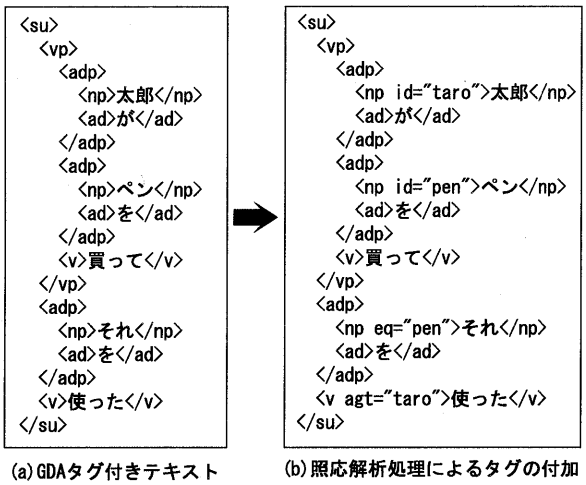
8M-08 不完全なタグを持つ文書の照応解析によるタグの洗練化

山本 浩司† 白松 俊† 伊藤 誠悟†
橋田 浩一‡ 奥乃 博†
† 東京理科大学 理工学部 情報科学科 ‡ 電子技術総合研究所

統語, 意味情報を明示できるように設計されたXMLタグセットであるGDA (Global Document Annotation) タグでは人手によるタグ付けのコストを抑制するため、さまざまな自動解析ツールが必要となる. 本稿では, 統語解析により一定レベルのタグの付いたGDA文書に対して照応解析を行うことによって照応関係のタグの洗練化を行うツールを開発した. 具体的には文書中の指示詞や代名詞が何を指すのかを, 候補を挙げ, 先行詞の候補となる主題, 焦点の重みや, それらから指示詞までの距離等を考慮して重み付けすることによって照応関係を推定し, 照応関係を明示するタグを自動的に付加する。

1. GDAについて

GDAプロジェクト[1]は, 多言語間に共通の統語・意味等に関するXMLタグの標準を作って普及させ, GDAタグを用いた情報検索, 機械翻訳, 要約などの技術の実用化を狙っている. タグ付けには曖昧性の解消が必要であり, そのためには人間がタグを付けることになるが, 人手による厳密なタグ付けには多大な労力を要する. その労力を軽減するために, 構文解析と照応解析により自動的に一定のレベルでタグを付与するツールが必要である. 本稿では, 構文解析によるタグ付けツール[6]によってタグ付けされた文書を, 照応解析することによって文章中の代名詞, 指示詞, ゼロ代名詞の指示対象を推定し, その照応関係のタグの洗練化を行う研究について述べる (図1).



(a) GDAタグ付きテキスト (b) 照応解析処理によるタグの付加

図1 照応解析によるタグの洗練化

Anaphoric-analysis-based refinement of GDA documents
by Koji Yamamoto†, Shun Siramatsu†, Seigo Ito†, Koiti Hasida‡, Hiroshi G. Okuno†
†Science University of Tokyo, ‡Electrotechnical Laboratory

2. 照応解析の手順

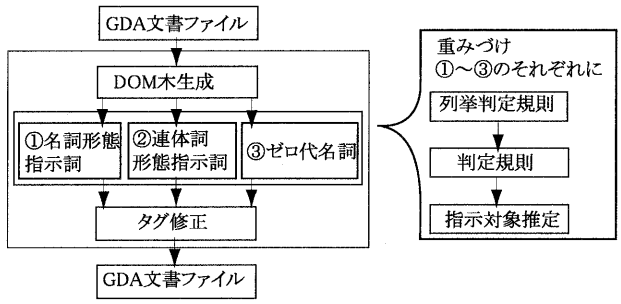


図2 照応解析処理の流れ

照応解析は, 名詞形態指示詞, 連体詞形態指示詞, ゼロ代名詞のそれぞれの場合に分けて行う (図2). 解析手順を以下に示す.

1) GDA文書をXMLパーサにかけDOM (Document Object Model) 木 (図3) にする
2) 照応詞を抽出
3) 照応詞の先行詞の候補の抽出
4) 列挙判定規則による重み付け
5) 列挙判定規則で得点が付いた候補に判定規則を用いさらに重み付けをする
6) 最も得点の高い候補を指示対象とし, それを反映するようタグを修正する

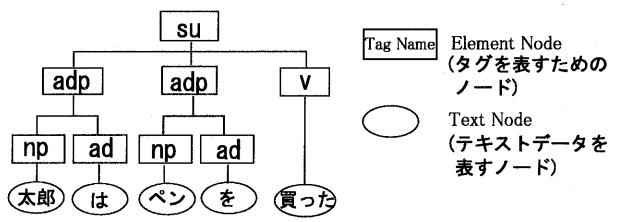


図3 DOM木

3. アルゴリズム

DOM 木生成後

指示詞の抽出

入力は DOM 木のすべてのノードを格納したベクターである. 構文解析によるタグ付けツール[6] により, 指示詞には

```
<ad hinsi="連体詞形態指示詞">その</ad><n>人</n>
```

のように品詞の情報を持つ属性が付いているので, 入力のそれぞれのノードに対し, 品詞情報の属性があれば参照し, 指示詞であれば, 配列に格納する. 出力はその配列である.

先行詞の候補の抽出

入力は DOM 木のすべてのノードを格納したベクターである. ここでタグ名が ad(助詞, 副詞等) であるもの以外のエレメントが持つテキストノードを配列に加える. 出力はその配列である.

列挙判定規則と判定規則

代名詞などの指示対象を決定するための情報はいろいろ絡みあっており, それぞれ単独で指示性を決定することは困難である. そのため推論規則に確信度の得点づけをして, 規則が適用されるたびにその得点を足すということをする. 列挙判定規則, 判定規則については文献[2)3)] を参考にした. 抽出された指示詞それぞれに対し, まず列挙判定規則を適用し先行詞の候補に重み付けをする. 以下に列挙判定規則の一例を示す.

主題・焦点に重みを与える規則

名詞が指している先行詞の候補としては, 文章中に既に現れた主題や焦点が話の中心を担うため, 先行詞になりやすい. 主題と焦点の定義と与える重みは表 1, 表 2 のとおりである.

表 1 主題の重み

表層表現	例	重み
ガ格の指示詞・代名詞・ゼロ代名詞	(太郎が) した.	21
名詞 は/には	太郎はした.	20

表 2 焦点の重み

表層表現 (「は」がつかないもので)	例	重み
ガ格以外の指示詞・代名詞・ゼロ代名詞	(太郎に) した.	16
名詞 が/も/だ/なら/こそ	太郎がした.	15
名詞 を/に/, /.	太郎にした.	14
名詞 へ/で/から/より	学校へ行く.	13

列挙判定規則でポイントがついた候補にのみ, 続いて判定規則を適用する. 以下に判定規則の一例を示す.

指示詞は人を指しにくいという性質を利用した規則

名詞形態の指示詞の場合で, 指示対象の候補となった名詞に人名を意味するタグ<persname>が付いているとき, その候補に-10 点を与える. ただし, 指示詞が「彼」や「彼女」のときは, 人を指すと考えられるので 10 点を与える.

「ここ」「そこ」などは場所を指しやすいという性質を利用した規則

照応詞が「ここ」/「そこ」/「あそこ」の場合で, 指示対象の候補となった名詞に地名を意味するタグ<placename>が付いているとき, その候補に 10 点を加える.

選択 (指示対象推定)

最も得点の高いものを指示対象とする.

タグ修正

先行詞, 照応詞の対に対しそれぞれ照応関係を示す id, eq 属性を付加する

4. ゼロ代名詞の抽出のための格解析

ゼロ代名詞の検出のために, 格解析を行う. まず, EDR 日本語動詞共起パターン副辞書[5] から動詞について概念関係子を抽出し, タグに付いている二項関係の情報と比較して不足している部分を検出する.

```
<adp opr="agt">
  <n opt="たろう">太郎</n>
  <ad hinsi="副助詞">は</ad>
</adp>
<vp syn="fc">
  <n opt="いし">石</n>
  <ad hinsi="格助詞">を</ad>
  <v opt="ひろって" bfm="拾う">拾って</v>
</vp>
<v opt="なげた" bfm="投げる">投げた</v>
```

図 4 二項関係付き GDA 文書

図 4 では<adp opr="agt">という二項関係の情報がタグ付けされている. そこから

投げる agt 太郎 (投げる の主格は 太郎)

という二項関係を抽出する. EDR 日本語動詞共起パターン副辞書から「投げる」は, agent と object という概念関係子を持ち, 主格と目的格を要することがわかる. ここで, 主格についての二項関係はあるが目的格に関するものが見当たらないため, その部分が不足しているとわかる. 不足部分が検出できれば普通の指示詞などと同様にその部分の先行詞を探す. 見つかればゼロ代名詞の照応であるとし, 見つからなければ単なる省略であるとする.

5. 評　　価

実装は Java で行った. また, 新聞記事から 20 の例文を用意し実験したところ, 13 の文に対して正しい指示対象を推定した. 指示詞と先行詞との距離が遠い場合等に推定の誤りが見られた.

6. ま と め

本稿では, 照応解析による GDA タグの洗練化について述べた. 今後の課題としてはシソーラスなども利用し推定の精度を高める予定である.

参 考 文 献

1) 橋田 浩一:大域文書修飾 Global Document Annotation (GDA) http://www.etl.go.jp/etl/nl/gda/
2) 長尾 真: 岩波講座ソフトウェア科学 15 自然言語処理, 岩波書店,1996.
3) 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, 自然言語処理 (言語処理学会誌),1997.
4) W3C Document Object Model (DOM): http://www.w3.org/DOM/
5) EDR 電子化辞書:http://www.iijnet.or.jp/edr/
6) 横山 他:不完全なタグを持つ文書の構文解析によるタグの詳細化, 8M-06, 本大会