LE-003

ゼロ照応解析のための統語的パタンの学習
Learning Sytactic Patterns for Zero-Anaphora Resolution

飯田龍 Ryu Iida　乾健太郎 Kentaro Inui　松本裕治 Yuji Matsumoto

1 はじめに

文章中の省略された格要素（**ゼロ代名詞**）を検出し，その先行詞を同定する**ゼロ照応解析**は情報抽出や機械翻訳など，多くの応用分野で必須の処理である．例えば，文章 (1) では，述語「増員する」のガ格が省略されており，ゼロ照応解析の処理ではこのゼロ代名詞 ϕ_i を検出し，先行詞候補集合から ϕ_i の適切な先行詞「団体 A_i」を同定する．

(1) 団体 A_i は約 100 人の作業者を雇っている。
（ϕ_i ガ）さらに 10 人を 増員する だろう。

近年では機械学習に基づく照応解析の手法 [7, 6, 10, 3] が発展し成果をあげているが，これらの手法では (i) 同一文内に先行詞が出現する場合と (ii) 文を越えて先行詞が出現する場合を区別せずに扱っている．しかし，(ゼロ) 照応解析の問題では，先行詞が照応詞（ゼロ代名詞）と同一文内に出現する場合は節間の関係など文の構造情報が重要なのに対し，先行詞が照応詞と異なる文に出現する場合には文章全体の談話構造における位置や，談話片の挿入を捉えるなど，考慮すべき特徴が異なる．名詞句照応解析の場合は，これらの情報を捨象しても，先行詞候補と照応詞の文字列の一致情報や意味の整合性などを利用して適切に解析できる場合が多いが，ゼロ照応の場合は照応詞に文字列の情報がないため，名詞句照応と比較して一般的に精度が低い．そこで，本稿では，特にゼロ代名詞と先行詞が同一文内に出現する現象（**文内ゼロ照応**）に着目し，この問題を精度良く解くことで文章全体のゼロ照応の精度向上を目指す．

我々が考慮すべき文の構造について具体例を示そう．南 [5] は隣接する節間に出現する接続助詞によって，二つの節の主語が一致するか否かを議論している．例えば，図 1(2) では接続助詞「て」で結ばれた二つの節の主語（「彼 $_i$」と「ϕ_i」）が一致しているのに対し，図 1(3) では接続助詞「ので」で節が連結されているが，こちらの場合では二つの主語（「先生 $_i$」と「ϕ_j」）は一致していない．ただし，手がかりとなる接続助詞で結ばれている場合に，必ずしも主語の（不）一致が保証されるわけではなく，他のゼロ照応の解析に有効な情報（選択制限など）などによって振舞いが変わる．人手で作成した規則に基づく手法 [9] でこれらの情報はすでに利用されているが，構造の情報とそれ以外の情報の組み合わせを網羅的に人手で記述することは容易ではない．

さらに，図 1(4) のような対象となるゼロ代名詞 ϕ_i が連体修飾節の中に出現しており，複雑な構造の場合にも図 2 のような一部語彙化した統語パタンが有効であると

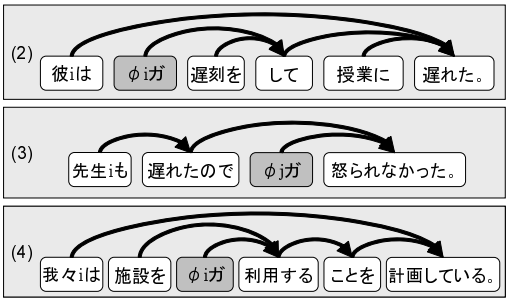

図 1: 文内ゼロ照応の例

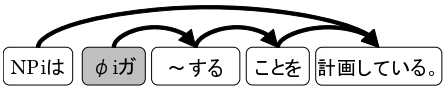

図 2: 文 (4) のゼロ照応解析に有効な統語パタン

考えているが，上述の節間の関係と同様にパタンを人手で網羅的に記述するのは困難である．そこで本稿では，ゼロ照応解析において重要な手がかりとなる文の構造情報を学習ベースの照応解析手法で利用する手法を提案し，評価実験を通じて構造情報の導入がどの程度解析の精度に貢献するかを示す．

以下，2 節で照応解析の先行研究を概観し，3 節で文の構造情報を効果的に既存の学習ベースの照応解析手法に導入する手法を提案する．次に，提案手法の有効性を調査するために行った評価実験の結果を 4 節で報告する．最後に 5 節でまとめる．

2 先行研究

照応解析の 2 つの部分タスクである先行詞同定と照応性判定の先行研究について説明する．

2.1 先行詞同定

先行詞同定の処理では，与えられた照応詞（ゼロ照応解析のタスクではゼロ代名詞）が与えられたときに，対象範囲内の先行詞候補集合から適切な先行詞を選択する．

従来の機械学習に基づく先行詞同定手法は，(i) 先行詞候補が真の先行詞か否かを分類する手法 [7, 6] と，(ii) 候補間の選好に基づく手法 [10, 2] の 2 つに分類できる．前者の手法では，与えられた照応詞 *TA* に対して各先行詞候補の絶対的な先行詞らしさを求め，最も先行詞らしい候補を *TA* の先行詞に決定する．ただし，どの候補も設定された閾値未満の場合は *TA* は先行詞を持たないと出力する．

これに対し，候補間の選好に基づく手法 [10, 2] では，先行詞同定のタスクを先行詞候補間の比較を行う問題に分解し，最も先行詞らしい候補を先行詞として決定する．例として，飯田ら [2] は，2 つの候補の間で勝ち抜き戦を行い最終的に勝ち残った最尤の候補を先行詞として出力するトーナメントモデルを提案している．(i) の手法が他

奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

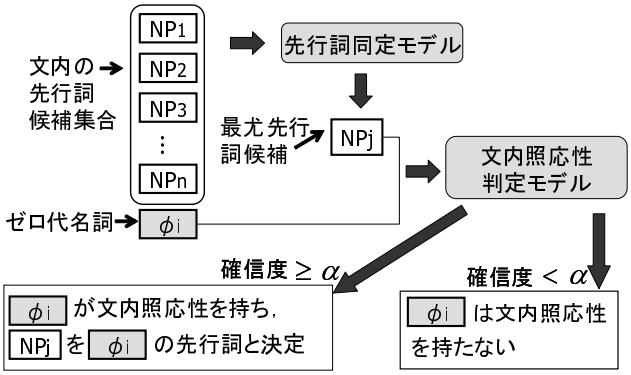

図 3: 探索先行分類型モデル

の候補に独立に先行詞らしさを出力するのに対し，(ii) の候補の選好に基づく手法では候補間の相対的な選好を学習できる．この相対的な選好を利用することによって解析精度が向上することが経験的に示されている [2].

2.2 照応性判定

照応性判定とは，文章中の任意の照応詞（ゼロ代名詞）の先行詞が文章内に存在するかを判定するタスクである．照応詞 Ana が文章内に先行詞を持つ場合，Ana は**照応性がある**と表現する．

照応性判定の手法はおおきく 2 つの手法に分類できる．Soon ら [7] や Ng ら [6] の手法では，先行詞同定処理と同時に照応性判定を行う．彼らのモデルではある照応詞候補に対して適切な先行詞が同定できれば，照応詞候補を照応詞に分類し，それ以外の場合には非照応詞に分類する．この手法では照応詞候補に対する先行詞候補の絶対的な先行詞らしさの値が必要となるため，候補間の選好に基づく手法を導入できないという欠点を持つ．

これに対し，我々が提案した探索先行分類型の解析手法 [3] では，照応性判定と先行詞同定の 2 つを分けて処理する（詳細は 3.1 で後述）．前者の手法が非照応詞のクラスを明示的に学習できないのに対して，後者の手法の照応性判定モデルでは非照応詞の訓練事例から直接的に非照応詞のクラスの振舞いを学習できるという利点を持つ．日本語名詞句照応解析タスクを対象とした評価実験において，探索先行分類型の手法を利用することにより，前者の手法よりも解析精度が向上するという成果を得ている．

3 提案手法

文の構造情報を学習ベースの手法に導入するためには，(i) 文の構造の表現方法と (ii) 表現された構造から有益な特徴をどのようにして抽出するかの 2 点を考える必要がある．この節では，まず提案手法で利用する探索先行分類型モデル [3] の概要を示し，次にこのモデルでどのように構造情報を利用するかを説明する．

3.1 探索先行分類型モデルの概要

探索先行分類型モデル（図 3）では，与えられたゼロ代名詞 ϕ_i に対し，まず探索範囲内（今回のタスクでは同一文内）に含まれる先行詞候補の集合 (NP_1,NP_2,...,NP_n) から尤も先行詞らしい候補（**最尤先行詞候補**）NP_j を選択し，次に，NP_j と ϕ_i の対が文内照応性を持つか否かを判別する．最尤先行詞候補の選択では，2 つの先行詞候補間で先行詞らしさの比較を行い勝ち抜き戦を行うことで最尤先行詞候補を決定するトーナメントモデル [2] を利用する．文内照応性判定の処理では，ゼロ代名詞と最尤先行詞候補の文内照応性を判定する分類モデルが必要となる．このモデルを訓練するための学習事例は以下のように作成する．

- 正例: 訓練コーパス中の文内に先行詞を持つゼロ代名詞 ϕ_p とその先行詞 NP_p の対 $\langle \phi_p, NP_p \rangle$.
- 負例: 文内に先行詞を持たないゼロ代名詞 ϕ_n と，ϕ_n について先行詞同定モデルが出力した最尤先行詞候補 NP_n の対 $\langle \phi_n, NP_n \rangle$.

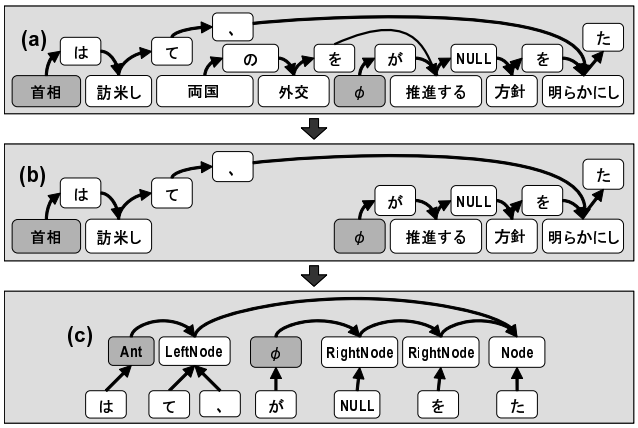

図 4: 文 (2) のゼロ代名詞と先行詞の間の部分構造

3.2 文の構造の表現

文の構造を表現する形式には節間の関係などさまざまな表現が考えられるが，今回の実験では試験的に文節を単位とした係り受け構造で文全体を表現した．具体的には，各文節は文節に含まれる機能語を子供として持ち，また文節間は係り受け関係で結ばれるような木構造で表現した．例えば，文 (2) を文節係り受け構造で表現すると図 4(a) のようになる．

(2) 首相は訪米して、両国の外交を（ϕ ガ）推進する方針を明らかにした。

今回は，この文全体の係り受け構造内のゼロ代名詞と先行詞候補の間のパス（図 4(b)）から統語パタンを学習することを考える．ただし，このパスの情報をそのまま学習に利用した場合，文字列情報（例えば,「首相」や「方針」など）をそのまま使用するため，訓練事例が疎になる可能性がある．そこで，最終的には，文節の内容語の情報を捨象し，機能語列を各文節ノードの子供ノードとすることで図 4(c) のような部分木に置き換え，この部分木から有効な統語パタンの学習を行う．

3.3 先行詞同定

トーナメントモデル [2] では，二つの先行詞候補間でどちらが先行詞らしいかという先行詞同定に有益な情報を利用できる．このモデルを訓練するために，(a) 与えられたゼロ代名詞，(b) 真の先行詞, (c) 偽の先行詞候補の三者の位置的な関係を次の 3 つのパスに分解して表現することを考える．

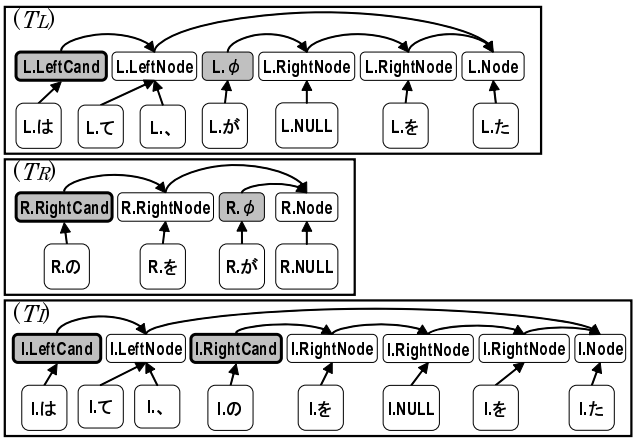

図 5: トーナメントモデルのための 3 つの部分構造

(1) 真の先行詞とゼロ代名詞の間のパス.
(2) 偽の先行詞候補とゼロ代名詞の間のパス.
(3) 真の先行詞と偽の先行詞候補の間のパス.

最終的には，この 3 つのパスを 3.2 で示した方法で木構造に変換して解析に利用する．例えば，図 4(a) の 2 つの先行詞候補「首相」と「両国」からは，図 5 のような部分木が抽出される．これらの部分木と 3.6 に示す 2 値素性の集合から先行詞同定に有効な統語パタンと素性の組み合せを学習し，解析に利用する．

3.4 文内照応性判定

文内照応性判定には，ゼロ代名詞と最尤先行詞候補の対の間のパスを用いた．ただし，学習事例の作成方法は 3.1 に従う．

3.5 分類器

構造情報を明示的に利用した分類手法には，Collins ら [1] の Tree Kernel や鈴木ら [8] の HDAG Kernel などのカーネル法を利用した学習手法，工藤ら [4] の部分木を素性とするブースティングを利用した分類手法などがある．今回の実験では，文献 [4] のアルゴリズムが実装された分類器 BACT[1]を使用した．先行詞同定と文内照応性判定の各処理では，図 5 に示した部分木の集合（T_L, T_R, T_I）とその他の素性集合（$f_1, f_2, \ldots, f_n$）を一つのおおきな木構造（$T_L, T_R, T_I, f_1, \ldots, f_n$）で表現し，その木から分類に有効な規則集合を学習する．解析の際には，同様に解析対象となるゼロ代名詞と先行詞候補の対を木構造で表現し，学習した規則集合を適用することで先行詞同定を行う．また，文内照応性判定の場合は，先行詞同定で求めた最尤候補とゼロ代名詞の間の部分木（T_{ANT}）とその他の素性集合（$f_1, f_2, \ldots, f_n$）で表現された木構造（$T_{ANT}, f_1, f_2, \ldots, f_n$）を利用する．

3.6 素性

解析には 3 節で示した文の構造情報に加え，ゼロ照応の解析に一般的に利用される以下の 3 種類の素性を利用した．素性の詳細については文献 [2] を参照されたい．

- 対象となるゼロ代名詞を持つ述語の語彙，統語情報に関する素性.
- 先行詞候補に関する語彙，統語，意味（名詞の意味属性），位置情報に関する素性.
- ゼロ代名詞を持つ述語と先行詞候補の対から抽出可能な情報（例えば，選択選好や述語と先行詞候補の距離など）に関する素性.

ただし，構造情報を含むすべての素性の抽出は，茶筌[2]と南瓜[3]で形態・構文解析して得られた結果を利用した．

4 評価実験

日本語新聞記事コーパスを対象に Ng らの解析モデル [6]（BM）と探索先行分類モデル [3]（SCM）の 2 つのモデルを利用し，以下の 3 種類の条件で比較実験を行い，提案手法の有効性を調査した．

- BM や SCM を構造情報を用いずにそのまま利用する（ORG）.
- 構造情報として図 4(b) に示したゼロ代名詞と先行詞候補の間のパスの情報を直接素性に加える（PATH）.
- 構造情報として図 4(c) に示した部分構造を利用する（TREE）.

4.1 評価事例

照応関係タグ付きコーパス[4]の一部 347 文章を，訓練用に 137 文章，パラメタ推定用に 60 文章，評価用に 150 文章に分割して実験を行った．今回は評価対象を文内に先行詞を持つガ格のゼロ代名詞に限定し，それぞれのモデルでどの程度適切に解析できるかを見る．訓練事例には 1,229 事例のゼロ代名詞が含まれており，パラメタ推定用には 846 事例，評価用には 1,104 事例を利用した．評価用事例のうち 524 事例（全体の 47.5%）のゼロ代名詞が文内に先行詞を持つ．この状況で，ゼロ代名詞が文内に先行詞を持つ場合は先行詞を同定し，それ以外（文外に先行詞を持つ，もしくは文章内に先行詞を持たない）の場合は棄却するという問題を解く．今回の実験では，純粋に対象とするゼロ代名詞に関する精度を求めるために，ゼロ代名詞の出現箇所は人間が与え，さらに対象とする箇所以外は正しい格関係，連体修飾関係を与えて評価を行った．

4.2 実験結果

先行詞同定と文内の照応性判定のそれぞれの処理で，文の構造を利用することによりどの程度解析精度が向上するかを調査した．まず，先行詞同定の解析結果を表 1 に示す．表 1 の BM_ORG と BM_TREE，SCM_ORG と SCM_TREE をそれぞれ比較すると，構造情報を加えることで解析精度が向上していることがわかる．

次に，文内の照応性判定の閾値（BACT の出力した判別関数の値）を動かして再現率-精度曲線を描いた（図 6）．精度，再現率は以下の式に従う．この結果より，閾値が適切に推定できた場合の上限値を見積もることができる．

$$\text{再現率} = \frac{\text{ゼロ代名詞の先行詞を適切に同定できた数}}{\text{文内に先行詞を持つゼロ代名詞の総数}},$$

[1] http://chasen.org/~taku/software/bact/

[2] http://chasen.naist.jp/hiki/ChaSen/

[3] http://chasen.org/~taku/software/cabocha/

[4] 詳細は http://cl.naist.jp/~ryu-i/coreference_tag.html を参照.

表 1: 文内ゼロ照応の先行詞同定の結果

	BM	SCM
ORG	0.523 (274/524)	0.712 (373/524)
PATH	0.536 (281/524)	0.693 (363/524)
TREE	0.656 (344/524)	**0.740** (388/524)

BM: Ng らの解析モデル, SCM: 探索先行分類型モデル, ORG: 各モデルで構造情報を利用しない，PATH: 図 4(b) のパスの情報を利用する，TREE: 図 4(c) の部分構造を利用する．

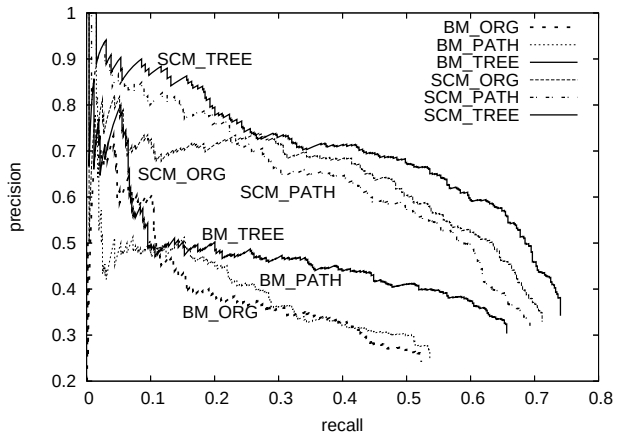

図 6: 文内ゼロ照応解析の再現率-精度曲線

$$\text{精度} = \frac{\text{ゼロ代名詞の先行詞を適切に同定できた数}}{\text{文内に先行詞を持つと判定したゼロ代名詞の総数}},$$

$$\text{F 値} = \frac{2*\text{再現率}*\text{精度}}{\text{再現率}+\text{精度}}.$$

図 6 より，Ng らのモデルと探索先行分類型モデルのどちらのモデルでも，構造情報を利用することで文内の照応性判定の精度が向上しており，すでに精度が良かった探索先行分類型モデルでも構造情報が有効に働いていることがわかる．また，PATH（パスの情報をそのまま学習したモデル）と TREE（内容語の情報を捨象して学習したモデル）の結果を比較することで，パスの情報を加工することがどの程度精度に貢献しているかを見ることができる．図 6 に示した結果より，BM と SCM のどちらのモデルでも TREE を利用した場合に精度が向上している．パスの情報からどのように情報を抜き出すかにはさまざまな工夫の余地があり，工夫次第ではさらに精度が向上する見込みがある．

最後に，パラメタ推定用の事例を利用して照応性判定の閾値を推定し，再現率，精度，F 値を求めた結果を表 2 に示す．この結果より，閾値を推定し実際に問題を解く状況を仮定した場合でも，提案手法（SCM_TREE）が最も高い精度で解析できていることがわかる．

紙幅の制約で詳細は述べないが，文章全体のゼロ照応解析の問題に関しては，文内で先行詞が見つからない場合には，構造情報を利用しない探索先行分類型モデルで文間の先行詞を探索を行った．提案手法を用いることで，文内ゼロ照応の精度が向上しているため，最終的に文章全体のゼロ照応解析の精度は良くなる．

4.3 誤り分析

先行詞同定と文内照応性判定の各処理で解析を誤った事例を分析した結果，最も解析誤りに関係する現象は直接引用であることがわかった．例えば，次の文で，直接引用の中に出現しているゼロ代名詞 ϕ_i の先行詞は引用の外にある「古前田監督」である．しかし，解析モデルは，述語の前方文脈に出現しており，かつ格助詞「は」を持つなど，いくつかの先行詞となる手がかりを持つ候補「選手」を先行詞として出力した．

表 2: 推定した閾値を利用して得られた実験結果

	再現率	精度	F 値
BM_ORG	0.426 (223/524)	0.308 (223/724)	0.357
BM_PATH	0.439 (230/524)	0.311 (230/740)	0.364
BM_TREE	0.573 (300/524)	0.382 (300/786)	0.458
SCM_ORG	0.536 (280/524)	0.580 (280/483)	0.557
SCM_PATH	0.600 (314/524)	0.494 (314/636)	0.542
SCM_TREE	0.649 (339/524)	0.577 (339/588)	**0.610**

「**選手**はそのときの経験を生かしてくれた。(ϕ_i **ガ**) 言わなくても分かっていた」と**古前田監督** $_i$。

直接引用のような談話の埋め込みの構造は，今回捉えようとした文の中の構造より文の間のゼロ照応関係を捉える問題に近く，直接引用内のゼロ代名詞には特別な処理を施すといった工夫が必要になると考えられる．これについては今後の課題としたい．

5 おわりに

本稿では，探索先行分類型モデルの先行詞同定，文内照応性判定の各処理に文の構造情報を導入する手法を提案した．従来手法と比較を行い，先行詞同定と文内照応性判定のそれぞれの処理で構造情報が精度向上に貢献することを示した．また，誤り分析の結果より，引用の外に出現している文内ゼロ照応の問題は質良く解析できているが，引用の中，つまり談話が埋め込まれた状況での解析精度が低いことがわかった．そこで，今後はこの引用の問題を文間のゼロ照応の問題の足掛かりとし，ゼロ照応解析に必要となる談話構造について考えたい．

参考文献

[1] Collins, M. and Duffy, N.: Convolution Kernels for Natural Language, *Proceedings of the Neural Information Processing Systems (NIPS)*, pp. 625–632 (2001).

[2] 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, pp. 906–918 (2004).

[3] 飯田龍, 乾健太郎, 松本裕治: 照応性判定を含む名詞句照応解析の実験と分析, 情報処理学会研究会報告 (自然言語処理研究会) NL-169-15, pp. 93–100 (2005).

[4] 工藤拓, 松本裕治: 半構造化テキストの分類のためのブースティングアルゴリズム, 情報処理学会論文誌, Vol. 45, No. 9, pp. 2146–2156 (2004).

[5] 南不二男: 現代日本語の構造, 大修館 (1974).

[6] Ng, V. and Cardie, C.: Improving Machine Learning Approaches to Coreference Resolution, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 104–111 (2002).

[7] Soon, W. M., Ng, H. T. and Lim, D. C. Y.: A Machine Learning Approach to Coreference Resolution of Noun Phrases, *Computational Linguistics*, Vol. 27, No. 4, pp. 521–544 (2001).

[8] 鈴木潤, 佐々木裕, 前田英作: 階層非循環有向グラフカーネル, 電子情報通信学会論文誌, Vol. 88, No. 2, pp. 230–240 (2005).

[9] 田村浩二, 奥村学: センター理論による日本語談話の省略解析, 情報処理学会研究会報告 (自然言語処理研究会) NL-107-16, pp. 91–96 (1995).

[10] Yang, X., Zhou, G., Su, J. and Tan, C. L.: Coreference Resolution Using Competition Learning Approach, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 176–183 (2003).