1C−3

# 話し言葉に対する形態素解析*

竹下敦 take@ntthli.ntt.jp
福永博信† fuku@ntthli.ntt.jp
NTTヒューマンインタフェース研究所‡

## 1 はじめに

自然言語処理に関する従来の研究は書き言葉を対象としていた。しかしながら、音声認識の性能が向上したことにより音声対話等における話し言葉も研究対象とすることが可能となってきた。

自然言語処理ではまず形態素解析を行なうが、形態素レベルの言語現象として話し言葉に特有あるいは顕著なものがあるために、書き言葉を対象とした従来の形態素解析システムでは正しく解析できない。

本稿では話し言葉用の形態素解析システムを構築する際の問題点と我々がとった解決方法について述べる。

## 2 話し言葉への従来の形態素解析システムの適用

話し言葉データを人手で文字化した約3,400文に対して書き言葉用の形態素解析システムを適用した。解析が失敗したもののうち、話し言葉に特有な言語現象が原因であるものは1,664件あった。それらの原因を大まかに分類すると以下の様になるが、解析失敗は必ずしも1つの原因によるものではなく、幾つかの原因が複合して失敗を招いている場合もある。

(1) **話し言葉に特有の接続関係がある場合**
「もですね」、「ですかね」

(2) **語尾伸ばしや音便によって単語が変形する場合**
「えー」、「うーん」、「だくさん(= たくさん)」「なきゃ」「ちゃう」

(3) **話し言葉が辞書に未登録の場合**
「まあ」、「かなあ」、「ねえ」、「ほう」

(4) **話者の言い間違い、いいよどみなど言葉にならない単語を発話した場合**
「あざかな」=「あざやかな」
「前半・ね・え・と・ま・あ・中盤」

以下では上記のような話し言葉に特有な言語現象を含んだ発話文を正しく形態素解析できるようにするための連接表と辞書の構築法について述べる。例えば、「技術中で替わってもらわないと〜」を解析すると、「と」は正しくは接続助詞であるのに格助詞として解析される、というような形態素レベルでは避けられない純粋な解析ミスは扱わない。

*Morphological Analysis for Spoken Language
†TAKESHITA, Atsushi FUKUNAGA, Hironobu
‡NTT Human Interface Laboratories

## 3 話し言葉形態素解析システム

### 3.1 基本方針

従来からある書き言葉用形態素解析システムは、コスト最小法で実現されており、ヒューリスティックスとしては文節数最小法を用いている[1)]。

話し言葉用のものを実現する際の基本方針として、書き言葉用のものと共存させることとした。そのため、書き言葉用の**通常解析モード**に、話し言葉用の連接表と単語辞書をロードし、それらの言語データを優先して使うことによって**話し言葉解析モード**を実現するという形態を取った。

### 3.2 話し言葉特有の接続関係に対する対処

話し言葉では書き言葉では起こりにくい接続が頻繁に発生するので、それらの接続も許すように連接表を修正する必要がある。これらの接続は以下の3種類に分類することができる。

#### 3.2.1 口調を整えるための語が入る際の接続

接続規則の例を以下に示す。

(1) 係助詞「も」 + です
(例)「守るほうもですね」

(2) 付属語 + 助詞「ね」
(例)「〜た方がいいですかね。」

(3) 付属語 + 接続詞
例えば、係助詞「は」 + 接続詞
(例)「それはだから本人に聞いて…。」

#### 3.2.2 文が途中で終ったり倒置したりする際の接続

接続規則の例を以下に示す。

(1) 格助詞「に」 + 句点/文末
(例)「巨人は、ほんとうに。」

(2) 自立語 + 句点/文末
例えば、副詞 + 句点/文末
(例)「訓練のことでちょっと。」

#### 3.2.3 文が途中から始まる際の接続

接続規則の例を以下に示す。

文頭 + 格助詞「と」
(例)「と言うことは」

### 3.3 音便形への対処

音便形は、既に登録されている付属語の活用形として吸収できるものは話し言葉特有の活用語尾として吸収し、それ以外のものは話し言葉特有の連語として新たに辞書に登録した。言語知識の記述体系は書き言葉用のものと同じにした。

(1) 付属語の活用形で吸収した例
"**なきゃ**": 打消の助動詞「ない」の仮定形
"**せりゃ**": 使役の助動詞「せる」の仮定形
"**いきゃ**": 補助動詞「いく」の仮定形

(2) 活用のない連語として登録した例
「ては」がなまったもの。
"**ちゃ**": 接続助詞「ちゃ」
"**ちゃあ**": 接続助詞「ちゃあ」

(3) 活用のある連語として登録した例
「てしまう」の縮約されたもの。
"**ちゃわ**": 補助動詞「ちゃう」の未然形
"**ちゃい**": 補助動詞「ちゃう」の連用形
...
"**ちゃえ**": 補助動詞「ちゃう」の仮定形

### 3.4 話し言葉の辞書への追加と修正

#### 3.4.1 感動詞の細分化

感動詞は下に示すように機能的に細分して辞書に登録し直した。これは、対話理解のためには発話行為やその上位の談話プランを認識することは不可欠であり、そのために発話行為とそれを遂行するための言語表現の対応をリストアップしておくことは、話し言葉においては頻出する感動詞に対しては有効であるからである[2)]。

- 感動: (例)「ああ」、「あら」、「もお」
- 応答: (例)「はい」、「いーえ」、「ええ」
- 間投声: (例)「あのー」、「そのう」、「えー」
- 挨拶: (例)「こんにちは」、「こんばんは」
- 号令: (例)「気をつけ」、「ばんざい」

#### 3.4.2 その他の単語の追加

話し言葉でのみ現れる語彙を辞書に追加した。以下に例を示す。

終助詞「ねえ」、接続助詞「けども」、副助詞「なんか」

### 3.5 対処の困難な話し言葉現象

以下のような言語現象は対処が困難であるので、今回は対象外とした。

#### 3.5.1 語尾等の伸ばし

語尾等の伸ばしは頻繁に起こるが、一般的手法で対処したり、辞書に登録したりすることは困難である。したがって、以下のようにごく少数の頻出するものについてのみ辞書に登録することにより対処した、それ以外のものに対しては対処しなかった。

"**あー**": 話し言葉に多い間投声
"**うーん**": 話し言葉に多い間投声

#### 3.5.2 意味不明な発話

いいよどみや言い間違いが原因で人間が聞いても分からない以下のような発話は対処しようがないので、対象外とした。

(例1)「それを、あのー、せいじ、あのー」
(例2)「おお、大きい窓口の方に」
(例3)「今までは、だから、に、あのー」

### 3.6 話し言葉形態素解析システムの評価

テスト用の話し言葉データ1,700文に対して形態素解析システムの話し言葉解析モードを適用した。

解析失敗のうち、話し言葉に特有な言語現象が原因であるものは107件であり、話し言葉に対する対処を行なう前が約3,400文に対して1,664件であったことと比較すると、解析の精度が大きく改善されていることが分かる。

## 4 まとめ

話し言葉に特有な形態素レベルの言語現象の整理と、形態素解析システムにおいてそれらを扱う手法について述べた。書き言葉用の形態素解析システムに対して辞書の一部と連接表を話し言葉用に置き換えることにより実現した。

また、話し言葉データ1,700文に対して話し言葉形態素解析システムの評価を行なった結果、有効性が確認された。

謝辞
本研究にあたり、解析結果のチェック等を手伝って頂いたNTTアドバンステクノロジ(株)の木村淳子氏に深く感謝します。

## 参考文献

1) 吉村、日高、吉田: 「文節数最小法を用いたべた書き日本語文の形態素解析」、情報処理学会論文誌 Vol.24,No.1,pp.40-46(1983)

2) 竹下: 「プラン認識に影響を与える対話現象」、人工知能学会 第4回全国大会 (1990)