話し言葉における引用部分を抽出するシステム*

横山　晶一†　　坂ノ上　剛‡　　橋本　力†
（†山形大学大学院理工学研究科）（‡山形大学工学部）
yokoyama@yz.yamagata-u.ac.jp

１．はじめに

話し言葉は、時系列的な発話になるために、書き言葉とは違う性質を持つことはよく知られている。たとえば、言い淀み、句の挿入、倒置などである。

話し言葉のデータとしては、日本語話し言葉コーパス[1]がよく利用されており、このデータを用いて多くの研究が行われている（たとえば[2]）。

我々の研究グループでも、主として自然言語処理の観点から、このデータベースを用いていくつかの研究を行ってきた[3,4]。

本稿では、話し言葉の書き起こしデータを用いて、話し言葉に含まれる引用部分の抽出を試みた結果を報告する。話し言葉における引用部分の詳細な定義は、次節で行うが、たとえば次の文でカギカッコでくくられた部分である。

(文 1) 私は「嬉しい」と思った。

この文では、「嬉しい」の部分が、自分の感覚を言い表しており、引用と判断される。ここで抽出する引用部分とは、自分の思考、他人の発話などで、書籍や慣用句を引用して述べるといったものは含んでいない。このような引用部分を抽出することによって、書き言葉と同じように、どの部分が話者自体の発話で、どの部分が引用かを区別することができ、書き起こし文の読みやすさにつながると考えられる。

本研究では、244 の文について、引用部分がどのように現れるかを調査し、それに基づいて引用部分を自動的に抽出するシステムを作成した。評価用として、上記とは異なる 185 の文に対してこのシステムを適用したところ、元の文とほぼ同じ約 60%の精度が得られた[5]。

* Extraction of quotation parts from spontaneous speech, by YOKOYAMA Shoichi , SAKANOUE Takeshi, and HASHIMOTO Chikara, Yamagata University

２．話し言葉と書き言葉

日本語話し言葉コーパスの完成によって、大規模な話し言葉データが集められ、研究に供されていることは周知の通りである。これに基づく研究も多く行われており、音声認識や音声と言語の観点から論じられている。

我々のグループでは、即時的、時系列的な話し言葉であっても、何らかの文法的側面が存在しているのではないかという観点から、いくつかの研究を行ってきた。

具体的には、話し言葉の中から文法的な側面を少数のデータについて調査したり[6]、話し言葉に多く含まれる、並列でない助詞「とか」について調査、分類したり[3]、話し言葉の倒置現象を分類して、倒置表現を修正する簡単なシステムを構築したりした[4]。

本研究では、書き起こされた話し言葉を読みやすくするために、話し言葉の中に含まれる自分の思考や、他人の言動を引用した部分を抽出するシステムについて述べる。

話し言葉における引用文抽出の研究については、すでに、引用文や挿入節を検出することによって、話し言葉の係り受け解析の精度を向上させる研究がある[7]。本研究では、主として形態素解析を用いて、引用部分のみを抽出することを目的とし、挿入節は取り扱わなかったので、実験に際して比較することはしていない。

３．引用文の定義と資料

３．１．引用文の定義

引用文を抽出するためには、それぞれ文の始まりと終わりを抽出する必要がある。しかしながら、話し言葉における引用文は、書き言葉に比べて非常に曖昧であるので、ここでは次のように定義し、それに準じた細かい規則を定める。

(a) 引用文の中身が口語である
(b) 引用文の中身が口語ではないが、内容が過去である
(c) 複数の候補があったときは、最も短い範囲を引用文とする

この定義に従うと、上述の文も含めて、引用が次のように定義される。

(文 1) 私は「嬉しい」と思った。　○
(文 2) 私は嬉しいと思う。　×
(文 3) 私は「嬉しいな」と思う。　○

文 1 は定義(b)、文 3 は定義(a)によってカギカッコの部分が引用と判定されるが、文 2 は引用を含まない文と判定する。

また、定義(c)については、次のようになる。

(文 4) 従兄弟が犬を飼い始めたので、私も「犬が飼いたい」と言った。

この文では、引用の始まりについて、「私も…」、「犬を…」、「従兄弟が…」という他の可能性が考えられるが、上の定義に従って、最も短い上記の部分を引用とする。

３．２．資料

日本語話し言葉コーパスの中で、引用文が多く含まれている 24 個の書き起こしデータファイルを使用する。そのうち 12 を訓練用で、残りの 12 を評価用とした。1 個のファイルには、約 4,000 文字含まれている。コーパスには、読み仮名などの不要な情報が含まれているので、これらを除外し、ポーズの部分を、読点や句点に変換する。フィラー等は引用文として機能することもあるので除外しない。

４．引用文抽出アルゴリズム

引用文の抽出は、まず、終りの検出から行う。引用文の終わりには、いくつかの検出要素が存在するが、始まりには特に著しい要素がないためである。

４．１．引用文の終わりの抽出

(a)「って」、「と」、直前が名詞でない「とか」を検出した場合

この場合には、次の条件を判定して、引用文かどうかを判断する。

(a1) 直前が感動詞またはフィラーの時、引用文と判定する
(a2) 語尾を抜き出し、「だよ」、「かな」など、口語に現れる語尾（15 個程度）かどうかを判定し、そうならば引用文と判定する
(a3) 文末が「た」で終わっている場合、文全体を過去と判定する。また、その時の動詞が「言う」、「思う」などの引用に用いるものかどうかを判定（現在 12 の動詞）し、両者を満足する場合を引用文と判定する

(文 5) 母はいつも「あんたはバカだね」と言う　○
(文 6) 母はいつもあんたはバカだと言う　×

上の文で、文 5 は、「だね」という口語形を含んでいるので引用文を含むが、文 6 ではそうでないので、引用文を含まないとする。

(b)「という」、「っていう」を検出した場合

(b1) 直前が感動詞またはフィラーの場合、引用文の終わりと判定する
(b2) 上記(a2)と同様に口語判定を行い、口語ならば引用文と判定する

(c)「と言うか」などの言い直しの前は引用と判定しない

(文 7) 知り合いと言うか、友達なんですけど。

(d) 口語表現を含まず、動詞が基本形の場合は、引用文でないことが多いので除外する

(文 8) ジャイアンツと言う球団があるんですが

４．２．引用文の始まりの検出

引用文の終わりと比較して、始まりの部分は余り明確ではない。たとえば、文頭、名詞＋格助詞または係助詞の直後、接続詞または接続助詞の後などである。そこで、次のようないくつかの規則を設定する。これはこのシステムで用いたものの代表的なものである。

(a)「名詞」＋「格助詞または係助詞」＋「読点」の直後から引用文とする

(文 9) 老人が、「写真を撮ってあげるよ」と言って

(b)「人」[8]＋「に」で、動詞が「言う」の場合、その直前から引用文とする

(文 10) 結婚した友人に「あなたの奥さんとても奇麗よね」と言って

(c) 引用文の直前が「もの」＋「格助詞」の場合、「もの」の前まで始まりをさかのぼらせる

(文 11)「これが、究極の水割りだよ」と言われたことと、

この文では、「これが、」の部分まで引用に含める。「人」や「もの」は、意味判定[8]を行っている。

(d) 接続詞の直後から引用文とする

(文 12) でも、「何か違うな」って思って

(e) 直後が動詞ではない接続助詞の直後から引用文とする

(文 13) 嫌だったんですけど「京都に行ってみよっかな」とかいう気になって

(a)～(d)のアルゴリズムを用いると、「行って」と「みよっかな」の間にカギカッコの始まりが来るが、上記の規則で、上に記載した場所にカッコの始まりが来る。

(f) 近くに「じゃ」や「いや」があった場合、その直前から引用文とする

(文 14) だから、「じゃ、犬にして、猫はやめようかな」って思って

(g) 直前が感動詞またはフィラーだった場合、その語を引用文とする

(文 15) でも行くと「（F あーあ）」と言われるから

上の文の F はフィラーを示す。

(h) 上記のすべてに当てはまらない時、文頭を引用文の開始とする

５．引用文抽出システム

前節の規則に基づき、引用文抽出システムを作成した。システムの流れは次のようになる。

(1) 話し言葉コーパスから文を入力
(2) 不要部分の除去や句読点の付加
(3) 文を形態素解析システム「茶筌」[9]で形態素に分ける
(4) 話し言葉用に一部の誤りを修正
(5) 引用文の終わりを検出
(6) 引用文の始まりを検出
(7) 出力

「茶筌」は、もともと話し言葉には対応していないので、口語的な言い方の場合、解析を誤る場合がある。たとえば、「もう二度と来ないかなって思ってたんですけど」を茶筌で解析すると、「来／ない／か／なっ／て」のように区切られる（「か」：副助詞、「なっ」：動詞、「て」：接続助詞）。このような、引用に関わる部分の誤りを、パターンで「来／ない／かな／って」と書き換えるのが、上記(4)の部分である。現在 15 のパターンで修正を行っている。

図 1 に、実際の抽出過程を示す。文 16 に対して、以下の順番に従って、抽出が行われている。番号の後には、前節で示したアルゴリズムのどの部分が使われたかを示している。システムでは、カギカッコは二重のもの“『 』”を用いる。

(文 16) 私は母に、じゃ、明日は晴れて遊びに行けるんだと言いました。

(1) 茶筌による形態素解析
(2) 引用文の終わりの検出開始
(3) 引用の助詞「と」を検出
(4) 語尾が「だ」なので、口語ではないと判断（終わり a2）
(5) 文末に注目し、過去を表わす助詞「た」を検出（終わり a3）
(6) 引用の助詞「と」の直後の動詞「言う」を抜き出し、引用に使用される動詞と判断（終わり a3）
(7) 引用の助詞「と」の前に終りのカギカッコ“』”を挿入

→ 私は母に、じゃ、明日は晴れて遊びに行けるんだ』と言いました。

(8) 引用文の始まりの検出開始
(9) 「晴れて」の接続助詞「て」を検出（始まり e）
(10) 接続助詞「て」の後にカギカッコ挿入

→ 私は母に、じゃ、明日は晴れて『遊びに行けるんだ』と言いました。

(11) さらにさかのぼって、接続詞「じゃ」を検出（始まり d）
(12) 「じゃ」があるので、上のカギカッコの始まりを消して、「じゃ」の前にカギカッコを挿入（始まり f）

→ 私は母に、『じゃ、明日は晴れて遊びに行けるんだ』と言いました。

(13) 上記を出力して終了

図 1　システムの解析例

上の図と、システムの流れに示した手順で明らかなように、入力文に対して、まず形態素解析を行い、引用文の終わりが存在するかどうかを判定する。

引用文の終わりが検出されたら、そこからさらにさかのぼって、引用文の始まりを検出する。上の例では、一度は誤った始まり部分（図の(10)のところ）にカギカッコをつけているが、さらにさかのぼることによって、「じゃ」を検出し、その前にカギカッコの始まりを移すことによって、この誤りを修正している。

表１　訓練用 12 個のファイルに対する抽出結果

正しく検出	開始位置の誤り	引用文以外の検出	未検出	引用文合計
148	67	23	29	244

表２　表１の評価

再現率	適合率	F 値
61.1%	62.6%	61.8%

表３　評価用 12 個のファイルに対する抽出結果

正しく検出	開始位置の誤り	引用文以外の検出	未検出	引用文合計
113	55	12	17	185

表４　表３の評価

再現率	適合率	F 値
61.1%	62.8%	61.9%

６．システムの評価

このシステムを評価するために、次のように、再現率、適合率、F 値を定めた。

・再現率 = 正しい引用文の抽出数／正しい引用文の総数
・適合率 = 正しい引用文の抽出数／抽出総数
・F 値 = 2×正しい引用文の抽出数／（抽出総数＋正しい引用文の総数）

最初に、12 個のファイルに基づいて、アルゴリズムを作成した。この 12 個のファイルに対してシステムを適用した結果を表１に、その評価を表２に示す。全体としてだいたい６割程度である。

次に、アルゴリズムを作成したのとは別の 12 個のファイルに対する同様の結果を表３、表４に示す。これらの表から分かるように、アルゴリズムを作成したファイルとは異なるファイルでもほぼ同様の結果（F 値ではわずかに勝っている）が得られており、システムの有効性が示されている。

７．問題点と今後の検討

現在のシステムでは抽出できなかったり、誤った抽出を行うケースを以下に示す。カギカッコは人間が付したものである。

(a) 倒置などによって、語尾が口語でなくなる

(文 17)「（F え）どうなっちゃうんだろうね この先」って予想するのが

(b) 話し言葉の特徴である冗長性のために文が途切れず、過去と現在の話が混在する

(文 18)「それじゃ、一緒に行こう」って言ったんですけど、今も行ってません。

(c) 引用文が長すぎる

(文 19) 友達は、「それは、きっと忙しいから、付き合うことはできないかもしれないけど、頑張れば付き合えるかもしれないから、もう少し頑張んなよ」って言って

(d) 引用文が複数の文にまたがる

(文 20) 医者は、「危険な状態です。ですので、毎日病院にきてくださいね」と言った。

今後は、まず、比較的短い文の抽出方法をさらに改良した後、長い文の抽出法を改善していく予定である。

参考文献

[1] 日本語話し言葉コーパス, Vol.1～18、国立国語研究所(2004)

[2] 話し言葉の科学と工学ワークショップ講演予稿集(2001～2004)

[3] 栗田侑美、横山晶一：話し言葉における簡略化と助詞「とか」の分類、情報処理学会東北支部平成 17 年度第５回研究会 (2006) A3-1

[4] 坂口正紘、横山晶一：話し言葉における倒置表現の分類と修正、情報処理学会東北支部平成 18 年度第６回研究会(2007) B2-2

[5] 坂ノ上剛：話し言葉における引用文抽出システム、山形大学工学部卒業論文(2008)

[6]山口裕之：話し言葉の文法構築に関する研究、山形大学工学部卒業論文 (1999)

[7] Ryoji Hamabe, Kiyotaka Uchimoto, Tatsuya Kawahara, Hitoshi Isahara: Detection of Quotations and Inserted Clausee and its Application to Dependency Structure Analysis in Spontaneous Japanese, Proceedings of the COLING/ACL (2006) pp.324-330

[8] 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）：日本語語彙大系、岩波書店 (1997, 1999)

[9] 形態素解析システム　茶筌、奈良先端科学技術大学院大学松本研究室、http://chasen-legacy.sourceforge.jp/