文の構造を利用した文内ゼロ照応解析*

飯田 龍　乾 健太郎　松本 裕治
奈良先端科学技術大学院大学
{ryu-i,inui,matsu}@is.naist.jp

1 はじめに

文章中のゼロ代名詞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．近年では機械学習に基づく解析手法 [12, 10, 15, 5] 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) 文を越えて出現する先行詞を区別せずに扱っている．しかし，ゼロ照応の問題では，先行詞が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解析に有効であるのに対し，異なる文に出現する場合は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，捉えるべき特徴がおおきく異なる．このため，この二つを分けて処理すれば，解析の精度が向上する可能性がある．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（以下，**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解析の精度向上を目指す．

文内ゼロ照応解析の処理には，**文内照応性判定**と**先行詞同定**の問題が含まれており，この二つの部分問題を意識ながら全体の処理を考えることが重要である．文内照応性判定とは，与えられたゼロ代名詞が文内に先行詞を持つ（**文内照応性**がある）か否かを判定するタスクであり，また，先行詞同定は，文内照応性判定で先行詞を文内に持つと判定したゼロ代名詞を対象に文内から先行詞を同定する処理である．

これまでのゼロ照応解析の先行研究から，文内ゼロ照応の解析には，南 [8] の主張する節間の主語同一性に代表される文の構造情報が有用であると考えられるが，文の構造情報とその他のゼロ照応解析に有効な情報（例えば，選択選好など）の組み合わせを人手で規則として書き尽くすことは困難である．そこで，本稿では，これら二つの情報を併用する機械学習に基づくゼロ照応解析モデルを提案し，その有効性について議論する．

2 節で文の構造を利用したゼロ照応解析の先行研究とその問題点を示し，3 節で文の構造を効果的に既存の学習ベースの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．

*Intra-sentential zero-anaphora resolution using dependency structure information
Ryu Iida, Kentaro Inui, and Yuji Matsumoto
Nara Institute of Science and Technology

表 1: 節間の主語同一性（南 [8]）

	隣接する節間の主語の関係	接続助詞の例
A 類	主語が一致しやすい	たり，ながら，つつ，て
B 類	主語が一致しにくい	ても，ので，けれど，ば，と，から，のに
C 類	文脈に依存	が

2 先行研究

文内のゼロ照応解析では，並列構造や連体修飾の関係など，文の構造情報が解析の手がかりとなる場合が多い．こうした構造情報の例に，南 [8] の節間の主語同一性に関する分析（表 1）がある．田村ら [14] や中岩ら [9] はこの関係を利用してゼロ照応の解析精度が向上したと報告している．例えば，田村ら [14] の手法では，複文を節単位で分割し，文章を単文の列で表現することで, 表 1 の節間の関係に加え，局所的な主題の遷移の特徴を説明するセンタリング理論 [3] における先行詞らしさの規則を記述することでゼロ代名詞の先行詞同定を行っている．また，白井ら [11] も同様に南の分類に着目し，節間の係り関係の解析に表 1 を細分化し適用することで解析精度の向上に寄与することを示している．このように，節間の関係に代表される文の構造情報は文内のゼロ照応解析に有効に役立つと考えられる．ただし，選択選好などのゼロ照応解析に有効なさまざまな情報との組み合わせを人手で考えるのは現実的でない．

これに対して，近年，機械学習に基づく照応解析手法 [12, 10, 15, 5] が注目を集めており，特に名詞句の照応解析において成果を得ている．日本語名詞句照応解析では，我々が提案した探索先行分類型モデル [5] を利用することで既存手法よりも先行詞同定，照応性判定ともに解析精度が向上するという結果を得た．ただし，この手法でも同一文内とそれ以外の区別をせずに処理しているため，このモデルに上述の文の構造情報を導入することで，文内のゼロ照応解析の精度が向上することが期待できる．

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 [5] 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．

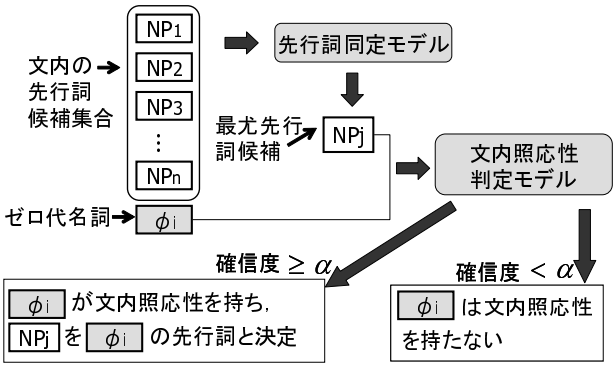


図 1: 探索先行分類型モデル

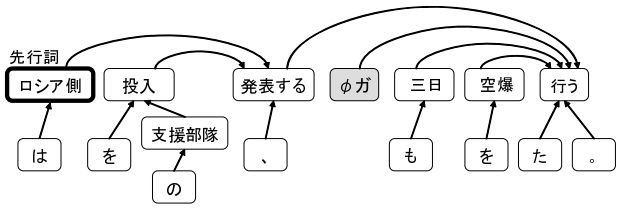


図 2: 文 (1) の文節構造

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 1）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 [4] 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 $\langle \phi_n, NP_n \rangle$.

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 (1) を文節係り受け構造で表現すると図 2 のようになる．

(1) ロシア側 $_i$ は支援部隊の投入を発表し、三日も首都空爆を（φ $_i$ ガ）行った。

3.3 先行詞同定

先行詞同定で利用するトーナメントモデルでは先行詞候補間でどちらが先行詞らしいかを判別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には，例えば，文 (1) の真の先行詞「ロシア側」は右側にある偽

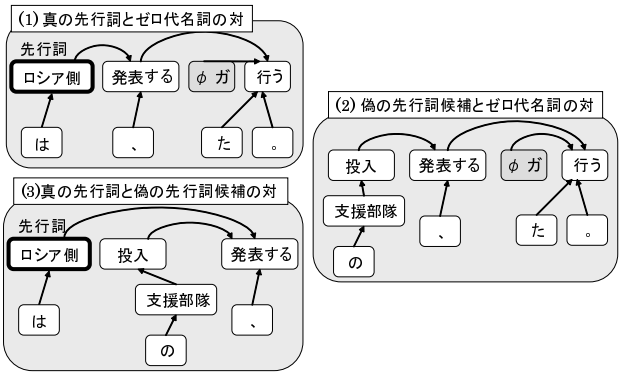


図 3: トーナメントモデルのための 3 つの部分構造

の先行詞候補「支援部隊」より先行詞らしいという傾向を学習しなければならない．この傾向を効果的に学習するために，図 2 から 2 つの候補と対象とするゼロ代名詞の 3 つ組の情報を抽出することを考える．今回はこの 3 要素の間の関係を以下の 3 つの部分木に分解して表現した．

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

例えば，図 2 の木からは図 3 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規則を抽出し，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構造（図 3，(1) に相当）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[1] の Tree Kernel や鈴木ら [13] の HDAG Kernel などをカーネル法を利用した学習手法，工藤ら [7] 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 [7] 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，学習時に抽出された部分木の集合 (T_1,T_2,...,T_m) とその他の素性集合 (f_1,f_2,...,f_n) を一つのおおきな木構造 (T_1,...,T_m,f_1,...,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（もしくは文内照応性判定）を行う．

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[2].

- 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.

[1] http://chasen.org/~taku/software/bact/
[2] 素性の詳細は文献 [4] を参照.

- 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.

選択選好の情報には日本語語彙大系 [6] の構文体系の情報に加え，藤田ら [2] の提案する動詞格構造の適格性モデルを利用した．このモデルでは，対〈名詞，格助詞-述語〉に関して大規模なコーパスから抽出した用例の頻度から計算した相互情報量を出力するが，今回はこの数値を離散化して素性に加えた.

構造情報を含むすべての素性の抽出は，茶筌[3]とCaboCha[4]で形態・構文解析して得られた結果を利用した．また，内容語に特化した学習を防ぐため，文の構造情報については図 3 に示した部分木のうち，内容語を表現しているノードの表層情報を捨象して利用した．内容語を加えた実験は現在作成中の大規模な照応関係タグ付きコーパスを利用して再度実験を行う予定である.

4 評価実験

提案手法の有効性を評価するために，日本語新聞記事コーパスを対象に以下の 4 つのモデルを比較評価した.

1. BM: Soon ら [12] の提案する探索型モデル．このモデルでは，文内照応性判定と先行詞同定の問題を同時に解く.
2. BM_STR: 図 3(1) に相当する先行詞（候補）とゼロ代名詞の対の構造を加えた BM.
3. SCM: 構造情報を利用しない探索先行分類型モデル.
4. SCM_STR: 構造情報を利用した SCM（提案手法）.

4.1 評価事例

照応関係タグ付きコーパス[5]の一部 60 文章から抽出したガ格ゼロ代名詞 896 事例を対象に 5 分割交差検定を行った．896 事例のうち 395 事例（全体の 44.1%）のゼロ代名詞が文内に先行詞を持つ．この状況で，文内に先行詞を持つゼロ代名詞の場合は先行詞を同定し，それ以外の場合は棄却するという問題を解く．なおここでのゼロ代名詞とは，述語の項（名詞句）が連体修飾関係を含む係り受け関係にない場合に，その項を指す省略された要素をいう．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 2 に示す．表 2 の BM と BM_STR，SCM と SCM_STR 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 4）．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.

表 2: 文内ゼロ照応の先行詞同定の結果

	BM	BM_STR	SCM	SCM_STR
精度	61.3% (242/395)	66.1% (261/395)	69.6% (275/395)	75.2% (297/395)

BM: ベースラインモデル, BM_STR: BM + 構造情報, SCM: 探索先行分類型モデル, SCM_STR: SCM + 構造情報.

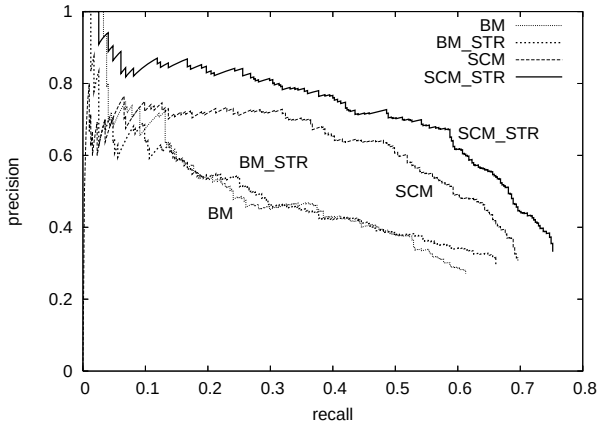


図 4: 文内ゼロ照応解析の再現率-精度曲線

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

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

図 4 より，文内照応性の判定においても構造情報が有効にはたらいていることがわかる．特に，探索先行分類型モデルのように文内照応性判定を個別に処理する場合により効果的に精度に貢献している．これは，BM(_STR) が文内に先行詞を持つゼロ代名詞のみから負例を作成しているのに対し，SCM_STR では文内照応性判定の訓練事例の負例の一部（181 事例）に外界照応となるゼロ代名詞を利用し，外界を指す場合の典型的なパタンを学習することで，解析時にそれらを適切に棄却したためだと考えられる.

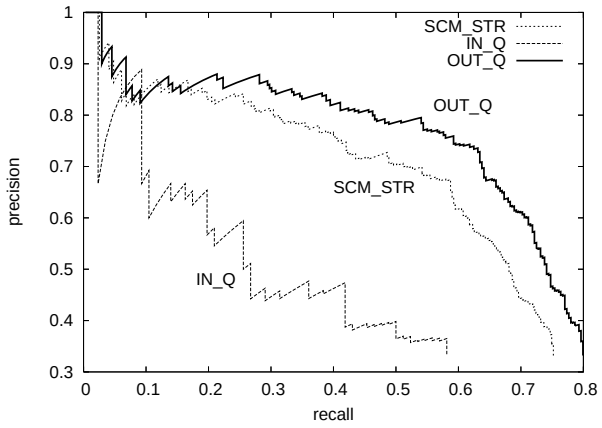
4.3 誤り分析

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.

「**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** $_i$。

引用の問題が文内ゼロ照応解析にどの程度影響を及ぼしているかを調査するために，図 4 の結果のうちゼロ代名詞が直接引用の外に出現している場合と中に出現している場合で分けて文内照応性判定を評価した（図 5）．図 5 より，引用の外にゼロ代名詞が出現している場合

[3] http://chasen.naist.jp/hiki/ChaSen/
[4] http://chasen.org/~taku/software/cabocha/
[5] 詳細は http://cl.naist.jp/~ryu-i/coreference_tag.html を参照.



IN_Q: 引用の中の事例，OUT_Q: 引用の外の事例.

図 5: ゼロ代名詞の出現位置によって分けて評価

は，高い精度で解析できているのに対し，引用の中にあるゼロ代名詞は極端に解析精度が低くなる．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関係を捉える問題に近く，この問題を考慮して解析するには別の枠組みを考える必要がある．これについては今後の課題としたい．

4.4 文章全体を通した評価

次に，今回の提案手法が文章全体のゼロ照応解析の向上にどの程度貢献するのかを調査した．ただし，解析は以下の手順で行った．

1. 文内ゼロ照応解析モデルを用い文内の候補集合から先行詞 NP_i を同定する．文内照応性判定の確信度（BACT が出力した判別関数の値）が閾値 α_i 以上の場合，NP_i を出力する．
2. 確信度が α_i 未満の場合，探索先行分類型モデルを用い，1. の解析範囲を除いた前方文脈から先行詞 NP_j を同定する．照応性判定の確信度が閾値 α_j 以上の場合は NP_j を出力する．
3. 確信度が α_j 未満の場合は照応性無しと出力する．

α_i，α_j を動かすことによって，図 6 に示す文章全体のゼロ照応解析の再現率-精度曲線を得た（SCM_STR）．比較対象として，文内，文間を区別せずに探索先行分類型モデルで解析した結果の再現率-精度曲線も示す（図 6, SCM）．図 6 より，構造情報を利用し文内の解析を行うことで全体の精度が向上していることがわかる．ただし，実際に利用する場合には提案手法では閾値を 2 つ推定する必要があり，その最適化がどの程度現実的なのかについては今後調査する必要がある．また，解析の順序を (1) 文内の最尤先行詞 C_{intra} を決定，(2) 文間の最尤先行詞 C_{inter} を決定，(3)C_{intra} と C_{inter} の間の比較を行い，最終的な出力を決めることでパラメタを一つ減らすことも考えられる．このように文間，文内の結果をどのように扱うのかについては工夫の余地があり，これについても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，評価実験の結果より，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．

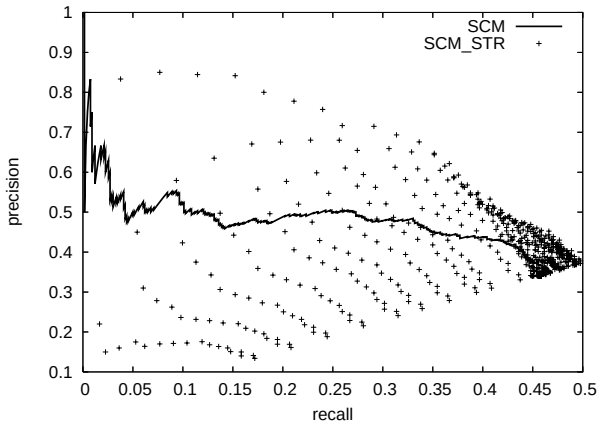


図 6: ゼロ照応解析の再現率-精度曲線（文章全体）

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] 藤田篤, 乾健太郎, 松本裕治: 自動生成された言い換え文における不適格な動詞格構造の検出, 情報処理学会論文誌, Vol. 45, No. 4, pp. 1176–1187 (2004).

[3] Grosz, B. J., Joshi, A. K. and Weinstein, S.: Centering: A framework for modeling the local coherence of discource, *Computational Linguistics*, Vol. 21, No. 2, pp. 203–226 (1995).

[4] 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).

[5] 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).

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

[7] 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).

[8] 南不二男: 現代日本語の構造, 大修館 (1974).

[9] 中岩浩巳, 池原悟: 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, Vol. 3, No. 4, pp. 49–65 (1996).

[10] Ng, V.: Learning Noun Phrase Anaphoricity to Improve Coreference Resolution: Issues in Representation and Optimization, *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 152–159 (2004).

[11] 白井諭, 池原悟, 横尾昭男, 木村淳子: 階層的認識構造に着目した日本語従属節間の係り受け解析の方法とその精度, 情報処理学会論文誌, Vol. 36, No. 10, pp. 2353–2361 (1995).

[12] 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).

[13] 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).

[14] 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).

[15] Yang, X., Su, J. and Tan, C. L.: A Twin-Candidate Model of Coreference Resolution with Non-Anaphor Identification Capability, *Proceedings of the 2nd International Joint Conference on Natural Language Processing* (*IJCNLP05*), pp. 719–730 (2005).