講演録作成を目的とした 話し言葉への引用符の自動付与

浜辺 良二† 内元 清貴‡ 河原 達也† 井佐原 均‡

† 京都大学 情報学研究科

‡ 独立行政法人 情報通信研究機構

1 はじめに

近年，音声認識技術の進展により，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，話し言葉にはフィラーや言いよどみ，言い直しなど，書き言葉に現れない冗長な表現が含まれているため，削除する必要がある．これら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なわれている [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引用符が付与されることが多い．一方，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．単純に音声を書き起こした場合，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節を認定した後，それらに引用符を付与すべきであるか否かを判定する．Support Vector Machine (SVM) を用いた機械学習により，これらの処理を行なう．

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．その中の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 において引用符を付与すべき箇所の判定基準を定めた上で分類を行なう．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固有名詞や強調箇所にも引用符が付与されるが，これらは引用符を付与すべきかどうかの曖昧性が大きいため本研究では対象外とする．

2.1 CSJ における引用と節境界

発話の引用箇所は文内の節となっており，その終端は CSJ で定義されている節境界と一致する．引用箇所の終端となりうる節境界は引用節・トイウ節・トカ節である．以下にそれぞれの例を示す．

♢ **「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** (引用節)

♢ **「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** (トイウ節)

♢ **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** (トカ節)

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．

ここで，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．

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こととした．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**

今回分類を行なった 111 講演には，引用符の付与対象となりうる節が 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．

ここで，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．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

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，上述のように代名詞など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．

2.3 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．

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否かを判定する．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，SVM に基づく 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．

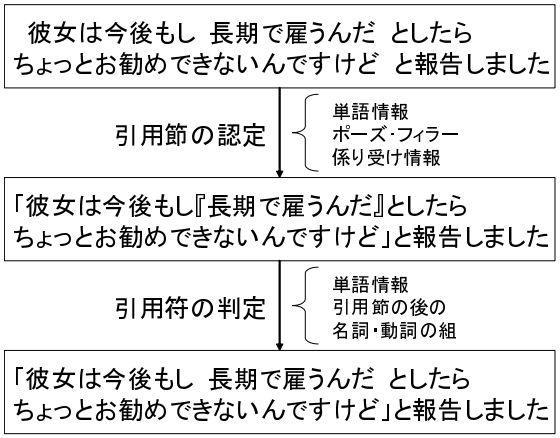

図 1: 提案手法の概要

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

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ない，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

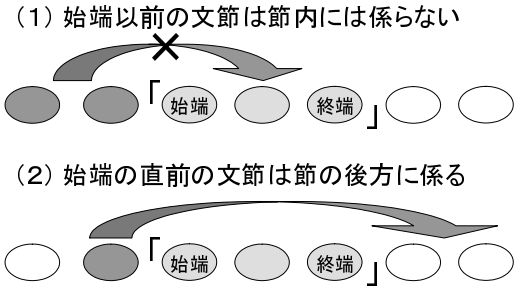

図 2: 引用節の始端以前の係り受けの制約

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否かという基準のもとに引用符の付与判定を行なう．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) 引用節終端の単語情報
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ける引用節の終端の認定は，形態素解析を行なった結果に対して以下のパターンとのマッチングを行なうことで実装した．
{ 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，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．

表 2: 新聞記事コーパスにおける引用節の直後の表現の例

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

4 評価実験

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．テストデータには模擬講演のうち 11 講演を用いた．以下では，引用節が既知である場合，および引用節を自動認定する場合のそれぞれについて評価を行なう．

4.1 引用符の付与結果 (引用節が既知の場合)

まず，すべての引用節が正しく認定されたと仮定し，それぞれについて 3.2 節で述べた手法で引用符を付与すべきか否かの判定を行なった．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 3 に示す．ここでは，3.2 節における素性 (1)〜(3) を順に追加した場合の精度について比較を行なった．

表 3 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
♢ 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．

♢ そっから娘達が「そいじゃこの犬はスリーピーと付けよう」ということで名前も貰った次第です

この例では，引用節の終端前後の表層表現に発言の引用であるという情報は含まれていない．このような場合に引用符を正しく付与するためには,「娘達」が引用

表 3: 引用符の付与精度 (引用節が既知の場合)

	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) 引用節終端の単語情報
(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

表 4: 引用符の付与精度 (引用節を自動認定する場合)

	再現率	適合率	F 値
係り受けを利用しない	14.3 % (8/56)	21.1 % (8/38)	17.0
係り受けを利用 (open)	17.9 % (10/56)	28.6 % (10/35)	22.0
係り受けを利用 (closed)	32.1 % (18/56)	50.0 % (18/36)	39.1
係り受けを利用 (正解)	48.2 % (27/56)	75.0 % (27/36)	58.7
終端のみの精度	53.6 % (32/56)	81.6 % (32/38)	68.1

節の内容の発言者であり「この犬」という指示表現が「娘達」の視点によるものであるなどといった解析が必要となるため，困難である．

4.2 引用符の付与結果 (引用節を自動認定する場合)

次に，3.1 節の手法により引用節を自動認定した結果に対して，引用符を付与すべきか否かを判定し，引用符の自動付与を行なった結果を表 4 に示す．引用符の付与判定では，3.2 節で述べた素性 (1)〜(3) のすべてを利用した．表 4 には引用節の認定方法に関して，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)
- 係り受けを利用した場合 (2 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
  - closed テストで得られた係り受けを利用
  - 正解の係り受けを利用
- 引用符の終端を正しく付与できた割合

なお，引用節の認定および係り受け解析の学習の際には，学会講演もあわせて利用した．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．closed テストでは，テストデータも学習に利用している．

表 4 から，引用節を自動認定する際の誤りによって，引用符の付与精度が大きく低下していることがわかる．特に，引用節の始端の認定誤りが多いことが，引用符の付与精度が低下する最大の要因となっている．これは，日本語では引用の開始を表すような特定の表層表現がなく，引用符の始端の前後においても特定の品詞や活用のパターンが現れないため，機械学習による始端の認定が困難であることが原因と考えられる．

引用符の始端を決定するためには，大域的な情報を考慮する必要がある．表 4 から，係り受け情報を利用することによって引用符の付与精度が向上することがわかった．しかし，open テストで得られた係り受けを用いて引用節の認定を行なった場合では，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く始端になりうる箇所の候補が少ない場合が多かった．

♢ 「ここにとどまってろ」と言えばとどまってる
♢ 「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

また，closed テストで得られた係り受けや正解の係り受けを利用することで，引用符の付与精度は，終端が正しく付与される精度に近づいていくことがわかる．このことから，引用符の付与精度向上のためには，係り受け解析の改善によって引用符の始端の認定精度を向上させることが重要であるといえる．

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な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，引用符の付与判定において，表層表現のみでは判定できない場合にも利用できる他の情報について検討することなどが挙げられる．

参考文献

[1] 浅原正幸, 松本裕治. 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] 下岡和也, 河原達也, 内元清貴, 井佐原均. 『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] 下岡和也, 内元清貴, 河原達也, 井佐原均. 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, Vol. 12, No. 3, pp. 3–18, 2005.

[4] 古井貞熙, 前川喜久雄, 井佐原均. 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, Vol. 56, No. 11, pp. 752–755, 2000.

[5] Ryoji Hamabe, Kiyotaka Uchimoto, Tatsuya Kawahara, and Hitoshi Isahara. Detection of Quotations and Inserted Clauses and Its Application to Dependency Structure Analysis in Spontaneous Japanese. In *Proceedings of COLING/ACL*, pp. 324–330, 2006.

[6] Taku Kudo and Yuji Matsumoto. Chunking with Support Vector Machines. In *Proceedings of NAACL*, pp. 192–199, 2001.

[7] 内元清貴, 村田真樹, 関根聡, 井佐原均. 後方文脈を考慮した係り受けモデル. 自然言語処理, Vol. 7, No. 5, pp. 3–17, 2000.