丸括弧解析システムの構築*

菅野紘平　　横山晶一　　西原典孝

（山形大学大学院理工学研究科）

１．はじめに

文章中に使われる様々な記号は、その定義に必ずしも従わず多様な使われ方をしている。機械処理において記号がどのように使われているかを判断することは難しく、記号部分が無視されるなど、まともに扱われないことが多い。

しかし記号の中には文章中に必要な情報が含まれる場合も多々あり、それらを機械的に解析することは非常に重要である。

これまでの研究[1～3]では、様々な記号の中で丸括弧について分類し、その機械処理の可能性について考察した。一般に丸括弧内の機械処理は難しいが、その使われ方によっては機械処理が可能である。

本稿では、丸括弧を含む文の機械処理について取り扱う。丸括弧はその使われ方によって「読み」や「略語」、「言い換え」等、幾つかに分類することができる。この分類に基づき、既に括弧の働きを明示するシステムを構築した[3]。

丸括弧が文章中でどのような役割を果たしているかを判断する方法としては、括弧の文章中での位置、括弧内外の文字種や文字数などの情報による判断や、括弧内外の情報の比較、データベースへのマッチング等が考えられる。

本研究ではこれらの情報に加え、略語辞典[4]、日本語係り受け解析システム「南瓜」[5]、分類語彙表[6]等の新たなデータやシステムを用いることにより、既発表のシステムでは判別できなかった分類の括弧についても処理を可能にした。更に括弧内外の関係についてタグを用いた表示を行うことで、既存のシステムと比較してより明快で精度の高い結果を得ることに成功した。

２．丸括弧の分類

高木の研究[1,2]では、新聞記事を中心に丸括弧を含む文を約 1600 文抽出して、括弧の使用法別に分類を行った。また、機械化に向けて括弧の分類に基づきアルゴリズムを考察した。以下に例文を挙げ、括弧の分類について述べる。

２．１．記号

慣用的な使われ方をするもので、丸括弧を含む部分をひとつの語として登録する必要がある。たとえば「(？)」、「(株)」等は１つの記号として登録する。

２．２．括弧内が固有名詞

括弧内が地名、人名、組織名など、様々な形を取る。固有名詞の意味情報を参照できれば、括弧外の語と比較することで判別可能な場合がある。

(1) 女子でメダル候補の上村愛子（北野建設）と里谷多英（フジテレビ）の取材に１００人近い日本人プレスが詰め掛けかけたため、空港は一時騒然とした。

文(1)では括弧内に組織名の固有名詞が存在している。括弧直前語に注目すると人名があり、括弧内外の比較により、括弧内が所属する組織を示すことが判明する。

２．３．括弧内が数字や単位

括弧内が数字の場合、箇条書きの他に年齢、年号、物量等があり、助数詞や数値、括弧直前語等に注目して処理を行う。

(2) 合弁会社の資本金は２１０万ドル（約２億８０００万円）で、吉野家の出資比率は５０％。

文(2)では助数詞が共に通貨の単位であり、括弧が単位の換算を示すことが判別できる。

２．４．読み仮名

同義語とも取れるが、括弧内が平仮名や片仮名のみで構成される場合に多く見られる。

２．５．括弧内・括弧直前いずれかが略語

新聞記事等では多くの略語が使われ、しばしば正式名称と共に括弧内外に併記される。

(3) 日本やＥＵなどは今回の措置に強く反発し、世界貿易機構（ＷＴＯ）に提訴する構えで…

---

* A Parentheses Analysis System
SUGANO Kohei, YOKOYAMA Shoichi,
NISHIHARA Noritaka (Yamagata University)

この分類は次に述べる同義・言い換えとも取れるが、基本的に略語単体で単語としての意味を持たない造語である場合の分類である。

２．６．同義・言い換え

括弧内外が交換可能であるものを指す。ただし読み仮名や数字、略語等、先に述べた分類の括弧は除く。

(4) どういう形で国際社会に対して関わっていくのかという根本的なポリシー（戦略）を、平素からしっかりと定めておかなければ駄目だ。

判別には括弧内外の意味を比較する必要がある。

２．７．情報補足

括弧前後に対して情報を補足しているものである。広義では、全ての括弧は情報補足に当たるが、ここでの情報補足は、括弧を取り除いても意味の重複が起こらない、すなわち完全に新しい情報を加える種の括弧を指す。

(5) …多くの都市では意外な（といっては失礼だが）人物が現地社会に食い込んでいた。

この種の括弧は実に多様な使われ方をするが、括弧内の品詞の形式に注目することで、判別可能なものも存在する。

同義・言い換えや情報補足は、既発表のシステムでは解析対象外という扱いであったが、本システムより新たに解析対象となった。

３．丸括弧解析システム

本システムの構成を図１に示す。入力文に対し、形態素解析、括弧部分の分別、係り受け解析、括弧判別、タグ付けの順に処理を行う。各処理について以下の節で説明する。

３．１．形態素解析による括弧抽出

入力文補正（半角→全角処理や改行の置き換え等）の後、補正した入力文に対し茶筌[7]による形態素解析を行う。形態素解析の目的は、入力文を形態素単位で区切り、各形態素の品詞や意味情報を得ることで括弧内外の意味比較に用いること、また括弧部分を分別することである。

３．２．係り受け解析

形態素解析により括弧部分を除いた入力文を抽出し、南瓜[5]による係り受け解析を行う。係り

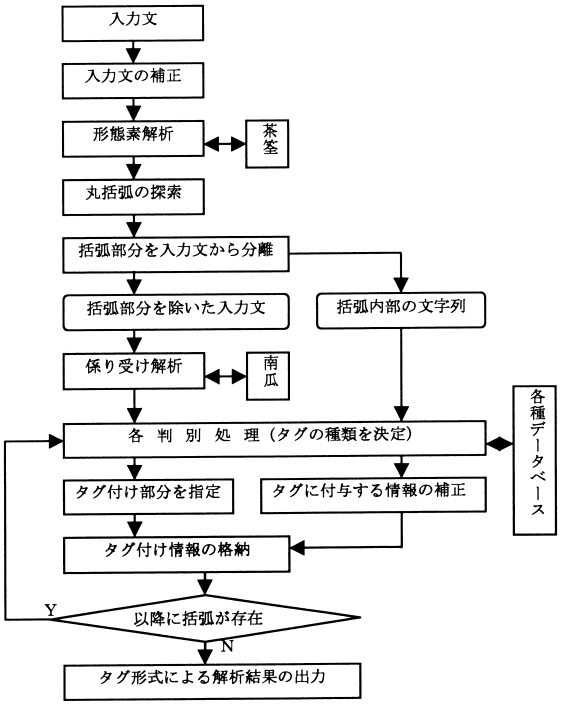

図１　システム構成

受け解析結果は、括弧内に対する括弧外比較対象の限定やタグ付け箇所の決定に役立てられる。

既発表のシステムでは、文型に関わらず括弧直前から一定数の形態素を取り出して括弧内と比較していた。そのためしばしば括弧内と無関係の語にまで比較がおよび、誤った結果が生じる問題があった。

(6) インターネットがテロリストの次の標的にされるかもしれないとして､米下院は７日(米国時間)､新世代の「サイバー戦士」を養成する法案を可決した。

○括弧対応箇所の判別
既発表システム（係受解析無）：米下院は７日
本システム（係受解析有）：７日

文(6)において､形態素単位の意味比較を行った際、既発表のシステムでは括弧直前の任意の形態素を比較対象とするため、括弧外“米下院は”と括弧内“米国時間”の２文節間の結びつきが強いと判断した。しかし、括弧部分を除いた本文に係り受け解析を行うと、“米下院は”の係り先は括弧直線文節“７日”ではなく、“可決した”であることが判明する。

基本的に括弧は直前の句または文に対して注記や補足を行うものであるため、括弧内外の意味の近さを調べるには括弧直前文節に直接的、また

は間接的に係り受け関係にある語のみを抽出し、比較すればよい。本システムでは係り受け解析を用いることで、余分な比較対象を排除し、誤判別の防止に役立てている。

係り受け解析を用いたタグ付けについては、３．４で述べる。

３．３．判別処理部：パターン解析、および形態素間の意味の近さの数値化

一部の種類の括弧（主に言い換えや情報補足）では、その判別に括弧内外の意味比較が必要不可欠である。前研究では括弧内外の意味比較を日本語語彙大系[8]の意味体系を参照することで行った。

(7) 今なお生業として鷹狩りを行っている「最後の鷹匠」松原英俊氏（山形県朝日村）と、彼の家族の姿を追跡取材した、写真も豊富なホームページ。

文(7)では括弧内外の名詞“氏”と“村”に注目した意味比較により、“人物（地名）”という形を取ることから括弧内が出身地、もしくは在住地を表した情報補足であると判別できる。

しかし言い換えの場合、括弧内外が同様のことを述べているという前提があるために、上記の例のようなパターン解析を用いるよりも、括弧内外の意味の近さを数値化し、数値に基づいた判別を行うことで、より幅広い解析を行うことができる。

本研究では日本語語彙大系[8]、分類語彙表[6]の２つのシソーラスを参照して２形態素間の意味の近さの数値化を行った。

日本語語彙大系[8]による点数（score1）は、２単語の意味属性を調べ、上位の意味カテゴリから順にカテゴリが一致しているか判定し、カテゴリの一致度を点数にしたもので、分類語彙表[6]による点数（score2）は、分類番号の近さを点数としたものである。点数付けは下記の計算式に従って行われる。

score1＝100×[カテゴリ一致数]÷[浅い階層にある名詞の階層数]
score2＝（10000－[分類語彙表の“品詞番号を除く分類番号”の差]）÷100

この計算値を用いることにより、パターン解析を用いることなく言い換えの判別が可能となる。

(8) 第２次補正予算によって、今年度予算の国債依存度（歳入に占める国債の割合）は最終的に４３％にものぼることになった。

文(8)で注目する形態素は、括弧内最終形態素“割合”と括弧直前形態素“度”である。これらの形態素について、シソーラスで意味を検索すると、抽象的な量を表すものであることが判明する。しかしこの例の判別を行うために、“抽象的な量（抽象的な量）＝言い換え”というパターンを用意する必要は無く、単に括弧内外の意味の近さを分析することで判別が可能となる。

“度”と“割合”についてscore1およびscore2をそれぞれ計算すると、次の値が算出できる。

score1＝100×4÷5＝80
score2＝（10000－1970－1960）÷100＝99.9

本システムでは意味が近いと判断する境界を
score1≧75 or score2≧99
に設定しているので、この２形態素は非常に近い意味を持っていると判断できる。

括弧内外は同様のことについて述べているという結果を受けて、システムの出力部において以下のような言い換えを表すタグ付けが行われる。

(9) 第２次補正予算によって、今年度予算の〈para=″歳入に占める国債の割合″〉国債依存度〈/para〉は最終的に 43%にものぼることになった。
（para＝paraphrase：言い換え）

３．４．タグ形式による解析結果の出力

文(9)で示したように、本システムでは解析結果をタグ付けによって行う。

（例１）
入力：日本など先進十八カ国が政府開発援助（ODA）を拠出する。
出力：（ODA）⇒略語

例１に示す通り、既発表のシステムでは括弧の種類のみを出力するに留まった。しかし、出力から“ODA”が括弧外のどの部分と対応しているかを判断できず、括弧の持つ曖昧性が解消されていないという問題があった。

本システムでは出力にタグを用いることにより、括弧内に対応する括弧外の範囲を明確に表現することを可能にした。

例２の出力では、タグにより括弧内“ODA”に対応する語が“政府開発援助”であることが明確に表現されている。

（例２）
入力：日本など先進十八カ国が政府開発援助（ODA）を拠出する。

出力：日本など先進十八カ国が<abbr="ODA">政府開発援助</abbr>を拠出する。
（abbr＝abbreviation：略語）

（例３）
入力：息子はエリート校であるニューヨーク市内の私立学校（幼稚園から高校まで）に通っています。
出力：息子は<ela="幼稚園から高校まで">エリート校であるニューヨーク市内の私立学校</ela>に通っています。
（ela=elaboration：情報補足）

例３はタグ付けに係り受け解析を用いる例である。括弧直前の係り受け関係が“エリート校である→ニューヨーク市内の→私立学校に”であるのに対し、“息子は→通っています。”であるので、“息子は”は括弧内に関係の無い文節であると判別し、タグの範囲には含まれない。

４．評価・考察

本研究により、既発表のシステムでは処理できなかった括弧の多くが処理可能となった。しかし、未だ未対応の括弧文も多く残っており、これらの処理方法を考えることがこれからの課題となる。

表１　システム評価

	既発表システム	本システム	total
基本括弧精度	46.65%	72.07%	1031
拡張括弧精度	3.95%	41.02%	902
全体精度	26.92%	57.58%	1933

表１は高木の研究[1,2]において新聞記事から抽出した約1600文の丸括弧文と、今回新たにweb上から収集した約300文の丸括弧文をシステムに入力した際の解析精度である。

基本括弧は前研究[3]で研究対象となった種類の括弧、拡張括弧は本研究で新たに処理対象に加えた“同義・言い換え”、“情報補足”の括弧を表している。

既発表のシステムではシソーラスから単語の意味を検索することに重点を置いていたが、本研究ではパターン処理に注目し、多くの解析パターンを取り入れることで基本括弧の処理精度向上に繋がった。

更に既発表のシステムではほとんど処理できなかった拡張括弧についても、ある程度の処理が可能となるシステムが完成した。

（10）租特法の強みはアメ（減税）とムチ（増税）がひとつの法案で国会に提出される点だ。

ただし、文(10)のように括弧内外がシソーラス上意味的に近いと判断できず、また形式的な特徴も無い括弧文は処理が不可能である。

５．おわりに

本研究で評価に用いた文は新聞記事が中心となるため、一定のルールのもとに使われている括弧が多い。しかし実際には、括弧は使用者によって自由な形式をもって使われるものであるため、こういった使用者独自の使われ方をする括弧についてどのように処理するべきかを考える必要がある。

また、本システムの出力を受けてどのような応用ができるか、またどのように入力文を書き換えるかを引き続き検討し、実用的なシステムの構築を目指す。

参考文献

[1] 高木美紀：機械処理のための丸括弧の分類、山形大学工学部平成 13 年度卒業論文（2002）
[2] 横山晶一・高木美紀：丸括弧の分類とその機械処理のための考察、言語処理学会第８回年次大会発表論文集（2002）
[3] 菅野紘平：丸括弧解析システムの構築、山形大学工学部平成 14 年度卒業論文（2003）
[4] Yoshio Kobayashi：略語辞典、http://www.inv.co.jp/~yoshio/
[5] 奈良先端科学技術大学院大学：日本語係り受け解析システム「南瓜」
[6] 独立行政法人国立国語研究所：分類語彙表増補改訂版、大日本図書（2004）
[7] 奈良先端科学技術大学院大学：日本語形態素解析システム「茶筌」
[8] NTT コミュニケーション科学研究所：日本語語彙大系、岩波書店（1997）