4P−10

かな漢字変換における共起情報の適用方式の拡張

上原 龍也　野上 宏康　齋藤 佳美
相原 義弘　天野 真家

(株) 東芝 総合研究所

1．はじめに

かな漢字変換の誤りの原因としては、同音語の誤り、文節切り誤り、未登録語に起因する誤りなどがある。特に、この内、同音語の誤りは発生率が高く、改善を要する課題となっている。この課題を解決する手法として、語と語の共起関係に着目し、これを利用して変換を行なう方式が提案されている。

しかしながら、この方式では、共起関係のデータベースが必要であるが、すべての共起関係を調べるためには、単語の膨大な数の組合せを調べなければならず、実際問題として不可能である。また、共起関係のデータベースが増大すると、同じ読み入力に対して複数の共起関係が適用でき、共起関係の競合が生じる。この場合、一文だけでなく、文章全体の文脈から同音語を選択する必要がある。

この問題を解決する手段としては、共起関係を適用する範囲を広げて、共起情報の適用可能な数を増加させる方法が考えられる。本稿では、共起関係を複数文にわたって適用することにより、文脈に適した同音語を優先するかな漢字変換の方法について述べる。

2．共起関係適用方式の拡張

従来、提案されていた共起関係を用いたかな漢字変換では、係り受け関係にある単語対に対して、あらかじめ記憶された共起関係と一致すれば、その単語を優先していた。例えば、「歌舞伎」と「公演」という共起関係が記憶されていたとする。このとき、「歌舞伎が昨日コウエンされた」の場合、「歌舞伎」と「コウエン」は係り受け関係にあるので、上記の共起関係を適用することにより、「公演」という同音語を優先することができる。これに対して、異なる文に含まれている2単語は係り受け関係になりえないので、従来方式では共起関係を適用していなかった。例えば、「ヨーロッパでは、歌舞伎が人気を呼んでいる。昨日行なわれたコウエンは満員であった。」という入力に対しては、「歌舞伎」と「コウエン」は別の文にあるので、従来の方法では、上記の共起関係を適用しない。

しかしながら、この入力の場合、「歌舞伎」が話題となっているので、「コウエン」は、「公演」である可能性が高い。これは、「歌舞伎」と「公演」という2単語間の関係が、係り受け関係という統語的な関係だけでなく、意味的に強く結び付いている関係であるために、係り受け関係にない2単語に対しても影響を及ぼしているからであると考えられる。このような意味的に強く結び付いている関係を、従来の係り受け関係に基づいた共起関係と区別して、以降、意味共起関係と呼ぶ。共起関係を意味共起関係として用いることにより、共起関係を複数文にわたって適用することが可能となる。つまり、複数文にわたるような係り受け関係にない単語間にも共起関係が成り立てば、その単語を優先する。これにより、文脈に適した変換が可能になる。例えば、上述の例文の場合、正しく「公演」を優先することができる(図1)。

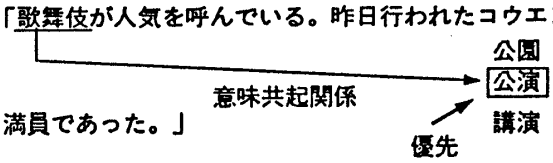

図1 意味共起関係の例

3．意味共起関係

上述したように共起関係を意味共起関係として適用することによって文脈に適した単語を優先することができるが、共起関係であるものがすべて意味共起関係であるという訳ではない。例えば、「子供」と「泣く」は共起関係であるが、「子供が部屋に入ってきた。その時、カナリヤがナイタ。」という文章に適用した場合、もし、「カナリヤ」が「鳴いた」という共起関係がデータベースに登録されていなければ、上記の共起関係により、「ナイタ」が、誤って「泣いた」が優先される(図2)。

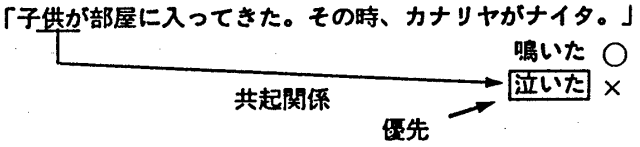

図2 意味共起関係とならないの例

Kana-Kanji Conversion Method using Semantic Cooccurrence Restriction
Tatsuya UEHARA　Hiroyasu NOGAMI　Yoshimi SAITO　Yoshihiro AIHARA　and Shin-ya AMANO
TOSHIBA Corp.

この理由は、「子供」と「泣く」は、統語的な結びつきが強く、意味共起関係ではないためである。

したがって、上述した共起関係の適用方式を実現するためには、共起関係が意味共起関係となる条件を分析しなければならない。

意味共起関係は、２つの単語間に統語的な関係が存在しない場合でも成立する必要があるので、共起関係にある２つの単語のうち、以下の条件をひとつでも満たす関係は意味共起関係ではない。

ａ）　２つの単語の一方が機能語(functional word)である場合

この場合、共起関係にある単語は機能語であり、主に統語的な役割を担っているので、その共起関係は統語的な関係であって、意味的な関係ではない。したがって、このような共起関係を用いても文脈に適した単語を優先することができない。この条件に従えば、以下のような機能語を含む共起関係は意味共起関係とならない。

- 接続詞　（つまり、したがって　など）
- 代名詞　（これ、この　など）
- 形式名詞　（もの、こと　など）
- 補助動詞　（できる、始める　など）
- 副詞句や副詞節を形成する名詞　（場合、際　など）

ｂ）　一方の単語を同じ品詞の単語と置き換えても共起関係になる場合

例えば、「よく」と「食べる」という共起関係に対して、「食べる」を他の動詞に置き換えても共起関係になる。したがって、「よく」が出現することによって、その周辺に特定の単語が出現しやすいという制約を与えることができないので、文脈に影響を与えないと考えられる。この条件に従えば、以下の単語を含むものは、意味共起関係から排除できる。

- 程度の副詞　（よく、かなり　など）
- 時の副詞として使われる名詞　（昨日、今年　など）

ｃ）　２つの単語が係り受け関係にある時のみ、強い関係にある場合

例えば、「子供」と「泣く」は、「子供が泣く」という係り受け関係にある場合にのみ、強い関係があるが、係り受け関係のない場合は、関係が弱いと考えられる。したがって、このような関係は、意味共起関係ではなく、統語的な共起関係である。この条件に適合するのは、以下の品詞の組合せである。

- 「名詞」と「動詞」
- 「動詞」と「名詞」
- 「形容詞」と「名詞」
- 「名詞」と「形容詞」
- 「形容動詞」と「名詞」
- 「名詞」と「形容動詞」

ｄ）　２つの単語の少なくとも一方が、単独では使用されない単語である場合

この条件に適する単語は、ある単語と結び付いて一つの意味を持つので、その単語と離れて出現する場合は共起関係が無効になると考えられる。この条件に従えば、次のような単語を含む共起関係は意味共起関係とならない。

- 接頭語・接尾語（不、性　など）

４．評価実験

本稿で述べた共起情報の適用方式をかな漢字変換システムに組み込み、評価実験を行なった。図3に実験結果を示す。入力データとしては、ビジネスに用いられる文書50ページを用いた。図3の縦軸は、共起情報を用いない時に誤変換であった文節が正解になった数を表し、横軸は、共起情報の数を表している。また、破線は、従来の係り受け関係のみに共起関係を適用した時の効果を、実線は、本方式の効果を表している。

図3に示すように、共起情報の数にかかわらず、係り受け関係のみに共起関係を適用した時に比べて、本方式によって1.8～2.2倍の効果が得られることを確認した。

５．おわりに

本稿では、共起関係が意味的に強い結び付きを持つ関係(意味共起関係)となる条件について述べ、意味共起関係を複数文にわたって適用することによって、文脈に適した同音語を優先するかな漢字変換方式を提案した。本方式を用いることによって、係り受け関係のみに共起関係を適用した時の約２倍の効果が得られることを確認した。なお、今後、意味共起関係の条件をさらに精密化してゆく予定である。

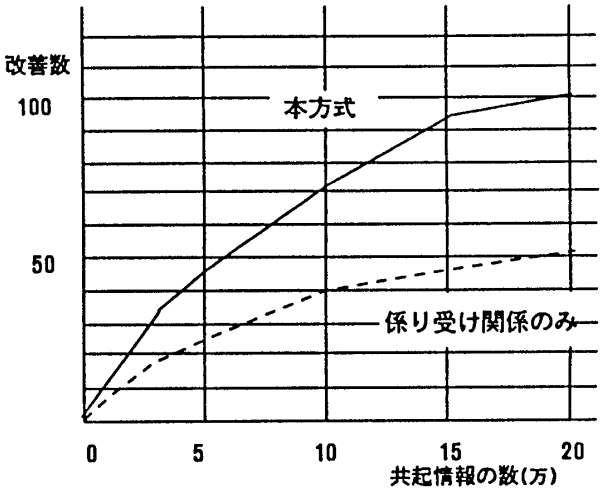

図3 実験結果