括弧表現の抽出・分類に関する研究

中山悟　　森田和宏　　泓田正雄　　青江順一
徳島大学大学院 先端技術科学教育部

1. はじめに

我々が記述する文章や新聞記事には，丸括弧や鉤括弧などの括弧表現が使われる．丸括弧には，“欧州連合(EU)”という言い換え的用法や，“(彼の)ノート”という補完的用法，他にも読み的用法，補足的用法，所属的用法などがある．また，鉤括弧においても，“「私は雨が好きです」と話した”という会話的用法や，“小説「城」を読む”という題目的用法など，括弧表現は様々な用法が存在する．1 つの記号が複数の表現に使用されているため，日英翻訳などの機械処理をおこなう上で，括弧表現の種類を識別する必要がある．他にも，文章中に存在する括弧表現の用法を特定することができれば，括弧表現の省略判定，括弧に着目した未知語の抽出，人名抽出などさまざまな情報抽出に応用できる．

括弧表現の出現数として，2004 年の毎日新聞記事によると，861,324 記事中，丸括弧表現が 383,032 回，鉤括弧表現が 328,504 回出現している．つまり，新聞記事において丸括弧，鉤括弧表現の出現はおよそ 5 記事中 4 記事に出現することになり，新聞記事をコーパスとした日英翻訳などの機械処理において括弧表現の抽出，分類は非常に重要であるといえる．

そこで本稿では，新聞記事から 2 種類の括弧表現(丸括弧，鉤括弧)を抽出し，形態素解析結果をもとに，字種情報や体言概念[1]，係り受け関係などを用いて，自動で分類する手法を提案する．また，この手法を用いて実験をおこない結果と考察について述べる．

2. 関連研究と括弧表現の用法

2.1 関連研究

これまでに，括弧に着目したさまざまな研究がおこなわれている．白井ら[2]は，丸括弧，鉤括弧のそれぞれの用法を分析し，日英翻訳をおこなう際に有効となる加工方法について述べた．久光ら[3]は，“欧州連合(EU)”のような丸括弧表現“X(Y)”における“X”と“Y”の共起頻度やルールを組み合わせることで，言い換え可能な“X”と“Y”を抽出する手法を提案した．岡崎ら[4]は，言い換え発生率を提案し，久光らの手法と組み合わせたものを素性として機械学習をおこなった．しかし，言い換え可能な丸括弧表現の抽出精度は，80%と満足できるものではない．三宅ら[5]は，対話システムにおける音声認識率の改善を目的とし，流行語や新語を中心に自動的に読みを付与する手法を提案した．読み付与の際に，“iPhone(アイフォン)”などのような括弧表現の読み的用法に着目し，結果として，90%以上の精度が報告されている．また，後藤ら[6]は，機械翻訳などに必要な文書短縮に着目し，鉤括弧の用法をニュース原稿から，引用と強調の 2 種類に分けて自動判別をおこなった．実験結果として，適合率 93%，再現率 82%という精度が報告されている．

これらの報告から，丸括弧においては，言い換えや読みといった 1 つの用法に絞って抽出する研究はおこなわれているが，全ての丸括弧を抽出し，分類する研究はおこなわれていない．また，鉤括弧においても，関連研究では，2 種類に分類しているが，本稿ではさらに細分化をおこなう．

2.2 括弧表現の用法

前節で述べた関連研究と2004年の毎日新聞記事に含まれる括弧表現から，丸括弧と，鉤括弧の用法をそれぞれ 16 種類と 3 種類に分類した．表 1，表 2 に用法と例文を示す．

丸括弧において，「頭文字」，「外来語の頭文字」，「その他の換言」は，言い換え可能な表現である．「著者」は新聞，雑誌などに用いられるものであり，主に文末に表記される．「複合」は括弧の中に 2 つ以上の用法が存在するものである．例文の“プリンスホテル(都内渋谷区，田中社長)”の場合は「複合(場所，補足・注釈)」となる．「その他」は，丸括弧表現 15 種類の用法に該当しないものであり，“(笑)”，“(中略)”など文章の内容とは直接関係しない表現となる．

また，鉤括弧において，「会話」は会話の引用文である．「強調」は“鉤括弧”を省略しても文が成り立つものであり，「題目」は“鉤括弧文”を省略しても成り立つものとする．例えば，“小説「城」を読む”の“「城」”を省略すると“小説を読む”となり“鉤括弧文”を省略しても文が繋がるので「題目」である．この「題目」は鉤括弧直前に“会社”や“小説”などの名詞になることが多く，会社名や小説のタイトルなどの未知語が抽出できる．

表 1　丸括弧の用法と例文

用法	例	用法	例
頭文字	東京大学(東大)	補足・注釈	旧正月(春節)
外来語の頭文字	非政府組織(NGO)	発言者	「メジャー入りを目指す」(防衛庁幹部)
その他換言	所得税法違反(脱税)	数字	(1)，(2)，(3)
読み	翁(おきな)	年月	第 45 回芸術賞第(03 年度)
場所	オマーン戦(埼玉)	職業	ブッシュ(米大統領)
所属	イチロー(マリナーズ)	著者	カント没後 200 年である。(専門編集委員)
年齢	イチロー(30)	複合	プリンスホテル(都内渋谷区，田中社長)
補完	(国際社会の)メジャー入り	その他	(笑)，(中略)

表 2　鉤括弧の用法と例文

用法	例
会話	「私は雨が好きです」と話した
強調	「みかん」に似ている
題目	小説「城」を読む

3. 提案手法

3.1 丸括弧の分類法

丸括弧を含む文に対し，形態素解析をおこなう．解析結果をもとに以下の手順で分類する．

Step1：丸括弧前後との依存関係を判定

文における丸括弧の位置，丸括弧文の品詞，丸括弧前後の品詞を取得する．取得することによって，語 X，Y に対して，“東京大学(東大)”のような“前方依存型：X(Y)”，“(国際社会の)メジャー入り”のような“後方依存型：(Y)X”，“(笑)”のような“単独：(Y)”の 3 種類に分類することができる．これにより，「頭文字」は“X(Y)”，「補完」は“(Y)X”のように依存関係を判定することでおおまかな分類をおこなう．

Step2：単独，後方依存型の処理

“単独：(Y)”，“後方依存型：(Y)X”は“前方依存型”に比べ，用法が限定される．“単独”は，「著者」，「その他」，“後方依存型”は，「数字」，「補完」となる．特徴を以下に示す．

- 著者
  Y の概念が“人”
- 数字
  Y が数字
- 補完
  括弧を省略しても文が繋がる
- その他
  新聞記事において「その他」の表現は固定されているため「その他」表現に関するルール作成により分類できる

これらの特徴をもとに単独，後方依存型の分類をおこなう．

Step3：前方依存型の処理

“前方依存型： X(Y)”の用法は多様であるが，用法ごとの特徴をもとに分類することができる．以下に Step2 に属さなかった 12 用法の特徴を以下に示す．

- 頭文字
  1. Y と X の先頭文字が一致
  2. Y の各文字が X にそれぞれ存在する
- 外来語の頭文字
  X，Y のうち一方が体言(英語を含まない)，もう一方が英語のみの表記
- その他換言
  丸括弧の言い換え表現は，“所得税法違反(脱税)”，“脱税(所得税法違反)”のように X と Y を入れ換えても文中に存在することを利用し，以下のように分類する
  1. Google N グラム[7]を用いて“X(Y)”，“Y(X)”それぞれを検索
  2. “X(Y)”，“Y(X)”両方が存在するものを抽出
  3. 2 を満たさないものでも，X が“ここ”，“あれ”などの指示代名詞であれば抽出
- 読み
  1. X と Y の読みが同じ
  2. Y に英語，漢字，数字が含まれない

- 場所
  X が体言，Y の概念が“地名”
- 所属
  X の概念が“人”，Y の概念が“学校”または，“組織”
- 年齢
  1. X が漢字のもの，または概念“人”
  2. Y は数字のみ，または“歳”，“才”を含む
  3. Y は 0～130 までの値
- 補足・注釈
  補足や注釈の内容を一意に決める法則性はない
- 発言者
  X が鉤括弧の終端記号，Y の概念が“人”
- 年月
  X の語尾が“年”，“月”，“日”，“年度”または，X が 130 以上の数字
- 職業
  X が体言，Y の概念が“職業”
- 複合
  「複合」は，語 A，B に対し，“X(A,B)”とすることができるため，“X(A)，X(B)”として処理をおこなうことで判定

括弧はいずれかの用法に分類できるため，特徴に属さないものは「補足・注釈」となる．これらの特徴をもとに前方依存型の分類をおこなう．

3.2 鉤括弧の分類法

丸括弧同様に，鉤括弧を含む文に対し，形態素解析をおこなう．解析結果をもとに以下の手順で分類する．

Step1：鉤括弧の位置を特定

鉤括弧の位置を特定する．鉤括弧が文中に存在するときは Step2，文末に存在するときは Step3 へ．また，1 文が“▼”，“■”などの記号と鉤括弧文のみの場合は「題目」に分類し，終了する．

Step2：鉤括弧文と係り先の関係を判定

鉤括弧直後の品詞に着目し，鉤括弧直後の品詞が助詞“と”，“などと”となる場合は係り先を係り受け解析を用いて特定する．特定された係り先が“話す”や“言う”などの発話に関する表現であれば「会話」に分類し，終了する．適合しない場合は Step3 へ．

Step3：鉤括弧文の最後が用言か体言かを特定

判定対象となる鉤括弧文の最後が用言で終わるか，体言で終わるかを特定する．最後が用言で終われば Step4，体言で終われば Step5 へ．

Step4：鉤括弧文の最後が用言で終わる場合の分類判定

鉤括弧内が用言で終わるときは「会話」となるが，一部の「強調」，「題目」を特定する必要がある．

そこで，「会話」の特徴から，鉤括弧直前が名詞であることは稀であるため，鉤括弧直前が名詞以外であれば「会話」に分類し，終了する．鉤括弧直前が名詞の場合は，Step5 へ．

Step5：鉤括弧文の最後が体言で終わる場合の分類判定

鉤括弧内が体言で終わるときの会話文は Step2 で「会話」に分類されるため，殆どの文が「強調」，「題目」となる．

「強調」と「題目」は文を省略できるかどうかで決まるため，鉤括弧前後の品詞や，体言概念を用いて省略判定をおこなう．例えば，“小説「城」を読む”であれば，鉤括弧前後の品詞が“名詞+助詞”となり，省略可能，“あれは「りんご」のようだ”であれば，前後の品詞が“助詞+助詞”となり省略不可能と判定される．判定結果により，鉤括弧自体が省略可能となれば「題目」，そうでなければ「強調」に分類し，終了する．

4. 実験・考察

4.1 実験設定

提案手法の有効性を示すために実験をおこなった．丸括弧用に 2006 年の毎日新聞記事から 1,000 個の丸括弧表現を含む 625 記事，鉤括弧用に 2006 年記事 1,000 個の鉤括弧表現を含む 650 記事をそれぞれ用いた．

分類結果の適合率，再現率を人手で評価した．適合率，再現率の算出式は以下のとおりである．

$$\text{適合率} = \frac{\text{正解に分類できた数}}{\text{抽出した数}} \times 100$$

$$\text{再現率} = \frac{\text{正解に分類できた数}}{\text{正解となる数}} \times 100$$

4.2 実験結果

丸括弧，鉤括弧それぞれの分類結果を表 3，表 4 に示す．各用法の正解となる数を全正解数，提案手法によって抽出した数を抽出数，正解に分類できた数を正解数と表記した．表 3，表 4 より，丸括弧は適合率，再現率共に 83.0%，鉤括弧は約 87%となり，良い精度を示すことができた．これにより，従来手法に比べ細分化することができたといえる．

丸括弧に着目した人名抽出(「所属」，「年齢」，「発言

表 3　丸括弧を含む記事の分類結果

用法	全正解数	抽出数	正解数	適合率	再現率
頭文字	1	1	1	100.0%	100.0%
外来語の頭文字	35	36	34	94.4%	97.1%
その他換言	19	31	17	54.8%	89.4%
読み	87	75	73	97.3%	83.9%
場所	99	90	84	93.3%	84.9%
所属	33	18	17	94.4%	51.5%
年齢	117	134	115	85.8%	98.3%
補完	41	25	24	96.0%	58.5%
補足・注釈	378	440	334	75.9%	88.4%
発言者	16	15	12	80.0%	75.0%
数字	118	85	84	98.8%	71.2%
年月	22	15	11	73.3%	50.0%
職業	7	5	5	100.0%	66.7%
著者	9	5	4	80.0%	44.4%
その他	5	4	4	100.0%	80.0%
複合	13	21	11	52.4%	84.6%
計	1,000	1,000	830	83.0%	83.0%

表 4　鉤括弧を含む記事の分類結果

用法	全正解数	抽出数	正解数	適合率	再現率
会話	499	508	462	90.9%	92.6%
強調	401	360	327	90.8%	81.6%
題目	100	129	84	65.1%	84.0%
計	1,000	997	873	87.6%	87.3%

者」,「著者」)の精度は，適合率 86.1%，再現率 84.6%，省略判定(「頭文字」,「外来語の頭文字」,「その他換言」,「補完」)の精度は，適合率 81.7%，再現率 79.2%となった．鉤括弧に着目した未知語抽出(「題目」)においては，適合率 65.1%，再現率 84.0%となった．

4.3 考察

丸括弧における「所属」,「職業」,「著者」などの再現率が低くなった原因として,体言概念のみに依存していることが挙げられる．例として，“佐々木(ガーラ湯沢)”は,「所属」が正解となるが,「補足・注釈」と誤解析してしまった．これは，“ガーラ湯沢”を概念“組織”と判定できなかったためである．このように，体言概念のみの判定では抽出できない例がみられた．体言概念は日々増えていく新語などには対応できない上に,既知語であっても全ての語に概念を付与するには,莫大なコストなどがかかるため困難である．つまり，体言概念に依存しない処理の考案が必要であるといえる．

鉤括弧における「題目」の適合率が低い原因として，“その結果「ボール」を投げることになった”のように「強調」が正解であるのに鉤括弧前後が“名詞+助詞”となるものをすべて「題目」と判定していることが挙げられる．文章を省略できるかどうかで「強調」と「題目」を判定しているため，“小説「ゼブラ」は人気だ”などの文は「題目」となり，正解となる．これらは，体言概念である程度の判定は可能であるが，前述したように，鉤括弧も丸括弧同様,体言概念に依存しない処理を新たに考案する必要がある．

5. まとめと今後の予定

本稿では，丸括弧，鉤括弧の 2 種類の括弧表現の抽出，分類について述べた．精度実験において，丸括弧は適合率，再現率共に 83%，鉤括弧は約 87%となった．

考察で述べたように体言概念に頼らない処理,または,コストなどの問題もあるが,体言概念の拡張によって精度が向上すると考えられる．今後は，体言概念に依存しない処理を考案し，精度向上を図りたい．

参考文献

[1] 池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦（編）:“NTT コミュニケーション科学研究所（監修），1997 :『日本語語彙大系』”

[2] 白井諭，矢部孝幸，松尾三津恵，西垣万亀子，大山芳史 :“新聞記事文における括弧書き表現の分析とその処理について”，情報処理学会第 53 回全国大会, 2L-9, Vol.2, pp.31-32, 1996.

[3] 久光徹，丹羽芳樹 :“統計量とルールを組み合わせて有用な括弧表現を抽出する手法”,情報処理学会自然言語処理研究会,NL-122,pp.113-118, 1997.

[4] 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.

[5] 後藤功雄，熊野正，江原輝将 :“かぎ括弧で囲まれた表現の種類の自動判別”，言語処理学会第 6 回年次大会，pp.35-38，2000.

[6] 三宅純平，竹内翔大，川波弘道，猿渡洋，鹿野清宏 :“括弧表現に基づく Web テキストマイニングを用いた流行語への自動読み付与の提案”，電子情報通信学会技術研究報告，Vol.108,No.422, SP2008-126,pp.1-6.

[7] 工藤拓，賀沢秀人著 : Web 日本語 N グラム第 1 版，言語資源協会発行