深層共起技術とニューラル技術を適用したかな漢字変換

4C−6

小林 勉　中里 茂美　斎藤 裕美　大黒 和夫
(株)東芝 情報処理・機器技術研究所

1．はじめに

かな漢字変換における研究課題の1つとして、同音異義語の問題がある。かな文字列から漢字かな混じり文字列への変換は、一般に1対多対応であるため、作成中の文章に即した候補を同音異義語の中からいかにして選択するかが、かな漢字変換の重要なポイントとなってくる。

この同音異義語の問題を解決しようと、今まで様々な手法が試みられてきた。たとえば、語の使用頻度を用いて優先順位を決定する方法、最近選択した語を優先する方法、あるいは特定の文法関係をともなって生起する語のグループを登録しておき、変換結果にそのようなグループがみつかれば優先するという、いわゆる『共起情報』を利用した方法などがよく知られている。またニューラル技術を応用して、話題に即した同音異義語を優先するという試みも報告されている[1]。

これらの手法は、どれかひとつが同音異義語の問題を完全に解決するものではなく、実際には適宜組み合わせて用いられている。我々は今回、深層共起技術およびニューラル技術を組み合わせて、同音異義語を選択する方式を試作した。

2．深層共起技術

我々は、従来の共起情報を用いた同音異義語の優先度付けを発展させて、深層格を用いて共起情報を記述する試みを既に報告した[2]。この方式では単語を意味的な観点から分類することによって、深層格を用いた共起情報を効率よく記述することができた。本論文では、深層格を用いた共起情報を『深層共起情報』と呼ぶ。また簡略のために、深層共起情報を用いた同音異義語の優先度付けに関する技術をまとめて『深層共起技術』と呼ぶことにする。

深層共起情報を用いた同音異義語の優先度付けの概略を図1に示す。この例では『しゃいんがじこにあった』というかな文字列をかな漢字変換して『社員が事故に遭った』という漢字かな混じり文字列を第1候補として出力している。

入力文字列は、辞書や文法知識を用いて文節に区切られる。その結果『しゃいん(が)・じこ(に)・あ(った)』

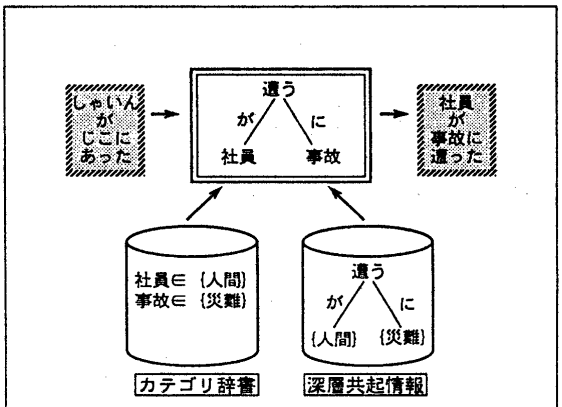

図1 深層共起情報を用いた同音異義語の優先度付け

のようになる。ここで、『しゃいん』,『じこ』,『あう』についてそれぞれ

社員 社印
自己 事故
会う 合う 遭う 逢う

という同音異義語が存在するため、候補の組合せは16通りある。これら16通りの候補をすべて深層格のレベルで格解析し、カテゴリ辞書を参照しつつ、深層共起情報とマッチングをとる。その結果、最もマッチング度の高いものが第1候補となる。また、マッチング度がある値よりも低い場合には、候補から除去してしまう。

このように深層共起技術では候補を深層格レベルまで解析して、それぞれの候補の妥当性をチェックするため、従来の方式では扱いが難しかった共起情報も、比較的安全に取扱うことができる。

3．ニューラル技術

深層共起技術が1文中の文法的解析に主眼を置いているのに対して、作成中の文章全体というもっと広い範囲の情報を元にして、話題に即した同音異義語を優先しようというアプローチがある。この目的のためにニューラル技術を用いる。この方式については、文献[1]で詳しく述べられているが、ここで簡単に概観しておく。

図2に示すように、単語間にネットワークを構成し、

The Kana-Kanji conversion using deep-case-analyzing technology and neural technology
Tsutomu KOBAYASHI　Shigemi NAKASATO　Hiroyoshi SAITOU　Kazuo OOGURO
TOSHIBA Corp. Information Systems & Engineering Laboratory

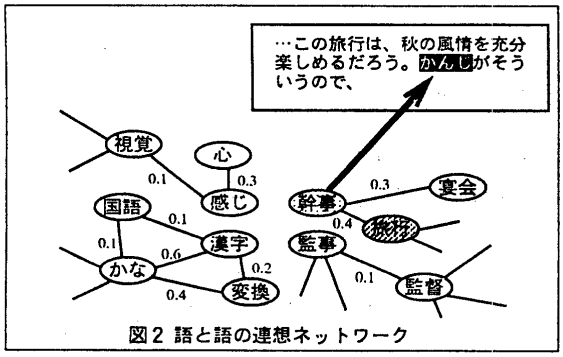

図2 語と語の連想ネットワーク

単語と単語の『使用上の近さ』に応じてリンクの重みを設定する。ある単語と単語が、統計的にみて文章中の近い距離に出現するならば、それらの単語間のリンクの重みは大きくなる。

それぞれのノード(単語)は活性値O_jをもつ。このO_jは次のように計算する。

$$O_j \leftarrow f(n_j)$$

$$n_j \leftarrow (1-\delta)n_j + \delta\,(\Sigma\, w_{ji}O_i + I_j)$$

ただし

f :単調増加関数

δ :$0 < \delta < 1$の実数

w_{ji}:単語iと単語jの間のリンクの重み

$(w_{ij} = w_{ji},\ w_{ii} = 0)$

I_j :ノードjへの外部入力

関数fの形式およびδは経験的に最適値が決められる。ノードへの外部入力I_jは、ノードに対応する単語が同音異義語の中から選択されると正の値になり、徐々に減衰するようになっている。また図2においてリンクがつながっていないものに対しては、仮想的に負のリンク値でつなげている。これは他のノードへの抑制項として働き、話題の移り変わりに応じてスムーズに活性値の分布が移行するのを助ける。

このような動作をさせることによって、入力中の文章の話題に関係する単語の活性値が高くなる。これを利用して、同音異義語の選択時に活性値の高い単語を優先することによって、話題に適した候補を出すことができる。

4. 深層共起技術とニューラル技術の組合せ

文献[1]でも述べられているように、ニューラル技術は従来のかな漢字変換の欠点を補うものであって、従来のかな漢字変換に取って替わるものではない。

従来のかな漢字変換とニューラル技術を組合せることによって、従来のかな漢字変換では不可能だったことができるようになる。その結果、全体的に変換率は向上するが、ニューラル技術による悪影響もあった。そのような悪影響を調査した結果、従来よりも強力な、かな漢字変換候補の文法的妥当性のチェックを入れることで、悪影響のほとんどをなくせることがわかった。

たとえば、社員旅行に関する文章を入力していたとすると、『幹事』という単語の活性値が上がってくる。このとき『かんじがかんじをかく』という入力文字列をかな漢字変換することを考える。

まず始めに辞書や文法知識を利用して、入力文字列を文節に分解する。その結果

『かんじ(が)・かんじ(を)・かく』

となる。ここで『かんじ』,『かく』にはそれぞれ

感じ 漢字 幹事 監事

書く 欠く 掻く 描く 各 核 .....

などの同音異義語が存在する。

もしこの状態で、活性値の高い単語を優先して出力すると

『幹事が幹事を書く』

という第1候補になってしまう。

一方、深層共起技術を使うと候補は

『幹事が漢字を書く』

『幹事が感じを欠く』

『監事が漢字を書く』

『監事が感じを欠く』

などに絞られてくる。この結果に対してニューラル技術を適用することで

『幹事が漢字を書く』

『幹事が感じを欠く』

を優先して出力できるようになる。

上に述べた深層共起技術による候補の絞り込みは、従来の共起情報でも、ある程度までは実現可能であるが、無意味な候補をすべて落とすためには、深層共起技術のような、膨大な量の共起情報を効率よく扱う仕組みと、従来よりも強力な文法的妥当性のチェックが必要となってくる。

5. まとめ

深層共起技術を用いることによって、文法的・意味的に妥当な候補を優先するとともに、妥当でない候補を落とすことができる。

ニューラル技術を生かすためには、従来よりも強力な文法的妥当性のチェックが必要である。そこで、深層共起技術を用いた結果得られた、文法的・意味的に妥当な候補に対して、ニューラル技術を適用すると効果が期待できる。

現在、同音異義語の選択をする手法をいかに組合せるかによって、効果にどの程度の違いが出てくるのかを測定する準備を進めている。

参考文献

[1] 鈴岡, 木村, 伊藤, 天野,「神経回路網を用いたかな漢字変換方式」、情報処理学会第40回全国大会(1990)

[2] 後藤, 横田, 中里, 大黒、「深層格を用いた仮名漢字変換」、情報処理学会第41回全国大会(1990)