最大エントロピーモデルを用いた括弧表現の分類

情報科学科　落合里咲　　指導教員：山村 毅

1　はじめに

日本語で記述される文章には，丸括弧や鉤括弧などの括弧表現が多く使われる．括弧表現には様々な用法が存在し，1 つの記号が複数種類の表現に使用される．記述された文章を読む際，我々はその括弧表現を無意識のうちに用法ごとに扱いを変えて読んでいる．用法の違いは文法的な役割の違いでもあるので，例えば，文章に対して構文解析等の機械処理を行う際にも，括弧表現の扱いを用法ごとに変える必要がある．本研究では，新聞記事の文章に対して構文解析等の機械処理を行うことを前提に，括弧表現が本文とどのような関係を持つのかを明らかにして分類することを目指す．

括弧表現の分類を扱った研究として，中山ら [1] は，新聞記事から丸括弧と鉤括弧の括弧表現を抽出し，形態素解析結果をもとに，それぞれ 16 種類と 3 種類の用法があることを述べ，それらに分類する分類器を構築している．また岡崎ら [2] は，括弧表現「X(Y)」において，表現「X」と「Y」に言い換えの関係が成立するかどうかを調べる言い換え発生率 PR(X,Y) という指標を提案し，ほかの複数の指標とともに SVM の素性として教師あり機械学習を行うことで，言い換え可能性に関する分類器を構築している．

2　括弧表現の分類カテゴリ

本文に対して括弧記号が必要か不必要か，括弧表現が必要か不必要か，またその処理方法の違いに基づいて以下の 6 つの分類カテゴリを設けた．

- 補足情報…本文に対して補足的な情報を付加するもの．
  例) 第４６回毎日芸術賞（０４年度）の受賞者が…
- 引用…文章もしくは単語を基本的にそのまま引用するもの．
  例) ９２年に「君がいるだけで」で日本レコード大賞。
- 独立文…括弧表現のみで独立した文として成立するもの．
  例)「因習的な、白人、男性、勝者の視点だ」。
- 小見出し…連番・小見出しなど，リスト表現に用いるもの．
  例)（１）石油収入の配分，（２）軍の統合…
- 補完…文章の内容理解のために，補われた表現を表すもの．
  例) 今すぐ（国連の）崩壊というのは避けたいと。
- 記号表現…省略表記など，表現の一部として扱われるもの．
  例) 角交換から［後］４九角の打ち込みが丸山期待の反撃だ。

3　分類手法

毎日新聞の 5 日分の記事から括弧表現を 7809 個抽出し，2 で述べた正解の分類ラベルを人手で付けたものを用いて，分類実験を行った．各カテゴリのデータ数を表 1 にまとめる．

表 1　実験データの分類

ラベル	補足情報	引用	独立文	小見出し	補完	記号表現
表中の表示	A	B	C	D	E	F
数	3382	2915	438	640	103	331

3.1　特徴抽出

文章から分類に有効な素性を選択し抽出することを特徴抽出という．今回，括弧表現の文における位置や括弧記号の形などに着目した素性を 24 個抽出した．

3.2　分類器

本研究では，分類器に最大エントロピーモデル [3] を用いている．最大エントロピーモデルは確率に基づいた分類器の一つであり，与えられた制約を満たすモデルの中でエントロピーを最大化するようなモデルを解として推定する．

4　実験と結果

抽出した素性 24 個全てを使用した分類実験と，分類の難しいカテゴリに着目した 6 個の素性を選択した分類実験を行った．

それぞれ 10 分割交差検定により実験を行った結果を表 2 と表 3 に示す．素性を 24 個使用した場合には 88.49%，素性を 6 個選択した場合には 88.71% の精度が得られた．

素性を 24 個使用した実験と 6 個選択した実験の正解率を比べると，6 個の方が約 2 ポイント高い．これは有意水準 5% において有意な差があり，単純に素性を増やすよりも，分類の難しいカテゴリに着目した素性を使用した方がよいということが言える．

表 2　分類結果 (素性 24 個)

	分類結果							
ラベル	A	B	C	D	E	F	正解数	正解率
A	3274	21	12	67	0	8	3274	96.81
B	17	2820	73	5	0	0	2820	96.74
C	96	194	120	28	0	0	120	27.40
D	203	0	8	385	0	44	385	60.16
E	101	0	0	2	0	0	0	0.00
F	155	0	1	20	0	155	155	46.83

表 3　分類結果 (素性 6 個)

	分類結果							
ラベル	A	B	C	D	E	F	正解数	正解率
A	3246	23	37	77	0	0	3246	95.95
B	14	2866	29	6	0	0	2866	98.32
C	37	31	233	137	0	0	233	53.20
D	157	10	0	432	0	41	432	67.50
E	51	0	0	52	0	0	0	0.00
F	176	0	5	0	0	150	150	45.32

5　おわりに

本研究では，6 つの括弧表現カテゴリを提案し，有効な素性を抽出し，最大エントロピーモデルを用いて分類を行った．その結果，素性 24 個使用した場合は約 86.5%，素性を 6 個選択した場合は約 88.7% の精度が得られた．

カテゴリ別に見ると補完は正解率 0% であり，これは実験データにおいて補完のデータが少ないためであると考えられる．

また，独立文を引用や小見出しに，小見出しや記号表現を補足情報に，それぞれ誤って分類したものが多かった．対応策として，似た傾向のカテゴリを同じカテゴリとして分類を行い，その後，カテゴリ内で更に細分化することが有効だと考えられる．

参考文献

[1] 中山悟，森田和弘，泓田正雄，青江順一:“括弧表現の抽出・分類に関する研究”，言語処理学会第 16 回年次大会，pp.379-382．2010.
[2] 岡崎直観，石塚満:“言い換え可能な括弧表現の抽出法”，言語処理学会第 13 回年次大会，pp.911-914．2007.
[3] 岡崎直観:“Classias - A collection of machine-learning algorithms for classification”，http://www.chokkan.org/software/classias/