チャットのための日本語形態素解析

風間 淳一[†]　光石 豊[‡]　牧野 貴樹[‡]　鳥澤 健太郎[‡]　松田 晃一[§]　辻井 潤一[‡]
[†] 東京大学理学部　[‡] 東京大学大学院理学系研究科
[§] ソニー株式会社 PSD センター UI 開発部

1 はじめに

本論文では, インターネット上のチャットで使用されるようなくだけた文章も解析可能な, 日本語の形態素解析器を提案する. 近年, インターネット等のオンライン環境が普及し, そこではチャットルームや掲示板などの活動が活発である. また, ユーザが仮想空間で自由に動き回り他のユーザや仮想生物とチャットを楽しむような環境も出てきた. そのような仮想空間の一つが PAW[2, 3] であり, 我々のグループは PAW 中でのユーザと仮想生物との対話機能を自然言語処理の技術を使って強化する研究を始めている [6].

その第一段階として必要になるのが形態素解析である. 従来, 様々な形態素解析器が提案されてきたが, チャットで使用される文章は, 次に挙げるような, それらの形態素解析器が主に対象としてきた新聞の文章とは大きく異なる性質をもつ.

1. 文字の挿入や置換が起こりやすい. 例) は~い, きょーかしょ
2. ニックネームや仮想空間内の地名など普通でない文字列の固有名詞が使用される. 例) たけぽん
3. 平仮名が多用される. 例) どようびはしずかだね
4. 文末表現や叫び声などで意味不明な文字列が使用される. 例) ほえ？

従来の形態素解析器は, これらチャットの文章に特有の性質に対応していないため, チャットの文章を十分な精度で解析することができない.

本論文では, 上記の問題のうち, 1. の文字の挿入や置換に対する解決策を提案する. 我々は, 品詞 bi-gram モデルを基にした確率的形態素解析器を作成し, これをチャットの文章が解析できるよう拡張することを試みた. まず, 文字の挿入や置換が, 直前の文字や元の文字に依存していると仮定し, それを考慮に入れるように品詞 bi-gram モデルを拡張した.

2 音声的変形

チャットでは, 次のような文字の挿入や置換によって形が変化した語が頻繁に使用される.

うん、学校からでーす。
きょうがっこーいく？

第一の例では「です」に「ー」が挿入され,「でーす」に変化している. 第二の例では「がっこう」の「う」が「ー」に置き換えられて,「がっこー」に変化している. これらの文字の挿入や置換は文字の発音と関係していると考えられるので音声的変形と呼んでいる. このように単語の形が変化してしまうと, 形態素解析は辞書検索の段階で失敗することになる. チャット文中の音声的変形を分析するとこれらの文字の挿入や置換には, 図 1 に挙げるような性質がある. これらの性質をみると, 文字の挿入や置換は直前の文字が何であるかに依存していると考えられる. そこで, 直前の文字と挿入される文字や置換前後の文字の間の依存関係を反映するように品詞 bi-gram モデルを拡張し, あり得る挿入や置換には高い確率が, あり得ない挿入や置換には低い確率が与えられるようにした. 次節では, この品詞 bi-gram モデルの拡張について述べる.

3 品詞 bi-gram モデルの拡張

我々が基本とした品詞 bi-gram モデルでは, 各単語 m_iの品詞が t_iである単語列 $m_1 m_2 \cdots m_h$からな

母音字が, それと同じ母音を持つ文字の後に挿入される.	ちょっと → ちょぉっと
小文字の方が挿入されやすい.	ちょぉっと > ちょおっと
「っ」が挿入される.	おしえて → おしえてっ, でかい → でっかい
同じ文字の挿入が連続しやすい.	ちょぉっっっと, はーーい
直前の文字の母音が「o」の場合, 「お, う, ー, ~」が互いに置換可能である.	がっこう → がっこー, こうかん → こーかん
直前の文字の母音が「e」の場合, 「え, い, ー, ~」が互いに置換可能である.	めいわく → めえわく
同じ母音の文字が直前にある時, 母音を表す文字は「ー, ~」で置換される.	しいたけ → しーたけ

図 1: 音声的変形の性質と例

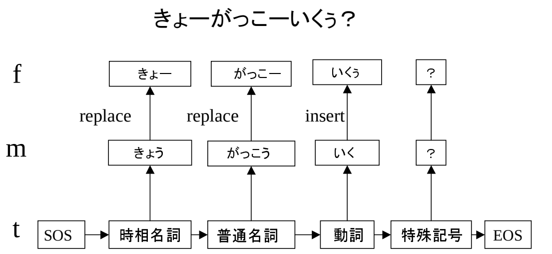

図 2: 拡張品詞 bi-gram モデルにおける文の生起

る文 Wは, 次のような確率で生起すると仮定されている[1].

$$P(W) = \prod_{i=1}^{h+1} P(t_i|t_{i-1})P(m_i|t_i)$$

品詞 bi-gram を基にした形態素解析は, 文 Wに対し, この確率を最大にする単語分割 $(m_1m_2\cdots m_h)$ と, 対応する品詞の列 $(t_1t_2\cdots t_h)$ を出力する [7]. 我々は, 2 節で述べた音声的変形を反映させるよう, この品詞 bi-gram モデルを次式のように拡張した (以下, 拡張品詞 bi-gram モデルと呼ぶ). 文は図 2 のように品詞 bi-gram モデルによって辞書にある拡張品詞 bi-gram モデルでは, 形態素 m_i が生成された後, ある確率で実際の文に現れている形 f_i が生成されると考える.

$$P(W) = \prod_{i=1}^{h+1} P(t_i|t_{i-1})P(m_i|t_i)P(f_i|m_i)$$

[1] t_0は文頭を表す, t_{h+1}は文末を表す特別な記号であり, それぞれ SOS, EOS と表す.

ただし,

$$P(f_i|m_i) = \begin{cases} 1 - \mathrm{TP} & \text{if } m_i = f_i \\ \mathrm{TP} \times P_t(f_i|m_i) & \text{if } m_i \neq f_i \end{cases}$$

$P(f_i|m_i)$ が新たに追加された項である. これは辞書中の単語 (m_i) の表記が実際に文中で現れている形 (f_i) に変形される確率である. TP は単語に何らかの変形が起きる確率であり, 現在のシステムでは定数[2]と仮定している. $P_t(f_i|m_i)$ は, 第 2 節で述べた文字の挿入や置換の直前の文字に対する依存性を考慮して, 次の式で計算する[3].

$$P_t(f_i|m_i) \approx \prod_{ins} L_{insert}(f_{i,ins}|f_{i,ins-1}) \times \prod_{rep} L_{replace}(f_{i,rep}|m_{i,o(rep)-1}m_{i,o(rep)})$$

$f_{i,k}$ は f_i 中の k 番目の文字, $m_{i,k}$ は語 m_i の表記中の k 番目の文字を表す. $o(rep)$ は $f_{i,rep}$ の置き換え前の文字の m_i の表記中での位置を表す. つまり, $m_{i,o(rep)}$ が置換された結果 $f_{i,rep}$ になったことを意味する. $L_{insert}(c_2|c_1)$ は, 文字 c_1 の後への文字 c_2 の挿入の起こりやすさであり, $L_{replace}(c_3|c_1c_2)$ は, 直前の文字が c_1 の時の, 文字 c_2 から c_3 への置換の起こりやすさである. 例えば,「がっこう」が「がぁっこー」に変形する確率 P(がぁっこー | がっこう) は次のように計算される.

$$L_{insert}(\text{ぁ}|\text{が}) \times L_{replace}(\text{ー}\,|\,\text{こう})$$

[2] 本研究の実験では $TP = 0.2$ としている

[3] この式は, L を確率とみると厳密な確率モデルとはいえないが, L_{insert}や $L_{replace}$はむしろペナルティとして働いている. 将来, 厳密な確率モデルに変更する予定である.

c_1	c_2	$log_e(L_{insert})$	例
あ	ぁ	-1.9	さあぁ
あ	っ	-1.9	さあっ
が	ぁ	-1.9	がぁっこう

表 1: $L_{insert}(c_2|c_1)$ の具体例

c_1	c_2	c_3	$log_e(L_{replace})$	例
お	う	ー	-1.8	おーさま
お	う	お	-1.8	おおさま
お	お	ー	-1.8	おーきい
お	お	う	-2.0	おうかみ
い	い	ぃ	-1.8	かっこいぃ
ふ	う	ー	-1.8	ふーせん

表 2: $L_{replace}(c_3|c_1c_2)$ の具体例

$$
\begin{aligned}
&= 0.150 \times 0.165 \\
&= 0.0248
\end{aligned}
$$

L_{insert}, $L_{replace}$は現在のところ, 直観で与えた値 (表 3,3) を使用しているが, 今後これらの音声的変形に対してタグ付けされたコーパスから推定する予定である.

上記のように文字が挿入されたり置換されたりした場合, 元の単語が辞書にあっても表記が変化したために辞書検索に失敗する. そこで, 我々のシステムでは, 入力文字中の文字を読み飛ばしたり置き換えたりしながら辞書検索することにより, 表記が変化してしまった語でも検索に成功するようにした.

4 実装

JUMAN[4] と同じ品詞体系, 活用体系を使用し, 辞書はJUMAN附属の辞書を変換して使用した. 辞書項目数は 783,603 であった. また, 品詞 bi-gram モデルのパラメータの推定には, 京都大学テキストコーパス [5] を使用した. これには SOS, EOS を含めて, 延べ 507,735 の形態素が含まれている. データスパースネスに対してスムージングなどは行なっていない. PAW のシステムは, ログインしているユーザのニックネームを知っているので, 起動時にニックネームリストを渡すことにより, それらを人名として辞書に追加する機能をもたせた[4]. また, 簡単な未知語処理として, 同種文字列[5]の抜き出しが実装してあり, 抜き出された形態素にはペナルティを与えて, 辞書中にある単語よりも優先されないようにした.

```
LINE = きょーがっこーはないよっ。
:ppr:きょー:きょう:きょう:*:名詞:時相名詞:
:pppr:がっこー:がっこう:がっこう:*:名詞:普通名詞:
:p:は:は:は:*:助詞:副助詞:
:pp:ない:ない:ない:ない:接尾辞:形容詞性述語接尾
辞:イ形容詞アウオ段:基本形:
:ps:よっ:よ:よ:*:助詞:終助詞:
:p:。:。:。:*:特殊:句点:
```

図 3: 音声的変形の解析成功例

5 実験

まず, 予備的な実験として我々の期待する動作をするかを調べた. 図 3 に我々のモデルがうまく働いた例を示す. 各行の第一の文字列は辞書検索ルーチンがどう文字を飛ばしたり置換して元の形に一致させたかを示すものであり, p は「何もせずに読み進む」, s は「読み飛ばし」, r は「置換」を意味する. 次に, 確率モデルの拡張によって, チャットの文に対する解析精度がどう改善されるかを実験した. テストコーパスには実際のチャットの文[6]を使用し, JUMAN, 我々のシステムで拡張していない品詞 bi-gram モデルの状態のもの, 拡張品詞 bi-gram モデルの状態のものの単語の切り分けの適合率を人手で測定した. 但し, チャットで多用される顔文字[7] が一つの単語として出力されない場合 (JUMAN ではほとんど出力されず, 我々のシステムにおいては全く出力されない) 誤りとした. また, 単語の最後にのみ文字が挿入されている場合, その文字を別の単語として分割 (JUMAN) しても, 単語の一部として出力 (我々のシステム) しても正解とした. これは, そのような場合 (特にそれが文末のときは) 他の部分の解析にあまり影響しないと考えられるからである.「あぁ～～～」などの叫び声の類は一つ

[4]ニックネームは辞書に登録されていない場合が多い. こうすることで未知語となってしまうことを避けることができる.

[5]カタカナ, アルファベット, 数字

[6]PAW のログを使用した.

[7](^0^) ^-^; 等

	N_{COR}/N_{SYS}	適合率
JUMAN	878/1086	80.8%
品詞 bi-gram	793/964	82.3%
拡張品詞 bi-gram	834/965	86.4%

表 3: チャット文の単語切り分けの適合率

にまとめられるものを正解とした. また, 辞書にない語で「まうまう」のように繰り返しがある場合は, 繰り返しの単位で切れるものを正解とした. 適合率は, システムの総出力単語数を N_{SYS}, そのうち切り分けが正しいものの総数を N_{COR}とすると, N_{COR}/N_{SYS}で計算される [1]. チャットの文, 300 文を解析した結果が, 表 3 である. 品詞 bi-gram のみの状態で JUMAN より精度が良いのは, ニックネームの追加によるところが大きいと思われる. N_{SYS}の値を見ると, チャットの文では, 一文あたりの単語数が平均 3 単語程度と, 非常に短い文が多いことが分かる. 拡張品詞 bi-gram モデルの状態の値がかなり上がっているのは, テストコーパス中に「は~い」などの我々の拡張に有利に働く単語が多く含まれていたためと考えられる.

6 まとめ

今回, チャットの文の実用的な形態素解析を実現するため, チャットの文に頻繁に現れる音声的変形に注目し, 確率的形態素解析器に対しそれを反映する変更を施した. 実際のチャットの文に対する実験によって, 我々の手法がチャットの文に対して有効であることを確認した. しかし, 十分な精度を達成しているとはいえず, さらなる改良が必要である. 改良については, 大きく分けて以下の 3 つが挙げられる.

- 文字の置換は 1 文字のみと仮定しているため, 「ます」から「ましゅ」への変形のような文字数が増える変形や, 「どうして」から「どして」への変形のような文字が省略される変形は扱えない. このような変形にも対処する必要がある.
- 音声的変形の確率モデルについては, まず, 数学的に厳密なものにする必要がある. また, 直前の 1 文字に注目しているのみで, 全ての音声的変形の現象を十分反映しているとはいえないので, 音声的変形をさらにうまく説明するように確率モデルを改良していく必要がある.
- 我々の形態素解析器では文字の挿入や置換を扱えるようにしたために, 単語候補が多数生成され, 解析時間が増大してしまうという問題がある. 上のような変形にも対応した場合, 単語候補の数はさらに増えると考えられ, 実際のシステムで使用するためには, 何らかの高速化が必要である.

謝辞　本研究では (株) 日本電子化辞書研究所, 京都大学の許諾を得て JUMAN 附属の辞書を利用させて頂きました. 心より感謝いたします.

参考文献

[1] Masaaki Nagata. A stochastic Japanese morphological analyzer using a forward-dp backward-A^* n-best search algorithm. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 201–207, 1994.

[2] ようこそ PAW へ. http://www.so-net.ne.jp/paw/index-j.html.

[3] 松田晃一. 不思議な島をペットと歩こう！インターネット上の共有仮想世界 PAW. *bit*, Vol. 30, No. 9, pp. 2–10, 1998.

[4] 黒橋禎夫, 長尾真. 日本語形態素解析システム JUMAN version 3.5, 1998.

[5] 黒橋禎夫, 長尾眞. 京大テキストコーパス・プロジェクト. 言語処理学会 第 3 回年次大会, pp. 115–118, 1997.

[6] 定政邦彦, 牧野貴樹, 光石豊, 鳥澤健太郎, 松田晃一, 辻井潤一. 「パーソナルエージェント用自然言語インターフェース」開発ツールキット (PANLI toolkit). 言語処理学会第 5 回年次大会発表論文集. 言語処理学会, 1999.

[7] 松本裕治, 影山太郎, 永田昌明, 齋藤洋典, 徳永健伸. 岩波講座 言語の科学 3 単語と辞書. 岩波書店, 1997. ISBN 4-00-010853-0.