5Q−3

大語彙かな漢字変換
—連語の効果について—

山田洋志　大山裕
(日本電気株式会社 C&Cシステム研究所)

1 はじめに

大語彙かな漢字変換システムは、大語彙の単語辞書と単語間の制約関係を記述した制約データを用い、広い分野で高精度の変換を実現することを目的としている[1]。

大語彙の単語辞書を使用することで、未登録語の出現率、単語の区切り誤り率が減少するという利点が得られることについては既に報告した[2]。

本稿では、単語間の関係を制限するための制約データとして連語(単語の共起)を導入した場合の変換率の向上について報告する。また、テキストから自動作成した連語辞書を用いた実験についても報告する。

2 連語

連語は、意味的に関連があり、同一の文や文章中に共起しやすい単語の組み合わせである。かな漢字変換では、同音語がある場合に、連語を優先して選択することで、変換誤りを減らすことが出来る。筆者らは、現在のところ隣り合った2単語に限定して使用している。また、連語適用の精度を増すために2単語間の位置関係に限定を設けている。単語の位置関係で現在使用しているものを、表1にあげる。

表 1: 連語の種類(単語の位置関係)

タイプ	例
AB	「情報—処理」
AをB	「情報-を-処理」
AのB	「情報-の-処理」
AがB	「事故-が-起こる」
AにB	「口-に-くわえる」
AでB	「あご-で-使う」
AとB	「需要-と-供給」
AがBされる	「効果-が-期待される」
AするB	「始まる-時間」

Kana-kanji translation method using large scale dictionary
— The effect of using collocation —

Hiroshi YAMADA and Yutaka OHYAMA
C&C Systems Research Laboratories, NEC Corporation

3 連語の効果に関する予備実験

大語彙単語辞書を用いたかな漢字変換システムに連語を導入することで、どのくらいの精度向上が望めるかについて実験を行った。

予備実験に用いた連語辞書は、連語を用いない場合の変換結果から誤変換個所を抜きだし、誤変換箇所を参照しながら手作業で連語を辞書に登録して作成した。登録した連語は1,468件である。従って、この実験で用いた連語辞書は、実験に使ったテキストを変換するのに都合がいい連語はほとんど収録している、という理想的な状態になっている。

予備実験用に作成した連語辞書を用いて変換率の測定を行った。単語辞書は31万語辞書を使用し、変換対象として高校生用教科書の抜粋を用いた。

全く連語を使用しない場合の変換率と、使用した場合の変換率を表2に示す。

表 2: 予備実験での変換率

文書	文節数	連語無	連語有	差
地理	2,119	87.8%	93.4%	5.6%
国語	1,135	85.3%	90.7%	5.4%
日本史	2,532	84.4%	90.3%	5.9%
理科	1,882	87.9%	92.9%	5.0%
社会	3,374	85.8%	90.8%	5.0%
指導要領	843	84.8%	89.7%	4.9%
合計	11,885	86.1%	91.4%	5.3%

表2から分かる通り、連語を用いることで平均5.3%の変換率向上が得られた。これは、変換に必要なデータが辞書にすべて蓄えられている場合の変換率向上分と考えられる。

実験に使ったテキストが少ないが、大語彙単語辞書と連語の組み合わせという方式で、良質の連語データを充分な量用意すると、90%以上の変換率実現の可能性があることが分かる。

4 連語辞書の試作と変換率の測定

中規模の連語辞書を実際に作成して変換率の測定を行った。

連語辞書は、「語と語の関係データ」[3]をもとにして作成した。「語と語の関係データ」は、朝日新聞の記事

およびJICSTの科学文献抄録から単語の組み合わせを抽出したものである。今回は、全データのうち出現頻度の大きいものをもとに約5万件の連語辞書を試作した。

試作した5万件の連語辞書を用いて変換率の測定を行った。単語辞書は、77万語のものを用い、変換対象として新聞記事、小説・評論、事務文書の3種類を用いた。

結果を表3に示す。

表 3: 試作した連語辞書による変換率

文書	文節数	連語無	連語有	差
新聞	12,025	80.2%	82.9%	2.7%
小説	13,793	81.2%	82.1%	0.9%
事務	7,775	79.5%	83.2%	3.7%
合計	33,593	80.4%	82.7%	2.3%

表3から分かる通り、試作した連語辞書による変換率の向上は2.3%であり、3節の5.3%とはかなり開きがある。そこで誤変換個所を調べたところ、同音語誤りが6.7%あった。より多くの連語が用意できれば、同音語誤りを減らすことができ、変換率を上げることができる。また、誤変換の中には、単語区切りの誤りが9.9%あった。区切り誤りを減らすためには、単語辞書、連語辞書の充実のほか、候補の評価関数や変換規則の見直しなどの作業が必要である。

5 連語辞書の自動作成と変換率の測定

4章の実験結果からも分かる通り、連語を使って幅広い文章で変換率を上げるには多くの連語が必要である。そこで、より大量の連語を効率良く集めるために、電子化されたテキストから自動的に連語を抽出する方法について検討を行い、基礎的な実験を行った。

5.1 連語辞書の自動作成

連語辞書の作成は、(1) テキストの形態素解析、(2) 連語の候補を抽出、(3) 連語マスタ辞書のフォーマットに変換、の手順で行った。連語データは隣り合った2自立語の組から品詞や接続関係で抽出した。

連語辞書の元データには新聞記事約200万文字を用いた。そこから選別された連語は186,248組で、読み・表記・品詞の等しいものをまとめて、109,202件の連語辞書を作成した。

連語辞書自動作成の方式については、別途に報告する予定である。

5.2 自動作成した連語辞書による変換率の測定

自動作成した連語辞書を用いて変換率を測定した。

単語辞書は77万語のものを用い、文書は、教科書、新聞記事(辞書作成用とは別の記事)、小説を用いた。結果を表4に示す。

表 4: 自動作成した連語辞書による変換率

文書	文節数	連語無	連語有	差
教科書	12,752	84.2%	84.5%	0.3%
新聞	12,025	80.2%	81.3%	0.9%
小説	13,787	81.2%	81.5%	0.3%
合計	38,564	81.9%	82.4%	0.5%

自動作成した連語辞書を使用すると、連語を使用しない場合に比べて変換率が0.5%向上した。連語によって良くなった個所だけを数えると1%程度の変換率向上になる。従って、変換率向上には、自動抽出する連語の質を高めることが重要である。また、新聞記事での効果が高いのは、連語辞書の作成にも新聞記事を使ったためと考えられる。分野による連語の片寄りについては今後の検討課題である。

現在、誤り個所の詳細な評価を行っている。

6 おわりに

大語彙単語辞書に加え、連語を用いることによる変換精度向上について実験による確認を行った。また、連語辞書の自動作成についても基礎的な実験を行った。

予備実験によって、大語彙の単語辞書と連語辞書を用いることで、90%以上の変換率を実現する可能性があることが分かった。

一方、実際に試作した5万件のデータを含む連語辞書による実験では、連語を用いない場合に対する変換率の向上は2.3%(変換率82.7%)にとどまった。変換結果の内には、連語の拡充によって解決できる誤りが多く見られた。

また、大量の連語データを用意するための方法として自動的に連語辞書を作成することを試み、10万件の連語辞書を自動作成した。自動作成した連語辞書を用いて変換を行ったところ、変換率の向上は0.5%にとどまった。

今後の課題としては、以下のものがある。

- 人手による作成および自動作成の両面からの連語辞書の充実
- 辞書以外の変換規則の改良
- 分野の影響の調査

さらに、連語方式の拡張、他の制約データの導入、文脈の利用の検討などを通して、大語彙方式に適したアーキテクチャの開発を行い、変換率が95%以上の高精度のかな漢字変換システムを実現する。

参考文献

[1] 山田他,情処40全大5P-1,1990
[2] 山田他,情処41全大3J-1,1990
[3] 田中他,情処40全大5F-1,1990