数学問題テキストにおけるさまざまな照応現象の解析

伊藤 巧　　松崎 拓也　　佐藤 理史
名古屋大学大学院 工学研究科 電子情報システム専攻

1 はじめに

2011 年に，人工知能技術で大学入試問題を解くことに挑戦する「ロボットは東大に入れるか」というプロジェクトが開始された [1]．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・共参照解析のために我々が昨年までに開発したシステム [4] を拡張し，さまざまな照応現象に対応したシステムについて述べる．

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 節でシステムの構成を示す．さらに，4 節でそのシステムの基本処理フローから外れる特殊な処理について説明し，5 節で構文解析との役割分担について述べる．そして，6 節でシステムの評価とその考察を行う．

2 照応表現の種類・出現数

これまで数学問題で観察した照応表現およびその出現数を表 1 にまとめる．なお，出現数の調査には 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 題を用いた．

連体詞形態指示詞や名詞形態指示詞，ゼロ代名詞は一般の文と同様に出現する．また，普通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…

上の例では，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．しかし，普通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」のように普通名詞は記号を伴って出現することが多いことである．

また，条件指示詞は，問題文の今までの条件を全て引き継ぐことを示すものと，ある特別な命題を指しているものの 2 種類がある．それぞれの例を以下に示す．

表 1: 照応表現の種類・例・出現数

種類	例	出現数	
		センター	二次試験
連体詞形態指示詞（単数）	その，この	26	12
連体詞形態指示詞（複数）	それらの，これらの	2	2
名詞形態指示詞（単数）	それ，これ	2	1
名詞形態指示詞（複数）	それら，これら	1	1
不飽和名詞の項となるゼロ代名詞	(φの) 斜辺，(φの) 半径	121	61
普通名詞	方程式	1	5
条件指示詞	そのとき，このとき	57	33
その他	それぞれ，もの	29	16

> 1. $AB = 1$，$BC = 2$，$CA = 3$ の $\triangle ABC$ がある．このとき，$\triangle ABC$ の面積を求めよ．
> 2. $f(x)$ の最大値と そのとき の x の値を求めよ．

1 の問題での下線部「このとき」は，$\triangle ABC$ に関して与えられた条件を引き継いでいるものであり，省略しても問題文の意味は変わらない．一方，2 の問題での下線部「そのとき」は「$f(x)$ が最大になるとき」という特別な条件を指しており，省略することができない．

3 システムの構成

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．

また，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，先行詞の数および意味タイプを推定する．

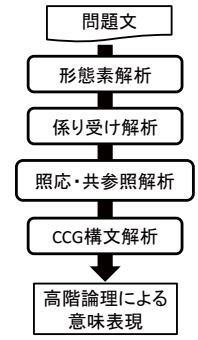

図 1: 言語処理部の構成

図 2: 照応解析システムの構成

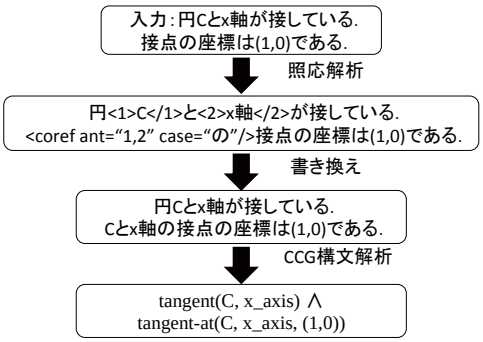

図 3: 言語処理部の解析の流れ

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，照応詞に最も近いものを先行詞として同定する．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．連体詞形態指示詞，名詞形態指示詞，ゼロ代名詞は，上述した処理フローで照応解析を行っている．

4 基本処理フローから外れる処理

2 節で述べた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．本節では，図 2 の処理フローから外れる「こと（命題・条件）」を指す照応表現および間接照応に対する処理について述べる．数学問題文では「こと」を指す照応表現として,「そのとき」,「(1) の場合」,「次の条件」などが出現する．日本語テキストに対して照応・共参照タグを付与したデータとしては NAIST テキストコーパス [5] があるが，同コーパスでは「こと」を指す照応表現はアノテーションの対象とはされていない．また間接照応とは，先行詞が文中に明示されない照応現象のことである．数学問題中では操作した結果が照応表現の指示対象となる場合があり，この現象は料理のレシピで表れる照応現象 [3] と類似している．例えば「肉と野菜と調味料をよく混ぜ，それを炒める」といった文の場合,「それ」の指示対象は「肉と野菜と調味料をよく混ぜたもの」であるが，文章中に明示されていない．

4.1 「こと」を指す照応表現

2 節で述べたように「そのとき」には照応解析が必要なものとそうでないものがある．現在のシステムでは,「最大」「最小」について述べている以下のような「そのとき」を解析対象としている．

1. $f(x)$ の最大値と そのとき の x の値を求めよ．
2. $f(x)$ が最大になるときの x の値と そのとき の $f(x)$ の値を求めよ．

上記のような表現を正規表現で抽出し，以下のようにタグ付けを行い，照応解析を行った．下記のタグは<coref>，</coref>で囲まれた部分を<alt>，</alt>で囲まれた部分で置き換えることを表す．

1. $f(x)$ の最大値と<coref>そのとき<alt>$f(x)$ が最大になるとき</alt></coref>の x の値を求めよ．
2. $f(x)$ が最大になるときの x の値と<coref>そのときの $f(x)$ の値<alt>$f(x)$ の最大値</alt></coref>を求めよ．

また「(1) の場合」の照応解析では，照応表現の検出をした後，(1) の問題文中で指示内容である命題・条件を同定する．指示内容の同定では「ような」「ための」「とき」「(と|に) なる」「とする」を検索し，それらに係る文節を根とする句を指示内容とする．出力例を以下に示す．なお,「<1>」のような番号のみのタグは先行詞あるいは指示内容を表し,「<coref ant="1" case="とき">」のようなタグは照応詞に用いる．ant 属性は照応する先行詞の番号を値とし，case 属性は書き換え処理の際に補うべき格助詞などを表す．

(1) <1>$f(x)$ の最大値が 3 になる</1>ように a の値を定めよ．
(2) <coref ant="1" case="とき">(1) の場合</coref>に $f(x)$ の最小値を求めよ．

「次の条件」の照応解析は，後方文脈から指示内容を探す．具体的には,「次の条件」が含まれる文の直後の文ないし数式を指示内容として同定する．ただし直後の文が「ただし」で始まる場合はその後の文ないし数式を指示内容とする．出力例を以下に示す．

実数 a を係数とする関数 $f(x)=ax^2+4$ について，<coref ant="1">次の条件</coref>を考える．<1>$0 \leq x \leq 2$で $f(x) \geq (x+1)^2$ が成立する．</1>

4.2 間接照応

操作の結果が指示対象となる間接照応の例を以下に示す.

> $y{=}x^2$ のグラフを x 軸方向に 2 だけ平行移動し，それを x 軸に関して対称移動したグラフを求めよ.

下線部「それ」の指示対象は「$y = x^2$ のグラフを x 軸方向に 2 だけ平行移動したグラフ」であるが，問題文中にはこれを明示する名詞句は登場しない．そのため，現在のシステムでは名詞句や数式に加え,「平行移動」や「かけあわせる」といった操作動詞を先行詞候補とする．上記の問題では下線部「それ」に最も近い先行詞候補「平行移動」を先行詞として同定する.

5 構文解析との役割分担

本システムでは，ゼロ照応のうち，ゼロ代名詞が不飽和名詞の項となる場合のみを解析の対象としている．そのためゼロ照応解析では，まず辞書に記載されている不飽和名詞を問題文から抽出する．その後，抽出した単語にノ格などでマークされた，項となる名詞句が存在せず（即ちゼロ代名詞が存在し）かつゼロ代名詞の指示対象が文法的（構文的）には決まっていない場合，照応解析が必要と判定する．照応解析が真に必要か否かに関わらず，抽出した不飽和名詞全てに対してゼロ代名詞の項の存在を仮定して，その指示対象の同定を行う手法も考えられる．例えば，CCG 構文解析時に照応解析結果を無視してもよいことを表す属性 skip を導入し，以下のように解析する方法が考えられる.

> <1>$f(x) = 0$</1>は<coref ant="1" case="の" skip="OK"/>実数解をもつ.

この場合，CCG 構文解析中に実質的には「$f(x) = 0$ は $f(x) = 0$ の実数解をもつ」「$f(x) = 0$ は実数解をもつ」の 2 通りの入力について解析することになるが，数学問題では一文中に多数の不飽和名詞が現れることも珍しくない．そのような場合，上記の方法では組み合わせ的に計算コストが大きくなる．そこで，本システムでは CCG 構文解析の処理の一部を照応解析のステップに組み込み，必要な部分のみ照応解析を行っている．本節では，その処理について述べる.

5.1 構文の適用

文法的に不飽和名詞の項が埋まると考えられる文構造の例を図 4 に示す．図 4 の 1 の下線部「実数解」は

> 1. NP は [不飽和名詞] をもつ
> 例：$f(x) = 0$ は 実数解 をもつ.
> 2. [点] から [図形] へ [不飽和名詞] を引く
> 例：点 P から円 C へ 接線 を引く
> 3. 不飽和名詞が連体修飾節の中にあり，その係り先が不飽和名詞の項の意味タイプと一致
> 例：半径 が 3 である円を C とする.

図 4: 文法的に不飽和名詞の項が埋まる文構造

「$f(x) = 0$」の実数解であること，2 の下線部「接線」は「円 C」の接線であること，3 の下線部「半径」は「円」の半径であることが文法的に決まる．したがって，パターンマッチングを行い，図 4 のような構文に当てはまる場合はゼロ照応解析を行わない.

5.2 項の数判定

不飽和名詞の中には，意味的に複数の項を要求するものが存在する．例えば「交点」は本来「〜と〜の交点」のように 2 つの項を必要とする．そのような不飽和名詞の項が埋まっているか判断するため，不飽和名詞に係っている項の数を判定する．項の数は，不飽和名詞にノ格がいくつ係るかによって判定する．ただし,「〜と〜の」といったようなノ格が係る場合は複数の項が埋まっていると判定し，一方「最大の」といったような「ノ形」の状詞 [2] が係る場合はその単語は項ではないと判定する．項の数の判定例を以下で述べる.

> 1. 円と直線が 1 点で交わり，交点 を A とする.
> 2. xy 平面上にある直線があり，$y = x$ との 交点 を A とする.
> 3. 円と直線の 交点 を A とする.

それぞれの下線部「交点」について考えると，1 は項が埋まっていない，2 は項が 1 つ，3 は項が 2 つ埋まっていると判定される．したがって，1 は先行詞を 2 つ，2 は先行詞を 1 つ同定する照応解析を行い，3 は既に項が埋まっているので照応解析を行わない.

5.3 不飽和名詞を項にとる不飽和名詞

そのほかに下流の CCG 構文解析の処理の一部を照応解析処理で先取りしている例として，以下のような不飽和名詞を項にとる不飽和名詞の例が挙げられる.

> 線分ABとCDは垂直で，長さの比を1 : 2とする.

上記の問題では，下線部「長さの比」で不飽和名詞「比」が不飽和名詞「長さ」を項にとっている．本来

表 2: ゼロ代名詞の検出

	Precision	Recall	F1
マーク模試	74% (32/43)	89% (32/36)	81%
東大プレ	88% (7/8)	100% (7/7)	93%

表 3: 先行詞の同定

	マーク模試	東大模試
正解率	85%(46/54)	60%(9/15)

「比」は項を 2 つ必要とし,「長さ」は項を 1 つ必要とする．5.2 節で述べた項の数判定を行いゼロ照応解析を行うと,「(ϕ_1 の) 長さ (と ϕ_2) の比」における ϕ_1 と ϕ_2 の先行詞を同定することになるが，これは誤りである．したがって,「比」の必要とする項の数 2 つを「長さ」に伝播させ,「長さ」の先行詞タイプに合致する「線分」を 2 つ先行詞として同定する．本節の冒頭で述べたような，仮に全ての不飽和名詞の項が埋まっていないと考えゼロ照応解析処理を行う手法では，このような表現に対応することは難しいと考えられる．

6 評価および考察

システム全体の評価として照応表現の検出と先行詞同定の精度を評価した．また，基本処理フローから外れる処理の評価として目視でその処理を確認した．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対象とせず，ゼロ代名詞の検出のみ評価対象とした．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調査した．また基本処理フローから外れる処理の評価は，現象の出現回数自体が少なく正規表現での検出が難しいものもあるため目視でその処理を確認した．

実験には，開発データのベネッセマーク模試『数学 IA』『数学 IIB』各 7 回分と評価データの代ゼミ東大模試 10 回分を使用した．なお確率，データ・統計の問題は除いてある．結果を表 2，表 3 に示す．表 2 の Precision が低いことから，不要なゼロ代名詞検出をしていることが分かる．これは係り受け解析ミスなどにより文法的に指示対象が決まる場合の判定誤りが原因であり，誤りの 12 問中 8 問を占める．また，表 3 より開発データのマーク模試では先行詞同定の正解率が高いが，評価データの東大模試では正解率が下がっていることが分かる．東大模試の誤りの 6 問中，後方参照の先行詞誤りが 3 問であるため，後方参照の先行詞同定アルゴリズムに改善の余地があると言える．

基本処理フローから外れる処理を目視で確認したところ，図 5 のようなものが確認された．なお，この実

1. $a > 0$ のとき，面積 $S(a)$ の最小値を求めよ．また，<coref>そのとき<alt>$S(a)$ が最小になるとき</alt></coref>の a の値を求めよ．
2. (1) <1> α, β が $\alpha \leq 1 \leq \beta$ をみたす</1>ようにk の値の範囲を定めよ．
   (2) <coref ant="1" case="とき">(1) の場合</coref>に $f(x)$ の最小値がとりうる値の範囲を求めよ．

図 5: 基本処理フローから外れる処理の実行例

験には先ほどの実験データに開発データである国公立の二次試験『数学』100 題を追加して調査した．図 5 の 1 では下線部「そのとき」の内容が問題文中に明示されていないが，正しく照応解析できている．このタイプは正規表現で照応解析を行っているので，照応解析を行った 6 箇所すべて正しく解析できていた．また，2 でも下線部「ように」に係る文節を根とする句を抽出することにより，正しく照応解析できている．なお，間接照応は今回の実験データでは出現しなかった．

7 結論

数学問題テキストに対する照応・共参照解析システムを改良し，評価を行った．より多様な照応・共参照タイプへの照応解析を実装することができたが，普通名詞による照応解析が未実装など拡張が十分であると言えない．今後は，普通名詞による照応解析や解析精度をより向上させることが課題だと考えられる．

謝辞

データを提供していただいた「ロボットは東大に入れるか」プロジェクトおよびベネッセコーポレーション，代々木ゼミナールの皆様に深く感謝いたします．

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.

[3] Hirokuni Maeta, Tetsuro Sasada, and Shinsuke Mori. A framework for procedural text understanding. In *Proc. IWPT*, pp. 50–60, 2015.

[4] 伊藤巧, 松崎拓也, 佐藤理史. 数学問題テキストに対する照応・共参照解析. 言語処理学会第 22 回年次大会, 2016.

[5] 飯田龍, 小町守, 井之上直也, 乾健太郎, 松本裕治. 述語項構造と照応関係のアノテーション: Naist テキストコーパス構築の経験から. 自然言語処理, Vol. 17, No. 2, pp. 25–50, 2010.