数学問題テキストに対する照応・共参照解析

伊藤　巧 †　　松崎　拓也 ‡　　佐藤　理史 ‡
† 名古屋大学 工学部 電気電子・情報工学科　‡ 名古屋大学大学院 工学研究科

1　はじめに

2011 年に，国立情報学研究所を中心とする「ロボットは東大に入れるか（略称: 東ロボ）」というプロジェクトが開始された [1]．このプロジェクトは人工知能技術で大学入試問題を解くことに挑戦するものである．

我々はこのプロジェクトにおいて，数学の言語処理部の開発に取り組んでいる．本稿では，その一部である照応・共参照解析の現状を述べる．これまで，照応・共参照解析に対する研究は多くなされているが，主として新聞等のテキストが対象であり，数学問題等のテキストを対象にした照応・共参照解析の研究は少ない．

以下では，2 節で数学問題中の照応表現の種類および出現数に関する調査結果を報告し，3 節で現在開発中のシステムの構成を示す．さらに，4 節でそのシステムの評価結果を示し，5 節で結果の分析を行う．

2　照応表現の調査

2.1　照応表現の種類

これまで数学問題で観察した照応表現を表 1 にまとめる．不飽和名詞とは，意味的に自立していない名詞である．例えば「斜辺」,「半径」などの名詞は，どの直角三角形の斜辺なのか，どの円，または球の半径なのかという情報が必要とされる．数学問題では，不飽和名詞の項（主としてノ格）がゼロ代名詞化する場合が多い．その例として以下のような問題が挙げられる．

> $\angle C = 90°$ の直角三角形 ABC がある．
> (φの) 斜辺の長さを c とすると，…

一方，述語の項がゼロ代名詞化する例は比較的まれであり，新聞等で項が頻繁に脱落するのと対照的である．

一般名詞による照応の例には次のようなものがある．

> k を自然数とし，方程式 $3x + y = k$ を考える．
> (1) 方程式 の正の整数解の組 (x, y) の個数は…

上例では，下線部の「方程式」は一般的な方程式を表しているのではなく，方程式 $3x+y = k$ を指している．

表 1: 照応表現の種類と例

連体詞形態指示詞（単数）	「その」「この」
連体詞形態指示詞（複数）	「それらの」「これらの」
名詞形態指示詞（単数）	「それ」「これ」
名詞形態指示詞（複数）	「それら」「これら」
不飽和名詞の項となるゼロ代名詞	「(φの) 斜辺」「(φの) 半径」
一般名詞	「方程式」
条件指示詞	「そのとき」「このとき」
その他	「それぞれ」「方」「もの」

表 2: 照応表現の出現数

	センター	二次試験
連体指示詞 (単数)	26	12
連体指示詞 (複数)	2	2
名詞指示詞 (単数)	2	1
名詞指示詞 (複数)	1	1
ゼロ代名詞	121	61
並列	23	7
連体修飾節の中	17	23
不飽和名詞を項とする述語	33	23
一般名詞	1	5
条件指示詞	57	33
その他	29	16

2.2　照応表現の出現数

1998 年度〜2014 年度の偶数年度のセンター試験『数学 IA』『数学 IIB』の問題 88 題と 1990 年度〜2014 年度の国公立および私立大学入試の二次試験『数学』の問題から無作為に選んだ 100 題に出現する照応表現の数を調査した．結果を表 2 に示す．

センター試験の方が 1 問あたりの問題文が長いため，照応表現の数も多くなった．また，一般名詞が照応的に使用される例は少ない．その一つの理由は,「円 C」のように一般名詞は記号を伴って出現することが多いことである．一方，ゼロ代名詞の出現数は非常に多い．これは，以下のような文法的に指示対象が決まる場合もカウントしたためだと考えられる．なお，以下の例では先行詞に下線が引いてある．

並列：f の最大値と (φの) 最小値
連体修飾節の中：(φの) 半径が 3 の 円
不飽和名詞を項とする述語：f(x)=0 は (φの) 実数解をもつ

3　システムの構成

3.1　数学解答システムの言語処理部

本研究で開発している照応解析システムを含む言語処理部全体の大まかな構成を図 1 に示す．照応解析システムは，構文解析の前に，照応表現を先行詞で書き換えるための，一種の前処理として実行される．言語処理部の解析の例を図 3 に示す．

ここで，照応解析の出力で用いる 3 種類のタグについて説明する．1 つ目は，先行詞を示す「<1>円</1>」のような番号のみのタグである．2 つ目は，指示詞が含まれる照応で用いる「<coref ant="1">これ</coref>」のようなタグである．ant 属性は照応する先行詞の番号を値とする．3 つ目は，ゼロ代名詞の照応で用いる「<coref ant="1"case="の"/>半径」のような空タグである．case 属性は，ゼロ代名詞の格を表す．

3.2　照応解析システムの構成

照応解析システムの構成を図 2 に示す．入力は，数式部分が MathML で書かれた xml 形式の数学問題である．システムでは，まず数式の内容に基づいてそれぞれの数式に，その意味タイプを表すタグを付与する．付与した意味タイプは，先行詞の決定の際に用いる．例えば，数式が $\triangle ABC$ のような形であれば，「三角形」を表すタグを付与する．次に，ゼロ代名詞を含めた照応表現の検出を行い，それぞれの照応表現に対して，先行詞の意味タイプを推定する．そして，タイプに適合する名詞・数式・記号 (以後，先行詞候補と呼ぶ) の中から，先行詞の同定をする．最後に，解析結果として得られた照応関係に基づき，照応表現を指示内容で書き換えた問題文を出力する．

以降では，まず，照応解析処理で用いる格フレーム辞書および用語タクソノミーについて述べ，次に，照応表現のタイプ毎に照応表現の検出および先行詞の同定アルゴリズムについて述べる．

3.3　格フレーム辞書と用語タクソノミー

本システムでは，照応詞の検出および先行詞の同定の際に格フレーム辞書と用語タクソノミーを用いている．格フレーム辞書は，動詞と不飽和名詞の 2 種類の辞書を用いている．動詞の格フレーム辞書は 1842 フレーム，不飽和名詞は 162 フレームから成る．それぞれの例を図 3.3 に示す．Vector はベクトル，Shape は

図 1: 言語処理部の構成

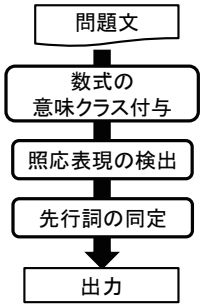

図 2: 照応解析システムの構成

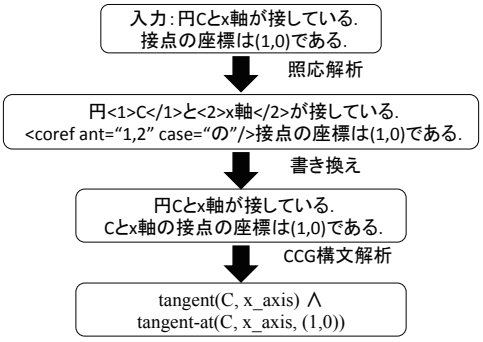

図 3: 言語処理部の解析の流れ

平面図形，Point は点を表し，1 つ目の例は「平面図形が点を通る」ということを意味している．

用語タクソノミーの一例を図 5 に示す．これは，「Shape」の下位概念に「有界な領域」や「直線」,「曲線」があることを示し，さらに「有界な領域」の下位概念として「円」や「多角形」があることを示している．用語タクソノミーは先行詞同定のステップで，先行詞タイプに適合するものを検出する際に用いる．使用している用語タクソノミーのノード数は，112 である．

3.4　照応表現別のシステム処理

上述のように数学問題テキストにおける照応表現には，様々な種類がある．そのうち，現在システムが対応しているのは連体詞形態指示詞，名詞形態指示詞，ゼロ代名詞である．以下では，照応表現のタイプ毎に照応の検出，先行詞タイプの決定および先行詞の同定に整理して，システムの詳細を述べる．

3.4.1　連体詞形態指示詞（単数）

問題中の連体詞形態指示詞「この」「その」を正規表現により抽出する．多くの場合,「その関数」のように連体詞形態指示詞の直後に名詞が表れるため，その名詞から先行詞タイプを決定する．そして，先行文脈中の先行詞候補のうち，照応表現に最も近いものを先行詞とする．

(Shape が)(Point を) 通る	(数列の) 初項
(Vector と)(Vector が) 直交する	(球 | 円の) 半径
(Shape と)(Shape で) 囲まれた	(方程式の) 重解

図 4: 格フレーム辞書 (左：動詞，右：不飽和名詞)

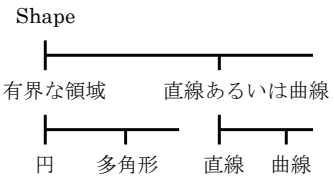

図 5: 用語タクソノミーの一部

3.4.2 連体詞形態指示詞（複数）

連体詞形態指示詞（単数）の場合と同様に，連体詞形態指示詞を抽出し，直後の名詞から先行詞タイプを決定する．先行詞同定の処理では,「その 2 本の直線」のように指示対象の数が明示されている場合と「それらの」のように明示されていない場合で処理が異なる．

指示対象の数が明示されている場合は，照応表現に近いものから順に先行文脈中の先行詞候補をその数だけ先行詞とする．数が明示されていない場合は，まず照応表現が含まれている文の先行文脈を調べ，先行詞タイプと適合する先行詞候補をすべて先行詞とする．適合する先行詞候補がない場合は，もう 1 つ前の文に対し同様の処理を行う．それでも先行詞が見つからない場合には照応解析処理を行わない．

3.4.3 名詞形態指示詞（単数）

問題中の名詞形態指示詞を正規表現により抽出する．次に，格フレーム辞書を用いて，指示詞の直後の動詞から先行詞タイプを推定する．例えば,「それと円が交わる」という問題の場合，格フレーム辞書より「それ」が「Shape」のタイプであると推定できる．先行詞の同定は，連体詞形態指示詞（単数）と同様に行う．

3.4.4 名詞形態指示詞（複数）

名詞形態指示詞（単数）の場合と同様に，名詞形態指示詞を抽出し，先行詞タイプを決定する．先行詞の同定は，連体詞形態指示詞（複数）と同様である．

3.4.5 不飽和名詞の項となるゼロ代名詞

まず，不飽和名詞の辞書に記載されている単語を問題文から抽出する．そして，以下のいずれにも当てはまらないとき，不飽和名詞の直前にゼロ代名詞があると判定する．

1. 項の意味タイプと合致する名詞句が不飽和名詞に係る場合　(例：「円の半径」)
2. 不飽和名詞の直前の単語が「ノ形」の状詞 [2] で，その単語に不飽和名詞の項と同じ意味タイプの名詞句が係る場合　(例：「 円の最大の半径」)
3. 不飽和名詞が連体修飾節の中にあり，連体修飾節の係り先が不飽和名詞の項の意味タイプと一致する場合　(例：「半径の長さが 1 より長い円」)
4. 直後に数式がくる場合 (一部の不飽和名詞のみ)
   (例：内積 $\vec{a}\cdot\vec{b}$)

1，2 は不飽和名詞の項が埋まっていると考えられるため，3 は文法的に指示対象が決定するため，4 は直後の数式から指示対象が決定するため除外した．4 では,「内積 $\vec{a}\cdot\vec{b}$」のように数式から「$\vec{a}$ と $\vec{b}$ の内積」と分かる場合と「半径 r」のようにこれだけでは何の半径か分からない場合があるため，一部の不飽和名詞のみ除外の対象とした．

次に格フレーム辞書を用いて先行詞タイプを決定する．例えば，不飽和名詞が「初項」の場合は，そのノ格の項のタイプ「数列」を先行詞タイプとする．先行詞の同定は，連体詞形態指示詞（単数）と同様である．

4 システムの評価

照応表現の検出と先行詞同定の評価を行った．照応表現の検出では，指示詞を含む照応表現は，正規表現で検出できるため評価の対象とせず，ゼロ代名詞の検出のみを評価の対象とした．また，先行詞同定の評価では，正しく検出できた照応表現の中で先行詞も正しく同定できた数を照応表現別に調査した．

実験には，1990 年度〜2014 年度の国公立および私立大学入試の二次試験『数学』の問題を使用した．ゼロ代名詞の検出で使用したデータセットは，照応表現の出現数調査に用いた 100 題と同一のものである．先行詞同定の評価は，連体詞形態指示詞，名詞形態指示詞では正規表現で照応表現を抽出した 20 箇所，ゼロ代名詞ではゼロ代名詞検出の評価に使用した問題の中で，正しくゼロ代名詞を検出できた 20 箇所を使用した．

ゼロ代名詞検出の評価結果を表 3 に示す．本来存在しないゼロ代名詞を誤って検出した箇所が 17 箇所あり，比較的多いことが分かる．先行詞同定の評価結果を表 4 に示す．他の照応表現と比較して，連体詞形態指示詞 (複数) の精度が低いことが分かる．

表 3: ゼロ代名詞の検出

Precision	Recall	F1
62%(28/45)	74%(28/38)	67%

表 4: 先行詞の同定

	正解数	正解率 [%]
連体指示詞 (単数)	16	80
連体指示詞 (複数)	11	55
名詞指示詞 (単数)	18	90
名詞指示詞 (複数)	18	90
ゼロ代名詞	16	80

5 結果の分析

4 章で行ったゼロ代名詞検出および先行詞同定の評価における誤り例の分析結果について述べる.

5.1 ゼロ代名詞の検出

存在するゼロ代名詞を検出できなかった箇所が 10 箇所あった. その原因の多くは, 格フレーム辞書のエントリ不足であった. 例えば以下のような例があった.

> 曲線 C_1 の接線のなかで, 点 (0,0) を通る 接線 の本数を求めよ.

下線部の「接線」は, 何の接線かという情報を必要とするが, 辞書に「接線」という単語が登録されていなかったため, ゼロ代名詞の検出が行われなかった. 今後, 漏れがないよう辞書を拡充していく必要がある.

一方, 本来存在しないゼロ代名詞を誤って検出した原因として, 数式の意味タイプを付与できていないことがある. 誤り例として, 以下のようなものがあった.

> 行列 A で表される 1 次変換を f とする. 原点以外の 1 点を $P_1(a_1, b_1)$ とし, P_1 が f により移された点を $P_2(a_2, b_2)$ とする. 原点以外の点 P に対して, 点 P を中心として原点を通る円を C_P で表す. C_{P_1} と C_{P_2} との 交点 の座標を求めよ.
> (1996 年 広島大学 一部改変)

この問題では, C_{P_1} と C_{P_2} に数式の意味タイプを付与できなかったため,「交点」の項がゼロ代名詞化されていると判定された.

この解決方法として, 数式の意味タイプ付与のアルゴリズムを改良することが考えられる. C_{P_1} と C_{P_2} に「円」のタイプを正しく付与できれば,「交点」の項が埋まっていると判断され, 照応解析が行われなくなる.

また, 不飽和名詞の項が埋まっていると判定するルールを弱める方法も考えられる. 現在は, ノ格の名詞句が不飽和名詞に係っており, かつ, その意味タイプが格フレーム辞書の項タイプと合致する場合にのみ, 項が埋まっていると判定している. これを弱め,「ノ形」の状詞以外のノ格が不飽和名詞の直前に存在するときは, それが項であると判定すれば, 上記の問題では,「交点」の直前に「C_{P_2} との」というノ格の文節があるので, 照応解析が行われなくなる.

5.2 先行詞の同定

連体詞形態指示詞 (複数) の先行詞同定の誤り例として, 指示対象の数と先行詞の数が一致しないケースが多かった. 以下にその問題例を挙げる.

> 円<1>C</1>に内接する<2>円</2>が 2 つ存在する. <coref ant="1,2">この 2 つの円</coref> をそれぞれ円 D,E とする.

この問題では,「この 2 つの円」という句から「円」をタイプとする指示対象を 2 つ同定する必要がある. しかし,「内接する円」が 2 つの円であるという情報を読み取れていないため, 上記のように先行詞として「C」および「(内接する) 円」の 2 つを出力し, 結果として指示対象が 3 つになった. この解決には, 問題文のより深い解析が必要になる. そこで, 現在の処理の順序を変更し, 先行詞の同定は, CCG 構文解析による問題文の分析の後で, 形式的な意味表現の上での処理として行うことが考えられる.

6 結論

本研究では, 数学問題テキストに対する照応・共参照解析システムを実装し, 評価を行った. 結果として, ゼロ代名詞検出や連体詞形態指示詞 (複数) の先行詞同定に課題があることが分かった. これらの課題に対処するため, 数式の意味タイプ付与のアルゴリズムの改良や, 形式表現上での照応解析を考える必要がある.

参考文献

[1] 新井紀子, 松崎拓也. ロボットは東大に入れるか?—国立情報学研究所「人工頭脳」プロジェクト—. 人工知能学会誌, Vol. 27, No. 5, pp. 463–469, 2012.

[2] 戸次大介. 日本語文法の形式理論. くろしお出版, 2010.