

4 P-11

共起グループを用いたかな漢字変換

山本 喜大 久保田 淳市

松下電器産業株式会社 情報システム研究所

1 はじめに

変換精度を向上させることを目的に、かな漢字変換の開発を進めている。かな漢字変換の誤変換を調査した結果、誤変換のうち同音語誤りが最も多いことがわかった。そこで、同音語誤りの削減に焦点を絞り、共起グループを用いて同音語を選択するかな漢字変換を試作し評価した。

2 共起グループ

「便宜を図る」の中の「便宜」「図る」のように、ほぼ決まり文句として、同一文中で直接的に修飾関係にある2単語の組を「関連語」と呼ぶ。

これに対し、例えば、ビジネス手紙文中の「拝啓」「貴店」「承る」のように、特定の文脈で高い頻度で出現する単語同士は、直接共起してはいないが間接的に互いに弱い共起関係にあると考えられる。そこで、この単語のグループを共起性に着目して「共起グループ」と呼ぶ。

本かな漢字変換は、共起グループ中の単語相互の共起性を使って関連語では表現できなかった共起関係を補い、変換対象文に対する共起関係の適用範囲を拡大することにより、同音語誤りを削減しようとするものである。

3 共起グループを用いたかな漢字変換方式

3.1 かな漢字変換の概要

本かな漢字変換方式の変換手順の概略を以下に記す。

1. 文節候補の生成
単語辞書と付属語辞書を検索し、文節候補を生成する。
2. 文節区切りの決定
三文節最長一致法を基本とし、補助規則に基づいて文節区切りを決定する。
3. 同音語順位の決定
規則に基づいて同音語順位を決定する。
4. 関連語を用いた同音語の選択
共起する2単語を保持する関連語辞書を検索し、文節候補列中に共起2単語がある場合には、当該文節の第1候補とする。

5. 共起グループを用いた同音語の選択

共起グループを用いて同音語を選択する。

ただし、関連語関係にある文節の順位は変更しない。

3.2 共起グループを用いた同音語の選択

本方式は、共起グループを識別するキーワードを設け、キーワードにより変換対象文に対応する共起グループを選択し、さらに共起グループ中の単語相互の共起性を同音語の選択に使う。

共起グループを用いた同音語選択アルゴリズムを以下に記す。

1. 共起グループの種類を選択

各文節の第1候補をキーにして、第1文節から順に第1候補が共起グループのキーワードかどうかを調べる。

2. 同音語の選択

キーワードである場合には、次のようにして同音語を選択する。

各文節の各同音語候補をキーにして、優先度の高い同音語候補から順に単語の属する共起グループの種類を調べる。共起グループの種類がキーワードと同じ場合には、その文節の第1候補とする。

図1に共起グループの処理の一例を示す。

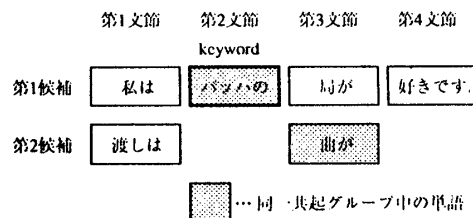


図1: 共起グループの処理

4 かな漢字変換実験

4.1 共起グループ辞書

主観的判断に基づき一般ビジネス文書と音楽関連文書から単語を抽出し、ビジネス文書と音楽関連文書に対応する共起グループ辞書を試作した。

試作した共起グループ辞書のグループ別語数を表1に示す。

表1: グループ別語数

共起グループの種類	単語数
ビジネス文書	555
音楽関連文書	316

4.2 キーワードの設定

かな漢字変換処理において、文節候補が唯一であり、その文節中の単語がある共起グループに属するなら、その共起グループが変換対象文に対応するグループとみなしうる。

そこで、共起グループに属する単語のうち同音語を持たない単語を、共起グループのキーワードとした。

キーワードのグループ別語数を表2に示す。

表2: キーワードのグループ別語数

共起グループの種類	キーワード数
ビジネス文書	219
音楽関連文書	229

4.3 実験1

ビジネス文書(23301文節)と音楽関連文書(944文節)に対し、関連語を使う場合と使わない場合の双方における本方式の変換率への寄与を、表3に示す。

表3: 共起グループ利用時の変換率への寄与

変換率への寄与(%)	ビジネス文書	音楽関連文書
効果1: 関連語未使用	0.14	1.59
効果2: 関連語使用	0.03	1.27

本方式により、変換率で0.03%～1.59%の効果を得た。本方式では、例文による効果の差は大きいですが、改悪される例が少ないという特徴がある。

なお、効果1より効果2の方が変換率への寄与が小さいのは、共起グループ中の一部の単語からなる共起対が、関連語として存在するためである。

4.4 実験2

変換対象文の内容に対応する共起グループをあらかじめ指定し、キーワードの有無に関わらず、グループ中の単語が文節候補中にある場合に、その文節候補を無条件に優先させた場合の変換率への寄与を表4に示す。

表4: 無条件優先時の変換率への寄与

変換率への寄与(%)	ビジネス文書	音楽関連文書
効果3: 関連語使用	0.01	1.48

ビジネス文書で、効果3が効果2より変換率への寄与が小さいのは、無条件に優先させた場合、改悪される副作用が多く生じるためである。

5 考察

1. 共起グループを用いたかな漢字変換の特性

音楽関連文書とビジネス文書では効果1で約10倍の差がある。これは、音楽関連文書はビジネス文書よりも変換対象文中に占める共起グループ中の単語の割合が大きいためである。この差は、共起グループの特殊性の度合に関係しており、ある程度対象を絞る方が効果を得やすいと考えられる。

2. キーワードの効果

特殊性の高い音楽関連文書では、効果3に対して効果2は変換率への寄与の比で約8割に留まるが、汎用性の高いビジネス文書では、改悪される副作用が多いため、効果2の方が効果3より変換率への寄与が大きい。キーワードは、汎用性の高い共起グループにおいては、誤った同音語順位の変更を抑制する効果があると考えられる。

6 まとめ

かな漢字変換の同音語誤りの削減を目的として、共起グループを用いて同音語を選択するかな漢字変換を試作した。

実験の結果、本方式が従来の分野辞書と比べ、汎用性の高い共起グループでも、キーワードにより副作用を抑制可能であることを確認した。

今後の課題として、適切な共起グループの種類選定と、客観的かつ効率的な共起グループの作成方法の開発が挙げられる。

今後、共起グループの数を増やし開発を進める予定である。