，照応詞と先行詞候補と照応詞の関係だけを用いて先行詞かどうかの 2 値分類のみ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

上記で示した問題に対して，さまざまな解決策が考えられるが，本稿では学習を用いた 2 つの手法を提案する．

3.2.1 センタリング素性

この問題のより直観的な解決法は，学習で扱う素性に，局所的な文脈情報を扱う素性を追加することである．一例として，対象としている先行詞候補が prefered center であるかどうかの 2 値をとる素性を導入する．この素性を用いることで，局所的な文脈の中での先行詞らしさを学習に考慮できる．さらにこの素性を，foward-looking center の中で性や数の制約を満たす最も上位の要素と改良することができ，それによって，前述の例の 2 つの“ジョン”を区別することができる．このような素性を導入するためには forward-looking center のリストを計算す

る処理が必要になり，Soon らや Ng らの学習モデルでは，これを扱っていない．このセンタリング理論の局所的な焦点の遷移を捉える素性をセンタリング素性と呼ぶことにし，4 節で実装方法について述べる．

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”であるかの比較を行うことである．そのような比較を実現する方法は，2 つの先行詞候補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．

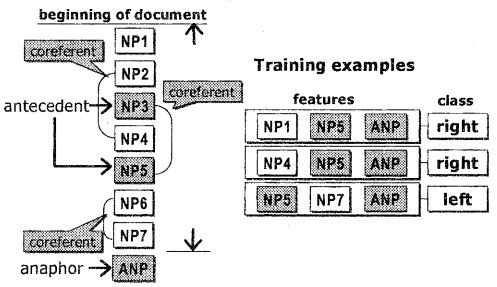

図 2: トーナメントモデル

図 1 の状況を調べ，同じ状況を図 2 に描き直す．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(と照応先の 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つい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 (先行詞となりやすい) という性質を学習できる．

4 評価実験

この節では，日本語ゼロ代名詞同定の実験を行うことで，先行研究と提案手法の比較を行う．

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．今回の実験では主格のゼロ代名詞の同定に問題を限定して，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．

4.2 素性

今回の実験で用いた 5 種類の素性 ((i) gramatical, (ii) semantic, (iii) positional, (iv) heuristic, (v) centering features) を表 1 に示す．(i) から (iv) までの素性は，以下の素性を除き Ng らが扱った素性に準拠している．

- SELECT_REST 素性，LOG_LIKE 素性: SELECT_REST 素性は，日本語語彙体系 (池原ら, 1997) を用いて選択制限を満たしているかどうかの 2 値をとる．LOG_LIKE 素性は，毎日新聞 (1991-1999) と日経新聞 (1990-2000) を CaboCha (工藤ら, 2002) を用いて構文解析したデータから，〈名詞-“が (助詞- 格助詞-一般)”-動詞〉の係り受けパタンで出現している共起デー

[2]GDA (Global Document Annotation (橋田, 2002)) 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．

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 の対に対して log-likelihood 係数を 5 段階に分け, その値を付与.
	ANIMACY	NP_i が PERSON もしくは 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.

タを抽出し,〈名詞-"が (助詞- 格助詞-一般)"-動詞〉のすべてのパタンに対して log-likelihood 係数を計算した値を 5 段階に分類した値を素性として用いた.

- CHAIN_LENGTH 素性: 今までの文脈で, よく照応された名詞句はよく照応されやすいという知見 (Ge and Charniak, 1998) に基づき, 対象とする名詞句が, 先行する文章の中でいくつの名詞句と照応関係にあるかを素性に用いた.

次に, センタリング素性を定義するために, Nariyama (2002) 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 (1986) 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. 実験では SRL_ORDER と SRL_ORDER_COMP の 2 つの素性を導入した. この素性の詳細は表 1 に示す. Nariyama のモデルでは Kameyama のモデルで扱っていない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に追加した (GA_REF 素性).

センタリング素性を用いることで, 局所的な文脈を考慮できる例を以下に示す. 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．

実験では，対象とする文章に対して茶筌 (松本ら, 2002) と CaboCha (工藤ら, 2002) を用い形態構文解析を行い，また yanee (山田ら, 2002) を用いて固有表現のタグを付与した．

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と評価されている Support Vector Machine (SVM) (Vapnik, 1998) を用いた．SVM はさまざまな自然言語処理のタスクにおいても高い精度を得ることができることが証明されている．

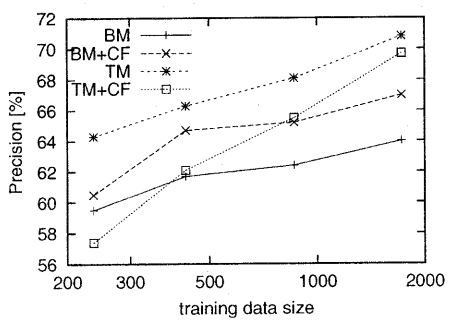

図 3: 学習曲線
BM: Ng らのモデル
BM+CF: センタリング素性を用いた Ng らのモデル
TM: トーナメントモデル
TM+CF: センタリング素性を用いたトーナメントモデル

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

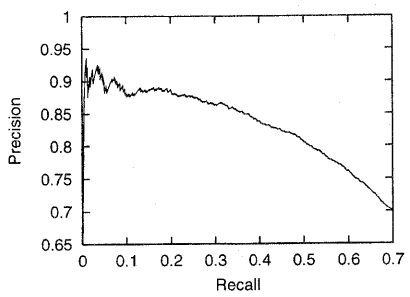

図 4: トーナメントモデルの Precision-recall 曲線

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 5%精度が向上していることがわかる．精度が良くなった実験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．トーナメントモデルの場合，1 回の候補間の比較に関する信頼度は，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，最後に勝ち残った候補が行った勝ち抜き戦のうち，最も小さな信頼度の値をトーナメント全体の信頼度をする．この信頼度に基づき，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，誤って照応先を確定するよりも少し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．

トーナメントの信頼度でソートした結果のうち，信頼度の高い事例でかつ解析に誤った事例の分析を行った．その結果，誤りを含む多くの事例は直接引用が関係していることがわかった．以下にその例を示す．

この例は，"**説明した**" の先行詞を同定する問題であり，省略されているガ格は**刑事法定**$_1$である．

ドイツのマンハイム地裁は十五日、ナチス・ドイツによるユダヤ人大虐殺は連合国や旧ソ連のでっち上げだとする「アウシュビッツのうそ」に理解を示す判決を下した**刑事法廷**$_1$の担当裁判長と判決文を執筆した裁判官の二人を解任した。**同法廷**$_1$は六月下旬、三年前に「アウシュビッツのユダヤ人収容所にガス室は存在しなかった」と公言して民衆扇動罪に問われた**ネオナチ指導者**$_2$に対し、執行猶予付き禁固一年の有罪判決を言い渡した。さらに今月九日に（**同法廷**$_1$ガ）発表した判決理由で「戦後半世紀たつ今もなお、ドイツはホロコーストを理由に、ユダヤ人の政治的、道徳的、金銭的要求にさらされており、**被告**$_2$はこれに対するドイツ民族の抵抗力を強化しようとした」と理解を表明。また「**被告**$_2$は意志強固にして知的な人物」と称賛し、執行猶予の理由を**説明した**。この判決理由に対してはユダヤ人社会を中心に内外から批判が続出、検察当局も、「判決自体が反ユダヤ主義をあおる民衆扇動罪に該当する」と、裁判官訴追の可能性をにおわせていた。

この状況で，直接引用を考慮せずにSRLを計算すると，省略されている**説明した**の最も近くにある**被告**$_2$はが最も優先度が高くなってしまう．そのため，誤ったセンタリング素性を学習に用いることになり精度の低下が予想される．

また，他の問題として視点の問題がある．以下の例では，“**聞いて**”の省略されているガ格は**堤義明オーナー**$_1$である．

西武の**森祇晶監督**は一日、東京・原宿のコクド本社に**堤義明オーナー**$_1$を訪ね、今季限りの退任を申し入れた。**堤オーナーも**$_1$これを了承。辞任が正式決定した。**森氏**$_2$は今後、野球解説者となる。後任監督の候補としては、石毛宏典内野手らが挙がっているが、石毛選手は態度を明確にしておらず、新監督決定まではなお曲折が予想される。会談は約1時間以上にわたった。会談後「9年間（**森氏**$_2$ガ）突っ走って、心身ともズタズタになり、日本シリーズ前の十月半ばに、休みを頂きたい、と（**森氏**$_2$ガ）申し入れた。『考え直せ』と慰留されたが、最終的にわがままを**聞いて頂いた**」と語った。堤オーナーは「長い間、よくやってもらった。次期監督も、今の森野球を踏襲していくのが一番いい。今後のことについても、いろいろアドバイスを受けた」と語り、今月末の球団の納会前に、森監督を特別表彰すると明らかにした。森監督は一九八五年オフに監督に就任し、以来9シーズンでリーグ優勝8度。優勝を逸したのは八九年だけで、この間、日本一も6度。通算成績は６７３勝４３８敗５９分だった。

この省略のSRLを計算すると，**森氏**$_2$が最も優先度が高くなるために，誤ったセンタリング素性を学習に用いてしまう．これは“**頂いた**”というモダリティを考慮して，“**聞いて**”のガ格が話者ではないという情報を考慮しなければならない問題である．

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代名詞の同定において効果的であることを示した．

今後の課題として，以下に3つの問題を述べる．

1. 主題と副主題の間の関係の同定
2. 複文や直接引用の文の分析
3. 選択制限の洗練

1に関しては，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．

新聞記事内の文の多くが複文であるために，GDAコーパスの中でタグ付けされた照応関係の半分以上が同一文内に先行詞を持つ．したがって，複文の正しい係り受け解析が必要となるが，我々が今回使用した依存構造解析器は高い頻度で複文の係り受けを誤ってしまう．そのため，照応解析で同定した格要素を，係り受けの問題を解く際の手がかりとするようなモデルも今後の課題としたい．また，4節で示したような直接引用の問題も，括弧の中にある文もしくは句が直接引用かを解析し，直接引用の場合は話者を同定した上で，それらを学習の枠組に組み込む必要がある．

最後に，ゼロ代名詞の同定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使ったが，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ことと同様にコーパスの質の向上も今後の課題としたい．

謝辞

GDAコーパスの利用を快く許可して下さった産業技術総合研究所の橋田浩一氏に深謝いたします．ま

た，本学学振特別研究員/メルボルン大学の成山重子博士には日本語ゼロ代名詞の解析に関して有用なご助言をいただきました．心より感謝申し上げます．

参考文献

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

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

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

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

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

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

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

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*.

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

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

J. R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

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).

V. Vapnik. 1998. *Statistical Learning Theory*. John Wiley.

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

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.

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.

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.

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.

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.

日経新聞社. 1990-2000. **日経新聞** *CD-ROM*.

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/

毎日新聞社. 1991-1999. **毎日新聞** *CD-ROM*.

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.

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.

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.

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.