環境による文脈変化を組み込んだかな漢字変換

２０２５年度　課題研究Ⅱ　お茶の水女子大学附属高等学校３年　友寄美尋

要旨

複数のかな漢字変換システムをデバイス上に実装し、入力環境に応じて切り替える方式の有効性を検証した。コーパスから同音異義語を抽出し、出現頻度の高い同音異義語を対象に、特定の語彙分布での変換実験を行った。得られた変換候補の傾向から、語彙分布の偏りと変換精度との関係を求め、全体をまとめて変換した場合と環境別に変換した場合の変換精度を予測・比較した。その結果、従来の方法よりも変換精度が向上したが、有意差は見られなかった。今後は、より高精度な実装方法を研究する予定である。

研究背景　新たな視点からかな漢字変換精度を上げたい！

・ひらがなを漢字へ変換する「かな漢字変換」
　→同音異義語を原因とする誤変換が多く発生[1]
・従来：文章の内容に基づいて文脈解析[2]
　→入力する「環境」による文脈変化[3]は考慮されていない

アプリや相手などの**環境に応じて変換システムを切り替える**
←従来のかな漢字変換システム(IME)に環境に応じた文脈変化を組み込む

実験　語彙数2~5の同音異義語の変換実験

語彙数が2~5の同音異義語のうち『現代日本語書き言葉均衡コーパス』で出現回数が多い5種類ずつ*を抽出し変換実験（図1）

①前回の変換で選択した語彙が最上位候補となる
②それまでの変換での出現回数に従って変換候補が並ぶ
傾向を発見
→傾向をもとに候補順位を予測
→実際の変換との予測一致度を算出（図2）

正規化エントロピー**を用いて結果を比較（図3）
→語彙分布の偏りを定量的に評価

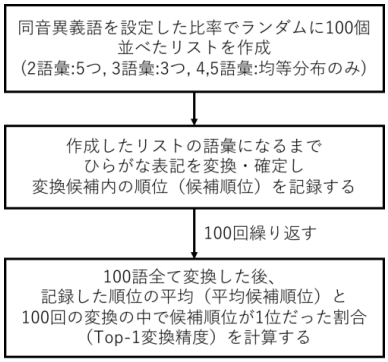

図1 実験操作

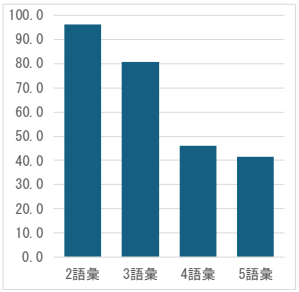

図2 変換候補予測一致度
（各語彙数での平均値）

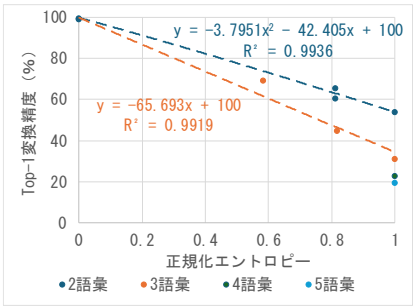

図3 実験1 結果（Top-1変換精度）

1. 直前の変換とそれまでの出現回数によって候補順位が推測できる

2. 語彙数が少ないほど・分布が偏っているほど変換精度が高くなる

＊2語彙：「ある」(或る/有る), 3語彙：「いる」(居る/要る/入る), 4語彙：「ご」(語/御/後/五), 5語彙：「じゅう」(銃/重/十/住/中) など

＊＊正規化エントロピー：分布の偏りの程度の指標。値が0に近いほど分布の偏りが大きい（特定の語彙が多く出現する）分布であることを表す。

要素数 n の各要素 i に対する確率分布 p_i に基づくエントロピー $H = -\sum_{i=1}^{n} p_i \log_2 p_i$ をその要素数での最大エントロピー $\log_2 n$ で割ることで0~1の範囲に値を収めた $H_{norm} = \frac{H}{\log_2 n}$ として算出

考察　実験を利用したシミュレーション

実験の結果から語彙分布の正規化エントロピーとTop-1変換精度の関係を散布図を用いて導出（図4）

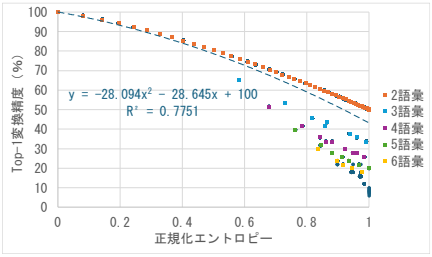

図4 正規化エントロピーとTop-1変換精度

1. 『現代日本語書き言葉均衡コーパス』内の12ジャンルのデータを取得
2. ジャンル別に正規化エントロピーを計算（図5）
3. 散布図で得た近似式に各ジャンルの正規化エントロピーの値を代入
4. 「ジャンル（＝入力環境）別にIMEを使い分けた場合」と「全てをまとめて変換した場合」の変換精度を予測（表1）

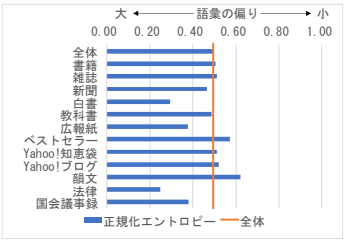

図5 ジャンル別
正規化エントロピー

表1 全体・ジャンル別変換の変換精度予測

	Top-1	候補順位
全体	**79.0%**	**1.30位**
ジャンル別	**80.7%**	**1.28位**
法律	91.2%	1.12位
韻文	71.4%	1.41位
全体・ジャンル別 T検定	**0.169**	**0.171**

ジャンル別変換の方が変換精度が向上したが、有意差は見られなかった

白書・広報紙・法律・国会議事録で効果◎

まとめと展望　環境の最適化・他の観点での検証

単語単位のかな漢字変換：環境の区別方法によって効果が分かれる
今後の課題：実装方法の改善、メモリ使用量・処理時間への影響の考慮、環境の区別方法の最適化
単語単位の変換でメモリ使用量・処理時間の測定、環境区別の最適化
→文節単位・文単位へ適用拡大

謝辞

本研究にあたり、お茶の水女子大学理学部情報科学科の戸次大介先生より、研究内容・手法や成果の伝え方など貴重な助言を数多くいただきました。また、お茶の水女子大学附属高等学校の山上通恵先生、十九浦美里先生、朝倉彬先生をはじめとする多くの方々にご相談に乗っていただきました。この場を借りて深く御礼申し上げます。

参考文献

[1] 山本喜大・久保田淳市. 共起グループを用いたかな漢字変換. 情報処理学会第44回全国大会. 1992, no.3, pp.189-190
[2] 加藤省三・荒木睦大・小越康宏・谷口秀次・森幹男. かな単語マルコフ連鎖モデルを用いたかな漢字変換法. 電気学会論文誌C. 2010, vol.130, no.6, pp.1054-1060
[3] 藤田悟・相田仁・猪瀬博・斉藤忠夫. 文脈理解を支援する知識の学習に関する研究. 情報処理学会第33回全国大会. 1986, pp.1675-1676
[4] 前川喜久雄（監修）・山崎誠（編）. 書き言葉コーパス―設計と構築―. 朝倉書店, 2014, 講座日本語コーパス2
[5] 国立国語研究所 コーパス開発センター. 『現代日本語書き言葉均衡コーパス』短単位語彙表 (Version1.1). 2021
[6] 国立国語研究所 コーパス開発センター. 『現代日本語書き言葉均衡コーパス』語彙表解説. 2021
[7] 国立国語研究所. “設計の基本方針 現代日本語書き言葉均衡コーパス（BCCWJ）”. 国語研コーパスポータル. https://clrd.ninjal.ac.jp/bccwj/basic-design.html, (参照 2025-08-04)