括弧表現の機械翻訳

麻野間 直樹　松尾 義博
日本電信電話株式会社　NTTサイバースペース研究所
{asanoma,yosihiro}@light.hil.ntt.co.jp

1　はじめに

新聞記事やWWW上の文書などのテキストには，引用文や補足説明，箇条書きなどを表す目的で括弧を含む表現がしばしば使われる．

括弧文字は，それを含む文の内容には直接関係なく，文の構造などの言語外情報を記述している機能的な文字[1]（以下，機能文字と呼ぶ）と言える．したがって実際の自然言語処理では，通常の文解析とは別に括弧表現用の特別な処理を設けなければならない[1]．

問題の重要性を計る意味で，日本語新聞記事1年分（「CD-毎日新聞 98 年版」のタイトルおよび本文を利用）のコーパスから，括弧文字を考慮した文区切りを行った[2]上で括弧文字を含む文の割合を調査した．

集計の結果，全体文数は 1,680,340 で，そのうち括弧を含む文数は 746,444 であった．全体の 4 割強もの文が括弧文字を含むことになり，括弧文字への対処が重要であることがわかる．

括弧表現に関連する従来の言語処理技術としては，括弧に関する用法の分類と簡単な処理法[1]，またその括弧分類の自動判別をする研究[2, 3]，括弧表現から言語知識を獲得する研究[4]などが行われてきた．しかしこれまでは，実際の自然言語処理システムへの適用とした網羅的な括弧の分類とその効果の測定が不十分であった．

本稿では，新聞記事などの記述文を対象とする機械翻訳システムへの適用を前提とした，括弧表現の翻訳方式について述べる．

以下ではまず，機械翻訳として必要な機能という観点で対処すべき括弧表現の意味分類を整理する．この括弧の分類に基づいて，括弧用の特別な処理について個別に説明する．さらに，この括弧用処理の効果を計る実験結果について述べる．

2　括弧の分類と傾向

括弧文字の用法は多種多様だが，ある程度のパタン化は可能である．その役割として，引用の括弧や補足説明のための括弧など一般的な分類は知られている（一例;[1]）．

ここでは，原言語を日本語とする機械翻訳を行うにあたって，検討対象の括弧文字と括弧の意味分類を再整理した．日本語テキストに含まれる括弧文字としては図2に示すものがあり，本稿ではこの全てを対象とする．

括弧表現の分類は表1に示すようになった[3]．この中でラベル参照は，[2]にもあるように，箇条書きされていた項目の内容を指す時に，代名詞のようにふるまう括弧表現を表す．これは他の括弧表現と異なり，括弧文字は内容語の一部となる．補足情報の括弧表現は，語句や文・文章の直前・直後に位置して補足的な説明を付与する．

次にこの分類が全て重要であることを示すために，まず括弧の分類毎の出現傾向を調べた．前節で用いた新聞記事コーパスから，括弧文字を含みなおかつ文として成り立っているもの[4]をランダムに 200 文抜き出した（これをテスト文Aとする）．この文中のすべての括弧表現について，その分類（用法）を人手で判別

[1] この定義から，句点などの文集短を表す記号（。！？▲），節や句のまとまりを示す読点なども機能文字の一種と言える．

[2] 1文の終端は，改行記号もしくは，括弧表現の内部に存在しない文終端記号までとする．

[3] 実際のテキストには以下のような意味の括弧表現が現れる．この中の括弧文字は単語の一部分と取れるため機能文字とは言いにくい．
数式　∫ｆ（ｘ）ｄｘ
顔文字　（ ´∀｀）

[4] 空白文字によって整形された表形式の断片などは評価対象外とした．

表１：括弧表現の分類と，新聞記事コーパス内の例文および出現頻度

分類		例文	件数
引用		a 「あの博物館では核兵器の悲劇は伝わらない」とカンダンさんは言う。	86
	複数文	b さらに、クマのぬいぐるみを示し「植物実験を手伝ってくれています。８番目の乗組員のクマちゃんに名前を付けて下さい」と子供たちに呼びかけた。	17
強調		c 優等生ゆえの“孤独”が垣間見える。	76
補完		d かつて行き過ぎ（た接待）があったことは、反省している。	3
箇条書き		e 条件とは（１）２ヶ月間以内の期限（２）各所１回ずつの回数制限（３）安保理構成国の外交官を同行させること――の３点。	5
		f 夏への参院選への対応を協議し、（１）比例代表は新党友愛を基軸に対応する（２）選挙区選挙では候補者の郵政事業への対応を重視する、との方針を決めた。	
ラベル参照		g 常識的には（２）のケースの可能性が高く …	1
補足情報（換言）		h どこでもゲシュタポ（ナチスの国家秘密警察）の目が光っていた。	115
		i そこで小学２年の長男（８）、長女（５）、二男（４）と暮らしていた。	
		j ［解説］国立大の独立行政法人化断念	
		k アジア太平洋経済協力会議（ＡＰＥＣ）エネルギー担当大臣会合	
読み		l 急に心筋梗塞（こうそく）で意識を失い …	14
		m つまり、彼らによって、私たちは蘇（よみがえ）るのだ。	

1. 括弧内が直前単語の読みに後方一致する か，もしくは
   （ひらがな、または"・" で構成された2文字以上の文字列 かつ 直前単語が漢字を含む）ならば 読み
2. 括弧内が （数字1文字 または 英小文字1文字 または ローマ数字）かつ 括弧表現が文末でなく，かつ
   - 括弧直後が(助詞 か 助動詞 か "、") かつ （括弧直前が名詞か接尾辞)でない ならば ラベル参照
   - 括弧直後が(助詞 か 助動詞 か 括弧記号以外の記号） でない ならば 箇条書き
3. （括弧文字が引用括弧 か 括弧内の最終単語が名詞ない）ならば 引用・強調・補完
4. そうでなければ 補足情報

図１：括弧表現分類の判定手順

し集計した．結果は表１の「件数」に示す．

この結果から補足情報，引用・強調の括弧は圧倒的に多く現れ，複数文を含む引用や読みも無視できない頻度で出現していることがわかる．このように頻出する括弧表現に対する対処は優先的に行うべきだと考える．

一方，機械翻訳への適用を考えると，翻訳品質を大きく低下させる原因についてはその対処も考慮する必要がある．

読み，箇条書き・ラベル参照，複数文の引用などの括弧表現の場合，文の構造の解析を誤ると翻訳文の品質が致命的に低下する可能性が高い．例えば，括弧用の処理を行わない場合，表１の例文ｌ，ｍのような文に対して不用意にひらがなを解析・訳出しようとし，著しく翻訳品質を下げる可能性がある．このような括弧表現は出現件数に関係なく対処すべきと考える．

引用括弧	‘ ’ “ ” 「 」『 』
その他の括弧	（ ） 〔 〕 ［ ］ ｛ ｝ 〈 〉 《 》 【 】 ＜ ＞ ≪ ≫

図２：JIS X 0208 の括弧相当文字

3 括弧分類の判定

冒頭で述べたように，括弧文字は機能文字なので，誤りのない翻訳処理やその性能向上を目指すためには，システムはそれ用に特別な処理を行う必要がある．

翻訳処理中では，まず前節で定義した括弧表現の分類を自動的に判定する．

括弧表現の分類の判定は，機械学習などを使わず決めうちの判定ルールによって行う．判定ルールでは，括弧文字や括弧内文字列の文字種，括弧表現の位置，括弧表現周辺単語の品詞

など局所的な情報を用いる.

この括弧分類の判定手順の一例を図１に示す．図中の条件は先頭から順に適用していく．括弧分類の判定では，文節の単位や品詞情報を用いるので，実装上は形態素解析直後にこの処理を行う．

4　括弧表現に対する操作

翻訳処理中に設ける括弧表現用の操作は，（1）部分文を分割して翻訳（2）句や節の係り受けを制限（3）括弧表現自体を無視して解析，の３つである．

括弧表現の分類が決定すると，それぞれどの操作を行うべきかの対応関係が決まる．表２に示す括弧表現の分類と操作の対応関係について，各分類に対して○印の操作[5]を行う．箇条書きについては，箇条ラベルの部分に加えて箇条書きの「項目」の部分の翻訳も考慮する．

表２：括弧表現の分類と操作の対応

分類		分割翻訳	係受制限	括弧内の翻訳
引用・強調・補完		×	○	○
	複数文	○	—	○
箇条書き	（ラベル）	○	×	○
	（項目）	○	○	—
ラベル参照		×	×	○
補足情報		○	×	○
読み		—	×	×

4.1　分割翻訳

分割翻訳は，入力文のうち括弧表現周辺の部分文がその他の部分と構造的に独立な時に，元の文と切り離して翻訳することを示す．別個に翻訳した結果は最後に元の文に埋め込む．

4.1.1　文や語句に対する補足情報

語句や文に付加される補足的な説明の括弧表現は，その括弧の内部は独立した文とみなせる．換言・略記の補足情報に加えて，直前語に対する訳語，箇条ラベルに対して分割翻訳の操作を行う．

4.1.2　箇条書き

文中に番号と箇条項目が羅列されている箇条書きの形式で，項目が一つの文として成り立っている場合（例文ｆ参照）は，元の文とは独立しているとみなせる．

4.1.3　複数文を含む引用

引用の括弧表現の中に複数の文が含まれている場合（例文ｂ参照）は，引用している元の文とは独立している．分割する時，基本的には句点などの文終端記号に基づいて一文を認定し抜き出せばよい[6]．

4.2　係り受け関係の制限

括弧表現が文の構造を記述している性質を利用すると，括弧表現部分内部もしくは直後にある部分文字列の係り受け関係に制限が加えられ，解析精度の向上が期待できる．

一文としては成立していないが，句や節の塊（係り受け制限文字列）が括弧表現によって判別できる場合にこの制限が適用できる．係り受け制限文字列は，引用などの場合は括弧の内部，箇条書きの場合は箇条書きの個々の項目に相当する．例えば，表１の例文ｅでは，３つの箇条項目が名詞句として成り立っており，それぞれに係り受け制限が適用可能である．

係り受けの制限として以下のような条件を採る．

> 「係り受け制限文字列内のヘッド以外の文節と，係り受け制限文字列外の文節との間の係り受け構造を認めない」

係り受けを制限する箇条書き項目の範囲は，箇条ラベルで挟まれた部分で決定できる．ただし，最後の箇条項目だけはある程度条件[7]を設けてその範囲を決めなければならない．

4.3　翻訳しない括弧表現

仮名で直前語の読みを表す括弧表現は，それ自体無視して処理するほうが正常な翻訳が可能となる．表１の例文ｌは単語の直後に読み括

---

[5] 実装上は，他にも読み括弧表現を無視する形態素解析や，「ａ）」のような閉じ括弧のみの括弧表現への対応といった機能が必要である．

[6] 文区切り記号を含むような単語では，その単語で文が区切られるとは限らないので工夫が必要である．単語の例として，“モー娘。”，“電車でＧＯ！”などがある．こういった問題は一般に文章から一文を切り出す際にも考慮すべきである．

[7] 文中の最後の箇条項目の終了点の条件は，例えば，文字列“――”の直前まで，あるいは読点＋助詞まで，などが考えられる．

表３：括弧表現分類の判定性能

分類	正解数	抽出数	適合率	再現率
引用・補完・強調	152	157	92%	97%
複数文	17	17	100%	100%
箇条書き	5	5	100%	100%
ラベル参照	1	1	100%	100%
補足情報	109	115	95%	95%
読み	14	14	100%	100%

表４：括弧表現用処理による翻訳品質の変化

訳文変化	件数
向上	58
不変	128
低下	4
差分無	10

弧がある例で，“（こうそく）”が読みの括弧表現と分類されたら，その後の処理では文構造から読みの括弧を削除し，“心筋梗塞で”という文節を形成する．

また表１の例文mは，単語の途中に読み括弧が入り込んでいる例で，“(よみがえ)”を削除した上でその前後を結合させた“蘇る”を一語の形態素として認識させる．このような括弧表現は形態素解析前に削除する必要がある．

5 実験

5.1 分類の判定性能評価

括弧表現の分類性能を測るため，テスト文Aに対して実際に括弧表現分類を自動判定し結果を検証した．解析系はルールベース型日英機械翻訳システムALT-J/E[5]を用いた．結果を表３に示す．

結果として，全分類とも高い判定性能を得られた．これによって，3節で示したような，比較的複雑でない判定手順でもかなり良い判定精度が得られることがわかった．

5.2 翻訳の品質評価

括弧表現用の処理を使う場合と，使わない場合とを比較して，テスト文AをALT-J/E[5]に走行させて翻訳品質の差異を調べた．評価方法は，「内容がより正確に伝わるようになったか」を判断基準とし，「翻訳品質の向上／低下／変わらない」を人手で評価した．表４に集計した結果を示す．「差分無」は翻訳文自体が変わらなかったものを示す．

この結果，全体として３割程度の文で訳文品質の向上が確認できた．内訳としては，複数文がある引用，補足情報，読みの括弧表現が含まれる文での翻訳品質の向上が目立った．逆に引用・強調の括弧表現では，単に従来の文に括弧文字が付与されただけで，文全体として翻訳品質は「不変」であることが多かった．

6 まとめ

機械翻訳に必要となる括弧表現の分類と，それに基づく自動判定手順および括弧表現に対する分割翻訳・係り受け制限・括弧表現削除の操作について述べた．アプリケーション上での効果を測るため，分類の判定実験および翻訳の品質評価実験を行い，双方とも括弧処理の十分な効果を確認できた．

本稿で取り上げた括弧文字以外にも，文書内にはHTML/XMLタグや太字，アンダーラインのような文字列の修飾といった，「機能文字」に相当する情報が埋め込まれている．今後は，このような言語外情報についても意味分類を整理し，同様のアプローチでシステムの性能向上を追究していきたいと考えている．

参考文献

[1] 白井，矢部，松尾，西垣，大山．“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-9，2-31,32，1996.

[2] 荻野．“リストのラベルとして使われる丸括弧とリストの範囲．” 計量国語学第19巻第4号，pp.208-215，1994.

[3] 後藤，熊野，江原．“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回年次大会発表論文集 pp.35-38，2000.

[4] 久光，丹羽．“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．” 情処研報 NL-122-17，pp.113-118，1997.

[5] 八巻，大山，白井，横尾．“日英機械翻訳システムALT－J/Eの研究開発．” NTT R&D，Vol.46，pp.1391-1398，1997.