著者・読者表現および外界ゼロ照応を考慮したゼロ照応解析

萩行 正嗣　　河原 大輔　　黒橋 禎夫
京都大学大学院 情報学研究科
{hangyo, dk, kuro}@nlp.ist.i.kyoto-u.ac.jp

1　はじめに

ゼロ照応解析は近年，述語項構造解析の一部として盛んに研究されつつある．ゼロ照応とは用言の項が省略される現象である．

(1)パスタが好きで、毎日 (ϕ ガ)(ϕ ヲ)食べています。

例えば，例 (1) の「食べています」では，ガ格とヲ格の項が省略されている．ここで，省略されたヲ格の項 (ゼロ代名詞) は前方で言及されている「パスタ」を照応しており (**文章内ゼロ照応**)，省略されたガ格の項は文章中では言及されていない文章の著者を照応している (**外界ゼロ照応**)[1]．外界ゼロ照応で照応されるのは例 (1) のような文章の著者や読者，不特定の人や物などがある[2]．

従来，日本語ゼロ照応解析の研究は，ゼロ照応関係を付与した新聞記事コーパス [1, 2] を主な対象として行なわれてきた．新聞記事は事件の内容などを客観的に伝えることが目的であり，著者や読者が談話構造中に登場することはほとんどない．一方，近年情報伝達の社会基盤となりつつある Web においては，著者や読者が談話構造中に登場することが多い．例えば，Blog や企業の宣伝ページでは著者自身の出来事や企業自身の活動内容が述べられており，通販ページなどでは読者に対して商品を買ってくれるような働きかけをする．このため，Web に対するゼロ照応解析では文書の著者・読者が重要な役割を持つ．また，著者や読者は，省略されやすい，敬語やモダリティなど省略を推定するための手掛りが豊富に存在する，などの特徴を持つため，談話中の著者や読者を明示的に扱うことは照応先同定で重要である．本研究では著者・読者に着目したゼロ照応解析システムを提案する．

[1]以降の例では，ゼロ代名詞の照応先を埋めた形で「パスタが好きで、毎日 ([著者] ガ)(パスタヲ) 食べています。」のように記述する．[著者] は文章内で言及されていない文章の著者を示す．

[2]一般に外界照応と呼ばれる現象には，現場文脈指示と呼ばれる発話現場の物体を指示するものも含まれる．本研究ではこのようなテキストの情報のみから照応先を推測できない外界照応は扱わない．

2　外界ゼロ照応

Web では著者・読者が談話中に多く登場するため，必然的に著者・読者に関するゼロ照応が多く発生し，その中には外界ゼロ照応も多く含まれる．Web コーパス [3] ではゼロ照応関係の約半数が外界ゼロ照応である．このため，Web テキストに対するゼロ照応解析では，特に外界ゼロ照応を扱うことが重要となる．

従来研究では，外界ゼロ照応をゼロ代名詞自体が出現しないものとして扱うことが多かった．例えば例 (1) の「食べる」において必須的なガ格が項をとらないもの扱っていた．しかし，外界ゼロ照応とゼロ代名詞自体を取らないもの (例 (1) のニ格など) は本質的に別の現象であり，これらを同一に扱うことは機械学習によるゼロ代名詞検出に悪影響を与えていた．

本研究では，ゼロ代名詞の照応先候補として [著者] や [読者] などの文章中に出現しない談話要素を設定する．外界ゼロ照応を扱うことにより，照応先が文章内にない場合でも，用言のある格がゼロ代名詞を項に持つという現象を扱うことができる．これにより，格フレームなどの用言が項を取る格の知識とゼロ代名詞の出現が一致するようになり，機械学習によるゼロ代名詞検出の精度向上を期待することができる．また，[著者] や [読者] などを区別して扱うことで，敬語表現などにおけるこれらの振る舞いの違いを表現することができる．

3　著者・読者表現

著者や読者は前述のような外界ゼロ照応の照応先だけでなく，文章内に言及されることも多い．

(2)僕 $_{\text{著者}}$ は京都に (僕ガ)行こう と思っています。
皆さん $_{\text{読者}}$ はどこに行きたいか (皆さんガ)(僕ニ)教えてください。

例 (2) では，文章中に言及されている「僕」がこの文章の著者であり，「皆さん」が読者である．本研究ではこのような文章中で言及される著者や読者を**著者表**

現，**読者表現**と呼び，これらを明示的に扱うことでゼロ照応解析精度を向上させる．

著者や読者は人称代名詞だけでなく固有表現や役職など様々な表現で言及される．本研究では人称代名詞に限らず，著者・読者を指す表現を著者・読者表現として扱うこととする．著者・読者表現は様々な表現で言及されるため，どの表現が著者・読者表現であるかを表層的な表記のみから判断することは困難である．本研究ではランキング学習 [4] により，文章中の著者・読者表現の同定を行なう (詳細は 5 節で述べる)．

文章中に出現する著者・読者表現が照応先となることを推定する際には通常の文章中の表現に利用する手掛りと著者・読者特有の手掛りの両方が利用できる．上述の例 (2) の 1 文目では「僕」が文頭で助詞「は」を伴ない,「行こう」を越えて「思っています」に係っていることから「行こう」のガ格の項は「僕」であると推測される．このようなゼロ代名詞との表層的な位置関係などは文章中の表現のみが持つゼロ照応解析での手掛りと言える．一方，2 文目の「教えてください」では，依頼表現であることからガ格の項が読者表現である「皆さん」であり，ニ格の項が著者表現である「僕」であると推測できる．このような依頼や敬語などに関する手掛かりは著者・読者特有の手掛りと言える．また，著者・読者特有の手掛りは外界ゼロ照応における著者・読者においても同様に利用できる．そこで本研究では，ゼロ照応解析において著者・読者表現は文章内ゼロ照応および外界ゼロ照応両方の特徴を持つものとして扱う (詳細は 4.1 節で述べる)．

4 提案手法

本研究では，ゼロ照応解析を用言単位の述語項構造解析の一部として扱い，以下の手順で解析する．

1. 共参照解析を行ないテキスト中に出現した談話要素を認識する．
2. 著者・読者表現の推定を行い，どの談話要素が著者・読者表現にあたるのかを推定する．(詳細は 5 節で説明する)
3. 推定された著者・読者表現から仮想的な談話要素を設定する．
4. 各用言について以下の手順で述語項構造を決定する．
   (a) 以下の手順で解析対象用言がとりえる述語項構造 (格フレームと談話要素の対応付け) の組み合わせを列挙する．
      i. 解析対象用言の格フレームを 1 つ選ぶ．
      ii. 解析対象用言と係り受け関係にある語と格スロットの対応付けを行う．
      iii. 対応付けられなかったガ格，ヲ格，ニ格，ガ 2 格の格スロットと，談話要素の対応付けを行う．
   (b) 学習されたランキングモデルにより述語項構造候補にスコア付けをし，最もスコアが高いものを述語項構造として出力する．

本研究では，ゼロ代名詞の照応先を談話要素という単位で扱う．談話要素とは文中の表現のうち共参照関係にあるものを一まとめにしたものである．本研究では，照応先として談話要素に加え，外界ゼロ照応に対応する [著者]，[読者]，[不特定-人]，[不特定-その他] を設定する．ただし，著者・読者表現が出現する場合には，[著者]，[読者] は設定しない．各述語項構造は格フレーム (cf) とその格フレームの格スロットとその照応先の対応付け (a) として表現される．本研究では，Web69 億文から Kawahara らの手法 [5] で自動構築された格フレームを用いる．手順 (4a) で列挙される述語項構造候補の例を図 1 に示す．

本研究では素性の重みの学習にランキング学習を利用する．学習に利用する順位データは各用言に対して以下のようにして作成したものを統合したものとする．

1. 述語項構造候補 (cf, a) のうち，a がコーパスにおける対応付け a^* と同じ述語項構造 (cf, a^*) に対して，確率的ゼロ照応解析スコア [6] を計算する
2. 計算されたスコアが最も高い (cf, a^*) を正解とし，それ以外の (cf, a^*) を削除する
3. (cf, a^*) だけが他の述語項構造候補 (cf, a) より高い順位とする順位データを作成する

図 1 の場合，計算されたゼロ照応解析スコアが【1 − 1】>【2 − 1】であったとすると，作成されるデータは【1 − 1】>【1 − 2】= ··· =【2 − 2】= ··· となる

4.1 素性による述語項構造の表現

入力テキスト t の解析対象用言 p に格フレーム cf を割り当て，その格フレームの格スロットと談話要素の対応付けを a とした述語項構造を表現する素性ベクトルを $\phi(cf, a, p, t)$ とする．$\phi(cf, a, p, t)$ は以下のように表現される[3]．

$$\begin{aligned}\phi(cf, a) = (&\phi_{overt\text{-}PAS}(cf, a_{overt}),\\ &\phi_{case}(cf, \text{ガ} \leftarrow e_{\text{ガ}}), \phi_{case}(cf, \text{ヲ} \leftarrow e_{\text{ヲ}}),\\ &\phi_{case}(cf, \text{ニ} \leftarrow e_{\text{ニ}}), \phi_{case}(cf, \text{ガ}\,2 \leftarrow e_{\text{ガ}\,2}))\end{aligned}$$

ここで $\phi_{overt\text{-}PAS}(cf, a_{overt})$ は直接係り受けがある述語項構造に関する素性ベクトルであり，$\phi_{case}(cf, c \leftarrow e)$ は格 c に談話要素 e が割り当てられることに関する素性ベクトルである．各格に対応する素性ベクトル

[3] 以降，p,t については適宜省略する．

僕 著者 は出町柳にあるラーメン屋によく行きます。すごく美味しいので、今日はその店を 紹介します。

談話要素

(a){ 僕 }，(b){ 出町柳 }，(c){ ラーメン屋，その店 }，(d){ 今日 }

列挙される述語項構造候補

【1-1】格フレーム:『紹介する (1)』，{ ガ:(a) 僕，ヲ:(c) ラーメン屋，ニ:[読者]，ガ 2:×，時間:(d) 今日 }
【1-2】格フレーム:『紹介する (1)』，{ ガ:(a) 僕，ヲ:(c) ラーメン屋，ニ:×，ガ 2:×，時間:(d) 今日 }
【1-3】格フレーム:『紹介する (1)』，{ ガ:[読者]，ヲ:(c) ラーメン屋，ニ:×，ガ 2:×，時間:(d) 今日 }
【1-4】格フレーム:『紹介する (1)』，{ ガ:[不特定-人]，ヲ:(c) ラーメン屋，ニ:×，ガ 2:×，時間:(d) 今日 }
⋮
【2-1】格フレーム:『紹介する (2)』，{ ガ:(a) 僕，ヲ:(c) ラーメン屋，ニ:[読者]，ガ 2:×，時間:(d) 今日 }
【2-2】格フレーム:『紹介する (2)』，{ ガ:(a) 僕，ヲ:(c) ラーメン屋，ニ:×，ガ 2:×，時間:(d) 今日 }
⋮

図 1: 談話要素と述語項構造の例

$\phi_{case}(cf, c \leftarrow e)$ は格 c に談話要素 e が対応付けられた場合の素性ベクトル $\phi_A(cf, c \leftarrow e)$ と何も対応付けられなかった場合の素性ベクトル $\phi_{NA}(cf, c \leftarrow \times)$ からなる．$\phi_{case}(cf, c \leftarrow e)$ は格 c がゼロ照応として対応付けられた場合にのみ考慮し，係り受け関係にある談話要素に対応付けられた場合には 0 ベクトルとする．

$\phi_{case}(cf, c \leftarrow e)$ は $\phi_{discourse}(cf, c \leftarrow e)$，$\phi_{[著者]}(cf, c \leftarrow e)$，$\phi_{[読者]}(cf, c \leftarrow e)$，$\phi_{[不特定:人]}(cf, c \leftarrow e)$，$\phi_{[不特定:その他]}(cf, c \leftarrow e)$，$\phi_{max}(cf, c \leftarrow e)$ からなる．$\phi_{discourse}$ は e が文章中の談話要素の場合のみに発火し，$\phi_{[著者]}$，$\phi_{[読者]}$ は e が [著者]・[読者] または著者・読者表現の場合のみ発火する．$\phi_{[不特定:人]}$，$\phi_{[不特定:その他]}$ は e が [不特定:人]・[不特定:その他] の場合にのみ発火する．ϕ_{max} は対応する各素性の最大値とする．

素性をこのように表現することで，著者・読者表現では $\phi_{discourse}$ および $\phi_{[著者]}$，$\phi_{[読者]}$ が発火することとなり，表層的な手掛りと著者・読者としての手掛りの両方を利用できる．また，ϕ_{max} はゼロ照応全体に関わる現象を扱うこととなる．

$\phi_{overt\text{-}PAS}$ には確率的格解析モデル [7] から得られる表層の係り受けの確率を用いる．ϕ_A を構成する各素性ベクトル ($\phi_{discourse}$ など) には Sasano ら [8] の素性に加え，用言のモダリティや敬語の種類を利用した．ϕ_{NA} には Sasano らの素性に加え，cf の c が必須格であるか，直前格であるかを用いた．

5 著者・読者表現推定

ゼロ照応解析の前処理として，著者・読者表現の自動推定を行なう．著者・読者表現の自動推定は談話単位で行なう．談話要素の周辺の語彙統語パターンを素性とし，著者・読者表現が他の談話要素より上位になるようにランキングを学習する．また，著者・読者表現が文章中に登場しない場合に対応する仮想的な談話要素として，以下の 2 つを導入する．

著者・読者表現なし (省略) 談話構造中に著者・読者が表現するが著者・読者表現は出現しない場合に対応し，ゼロ外界照応として著者・読者が登場する場合に上位とする．文章全体の語彙統語パターンを文書ベクトルとして素性にする．

著者・読者表現なし (談話) 談話構造中に著者・読者が登場しない場合に対応し，ゼロ外界照応としても著者・読者が登場しない場合に上位とする．0 ベクトルにより表現される．

各文書ごとにランキングデータを生成し，それを統合したものを学習データとする．自動推定の際は最上位となった談話要素を著者・読者表現とし，「著者・読者表現なし」が最上位の場合には，著者・読者表現なしとする．

6 実験

6.1 実験設定

実験では DDLC[3] の 1000 記事を利用し，5 分割交差検定により評価を行なった．述語項構造解析および著者・読者表現推定以外の解析結果が原因となる解析誤りを除くため，形態素情報，係り受け情報，固有表現情報，共参照関係はコーパスに人手で付与された正しい情報を利用する．著者・読者推定および述語項構造のランキング学習には SVM^{rank}[4]を用いた．

6.2 著者・読者推定実験結果

DDLC に対して，5 節で述べた手法により著者表現および読者表現を推定した結果を表 1 と表 2 に示す．

[4] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

表 1: 著者表現推定結果

		システム		
		出力あり		出力なし
		正解	誤り	
コーパス	あり	138	9	124
	なし	-	38	691

表 2: 読者表現推定結果

		システム		
		出力あり		出力なし
		正解	誤り	
コーパス	あり	50	2	32
	なし	-	30	886

ここでコーパスの「あり」「なし」はコーパスに著者・読者表現が出現したかを表し，システムの「出力あり」「出力なし」はシステムが著者・読者表現ありと判断したかを表す．この結果より，著者・読者表現なしを含めた精度では，著者・読者共に 8 割以上の精度で推定できることが分かる．一方，再現率はあまり高くなく，著者表現，読者表現共に約 6 割しか推定できなかった．これはコーパス全体で著者・読者表現が出現記事より著者・読者表現のない記事の方が多く，学習時に著者・読者表現なしを優先するように学習してしまったためと考えられる．

6.3 ゼロ照応解析結果

DDLC に対してゼロ照応解析を行なった．ベースラインは照応先として外界の談話要素を設定せず，著者・読者表現の情報も利用しないモデルである．提案モデル (推定) は自動推定した著者・読者表現を利用したものである．提案モデル (正解) は著者・読者表現についてはコーパスの正解を与えたものである．表 3 はベースラインとの比較のために，文章内ゼロ照応のみで評価した結果であり，表 4 は外界ゼロ照応を含めた全てのゼロ照応で評価を行なった結果である．この結果から，外界ゼロ照応および著者・読者表現を考慮することで，外界を含めたゼロ照応全体だけでなく，文章内ゼロ照応においても精度が向上することが分かる．文章内ゼロ照応での評価において，提案モデルは推定した著者・読者表現を利用した場合において，ベースラインのモデルよりも適合率，再現率ともに向上していることが分かる．適合率が向上している原因としては，必須的な格を埋める際に，無理に文章内から選択せず外界の談話要素を選択することができること，敬語などの表現の際に著者表現や読者表現を照応先として選択できることが考えられる．再現率が向上する原因としては，ベースラインモデルでは学習の際に外界照応をゼロ代名詞なしと学習してしまうため，必須的な格でも必ずしも対応付ける必要がないと学習してしまうが，提案手法では必須的な格はなるべく埋めるように学習することが考えられる．正解の著者・読者表現を与えた場合には，再現率，適合率ともに大きく向上している．このことから著者・読者表現の推定精度を向上させることで，ゼロ照応解析の精度がより向上すると考えられる．

表 3: 文章内ゼロ照応解析結果

	再現率	適合率	F 値
ベースライン	0.270	0.370	0.312
提案モデル (推定)	0.298	0.447	0.357
提案モデル (正解)	0.411	0.536	0.465

表 4: 全ゼロ照応解析結果

	再現率	適合率	F 値
ベースライン	0.115	0.370	0.176
提案モデル (推定)	0.356	0.458	0.401
提案モデル (正解)	0.423	0.535	0.472

7 まとめ

本研究では，文書の著者・読者に着目したゼロ照応解析システムを提案した．実験により，ベースラインシステムより高い精度を達成することを示した．

参考文献

[1] Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. Annotating a japanese text corpus with predicate-argument and coreference relations. In *Proc of the Linguistic Annotation Workshop*, pp. 132–139, June 2007.

[2] Daisuke Kawahara, Sadao Kurohashi, and Koiti Hasida. Construction of a japanese relevance-tagged corpus. In *Proc. of LREC 2002*, May 2002.

[3] Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. Building a diverse document leads corpus annotated with semantic relations. In *Proc. of PACLIC 2012*, pp. 535–544, November 2012.

[4] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proc of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142, 2002.

[5] Daisuke Kawahara and Sadao Kurohashi. Case frame compilation from the web using high-performance computing. In *Proc. of LREC 2006*, pp. 67–73, 2006.

[6] Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. A fully-lexicalized probabilistic model for japanese zero anaphora resolution. In *Proc. of Coling 2008*, pp. 769–776, August 2008.

[7] Daisuke Kawahara and Sadao Kurohashi. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proc. of HLT-NAACL 2006*, pp. 176–183, June 2006.

[8] Ryohei Sasano and Sadao Kurohashi. A discriminative approach to japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proc. of IJCNLP2011*, pp. 758–766, November 2011.