社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－SLP－64 (25)
2006／12／21

話し言葉における引用節の自動認定および引用符の付与

浜辺 良二† 内元 清貴†† 河原 達也† 井佐原 均††

† 京都大学大学院 情報学研究科 〒606-8501 京都市左京区吉田本町
†† 独立行政法人 情報通信研究機構 〒619-0289 京都府相楽郡精華町光台 3-5

あらまし 講演のような話し言葉の書き起こしや音声認識結果を，講演録などのアーカイブとして二次利用する場合，文章として適した形態にする必要がある．本研究では，話し言葉の中で発言の引用が行なわれている箇所に引用符を自動付与する手法を提案する．機械学習により，まず引用構造をとる節を自動認定し，それらに引用符が必要かどうかを判定する．引用構造の認定では，表層表現や音響的特徴に加え，係り受け情報を利用することで認定精度の改善を図る．引用符付与の判定においては，学習の際に新聞記事コーパスから得られる情報をあわせて利用する．『日本語話し言葉コーパス (CSJ)』に対して，引用符付与の基準を定めた上で本手法の実験的評価を行なった．

キーワード 話し言葉，引用節，係り受け解析，機械学習

Detection of Quotations and Insertion of Quotation Marks in Spontaneous Japanese

Ryoji HAMABE†, Kiyotaka UCHIMOTO††, Tatsuya KAWAHARA†, and Hitoshi ISAHARA††

† School of Informatics, Kyoto University Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan
†† National Institute of Information and Communications Technology
3-5, Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan

Abstract Transcriptions and speech recognition results of public speaking include many expressions peculiar to spoken language. Thus, it is necessary to transform them into document style for practical use of them. We focus on detecting quotations and enclosing them in quotation marks as written text. Quotations are detected with SVM-based text chunking method that considers information on morphemes, acoustic features, and dependency structures. Then, it is determined whether or not they need to be enclosed in quotation marks by machine learning method using the corpora of spoontaneous speech and newspapers. We defined how to classify quotations and evaluated our method on *the Corpus of Spontaneous Japanese (CSJ)*.

Key words spontaneous speech, quotation, dependency structure analysis, machine learning

1. はじめに

近年，音声認識技術の進展により，ディクテーションだけでなく講演や会議などの話し言葉を対象とした音声認識の研究が盛んとなっている．このような話し言葉の書き起こしや音声認識結果は，文書に用いる書き言葉とは異なる点が多く，テキストとしての可読性がよくない．そのため，講演録や会議録などのアーカイブとして二次利用する際には，文章として適切な形態に整形する必要がある．

例えば，話し言葉においてはフィラーや言いよどみ，言い直しが多く発生するが，これらは書き言葉に現れない冗長な表現であるため，削除する必要がある．フィラー・言いよどみ・言い直しを自動検出するための手法がこれまでに提案されている [1],[2]．また話し言葉では句読点がなく文の区切りが明確でないため，文境界を推定するための研究も行われている [3]．

本研究では，話し言葉に対して発言の引用箇所に引用符を自動付与することを目的とする．新聞記事などの書き言葉では発言の引用箇所には引用符が付与されていることが多い．話し言葉では引用はポーズや声の調子によって示されるが，書き起こしではそれらの情報が欠落してしまうため，可読性が著しく低下する．

本研究で提案する手法では，引用符の付与対象となりうる引用構造を検出した後，それらに引用符が必要であるかどうかを判定する．引用構造の検出は，SVM を用いたテキストチャンキングによって実現する．その際に，自動推定した係り受け情報を用いることで精度の改善を図る．引用符の判定においては，機械学習を行うが，その際に新聞記事コーパスに付与されている引用符の情報をあわせて利用する．

本稿では『日本語話し言葉コーパス (CSJ)』[4] を用いて分析・評価を行なう．CSJ は学会講演や模擬講演などのモノローグを対象として収集・構築されたコーパスである．話し言葉においては文の定義が明確でないため，CSJ では文に相当する単位として様々な種類の節が定義されており，コアと呼ばれる一部の講演の書き起こしには，形態素・係り受け・節単位などの言語的情報が付与されている．

2. 話し言葉における引用

本章では，CSJ において引用構造をとる節に対して，引用符が必要かどうかの判定基準を定めた上で分類を行う．また引用符の自動付与における問題点について述べる．なお，新聞記事では発言の引用以外にも固有名詞や強調箇所にも引用符が付与されているが，これらは引用符をつけるかどうかの曖昧性が大きいため本研究では自動付与の対象外とする．

2.1 CSJ における引用構造

CSJ では引用構造を表す節として「引用節」というラベルが定義されている．しかし，引用節のすべてが発言の引用となっているわけではない．以下の例文では {} 内が引用節に相当する．

◇ {彼女は今後もし {長期で雇うんだ} としたらちょっとお勧めできないんですけど} と報告しました

「彼女は今後もし長期で雇うんだとしたらちょっとお勧めできないんですけど」は報告内容の引用であり，引用符を付与できるが，「長期で雇うんだ」は引用ではないので，引用符を付与すると不自然な文章になってしまう．このように，引用符を付与すべき引用節とそうでないものを区別する必要がある．

また，引用節に加えて，トイウ節・トカ節についても同様の引用構造をとりうる．以下に，トイウ節・トカ節による引用の例を示す．

◇ 「実は普通の会社に勤めたんですけども辞めることにしたんですよ」っていう風に言ったら

◇ 病院ですから凄く看護婦さんとかに怒られて「病院の中で走っちゃいけません」とか いつも怒られてたんですけれども

以降は引用節・トイウ節・トカ節をあわせて引用節と表現することにする．これらを対象として引用符が必要かどうかの判定基準を定める．

2.2 引用符付与の判定基準

本研究では，発言内容の引用とみなせる引用節に対して，引用符を付与するものと定める．CSJ コアの模擬講演 111 講演に含まれる引用節について，引用符を付与すべきかの判定を人手により行なった．発言の引用かどうかが曖昧な場合には，前後の文脈や表層表現から明確に判断可能な場合のみに引用符を付与する．例えば，以下の文は過去に引用符内の発言が行なわれた可能性もあるが，文脈からは判断できないため，引用とはみなさない．

◇ で彼はですね×「この先何やりたいか分からないけどとにかく旅をするんだ」という感じで来てる人で

今回分類を行った 111 講演には，引用構造となる節が 3,676 個 (引用節 1,895 個，トイウ節 1,335 個，トカ節 446 個) 含まれており，そのうち発言の引用箇所と判定したものは 534 個 (引用節 294 個，トイウ節 162 個，トカ節 78 個) であった．

2.3 直接引用と間接引用

引用は直接引用と間接引用に大きく分けることができる．ここでは，前節で得られた引用箇所に対して，以下の 2 つの基準で直接引用と間接引用の分類を行なった．

(1) 文体による分類

丁寧体 (です，ます) や終助詞 (ね，よ) で引用節が終わる場合，また以下のような命令文・依頼文・質問文・感嘆文などは直接引用とみなせる．

◇ で私は「コーヒー買いに行ってこい」とか言われちゃったりもして (命令文)

◇ で「一応病院に行って検査してください」とそういう風に言われて (依頼文)

◇ で看護婦さんがね「子供を抱いてみますか」って言ったんですけど (質問文)

◇ で私は「はい」ってもう直立不動で聞いてました (感嘆文)

（2） 人称・時・場所に関する語での分類

直接引用と間接引用では，人称・時・場所に関する語(「私，あなた」「今」「この，そこ」など)の指す実体が異なる．これらのように発話の場面に縛られる要素は「ダイクシス」(deixis)と呼ばれる．引用節内にこれらの単語が含まれる場合，その実体を文脈から判断することで直接引用と間接引用の分類ができる．以下の文では，「私」は話者に当たるので，間接引用である．

◇ で聞くところによると「私の住んでいる荒川区は二十三区で一番お年寄りの多い区だ」と聞きました

また，以下のように引用箇所における具体的な内容が「何々」「こう」「あんな」などと置き換えられている場合も，間接引用とみなせる．

◇ それで木にこう「この木は何々科の何々という木だよ」というような表示がされています

◇ お互いの学園祭に行って「こう いうことやってんだよ」とか「あんな ことやってんだ」っていう感じで盛り上がり

分類の結果，引用箇所534個のうち，直接引用が425個，間接引用が44個あった．また，以下のように，直接引用の要素と間接引用の要素が混在している場合も17個あった．

◇ そのことを彼女に聞いてみると「私達が行こうとしている場所はまさしくそこよ」という答えでした

(文脈から，引用符内の「私」は「彼女」でなく話者を指しており間接引用となるが，引用符は終助詞「よ」で終わるので直接引用の形式が混ざっている．)

さらに，(1)(2)の基準に該当する要素がなく，直接引用か間接引用かの判定が不可能なものは48個であった．これには，以下のように話者が特定されていない場合や，引用節が名詞節のように用いられている場合などが含まれる．

◇ 例えば「マンションの隣り同士が挨拶しない」とか「隣りの人は知らない」とか最近よく聞きますけど

◇ 病院に行って「クリーニングの薬にかぶれたんじゃないか」とかいうことを色々説明したところ

書き言葉では，間接引用には引用符をつけないのが一般的であるが，直接引用と間接引用を自動分類するには，上述のようにダイシクスの照応を解析する必要があり，非常に困難である．本研究では，発言の引用箇所全てを対象として引用符を付与するものとし，直接引用と間接引用の自動分類については，今後の課題とする．

2.4 引用の直後の表現

引用節が発言の内容を表す場合，引用節の後には特定の動詞が現れることが多い．例えば，「言う」は発言の引用を表すためによく用いられる動詞である．しかし以下の例文のように，引用節の後に特定の動詞が現れていても，引用節が必ずしも発言の引用を表しているとは限らない．

◇ ×「これはどういうことか」と言うとですね

◇ ×「手に入れる」と言っても決して高い金を払ったのではなく

このようなパターンをルールベースで記述するのは非常に困難であるため，本研究では話し言葉のコーパスから機械的に学習することによって，引用符の判定を行なう．

また，引用を表す動詞は「言う」の他にも，「話す」「告げる」「述べる」「書く」「頼む」「報告する」などが挙げられる．さらに「声をかける」などの連語や「言い残す」などの複合動詞も引用を示すために用いられ，その種類は非常に多い．これらの表現は，引用符の判定を行なう上で大きな手がかりとなる一方，使われる頻度の少ない表現もあり，話し言葉のコーパスのみから学習するには膨大な量のコーパスが必要となる．そこで本研究では，新聞記事コーパスから引用を示す表現を収集し，引用符の判定に利用する．

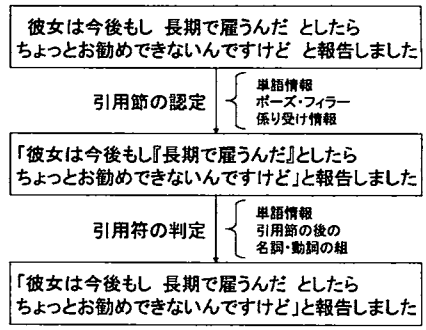

図1 提案手法の概要

3. 引用節認定と引用符判定のアプローチ

本研究では，話し言葉に対して引用符を自動付与する手順として，引用節の範囲を認定した後，それらに引用符が必要かどうかを判定する．図1に本手法の流れ図を示す．以下ではそれぞれの手法について説明する．

3.1 引用節の認定

引用節の認定は，著者らが研究を行なってきた手法[5]に従い，テキストチャンキングの問題として扱う．テキストチャンカには，Support Vector Machine に基づく

表 1　チャンクラベルの種類

ラベル	ラベルの説明
B	引用節の始端
E	引用節の終端
I	引用節の内部 (始端，終端以外)
O	引用節の外部
S	1 文節から成る引用節

YamCha [6] を用いる．チャンクラベルは表 1 に示したものを文節ごとに付与する．YamCha における多項式カーネル次数は 3，解析方向は Right to Left とし，後方 3 文節のチャンクラベルを動的素性として利用している．

チャンキングの素性としては，単語情報 (表層表現・読み・品詞情報・活用の種類・活用形) やポーズ長・フィラーの有無および話速を用いる．引用節の終端では「〜と思う」「〜っていう」「〜とか」などの表現が多く現れるので，単語情報が引用節の終端を認定する際の大きな手がかりとなる．しかし，これらの素性はすべて局所的な情報であるため，引用節の始端を同時に推定するのは困難である．始端を決定するためには，上述の素性に加え大域的な情報も必要となる．

そこで，始端を決める際に，自動推定した係り受けの情報をあわせて利用する．引用節の終端が得られている場合，始端より前の文節の係り受けには図 2 のような制約が成り立つ．本手法ではこの制約を利用し，チャンキングを 2 回にわたって行なう．1 回目のチャンキングでは上述の素性のみを用いて引用節の認定を行い，得られた終端の情報をもとに，図 2 における (1)(2) の係り受けの確率を素性に加えて，2 回目のチャンキングを行なう．係り受け解析には，最大エントロピーモデルによる手法 [7] を適用した．

図 2 において，(1) の確率が低く，(2) の確率が高い文節ほど，引用節の始端になりやすいといえる．以下の例文では，「男の子に」が「言おうと」に係ることから，直前の文節が引用節の後方に係っている「電話番号を」が始端になると推定できる．

(例)　その
男の子に
{電話番号を
教えてくれ} と
言おうと
思ったんですが

なお，トカ節による引用は並列して用いられることが多く，その場合は以下の例文のように，図 2 の制約が満たされない (2 つ目のトカ節の始端「隣の」の直前の文節「挨拶しないとか」は，このトカ節の内部「知らないとか」に係っている)．

図 2　引用節の始端以前の係り受けの制約

(例)　例えば
{マンションの
隣り同士が
挨拶しない} とか
{隣りの
人は
知らない} とか
最近
よく
聞きますけど

よって，トカ節については係り受けを利用せず，1 回目のチャンキングのみを行う．

3.2　引用符の付与判定

前節で得られた引用節に対して，それらが発言の引用であるかどうかという基準をもとに引用符の付与判定を行う．本手法では，引用符の有無をコーパスから SVM によって学習する．判定に用いた素性は以下の通りである．

(1)　引用節終端の形態素情報

引用節の末尾および直後の形態素の出現形・基本形・品詞情報・活用形を素性として用いる．2.2 節および 2.3 節で述べたように，引用節の末尾および直後の表層表現によって，引用符の付与が判定できることが多い．

(2)　引用節の後に現れる名詞と動詞の組

2.3 節で述べたように，発言の引用の後には「話す」などの特定の動詞が使われることが多い．動詞単体だけでなく，「報告する」「声をかける」など，名詞と動詞が複合した表現も扱えるように，引用節の後の名詞と動詞の組を素性として利用する．例えば，「話す」「報告する」「声をかける」に対しては，(*, 話す)，(報告, する)，(声, かける) という組を抽出する．

(3)　名詞と動詞の組に対して，新聞記事コーパスで引用符が付与される割合

上述の名詞と動詞の組が得られる引用節を新聞記事コーパスからも同様に取得し，それらの引用節に引用符が付与されている割合を素性として利用する．新聞記事コーパスにおいては，引用節の直後には格助詞の「と」

表 2　新聞記事コーパスから得られる
引用節の後の名詞と動詞の組の例

引用符の有無・割合				(名詞) (動詞)
有	5746	無	146	97%	* 話す
有	5146	無	331	93%	* 述べる
有	911	無	130	87%	* 答える
有	790	無	132	85%	説明 する
有	710	無	11474	5%	* 思う

が現れる場合が大半である．または,「と」の直前に「など」が入る場合も多い．そこで，新聞記事コーパスでの引用節の終端の認定は以下のパターンとのマッチングにより行う．

*** { 動詞 | 形容詞 | 形容動詞語幹 | 助動詞 | 終助詞 } (引用符閉)? (など)? (と)**

なお，新聞記事コーパスの形態素解析には Juman 5.1 を用いている．

毎日新聞 1995 年のデータから得られた引用節の後の名詞と動詞の組の例を表 2 に示す．表 2 にはそれぞれの名詞と動詞の組が得られる新聞記事コーパスの引用節に対して，引用符が付与されている個数，付与されていない個数，および引用符の付与されている割合を示している．

4. 評価実験および考察

ここでは，引用節の認定および引用符の判定を行なった結果と考察について述べる．実験に用いたコーパスは CSJ コア 188 講演 (模擬講演 111 講演+学会講演 77 講演) の書き起こしである．以下の各実験のテストデータには模擬講演のうち共通の 11 講演を用いた．

4.1 引用節の認定

3.1 節の手法を用いて，引用節の自動認定結果を行なった結果を表 3 に示す．学習データにはテストデータを除く模擬講演 100 講演に加えて，学会講演もあわせて利用した．表 3 には以下の 5 通りの実験結果を示している．

- 係り受けを用いない場合 (1 回目のチャンキング)
- open テストで得られた係り受けを利用した場合 (2 回目のチャンキング)
- closed テスト (テストデータも学習に利用) で得られた係り受けを利用した場合 (2 回目のチャンキング)
- 正解の係り受けを利用した場合 (2 回目のチャンキング)
- 引用節の終端が正しく認定できた割合

なお，係り受け解析精度は open テストで 79.8% , closed テストで 89.5% であった．

表 3　引用節の認定精度

	再現率	適合率	F 値
係り受けを利用しない	47.9 % (192/401)	53.0 % (192/362)	50.3
係り受けを利用 (open)	49.8 % (200/401)	55.7 % (200/359)	52.6
係り受けを利用 (closed)	60.8 % (244/401)	67.6 % (244/361)	64.0
係り受けを利用 (正解)	80.5 % (323/401)	90.0 % (323/359)	85.0
終端のみ一致	88.3 % (354/401)	98.1 % (354/361)	92.9

引用節の終端の 9 割以上は正しく認定できた．始端とともに正解した割合については，open テストでも自動推定した係り受けを利用することによって精度が向上した．これは 2 回目のチャンキングの際に素性として利用した係り受けが有効に作用したことを示している．なお，closed テストの結果や正解の係り受け情報を用いた場合には，認定精度は大きく向上しており，係り受け解析の精度が改善されるのに伴い引用節の認定精度も向上することがわかる．

4.2 引用符の付与判定

次に，すべての引用節に対して，引用符を付与すべきか否かの判定を行った．学習データにはテストデータ以外の模擬講演 100 講演を用いている．判定の結果を表 4 に示す．ここでは，3.2 節における素性 (1)〜(3) を順に追加していった場合の精度について比較を行った．素性 (3) の引用符の割合は，3.2 節で述べた方法により求めた．表 4 においては，すべての引用節が正しく認定できたと仮定した場合の引用符の判定結果を示している．

表 4 から，引用節の後の名詞と動詞の組，および，それらに対して新聞記事コーパスで引用符が付与される割合を素性として加えることで，判定の精度が向上していることがわかる．以下の例文では，話し言葉では頻度の少ない「明記する」や「言い張る」といった表現に対して，新聞記事コーパスで引用符の付与される割合が高いという情報を用いることで，正しく判定できるようになった．

◇ 「これは保健所に連れていく犬です」ということをちゃんとはっきり明記して書いてある訳ですよ

表 4　引用符の判定結果

	再現率	適合率	F 値
素性 (1) のみ	60.7% (34/ 56)	82.9% (34/ 41)	70.1
素性 (1)+(2)	64.3% (36/ 56)	81.8% (36/ 44)	72.0
素性 (1)+(2)+(3)	67.9% (38/56)	86.4% (38/44)	76.0

(1) 引用節終端の形態素情報
(2) 引用節の後に現れる名詞と動詞の組
(3) 新聞記事コーパスで引用符が付与される割合

◇ 暫く「逃げたのは姉のかんなの方でここにいるのはさくらだ」と言い張っていたのを覚えています

逆に以下の例では，素性 (1)(2) のみを用いた場合には，引用符が誤って付与されていたものの，新聞記事コーパスでは「覚える」という動詞に対しては引用符が付与されることが少ないことから，引用符が不要であると修正された．

◇ そんで何かその時に×「戻した方が気持ちが楽になるんだ」っていうことを覚えて

本手法で検出できなかった引用符には，以下のようなものがあった．

◇ そういう方に声掛けて「お前一緒にちょっと医者でも行くか」っちゅうことで色々調べて連絡し合ったりすんのはいつも決まって私の役目なんですよね

この例では，「お前」という単語によって直接引用であることが表されている．また，引用節の手前の「声掛けて」という表現も，引用符が付与できると判定するための手がかりとなっている．本手法では引用節の終端の情報しか利用していないが，上記のような引用符を判定するためには，さらに引用節の直前や内部の情報を判定に取り入れる必要がある．しかしながら，引用節の始端の周辺の形態素を素性に加えても，全体の精度が低下する結果となった．これは素性数の増大が原因であるとも考えられるため，より一層の検討が必要である．

次に，4.1 節の open の係り受けを利用した引用節の自動認定結果に対して引用符の判定を行った．その結果，引用符の有無と位置が正しく推定できた割合は再現率 14.3%，適合率 22.9%，F 値 17.6 であり，十分な精度は得られなかった．正しく引用符の有無と位置が推定できたものは，以下の文のように一文が短いものや引用節の終端が文頭に近い場合が多かった．

◇ 「肝臓がかなり痛んでいますね」と言われて確かにこう今年に入って一月二月とか凄いペースで飲んでいたなと思ってちょっとお酒の歴史を振り返ってみたんですけれども

◇ 「ここにとどまってろ」と言えばとどまってる

しかし実際には，一文が長く，文の途中から引用節が始まるような場合に，引用符を付与することによる可読性の向上効果が大きいと考えられる．そのような引用符を自動で付与するためには，係り受け解析精度の改善とともに，引用節の認定精度の向上が必要である．

また，誤った位置に引用符が挿入されると，かえって可読性が損なわれる．従って，引用符の判定手法においては，引用節の自動認定結果が誤っている場合に引用符が付与されないような，頑健な学習を行う必要がある．

5. おわりに

本稿では，CSJ を対象として，引用節の認定と引用符付与の判定を行う手法について述べた．今後の課題としては，係り受け解析の改善などによって引用節の認定精度を向上させることや，評価実験で明らかになった引用符の判定における問題点をもとに，判定に利用する情報について検討することが挙げられる．

文　献

[1] 浅原, 松本：“形態素解析とチャンキングの組み合わせによるフィラー/言い直し検出”, 言語処理学会 第 9 回年次大会 発表論文集, pp. 651-654 (2003).

[2] 下岡, 河原, 内元, 井佐原：“『日本語話し言葉コーパス』における自己修復部 (D タグ) の自動検出および修正に関する検討”, 情報処理学会研究報告 (2005).

[3] 下岡, 内元, 河原, 井佐原：“日本語話し言葉の係り受け解析と文境界推定の相互作用による高精度化”, 自然言語処理, **12**, 3, pp. 3-18 (2005).

[4] 古井, 前川, 井佐原：“科学技術振興調整費開放的融合研究推進制度－大規模コーパスに基づく『話し言葉工学』の構築－”, 日本音響学会誌, **56**, 11, pp. 752-755 (2000).

[5] R. Hamabe, K. Uchimoto, T. Kawahara and H. Isahara: “Detection of Quotations and Inserted Clauses and Its Application to Dependency Structure Analysis in Spontaneous Japanese”, Proceedings of COLING/ACL 2006.

[6] T. Kudo and Y. Matsumoto: “Chunking with Support Vector Machines”, Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics, pp. 192-199 (2001).

[7] 内元, 村田, 関根, 井佐原：“後方文脈を考慮した係り受けモデル”, 自然言語処理, **7**, 5, pp. 3-17 (2000).