２ちゃんねる解析用の形態素解析器の作成

早藤健　建石由佳

工学院大学情報学部
Email: j106097@ns.kogakuin.ac.jp,yucca@cc.kogakuin.ac.jp

１．　はじめに

本研究は、インターネット上の大型掲示板「２ちゃんねる」[1]で使われている日本語の形態素解析の解析精度の向上を目的とした研究である。

現在、日本のインターネット上のコミュニティでは従来の日本語とは異なる言葉の使い方をしている。そこで使われる日本語は、視覚的、感覚的に文字を使用し通常の日本語とは性質が異なるので、従来の日本語の文法を使った形態素解析では解析の精度が期待できない。本研究は、代表的なコミュニティである「２ちゃんねる」に対象を絞り、２ちゃんねる用語を含む文章の形態素解析の精度向上を目的とする。インターネット上の２ちゃんねる用語集[2]をもとに、日本語形態素解析器 MeCab[3]の辞書、品詞体系、接続コストを整備することで２ちゃんねる解析用の形態素解析器のプロトタイプを作成し、整備前と後の MeCab で解析精度の比較を行った。

２．２ちゃんねるで使われる特殊な文字や記号を使った表現

２ちゃんねるでは、通常の日本語では使われない、以下のような表現が多用される。

(1)複数の文字や記号を組み合わせて１つの文字を表現するもの

例１：組み合わせ文字

文字	説明
糸冬	糸と冬という漢字を並べて「終」を表現。
ネ申	カタカナのネと申という文字を並べて「神」を表現。
メ几 木又	メと几、木と又をいう文字を 2 行に渡り並べて「殺」を表現。
タヒ	カタカナのタとヒを並べて「死」を表現。上の横棒がないので不完全だが使用されている。

文字を大きくする代り、検索をよけるための当て字などの目的で使われる（例 1）。

(2)顔文字、アスキーアート

(2-1)顔文字

文字や記号を組み合わせて表情を表現している。アスキーアートの一種であり、その中でも１行で表現できるものを指す。文章にさまざまなニュアンスを加えるために用いる。最近では携帯電話のメールにもよく使われる(例２)。

例２：顔文字

(ﾟ∀ﾟ)　(・∀・)　(ﾟдﾟ)　(ﾟДﾟ)
ヽ(`Д´)ﾉ　＼(^o^)／　⊂ミ⊃＾ω＾）⊃

(ﾟдﾟ)や(ﾟДﾟ)といった微妙な違いのものでも意味や使い方に違いがあるものが存在する。また、例２下段のように、表情だけでなく身振りなどの動作を表わすものもある。

(2-2)アスキーアート

写真や画像を貼ることができない掲示板では、アスキーアートで代用して表現するといった工夫がされている。表現の幅が広がるので２ちゃんねるにおいても多用されており、アスキーアートから様々な２ちゃんねるのキャラクターが生まれている(例 3)。

例 3:アスキーアート

```
              ∧ ∧
～′￣￣(´ー`)        ∧＿∧
  UU￣￣  U U        (´・ω・)
                     (      )
                      u―u′
```

(3)カタカナやひらがな、全角や半角などを混合してパターンを変えた表現（例 4）

例４：文字種の変更

意味	表現
熱い	アツイ、アツい、ｱﾂｲ、ｱﾂい、あつい
落ち着け	おちつけ、オチツケ、ｵﾁﾂｹ、落ちつけ、おち着け

(4)似ている文字や記号を使ってパターンを変えた表現(例 5)

例５：形の似た文字の使用

意味	表現
アンパン	アソパソ（カタカナの「ン」を「ソ」で表記）
あやしい	ぁやιぃ（小文字とギリシャ文字に置き換えて表記）

(5)文字の順番を入れ替えてパターンを変えた表現(例 6)

例６：順序の入れ替え

意味	表現
落ち着け	おちけつ
ポイント	ポイトン

(6)感覚で単語のパターンを変えた表現（例７）

例７：感覚的な表現

意味	表現
笑った	ワロタ、ワラタ、ワロス
酷い	ヒドス
落ち着け	もちつけ

これらは視覚的効果、言い間違いに似せて面白くする、検索避け、などの目的で使われる。

２ちゃんねるで多用されるこれらの表現は、有志によって用語集などにまとめられている。また、２ちゃんねるの外でも、各種掲示板、メール等でも使われ、中には日常生活で使われて始めているものもある。

３． MeCab の単語登録機能を使ったユーザ辞書の作成

はじめにユーザ辞書に登録する単語を収集する。今回は 2 ちゃんねる用語のまとめサイトである 2 典 Plus[2]からデータを収集、その後、品詞別に分けてまとめていく。この際、登録不要と考えられる単語を取り除いておく（3-1-1 参照）。次に MeCab の登録形式に従い、表層形,左文脈 ID,右文脈 ID,コスト,品詞,品詞細分類 1,品詞細分類 2,品詞細分類 3,活用形,活用型,原形,読み,発音を決めてテキストファイルに保存する。ここで発音の後に情報を追加できるので、検索時にわかりやすくなるよう「２ｃｈ」「顔文字」「ＡＡ（アスキーアート)」などの情報を追加する。

（例．２ちゃんねる,1285,1285,8000,名詞,一般,*,*,*,*,*,２ちゃんねる,ニチャンネル,ニチャンネル,2ch）

その後、登録用の形式に直したすべての単語を CSV ファイルとして保存し、MeCab にユーザ辞書として登録する。

３－１．登録不要と考えた単語の種類

以下のような単語は登録しなかった。

(1)「サーバー」の意味で使われる「鯖」、「アカウント」の意味で使われる「垢」などのように、一般の単語として存在していて、本来の意味とは違った使い方をされているものは登録することによって一般の単語の形態素解析に支障がでると考え、登録しなかった。

(2)「オーディエンス」や「ゲートキーパー」など外国語をカタカナ表記したもの（外来語)、「荒らし」、「煽る」など一般でも使われている単語、作品名や組織名、人名などの固有名詞や専門用語など一般の単語だと考えられるものは登録しなかった。ただし、２ちゃんねる発祥のものや関係の深いもの、ニックネーム、略語は除外せず登録した。

(3)「実況スレ」(実況＋スレ[1])、「アニメヲタク」(アニメ＋ヲタク）のようにさらに細かい単語に分けられると考えられるものは単体では登録しなかった。ただし、「アニヲタ」(アニメヲタクの意）のように分けられないもの（この場合、アニ＋ヲタと分けることはできない）は１つの単語であると考えて登録した。

[1] 「スレ」は「スレッド」の意味

３－２．特殊な品詞の単語について

「禿同」という単語がある。これは「激しく同意」という意味であり「激しく同意」→「激同」→「禿同」と変化して現在に至る。品詞の異なる単語がまとめられたものなので品詞の判断が難しい。よって本研究ではこういった単語の品詞を成句として登録した。また、「ワロタ」（動詞であり「笑っ＋た」の複合語）など品詞は明らかであるが異なる品詞の単語から成る単語については成句という情報を品詞細分類の部分に記述した。文脈 ID の割り振りについては、その単語を普通の日本語に直した場合に一番左にくる単語の品詞を左文脈 ID、一番右にくる単語の品詞を右文脈 ID とした。（例.「ワロタ」を普通の日本語に直すと「笑った」なので、笑っ（動詞）＋た（助動詞）→ 左文脈 ID は動詞､右文脈 ID は助動詞）

４．作成したユーザ辞書の登録前と後での解析結果の比較

前節に挙げた方針で用語を登録する前と後で形態素解析の精度を測定した。

まず２ちゃんねる[1]や２ちゃんねるスレッドのまとめサイト[4]からスレッドを選択、文章をすべてテキストファイルに保存する。次に、手作業で文章中から２ちゃんねる用語と考えられるもの、特殊な表現（２－１の(1)～(6)）の前後を形態素に切り分けて、可能ならば品詞を付けていく。これを正解として、作成したユーザ辞書を登録する前と後の MeCab で形態素解析を行い、解析結果の精度を手作業で比較する。ただし、レス番号の行は解析の対象から除外する。（固定のパターンのものが多いため）

５． 結果

４節の手順で２回実験を行った。実験１は２ちゃんねる[1]からランダムにとった６０９レス、実験２は２ちゃんねるのまとめサイト[4]からとった２２０レスに対して行った。

実験１

ユーザ辞書の登録	登録前	登録後
①手動単語抽出数	570	
②切り出し成功	493(86.5%)	515(90.4%)
③品詞正解	70(12.3%)	349(61.2%)
④誤検出		57

実験２

ユーザ辞書の登録	登録前	登録後
①手動単語抽出数	138	
②切り出し成功	120(87.0%)	121(87.7%)
③品詞正解	26(18.8%)	78(56.5%)
④誤検出		0

結果の各項目は以下のようになっている。

①…MeCab を使用せずに手作業で抽出した２ちゃんねる特有の表現の抽出数。②…手動で抽出した表現について MeCab で形態素解析を行い、２ちゃんねる用語の切り出しが正しくできていた数。③…②の中で品詞の解析も正しく行うことができた数。④…２ちゃんねる用語登録後の MeCab で、レスの行を除く文全体を形態素解析した際、一般の単語を登録した単語だと誤認識してしまった数。

実験１の④の誤検出の原因の大半はリンクとして貼られているアドレスの一部を単語として認識してしまったパターンである。(例８)

例８：誤検出のパターン

http://yasai.2ch.net～とあった場合、登録した「2ch」という単語を検出。

４－２．実験の評価

実験１、２ともに②の切り出しにおいてユーザ辞書登録前と後であまり差がでなかったのは、MeCab の標準機能で半角やカタカナで記述している単語は１つの名詞として認識しているためと考えられる。そのため、ユーザ辞書登録前は切り出しに成功したが、品詞の判別に失敗したというパターンが多かった(例９)。ユーザ辞書に登録することで品詞も正しく判断できた。

例9：切り出しに成功したが品詞の判別に失敗したパターン

例文	おまえの発想にワロタ
登録前	おまえ / の / 発想 / に / ワロタ 名詞 助詞 名詞 助詞 名詞
登録後	おまえ / の / 発想 / に / ワロタ 名詞 助詞 名詞 助詞 動詞

登録した通りに記述されている単語については、登録することよって解析の精度を向上させることができたといえる。しかし、ユーザ辞書に登録したものと少しでもパターンが違うだけで認識することができなくなった(例１０)。

例10：微妙な違いにより認識に失敗するパターン

登録単語	認識に失敗するパターン
(･ω･)	(･ω･)(･ω･´)(;ω;)
	認識に成功するパターン
	(･ω･) d(･ω･)

このパターン以外にも例８のようにリンクとしてアドレスが貼られていて、その中に登録した単語が含まれている場合に誤認識してしまう事があった。

アスキーアートについては複数の行を視覚的に使っているため、ユーザ辞書への登録が難しい。また、1 行ごとの処理をベースとした解析では処理を行うことができなかった。

５． 考察

本研究で解析の精度が上げられなかったパターンを変えた表現、視覚的、感覚的表現は現代のコミュニティではたびたび使用されるものであり、今後さらに複雑化していくことが考えられる。また、２ちゃんねるは様々なジャンルのスレッドが存在し、スレッドによって文の表現や使われる用語が変わる。そのため、本研究で行った辞書登録だけでは不足していると考えられる。また、スレッドを変えて実験を行う度に違う結果がでると考えられる。すべてのパターンを推測して、登録することにより形態素解析の精度を向上させることも可能であるが、推測には限度があり、登録する単語数も膨大な数になるのであまり効率が良くない。文字の順番、全角半角ひらがなカタカナ、スペース、当て字などでパターンを変えて表現しているものに関しては、MeCab に登録した単語をもとにパターンが違っていても、ある程度推測する機能（Web 検索などで使われているようなもの）を追加することが良いと考えられる。また、アドレスなどのある程度パターンのある文字列（http～など）はそのパターンがでてきた場合、その文字列の終わりまでを１つの形態素として解析するように設定するのが望ましい。アスキーアートや複数の文字で１つの文字を表現するなど、視覚的な表現をしているものについては MeCab で形態素解析を行うことは難しい。MeCab での解析を行わず、画像認識など他の分野の技術で前処理を行うのが望ましい。

６．おわりに

本研究では、辞書登録で品詞の誤りが減らせることがわかった。また、２ちゃんねる特有の表現をしている文を形態素解析する際に失敗しやすいパターンを見つけることができた。このパターンに対して対策していくことでさらなる形態素解析の精度向上が期待できる。

参考文献

[1] ２ちゃんねる http://www.2ch.net
サンプル収集日 2009 年 10 月 1 日
[2]2 典 plus, http://www.media-k.co.jp/jiten/
サンプル収集日 2009 年 7 月 6～7 日
[3] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237 (2004.)
http://mecab.sourceforge.net/
[4]痛いニュース
http://blog.livedoor.jp/dqnplus/
サンプル収集日 2010 年 1 月 1 ～5 日