1X-8

日本語文章における直接照応および間接照応の統合的解析*

井之上 直也 † 飯田 龍 ‡ 乾 健太郎 †,§ 松本 裕治 †
† 奈良先端科学技術大学院大学 ‡ 東京工業大学 § 東北大学
{naoya-i,inui,matsu}@is.naist.jp ryu-i@cl.cs.titech.ac.jp

1 はじめに

ある言語表現が他の言語表現と同一の内容，または同じ対象を指すとき，これらの表現は**照応関係**にあるという．例えば，図 1(a) では「このデータ」と「資料」が照応関係になっている．このような照応関係を特定する処理は**照応解析**と呼ばれ，情報抽出や対話システムなどの応用分野で必須となる基盤技術である．照応関係の中でも，図 1(a) のように先行詞と照応詞が同義表現や上位下位関係であるなど，直接的な指示の関係にある場合を**直接照応**という．一方，図 1(b) のように，先行詞と照応詞が部分全体関係や属性関係であるなど，間接的な指示の関係にある場合を**間接照応**という．さらに，先行詞が同一文章内にない場合を**外界照応**という．本論文では，この 3 つの照応関係をまとめて**照応範疇**と呼ぶ．2 つの例からも分かるように，この照応範疇には，照応表現が同じでもどの照応範疇に属すかわからないという曖昧性が存在する．このため，照応解析では 3 つの照応範疇の可能性を同時に考え，先行詞の同定や照応範疇の分類手法の設計が必要になる．ところが 2 節で述べるように，従来の研究には照応範疇の曖昧性についての議論がほとんどない．本研究ではこの問題に注目し，日本語文章内の「この/その/あの＋名詞」のような指示連体詞をともなう照応詞について，先行詞の同定・照応範疇の分類に関する以下の 2 項目を調査する．

調査項目 A：直接照応と間接照応で，個別に先行詞同定モデルを作成すべきか，それとも 2 つの照応範疇を区別せずに 1 つの先行詞同定モデルを作成すべきか．

調査項目 B：照応範疇の分類には前方文脈のどの情報を利用すべきか．

本稿では，まず 2 節で関連研究を紹介し，3 節で詳しい調査内容について説明する．次に 4 節で評価実験の結果を報告し，5 節でまとめる．

2 関連研究

直接照応の先行詞の同定に関連する既存研究としては名詞句共参照解析があり，Message Understanding Conference や Automatic Content Extraction などの評価型タスクを通じて盛んに研究されてきた [4, 7, 他]．また，間接照応はおおむね英語の Briding Reference [1] に相当し，文献 [3] などの先行研究があるが，直接照応に比べて先行研究が少なく，いまだに多くの課題が残っている．先行詞の同定については，このように直接照応と間接照応がそれぞれ独立に研究されてきた．

*Anaphora Resolution for Japanese Definite Noun Phrases
Naoya Inoue†, Ryu Iida‡, Kentaro Inui†, and Yuji Matsumoto†
† Nara Institute of Science and Technology
‡ Tokyo Insititude of Technology

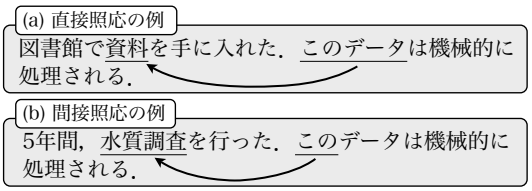

図 1: 直接照応と間接照応の例

照応範疇の分類に関する研究としては文献 [6] がある．Vieira らは英語の定名詞句について，前方文脈の全ての先行詞候補の情報により照応関係を 3 種類[⋆1]に分類したあと，先行詞の同定を行う手法を提案している．評価実験では，照応範疇の分類性能は再現率 57%，精度 70%であったと報告している．しかしながら，Vieira らの手法は人手で作られたルールに基づく発見的な手法のひとつであり，さらに照応範疇の分類に関する先行研究が非常に少ないため，どのような情報が分類に役立つのか十分に調査されているとはいえない．以上の背景から，我々は 1 節で述べた調査項目 A，B のそれぞれについて調査を行った．

3 先行詞同定と照応範疇分類

3.1 照応範疇ごとの先行詞同定

直接照応の先行詞の同定には，照応詞との主辞の一致や意味的類似度が解析の手がかりとなるのに対し，間接照応では先行詞と照応詞の部分全体関係などを認識することが重要だと考えられる．既存研究の多くは直接照応と間接照応のどちらか一方について先行詞の同定を行っているので，直接照応と間接照応の可能性を同時に考えた場合に，解析の手がかりをどのように用いて先行詞の同定を行うべきかは自明ではない．これを明らかにするため，調査項目 A では以下の 2 つの先行詞同定の枠組みを比較する．

混合法：直接照応と間接照応のそれぞれの解析に有効と考えられる情報を全て用いて 1 つの先行詞同定モデルを作成する．

分離法：直接照応の解析に有効と考えられる情報を用いた直接照応の先行詞同定モデルと，間接照応の解析に有効と考えられる情報を用いた間接照応の先行詞同定モデルの 2 つを作成する．

先行詞同定モデルの作成には，日本語の照応解析で高い成果を上げている飯田らの手法 [7] に倣い，トーナメントモデルの枠組みを利用する．トーナメントモデルは先行詞候補間の比較を行い，最終的に最も先行詞らしい候補を決定する枠組みである．訓練事例の作成方法などの詳細については文献 [7] を参照されたい．

[⋆1] 我々の 3 つの照応範疇とおおむね対応する．詳細は文献 [6] を参照されたい．

3.2 前方文脈情報を利用した照応範疇分類

1 節で見たように，照応範疇の分類では前方文脈の情報が解析の手がかりになると考えられるが，前方文脈の全ての先行詞候補の情報や，同定した先行詞の情報など，どのような情報が分類に有効なのかは自明ではない．そこで調査項目 B では，機械学習に基づく 2 つの照応範疇分類モデルの比較[⋆2]により，これを明らかにする．

広域文脈型：照応詞と前方文脈にある全ての先行詞候補の情報を使い，照応範疇を分類する．学習・分類では，照応詞の主辞や格助詞などの情報に加えて，前方文脈の各先行詞候補から抽出される品詞情報や，照応詞と同一主辞を持つ候補があるかなどの情報を抽出し，それらの全てを素性として用いる．これは，2 節で紹介した Vieira ら [6] の手法に準じたモデルである．訓練事例作成の際は，1 つの照応詞とその先行詞候補集合から 1 つの訓練事例を作成する．

分離法最尤先行詞型：照応範疇分類の対象となる照応詞に対し，3.1 節で示した分離法に基づく先行詞同定モデルを用いて，直接照応と間接照応の 2 つの解釈での最尤先行詞を同定し，照応詞とその 2 つの最尤先行詞の情報を使って照応範疇を決定する．学習・分類には，照応詞の主辞や格助詞などの情報に加えて，直接照応として同定された最尤先行詞については直接照応の先行詞同定に利用する素性を，また間接照応として同定された最尤先行詞については間接照応の先行詞同定で利用される素性を抽出し，それら全てを照応範疇分類に用いる．訓練事例作成の際は，(i) 照応詞が直接照応である場合，⟨照応詞，タグ付与された直接照応の先行詞，間接照応の先行詞同定モデルにより同定された最尤先行詞⟩の 3 つ組を，(ii) 照応詞が間接照応である場合には，⟨照応詞，タグ付与された間接照応の先行詞，直接照応の先行詞同定モデルにより同定された最尤先行詞⟩の 3 つ組を，(iii) 照応詞が外界照応である場合には，⟨照応詞，直接照応と間接照応の先行詞同定モデルにより選択した 2 つの最尤先行詞⟩の 3 つ組を訓練事例として抽出する．

4 評価実験

1 節で述べた調査項目 A，B のそれぞれを調査するため，3.1 節と 3.2 節で述べた先行詞同定モデルと照応範疇分類モデルについて，先行詞の同定と照応範疇の分類の評価実験を行った．照応範疇分類モデルの評価にあたっては，前方文脈情報の利用が分類に寄与することを確認するため，照応詞の主辞や品詞などの情報だけを使って照応範疇の分類を行うベースラインモデルを用意した．また，それぞれのモデルの 2 値分類器には多項 2 次カーネルの Support Vector Machines (SVM) [5] を使用し[⋆3]，照応範疇分類モデルの 3 値分類では，SVM による *one-versus-rest* 法を用いた．なお，紙面の都合により説明できなかった各モデルの素性の詳細については，文献 [2] を参照されたい．

4.1 評価事例

NAIST テキストコーパス [8] の報道 2,320 記事（19,669 文）を対象とし，記事中に出現した指示連体詞「この/その/あの」について人手による照応関係および照応範疇の付与を行った[⋆4]．作業の結果，1,749 の指示連体詞にタグが付与され，そのうち 600 が直接照応，901 が間接照応，248 が外界照応であった．評価の際は，直接照応と間接照応については先行詞が文節の主辞であり，かつそれが前方文脈にある事例のみを対象とし，最終的に 1,698 の指示連体詞に対して，記事単位で 10 分割交差検定を行った．

表 1: 先行詞の同定精度（正解率）

	混合法	分離法
直接照応	63.3% (362/572)	**65.4% (374/572)**
間接照応	50.5% (443/878)	**53.2% (467/878)**
全体	55.2% (801/1,450)	**58.0% (841/1,450)**

表 2: 照応範疇の分類性能（F 値）

	BM	C_CS	SC
直接照応	70.9	71.4	**76.1**
間接照応	83.7	80.7	**84.3**
外界照応	48.9	39.4	**57.2**

BM：照応詞の情報だけを利用するベースラインモデル，C_CS：広域文脈型，SC：分離法最尤先行詞型

4.2 実験結果

先行詞の同定結果を表 1 に示す．表 1 の結果より，調査項目 A について，直接照応と間接照応で個別に先行詞同定モデルを作成すべきだということがわかった．次に，照応範疇の分類結果を表 2 に示す．表 2 の結果より，調査項目 B について，照応範疇の分類にはあらかじめ選択した最尤先行詞の情報を前方文脈情報として用いることが有効だとわかった．また，ベースラインモデルより広域文脈型の分類性能が全体的に低いことから，全先行詞候補の情報を単純に素性に加えると解析に悪影響を及ぼすことがわかった．

5 おわりに

本稿では，日本語文章における直接照応と間接照応，外界照応がこれまで独立に研究されてきたことを指摘した．また，先行詞の同定と照応範疇の分類を統合的に扱う方法を提案し，その有効性の調査を行った．調査の結果，(i) 先行詞同定モデルは直接照応と間接照応を区別して作成すべきであり，(ii) 照応範疇の分類には，あらかじめ選択した最尤先行詞の情報を前方文脈情報として用いることが有効であることが分かった．今後は，特に精度の低かった間接照応の先行詞の同定について，名詞句間の意味関係の認識方法などを検討していきたい．

参考文献

[1] Clark, H. H.: Bridging, *Thinking: Readings in Cognitive Science* (1977).
[2] Inoue, N., Iida, R., Inui, K. and Matsumoto, Y.: Resolving Direct and Indirect Anaphora for Japanese Definite Noun Phrases, *Journal of Natural Language Processing*, Vol. 17, No. 1, pp. 221–246 (2010).
[3] Poesio, M., Mehta, R., Maroudas, A. and Hitzeman, J.: Learning to resolve bridging references, *Proc. of ACL*, pp. 144–151 (2004).
[4] Soon, W. M., Ng, H. T. and Lim, 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.: *The Nature of Statistical Learning Theory*, Wiley (1995).
[6] Vieira, R. and Poesio, M.: An Empirically Based System for Processing Definite Descriptions, *Computational Linguistics*, Vol. 26, No. 4, pp. 539–593 (2000).
[7] 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, pp. 831–844 (2005).
[8] 飯田龍, 小町守, 乾健太郎, 松本裕治: NAIST テキストコーパス：述語項構造と共参照関係のアノテーション, 情報処理学会研究報告（自然言語処理研究会）NL-177-10, pp. 71–78 (2007).

[⋆2]混合法の最尤先行詞を用いることも選択肢として考えられるが，あとの実験で混合法の先行詞同定モデルは低い精度を示すため，分離法の最尤先行詞を使うモデルの結果のみを掲載した．

[⋆3]SVM の実装には SVM*light* (http://svmlight.joachims.org/) を利用し，パラメタはデフォルト値を用いた．

[⋆4]公開されている NAIST テキストコーパスでは，厳密に同一実体を参照している場合にしか照応関係のタグが付与されておらず，我々の調査に必要な照応関係について網羅的にタグ付与されていないため，あらためて人手でのタグ付与作業を行った．