

|              |                                                                                 |
|--------------|---------------------------------------------------------------------------------|
| Title        | 教師なし学習で推定した確率分布によるゼロ代名詞の解析                                                      |
| Author(s)    | 杉原, 大悟                                                                          |
| Citation     |                                                                                 |
| Issue Date   | 2003-03                                                                         |
| Type         | Thesis or Dissertation                                                          |
| Text version | author                                                                          |
| URL          | <a href="http://hdl.handle.net/10119/1654">http://hdl.handle.net/10119/1654</a> |
| Rights       |                                                                                 |
| Description  | Supervisor:鳥澤 健太郎, 情報科学研究科, 修士                                                  |

# 教師なし学習で推定した確率分布によるゼロ代名詞の解析

杉原 大悟 (110065)

北陸先端科学技術大学院大学 情報科学研究科

2003 年 2 月 14 日

キーワード: 照応解析, ゼロ代名詞, 教師なし学習, EM アルゴリズム.

## 1 背景と目的

本研究では, 教師無し学習手法のみで推定された確率分布によるゼロ代名詞照応解析手法を提案する. 自然言語による情報の伝達の過程においては, 情報を受け取る側が容易に推測可能な事柄を指示する表現は代名詞で置き換えられることが多々発生する. ある言語表現が, 後に現れる言語表現と同一の対象を指す場合, これら表現は照応関係にあるといい, 前者を先行詞, 後者を照応詞という. ゼロ代名詞とは, 文中からは見えない照応詞のことである. ゼロ代名詞は日本語において一般的な現象であり, その照応解析は, 機械翻訳や要約など, 他の自然言語処理アプリケーションを作成する際に非常に有益である.

ゼロ代名詞照応解析のための様々な手法が, 先行研究において提案されている. 関ら [2001] や河原ら [2001] は, 大規模コーパスから自然言語に関する資源を自動的に取得し, それを用いてゼロ代名詞の照応解析を行っている. 河原らは格フレーム辞書を大規模コーパスから自動的に構築する手法を提案し, その格フレーム辞書をゼロ代名詞照応解析に応用している. 関らは, コーパスから推定した確率分布を用いてゼロ代名詞照応解析を行っている. しかし, 河原の手法では, 人手によるシソーラスを用いている. 関らの手法でも, 人手によるシソーラスを用いており, さらに, 素のコーパスとゼロ代名詞照応関係のタグのついたコーパスの両方から確率分布を取得している. このように, 先行のゼロ代名詞照応解析研究によるゼロ代名詞照応解析のための言語的資源の取得には, 何らかの人手によるコストがかかっている. 本研究では, これら先行研究に対し, ゼロ代名詞照応解析のために用いる言語的な資源を, タグ無しコーパスから教師無し学習のみによって得る手法を提案する.

## 2 本研究の手法

本研究で扱うゼロ代名詞照応解析のための言語的な資源は 2 つある. 1 めは Rooth [1999] の教師無し手法で取得された「名詞の動詞に対する係りやすさ」を表す確率分布である.

実際には、本研究では、Rooth の手法を日本語に対して拡張した鳥澤 [2001] の手法による確率分布を用いる。鳥澤の手法では、構文解析結果を用いて学習データに重み付けをしている。その構文解析器の学習には EDR コーパスが用いられているが、EDR コーパスにおいて人手で与えられている正解は、鳥澤の手法で推定される確率分布の正解にはなりえないので、鳥澤の手法を教師無し手法と考える。そして、2 つめは、Rooth の手法による確率分布を利用して本研究にて推定される「名詞と動詞との位置関係による名詞のゼロ代名詞の先行詞になりやすさ」を表す確率分布である。本研究では、テキスト中の名詞のゼロ代名詞の先行詞になりやすさは、名詞と動詞との間の意味的な関係だけでなく、名詞と動詞との距離、名詞に後置する助詞など、様々な統語的な要素によっても左右されると考える。本研究では、そのような要素を「名詞と動詞との位置関係」と呼ぶ。本研究では、 $R$  を名詞と動詞との位置関係を表すベクトルとし、以下のように表す。

$$R = \langle d, c1, rentai, head, eachid \rangle$$

$d$ : 名詞と動詞の文間距離.  $d \in \{0, 1, 2, \dots\}$

$c1$ : 名詞に後置する助詞.  $c1 \in \{0, 1, 2, 3, \dots\}$  これらの数字によって助詞を区別する。

$rentai$ : 名詞が連体修飾節に含まれているか.  $rentai \in \{0, 1\}$

$head$ : 名詞がテキストの先頭の段落に含まれているかどうか.  $head \in \{0, 1\}$

$eachid$ : 複数の名詞とある動詞が同じ  $\langle d, c1, rentai, head \rangle$  の値を持つ時、それらを区別する id.  $eachid \in \{1, 2, \dots\}$

本研究では、テキスト中に現れる様々な「動詞と名詞の間の位置関係」を上記の  $R$  の値として区別する。本研究では、テキスト中の名詞がゼロ代名詞になる時を表す確率モデルを構築し、それに対して EM アルゴリズムの一般的な手法を適用することで、「 $R$  の値を持った名詞が格  $c$  でゼロ代名詞の先行詞になる度合い」を表す確率分布を推定する方法を導出した。この  $R$  に関する確率分布は、素のコーパスから抽出可能な名詞の出現頻度および  $R$  の値の出現頻度、そして Rooth の手法による教師なし手法で推定された確率分布に対して、一般的な EM アルゴリズムの導出過程を適用することで推定される。本研究では、この  $R$  に関する確率分布と、Rooth の手法による確率分布を用いてテキスト中の名詞の「ゼロ代名詞の先行詞になりやすさ」を評価し、それを元にゼロ代名詞照応解析を行った。

### 3 実験

本研究では、読売新聞 3120 記事から、我々の推定手法に従って  $R$  に関する確率分布の推定を行った。そして、Rooth の手法による確率分布と、本研究にて推定された  $R$  に関する確

率分布を用いて、ゼロ代名詞の先行詞の特定に関する実験を行った。実験には、学習データには含まれていない読売新聞 10 記事を用いた。これらの記事には、あらかじめゼロ代名詞の出現個所と照応関係の正解を与え、正解コーパスとした。実験は、ゼロ代名詞の位置と格が正しく検出されたという前提で行われ、正解コーパス中の各ゼロ代名詞の先行詞候補を、確率の値によって評価し、値の高い先行詞候補から順番に出力させた。本研究では、ある格のゼロ代名詞の格要素の数が  $k$  個の時に、出力した先行詞候補の上位  $k$  位に正解が含まれているならば、システムは正しく先行詞特定できたとした。精度は、正しく特定できた先行詞の数を、正解コーパス中のゼロ代名詞の数で割ることで算出した。本研究での先行詞特定の正解率は 41.07% となり、残念ながら先行のゼロ代名詞照応研究に劣る結果になった。しかし、Rooth の手法による確率分布のみで先行詞の特定を行う場合の精度と、 $R$  に関する確率分布と Rooth の手法による確率分布の両方を用いて先行詞の特定を行う本研究手法の精度を比較したところ、本研究手法による精度が、Rooth の手法による確率分布のみを用いた場合の精度よりも 11% 高くなることが分かった。以上から、本研究では、本研究の手法によって取得された確率分布により、ゼロ代名詞照応解析の精度の向上が確かめられたと結論づける。