2005－SLP－57（15）
2005／7／16

独話データのポーズ単位を利用した節境界判定

柏岡 秀紀

ATR 音声言語コミュニケーション研究所

〒 619-0288 京都府相楽郡精華町光台 2 丁目 2 番地 2

hideki.kashioka@atr.jp

講演などの独話においては、発話者が文の切れ目を明確にせず発話し続ける傾向が強いため、文境界が明確でない。同時通訳や聴衆への理解支援などの処理の実現には、追従性が高く、文法的にもまとまった比較的短い処理単位の判定が必要である。その処理単位を節と考え、音声の入力において比較的容易に見出せるポーズにより区切られた入力に対して、節境界を検出する手法について検討する。本稿では、我々が開発した節境界検出プログラム CBAP を用い、CBAP が期待する文単位での形態素解析とポーズ単位での形態素解析の異なりと調べ、テキスト上での情報を利用した CBAP の出力を修正することによる効果とポーズ長を考慮することによる効果について調査した結果について報告する。

独話、ポーズ、節境界単位、形態素解析

Clause Boundary Detection with Pause Unit in Japanese Monologue

Hideki KASHIOKA

ATR Spoken Language Communication Research Laboratories

2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-0288 Japan

hideki.kashioka@slt.atr.co.jp

Sentences in a monologue generally tend to be long and complicated, and the sentence boundaries in utterances are not clear. They cause problems for spoken language processing (i.e., recognition, parsing, translation, and synthesis); therefore, a short constituent is required to process them. A clause is a syntactically and semantically sufficient constituent. Some tools are available for clause detection, but almost all of them require sentence input. A pause unit can be used as a constituent in monologues because they are segmented by acoustic information. This paper discusses a system in which input is a pause unit and output is a clause boundary annotated text.

monologue, pause, clause boundary unit, morphological analysis

1 はじめに

音声言語処理において講演などの独話の処理は、対話処理とともに重要な課題である。対話における処理単位は、話者交代等による比較的短い発話であるのに対し、講演などの独話の処理単位は、一人の話者が話し続けることから発話が長くなる傾向があり、明確な処理単位がない [1]。講演の発話では、文末と思われるところでも、接続表現等を利用し文末を明確にせず、発話を続けたり、文の形式が崩れた、ねじれた表現などが多く見受けられる。

講演などの独話において期待される音声言語処理のひとつに、同時通訳がある。同時通訳では、原発話に追従して訳出していく追従性 (同時性) が要求される。処理単位が長ければ、原発話を聞いているだけの時間が長くなり、追従性が損なわれ、処理単位が短ければ、追従性は高くなる。しかしながら、処理単位を単に短くした場合、適切な対訳を生成するための情報が欠落してしまう。そのため、翻訳における適切性や正確性が失われる。たとえば単語ごとの処理を考えると、訳語選択の手掛かりが失われ、また、原言語の語順で訳語が現れる。原言語になじみがあれば、聞き手が語順を考慮することで理解できるかもしれないが、一般には、理解しがたい訳出となる。そこで、処理の追従性が高くまた、適切な処理に必要な情報を保持している処理単位が望まれる [2]。節は、文法的なまとまりがあり、比較的短い単位であることから、処理単位として有望であろう。

正確な節の判定は、正確な構文構造を捉える必要があり、節の入れ子構造や並列構造等、複雑かつ曖昧な構造により困難である [3][4]。しかしながら、節末の境界は局所的な形態素列のパターンにより、ある程度容易に判定できる。我々は、実際に節境界を判定するツールとして、局所的な形態素列のパターンを利用した CBAP とよばれるツールを作成した [5]。CBAP は、様々なテキストに対して 97%の精度を達成している。また、CBAP は、形態素解析結果を入力としているが、この精度は、文を入力とした解析結果を前提としている。

実際の音声入力において、節境界を判定するために文末まで処理を保留することは、先ほど述べた追従性に反する。そのため、より短い入力単位で節末の判定を実現することが望まれる。入力音声において比較的容易に判定できると思われるポーズによる発話の分割を行い、分割された発話ごとに形態素解析した場合の CBAP の振る舞いを分析を行った。本稿では、文単位とポーズ単位での形態素解析の比較を行うとともに、各単位での形態素結果を入力としたときの CBAP の解析結果の比較について述べる。さらに、精度を向上させるために、入力をポーズ単位とした場合の CBAP の出力へのフィルターについて検討を加え、また、ポーズ長と文末との関係についても考察を加える。

2 課題設定

2.1 対象データ

本稿では、NHK で放送されている解説番組「あすを読む」を対象として、250 番組の書き起こしテキストを利用した。「あすを読む」は、月曜日から金曜日までの毎日 10 分間の番組であり、250 番組でほぼ 1 年間のデータとなっている。番組毎に扱っている話題を含む分野を専門とする解説委員が解説を行っている。

対象とする書き起こしテキストは、作業者の判断で書き起こしの際に、読点を含めた。読点の判断 (文の認定) は、書き起こし作業者の主観に任せた。また、200ms 以上のポーズおよびフィラーの前後に改行をいれ、ポーズ等による発話の分割点の情報を記述している。

1. そして九十年代になりますと、このゲノムとデジタル、これが互いに深い関係をもつようになって参りました。
2. 同じ理科系と言いましても、この生物系そして理工科系、相当違うところがあるわけなんですが、おもしろいことにこうした両極端が結び付いてきたというわけなんですね。

図 1: 文により分割されたテキスト

2.2 分割点

文末の判定は、先ほども述べたように書き起こし作業者の主観的な判断による。また、ポーズによる発話の分割は、書き起こし作業時にポーズ長をはかり、200ms 以上のポーズおよび、フィラーの前後で改行を行った。その結果、1 番組には、平均 58.6 文、194.2 個のポーズ単位が含まれている。一文に

対して、平均 3.3 個のポーズ単位が含まれていることになる。文単位でのテキストの CBAP による解析では、節境界は 1 番組中に平均 265.8 個、含まれていた。

1. そして
2. 九十年代になりますと
3. このゲノムとデジタル
4. これが
5. 互いに深い関係をもつようになって参りました
6. 同じ理科系と言いましても
7. この生物系そして理工科系、相当違うところがあるわけなんですが
8. おもしろいことにこうした
9. 両極端が
10. 結び付いてきたというわけなんですね

図 2: ポーズ単位により分割されたテキスト

文単位でのテキスト (図 1) とポーズ単位でのテキスト (図 2) の例を示す。文単位では 2 文のものが、ポーズ単位では、10 個に分割されている。文単位での形態素結果を入力とした CBAP の出力は、12 個の節に分割している (図 3) また、ポーズ単位の場合には、文末情報が欠けているという前提から、読点を削除している。

1. そして九十年代になりますと/条件節ト/
2. このゲノムとデジタル/体言止/
3. これが互いに深い関係をもつように/ヨウニ節/
4. なって参りました。/文末/
5. 同じ理科系と言いましても/譲歩節テモ/
6. この生物系/体言止/
7. そして/談話標識/
8. 理工科系相当違う/連体節-形式名詞/
9. ところがある/連体節-形式名詞/
10. わけなんですが/並列節ガ/
11. おもしろいことにこうした両極端が結び付いてきたという/連体節トイウ/
12. わけなんですね。/文末/

図 3: 節分割されたテキスト

おもしろい/こと/に/こうした/両極端/が/結び付い/て/き/た/と/いう/わけ/な/ん/です/ね/

図 4: 文単位での解析結果例

おもしろい/こと/に/こう/し/た/両極端/が/結び付い/て/き/た/と/いう/わけ/な/ん/です/ね/

図 5: ポーズ単位での解析結果例

3 実験

3.1 形態素解析

本稿では、形態素解析に茶筌 [6] を利用している。先の例の一部の解析結果は、文単位で処理した場合、図 4 のようになり、ポーズ単位で処理した場合、図 5 のようになる。

この例にあるように下線を引いた部分の解析が、文単位では、こうした（連体詞)」と解析されているのに対して、ポーズ単位では、「こう (副詞) ＋し (動詞 (する)) ＋た (助動詞)」と解析される。ポーズ単位では、「こうした」の後で処理単位が分割されているため、後続の「両極端」に係る連体詞として

処理されず、文末として捕らえやすいように動詞としての解釈が優先されたためである。

茶筌による解析の結果から、250 番組の書き起こしテキスト全体は、ほぼ 44 万語 (延べ語数) からなる。文単位での処理結果と、ポーズ単位での処理結果には、句読点を除いて、約 4,519 箇所の差異が見られた。これは、全体の語数から見れば 1%程度の差異である。特に目立った定型的な差異は見当たらなかった。

3.2 CBAP 結果

前節で示した茶筌の解析結果を基に CBAP で節境界の判定を行った。CBAP は、300 強の局所的な (節境界の前後 2,3 語の) 形態素列パタンの一致により節境界を検出し、150 種類弱の節ラベルを付与する。各節境界ラベルは、階層的に分類されており、表 1 のような階層を持つ。

表 1: 節のタイプ

並列節	並列節
連用節	条件節, 譲歩節, 連用節 (その他) 時間節, 理由節
補足節	補足節, 引用節, 間接疑問節
連体節	連体節
その他	従属文, 体言止, 文末 主題ハ, 談話標識, 感動詞, 間投句

文単位の形態素解析結果を入力とした場合、66,451 箇所の節境界が検出されており、ポーズ単位での形態素解析結果を入力とした場合、70,632 箇所の節境界が検出された。両者の間には、節タイプを含めた節境界判定に、21,873 箇所の差が見られた。表 2 に、その主なものを示す。

CBAP は文単位での形態素解析結果を入力した場合、精度が 97%を超えていることから、21,873 箇所の差異のほとんどは、ポーズ単位での形態素解析結果による誤りであると考えられる。この差異は、全体の 34.2%となる。また、形態素の差異は、句読点を除いて約 4,519 箇所であり、文単位で文末であった部分が、14,660 箇所で合わせて 2 万弱となり、CBAP での差異とほぼ同じになる。形態素の差異は節境界の判定に重要な部分で生じているといえる。

表 2: 主な節境界検出結果の差異

頻度	文単位の分割	ポーズ単位の分割
7,361	「文末」	「連体節」
6,004	「文末」	「従属文」
3,994	–	「談話標識」
710	「文末」	–
534	「文末」	「間接疑問節」
325	「連体節」	–
268	「文末」	「主題ハ」
221	「引用節」	「条件節ト」

4 考察

本節では、表 2 を元に、ポーズ単位での形態素結果を入力としたときの CBAP の精度向上について検討、議論する。

4.1 追加規則による修正

まず、テキストに含まれる情報を利用した改善について検討する。表 2 に示されるように、節境界判定における差異の主なものは「文末」にかかわる部分で生じていることがわかる。特に「文末」を「連体節」および「従属文」に誤っている。以下の単純なラベルの置き換えにより改善について調べた。表 3 にこの置き換えを行うことによる差異の変化を示す。

RULE1 「連体節」のラベルがポーズ単位の境界であれば「文末」に置き換える。

RULE2 「従属文」のラベルがポーズ単位の境界であれば「文末」に置き換える。

この表 3 から、RULE1 を利用することで、7,361 箇所の改善に対して、もともとの「連体節」を「文末」としてしまう誤りが、1,257 箇所増加している。また、RULE2 を利用することで、6,004 箇所の改善ができていることがわかる。RULE2 のほうは、大きな副作用は起こしていない。実際には、75 箇所の誤りが増加していた。以上のことから、これらのラベル置き換え規則により、約 12,000 箇所の修正、および、約 1,300 箇所の誤りの増加が確認された。誤りと思われる箇所の約 55%は、改善できたことになる。

表 3: RULE1, RULE2 による節境界単位判定の変化

頻度	RULE1	RULE2	文単位の分割	ポーズ単位の分割
7,361	**0**	7,361	「文末」	「連体節」
6,004	6,004	**0**	「文末」	「従属文」
3,994	3,994	3,994	–	「談話標識」
710	710	710	「文末」	–
534	534	534	「文末」	「間接疑問節」
325	325	325	「連体節」	–
268	268	268	「文末」	「主題ハ」
221	221	221	「引用節」	「条件節ト」
	1,257		「連体節」	「文末」
		75	「従属文」	「文末」

4.2 ポーズ長との関係

表 2 に現れる異なりの要因の主なものは、ポーズによる区切りからでは「文末」の情報をテキスト上では見つけられないということである。3 番目の「談話標識」が増加しているのは、ポーズで分割されたために接続助詞を接続詞とする解析誤り、およびポーズ単位による他のタイプの過剰な節境界の検出によるものと考えられる。これは、CBAP の談話標識の検出が、直前に節末を期待していることによる。現在、テキスト以外の情報で比較的利用しやすい情報は、ポーズ長である。図 6 は、実験に利用したコーパスの一部の 50 番組のデータから得られたポーズ長と文境界との関係を示すものである。横軸にポーズ長をとり、縦軸は文境界となるポーズの割合を示している。この図 6 から、見かけ上はほぼ 1000ms 以上のポーズであれば、文境界と考えられそうである。

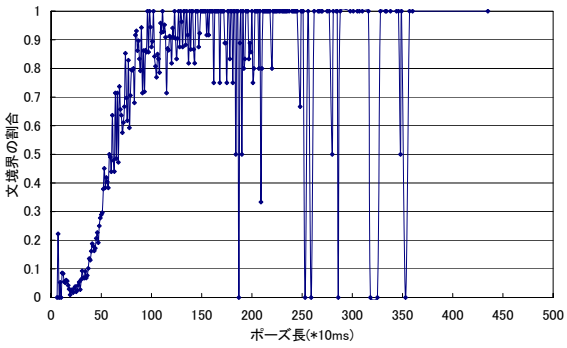

図 6: ポーズ長と文境界との関係

そこで、図 7 は、横軸にポーズ長を長いほうからとり、横軸の値を閾値としたときの文境界判定の精度を縦軸とするグラフである。このグラフから、最も精度良く文境界を判定する閾値は、1290ms で 93.6%となった。

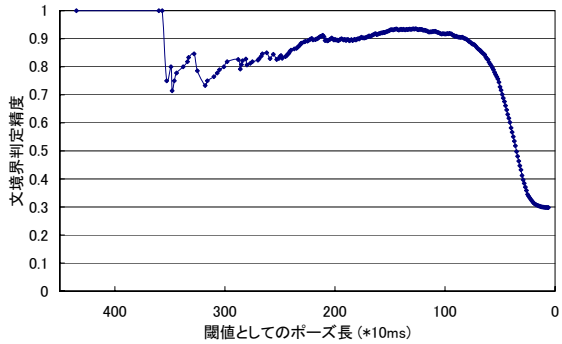

図 7: ポーズ長と文境界精度

このことから、ポーズ長を利用することで、文末の情報を得ることで、節境界判定の精度を向上させることが考えられる。

4.3 節境界判定修正による形態素解析の修正

節境界の判定をこれまでに述べたような手法で修正することにより、もともと誤りであった形態素の情報を修正することができる。たとえば、図 5 に示した “こうした両極端” の形態素解析結果から、 “こうした” の部分は「連体節」と判断されるが、節

境界ではないと判断できれば、図 4 に示す形態素列に修正することができる。形態素の変化は、高々1%であるが、翻訳や要約などの処理において、「文末」等の判断基準となる述部の形態素誤りは、その影響がおおきい。そのため、このような修正は、後段の処理において有効に働く。

5 おわりに

本稿では、講演など独話の音声言語処理に必要な処理単位の判定について、節をその処理単位としたときの追従性（同時性）を考慮した判定手段について議論した。我々が開発した節境界判定のツール CBAP を元に、文末を待たずに、ポーズ等で判断できる単位を入力としたときの形態素解析の精度と CBAP の精度について文単位の入力との比較を行った。その結果、形態素解析の誤り 1%に対して、CBAP の精度が 30%以上、落ちることがわかった。しかしながら、節ラベルの比較的単純な書き換え、およびポーズ長などを考慮することにより、少なくとも誤りの約半数を回復することができることがわかった。

今後は、音声認識誤りを含む結果に対して、高精度な処理単位の判定を行うための検討をするとともに、節単位での要素処理技術の高度化、および、翻訳のための課題を調査し、追従性の高い翻訳システムの構築を目指す。

参考文献

[1] Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H. and Yamamoto, S.: “Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world” Proceedings of LREC-2002, pp.147–152. (2002).

[2] Kashioka, H., Maruyama, T.: “Segmentation of Semantic Units in Japanese Monologues” Proceedings of ICSLT-O-COCOSDA-2004, New Delhi, India. (2004).

[3] 金淵培 and 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, (1994).

[4] Tjong Kim Sang, E. F. and Déjean, H.: “Introduction to the CoNLL–2001 Shared Task: Clause Identification” In Daelemans, W. and Zajac, R. (Eds.), Proceedings of CoNLL-2001, pp.53–57. Toulouse, France. (2001).

[5] 丸山 and 柏岡 and 熊野 and 田中: “日本語節境界検出プログラム CBAP の開発と評価” 自然言語処理, Vol.11, No.3, pp.39–68, (2004).

[6] Asahara, M., Matsumoto, Y.: “Extended Models and Tools for High-performance Part-of-Speech Tagger” Proceedings of COLING 2000, July, (2000).