照応, 省略, 共参照タグ付コーパスの構築

植田 禎子
日本システムアプリケーション
y-ueda@jsa.co.jp

荻野 孝野
日本システムアプリケーション
神戸大学
ogino@jsa.co.jp

飯田 龍
奈良先端科学技術大学院大学
ryu-i@is.naist.jp

乾 健太郎
奈良先端科学技術大学院大学
inui@is.naist.jp

奥村 学
東京工業大学
oku@pi.titech.ac.jp

1 はじめに

文章には多くの照応が含まれており, 自然言語処理においては, その解決が様々な分野で要求されている. たとえば, 機械翻訳, 文章要約, 日本語教育の分野が代表的なものである. こうした背景から, 本研究では照応解決の研究材料として照応, 省略, 共参照タグ付コーパスを作成した. これまで照応現象について, 照応タグ付コーパスを作成したプロジェクトには, MUC-6,7[1][2] や ACE[5], GNOME[3], GDA[4], 関ら [6], 河原ら [7] などがある.

本研究では, 新聞記事を対象に, 記事中のゼロ照応, 代名詞照応, 名詞句照応とその参照先を人手で同定し, タグ付けを行った. また複数記事間において名詞句照応のタグ付けを行った. 対象としたテキストは, テキスト自動要約の評価型プロジェクト TSC(Text Summarization Challenge)[8] で使用されたもので, すでに自動要約や談話構造 [9][10] に関するタグが付与されている. 同じテキストに多種のタグが付与されたことにより, 応用性の高いコーパスとなっている.

以下, 2 節では, コーパスの概要, 設計について報告する. ついで 3 節では, 作業結果の概要および作業上の問題点とその対処について, 最終節で展望について述べる.

2 コーパスの概要

コーパス作成の対象となった TSC データは, 毎日新聞記事データ (異なり 978 記事) である. TSC は過去 3 回開催され, 第 1 回目の TSC では, 単一文書要約タスク, TSC-2 では単一文書要約と複数文書要約タスク, TSC-3 では複数文書要約のタスクが行われた. 本研究はそのすべての記事に照応タグの付与を行った. 内訳は表 1 のとおりである[1].

表 1: TSC データの内訳

	TSC	トピック数	記事数
単文書要約	TSC-1	—	180
	TSC-2	—	60
複数文書要約	TSC-2	50	377
	TSC-3	65	390

作業はまず, 記事内でゼロ照応, 代名詞照応, 名詞句照応のタグ付けを行い, 複数文書要約のデータに対しては, その後トピックごとに記事間の名詞句照応のタグ付けを行った. 作業は形態素解析を行っていないテキストに対して行った.

2.1 参照先の指定

参照先はできる限り記事中に求めた. 後の文書間照応において, 他の記事にある ID 値を参照するため, 記事 ID と記事中固有の値を組み合わせて ID とした. 結果, ID 値は全データ内で唯一の値となっている[2].

(1) <ID id=”990423004_101”>香川大学<ID >では

記事中に参照先が存在しない場合は, あらかじめ設定した予約 ID 値を用いてタグ付けを行った. 予約 ID 値は表 2 の粒度で区別した. 参照先が 2 文以上にまたがっている (節を参照している) 場合も, テキスト中に ID 値を与えず, 予約 ID 値を用いた.

[1] TSC-2 の複数文書要約では, トピックにより記事に重なりがある. この表の値は延べ数である.

[2] 本稿中では, 便宜上 ID 値を適当な数に設定している場合もある.

表 2: 外界照応と節照応の予約 ID 値

カテゴリ	説明	予約 ID
外界一般	一人称・二人称以外すべて	1
外界一人称	書き手	2
外界二人称	直接引用の聞き手もしくは読み手	3
節	2 文以上にまたがった節	4

表 3: 代名詞照応

カテゴリ	タグ名	例
人称代名詞	代名詞	彼
指示代名詞	指示代名詞	それ
限定指示	限定指示	その
代行指示	代行指示	その
関係接頭辞	関係接頭辞	本大会, 同組織

2.2 文書内照応

記事中の照応関係は, 照応元が参照先の ID を ref 属性で参照することで示した.

参照先と照応詞の関係には, 指示照応 (reference anaphora) と語義照応 (sense anaphora) がある. 次の指示代名詞「それ」は, a. では指示照応, b. では語義照応である.

(2) a. 太郎は今話題の本を買った.
次郎は**それ**を借りた.

b. 太郎は今話題の本を買った.
次郎も**それ**を買った.

本研究ではこの両者について, ゼロ照応, 代名詞照応, 名詞句照応で方針を分けてタグ付けを行った.

ゼロ照応

述語において必須格の省略を作業者の直感で同定し, その格要素を補った. ゼロタグにおいて, 参照先の ID 値を ref 属性で参照し, case 属性で省略されている格を示した. 省略認定の対象にした格要素は, ガ, ヲ, ニ, カラ, トである.

(3) 香川で, 年次大会を<ゼロ ref=”1” case=”GA”>開く</ゼロ>

ゼロ照応の参照先は, 語義照応を優先し, 省略箇所と参照先の距離の近さに注目してタグ付けを行った. 具体的には, 参照先となり得る名詞句の中から, 省略箇所から記事の前方に向かってもっとも近いものを参照先として選択した. 省略箇所から前方に参照先がなかった場合は, 後方へ参照先を求めた. その場合も複数の候補がある場合は省略箇所からもっとも近いものを選択した. ゼロ照応が複数の実体を参照している場合は, ID 値の集合を参照した.

参照先に曖昧性がある場合は, case 値に”曖昧”と格をあわせて付与した.

(4) 学会の歴史を<ゼロ ref=”1” case=”GA”><ゼロ ref=”2” case=”曖昧 GA”>振り返る</ゼロ></ゼロ>

格が交替している述語に対しては, ゼロ交替タグを用いて, 格交替ごとに参照先を補った[3].

(5) 京都大学に
<ゼロ交替 ref=”1” case=”GA” ><ゼロ ref=”1” case=”WO”>発掘さ<ゼロ>れ</ゼロ交替>た.

複文において, 名詞句が従属節と主節の両方の述語にかかっている場合は, 従属節側で省略されているとして, 省略を補った.

間接照応 (bridging reference)

述語におけるゼロ照応だけでなく, 間接照応にもタグ付けを行った. 間接照応については, 日本語教育や談話解析に関する分野などでいくつか研究が行われている [7][11][12].

本研究では, ゼロ照応と同じくゼロタグを用いて case の値を”NO”とすることで, 間接照応関係を示した. 参照先の選択ではゼロ照応に順じて距離を優先させ, 語義照応関係にタグ付けを行った.

(6) <ゼロ ref=”1” case=”NO”>中身</ゼロ>はカラだった.

なお, 間接照応の参照先が外界一般となる場合については, タグ付けを行っていない. また動作性名詞については, 間接照応の照応詞として認定していない.

代名詞照応

照応詞を伴った照応では, 人称代名詞, 指示代名詞, 指示形容詞, 関係接頭辞にタグ付けを行った. 指示形容詞については限定指示と代行指示を区別した. 表 3 にタグ名と例の一覧をあげる.

参照先の選択においては, 基本的に指示照応を優先させた. 記事中に同一実体を指示している名詞句がある場合はその名詞句を参照し, ない場合は語義照応となっている名詞句を参照した. 両者を区別するために cl 属性を導入し, 同一実体を参照している場合は”同

[3]表層上は, 格交替を起こしているが, 原形に戻すことができない場合は, 専用のマークを用いた.

一”, 別実体の場合は”異同”という値を付与することとした. 指示照応関係にある名詞句が記事中に複数存在した場合は, 初出の名詞句を参照先とし, 初出の名詞句が照応表現であった場合には, 次の名詞句を参照先とした. 指示照応は, 定名詞間の同一指示関係のみを対象とし, 総称名詞, 不定名詞は参照先および照応詞として認定しなかった. 同様の議論には飯田ら [13] がある.

名詞句照応

裸名詞句による指示照応関係にタグ付けを行った. タグ名は裸名詞句とした. 代名詞照応と同様に, 参照先の ID 値を ref 属性で参照するが, cl 属性の値はすべて”同一”となっている. 参照先は, 指示照応関係にある名詞句中で初出の名詞句とした. ただし, 記事中に集合をあらわす表現とその集合のすべての要素が出現している場合には, 集合をあらわす表現を参照元とし, それぞれの要素に与えた ID を ref 属性に持つことで指示照応関係を示した. その結果, 初出の名詞句が参照先とならない場合も存在する.

2.3 文書間照応

文書間照応タグ付け作業は, 複数文書要約のデータを対象に行った. トピックごとに記事をまとめ, 前節までの文書内照応タグ付け作業を行った後, 記事を時系列順に並べて作業を行った. 他文書タグと KEY タグの 2 種類がある.

他文書タグは, ある記事中の名詞句が, ほかの記事中の名詞句と同じ実体を参照していることを示す指示照応のタグである. ref 属性を持ち, 他の記事にある参照先の ID 値を参照することで, 照応関係を示す. 参照先は, 指示照応関係にある名詞句中で, 時間的に若い記事の初出の名詞句とした. 文書内照応タグ付け作業で, 指示照応のタグがすでに付与されている名詞句群に関しては, 記事中で参照先となっている名詞句に他文書タグを付与することで, その記事の指示照応関係にある名詞句群全体にタグを付与したことと等しいとした.

(7) 記事:990423004
<ID id=”990423004_101”>アメリカ</ID>北西部の,.....
記事:990425003
<他文書 ref=”990423004_101”>米国</他文書>ワシントン州の,.....

TSC-2 の複数文書要約タスクでは, トピックごとにキーワードが与えられている. それらについては, 指示照応でなくても語義照応であれば, タグ付け作業の対象とした. 指示照応関係と区別するため, KEY タグを用い, 参照先は記事中に求めず, タグに含めた.

(8) <KEY key=”携帯電話”>携帯</KEY>の普及率は,.....

3 作業結果と問題点

3.1 タグ付け結果

述語のゼロ照応は 21310 箇所にタグが付与された. 表 4 にまとめる. 縦に省略された格, 横にゼロ照応と参照先の距離を文単位で示す. 間接照応は 4800 箇所にタグが付与された. 同じく表 4 に示す.

代名詞照応と名詞句照応の結果は, 表 5 に示す. 総計 15868 箇所にタグが付与された.

文書間照応では, 他文書タグで示された関係は 6351 箇所, KEY タグで示された関係は, 2579 箇所であった.

3.2 作業上の問題点

作業する上で問題となった点をいくつか上げる.

ゼロ照応のタグ付けにおいて, 述語の主体となるべき語が助詞ノや助詞デで出現している場合があった. この場合は省略とはみなさないこととした.

(9) a. 彼**の**行くところは, 大体わかる.

b. 県警**で**現場を捜索した.

ゼロ照応では, 参照先となり得る名詞句が記事中に数多く出現する場合があるが, 今回は候補となる名詞句の中から省略箇所と参照先との位置関係で参照先を決定している. その結果, 語義照応関係となっている場合があり, ゼロ照応の参照先が総称名詞や不定名詞だった場合には, 他の参照先候補との関係が示せず, 照応解決タスクで問題になる可能性がある. また作業上では, ゼロ代名詞は省略されていても, 文の意味が理解できるという性質から, 述語ごとに省略の有無を検証していかなければ, 簡単に見過ごしてしまう問題点もあった. 作業開始時点でアノテイタ間で互いの作業結果を見比べる場を設けたり, 作業結果を別のアノテイタが見直すなどの工夫が必要であった.

間接照応のタグ付けにおいて, 参照先が記事中に存在している場合は, 作業者間の一致をとることができたが, 外界一般となる場合について一致をとることが難しかった. どのような性質を持った名詞句を照応詞

表 4: ゼロ照応と間接照応の結果

	前方照応					後方照応					外界			
	4 以上	3	2	1	同文	同文	1	2	3	4 以上	一般	一人称	二人称	節
ガ	1712	683	1378	3195	5757	391	195	78	55	290	3587	310	7	15
ヲ	154	65	138	401	694	76	18	5	4	12	193	2	0	5
ニ	206	73	159	362	567	33	25	14	10	41	616	6	3	1
カラ	9	6	6	28	49	2	3	0	0	2	36	0	0	0
ト	10	3	9	29	49	1	1	3	0	0	7	0	0	0
ノ	777	307	470	1234	1727	101	55	24	20	72	12	1	0	0

表 5: 代名詞照応と名詞句照応の結果

	代名詞		指示代名詞		代行指示		限定指示		関係接頭辞		裸名詞句
	同一	異同	同一	異同	同一	異同	同一	異同	同一	異同	同一
前方	153	25	354	98	433	144	269	68	1095	126	11947
後方	9	5	6	3	3	1	12	4	22	10	192
一般	0	56	0	223	0	121	1	44	0	212	0
一人称	26	0	0	0	0	0	0	0	0	0	0
二人称	0	0	0	0	0	0	0	0	0	0	1
節	0	2	0	73	0	121	0	8	0	1	0

とするかについて作業者間で一致がとれていなかったためと考えられる.

代名詞照応および名詞句照応において, 時間・空間表現に対する指示照応の定義が問題となった. 次の例で「29 日」と「同日」は照応関係にあるが, 届出をした時刻と受理された時刻は正確には異なるかもしれず, 指示照応関係かどうかの判断が難しい.

(10) **29 日**に届出をし, **同日**受理された.

空間表現でも同様の問題が起こり得る. 今回は時間・空間表現はそれが指示し得る範囲が過不足なく重なった場合を指示照応とした.

4 展望

1 節で述べたように, 本研究でタグ付けの対象としたテキストには, 重要文抽出や, 談話構造についてのタグがすでに付与されている. これらに省略, 照応, 共参照タグが加わったことにより, 照応解決の研究材料となることはもちろんであるが, 重要文抽出や談話構造解析の分野において照応現象がどのレベルで関わっているのかを検討することが可能となるだろう.

参考文献

[1] MUC-6: http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html

[2] MUC-7: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

[3] GNOME: http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/

[4] GDA: http://i-content.org/gda/

[5] ACE: http://www.nist.gov/speech/tests/ace/ace04/

[6] 関 和広, 藤井 敦, 石川 徹也: ゼロ代名詞の検出と補完を統合した確率的照応解消モデル, 言語処理学会第 8 回年次大会発表論文集, 2002.

[7] 河原大輔, 黒橋禎夫, 橋田浩一: 「関係」タグ付きコーパスの作成, 言語処理学会第 8 回年次大会発表論文集, 2002.

[8] TSC: http://lr-www.pi.titech.ac.jp/tsc/

[9] 横山 憲司, 難波 英嗣, 奥村 学: Support Vector Machine を用いた談話構造解析, 情報処理学会研究報告, NL-15, 2003.

[10] 衛藤純司, 奥村学: 文書横断文間関係タグ付コーパスの構築, 言語処理学会第 11 回年次大会発表論文集, 2005.

[11] 村田 真樹, 長尾 真: 意味的制約を用いた日本語名詞における間接照応解析, 自然言語処理, Vol.4, No.2, 1997.

[12] 礒江 健史, 竹井 光子, 相沢 輝昭: ゼロ連体格代名詞の自動検出システム, 言語処理学会第 10 回年次大会発表論文集, 2004.

[13] 飯田 龍, 乾 健太郎, 松本 裕治, 関根 聡: 機械学習による日本語名詞句照応解析の一手法, 言語処理学会第 10 回年次大会発表論文集, 2004.