自然言語処理 143－7
音声言語情報処理 36－2
（2001．6．1）

話し言葉の形態素解析

松本 裕治
奈良先端科学技術大学院大学
情報科学研究科
matsu@is.aist-nara.ac.jp

伝 康晴
千葉大学 文学部
行動科学科
den@L.chiba-u.ac.jp

話し言葉研究のための基礎データとしてタグ付きコーパスの蓄積が進んでいる。言語データへの最も基本的なタグは単語わかち書きと品詞付与である。本稿では、書き起こされた話し言葉データへの形態素タグ付け自動化のための問題点について考察する。まず、書き言葉と対比して見られる話し言葉の特徴と問題点について整理する。次に、統計的学習に基づく日本語形態素解析システムを用い、少量のタグ付き話し言葉データが解析精度にどのように貢献するかを観察する。

[キーワード] 形態素解析、話し言葉、非流暢性、コーパス、統計的言語処理

reminder

Morphological ken Japanese

MATSUMOTO Yuji
Graduate School of Informatic
Nara Institute of Science and Technology
matsu@is.aist-nara.ac.jp

DEN Yasuharu
of Letters, Chiba University
den@L.chiba-u.ac.jp

Tagged corpora are indispensable resource for linguistic research. Several projects are now under way for constructing spoken language corpora. The foundamental annotation to corpora is segmentation and part of speech tagging. In this paper, we examine the issues peculiar to spoken language annotation compared with written language. First, we summarize the characteristics and issues of spoken language. We then report some experiments of automatic part of speech tagging based on statistical learning algorithm, through which we see how a small size of tagged corpus is effective in improving the accuracy of the automatic taggers.

[Keywords] Morphological Analysis, Spoken Language, Disfluency, Corpus, Statistical Language Processing

1 まえがき

言語研究には、客観的な研究材料として充分な量のコーパスの存在が重要であり、特に、様々な言語情報が付与されたタグ付きコーパスの蓄積が急務である。近年、音声言語コーパスの蓄積を目指したいくつかのプロジェクトが活動を始めており、その書き起こしデータの蓄積も進んでいる。それらの書き起こしデータに対して、単語への分割およびそれらへの品詞や発音を正確に付与する必要がある。これを整合性を保ちながら人手により行うのは極めて困難であり、計算機による自動化、あるいは、計算機による精度の高いタグ付け支援を行うことがこれからの重要な課題である。

本稿では、日本語音声対話の書き起こしデータを対象にして、形態素解析システムによって品詞タグ付けを試みた我々の経験に基づき、話し言葉の形態素解析についての様々な問題点について述べ、その対応策について考察する。また、統計的学習モデルに基づく形態素解析器を用いて行った話し言葉解析実験について報告する。形態素解析の実行および実験については、日本語形態素解析システム「茶筌」[6] とその学習プログラム [1] を使用したので、それに準じて説明を行う。

2 話し言葉の特徴と問題点

話し言葉には、書き言葉には稀にしか出現しない言い回しや言語現象が見られる。そのため、書き言葉を対象に開発された言語処理システムでは精度の高い解析を行うことができないことが多い。実際、3節で示すように、書き言葉コーパスによって学習した形態素解析システムを自由な話し言葉に適用した実験では、書き言葉に比べて遥かに低い精度の解析しか行えなかった。本節では、新聞記事や随筆等の書き言葉と比べ、話し言葉に特徴的に現れる現象を分類し、それぞれの問題点を整理する。

2.1 フィラー

話しの調子を整えたり、思考の中断や遅れに伴って、話し言葉の端々に挿入され、発話の続き具合を知らせる間投的な詞。「あの」「えー」など。IPA品詞体系 [5] に基づいて我々が整理した辞書では、「あ」「あー」「あの」「あのー」「うん」「うんと」「え」「えー」「えーっと」「えーと」「ええと」「えと」「そうですね」「その」「そのー」「と」「なんか」「ま」「まあ」

表 1: IPADIC2.5 に登録されているフィラーの一覧

「フィラー」という品詞を独立した閉じたクラスの品詞として設定し、有限 (個) の単語のみを登録している。表1に茶筌に標準添付の辞書 (IPADIC2.5) に登録されているフィラーの一覧を示す。

フィラーは、多くの場合、語彙的に同定できるが、「あの」「その」「なんか」「まあ」のように、他の品詞と曖昧な場合もあり、その識別はかなり難しい。

2.2 非流暢性

表現を発話する際の失敗等のために、言い淀んだり、同じ語や表現を言い直し、あるいは、繰り返したりする現象。次に現れる単語の一部を発話した後、改めて言い直す場合や、文の一部を途中で中断して言い直す場合がある。言い直しや繰り返しは、次に現れる単語の接頭の発音の一部であったり、その類似の音であったり、あるいは、意味的に類似の単語の一部の音であることが多く、それ自身単語の断片であり、閉じたクラスの詞として定義することができない。また、これらは、文法的には無意味な語断片であり、発話から取り去っても元の発話の意味内容に影響を与えない。

このような語断片が、書き起こされた話し言葉データに含まれてしまっている場合は、これを自動的に識別するのは大変難しい。茶筌による解析の場合は、そもそも「語断片」を単語として網羅的に登録しておくのは不可能なため、原理的に同定することができない。未知語の判定と同様の手法を用いることにより、語断片の同定をある程度行なえる可能性はある。しかし、ここで議論している非流暢性は、書き起こしの段階で作業者には語断片ということがわかっているのであるから、この時点でタグを付与していくのが自然である。現在、国立国語研を中心として進んでいる日本語話し言葉コーパスプロジェクト [7] でも、書き起こし段階でタグ付けを行なうようになっており、今後もこの段階でタグによってマーキングされるのが主流となると考えられる。

例えば、書き起こしの際に、次のように XML 形

式で語断片を表すタグを付与することが考えられる。

<NL appear="コ"/>今後はこれが<NL appear="シュ"/>主流に

「茶筌」では、コメント文字列の開始と終了を指定する機能が備えられており、この機能を使うことで、コメント文が存在しないものとして (ただし、語の境界はそこに存在するとして)、形態素解析を行うことができる。なお、語断片をともなう非流暢の出現傾向は、個人差および発話の環境によって大きく異なると考えられる。我々が取り扱っている音声対話データ [4] では、出現比率は全形態素中の 1% 弱である。

2.3 非文法性

話し言葉においては、発話そのものが文になっていない場合や、文が途中で打ち切られる場合など、必ずしも文法的に正しい表現が用いられるとは限らない。統計モデルによる形態素解析では、連続する高々数個程度の形態素の列しか見ないので、文全体の文法性は要求しない。しかし、格助詞や感動詞で発話が終ったり、助詞が文頭に現れるなど、一般の書き言葉では起こりえないか極めて稀にしか出現しない表現が散見される。統計的学習に基づくシステムでは、コーパスに出現しなかった品詞の用法を学習することができず、通常極めて低い確率しか仮定されないため、解析を正しく行なうことができない。ただし、書き言葉には稀であるが、話し言葉ではある程度の頻度で見られる現象であれば、話し言葉の学習データを蓄積することにより、多くの現象には対応することができるようになる。3 節では、これを実験により確認する。

2.4 話し言葉に固有の表現

日本語の場合は、書き言葉と話し言葉では、用いられる単語や表現が大きく異なることがある。例えば、文末に用いられる終助詞「ね」「な」などは、話し言葉特有の表現である。その他、話し言葉が書き言葉と異なる主な要因には次のようなものがある。

縮約: 2 つ以上の単語が音韻変化により 1 つの単語に縮退する現象。例えば、「(〜し) ちゃう」という単語は、「(〜し) て+しまう」という 2 つの単語が 1 つに縮退した表現である。その他にも、「(〜し) て+おく ⇒ とく」、「(〜し) て+いる ⇒ てる」、「(〜し) て+おる ⇒ とる」など、同様の現象が見られる。また、「これ+は ⇒ こりゃ」、「で+は ⇒ じゃ」のような例や、「〜すれ+ば ⇒ 〜すりゃ」のように、用言の活用語尾を含めて縮約する場合がある。

これらは、話し言葉独特の単語であり、かつ、これらを 1 つの単語と考えると独特の働きをもつ。形態素解析におけるこれらの取り扱いについては、文献 [1] を参照されたい。

方言、俗語: 方言にはその地方特有の単語、言い回し、文末表現、用言の活用等がある。多くの場合は、辞書に単語を登録することによって解決するが、用言の活用については、方言毎に活用一覧を持つ必要があり、単純には解決できない。例えば、「きい-ひん」(京都方言の「来-ない」) や「けえ-へん」(大阪方言の「来-ない」) などは、それぞれ活用形を追加する必要がある。また、これらと同様の現象だが、関西方言の「食べえな (食べな)」などは母音の長音化であるが、これを活用語尾の一部として定義するべきかどうかは議論を要するだろう。同様の問題に、話し言葉特有に現れる俗語をどのように収集し、辞書に登録すればよいかという問題がある。

また、以下に示す問題は、話し言葉に固有という訳ではないが、正しい発音情報を付与する上では重要な問題である。

数詞の音韻変化: 数詞と助数詞・接尾辞が連接した場合に、数詞の発音が変化する現象。たとえば、「匹」との連接では「一匹 (いっぴき)」「二匹 (にひき)」「三匹 (さんびき)」、「枚」との連接では「一枚 (いちまい)」「二枚 (にまい)」「三枚 (さんまい)」、「つ」との連接では「一つ (ひとつ)」「二つ (ふたつ)」「三つ (みっつ)」のように、後部要素によって数詞の変化パターンが異なる。このようなパターンは、IPA ディクテーションプロジェクト [2] によってある程度整理されており、ソフトウェア ChaWan として実現されているので、これを後処理として利用するのが適当である。

連濁: 自立語と接尾辞や自立語同士が複合する際に、後部要素の語頭の清音が濁音化する現象。「水-不足 (ぶそく)」「桧-造り (づくり)」など。発音を正しく付与するために、形態素解析辞書にすべての複合語を登録するのは得策ではない。連濁に

はある程度の規則性があり (たとえば、[3])、後部要素の語種の情報 (和語か漢語か外来語か) などから、後処理で導出できる場合がある。その一方で、連濁規則には多くの例外があることも知られており (たとえば、漢語は原則として連濁しないといわれているが、「株式-会社 (がいしゃ)」は漢語にも関わらず連濁する)、これらは複合語として辞書登録するしか方法がない。後処理で扱うべきものと辞書登録すべきものの切り分けについては、十分検討する必要がある。

その他の音韻変化: 別の種類の音韻変化として、助詞「の」「に」や形式名詞「の」が「ん」になる現象や、ラ行動詞未然形「〜ら (ない)」が「〜ん (ない)」になる現象などもある。これらは辞書エントリの追加で対処できるが、精度よく解析するためには、このような現象が十分に生起している学習コーパスを用いる必要がある。さらに、一段動詞未然形の「れ ⇒ ん」変化現象 (例:「くれ (ない) ⇒ くん (ない)」) は、一段動詞の語幹の定義 (現行では「くれ」が語幹) を変えない限り、うまく扱えない。

発音の曖昧性: 品詞、表記とも同一であるのに発音が異なる語がある。接尾辞の「町 (まち)」「町 (ちょう)」のように慣用的な違いによるもの (例えば、「河原町 (かわらまち)」と「高山町 (たかやまちょう)」の違い) は、辞書に登録することで対応することになろう。また、「末 (すえ)」「末 (まつ)」のようにどちらでもよい場合とそうでない場合の両方があり得るもの、「今日 (きょう)」「今日 (こんにち)」のように必ず区別しなければならないもの等、様々なバリエーションがある。発音の誤りは、話し言葉に限らず、書き言葉でも同じように見られるが、その比率はあまり高い訳ではない。我々が扱っている RWCP コーパス [5] では、発音のみによる誤りは全体の 0.2%以下である。

以上のような問題点を持つ話し言葉コーパスに対し、現在の統計的学習に基づく形態素解析システムがどの程度自動的にわかち書き、品詞、発音情報を付与することができるかを次に実験により検討する。なお、書き言葉コーパスに対して、交差検定による実験では、現在の「茶筌」の学習機能では、わかち書き精度 99.1%(再現率 98.9%)、詳細な品詞付与の精度 97.7%(再現率 97.5%) である。

3 話し言葉コーパスによる形態素解析実験

現状では、統計的学習に充分足りるだけの品詞タグ付き話し言葉コーパスは (すくなくとも我々の品詞体系では) 存在しない。そこで、タグ付き書き言葉コーパスに少量のタグ付き話し言葉コーパスを追加することにより、どの程度の精度向上が得られるかを実験により確認した。ここで用いたタグ付きコーパスは次の 3 種類である。

1. RWCP コーパス [5]: 95 年版毎日新聞から抽出された 3000 記事。約 36000 文、92 万形態素。書き言葉コーパス。
2. AI 学会対話コーパス [4]: 人工知能学会の談話・対話研究のためのタグ付きコーパス。ホテルの予約、地図課題などの対話集。約 1500 文、8000 形態素。フィラーや非流暢性を多く含んだ、極めて自然な対話データ。以後、対話データと呼ぶ。
3. シニア支援システム会話コーパス: NEDO の支援による音声ディクテーションプロジェクト (奈良先端大鹿野教授代表) の一環として収集された会話コーパスの一部。料理、観光、病院等の案内例文集。500 文、4300 形態素。フィラーや非流暢性が取り除かれた (きれいな) 会話データ。以後、会話データと呼ぶ。

後者の 2 つのコーパスの一部をそれぞれ付録 1、付録 2 に示す。いずれも話し言葉コーパスではあるが、一方は、対話をそのまま転記したものであり、もう一方は非流暢性等を除去したものであり、話し言葉としては両極端の性質をもつコーパスと考えられる。

これらのデータに対して、次のような 2 種類の実験を行った。一つは、書き言葉コーパスによって学習したシステムが、自由な対話や会話に対してどの程度の精度で形態素解析が行えるか。また、自分とはかなり性質の異なる (少量の) 話し言葉コーパスを学習データとして書き言葉コーパスに追加することにより、話し言葉解析においてどの程度の精度の改善が得られるかを確認する実験である (実験 I)。もう一つは、自分と同様の性質をもった (少量の) 話し言葉コーパスを書き言葉コーパスに学習データとして追加することにより、どの程度の精度の改善が得られるかを確認する実験である (実験 II)。

学習データ	テスト (対話) 精度/再現率	テスト (会話) 精度/再現率
RWCP	92.85/90.04 86.12/83.51 81.45/78.98	99.21/99.14 99.00/98.93 98.16/98.09
RWCP+対話	- - -	99.67/99.37 99.51/99.21 98.90/98.61
RWCP+会話	93.22/90.63 87.48/85.05 82.78/80.48	- - -

表 2: 実験 I の結果

学習データ	(対話 1) 精度/再現率	(対話 2) 精度/再現率
RWCP	92.96/90.35 85.92/83.51 81.32/79.03	92.15/89.68 85.66/83.36 80.79/78.63
RWCP+対話 1	- - -	95.29/94.83 92.59/92.14 89.84/89.40
RWCP+対話 2	95.50/94.48 91.76/90.78 88.61/87.66	- - -

表 3: 実験 II の結果

3.1 実験と結果

実験 I では、まず, RWCP コーパスだけを使って学習した茶筌を用いて, 対話データと会話データを形態素解析した場合の精度を測定した (表 2 の最初の 3 行)。表中の一つの欄の中の 3 行は、上から (1) わかち書き、(2) 品詞大分類 (名詞、動詞等) まで、(3) 品詞の細分類まで、を考慮したものである。各行の数値は、スラッシュの前が精度 (precision)、後ろが再現率 (recall) である。この結果から、対話データについては、書き言葉で学習したシステムの性能が極めて低いことがわかる。一方、会話データについては、サンプルが極めてきれいな会話文であるため、書き言葉による学習でも高い解析精度が得られていることがわかる。表の上から第 2 欄目は、RWCP コーパスに対話コーパスを追加し、両者を学習データとして用いたシステムによる会話データの精度である。改善の絶対値は少ないが、誤りのうち、精度においては約 50%、再現率においては 30%弱の改善が見られ、少量であっても話し言葉データの追加が効果をもつことがわかる。表の最も下の欄は、逆に会話コーパスを学習データに追加した場合の結果である。改善率は低いが、誤りのうち、精度で 5〜10%、再現率では 6〜9%の改善が見られる。会話データの解析精度の改善に比べると効果は低い。

実験 II では、書き言葉コーパスに対して、類似の性質をもった話し言葉コーパスを追加することにより、解析精度がどのように影響を受けるかを観察した。対話コーパスを 2 等分し、それぞれを RWCP コーパスに追加することにより交差検定を行った。第 1 行と第 2 行、第 1 行と第 3 行を比較することにより、精度の改善の傾向がわかる。この実験の結果では、約 4000 語程度の話し言葉データ (書き言葉データの 5%以下の量) を追加することにより、精度と再現率が 40〜50%も改善されていることがわかる。

3.2 実験結果に対する考察

実験 I で特徴的なのは、たとえ話し言葉であっても会話データのようにきれいな文は、書き言葉とかなり共通の性質をもっていると考えられることである。ただし、一部の助詞や助動詞の使用法は書き言葉では稀なため、話し言葉データを学習データに追加することにより、更に改善の余地があることがわかる。一方、対話データには、フィラーや語断片などの非流暢性が多く含まれるため、きれいな話し言葉コーパスだけを追加しても改善の余地が少ないことが実験からも示されている。

実験 II で特徴的なのは、試験対象と同じような性質をもったデータを学習データに追加することにより、大幅な精度向上が達成できるということである。書き言葉コーパスだけでは、フィラーや一部の非流暢性などがまったく学習できていないため、たとえ少量のデータであっても、これらの話し言葉特有の現象をもつデータの追加が効果的である。ただし、絶対的な精度は、書き言葉やきれいな話し言葉には遠く及ばず、さらなるタグ付きコーパスの蓄積、あるいは、学習モデルの改良が必要である。

4 あとがき

話し言葉特有の言語現象を整理し、話し言葉の書き起こしデータに対する形態素解析の問題点について考察した。また、統計的学習に基づく形態素解析システムを用いて行った実験について報告した。書き言葉に比べて、用いられる表現や単語の違いに対応するためには、今後、辞書およびタグ付きコーパスの蓄積が重要である。

参考文献

[1] Masayuki Asahara and Yuji Matsumoto, "Extended Models and Tools for High-performance Part-of-speech Tagger," *Proc. 18th International Conference on Computational Linguistics*, pp.-, 2000.

[2] 河原 達也, 他, "日本語ディクテーション基本ソフトウェア (98 年度版)," 日本音響学会誌, Vol.56, No.4, pp.255–259, 2000.

[3] 佐藤 大和, "複合語におけるアクセント規則と連濁規則," 講座 日本語と日本語教育, 第 2 巻: 日本語の音声・音韻 (上), 杉藤 美代子 (編), 明治書院, pp.233–265, 1990.

[4] 人工知能学会 談話・対話研究におけるコーパス利用研究グループ, "様々な応用研究に向けた談話タグ付き音声対話コーパス", 人工知能学会研究会資料, SIG-SLUD-9903, pp.19-24, 2000.

[5] 新情報処理開発機構テキスト・サブ・ワーキンググループ, "研究開発用知的資源:タグ付きテキストコーパス報告書," 1998.

[6] 松本裕治, 形態素解析システム『茶筌』, 情報処理, Vol.41, No.11, pp.1208-1214, 2000.

[7] 前川 喜久雄, 籠宮 隆之, 小磯 花絵, 小椋 秀樹, 菊地 英明, "日本語話し言葉コーパスの設計," 音声研究, Vol.4, No.2, pp.51-61, 2000.

付録 1: AI 学会対話コーパスの例

あのすいません
はい京都観光ホテルでございます
宿泊の予約をお願いしたいんですが
ありがとうございます
では宿泊の人数とお日にちをお願いいたします
はいえ人数は１人です
はいお１人様ですね
えそして日にちがえ８月１０日から２泊でお願いしたいんですが
はい
８月１０日ご到着で１２日までの２泊でございますね
はい
しばらくお待ちください
お待たせいたしましたあいにくですが今シングルルームがいっぱいとなっておりますが
ああそうですか
他で
他の分では空いていませんか
あはい
え和室でしたら１泊１８０００円こちらのほうは空きがございますまたツインルームのシングルユース１４０００円こちらも空いてございますがいかがいたしましょう
ああそうですか
えそしたらツインルームのシングルユースのほうでお願いいたします

付録 2: シニア支援システムコーパスの例

今日のおすすめメニューは何ですか？
何かおいしいものを教えてください
おすすめのものを教えてください
今日のおすすめは何ですか？
手軽なメニューを検索
新築祝いを買いたいんですけど、二万円くらいでよいのはないですか？
何かおすすめ商品はありますか？
商品券が一番安く手にはいるお店を知っていますか？
この近くで時計がいろいろと売っている店はどこですか？
テレホンカードを買いたいんですけど
ちょっとしたおひたしを作りたいんですけど
京都風の料理はありますか？
煮物料理について教えてください
おでんの作り方を教えてください
中華が食べたい