1R－2

国語辞書から構築した意味関連グラフに基づく仮名漢字変換の改良

松浦 司
東京大学理学系研究科情報科学専攻

1 はじめに

日本語には多数の同音異義語が存在するため、仮名漢字変換において文節分けの後にも複数の変換候補が存在する場合が多い。それらの候補を絞り込むために、従来、格フレーム情報[1]や共起情報[2][3][4]、またニューラルネットワーク[5][6]を利用する手法が研究されてきた。しかし、それらの情報は高品質に生成するには労力を要し、またどの様な文書から抽出すれば良いのか明らかではなかった。

本研究においては、一冊の国語辞書という非常に入手しやすい文書から、単語間の関連を意味関連グラフの形で自動的に取り出した。そして、複数の同音異義語の中から文脈に適合するものを意味関連グラフを用いて選択することにより、仮名漢字変換精度の向上を図った。

2 意味関連グラフ

意味関連グラフとは、単語間の意味的関連を表すグラフである。グラフ中の一ノードは一単語に対応しており、関連のある単語同士は有向の枝で結ばれる。

本研究に関わる実験では、国語辞書[7]の見出し語と、その語義文に現れる単語との間には意味的関連があるものと仮定して、見出し語のノードから語義文中の語のノードへ向かう有向の枝を作成した。国語辞書は、予め「茶筅」[1]による形態素解析を施した上で、使用した。

単独では意味が薄く、文全体の意味を整えるために使われるような単語を安易に語義文中の語と関連付けると、実際には無関係な多数の語を間接的に関連づけることになり、変換精度の低下を招くおそれがある。そのため、脚注[2]に示す品詞および単語は見出し語であっても、語義文中の単語と関連付けないこととした。また、語義分の用例中の単語も見出し語と関連が深いとは限らないので、関連付けの対象から除外した。

3 意味関連グラフを用いた仮名漢字変換

本研究では、変換候補中で最も入力中の文章の文脈に適合しているものを、最も尤度の高い候補、即ち第一候補と見なす。文脈は、先だって最近に変換された語および高頻度に変換されている語によって表されるものとする。意味関連グラフ中で、文脈を表す単語群と最も近い距離にある候補が第一候補として選択される。先行する複数の文から取り出された文脈情報が利用されるので、変換中の文が文脈を特定する情報を含んでいなくても、適切な候補を選択し易くなっている。

4 評価

第一候補が適切な候補であった場合に正しい変換が行われたものとし、以下の様に変換効率を定義する。但し、変換結果が正解の別表記方であった場合にも、正しい変換であるとみなす。

変換効率 ＝ 正しく変換した単語数 ／ 変換した全単語数

評価用文章として、毎日新聞社がWWW上で提供している新聞記事の1998年6月分を使用した。全645記事をまず茶筅で形態素解析し、漢字表記もしくは漢字仮名交じり表記されている単語だけを選び出した。このようにして得られた117091語を茶筅によって平仮名読みに変換し、再び本研究で作成したシステムで

[1] 茶筅 ver.1.0、1997年、奈良先端科学技術大学院大学 松本研究室
[2] 「茶筅」で特殊記号、副詞的名詞、形式名詞、数詞に分類される語及びサ行変格動詞「する」

Improvement of the Kana-kanji Translation Based on the Semantic Relation Graph Constructed from a Japanese Dictionary
Tsukasa Matsuura
Kanada Lab., Department of Information Science, University of Tokyo
2-11-16 Yayoi, Bunkyo, Tokyo 1113-8685, Japan

仮名漢字変換を行った。元の新聞記事中で平仮名のみを用いて表記されていた単語は、正しい変換結果を決定できないので変換の対象から除外した。

評価実験では、文脈を表す単語の保持数と、関連語間の最大距離とを変更して変換効率を測定した。意味関連グラフ中で他のノードを介して間接的に結びついている単語同士も関連語とみなすが、最大距離より離れたノード同士は無関係であると考える。

グラフ1が示すように、文脈を表す単語保持数と関連語間最大距離とを共に大きくすると、より変換効率は改善されている。この結果から、変換効率は未だ市販システムに及ばないものの、当手法の有効性が確かめられた。

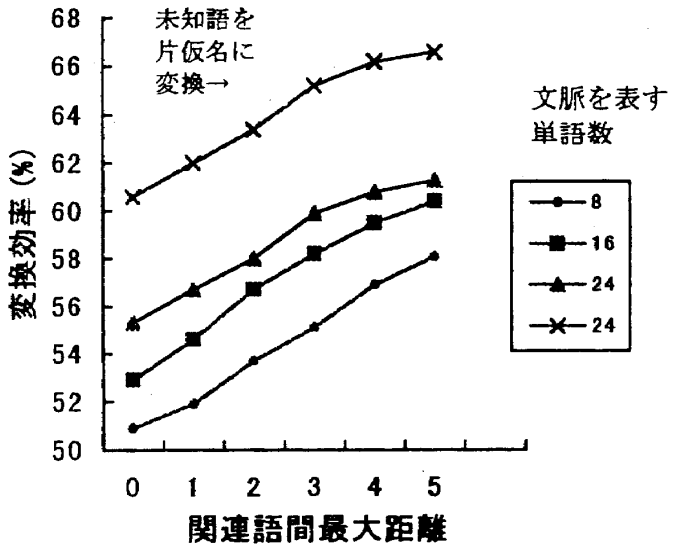
変換された語の、19.7%は意味関連グラフに未登録の語であった。より見出し語数の多い国語辞書を使用する、或いは複数の国語辞書を組み合わせて使用するなどの方法で、未登録語を減らすことによって変換効率の向上が期待できる。参考のため、未登録語には外来語が多いという経験則に基づいて未登録語を片仮名に変換した場合[3]の変換効率をグラフ1に示す。

5 まとめ

国語辞書から構築した意味関連図を用いて、仮名漢字変換の性能を向上させられることが確かめられた。この結果から意味関連グラフは、人間が言葉を記述する際に用いている語間の関係を部分的に表し得るものと考えられる。更に、本研究に関わる実験で作成した意味関連グラフのサイズは3.3MB[4]であり、また一語の変換に要する平均時間は0.095秒[5]と、実用に耐えられる範囲にある。

今回は見出し語数約5万3千語の辞書を用いたが、将来はより大規模な辞書や複数の辞書を組み合わせて使うことによって更に性能向上を図る予定である。また、現段階では語の使用頻度を全く考慮していないが、一般に、或いはユーザによって、高頻度に使用される語を優先して変換候補として選択することによって変換効率の向上が期待できる。

グラフ1
変換効率の改善



参考文献

[1] 牧野寛, 他: べた書き文の仮名漢字変換システムとその同音語処理, 情処学論, Vol22, No.1, pp.59-67 (1981)

[2] 高橋雅仁, 他: 単文内での共起情報を用いた同音語処理, 情処学論, Vol.37, No.6, pp.998-1006(1996)

[3] 上原龍也, 他: かな漢字変換における共起情報の適用方式の拡張, 情処学会44回全国大会講演論文集(3), pp.187-188(1992)

[4] 山本喜大,, 他: 共起グループを用いたかな漢字変換: 情処学会44回全国大会講演論文集(3), pp.189-190(1992)

[5] 鈴岡節, 他: 神経回路網の連想機能を用いたかな漢字変換方式, 情処学会40回全国大会, 1C-3(1990)

[6] 小林勉, 他: ニューロ仮名漢字変換の実現, 東芝レビュー, Vol.47, No.11, pp.868-870(1992)

[7] 山岸徳平 編: 清水新国語事典, 清水書院 (テグレット社 光の辞典 ver.3 より)

[3] 片仮名表記の未知語数は、文脈を表す単語数および関連語間最大距離に依存しない。

[4] 単語の表記や読みは登録されているが、活用や品詞情報は含まない。

[5] 文脈を表す単語数24、関連語間最大距離5の条件で、Pentium 200MHz及びメモリ64MBを搭載したPCにて測定。