動詞と格を用いたパターン分類によるゼロ照応問題の解決

情報科学科　近藤 駿佑　　指導教員：山村 毅

1　はじめに

人々が生活するために欠かせないものの一つに，文章がある．文章は何かを人に伝えるための重要な手段の一つであり，例として新聞，雑誌などからインターネットを使った記事，ブログ，ニュースまで様々な場面で使われる．このような文章では，何回も同じ単語を使う場合や，話の流れから書かなくても理解できるような場合にはそれを省略したりすることがある．このような表現を照応表現という．自然言語処理システムを構築していく上でこのような照応表現が何を表しているのかを明らかにすることは不可欠である．

本研究では照応解析を機械学習を用いて行うことで省略部分を明らかにしていくシステムを考える．

2　ゼロ照応解析

ゼロ照応解析とは，文の中で省略されている部分を明らかにする解析のことである．そして，省略部分が指している内容を先行詞という．

太郎は朝ごはんを食べた。そして，(ϕ) シャワーを浴びた。

この文において ϕ は省略部分を表すが，この指している内容は太郎なので，先行詞は太郎ということになる．

3　提案手法

3.1　概要

まず本研究では，扱うゼロ照応の条件として，先行詞が省略部分の一文前にあるものに限定して解析を行うものとする．この際，先行詞が一文前のどの格に属するのかを判断できた時点で先行詞の特定に成功したものとする．また，取り扱う格は「ガ」「ガ２」「ニ」「ト」「ヲ」「デ」の 6 つに限定する．「ガ２」格は主格以外で「ガ」格が使われている場合の主格を表すときに用いる．本研究で提案するゼロ照応問題解決手法は，省略部分の存在する文脈を特徴表現し，そこから先行詞への写像をパターン分類問題をとらえて，機械学習で行うものである．

3.2　特徴量

特徴量は省略部分が含まれる文の動詞またはサ変名詞，省略部分の格，一文前の動詞またはサ変名詞を用いる．

太郎は学校に行った。そして (ϕ) 給食を食べた。

この文から次のように特徴量を表現する．

食べた　ガ　行った　ガ

また，次のように動詞とサ変名詞が混在する文の場合，区別をつけるために特徴量を追加したものも考える．

太郎は学校に行った。そして (ϕ) 授業に参加した。

この文から次のように特徴量を表現する．

食べた　動詞　ガ　参加　サ変名詞　ガ

3.3　分類器

3.2 で述べた特徴表現を用いて，その先行詞が一文前の「ガ」「ガ２」「ニ」「ト」「ヲ」「デ」の 6 つの格のどれに当てはまるかを決定する分類器を学習する．

4　実験と評価

3 で述べた手法を実装し，その性能を評価した．学習データ及びテストデータには京都大学テキストコーパス [1] を用いた．また，分類器の実装には weka を用いた．分類器としては，SVM，ニューラルネットワーク，ナイーブベイズを用いた．そして評価には，10 分割交差検定を用いた．

5　結果

特徴量の追加をする前の結果を表 1 に示し，特徴量を追加した後の結果を表 2 に示す．表 1 と表 2 より，3 種類の分類器の中で SVM が最も正解率が良いことが示された．また，特徴量を追加すると SVM のみ正解率が上がることが示された．

表 1　特徴量追加前の正解率

分類器	SVM	NN	bayes
正解率	77.5	77.4	72.8

表 2　特徴量追加後の正解率

分類器	SVM	NN	bayes
正解率	78.1	76.4	71.2

6　考察とまとめ

動詞と格を特徴量として用いると一文前のゼロ照応は精度よく解析できることがわかった．その中でも SVM を用いる解析が 8 割近い正解率を誇ることが示された．特徴量を追加した場合でも，その効果は薄く，NN と bayes を用いた解析では逆に正解率が下がってしまうという結果が得られた．このことから動詞とサ変名詞の区別は有用ではないことがわかった．今後正解率を更に上げていくためには，特徴量をもっと増やすということと，正解が「ガ」格以外となるものの学習データが少ないので，そのデータも増やすことでより精度が高まると思われる．

参考文献

[1] 黒橋禎夫, 長尾眞. 1997. 京都大学テキストコーパス・プロジェクト. 言語処理学会第 3 回年次大会予稿集, pp.115−118.