

深層格を用いたかな漢字変換における共起関係の自動学習

5 L - 5

松下慎一、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同音異義語の決定のための一手法として単語同士の共起関係を利用したものがあつた。その共起関係を意味的な観点から分類したカテゴリにまで発展させた情報を使用するかな漢字変換 [1] (以後「深層共起かな漢字変換」と略す) の開発を我々は試みてきた。共起関係の記述をカテゴリに対して行うことにより、効率良く共起情報を記述することが可能となり、辞書中のすべての単語に対して共起情報の記述が可能となった。しかし逆に共起の例外的な使われ方に対しては障害になるといった側面も合せ持つ。そこでそのような場合に一度確定された単語の共起関係を学習することのできる方式が必要となつてくる。

2. 深層共起かな漢字変換の概要

我々は、従来の共起情報を利用した同音異義語の優先度決定法を発展させ、深層格を用いて共起情報を記述したかな漢字変換についての研究を行つてきた。この方式は単語を意味的な観点から分類し、その分類されたカテゴリに対して共起情報を記述することにより大量の共起情報を効率良く記述することに成功した。

深層格を用いた共起情報による同音異義語の優先度決定の概要を図1に示す。

この例では「しゃいんがじこにあつた」という文字列をかな漢字変換して「社員が事故に遭つた」という漢字かな混り文字列を出力している。

入力文字列は形態素解析により文節ごとに区切られる。その結果「しゃいんが/じこに/あつ

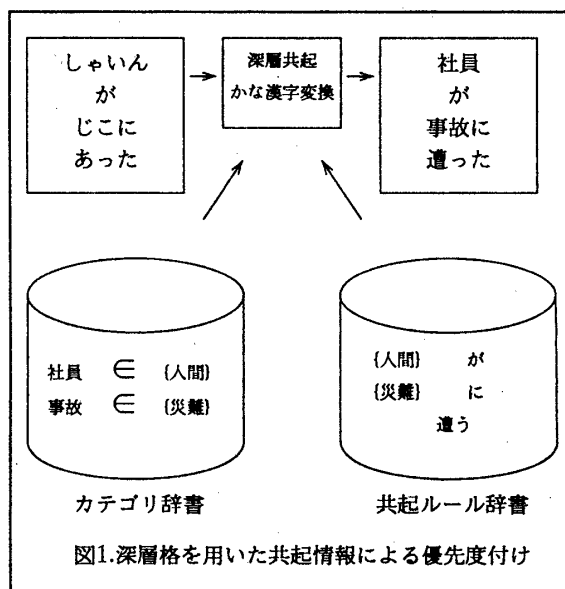


図1. 深層格を用いた共起情報による優先度付け

た」というかたちとなる。ここで、「しゃいん」、「じこ」、「あつた」のそれぞれについて次のような同音語が発生したものとする。

- 社員、社印
- 自己、事故
- 会う、合う、遭う、逢う、遇う

するとかな漢字変換部は共起ルール辞書とカテゴリ辞書を使用し、その中の情報から最も優先度の高い候補の組み合わせを選び出す。また逆に優先度があらかじめ定めたある水準より低かつた場合は候補としても出力しないという処理を行う。

深層共起かな漢字変換の主な特徴は以下の通りである。

- (1) 効率良く共起情報を辞書に収容できる
- (2) 使用者の新語対応がやりやすい
- (3) ツリー構造の構造のマッチングを行うの

で、表層格が変わるような場合でも共起ルールは1つで良い

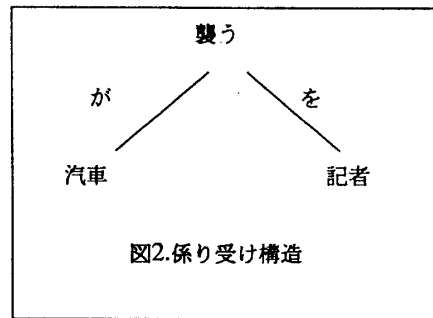
(4) 共起ルールを用いて候補を落とすことができる

3. 共起関係の自動学習

深層共起かな漢字変換は、かな漢字変換で使用する辞書中の単語すべてをカテゴリに対する共起ルールとして辞書に記述しているという前提を持っている。しかし、これは共起ルール辞書に記述されていない例外的な使われ方に対してはうまく働かない。このような例外的な使われ方に対しても深層共起かな漢字変換がうまく働くようにするための一方策として共起関係の自動学習（以後「深層共起学習」と略す）がある。これは、出現した共起関係を学習し、次に同じ関係が出現した際にその学習結果を共起関係の解析に利用するという方法である。従来の技術として単語同士の共起関係の学習といったものがあるが、深層共起学習はその機能を発展させ深層共起かな漢に対応すべくカテゴリレベルでの学習を目的としている。

深層共起学習の学習動作は二段構成になっている。まず入力された文字列をかな漢字変換が変換し確定する。そして、その確定された文の解析を行い係り受けの関係を抽出し、単語レベルで関係を学習する。次に、この学習されたデータの中に同じカテゴリへの偏りが見られるようなデータを見付け出し、それをそのカテゴリへと纏める。

例えば次の入力「車が記者を襲う」が確定されたとする。この文はかな漢字変換によってすでに単語ごとに分けられ更に各種の情報が付加されている。これらの情報を元に係り受けの解析を行い図2に示す構造を作成し解析結果を学習データとして記憶する。同様に「車が襲う」「バスが襲う」「地下鉄が襲う」などの組み合わせが確定されデータが蓄積されるとデータのカテゴリによる分類が行われる。そしてそれらのデータを纏め上



げてカテゴリレベルでの学習データとして記憶する。この例の場合は「{乗り物}が襲う」という形となる（‘{}’はカテゴリであることを表す）。

深層共起学習は以下のような特徴を持つ。

- (1) 単語同士の共起関係を学習する
- (2) カテゴリレベルで共起関係を学習する
- (3) 一文内の単語の関係を学習する
- (4) 深層格が同じであれば学習時と表層格の違う型式の文にも使用可能

4. おわりに

深層共起学習は、辞書に登録されていない共起関係の新たな学習を行うことを目的としている。更に、共起関係の使用頻度データの記憶、深層レベルでの共起関係の学習といった機能を合せ持ち、深層共起かな漢字変換の変換率を向上させることを期待できる。

現在、学習したデータをカテゴリレベルへと上げる統計処理、係り受け解析の効率の向上といった問題の解決のための調査を行っている。

参考文献

- [1] 後藤他「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）
- [2] 小林他「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」、情報処理学会第45回全国大会（1992）