指定指示・代行指示を区別した指示連体詞の照応解析*

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 情報科学研究科
{naoya-i,inui,matsu}@is.naist.jp
‡ 東京工業大学 大学院情報理工学研究科
ryu-i@cl.cs.titech.ac.jp

1 はじめに

ある言語表現が文章内の他の表現を指す時、それらの表現は**照応関係**にあるという。このとき、指示する側を**照応詞**、指示される側を**先行詞**と呼ぶ。例えば、例 (1) では照応詞「このデータ」と前方文脈の先行詞「資料」が照応関係にある。

(1) 図書館で 資料 を手に入れた。このデータ は機械的に処理される。

このような照応関係を特定する処理を照応解析といい、質問応答や情報抽出システムなどの実現に必要となる重要な要素技術である。照応関係の中には例 (1) の指示連体詞「この」のような指示表現を伴う場合があり、ここでは「指示連体詞＋名詞」が照応詞となっている。このように、名詞句全体が照応詞となる場合を**指定指示 (限定指示)** という。一方、例 (2) では指示連体詞「この」が前方文脈の先行詞「水質調査」と照応関係にある。

(2) 5 年間、水質調査 を行った。この データは機械的に処理される。

このように、指示連体詞のみが先行詞と照応関係となる場合を**代行指示**という。また、本稿では指定指示・代行指示を総称して**指示関係**と呼ぶ。この指示関係には、上述の例のように照応詞側の表現が同じ場合でも、指定指示となるか代行指示となるかわからないという曖昧性が存在する。したがって指示連体詞の照応解析には、先行詞同定に加えて指示関係の分類が必要となる。

照応解析の既存研究では近年成熟してきた機械学習を導入することで解析精度を向上させてきたが [7, 5]、そのほとんどが指定指示と代行指示を独立に扱っており、指示関係の問題をどうモデル化するかについての議論は少ない。我々も機械学習に基づく照応解析手法を採用するが、本研究では特に指示連体詞の先行詞同定と指示関係分類について以下の 2 つの項目について調査を行う。

調査項目 1 先行詞同定と指示関係の区別
先行詞同定は指定指示と代行指示で共通のモデルを利用すべきか、指示関係ごとに独立に先行詞を同定するモデルを利用すべきか。

調査項目 2 指示関係分類に有効な先行文脈情報
前方文脈の全先行詞候補から抽出できる情報や、明示的に同定した先行詞から得られる情報のうち、どのような情報が指示関係分類に役立つのかを調査する。

本稿では、まず 2 節で 2 つの調査項目の詳細について説明し、3 節で関連する先行研究を紹介する。4 節では実験に用いるデータセットについて説明し、5 節で実験結果について報告する。最後に、6 節でまとめを行う。

2 先行詞同定と指示関係分類

2.1 先行詞同定と指示関係の区別

指定指示の先行詞同定では、先行詞と照応詞の文字列一致情報や 2 つの表現が同義関係となるか否かなどが重要な解析の手がかりとなるのに対し、代行指示の場合は「(先行詞) の (照応詞)」のような表現で表される意味関係が手がかりになると考えられる。ここでは、それぞれの指示関係ごとに利用する情報を明示的に分けることが有効かどうかを調査するため、以下の 2 つの先行詞同定モデルを比較する。

共通モデル 指定指示と代行指示を区別せずに 1 つの先行詞同定モデルを作成する。表 1 で後述するように、このモデルでは 2 つの指示関係の解析のそれぞれに重要だと考えられる素性を、区別せずに全て利用して学習・分類を行う。

独立モデル 指示関係が指定指示の場合と代行指示の場合で訓練事例を分けて学習し、指定指示用の先行詞同定モデルと代行指示用の先行詞同定モデルの 2 つを作成する。表 1 からわかるように、2 つのモデルで利用する素性は一部異なる。

それぞれのモデル作成には、候補間を比較し最終的に最も先行詞らしい候補を同定する、飯田ら [12] のトーナメントモデルの枠組みを利用する。訓練事例の作成方法などについての詳細は、文献 [12] を参照されたい。

2.2 指示関係分類に有効な先行文脈情報

2.1 で述べた先行詞同定の調査に加えて、指示関係の分類についても有効な先行文脈情報を検証する。以下で、比較する指示関係分類モデルについて図 1 を用いて説明する。

指示関係分類先行型 (前分類型) このモデルでは先行文脈の全ての先行詞候補の情報を用いて指定指示か代行指示かを決定した後、その指示関係に対応した先行詞同

*Resolving direct and indirect anaphora for Japanese demonstrative determiners
Naoya Inoue†, Ryu Iida‡, Kentaro Inui†, and Yuji Matsumoto†
† Nara Institute of Science and Technology
‡ Tokyo Insititude of Technology

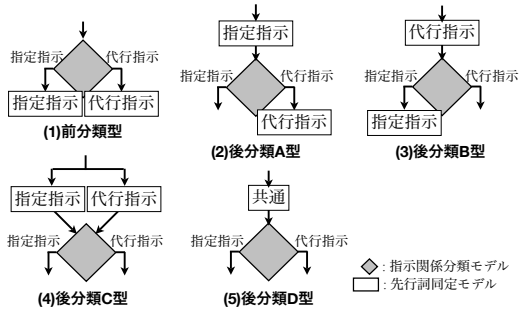


図 1: 比較する照応解析モデル

定モデルで先行詞同定を行う (図 1-1)。指示関係分類の際は、先行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か否か、照応詞との類似度などの情報を抽出し、それらの全てを先行文脈の素性ベクトルとする。ただし、先行詞候補と照応詞の類似度など 2 値素性でないものについては、候補中の最大値を各素性の値とする。

先行詞同定先行型 (後分類型) 先行文脈中の候補全体を利用する代わりに、先行詞同定モデルで同定した先行詞の情報を用いて指示関係分類を行う。これは、名詞句共参照解析手法において名詞句が照応詞か否かを分類する際に、先行文脈全体の情報よりあらかじめ同定した先行詞の情報を使って分類したほうが、高い分類精度を得られたという飯田ら [12] の報告に基づく。ここでは、指定指示を仮定して先行詞を同定し、その結果が指示関係分類モデルによって代行指示と分類された場合は代行指示の先行詞同定を行う **A 型**(図 1-2)、その逆の **B 型**(図 1-3)、各指示関係の先行詞同定を並行に行い、その両方の結果を用いて分類を行う **C 型**(図 1-4)、共通モデルで先行詞を同定し、その結果から分類を行う **D 型**(図 1-5) の 4 つの指示関係分類の精度を調査する。訓練事例作成の際は、利用する先行詞同定モデルと指示連体詞の指示関係が同じ場合はタグ付与された先行詞と照応詞の対、指示関係が異なる場合は先行詞同定モデルが出力する先行詞と照応詞の対から素性を抽出し、指示連体詞が指定指示ならば正例、代行指示ならば負例とする。

3 先行研究

指定指示に関連する先行研究としては、文章内の名詞句同士が同一実体を指示するかを同定する名詞句共参照解析があり、Message Understanding Conference(MUC)[*1] や Automatic Content Extraction(ACE)[*2] などの評価型タスクを通じて研究が盛んに行われ [7, 5]、MUC-6 の問題については F 値で約 7 割という結果が報告されている。代行指示は、英語の Bridging Reference [1] に相当するが、その解析には照応詞と先行詞の関係に同義関係、上位下位関係、部分・全体関係などの様々な関係を捉える必要があり、照応解析の中でも難しい問題とされている。主な関連研究には、Poesio ら [6] の研究などがある。

指定指示と代行指示の両方を扱う関連研究としては、Vieira ら [9] の英語の定名詞句に対する照応解析の枠組みがある。彼らは定名詞句の照応現象を (i) 先行詞と照応詞の主辞が一致する照応、(ii) 主辞が一致しない外的知識を介した照応、(iii) 新出談話要素の 3 つに分類しており、2.2 で述べた前分類型で、つまり先行文脈全体の情報から指示関係を分類したのち先行詞を同定するという手法を用いている。また、村田ら [11] の手法では、人手で作成した 100 の規則を用いて日本語の指示表現・ゼロ代名詞の照応解析を行っている。彼らの手法では指示連体詞の解析は指定指示から代行指示の順で決定的に行われ、指定指示の規則で解析できない場合に代行指示の先行詞を探索している。この手法は、2.2 の後分類 A 型に相当する。

4 人手によるタグ付与

本評価実験は、NAIST テキストコーパス [13] の 2,929 記事 (38,384 文) 中 の指示連体詞を対象にする。ただし、NAIST テキストコーパス中の照応関係は厳密に現実 (仮想) 世界の同一実体を参照しなければならないという強い制約のもとでタグ付与されているため、本研究で対象とする指示連体詞の照応関係について網羅的にタグ付与されていない。そこで、今回我々は改めて指示連体詞「この」「その」「あの」に対して照応関係[*3]および指定指示・代行指示の区別を人手で付与した。照応関係の付与においては、「その朗読」が「読み上げた」を指示するような、先行詞が述語の場合もタグ付与の対象とし、先行詞が文章内に無い場合には外界指示とみなし、その情報も付与した。指示関係の付与については、指定指示とも代行指示とも解釈できる場合には指定指示を優先した。このような事例の多くは、例えば「この大地」が「日本」を指示するような地政学的実体が多く、ここでは ACE の仕様にも採用されているように指定指示に分類した。作業の結果、1,449 記事 (報道記事 869、社説記事 580) 中の 4,087 の指示連体詞にタグ付与が行われ、このうち指定指示は 31.0%(1,269/4,087)、代行指示は 57.2%(2,336/4,087)、外界指示は 11.8%(482/4,087) であった。

5 評価実験

タグ付与された報道記事 869 記事中の 1,754 の指示連体詞を対象に、10 分割交差検定による評価実験を行った。2 値分類器には線形カーネルの Support Vector Machine(SVM) [8] を使用し、SVM の実装には SVM*light*[*4] を用い、実験を通じてパラメータは c=1.0 とした。

評価の対象とする指示連体詞は、先行詞が前方文脈にあり、かつそれが名詞句の主辞のみの場合の 1,363 事例とした (指定指示が 40.6%(553/1,363)、代行指示が 59.4%(810/1,363) であった)。事例抽出の過程では、後方照応 0.6%(11/1,754)、外界指示 14.7%(258/1,754) の事例を除外し、さらにこのうち先行詞が複合名詞の主辞以外の一部である 8.2%(122/1,485) の事例を除外した。

[*1] http://www-nlpir.nist.gov/related_projects/muc/index.html
[*2] http://www.nist.gov/speech/tests/ace/
[*3] 照応関係であるが共参照関係ではない identity-of-sense anaphora [4] の関係も含む
[*4] http://svmlight.joachims.org/

表 1: 先行詞同定と指示関係分類に用いる素性

種類	素性名	詳細
(a)	CANDIDATES_SENTDIST	$CAND_i$ 同士の文距離。
	CAND_ANP_SENTDIST	$CAND_i$ と ANP の文距離。
	CAND_ANP_COMB_{HEAD, CASE, POS_NE}	$CAND_i$ と ANP の主辞の組合せ、格助詞の組合せ、品詞と固有表現種類の組合せ。
	CAND_DEPENDED	$CAND_i$ に対して係る文節数。
	CAND_{POS***, NE*, CASE}	$CAND_i$ の品詞、固有表現の種類、格助詞。
	CAND_PARRAREL	$CAND_i$ の右方向で最も近い句読点が「、」ならば 1。それ以外は 0。
	CAND_DESCRIPTIVE	$CAND_i$ が「な」または「の」で係る場合は 1。それ以外は 0。
	CAND_EOS_ANP_BOS	$CAND_i$ が文内の最終文節内にあり、ANP が文頭にある場合は 1。それ以外は 0。
	CAND_ANP_NOUN_SIM*	$CAND_i$ と ANP の名詞類似度。
	CAND_ANP_HYPONYM_OF_ANP*	$CAND_i$ が ANP の下位語ならば 1。それ以外は 0。
	CAND_ANP_SYNONYM_OF_ANP*	$CAND_i$ と ANP が同義関係ならば 1。それ以外は 0。
	CAND_ANP_STR_LAST_MATCH(_HEAD)*	$CAND_i$(の主辞) が ANP と後方一致すれば 1。それ以外は 0。
	CAND_ANP_STR_PART_MATCH(_HEAD)*	$CAND_i$(の主辞) が ANP と部分一致すれば 1。それ以外は 0。
	CAND_ANP_STR_COMP_MATCH*	$CAND_i$ が ANP と完全一致すれば 1。それ以外は 0。
	CAND_ANP_PMI**	$CAND_i$ と ANP の共起関係の強さを表す相互情報量。
	ANP_TYPE***	「この」「その」のような指示連体詞の種類。
	ANP_{POS***, CASE}	ANP の品詞、格助詞。
(b)	ANP_HEAD	ANP の主辞。
	MAX_PMI	先行詞候補中で最大となる、$CAND_i$ と ANP の共起関係の強さを表す相互情報量。
	MAX_NOUN_SIM	先行詞候補中で最大となる、$CAND_i$ と ANP の名詞類似度。
	HAS_SYNONYM_OF_ANP	ANP と同義関係となる $CAND_i$ があれば 1。それ以外は 0。
	HAS_HYPONYM_OF_ANP	ANP の下位語となる $CAND_i$ があれば 1。それ以外は 0。
	HAS_STR_LAST_(HEAD)_MATCHED	ANP(の主辞) と文字列一致する $CAND_i$ があれば 1。それ以外は 0。
	HAS_STR_COMP_MATCHED	ANP と完全に文字列一致する $CAND_i$ があれば 1。それ以外は 0。

(a) 先行詞同定モデルの素性。$CAND_i$ は比較する 2 つの先行詞候補、ANP は照応詞を表す。'*'は指定指示の先行詞同定モデルと共通モデル、'**'は代行指示の先行詞同定モデルと共通モデルで用いる。(b) 前分類型の指示関係分類モデルの素性。a の '***'とともに用いる。いずれも $CAND_i$ は先行文脈中の先行詞候補を表す。後分類型の指示関係分類モデルには用いる先行詞同定モデルの素性をそのまま用い、C 型の場合は 2 つのモデルの素性に接頭辞を付け区別して扱う。

5.1 素性

今回の実験で先行詞同定と指示関係分類に使用した素性を表 1 に示す。指定指示の先行詞同定には、品詞や格助詞、文字列一致の素性のほか、例 (3) に示すような上位下位関係・同義関係を捉えるための外部知識として、隅田ら [10]、井佐原ら [3] の辞書を用いた。

(3) ...新型車 が発売された。このミニバン は...

さらに先行詞と照応詞の同義関係を捉える指標として、新聞記事約 20 年分から抽出した、名詞と格助詞・動詞の共起情報を利用した。共起データは pLSI [2] でスムージング (隠れクラス数 1,000) し、隠れクラスへの帰属確率を素性として学習したベクトル間のコサイン類似度を名詞間距離として素性に組み込んだ。

代行指示の先行詞同定では、例 (4) に示すような「(名詞 A) の (名詞 B)」の共起関係の強さを捉えるため、新聞記事約 20 年分から抽出した共起データを用いて先行詞と照応詞の自己相互情報量を素性として組み込んだ。

(4) ...分子構造$_A$ である。この 変化$_B$ が...

5.2 実験結果

先行詞同定モデルの精度を表 2 に示す。次に、指示関係分類に用いる情報の有効性を評価するため、以下に定義した正解率を用いて各モデルの解析精度を比較する。

$$\text{正解率} = \frac{\text{指示関係・先行詞ともに正しく解析できた個数}}{\text{指示連体詞の総数}} \quad (1)$$

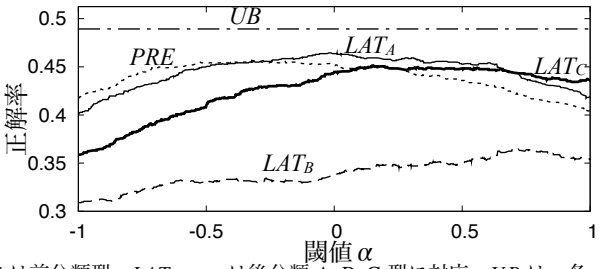
また、ここでは指示関係分類器の挙動を見るため、分類器が出力するスコア (分離平面からの距離) に対して α 以上なら指定指示、それ以外なら代行指示とする閾値 α を変化させ、正解率の最大値を見積もる。ただし紙面の都合上、後分類型は精度の高かった独立モデルを先行詞同定に用いる A,B,C 型のみを対象とし、実験結果を図 2 に示す。

表 2: 先行詞同定モデルの精度

モデル	指定指示	代行指示	全体
独立	58.2%(322/553)	42.6%(345/810)	48.9%(667/1,363)
共通	59.5%(329/553)	28.8%(233/810)	41.2%(562/1,363)

さらに各モデルの指示関係分類精度の上限値を見積もるため、先行詞同定が常に正しく行われると仮定した基での正解率を比較する。すなわち、先行詞同定の結果には、利用する先行詞同定モデルと指示連体詞の指示関係が同じ場合はタグ付与された正解の先行詞を、指示関係が異なる場合は学習した先行詞同定モデルが実際に出力する先行詞を用いる。この実験結果を図 3 に示す。以下に、これら 2 つの評価実験の結果から分かったことを 1 節の 2 つの調査項目に対応付けて説明する。

- 調査項目 1 については表 2 の結果より、先行詞同定は指定指示・代行指示を区別して独立に学習・分類するモデルを用意すべきだということが分かった。
- 調査項目 2 については、図 2 と図 3 から分かるように、現状の先行詞同定精度では指示関係分類に指定指示の先行詞同定結果のみを使う後分類 A 型、先行詞同定が正しく行える状況では両方の指示関係の先行詞同定結果を利用する後分類 C 型が、最も高い正解率を得た。これらの結果より、先行文脈の先行詞候補全てを用いるよりも、明示的に同定した先行詞を用いる方が正しく指示関係を分類できることがわかった。特に、正しい先行詞を指示関係分類器に提示できる場合は、指定指示・代行指示の両方の解析結果を用いるのが有効だとわかった。



PRE は前分類型、$LAT_{\{A,B,C\}}$ は後分類 A, B, C 型に対応。*UB* は、各指示連体詞について指示関係は正しく分類できたと仮定し、また先行詞同定については独立モデルを用いて先行詞同定を行った場合の正解率を表す。

図 2: 照応解析の正解率

- 先行詞が正しく同定できると仮定した場合でも後分類 B 型の正解率が低い原因は、先行詞情報の抽出に用いる代行指示先行詞同定モデルの素性が代行指示の特徴を適切に捉えられなかったためだと考えられる。
- 後分類 B 型に比べ、先行詞同定モデルの精度が高い後分類 A 型のほうが高い正解率を得たという結果より、先行詞同定の精度を向上させることで最終的な正解率が向上することを期待できる。

5.3 先行詞同定の誤り分析

指定指示の先行詞同定の誤り事例のうち、10 分割交差検定の結果から SVM の返したスコアが最も高い 5 つを抽出し、50 事例について誤り分析を行った。この結果、誤り原因の半分 (26/50) は、文字列一致素性が過剰に働いたため正解の先行詞が棄却されたことだとわかった。例えば、例 (5) では照応詞「この右ハンドル型ミニバン」の先行詞は「新型車」であるが、先行詞候補「ミニバン」と「新型車」を比較する際に文字列一致素性が過剰に働き「ミニバン」を誤って先行詞と同定してしまう。

(5) ... クライスラーの 新型車 だ。商用バンをベースにし、スタイルが角張っているとの不評もあった従来の ミニバン から大きく脱皮、「乗用車を思わせる画期的なスタイル」となった。クライスラーのイートン会長は「市場アクセスの改善いかんだ」としながらも、この右ハンドル型ミニバン の日本への投入...

また、文字列一致素性の誤りの中には照応詞「この日」に対して「21 日」「4 日」などの同一接尾辞の先行詞候補が複数出現し、同定を誤るものも見られた。

代行指示の先行詞同定では、約 9 割の解析結果が照応詞の直前に位置する先行詞候補を選択していた。また、本実験で重要な解析手がかりとした先行詞候補と照応詞の相互情報量の素性には、学習の結果、非常に小さな重みが割り当てられていた。そこで相互情報量の推定に用いた共起事例を見てみると、一般に出現すると思われる「少女 の 成長」のような共起が 1 度も出現しないことがわかった。このことから、先行詞候補と照応詞の関係が捉えられず代行指示の特徴を適切に学習できなかったと考えられるため、今後は先行詞が述語の場合も含め、解析の手がかりとする指標をさらに検討する必要がある。

6 おわりに

指示連体詞の照応解析に必要となる、先行詞同定と指定指示・代行指示の指示関係分類に対する調査を行った。

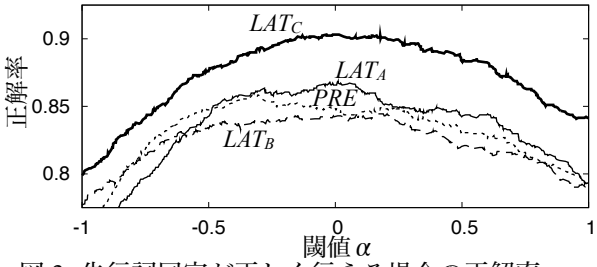


図 3: 先行詞同定が正しく行える場合の正解率

この結果、先行詞同定については指示関係を区別し独立に解析モデルを構築したほうが、区別しない場合に比べて精度がよいことがわかった。また、指示関係分類については、指定指示・代行指示それぞれについてあらかじめ最も先行詞らしい候補を同定しておき、それらを手がかりとすることが分類精度に最も貢献できることを明らかにした。先行詞同定の精度向上が指示関係分類の精度向上に繋がる事も確認できたため、今後は特に精度の悪かった代行指示の先行詞同定に関して、名詞句間の意味関係を捉える指標など、どのような情報が必要になるかをさらに吟味する必要がある。さらに、実際の解析では文章内に指示対象を持たない外界指示が存在するため、外界指示を含めた 3 種類の指示関係分類についても今後取り組む予定である。

参考文献

[1] Clark, H. H.: Bridging, *Thinking: Readings in Cognitive Science* (1977).

[2] Hofmann, T.: Probabilistic Latent Semantic Indexing, *Proceedings of the Twenty Second Annual International SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57 (1999).

[3] Isahara, H., Bond, F., Uchimoto, K., Utiyama, M. and Kanzaki, K.: Development of the Japanese WordNet, *In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)* (2008).

[4] Mitkov, R.: *Anaphora Resolution*, Studies in Language and Linguistics, Pearson Education (2002).

[5] Ng, V. and Cardie, C.: Improving machine learning approaches to coreference resolution, *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 104–111 (2001).

[6] Poesio, M., Mehta, R., Maroudas, A. and Hitzeman, J.: Learning to resolve bridging references, *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 144–151 (2004).

[7] 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).

[8] Vapnik, V. N.: *The Nature of Statistical Learning Theory* (1995).

[9] Vieira, R. and Poesio, M.: An Empirically Based System for Processing Definite Descriptions, *Computational Linguistics*, Vol. 26, No. 4, pp. 539–593 (2000).

[10] 隅田飛鳥, 吉永直樹, 鳥澤健太郎, 萬成賢太郎: Wikipedia からの大規模な上位下位関係の獲得, 言語処理学会第 14 回年次大会, pp. 769–772 (2008).

[11] 村田真樹, 長尾真: 用例や表層表現を用いた日本語文章中の指示詞・代名詞・ゼロ代名詞の指示対象の推定, 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, Vol. 95, No. 168, pp. 43–50 (1995).

[12] 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).

[13] 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報告, No. 2007-NL-177, pp. 71–78 (2007).