間接引用から直接引用への変換

池崎 健一郎　江原 暉将　加藤 直人
NHK放送技術研究所ヒューマンサイエンス
{ikezaki,eharate,katonao}@strl.nhk.or.jp

1.はじめに

放送ニュースの読み原稿として使われるニュース記事は、その用途から引用文が非常に多く、その中にはかぎ括弧（“「”と“」”）を用いない、引用文（間接引用文）もかなりの割合で含まれている。この間接引用文を、係り受け解析させてみると、主語と述語が離れているため、間違った解析結果が多く発生する。しかし、間接引用部分をかぎ括弧で囲むこと（直接引用文に変換する）でこのような誤りを減らせる可能性がある。そこで、間接引用文を直接引用文に自動的に変換するツールを考案した。本稿ではこの間接引用→直接引用変換の手法について述べ、さらに評価を行ったので報告する。

2.間接引用→直接引用変換

2.1 基本的考え

間接引用文を直接引用文に変換するには、間接引用文をデータから探し出し、かぎ括弧の挿入に適切な位置を見つけ、しかるべき後処理を加えるという三段階の処理が必要となってくる。このうち、最も重要なのは、かぎ括弧の適切な挿入位置の推定である。そこで、間接引用の特徴的なパターンを探してみると、いくつかの決定的なパターンが存在することがわかった。

このパターンを
①：話者特定部(話者が特定できる文節)
②：引用開始部(発話開始直前の文節)
③：引用終了部(発話終了直後の文節)
④：引用表明部(引用文であることを示す文節)
の四つに分類した。

図１は、2000年１月のニュース記事データベースに出現した間接引用文の一例である。

前述の間接引用パターンと照らし合わせると、①の話者特定部が、『宮沢大蔵大臣は、』にあたり、以下、②の引用開始部が『記者会見で、』、③の引用終了部が『という』、④の引用表明部が『示しました。』となる。

そして②の直後から③の直前までをかぎ括弧で囲むと、図２に示すように間接引用文を直接引用文に変換することができる。間接引用文のそれぞれのパターンに属する言語表現例を図３に示す。

かぎ括弧で囲むことにより、係り受け関係が、かぎ括弧の外と中で２つに分割されることになる。よって、図４と図５に示すように、係り受け解析結果の誤りを減らすことができる。

2.2 変換アルゴリズム

実際には、前節に示す四つのパターンすべてが

宮沢大蔵大臣は、きょうの閣議のあとの記者会見で、悪化している国の財政について平成13年度予算では、国債発行額の増加傾向に歯止めがかかるという見通しを示しました。

図１：間接引用文の一例

宮沢大蔵大臣は、きょうの閣議のあとの記者会見で、「悪化している国の財政について平成13年度予算では、国債発行額の増加傾向に歯止めがかかる」という見通しを示しました。

図２：「」挿入後

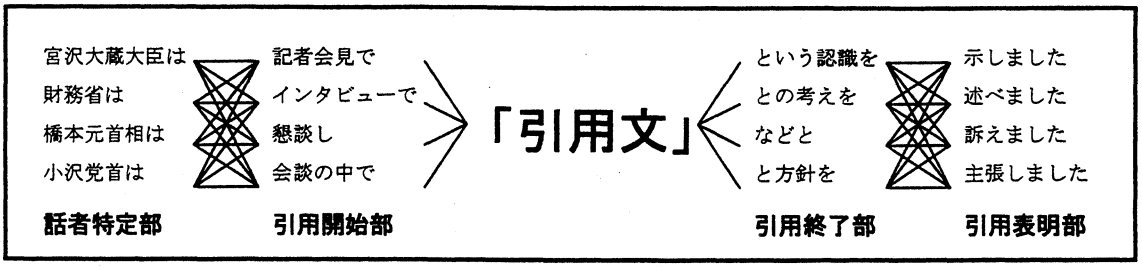

図３：間接引用文の言語表現の一例

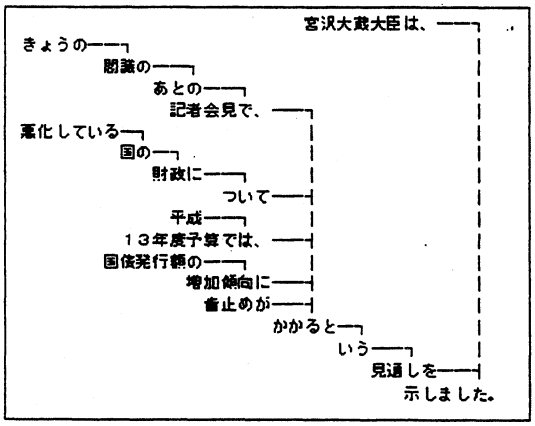

図４：図１の係り受け解析結果

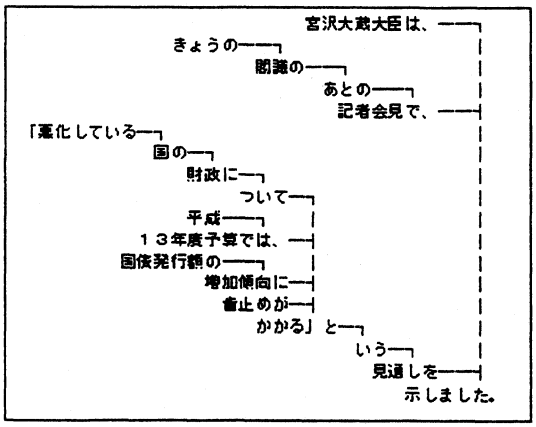

図５：図２の係り受け解析結果

出現することはまれで、④単独、①・②の二つの組み合わせ、①・③・④の組み合わせなど様々な場合がある。

かぎ括弧の挿入位置推定が、もっとも確実なのは全ての部分が出てくる場合であるが、それのみでは再現率が低くなってしまうので、パターンマッチの優先度を次のように定めた。

優先度１：④

引用文であるのだから、引用であることを示す動詞が文中になければならない。そこで、はじめに④を文中から探し出すことにした。

優先度２：①

引用において、「どこが」、「だれが」という情報は必須である。引用変換を行う主要な目的の一つは係り受け関係をはっきりさせることであり、無意味なかぎ括弧の挿入は避けるべきである。しかし、「それによりますと」「これまでの調べによりますと」など、文中に主語を含まない引用文があるので、これら特殊な表現は例外規則として記述しておくことにする。また、話者特定部は固有名詞を含むことが多いため、他の部分に比べて種類が膨大なものとなり、規則にすべてを記述することは不可能である。そこで、最初から三文節までの間に、係助詞“は”を持つ文節を話者特定部と推定するモジュールを作成した。

優先度３：②、③

②と③は、必須条件ではない。②や③が存在しない場合、①が②の役目を、④が③の役目を兼ねることになる。つまり、①の直後から④の直前までが引用部分となる。これらは、①と④のパターンマッチが終了した後に検索する。

以上をまとめると、変換アルゴリズムは図６の

ようになる。

step1：文中にかぎ括弧で囲まれた発話がある文章を探す。あれば終了する。
step2：step1 に④の部分が存在しているかを検索する。なければ終了する。
step3：step2 の結果に①の部分、もしくはあらかじめ記述してある例外規則があるかを検索する。存在すれば step5 に、なければ step4 に行く。
step4：一番前から三番目の文節まで係助詞“は”を持つ文節を探してゆき、あればその文節を①とし、なければ終了する。
step5：①“「”引用候補文“」”④の形を引用候補文とする。
step6：引用候補文のなかに、前から見て②はないか、後ろから見て③はないか探す
step7：もし、step6 があれば、“「”の位置を②の直後に、“」”の位置を③の直前にずらして出力する。なければ step5 の形をそのまま出力する。

図 6：変換アルゴリズム

3.評価実験

3.1 間接引用パターンの言語表現の収集

今回、NHK ニュース記事データベースの 1998 年 11 月分のデータ 3642 記事、19044 文を用いて、間接引用パターンの言語表現を人手で収集した。その結果、計 78 例を得た。

さらに 19044 文中に出現した直接引用文を利用してパターン規則を追加した。この際には、直接引用文から、かぎ括弧直前の文節と直後の文節をそれぞれ抽出し、出現頻度 10 以上の文節をパターンとして認定した。こうして、自動抽出した言語表現例は計 73 例となった。人手と自動的に抽出した言語表現例は合計で 151 例である。これらをパターン例として登録して、前章で述べた変換アルゴリズムを用いて、間接引用→直接引用変換プログラムを作成した。

3.2 実験 1:間接引用から直接引用への変換精度

評価データは、2000 年と 2001 年の政治記事と経済記事のデータの中から間接引用が含まれている文を各 200 例、計 800 例用意した。評価データすべてに人手でかぎ括弧を付与して正解データとし、本手法でどれだけ正確にかぎ括弧の挿入位置が推定できるか実験を行った。

評価は再現率と適合率で行った。再現率および、適合率は以下のように算出する。

再現率(*Recall*)

$$=\frac{\text{変換プログラムの出力と正解データの一致数}}{\text{評価データ数}}\times 100$$

適合率(*Precision*)

$$=\frac{\text{変換プログラムの出力と正解データの一致数}}{\text{変換プログラムが出力したデータ数}}\times 100$$

その結果を表 1 に示す。

		再現率	適合率
2000	政治	68.5%	87.2%
	経済	47.5%	87.9%
2001	政治	70.5%	89.8%
	経済	60.5%	94.5%

表 1：間接引用から直接引用への変換精度

表 1 を見ると、適合率はジャンルの違いによる影響は少なくほぼ 90%前後であった。しかし、再現率に関しては、精度が多少落ち、またジャンルによって 10%以上の差が生じる現象が見られた。

ジャンルによって再現率の精度が違う理由としては政治が特定の人物・政党・話題に偏りやすく、特定のパターンマッチングが比較的しやすいのに比べ、経済は話題が人物・会社等に偏ることが無いため、再現率が下がってしまったことが挙げられる。

再現率の絶対値が低かった理由は、一ヶ月間だけの狭い範囲で収集したため、言語表現が足りなかったからことが考えられる。また、主語判定の誤りやパターンマッチのルールが矛盾する場合もあった。図 7 は、変換に失敗した一例である。

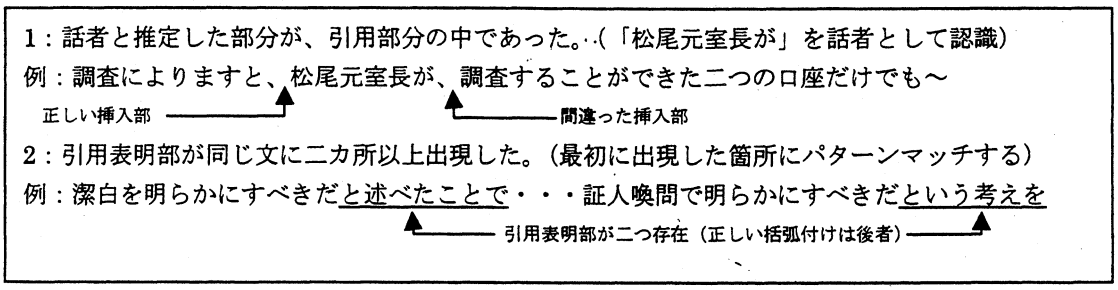

図 7：変換誤りの例

引用表明部が二カ所以上出現したり、話者と推定した部分が引用文の中であったことがあった。

3.3 実験 2:主動詞に対しての係り受け精度

間接引用から直接引用に変換することにより、係り受け解析の精度が向上するかについても実験を行った。直接引用に変換した結果の出力文から 40 例をランダムで抜き出し、主動詞に対する各文節の係り受け関係が正しくなされているか評価した。評価ツールとして形態素解析に JUMAN3.61[1]を、係り受け解析に KNP2.06[2]を使用した。

評価には、次のような評価基準を設けてランク付けを行った。

A：解析誤りがすべて正された。
B：解析誤りは一部あるが、精度は向上した。
C：変化しなかった。
D：直接引用自動変換により、精度が悪化した。

その結果を表 2 に示す。

評価基準	A	B	C	D
データ数	31	3	4	2

表 2：係り受け解析の精度

表 2 を見ると、係り受け解析の向上率は 85%（『向上した』に A と B を含める）であり、目的の一つであった係り受け解析の精度向上は果たせたといえる。評価基準 D の 2 例はいずれも間接引用→直接引用変換の失敗が原因であり、間接引用から直接引用の変換が、係り受け解析に与える副作用はさほどないと言える。

4.おわりに

間接引用文を直接引用文に変換する手法を提案した。この手法では再現率は 60%前後、適合率は 90%前後であり、言語表現例を増やせばこれをさらに上回ることが予想される。

また、直接引用文に変換することによって、係り受け関係を明確化でき、解析率向上にもつながることを検証した。

さらに本手法は字幕制作や機械翻訳における自動短文分割[3]にも応用できる。

今後としては、より広いジャンルの比較検討、言語表現の収集、アルゴリズムの効率化などによって、再現率・適合率の向上をはかっていきたい。また、適用範囲を広げて連体修飾節抽出など、文から意味を細分化して切り出す手法に進めていく予定である。

参考文献

[1]黒橋 禎夫、長尾 真(1998)：日本語構文解析システム JUMAN ver3.61 http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[2]黒橋 禎夫(1998)：日本語構文解析システム KNP ver2.0.b6. http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html

[3]江原ほか(2000)：聴覚障害者向け字幕放送のためのニュース文自動短文分割 自然言語処理 138-3