

仮名漢字変換 S/W における口語表現入力機能の改良

3H-6

相川 勇之 高山 泰博 鈴木 克志

三菱電機株式会社 情報技術総合研究所

1. はじめに

i モードをはじめとする携帯機器の普及により、電子メールなどで日本語を入力する機会が急増している。日本語入力機能としては仮名漢字変換方式が主流であり、携帯電話をはじめとして今後普及が進むと予想されるデジタルTVやカーナビなどに組み込んで利用される組込用の仮名漢字変換 S/W に対するニーズが高まっている。われわれは二文節最長一致法[1]に基づく、組込用途に特化した省メモリの連文節変換 S/W を開発した。

多様な文書を作成するために利用されるパソコン上の仮名漢字変換とは異なり、携帯機器では電子メールの入力が主要な用途となっている。携帯機器で入力するメールはプライベートなものが中心であるため口語的な表現が多くなる。とくに i モードなど携帯電話の利用者には若年層が多く、電子メール、チャット、掲示板への書き込み等において口語調のくだけた表現が多用される。

本稿では、仮名漢字変換 S/W における口語表現入力機能の改良について述べる。携帯機器ユーザによる書き込みが多いと思われる掲示板のデータを収集し、この収集データをもとに付属語辞書を改良した。また主観評価による変換精度測定実験を行ない、改良に効果のあることを確認した。

2. 口語表現入力における課題

電子メールなどでの口語表現入力における仮名漢字変換の課題として以下の 2 点があげられる。

- (1) 文節末表現
- (2) 自立語の変形

前者の例として、「やったあ」など文節末の音を伸ばすことを表現するために付与される「あ」や「…だよん」の「よん」などがある。後者の例として、

An improvement of Kana Kanji Conversion program concerned with colloquial style inputs

Takeyuki AIKAWA, Yasuhiro TAKAYAMA, Katsushi SUZUKI
Mitsubishi Electric Corporation.

5-1-1 Ofuna, Kamakura, Kanagawa 247-8501, JAPAN

「すごい」「いつも」を強調した表現である「すんごい」「いっつも」などがある。

上記の口語表現には平仮名表記が多く、単文節変換では読みを入力してそのまま確定すれば良いのであまり問題にはならない。しかし連文節変換では、これらの口語表現が文の一部として入力されるので、解析に失敗するとその前後にまで影響が及び、深刻な誤変換の原因となる。

次章では、課題 (1) を解決するための付属語辞書の改良について述べる。

3. 口語表現入力機能の強化

図 1 に口語表現入力機能の強化手順を示す。携帯機器ユーザによる書き込みが多いと思われる掲示板のデータを収集し、この収集データをもとに付属語辞書を改良する。字面の統計情報を用いてコーパスから口語表現を自動抽出する方式[2]も提案されているが、自動抽出には課題が多く精度も不十分であるため本稿では手作業を中心とした口語表現抽出を行なう。

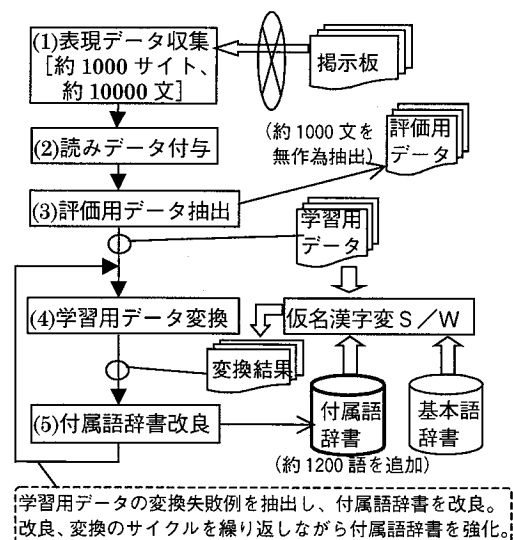


図 1 口語入力機能の改良

(1) 掲示板データの収集

携帯機器による生の入力データが数多く得られる掲示板に書き込まれた文を収集対象とする。掲示板への書き込みは不特定多数へのメッセージであるため電子メールとはやや異なる部分もあるが、ネット語とでも言うべき特有の表現が数多く見られ、電子メール入力と文体が共通すると思われる。また、もともと不特定多数向けに書かれた文であるため、プライバシーに関する配慮が欠かせない電子メールに比べて大量のデータを収集しやすいという利点がある。データの偏りを防ぐため収集サイト数を多めに設定し(約 1000 サイト)、各サイトごとに 10 文ずつ約 10000 文を収集した。

(2) 読みの付与

仮名漢字変換 S/W の口語入力機能強化のためには、上記データの変換結果を分析する必要がある。チューニングを繰り返すために、上記(1)に対する読みデータを手作業で作成した。

(3) 評価用データの抽出

評価用データとして、収集データから 1 割(約 1000 文)を無作為に抽出した。このデータは次章で述べる主観評価で用いる。

(4) 学習用データの変換

上記(3)で抽出した以外の約 9000 文を学習用データとして仮名漢字変換処理を行ない、その誤変換結果を分析した。

(5) 付属語辞書の改良

誤変換結果の分析により、口語的な文節末表現を抽出する。学習用に使用した約 9000 文に含まれる口語的表現には重複があるので、図 1 に示したように付属語辞書の改良は段階的に行なった。最終的に約 1200 語を付属語辞書に追加し、付属語接続表も口語体に対応できるよう整備した。

4. 実験

付属語辞書改良の効果を確認するため、改良の前後で主観評価による変換精度測定実験を行なった。評価対象として前節の(3)で抽出した約 1000 文を使用するが、これには非常に長い文も含まれる。携帯機器では入出力デバイスの制限により、長い読みを一度に入力することはできない。そこで、読みが 30 文字以内となるように句読点などで自動的に文を分割した。また、携帯電話では操作仕様の関係で読み入力が困難な記号を含む文を除去した。

主観評価は文単位で行ない、○か×かの 2 値判定とした。主観評価の基準を以下に示す。

(1) 表記ゆれは原則許容(○と判定)

カタカナと平仮名のゆれ、漢字表記と平仮名表記のゆれなどについては、自然な日本語であれば許容して○と判定する。

(2) 同音語の違いについてはある程度許容

入力文が短い場合、正解データ(元文書)とは異なる変換結果であっても、それが自然な日本語となる場合がある。たとえば「かえるんじゃない」という読みに対して「買えるんじゃない」が正解だとしても、「帰るんじゃない」という変換結果は誤りとはいえない。このような違いについてはある程度許容し、正解と異なる同音語であっても○と判定する。ただし、一般的とは思われない単語が第一候補となる場合については×と判定する。

表 1 に実験結果を示す。精度の絶対値が低いのは、長い文についてはその一部が誤っていても×とする厳しい評価としているためである。

表 1: 主観評価による変換精度

	精度	○	×
改良後	50.8 %	514 文	498 文
改良前	42.1 %	426 文	586 文

5. 評価

変換精度が約 8% 向上したことにより、掲示板データの分析に基づく付属語辞書強化による改良の効果を確認できた。改良後の変換結果について分析したところ、基本語辞書に登録されていない未知語が誤変換の原因の半分以上を占めていることがわかった。未知語については、2 章で述べたような自立語の変形と、芸能人名などの新語・流行語とに大きく二分できる。今後はこれらの未知語を効率よく収集して辞書登録する仕組みが必要である。

6. まとめ

大量の掲示板データの分析に基いて付属語辞書を改良し、仮名漢字変換 S/W の口語入力機能を強化した。主観評価による精度比較実験を行ない、改良の効果を確認した。

参考文献

- [1] 牧野「べた書き文の分かち書きと仮名漢字変換」, 情報処理学会論文誌 Vol. 20 No. 4 pp. 337-345, 1979.
- [2] 延澤他「文字間統計情報に基づく口語文字列の自動抽出」, 自然言語処理 Vol. 8 No. 3 pp. 39-57, 2001.