E_016

シソーラスによる意味処理を用いた派生語の子音入力方式
The Effect of Thesaurus-Based Semantic Preference on Consonant-Kanji Conversion of Japanese Derivative Word

市丸 夏樹*
Natsuki Ichimaru

1. はじめに

携帯端末等の少数キーによる言語入力方式の開発は世界的な課題となっている．特に欧米では，携帯電話の１つにキーに３つ程度のアルファベットを割り当てて曖昧な形で入力し，単語辞書によって曖昧性を絞り込む入力方式(T9 input method)が広く普及している．日本語の場合はローマ字表記の子音部分のみを入力する方式が一般的であり，同一キーの連打を除去するものとして期待される．日本語の子音入力に関する関連研究としては，米 Tegic Communications 社の T9(日本語版)や，田中らによる TouchMeKey[4]等が挙げられる．子音による日本語入力は，携帯端末の使い勝手の向上や，親指の使いすぎによる反復運動過多損傷(RSI)の予防などに貢献するものと期待される．

しかし日本語の子音入力システムは，これまでに幾つかのメーカーが携帯電話に搭載してはいるものの，あまり一般的には広まっていないようである．これは，他の言語に比べて日本語の同音異義の曖昧性が非常に高く，打鍵数を削減すると曖昧性が増大する場合があるためであると考えられる．特に派生語では曖昧性が顕著である．例えば「kkak」と打鍵した場合に名詞と接尾語の全ての組み合わせを考慮すると，「機械化」「下降器」など 5,500 通りを越える変換候補が得られる．それらの変換候補から妥当なものを選び出すためには，強力な優先付けを実現する機構が必要となる．そこで本研究では従来のインターフェース上のアプローチとは異なり，特に日本語の単語の意味の側面に着目して，子音のみで派生語を入力した場合に生じる曖昧性に対し計算機側で自動的に優先順位付けを行う手法を開発した．

本稿では，我々が提案した派生語モデルを子音-漢字変換に適用することを試みた．その結果，派生語語基と接尾語を同時に変換した場合の正解率は約 86%であり，意味分類を用いた場合と比較して約 8%の向上がみられた．また，派生語を語基と接尾語に分割し互いに制約を加えながら変換した場合には，最尤解では語基と接尾語でそれぞれ 89%，94%，10 位解では両者とも約 99%の正解率が得られた．子音入力では一般の単語に対してもある程度の数の変換候補が出現することは避けられない．その中で上記のような高い正解率が得られているということは，派生語の子音入力における提案手法の有効性を示していると言える．

2. 従来の手法

T9 や TouchMeKey などの従来の子音入力システムでは予測変換方式が用いられている．そこでは一度使用された語と語の間の連接関係とそれに関する統計情報が主に利用されているものと思われる．しかし，派生語の場合にはこの予測変換が有効に機能しない．派生語は膨大であるため，その全てを予め収集しておくことは困難であり，学習済みでない語の出現の予測が必要となるからである．

*鳥取環境大学環境情報学部情報システム学科
ichimaru@kankyo-u.ac.jp

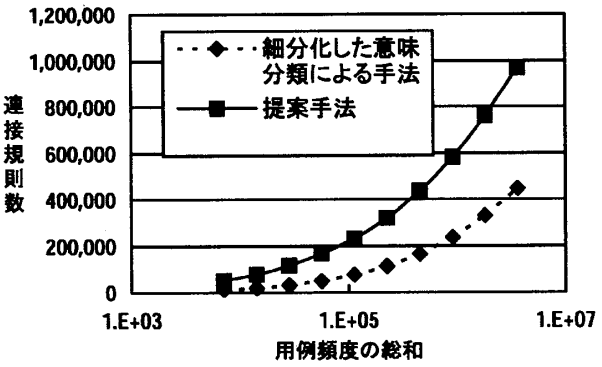


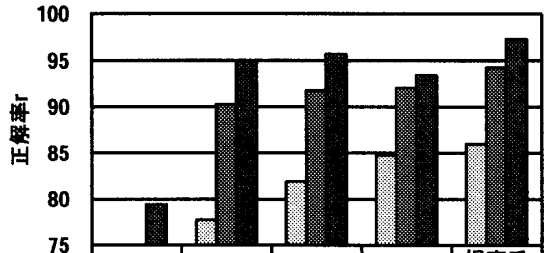
図 1: 用例の汎化により獲得された連接規則数

一方，複合語翻訳[6]の分野においては，従来数千分類程度までの意味分類を用いた選択制限が利用されてきた．しかし従来の意味分類による手法を派生語処理に適用した場合には，次のような問題点が生じることがこれまでの我々の研究の過程において明らかになった[1][2]．第一に，粗い意味分類を用いた場合には最尤解の正解率が上がらない．第二に，意味分類を過度に細分化した場合には解候補数が減少し，正解が出力されなくなることが多い．第三に，正解率を最大にする最適な意味カテゴリ数は学習サンプル数の増加に伴って変動する．さらに，1 位解と 10 位解では最適な意味カテゴリ数が異なるため，両者を両立することができない．これらのことから意味分類による手法は，使用中に学習を伴う日本語入力方式への応用にはあまり適さない．

3. 派生語のモデル

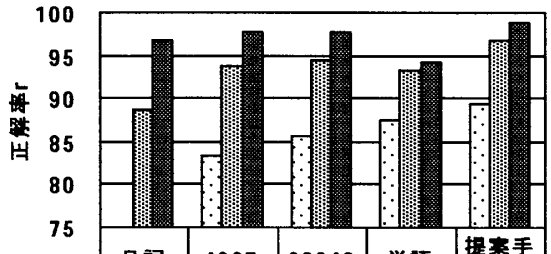
名詞と接尾語からなる派生語には同音異義の曖昧性がとりわけ多い．その上，曖昧性を解消するための表層的な手がかりが乏しい．そのため，日本語入力時の仮名漢字変換の失敗の原因になりやすい．そこで我々は，意味処理と統計処理を統合した確率派生語文法を提案した．

提案手法[1][2]では，まずシソーラス中の全ての中間概念ノードと全ての接尾語間の連接を表す生成規則の存在を仮定する．次に，テキストコーパスから収集した派生語用例を構文解析し，全ての構文木を求める．そして，各々の構文木に対して語基部分の概念ノードの子孫である単語とそのうち用例に含まれるものの割合を求め，各用例の頻度を(子孫用例数/子孫単語数)の比によって重み付けながら各構文木に分配する．最終的にこの構文木集合に対して最尤



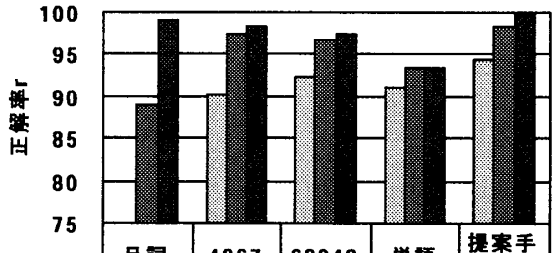
	品詞	4067	68048	単語	提案手法
1位	47.69	77.77	81.88	84.70	85.97
3位	67.30	90.24	91.78	92.09	94.30
10位	79.40	94.99	95.68	93.40	97.31

図2: 派生語の子音-漢字変換の正解率



	品詞	4067	68048	単語	提案手法
1位	69.11	83.37	85.74	87.42	89.33
3位	88.61	93.80	94.33	93.29	96.65
10位	96.73	97.67	97.64	94.28	98.86

図3: 分割変換時の語基変換の正解率



	品詞	4067	68048	単語	提案手法
1位	66.11	90.30	92.24	91.16	94.48
3位	89.09	97.48	96.80	93.47	98.32
10位	98.98	98.33	97.38	93.50	99.74

図4: 分割変換時の接尾語変換の正解率

推定法を適用して学習し，PCFG を構築する．シソーラス中の上位下位関係も生成規則化し，導出パスが長いほど構文木の生起確率を減衰させる．解析時は通常の PCFG による構文解析と同様であり，構文木の生起確率によって変換候補に優先付けを行う．構文木に使用する生成規則の抽象度は構文木によって様々であり，変換候補毎に動的に決定される．以上により，曖昧性を絞り込まず，語基と接尾語の全ての組み合わせを受理しながら，望ましい派生語候補を優先解として出力できる．

この手法では，従来の意味分類による手法の主要な問題点が解決され，十分な量の学習サンプルを与えれば 1 位解と 10 位解の正解率をほぼ両立できる．なお，提案手法では汎化を制限しないことから連接規則数の増加が懸念されたが，そもそも意味分類を細分化した段階で連接規則が増加しているため，汎化による増加分は約 2 倍程度に過ぎない(図1)．

4. 評価結果

文献[2]に示した仮名漢字変換の場合と同様に，派生語のローマ字表記から求めた子音文字列を入力とし，コーパス中に出現した表記を正解とする子音-漢字変換実験を行った．派生語学習サンプルとしては EDR[5]日本語単語辞書，日本語コーパス，新聞記事 6 年分[7]から抽出したのべ約 360 万語の派生語用例を使用した．シソーラスとしては EDR 概念体系辞書の全ノードを使用した．試験サンプルには人手チェック済みの形態素データ[3]より抽出した派生語正例のべ 14,823 語を使用した．

この場合の正解率を図 2 に示す．ここでは比較する従来手法として，名詞と接尾語の連接を品詞間の連接として捉えた場合(品詞)，m 分類の意味分類を用いて選択制限を行った場合(m)，用例を単語辞書に直接登録した場合(単語)の正解率を併記している．

また，語基と接尾語の分割変換の場合の正解率を図 3, 4 に示す．派生語単体で変換すると最大変換候補数は(|語基候補数|×|接尾語候補数|)となるが，語基と接尾語を個別に変換・確定すると高々(|語基候補数|+|接尾語候補数|)で済む．語基変換時には仮名表記の接尾語との連接可能性が使用でき，接尾語の変換時には確定済みの漢字表記の語基との連接可能性が利用できる．これにより，子音入力においても高い正解率を得ることが可能となり，提案手法の有効性を示すことができた．

参考文献

[1] 市丸夏樹. 用例とシソーラスに基づく派生語処理に関する研究. 九州大学博士論文, 2006

[2] 市丸夏樹, 中村貞吾, 日高達. 汎化用例とシソーラスを用いた派生語の仮名漢字変換の特性. 自然言語処理, Vol. 12, No. 2, pp. 189-207, 2005

[3] 新情報処理開発機構. RWC テキストデータベース. メディアドライブ, 1998

[4] 田中久美子, 犬塚祐介, 武市正人. 携帯電話における日本語入力-子音だけで日本語が入力できるか. 情報処理学会論文誌, Vol. 43, No. 10, pp. 3087-3096, 2002

[5] 日本電子化辞書研究所. EDR 電子化辞書. 1999

[6] 藤井 敦, 石川 徹也. 技術文書を対象とした言語横断情報検索のための複合語翻訳, 情報処理学会論文誌, Vol.41, No.4, pp. 1038-1045, 2000

[7] 毎日新聞社. CD-毎日新聞 '91--'95, '98. 日外アソシエーツ, 1991-1995, 1998