口語的表現を含む日本語文の形態素解析

1B−2

竹元義美　福島俊一
(NEC C&Cシステム研究所)

1 はじめに

従来の形態素解析[1]は、書き言葉を中心に研究されていたため、話し言葉に特有の表現への対処が十分に行われていないという問題があった。竹下ら[2]は、話し言葉に対する形態素解析手法として、話し言葉に特有な言い回しを整理し、辞書登録を行っている。しかし、広い分野のテキストを対象とした自然言語処理を想定した場合に、話し言葉特有の言い回しへの対処だけではまだ十分とは言えない。

本稿では、話し言葉特有の言い回しおよびテキスト特有の表記による強調表現を口語的表現として捉え、口語的表現を含む日本語文の形態素解析手法とその評価について述べる。

2 解析誤りを引き起こす口語的表現の分類

口語的表現に対する処理を特別に行っていない形態素解析を、口語的表現が頻出するテキストを対象に実行し、口語的表現が引き起こした誤り箇所を分析した。分析の対象には、週刊誌から特に口語調エッセイや対話文を取り出したテキスト(約1.6万字)を用いた。形態素解析は、字種の変化に基づいて定めた区間を最小文節数～最小文節数+1でカバーする候補の中から、ヒューリスティックスにより第一候補を決定する方式をとった[3]。また、解析用辞書には50万語の辞書を用いた。解析誤りとしては、特に文節区切り誤りに着目し、それらは人手で作成した文節区切り正解と形態素解析による文節区切り結果を比較して抽出した。文節区切り精度は図1(a)のようになり、文節区切り失敗は口語的表現によるものが大半であった。文節区切りに失敗した口語的表現を調査して以下に示すように3項目に分類した。

1. 話し言葉に特有な言い回し
   - 音の簡便化　(例)困っちゃう
   - 特有な語の接続　(例)なるほど【ねえ】
2. 表記による強調表現
   - 意図的な片仮名表記　(例)ガンバって
   - 特殊な文字の挿入・追加
     (例)ず【ー】っと、冷た【ぁ】い
3. その他
   - 方言・古語　(例)かっこいい【じゃん】
   - 擬音語・擬態語　(例)わははははは

話し言葉特有の言い回しとして、まず、音の簡便化が起こりやすい。例えば、書き言葉における“困ってしまう”が話し言葉では“困っちゃう”に変化する。また、“なるほどねえ”の“ねえ”のように話し言葉に特有の語が接続しやすい。これらの例に対して話し言葉特有の言い回しに対応していない形態素解析を実行すると、解析に失敗して文節区切り誤りを引き起こす。

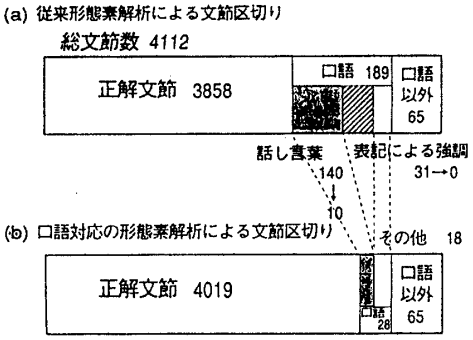

図1:文節区切り評価結果

一方、表記による強調表現もテキスト特有の口語的表現として捉える。テキスト上では、通常は平仮名で表記する単語を書き手が意図的に片仮名で表記したり、長音など特殊な文字を単語に挿入または追加したりして読み手の印象を強める場合がある。特殊な文字とは、具体的には、長音(“ー”、“～”)や“ぁ”、…、“ぉ”などの文字である。このような場合、通常の表記は辞書に登録されていても未知語となってしまう。

本稿では、上記の1と2に対処し、3は件数が少なかったため対処しなかった。以下、1と2の対処方法を3節で述べ、その評価結果を4節に示す。

3 口語的表現への対処方法

3.1 話し言葉特有の言い回しへの対処

比較的大規模な週刊誌テキスト(約29万文字)を対象とした文節区切り誤りの分析に基づいて話し言葉に特有な言い回しを整理し、それらを[2]と同様に助動詞相当語あるいは助詞相当語として辞書登録した。登録した単語は、活用を含めて138件である。登録した単語の例を表1に示す。

3.2 表記による強調表現への対処

意図的な片仮名表記や特殊文字を用いた表記による強調表現は、3.1節のような辞書登録で対処しようとすると、登録しなければならない表記パターンが多すぎるため現実的でない。そこで、3.2.1、3.2.2節で説明するような片仮名列置換検索処理と特殊文字置換検索処理とを導入する。これらの処理を導入する前の形態素解析には、句読点などで区切られた区間ごとに一括して辞書検索を行い、単語の候補をすべて辞書から取り出してしまってから、接続検定・候補選択を行うような流れを想定している。そして、片仮名列置換検索処理と特殊文字置換検索処理は、通常の辞書検索処理の直後に組み込む。すなわち、辞書検索→「片仮名列置換検索」→「特殊文字置換検索」→接続検定→候補選択という流れとする。片仮名列置換検索処理と特殊文字置換検索処理は、辞書検索結果としての単語の候補を追加す

A morphological analysis method
for colloquial Japanese text

Yoshikazu TAKEMOTO and Toshikazu FUKUSHIMA
C&C Systems Research Laboratories, NEC Corporation

	書き言葉との対応		例
助動詞相当語	ちゃう	接続助詞「て」＋補助動詞「しまう」	困っちゃう
	ちまう	接続助詞「て」＋補助動詞「しまう」	捨てちまう
	てる	接続助詞「て」＋補助動詞「いる」	来てる
	じゃう	接続助詞「で」＋補助動詞「しまう」	悩んじゃう
	じまう	接続助詞「で」＋補助動詞「しまう」	死んじまった
	でる	接続助詞「で」＋補助動詞「いる」	泳いでる
	じゃ	断定助動詞「だ」	冗談じゃない
助詞相当語	ちゃ	接続助詞「て」＋係助詞「は」	急がなくちゃ
	じゃ	格助詞「で」＋係助詞「は」	彼じゃ無理だ
		形容動詞連用形「で」	穏やかじゃない
	なきゃ	打ち消し助動詞「ない」仮定形＋接続助詞「ば」	行かなきゃ
	ねえ	終助詞「ね」	なるほどねえ

表1:登録した話し言葉の例

るように働くので、接続検定以降の処理は変更の必要はない。

3.2.1 片仮名列置換検索処理

片仮名列置換検索処理は、辞書から単語が検索されていない片仮名列を平仮名に置換し、辞書を再検索する処理である。例えば、単語辞書に“がんば(る)”はあるが“ガンバ(る)”がなかったとしても、“ガンバった”を“がんばった”に直して再検索するので、“ガンバ(る)”という単語の候補が得られる。以下では、片仮名列置換検索の適用条件・適用範囲などの詳細を述べる。

1. テキスト中の片仮名列のうち、長さが3文字以上の単語の組み合わせでカバーされないものに適用する。
2. 片仮名列を平仮名列に置換する範囲は、片仮名列の先頭から最長一致で単語をつないでいったときに3文字以上の単語がとれなかった部分とする。
3. 再検索では、平仮名に置換した範囲より前方の字種境界位置(平仮名から漢字や片仮名に変化した位置、または、句読点の直後など)から置換範囲の末尾までの各文字位置を先頭とする単語を検索する。検索された単語の末尾位置は置換範囲を越えてもよい。
4. 置換範囲の前方より再検索した場合、置換範囲に届かない単語は無効とする(通常検索で既に得られているため)。

1、2において3文字以上としているのは、「拳銃」の意味の“ガン”が辞書登録されていた場合に例えば“ガンバった”に対して上記の処理が行われなくなる(つまり、“バ”についてのみ平仮名に置換して再検索されることになる)ことを防ぐためである。

3は、片仮名列が漢字・平仮名混じりの単語の一部になっている場合（例:“捨てゼリフ”が1単語として登録されている場合など）に備えるためである。

本処理は、辞書から検索される単語の候補を増やすことを目的とするため、本処理を用いて検索された単語はすべて意図的に片仮名されているものという保証はない。従って、本処理で検索された単語は、評価値を悪くすることによって文節候補選択時への悪影響を抑える必要がある。

3.2.2 特殊文字置換検索処理

特殊文字置換検索処理は、2節で述べた特殊文字の位置で、その文字を含む単語が検索されなかった場合、以下のように特殊文字を削除または置換して辞書を再検索する処理である。

1. 特殊文字を削除して再検索する。再検索は、削除した位置より前方の字種境界位置から削除位置までの各文字位置を先頭とする単語を検索する。ただし、削除位置をまたがる単語でない場合は無効とする。
2. 直前の文字との音韻的な規則に基づいて特殊文字を置換して再検索する。再検索は、置換した位置より前方の字種境界位置から置換位置までの各文字位置を先頭とする単語について行う。ただし、置換位置を含む単語でない場合は無効とする。

1の処理では、例えば“ずーっと”を“ずっと”に直して再検索できる。2の音韻的な規則の例としては、え段の直後の長音を“い”または“え”に置換する、お段の直後の長音を“う”に置換するなどがある。2の処理では、例えば“どー する”を“どうする”に直して再検索できる。

4 評価

4.1 評価方法

2節の分析に用いたのと同じテキストを対象として、提案した手法の効果を評価した。まず、3.1節で述べた話し言葉に特有な言い回しを辞書に追加登録して形態素解析を実行し、追加登録前と文節区切り誤り箇所の変化を調べた。次に、テキストから片仮名列の出現箇所と特殊文字の出現箇所を抽出し、その各々について3.2節で述べた置換検索処理を適用した場合の結果を机上シミュレーションで求めた。

4.2 評価結果および考察

図1に、口語的表現に対処する前(2節の分析結果)と比較した対処後の文節区切り精度とその内訳を示す。文節区切り正解率は93.8%から97.8%に向上した。話し言葉に特有の言い回しによる誤りは140件から10件に減少し、強調表現による誤り31件はなくなった。

話し言葉に特有の言い回しによる誤りで、今回の辞書登録で対処できなかった例には次のようなものがある。

1. “作っちゃ／うんだ” (正解:作っちゃうんだ)
2. “いう／かさ” (正解:いうかさ)
3. “外はねーだろ” (正解:外は／ねーだろ)

1、2は、辞書登録や接続表修正が不十分なのが原因であった。また、助動詞相当語や助詞相当語だけでなく、3のように話し言葉特有の自立語についても登録する必要がある。なお、今回評価したテキストでは、置換検索処理によって、それまで解析に成功していた箇所が失敗するという副作用はなかった。

5 まとめ

口語的表現を含む日本語文の形態素解析手法を提案し、評価した。提案する手法では、話し言葉に特有な言い回しを助動詞相当語・助詞相当語として辞書登録し、意図的な片仮名表記や長音などの特殊文字を用いた表記による強調表現を適当な文字列に置換して辞書を再検索する。特有な言い回しは実際に辞書登録し、置換検索は机上シミュレーションにより評価し、その有効性を確認した。

今後は、3.2節で述べた置換検索処理を実現して、より大規模な週刊誌テキストで評価する予定である。

謝辞 評価用テキストを提供してくださった、株式会社 講談社に深く感謝致します。また、日頃から有益な助言をくださるNEC C&Cシステム研究所 山田氏に感謝します。

参考文献

[1] 長尾監修,日本語情報処理,電子通信学会編,1984
[2] 竹下他,情処42全大1C-3,1991
[3] 福島他,情処45全大6C-1,1992