かぎ括弧で囲まれた表現の種類の自動判別

後藤功雄　熊野正　江原輝将
{igoto, kumano, eharate}@strl.nhk.or.jp
ＮＨＫ放送技術研究所

1　はじめに

日本語を英語に機械翻訳する場合において、１文の長さが長い場合には翻訳精度が下がるという問題がある。そこで機械翻訳の前処理として、文を短い複数の文に分けることが、翻訳精度の向上に有効である[1]。

また、ニュース原稿から放送用の字幕を作成する場合においても、短い文への分割が行われることがある[2]。

このような短文分割の手法の一つとして、かぎ括弧で囲まれた引用文を主文と分離する方法がある。引用部分を新たな１文としてまとめることで、元の文を短くすることができる。

しかし、かぎ括弧（“「”, “」”, “『”, “』”）は引用の他に単語を強調するなど、他の用途でも使われている。また、引用の中でも、文の分割のパターンが異なってくる場合や、文の分割が困難な場合がある。これらのように「かぎ括弧で囲まれた表現（以降「かぎ括弧表現」と略称する）」は、複数の種類に分けることができ、その種類に応じて処理を変更する必要がある。１文を複数の短い文に分けられる場合とそうでない場合がある。

本稿ではこのような「かぎ括弧表現」の分類について考察する。まず、ＮＨＫのニュース原稿を対象として、「かぎ括弧表現」の種類にどのようなものがあるかについて述べる。さらに、「かぎ括弧表現」の種類の判別を自動でおこなう手法について述べ、この手法を用いた実験の結果を報告する。

2　ニュース原稿に現れるかぎ括弧に囲まれた表現の特徴

文中にかぎ括弧が用いられる用途としては、大きく分けると「引用」と「強調」の２種類が考えられる。本稿では「引用」の定義としては「他者の発話や認識・感情などの表出をそのまま自分の発話の一部とすること」とする。ただし、発話のうち引用する部分が少なければ少ないほど「引用」であると同時に「強調」の意味合いが増し、他者の表現が自分の発話の中に構文的に取り込まれていく。そのように引用する部分が少なく、他者の表現が自分の発話の中に取り込まれているものは、文の分割が困難である。

そこで、「引用」として用いられている「かぎ括弧表現」の中で、それを手がかりとした文の分割が可能なものについて９８年のニュース原稿を用いて調べた。

「引用」であるかどうかを決める主な要因は、かぎ括弧部分が引用であることを示す言葉（「述べる」など）に係っていることである。ニュース原稿の特徴には、かぎ括弧よりも前に引用であることを示す述語があることはあまりなく、ほとんどが後にきているということがある。また、前に引用を示す述語がある場合には、その後のかぎ括弧を利用した文の分割は困難である。そこで、かぎ括弧の後の表現について調べた。その結果、ニュース文中の引用文を主文と分離して文の分割を行う事ができるのは、５つのパターンがあることが分かった。以下に各パターンと文例、分割例を示す。

ｉ：　かぎ括弧部分が係る叙述を示す述語があり、その後に文が続かない場合で、かぎ括弧部分を「次のように」で置き換えて、かぎ括弧部分をそのままもう１つの文とすることができる場合。

＜文例＞

また野田幹事長は、この後の記者会見で、平成十年度予算案について、「歳出の削減を目指した財政構造改革法は、景気をさらに冷え込ませるもので、それに縛られている予算案は間違っている」と述べました。

＜分割例＞

また野田幹事長は、この後の記者会見で、平成十年度予算案について、次のように述べました。「･･･」。

ii：　かぎ括弧部分が係る語が叙述を示す言葉では

ないが、叙述を示す言葉を補って、かぎ括弧部分をそのままもう１つの文とすることができる場合。
＜文例＞
今年一年間の見通しについて自動車工業界では、「アジア向けの輸出の落ち込みを、自動車の販売が好調なヨーロッパ向けなどが補う形で推移すると見られ今年並みの水準になるのではないか」としています。
＜分割例＞
今年一年間の見通しについて自動車工業界では、次のような考えを示しています。「・・・」。

iii：　かぎ括弧部分が係る叙述を示す述語があり、その後にも文が続き、１つの文を３つの文に分割することができる場合。
＜文例＞
また加藤幹事長は、「自民党も経営者側に配慮するだけの政党としてやっていくわけにはいかない。働く人々と直接話し合いながら、政策を進めていきたい」と述べ、全電通など労働組合と政策をめぐる協議を行っていきたいという考えを示しました。
＜分割例＞
また加藤幹事長は、次のように述べました。「･･･」。そして、全電通など労働組合と政策をめぐる協議を行っていきたいという考えを示しました。

iv：　かぎ括弧部分が係る叙述を示す述語がなく、その後にも文が続き、叙述を示す言葉を補って１つの文を３つの文に分割することができる場合。
＜文例＞
小泉厚生大臣は、「赤字国債は、現在でも大量に発行しており、特別減税のために、さらに国債を上乗せするというのは財政構造改革には結びつかず、無責任な政策だ」として、景気対策のため新たに恒久的な減税を求める野党側の姿勢を批判しました。
＜分割例＞
小泉厚生大臣は、次のような考えを示しました。「・・・」。そして、景気対策のため新たに恒久的な減税を求める野党側の姿勢を批判しました。

v：　かぎ括弧部分が係る言葉が叙述を示す名詞であり、かぎ括弧部分を「次のような」で置き換えて、かぎ括弧部分をそのままもう１つの文とすることができる場合。
＜文例＞
さらに、会談では、政府・自民党の法案は、大手の金融機関が破綻した場合には対応できないなど問題点が多く、十分な審議が必要だという認識で一致し、「自民党が目指している来月の小渕総理大臣の訪米までに成立をはかれるような情勢ではない」という意見も出されました。
＜分割例＞
さらに、会談では、政府・自民党の法案は、大手の金融機関が破綻した場合には対応できないなど問題点が多く、十分な審議が必要だという認識で一致し、次のような意見も出されました。「・・・」。

3　かぎ括弧の種類の定義

かぎ括弧で囲まれた表現を調査したところ、２章で述べた分割の手がかりとなる５種類の引用表現の他に、３つの種類があることが分かった。そこでかぎ括弧表現を、以下の７種類に分類することとした。
○　種類１　２章のｉに分類されるもの
○　種類２　２章のiiに分類されるもの
○　種類３　２章のiiiに分類されるもの
○　種類４　２章のivに分類されるもの
○　種類５　２章のｖに分類されるもの
○　種類６　引用的意味を含むがそれを手がかりとした文の分割が困難、または不適当な場合。
＜文例＞
これに対する自見郵政大臣の意見は、ＮＨＫの予算を「おおむね適当なもの」と認め、引き続き事業運営の刷新や効率化を徹底するとともに、放送のデジタル化に向けて、先導的な役割を果たすべきだとしています。

○　種類７　文中の言葉を単に強調する目的だけのためだけで、引用的意味を含まずにかぎ括弧が用いられている場合。
＜文例＞
この「のど自慢」の模様は四月十二日に総合テレビでお伝えするほか、四月五日のＮＨＫスペシャルでもお伝えする予定です。

○　種類８　社会調査等の統計を取るためのアンケートに対する回答や解答としての選択肢にかぎ括弧が用いられている場合。
＜文例＞

また今回の減税が今年限りの特別減税となっていることについて、▼「来年度以降も減税を続けるべきだ」と答えた人が四十九％で最も多かったのに対して、▼「今年度限りの特別減税でよい」が十％▼「どちらともいえない」が三十五％でした。

種類８は、意味的には種類６に分類されるが、これはニュース文に独特のパターンであり、よく用いられるため、別の種類とした。

4　各種類における特徴の抽出

９８年のニュース文からかぎ括弧に囲まれた表現を約２万件を抽出してコーパスを作成し、各種類の特徴を調べた。

○　引用を示す特徴抽出

引用を示す表現のほとんどはかぎ括弧の直後に「と」または「などと」が続くことが分かった。しかし、種類７のかぎ括弧の直後にも続く場合があり、それだけでは分類することはできない。そこで、表１に示す文の構成を判別基準とした。ここで、丸括弧は必須でないものを示す。

また、各種類別についてコーパスから各種類の特徴を示す文字を抽出した。この特徴の基準を表２に示す。また、引用である「かぎ括弧表現」の末尾は、多くの場合ある特定の述語であることが多く、最後の１文字だけでも引用としての特徴を備えているので、「かぎ括弧表現」の末尾の１文字について全てのパターンを抽出し、これを引用の特徴とする。(パターン数７３個)

さらに、「かぎ括弧表現」の内部に句読点が含まれる場合のほとんどは引用であることが分かった。そこで、句読点の有無も引用の特徴とする。

○　引用的意味を含まない強調を示す特徴抽出

「かぎ括弧表現」の直前の１文字と直後の１文字について調べたところ、ほとんどの場合、「引用的意味を含まない強調表現」にのみ用いられている固有

判別基準	特徴を示す文の構成
種類1	「・・・」（など）と（形容詞等の叙述述語に係る言葉）＜叙述述語＞。
種類2	「・・・」（など）としています。
種類3	「・・・」（など）と（形容詞等の発話述語に係る言葉）＜叙述述語＞、＜続きの文＞。
種類4	「・・・」（など）と（して）、＜続きの文＞。
種類5	「・・・」（など）という＜叙述名詞＞・・・。

表１　文の構成による判定基準

判別基準	特徴を示す文字列のパターン	パターン数
種類1	かぎ括弧の後の発話を示す文字のパターン全て	312
種類2	かぎ括弧の後に続く文字のパターン全て	4
種類3	かぎ括弧の後の発話を示す文字のパターン全て	440
種類4	かぎ括弧の直後の「と」または「などと」の後につづく「して、」といった種類4の特徴を示す文字のパターン全て	21
	種類4を示す特徴的な＜続きの文＞を全て	242
種類5	かぎ括弧の直後の「と」または「などと」の後につづく「いう」といった種類5の特徴を示す文字のパターン全て	6
	かぎ括弧の後の発話を示す文字のパターン全て	375

表２　種類１～５の特徴を示す抽出文字

	種類1	種類2	種類3	種類4	種類5	種類6	種類7	種類8	Precision(%)
判別種類 1	898	5	6	5	0	1	0	1	98.03
判別種類 2	0	33	0	1	0	0	0	1	94.29
判別種類 3	13	0	763	4	0	3	0	9	96.34
判別種類 4	2	0	2	120	7	6	5	0	84.51
判別種類 5	0	0	1	0	126	13	7	0	85.71
判別種類 6	41	0	28	10	13	13	1	0	12.26
判別種類 7	0	0	1	0	5	9	1071	14	97.36
判別種類 8	0	0	0	0	0	0	13	54	80.60
判別不能	50	0	52	12	37	68	175	21	
合計	1004	38	853	152	188	113	1272	100	
Recall(%)	89.44	86.84	89.45	78.95	67.02	11.50	84.20	54.00	

表3　実験結果

の文字があることが分かった。その文字は以下の通りである。

・直前の1文字　:「の」「る」
・直後の1文字　:「を」「の」「で」「に」「や」

また、ほとんどの場合に、種類7と種類8にのみ用いられている直後の1文字は以下の通りである。

・直後の一文字　:「は」「が」

○　アンケートの回答の選択肢・回答の特徴抽出

この種類の場合は、かぎ括弧の直後に「が？%」や「は？人」といった特徴を示す言葉が続く場合が多い。そこでかぎ括弧の後に続く言葉のパターン全てを抽出し、数字を示す部分を数字であれば何でも可であると変更したパターンの集合を作成し、これを特徴とした。

上記の特徴を用いてかぎ括弧表現の自動分類を行うアルゴリズムを構成した。

5　実験

4章で述べたニュース原稿の98年のデータから抽出したルールセットを用いて、99年1月の分類「政治」「経済」「社会」「国際」のニュース原稿1ヶ月分全てのかぎ括弧表現を含む文をテストデータとして、かぎ括弧表現の自動分類アルゴリズムにより、かぎ括弧表現の自動判別を行った。実験結果を表3に示す。全体の Pricision は 93.13%、Recall は 82.74%であった。種類6は、特徴的なパターンがなく、引用の特徴を備えているが種別1～5に判別されなかったものを種類6と判別したため、判別精度が低くなっている。

6　おわりに

本稿では、かぎ括弧の種類を表層的な情報のみで自動判別する手法について述べ、その有効性を評価実験により確認した。

今後は、パターン集合の生成の自動化や、各パターンの持つ確実さの情報を用いたパターンの適用などを行う必要がある。決定木を用いた判別[3]とも比較したい。また、この判別結果を基にして文の自動分割を実際に行ったり、かぎ括弧部分の用途に合わせた適切な翻訳処理の制御、また、日本語解析段階への判別結果の利用についても検討していく予定である。

参考文献

[1] 金 淵培，江原 暉将：日英機械翻訳のための日本語長文自動短文分割と主語の補完，情報処理学会論文誌，Vol.35，No.6，pp.1018-1028 (1994).
[2] 福島 孝博，江原 暉将，白井 克彦：短文分割の自動要約への効果，自然言語処理，Vol.6，No.6，pp.131-145 (1999).
[3] 張 玉潔，尾関 和彦：決定木による日本語長文の短文分割，自然言語処理，Vol.7，No.1，pp.13-30 (2000).