文脈解析を用いた仮名漢字変換システムの開発

北澤 遼太 † 奥村 紀之 †
† 長野工業高等専門学校 電子情報工学科

1 はじめに

現在使用されているかな漢字変換システムは，変換履歴や統計的手法から変換効率を上げる方法が取られている．しかし，まだまだ初出の単語については誤変換が行われる確率が高い．本研究では，一つのテーマを持つ文章を書く際に，関連の強い単語同士が頻繁に使われると仮定し，かな漢字変換システムに関連度計算方式を用いる手法を提案し評価する．

2 かな漢字変換について

かな漢字変換を行うには，まずユーザからの入力に対して，形態素解析を行い，単語に分割する．それらの単語と同じ読みを持つ漢字を辞書から抽出する．抽出した表記を候補とし，ユーザが選択することによって変換が終了する．これらの各段階について以下に詳細を述べる．

2.1 形態素解析部分

ユーザの入力に対して行われる処理が形態素解析である．ここでは，解析器 MeCab[1] により，かなで書かれた文を文節に区切る分かち書きのみを行う．

2.2 辞書引きと候補選択部分

辞書引きアルゴリズムには，Double-Array やパトリシア木などといった方法がある [2]．本研究では高速化などは行わないため辞書引きアルゴリズムについての検討はしない．

変換候補については，変換したい候補が1番最初に出てくることが望ましい．

3 提案するシステム構成

図1に提案するシステムの構成を示す．入力から出力までの流れはかな漢字変換の基本となる部分である．本研究では一つのテーマに対する長文を執筆する際の初出単語の変換効率を上げるシステムを提案する．

文節区切りにおいては，入力がひらがな文であるため，コーパスによりコスト値 (連接コスト，生起コスト) を推定し直した辞書を利用する．単語の選択では，その文章から抽出した単語やその文章を書かれるに至った変換履歴との関連度 [4] を用いることで，文章の意味を捉えた上での変換候補の抽出を行う．

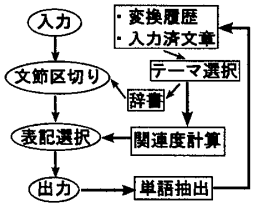

図 1: システム構成

以下にそれぞれの部分について詳しく説明する．

3.1 コーパスを使用したコスト推定

本研究では，コスト推定のシステムとして MeCab を使用している．MeCab ではあらかじめコストを推定した辞書を保持しており，コストの設定により，文節区切りの精度が変化してくる．本研究では一つのテーマについての長文を前提としているため，それらに適応した新たなコストを得ることが必要となる．

3.1.1 実験

実験として，読売新聞の記事1ヶ月分を用いて作成したコーパスと，入力として被験者30名から収集した378文を用意した．この文を MeCab 標準の辞書で解析した場合と，コーパスを用いて，新たなコスト値を推定した辞書で解析した場合を比較した．

結果として，378文中 MeCab 標準辞書では240文が，新たなコスト値を用いた辞書では235文が分かち書きの分割数が一致した．この中で分割位置が期待する箇所と違ったものは17文と20文であった．このように文節区切りの精度に大きな差は見られなかったが，正しく分割できる文に変化が見られた．例として以下のような結果が得られた．

例文

朝顔/に/毎朝/水/を/あげる/。(計 378 文)

標準辞書による結果

あさがお/に/まい/あさみず/を/あげる/。

コスト推定後の辞書による結果

あさがお/に/まいあさ/みず/を/あげる/。

A Development of Kana-Kanji Conversion System based on Context Analysis

† Ryota KITAZAWA
† Noriyuki OKUMURA(noriyuki_okumura@ei.nagano-nct.ac.jp)
Nagano National College of Technology, Department Electoronics and Computer Science (†)

この結果より，コストを変化させることにより，単語の区切り位置が変化することは確認できた．またコーパスの設定により 90%を超える変換精度が得られる事が知られている [3]．以降の節では，正しく文節に区切れることができる前提として評価を行う．

3.2 関連度計算方式を用いた変換候補選択

変換候補選択部分については関連度計算方式 [4] を用いる．編集している文章中や変換履歴から単語を抽出しておき，変換対象となる変換候補と関連度を計ることにより，変換候補を推定する．

本研究では，テーマを仮定し，特徴を表すような単語をあらかじめ抽出したとし，関連度計算により，期待する変換候補が得られるかの実験を行った．以下に詳細を述べる．

3.2.1 実験 1

はじめに，「宇宙」に関する論文を書いているとし，以下に示すテーマから連想される単語を抽出したとする．

抽出単語

地球　太陽　惑星

ここで，変換したい読みを「きんせい」とする．ここで変換を期待する単語は「金星」である．辞書からこの読みに一致する単語を抜き出し，抽出単語とそれぞれ関連度を計算する．その結果を表 1 に示す．

表 1: 関連度計算結果 (宇宙)

	地球	太陽	惑星
金星	0.031177	0.026703	0.082209
近世	0.005034	0.002529	0.009776
均整	0.002563	0.002433	0.001627

次に，テーマを「歴史」とし，抽出単語を新たに設定し，同様の実験を行った．その結果を表 2 に示す．

表 2: 関連度計算結果 (歴史)

	中世	ヨーロッパ	革命
近世	0.373854	0.045363	0.024702
金星	0.006202	0.005108	0.004735
禁制	0.004995	0.001756	0.003961

3.2.2 実験 2

次に，一般的に使用されている入力システムとの比較結果を表 3 に示す．

評価方法は，各入力システムについては，読みをすべて入力し第一変換候補を結果とした．$tf \cdot idf$ と関連度計算方式は，「きみ」の変換のみを行った．また，テーマは食べ物や料理として，抽出単語を設定した．

表 3: 各入力システムとの比較

原文	黄身が好きだ
MS-IME	キミが好きだ
GoogleIME	君が好きだ
$tf \cdot idf$	君
関連度計算方式	黄身 (白身、卵、料理)

() 内は抽出単語

3.2.3 考察

実験 1 の結果から，「宇宙」というテーマでは，「金星」という表記がそれぞれの抽出単語に対し最も関連度が強く，「歴史」というテーマでは「近世」という表記が最も関連度が強いことがわかる．よって，これらを変換の第一候補とすればよいことがわかる．

実験 2 の結果から一般的には「きみ」の表記として「君」が選ばれる可能性が高いことがわかる．しかし，原文では「黄身」と変換されることを期待しているので，誤変換という結果になってしまう．

この他に 10 種類 (きみ，かいとう，ほんこう，きんこう，こうたい，ちゅうせい，せいそく，せんたく，いこう，こうせん) の同音異表記語を比較した結果，関連度計算方式を用いることにより，こちらが意図した表記を候補としてあげる結果となった．

4 おわりに

本研究では，文脈解析を用いたかな漢字変換システムの手法として関連度計算方式を提案した．実験により，テーマに沿った単語を設定することで意図する変換結果を得られることがわかった．ただし，テーマの設定方法や単語抽出方法は今後の検討課題である．

参考文献

[1] 工藤拓：汎用形態素解析器 MeCab

[2] 望月久稔，中村康正，尾崎拓郎：ダブル配列によるパトリシアを拡張した基数探索法，日本データベース学会 Letters Vol.6，No.1，pp.9-12(2007).

[3] 森信介，土屋雅稔，山地治，長尾真：確率的モデルによる仮名漢字変換，情報処理学会論文誌，Vol.40，No.7，pp.2946-2953(1999).

[4] 渡部広一，奥村紀之，河岡司：概念の意味属性と共起情報を用いた関連度計算方式，自然言語処理，Vol.13，No.1，pp.53-74(2006)