先行文脈と局所文脈を併用した照応性判定モデルの学習*

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

文章中の照応関係を同定する照応解析は機械翻訳や自然言語での質問応答など自然言語処理の応用分野で必須の処理である．照応解析の処理は，おおきく照応性判定（anaphoricity determination）と先行詞同定の二つの処理に分解できる．照応性判定は，文章中の名詞句が先行文脈に対となる先行詞を持つ**照応詞**か，もしくはそれ以外（**非照応詞**）であるかを分類するタスクである．また，先行詞同定は，照応性判定で検出した照応詞に対して先行詞を同定する処理である．

初期の照応解析に関する研究 [3, 4, 1] では統語的な特徴から照応詞と判断できる代名詞や指示詞についてのみ研究の対象としている．つまり，文章中のどの要素が照応詞となるかはあらかじめ与えられた上で先行詞同定処理の精度の向上を目指すことが照応解析の主要な目的のように考えられてきた．しかし，以下の二つの理由により，照応性判定処理も視野に入れた照応解析処理にも研究者の関心が集まってきている [2, 6, 9, 7].

- 照応性判定に定冠詞を手がかりとして利用できる英語などの言語においても，照応性を判定することはそれほど簡単な問題ではない．
- 照応解析全体の精度は照応性判定の結果に依存する．

もちろん，日本語のように定冠詞を手がかりとして利用できない言語の場合は，照応性判定の問題はさらに重要である．

これまでの照応性判定に関する研究から以下の二つの点が明らかになってきた．

- 照応性判定の重要な手がかりの一つは先行詞候補を探索することによって得ることができる．なぜなら，適当な先行詞候補が先行文脈に存在することが照応詞候補が照応詞と判断される必須条件となるためである．
- 照応性を効果的に判定するためには，照応詞の振舞いだけでなく非照応詞の振舞いについても学習する必要がある．

しかし，2 節で述べるように，これらの二つの情報をともに利用する解析モデルはこれまでのところ報告されていない．そこで，本稿では，これら二つの利点を併用する照応性判定モデルを提案する．さらに照応性判定の問題は名詞句照応だけでなく，ゼロ照応に関しても問題となる．そこで，ゼロ代名詞が前方照応であるかそれ以外（後方照応，外界照応）であるかを分類する問題にも同じ方式を適用し，名詞句照応とゼロ照応での照応性判定の結果を比較する．2 節では先行研究の手法とその問題点を示し，3 節で上述の二つの情報を組み合せた照応性判定モデルを提案する．次に，提案手法の有効性を調査するために 4 節で日本語名詞句とゼロ代名詞の照応性判定の評価実験を行い，結果について考察する．最後に 5 節でまとめる．

*Learning An Aanaphoricity Determination Model Combining Preceding and Local Contextual Information
Ryu Iida, Kentaro Inui, and Yuji Matsumoto
Nara Institute of Science and Technology

2 先行研究

従来の照応詞性判定手法はおおきく**探索型手法**と**分類型手法**に分類できる．探索型手法では，文章中の任意の名詞句（照応詞候補）に対して先行文脈から先行詞となる候補が存在するか否かを探索することにより間接的に照応性判定を行う．つまり，もし先行詞となる適切な候補がみつかれば，照応詞候補は照応詞と判定され，それ以外の場合には非照応詞と判定される．探索型の代表的な手法である Soon ら [8] の手法では，照応詞候補に対して先行文脈中の各名詞句をそれぞれ先行詞候補とみなし，照応詞候補に近い先行詞候補から順に先行詞候補と照応詞候補の対が照応関係にあるか否かの分類問題を解く．ひとつでも照応関係にあると分類された場合は照応詞候補は照応詞であると判定され，逆にどの先行詞候補とも照応関係にないと分類された場合には照応詞候補は非照応詞と判定される．この探索型手法では，先行文脈に適切な先行詞候補を持つか否かを調査するため，照応詞候補の照応性を判定する際に先行文脈情報を利用できるという利点を持つ．

探索型手法ではこのような利点があるが，非照応詞の振舞いを学習するように設計されていないため，非照応詞を適切に棄却できるとは限らない．例えば，図 1 に示した Soon のモデルでは，訓練事例を作成する際に照応詞 *ANP* に対して最も近い先行詞との対 NP_5-*ANP* を正例に，また照応詞と先行詞の間の名詞句それぞれと照応詞の対（NP_6-*ANP*,NP_7-*ANP*）を負例とする．そのため，非照応詞と任意の先行詞候補との対は訓練時に考慮されない．この事例作成方法は Ng ら [5] や Yang ら [11] の手法でも採用されているため同様に問題となる．

これに対し，分類型手法 [6, 7] では，照応性判定の問題を先行詞同定の処理と切り離して考える．この手法は，探索型手法で利用できていない非照応詞の情報を利用して照応性判定の分類器を作成するという利点がある．Ng ら [7] の評価実験によると，最初に照応性判定を行うことで照応詞とする名詞句を明示的に削減

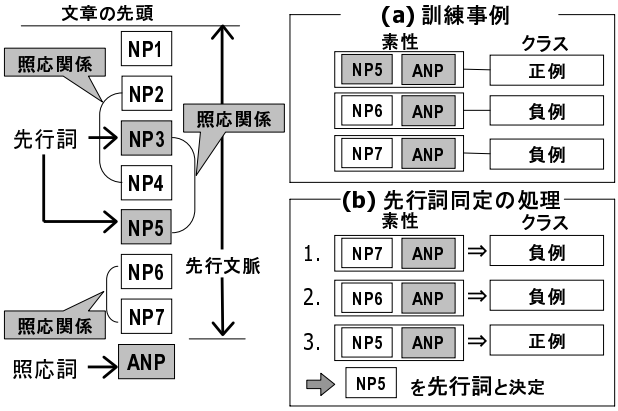

図 1: Soon らの探索型モデル

し，その後に照応詞と判定された名詞句のみを対象に先行詞同定を行うことで照応解析全体の精度が向上したと報告されている．

しかしながら，分類型手法には逆に，先行詞候補の情報が利用できないという欠点がある．Ng らは，英語の名詞句を対象とした照応性判定の実験結果より，語彙，統語や名詞意味属性などの局所文脈の情報のみで照応性を判定できると述べているが，それが他言語において成り立つとは限らない．例えば，日本語名詞句の照応性判定の場合，4 節で後述する我々が行った実験によると，分類型手法では平均精度が 49.2%にすぎなかったのに対し，先行文脈と局所文脈情報を組み合わせた我々の手法では 81.1%に精度が向上した．

上述の手法はいずれも名詞句の照応性を判定するものだが，日本語ゼロ代名詞についても同様に照応性の判定が問題となる．山本ら [13] は日本語対話文章を対象に，決定木を用いて文章内のゼロ代名詞が照応性を持つか否かを分類している．学習に利用した情報の多くは局所文脈から抽出されるものであるため，この手法は分類型手法に分類できる．ただし，ゼロ代名詞についても先行文脈に先行詞となり得る候補が存在することが照応性判定の有効な手がかりとなることは名詞句の場合と共通であると考えられる．そこで，後述する 4 節では照応性判定の実験を通して名詞句とゼロ代名詞の照応性判定それぞれを評価し，二つの間の共通性を経験的に示す．

3 提案手法

提案手法では，先行文脈の情報を分類型手法に導入する．この手法では各照応詞候補に対して次の 2 段階の処理で照応性を判定する．

1. 照応詞候補 NP_i に対して先行詞同定モデルを用いて先行詞候補集合から最も先行詞らしい候補（**最尤先行詞候補**）AC を同定する．
2. AC-NP_i の対が照応関係にあるか否かを分類する．もし AC-NP_i が照応関係にあると分類された場合は NP_i は照応詞と決定される．そうでなければ，NP_i は非照応詞と判断される．

先行詞同定には，各照応詞候補に対して最尤先行詞候補を探索できるモデルであれば任意の解析モデルを利用することができる．例えば，Ng らのモデル [5] はすべての先行詞と照応関係に無いと判定された場合でも先行詞らしさの値が最もおおきい候補を出力するので，このモデルを使ってもよい．

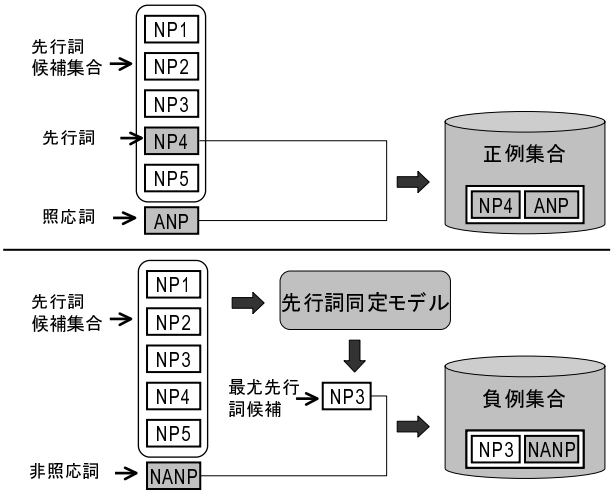

図 2: 照応性判定モデルのための訓練事例作成

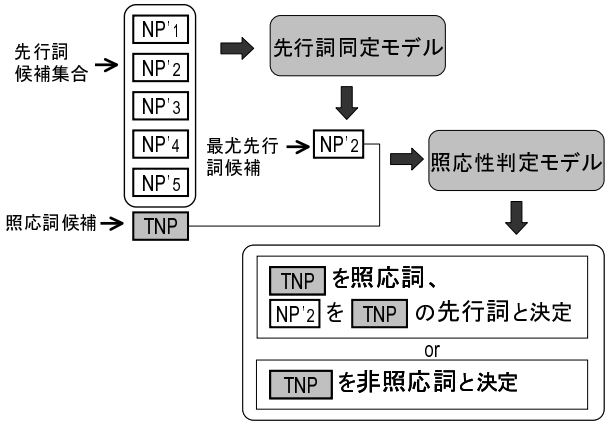

図 3: 照応解析処理の流れ

提案手法では，照応性判定のために分類型手法のモデルを拡張し，照応詞候補と最尤先行詞候補の対を用いて照応性の分類問題を解く．訓練時には，照応詞を正例，非照応詞を負例とし，以下の方法で訓練事例を作成する．正例については，訓練用コーパスに出現する各照応詞に対して先行詞との対を正例集合に加える．図 2 の上部では，照応詞である ANP とその先行詞である NP_4 の対を正例集合に加えている．一方，負例については，非照応詞に対して先行文脈の先行詞集合から先行詞同定モデルを用いて最尤先行詞候補を決定する．この最尤先行詞候補と非照応詞の対を負例集合に追加する．図 2 の下部の例では，非照応詞である $NANP$ に対し先行文脈に出現している先行詞候補（$NP_1,\ldots,NP_5$）の中から最尤先行詞 NP_3 を決定し，NP_3-$NANP$ を負例集合に追加する．この手続きにより，非照応詞に対し先行文脈の情報を明示的に加えて最終的な照応性判定の分類器を作成することができる．

図 3 に提案手法を用いた照応解析処理の全体像を描く．解析の際は，まず対象となる名詞句 TNP に対して，先行文脈に出現する先行詞候補集合（$NP'_1,\ldots,NP'_5$）から先行詞同定モデルを用い最尤先行詞候補を選択する．

表 1: 各モデルが使用する素性

	探索型モデル	分類型モデル	提案モデル	
			先行詞同定	照応性判定
TNP	√	√	√	√
ANT	√		√	√
ANT-ANT			√	

ここでは仮に NP'_2 が最尤先行詞候補として選ばれたとすると，次に解析モデルは NP'_2-*TNP* が照応関係にあるか否かを分類する．もしモデルが対を照応関係にあると分類した場合には *TNP* を照応詞，最尤先行詞候補 NP'_2 を先行詞として出力する．そうでない場合は *TNP* を非照応詞と判断する．

上述のように照応性判定モデルを作成することで，分類型手法で利用できていない先行文脈情報を照応性判定に導入でき，また探索型手法の問題であった非照応詞の情報を明示的に利用できる．つまり，探索型手法と分類型手法のそれぞれの欠点を補いつつ，二つの手法の利点を兼ね備えている．

4 評価実験

Ng ら [5] の探索型モデル，Ng ら [6] の分類型モデル，提案手法のモデルの 3 つのモデルを比較するため，名詞句とゼロ代名詞それぞれについて照応性判定の評価実験を行った．提案手法の先行詞同定の処理には任意のモデルが利用可能だが，今回の実験では我々が提案したトーナメントモデル [18] を利用した．また，3 つのモデルでは分類器として共通に Support Vector Machine[10] を使用した．

4.1 素性

学習には，以下の 3 種の素性を導入した[1]．

- *TNP*: 照応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．
- *ANT*: (i) 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，(ii) 照応詞候補と先行詞候補の関係から抽出可能な情報（例えば，意味的な整合性や二つの候補の距離など）に関する素性．
- *ANT-ANT*: 先行詞候補間の情報（例えば，二つの候補間の距離）に関する素性．

素性 *TNP* と *ANT* は Ng ら [5] の探索型モデルで使用できるが，素性 *ANT-ANT* は先行詞候補間で比較を行うトーナメントモデルを利用する提案手法のモデルでしか使用できない．表 1 にどのモデルがどの種類の素性が使用可能かをまとめる．

実験では，茶筌 [17] と CaboCha[15] を用い形態素解析，固有表現タグ付与，係り受け解析を行い，すべての素性は自動的に抽出した．

4.2 名詞句の照応性判定実験

4.2.1 評価事例

評価実験のために日本語新聞記事 90 記事に照応関係のタグを付与し，照応関係タグ付きコーパスを作成した．このコーパスは 876 の照応詞と 6,292 の非照応詞，あわせて 7,168 の名詞句を含み，照応詞については先行詞がどの名詞句に相当するかのタグも付与されている．実験では，このコーパスを記事単位で分割し，10 分割交差検定を行った．

[1]素性の詳細については文献 [18, 19] を参考にされたい．

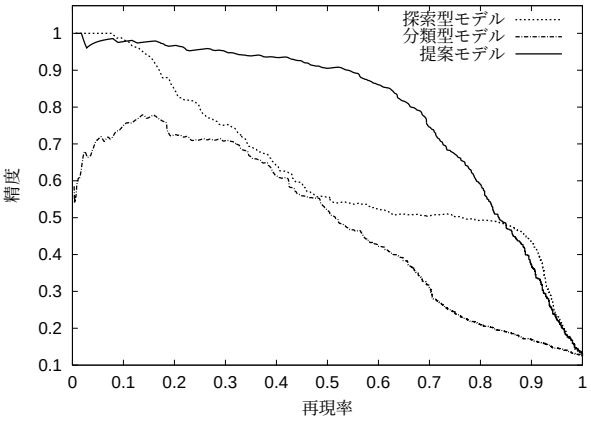

図 4: 名詞句の照応性判定における再現率-精度曲線

表 2: 名詞句の照応性判定における 9 点平均精度

	探索型モデル	分類型モデル	提案モデル
平均精度	63.6%	49.2%	81.1%

4.2.2 実験結果

今回の実験の目的は 3 つの解析モデルの照応性判定の精度を比較することにある．そこで照応詞を正しく検出できた場合を正解とし，再現率，精度を以下の式を使って求める．

$$\text{再現率} = \frac{\text{照応詞を照応詞として正しく検出できた数}}{\text{照応詞の総数}},$$

$$\text{精度} = \frac{\text{照応詞を照応詞として正しく検出できた数}}{\text{システムが検出した照応詞の総数}}.$$

この再現率と精度を使って再現率-精度曲線を図 4 に描く．また再現率がそれぞれ $0.1, 0.2, \ldots, 0.9$ のときの精度から平均精度を求めたものを表 2 に示す．提案モデルと探索型モデルを比較することにより，非照応詞を訓練事例に導入することがどの程度精度向上に貢献しているかを調べることができる．表 2 の 9 点平均精度で比較した場合，約 17%の精度向上が見られ，非照応詞の導入が有効であることがわかる．また，提案モデルと分類型モデルを比較した場合は，先行文脈の情報を利用することがどの程度照応性判定に役立つかを調べることができる．表 2 より，日本語名詞句照応に関してはこの先行文脈の情報を利用した方が非照応詞を利用した場合よりも明らかに精度向上に貢献していることがわかる．

4.3 ゼロ代名詞の照応性判定実験

4.3.1 評価事例

ゼロ代名詞照応性判定の実験には省略タグ付きコーパス [14] の一部 5,682 文からガ格のゼロ代名詞 6,182 事例を抽出し，この事例を用いて 10 分割交差検定を行った．ゼロ代名詞に関しては照応詞が 4,225 事例に対し，非照応詞が 1,957 事例と，名詞句の場合と出現傾向が異なるため，非照応詞を正しく非照応詞として検出できた場合を正解として再現率，精度を算出した．

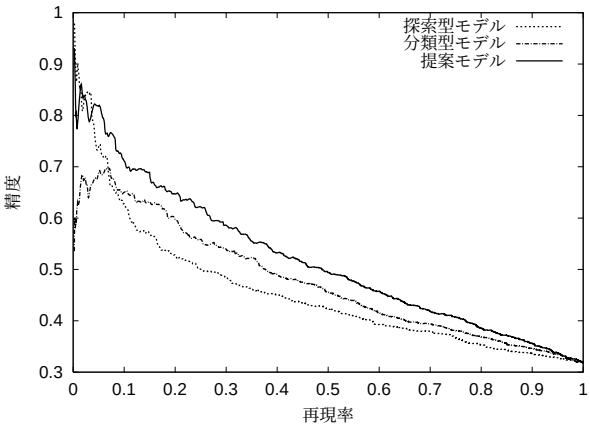

図 5: ゼロ代名詞の照応性判定における再現率-精度曲線

表 3: ゼロ代名詞の照応性判定における 9 点平均精度

	探索型モデル	分類型モデル	提案モデル
平均精度	44.2%	47.3%	50.9%

4.3.2 実験結果

名詞句の場合と同様に，再現率-精度曲線を図 5 に，9 点平均精度を表 3 に示す．結果より提案モデルが他のモデルと比べて精度が若干よいが，名詞句の場合と比較すると他の二つのモデルとの精度の差は小さい．この原因として，名詞句照応とゼロ照応の照応性判定で利用する情報の違いがあげられる．名詞句照応の場合には，照応詞「首相」と先行詞「村山首相」のように照応詞と先行詞の文字列の一致情報を利用することで，先行詞同定の質が向上するとともに，照応性判定においても照応詞となるか否かの良い手がかりとなる．一方，ゼロ照応の場合は文字列一致情報が利用できない．代わりに述語と先行詞候補が選択制限を満たすか否かの情報を利用することになるが，語彙大系 [16] や分類語彙表 [12] の述語辞書から得られる選択制限は粒度が荒く，文字列の一致情報より制約として緩いものになってしまい，その結果照応性判定の精度が低下していると考えられる．

5 おわりに

本稿では，従来の照応性判定モデルがそれぞれ利用していた (i) 非照応詞の情報と (ii) 先行文脈の情報の二つを併用した照応性判定モデルを提案し，名詞句とゼロ代名詞を対象にした照応性判定の評価実験を通じて提案手法の有効性を示した．ただし 4 節に示した実験結果を見ると，名詞句照応，ゼロ照応ともにさらなる精度向上を目指す必要がある．名詞句照応に関しては，我々のこれまでの調査 [19] によると，名詞の定性（definiteness）推定の誤りが名詞句の照応性判定の精度低下のおおきな原因であることがわかっており，今後はこの定性の分析に焦点をあて名詞句の照応性判定の改善を目指したい．また，ゼロ照応に関しては選択制限の質の向上に加え，文章の構造や談話の流れなどさまざまな観点から問題を分析を進め，照応性判定に必要となる情報を考えたい．

参考文献

[1] Baldwin, B.: *CogNIAC: A Discourse Processing Engine*, PhD Thesis, Department of Computer and Information Sciences, University of Pennsylvania (1995).

[2] Bean, D. L. and Riloff, E.: Corpus-based Identification of Non-Anaphoric Noun Phrases, *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 373–380 (1999).

[3] Hobbs, J.: Resolving Pronoun References, *Lingua*, Vol. 44, pp. 311–338 (1978).

[4] Lappin, S. and Leass, H.: An Algorithm for Pronominal Anaphora Resolution, *Computational Linguistics*, Vol. 20, No. 4, pp. 535–561 (1994).

[5] 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 (2002a).

[6] Ng, V. and Cardie, C.: Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution, *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pp. 730–736 (2002b).

[7] 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).

[8] 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).

[9] Uryupina, O.: High-precision Identification of Discourse New and Unique Noun Phrases, *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL) Student Research Workshop*, pp. 80–86 (2003).

[10] Vapnik, V. N.: *Statistical Learning Theory*, Adaptive and Learning Systems for Signal Processing Communications, and control, John Wiley & Sons (1998).

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

[12] 国立国語研究所: 分類語彙表, Vol. 国立国語研究所資料集 6, 秀英出版 (1993).

[13] 山本和英, 隅田英一郎: 決定木学習による日本語対話文の各要素省略補完, 自然言語処理, Vol. 6, No. 1, pp. 3–28 (1999).

[14] 植田禎子, 荻野孝野, 飯田龍, 乾健太郎, 奥村学: 照応, 省略, 共参照タグ付コーパスの構築, 言語処理学会 第 11 回年次大会 発表論文集 (2005).

[15] 工藤拓, 松本裕治: Support Vector Machine を用いた Chunk 同定, 自然言語処理, Vol. 9, No. 5, pp. 3–21 (2002).

[16] 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).

[17] 松本裕治, 北内啓, 平野善隆, 松田寛, 高岡一馬, 浅原正幸: 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 (2002).

[18] 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).

[19] 飯田龍, 乾健太郎, 松本裕治, 関根聡: 最尤先行詞候補を用いた日本語名詞句同一指示解析, 情報処理学会論文誌, Vol. 46, No. 3 (to appear).