

機械学習による日本語名詞句照応解析の一手法*

飯田 龍[†] 乾 健太郎[†] 松本 裕治[†] 関根 聡[‡]
奈良先端科学技術大学院大学[†] ニューヨーク大学[‡]
{ryu-i,inui,matsu}@is.aist-nara.ac.jp sekine@cs.nyu.edu

1 はじめに

文章内の同一指示対象を同定する照応解析は、対話モデルや高品質な機械翻訳システムを実現するために必要とされ、情報抽出や自然言語での質問応答タスクなどの応用分野で特に重要である。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 [1, 9, 10]。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 [2] のような言語学的な知見をもとに規則を記述する。MUC-7¹における照応解析のタスクでは、約70%の精度と約60%の再現率が報告されているが [1]、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 [3, 4, 6] 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、日本語を対象とした名詞句照応解析では、冠詞の情報が無いため、名詞句の指示性の推定が困難であると考えられる。

そこで本稿では、提案する日本語名詞句照応解析モデルを用い、どのような現象がどの程度解析可能かを調べることで、今後我々が真に取り組むべき問題を明らかにする。2節では村田ら [9, 10] の名詞句照応解析の手法について簡単に述べる。3節では新しい名詞句照応解析モデルを提案する。このモデルは、我々が以前提案した先行詞候補間の先行詞らしさを捉えるモデル(トーナメントモデル) [12] を拡張したものである。次に4節では、日本語の名詞句照応解析の実験を行い、解析結果について報告する。最後に5節で解析結果から考察した問題と今後の方針について議論する。

2 先行研究

日本語の名詞句照応解析手法として、村田ら [9, 10] の規則ベースの手法があげられる。

村田らは、人手で作成した86個の規則を用い文章中の名詞句の指示性を可能性と得点という2つの側面ですコアリングすることにより、その指示性を総称名詞句、不定名詞句、定名詞句のいずれかに分類する。次に、定名詞と分類した名詞句のみを対象に名詞句間の同一指示関係を規則に基づいて推定する。

村田らの手法のように最初に名詞句の指示性を分類し、次に検出した定名詞のみを対象に照応関係を推定する処理の流れは一見適切であるように思われる。しかし、日本語では冠詞などの情報が無いため名詞句の指示性を推定することが困難であり、指示性の推定の処理を誤ることで正しく照応関係を認定できない可能性がある。そこで、個々の名詞句が定名詞か否かを推定することなく名詞句間の照応関係を認定する手法を提案する。

3 提案手法

本提案手法では、名詞句照応解析の問題を2つに分割して処理する。つまり、(i) 任意の名詞句(照応詞候補)が与えられた場合に、先行文脈中の名詞句の中で最も先行詞らしい名詞句(最尤先行詞候補)を同定し、次に(ii) その最尤先行詞候補を参照しながら照応詞²を認定することで、最尤先行詞候補と照応詞候補の間の照応関係を決定する。この処理を行うことで、個々の名詞句の指示性を推定するのではなく、照応詞候補と最尤先行詞候補の両方の情報を参照しながら照応詞候補が真に照応詞かそれ以外、つまり先行文脈のどの名詞句とも照応関係にない名詞句(非照応詞)かを判別できるという利点がある。

3.1 最尤先行詞候補の同定

照応詞候補に対して最尤先行詞候補を決めるモデルとして、これまでに我々が提案したトーナメントモデルを用いる。このトーナメントモデルでは照応詞候補に対して、先行するすべての名詞句のうちでどれが最も先行詞らしいかを決定するために先行詞候補間の勝ち抜き戦を行い、最尤先行詞候補を決定する。詳細については文献 [12] を参考にされたい³。

*A Machine Learning-based Method for Resolving Japanese NP Coreference

Ryu Iida[†], Kentaro Inui[†], Yuji Matsumoto[†], and Satoshi Sekine[‡]
Nara Institute of Science and Technology[†] New York University[‡]

¹The Seventh Message Understanding Conference (1998): <http://www.itl.nist.gov/iaui/894.02/related.projects/muc/>

²本研究では、文脈内照応の関係にある名詞句のみを照応詞とする。外界照応を対象としない理由については4.1で後述する。

³[12]では、トーナメントモデルをゼロ代名詞の先行詞同定に用いたが、本手法では名詞句照応解析のために新たに素性を導入して

3.2 照応詞の認定

トーナメントモデルを用いて照応詞候補と対となる最尤先行詞候補を決定したのち、次にその対の情報を参照しながら照応詞候補が照応詞か非照応詞かの分類問題を解くことで照応詞を認定する。そのため、照応詞候補が真に照応詞であり、最尤先行詞候補がその先行詞である場合は正例として分類し、照応詞候補と最尤先行詞候補の対が照応関係にない(照応詞候補が非照応詞である)場合は棄却する分類モデルを作成する必要がある。訓練事例の抽出を図1を用いて説明しよう。図1では、照応詞 ANP に対して4つの名詞句 (NP_1, \dots, NP_4) が先行文脈に出現している状況を仮定している。ANP の先行詞は NP_2 とする。この状況で、分類器は照応詞候補が照応詞か否かの2値分類問題を解く。訓練時には、照応詞とその先行詞(先行詞が複数ある場合は最も近い先行詞)の対 (NP_2 -ANP) を正例とする。また、非照応詞である照応詞候補とその照応詞候補の最尤先行詞候補の対を負例とする。図1の例で NP_5 が非照応詞だったとすると、 NP_5 に対してトーナメントモデルを用いて最尤先行詞候補 NP_3 を決定し、その対 (NP_3 -ANP) を負例として訓練事例に追加する。このように訓練事例を作成することで、3.1の段階で照応詞に対して先行詞を正しく決定できた場合はその照応関係を認定し、また非照応詞の場合には、非照応詞と最尤先行詞候補の対を適切に棄却できると考えられる。

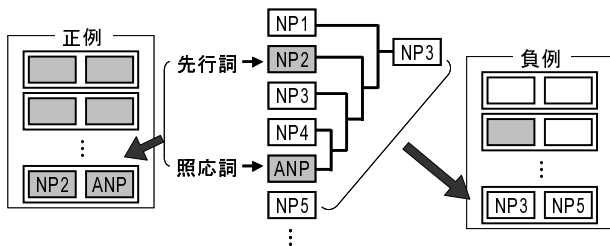


図1: 訓練事例の作成

4 評価実験

4.1 問題設定

作成した照応解析モデルを評価するために、名詞句照応関係タグ付きのコーパスが必要となるが、名詞句照応関係のタグを付与する際、以下に示すような問題を考慮する必要がある。

総称名詞や不定名詞の照応関係: 総称名詞や不定名詞に関しても照応関係と認定可能な場合がある。例えば、次の例の総称名詞「図書館₁」と総称名詞「図書館₂」は同一の概念を指しているために照応関係として認定すべきかもしれない。

図書館₁には本₁が置いてある。
図書館₂の本₂は借りることができる。

これに対し、総称名詞「本₁」と総称名詞「本₂」の場合は、「本₁」が「本を意味する類に属するすべての要素」を指すのに対し、「本₂」は「図書館の本(図書館このモデルを用いる。

に置いてある本₁)」を指し、「本₁」と「本₂」という包含関係が成立するため「本₁」と「本₂」を同一指示の関係として認定するか否かの判断が困難となる。また不定名詞においても同様の現象が起こる。

外界照応: 指示代名詞(「それ」など)や人称代名詞(「私」など)、指示連体詞(「その」など)は文章外の要素と照応関係にある場合がある。例えば、次の例文において「その植物」は文章外のある植物を指示しているため外界照応の関係にある。

庭師はその植物に水をやりましたか。

この外界照応の関係を名詞句照応解析の対象とするか否かが問題となるが、指示連体詞などの明示的な手がかり語が無い名詞句、例えば定名詞句「村山首相」のような場合でも、外界のある人物と外界照応の関係があるとみなすことができるため、揺れなく外界照応を判定することは困難であると考えられる。

複合名詞句の構成素: 照応詞もしくは先行詞の候補となる名詞句が複合名詞句である場合、その名詞句の構成素を解析の対象とするか否かが問題となる。例えば、「[八重洲 東][駐車場]」という4つの形態素を含む複合名詞句を考えた場合「八重洲」「[八重洲 東]」「[駐車場]」などの構成素は解析対象としたいが「[八重洲 東][駐車]」のように複合語の内部構造を無視した名詞句の場合は解析の対象外としたい。しかし、この複合語の内部構造を手で揺れなく同定することは困難である。

そこで上の3つの問題点に対処するため、名詞句照応関係のタグ付与の基準⁴として以下の3つを設定した。

- 総称名詞と不定名詞は照応詞、先行詞として考えない。
- 談話内に出現した名詞句のみを先行詞とする。
- 照応詞は文節の主辞(最右の名詞自立語)を対象とする。

このような基準を採用することで、3節で示した名詞句照応の解析モデルにより適した問題設定となる。上の基準にしたがうと照応関係のタグが付与されていない名詞句(非照応詞)は(a)総称名詞もしくは不定名詞である名詞句、もしくは(b)文脈の前方に先行詞が無い定名詞である名詞句のいずれかとなる。(a)の場合は、照応詞か否かの判別(3.2)の段階で照応詞候補とその最尤先行詞候補の対の両方の情報を参照することにより、総称名詞(もしくは不定名詞)であることを推定し棄却可能である。また(b)の場合も、定名詞と最尤先行詞候補のそれぞれの意味属性が明らかに異なる場合は棄却可能であると考えられる。

4.2 訓練・評価データ

4.1のタグ付けの基準にしたがい、京大コーパス[8]の報道90記事に対して名詞句照応関係のタグを付与した。今回の実験では、883の名詞句間の照応関係を抽出し10分割交差検定を行った。

⁴照応関係タグ付与の基準の詳細は <http://cl.aist-nara.ac.jp/~ryu-i/coreference.tag.html> を参照。

実験では、対象とする文章に対して茶釜 [13] と CaboCha[11] を用い形態構文解析を行い、固有表現のタグを付与した。また、学習器として Support Vector Machine (SVM)[5] を用いた。

4.3 素性

実験では以下に示す 4 種の素性を導入した。

語彙的な情報を用いた素性: 照応詞候補と最尤先行詞候補の 2 つの文字列の“完全一致”, “前方一致”, “後方一致”, “主辞(最右の内容語)一致”, “部分一致”, “構成文字列の一致”などの情報を素性として導入した。これらの素性を用いることにより, 例えば, 最尤先行詞候補「村山富一首相」と照応詞候補「村山首相」の場合は, “主辞の一致”と“構成文字列の一致”の 2 つの一致情報を照応関係の根拠として学習・分類することができる。

形態・統語的な情報を用いた素性: 照応詞候補と最尤先行詞候補それぞれの品詞, 指示詞, 助詞, 対象とする名詞句の連体修飾要素の時制を素性として導入した。これらは, 例えば指示連体詞「その」が名詞句に係る場合は, その名詞句は定名詞である可能性が高いなどの村田らの規則に基づく。

意味的な情報を用いた素性: 照応詞候補と最尤先行詞候補の意味属性が異なる場合は照応関係とならない可能性が高い。今回の実験では, CaboCha が出力した固有表現のタグや分類語彙表 [7] の意味属性を素性として用いた。

名詞句間の距離情報を用いた素性: 照応詞候補と最尤先行詞候補の距離が離れるほど照応関係とならない可能性が高い。そこで, 照応詞候補と最尤先行詞候補の文間の距離を素性として導入した。

4.4 実験結果

まず, 真の照応詞に対してどの程度正しく先行詞を同定できるかを調査したところ, 解析の精度は 86.6% (765/883) であった。次に照応詞の認定を含めた実験結果を表 1 に示す。結果より, 先行詞同定の精度が 86.6% で

表 1: 名詞句照応解析の結果

種類	精度
照応詞の認定と先行詞の同定	65.9% (582/883)
非照応詞の棄却	97.4% (6042/6202)
再現率	65.9% (582/883)
精度	78.4% (582/742)

あるのに対して, 照応詞の認定の処理まで加えると解析の精度 65.9% となり, 照応詞そのものの認定が困難であることがわかる。また, 非照応詞については 97.4% と精度良く棄却することができた。

次に, 表 1 に示した「照応詞の認定と先行詞の同定」の結果を照応詞の種類で分類した(表 2)。ただし, 固有表現には CaboCha の出力する IREX の 8 種の固有表現を, また代名詞には茶釜の解析結果を用いた。固有表現, 代名詞に分類しなかった名詞句を普通名詞とした。表 2 より, 照応詞が固有表現である場合に解析精

度が最も良いことがわかる。固有表現の場合は照応詞と先行詞の文字列が一致するケースが多いため, その特徴をうまく学習できたと考えられる。代名詞に関しては, 学習事例が少なく, またセンタリング理論 [2] などの言語学的手法がかりを直接的に用いていないために, 代名詞がどのような場合にどの先行詞候補と照応関係になりやすいかの傾向が学習できなかったと考えられる。

表 2: 照応詞の種類と精度の関係

種類	(a) 先行詞同定	(a) + 照応詞の検出
固有表現	94.8% (368/388)	84.3% (327/388)
普通名詞	81.5% (392/481)	52.8% (254/481)
代名詞	35.7% (5/14)	7.1% (1/14)

次に, 提案する名詞句照応解析モデルの解析の信頼度について議論する。照応解析においては, 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。そこで, 提案する解析モデルの信頼度を導入し, この信頼度を用いて解析結果の取舍選択をすることを考える。解析の信頼度には SVM のモデルが出力する分離平面からの距離を用いた。導入した信頼度に基づき, 評価事例を信頼度の高いものから順に並べ, Precision-Recall 曲線を描いた結果を図 2 に示す。図 2 を見てわかるように, 再現率を犠牲にすることで高い精度を得ることができた。このことから, 導入した信頼度が有用であるといえる。図 2 より, 提案するモデルでは再現率を約 50% に抑えることで, 精度を約 90% まで上げることができていることを示している。

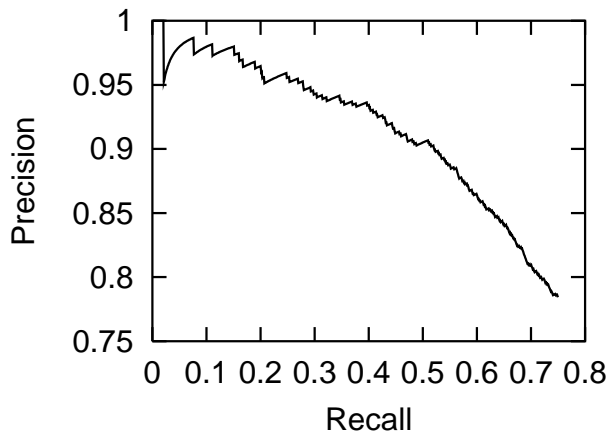


図 2: Precision-Recall 曲線

4.5 誤り分析

実験の結果より, 先行詞同定の解析誤りと照応詞認定の解析誤りそれぞれについて人手で分析した結果を報告する。

4.5.1 先行詞同定の解析誤り

真の照応詞に対する先行詞同定を誤った 118 事例を分析した結果を表 3 に示す⁵。

表 3: 先行詞同定の誤りの割合

誤りの原因	割合
(1) 名詞意味属性の粒度	35.6% (42/118)
(2) 特徴的な語の過剰な重み	16.9% (20/118)
(3) 文字列素性が過剰に働く	18.6% (22/118)
(4) 文章内外の情報が必要	15.3% (18/118)
(5) 定名詞の推定誤り	9.3% (11/118)
(6) その他	22.9% (27/118)

表 3 のうち、典型的な誤りについて以下にまとめる。まず、(1) の名詞意味属性の粒度について述べる。照応詞と先行詞は同じ意味属性を持つことが望ましいが、素性として導入する意味属性の粒度は用いる言語資源に依存する。質の良い意味属性を用いれば、照応詞「今年」に対して名詞句「事業」のような明らかに異なる意味属性に分類されている名詞句を先行詞候補から除外できる可能性がある。しかし、「兄」と「妹」という明らかに照応関係にならない対でも、ある言語資源の体系においては、同一の意味属性に含まれることが多い。そのため、単純に意味属性の一致情報を機械学習に導入することは誤った学習結果を導く恐れがある。また、今回導入した素性集合と事例の作成方法では、先行詞候補が指示代名詞や人称代名詞、接頭辞「同」を含む名詞句である場合は、その名詞句自体に情報がほとんど含まれないために解析を誤る傾向にあることがわかった(2)。先行詞候補がさらに前方の他の名詞句と照応関係にある場合、その関係から推測できる情報を用いることも考えられるが、逆に解析誤りが連鎖的に起こる恐れも出てくる。さらに、今回導入した文字列一致の素性が過剰に働くために解析を誤る場合も見られた(3)。文字列の素性は固有表現の特徴を捉えるためには有効であると考えられるが、逆に照応詞「キリスト教会」に対して「キリスト教会色」のように明らかに照応関係にない最尤先行詞候補を先行詞として決定してしまう。また、表層文字列が同じ名詞句(例えば、「二人」など)が複数回出現して、かつそれらが異なる談話実体を指す場合も誤って先行詞とする傾向が見られた。

4.5.2 照応詞認定の解析誤り

照応詞認定の解析誤りを分析する際には 4.4 で導入した解析の信頼度を用い、その信頼度が高く、かつ解析を誤った 50 事例を分析した。分析結果を表 4 に示す。照応詞認定の誤りの多くは照応詞候補(もしくは最尤先行詞候補)が定名詞か否かの特徴を捉えることができず、総称名詞(もしくは不定名詞)である場合でも照応関係を認定するという誤りである。そのため、さらに名詞句の指示性の特徴を捉えるための手がかりを調査することこそ今後我々が取り組むべき課題の一つであるといえる。

⁵これらの問題は同時に起こる場合もあるため、その場合は 1 つの事例を複数のカテゴリに分類した。

表 4: 照応関係推定の誤りの割合

誤りの原因	割合
(1) 定名詞の推定誤り	50.0% (25/50)
(2) 文字列素性が過剰に働く	14.0% (7/50)
(3) 文章内外の情報が必要	12.0% (6/50)
(4) その他	22.0% (11/50)

5 おわりに

本稿では、最尤先行詞候補を同定した上で照応詞を認定する名詞句照応解析手法を提案した。新聞報道記事に対して名詞句照応解析の実験を行い、再現率 65.9%、精度 78.4% を得た。照応詞が固有表現である場合に解析精度 84.3% を得たが、普通名詞の指示性については表層文字列から得られる単純な素性を用いただけでその特徴を捉えることが困難なことが結果よりわかる。今後は、名詞句照応解析においてもセンタリング理論で導入されている局所文脈の情報をより直接的に用い、かつ照応関係を同定するためににより適切な意味の粒度を調査する必要もある。さらに以前から議論されている名詞句の指示性の推定の問題にも取り組みたい。

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] Grosz, B. J., Joshi, A. K. and Weinstein, S.: Centering: A framework for modeling the local coherence of discourse, *Computational Linguistics*, Vol. 21, No. 2, pp. 203–226 (1995).
- [3] 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).
- [4] 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).
- [5] Vapnik, V. N.: *Statistical Learning Theory*, Adaptive and Learning Systems for Signal Processing Communications, and control, John Wiley & Sons (1998).
- [6] 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).
- [7] 国立国語研究所: 分類語彙表, 国立国語研究所資料集 6, 秀英出版 (1993).
- [8] 黒橋禎夫, 長尾眞: 京都大学テキストコーパス・プロジェクト, 言語処理学会第 3 回年次大会 発表論文集, pp. 115–118 (1997).
- [9] 村田真樹, 黒橋禎夫, 長尾眞: 表層表現を手がかりとした日本語名詞句の指示性と数の推定, *自然言語処理*, Vol.3, No.4, pp. 31–48 (1996).
- [10] 村田真樹, 長尾眞: 名詞の指示性を利用した日本語文章における名詞の指示対象の推定, *自然言語処理*, Vol.3, No.1, pp. 67–81 (1996).
- [11] 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).
- [12] 飯田龍, 乾健太郎, 高村大也, 松本裕治: 文脈の手がかりを考慮した機械学習によるゼロ照応解析, 言語処理学会第 9 回年次大会 発表論文集, pp. 585–588 (2003).
- [13] 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『茶釜』 version 2.2.9 使用説明書, 奈良先端科学技術大学院大学 (2002).