4W－10

自動短縮登録システムにおけるテキスト辞書の選択機能の検討

蓮井洋志　西野順二　小高知宏　小倉久和
福井大学情報工学科

1　はじめに

仮名漢字変換システムにおける変換機能はユーザの入力した仮名文字列を普段読み書きしている表記に置き換える．変換機能は，システムが仮名文字列を単語に分割し，辞書を検索してその語の変換表記を取り出す．日本語の語すべてを検索の対象とすると，競合する変換結果が増えるためにユーザがその候補群から意図した候補を選択する手間が大きくなる．そこで，従来の仮名漢字変換システムでは分野などを基準として複数の辞書に分割し，ユーザが入力テキストの内容にあわせて辞書を選ぶことで，変換結果の中で明らかに選ばれる可能性のない語が競合することを防いでいる．

しかし，辞書を分割した結果，ユーザがテキストの内容にあった辞書のグループを選択する必要が生まれた．書いているテキストの内容にあった辞書グループ，つまり (i) テキストに出現するが辞書には未登録である語の数が最も少なくて，(ii) 総登録語数ができる限り少ない，辞書グループを選ぶことは難しい．

本稿では，自動短縮登録システム [1] においてテキスト辞書を選択する機能について検討する．自動短縮登録システムは仮名漢字変換システムに，以前に入力した語を短縮した読みでテキスト辞書に自動登録する自動短縮登録機能と，テキスト辞書を検索してユーザが短縮した読みを変換する短縮変換機能を付加したシステムである．テキストの内容にあったテキスト辞書を選択することで，短縮変換での変換結果の競合を減らしたままで，辞書内にそれ以降の文章で出現する語の数を増やす効果を期待する．

本研究の選択では，テキストの内容にあったテキスト辞書を選ぶために，現在入力中のテキストに対応したテキスト辞書と類似したテキスト辞書を選択する．類似したテキスト辞書を選択する基準として，テキスト辞書間の類似度を定義する．

Discussion about Function for Selection of Text Dictionary in System of Dynamic Registration of Shortening Hiragana Representation
Hiroshi Hasui,Junji Nishino,Tomohiro Odaka,
Hisakazu Ogura at Department of Information Science in Fukui University

2　テキスト辞書の選択

2.1　辞書の選択

仮名漢字変換システムにおいて辞書の分割や選択は，テキストの内容にあった登録語の集合を作る目的がある．ここで，辞書グループが入力中のテキストの内容にあうとは，登録語の集合がテキストに出現する単語をほぼ網羅していて，かつ辞書グループの総登録語数がなるだけ小さい状態を指す．テキストの内容に適切な辞書グループを選択できれば，変換結果の競合が減るために，変換時の第一候補の正解率が向上する．

自動短縮登録システムにおいては，短縮変換のためのテキスト辞書をテキストごとに用意する．テキスト辞書をテキストごとに用意することは，テキストの入力において変換候補の中に不要な候補を少なくする効果を狙っている．しかし，テキストごとに辞書を分割したためにテキスト辞書の数が多く，ユーザには分割した辞書の中のどの辞書を選択すればよいかは容易に分からない．

本研究では，仮名漢字変換システムに自動短縮登録機能，短縮変換機能を付加した自動短縮登録システムにおいて，入力中のテキストの内容にあわせてテキスト辞書を選択する機能を検討する．現在入力中のテキストのテキスト辞書に対応した類似したテキスト辞書を選ぶ．

2.2　テキスト辞書間の類似度

テキスト辞書が類似している度合をはかるために，テキスト辞書間の類似度を以下のように定める．ただし，類似度は α，n は共通語の数，N はテキスト辞書の単語の総数，共通語の変換表記の長さを e_i，テキスト辞書の各々の語の変換表記の長さを f_j とする．$\lfloor\ \rfloor$ の中の数式の小数点以下を切り上げる演算である．

$$\alpha = \frac{\sum_{i=1}^{n} \lfloor e_i/2 \rfloor * 2}{\sum_{j=1}^{N} \lfloor f_j/2 \rfloor}$$

複合語を構成する単語の長さは２文字であることが多い．$\lfloor e_i/2 \rfloor$ は近似的に共通語を構成する単語数を表す．構成する単語数が多い語は，短い語と比較してより専門化した内容を表すことが推測できる．テキストの特徴を表した語が共通に登録されていることは，テ

表 1: 論文間の類似度

	A	B	C	D	E
A	1.000	.3806	.1838	.09100	.02458
B	.3806	1.000	.09410	.06562	.01241
C	.1838	.09410	1.000	.07004	.02126
D	.09100	.06562	.07004	1.000	.01653
E	.02458	.01241	.02126	.01653	1.000

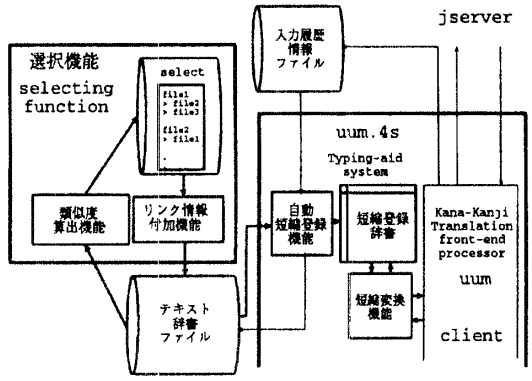

図 1: テキスト辞書の選択機能の概要

キスト辞書の性質がより類似していることを意味する．

自動短縮登録システムを用いて情報処理関係の論文を 4 編入力し，その結果作られたテキスト辞書間の類似度をはかった．論文 A は自動短縮登録に関する論文で，論文 B は自動短縮登録に関する研究発表の予稿である．論文 C は予測を利用した入力補助方式に関する論文である．論文 D は推敲支援システムに関する論文である．テキスト E は EDR 電子化辞書の朝日新聞の記事のコーパスから作成したテキスト辞書である．総当たりでこれらの文章のテキスト辞書間の類似度をはかった．その結果を図 1に示す．

同一ファイルの類似度を除けば，A と B の類似度が 0.3806 で最も大きく，次に大きいものは A と C の類似度で値は 0.1838 である．テーマも研究分野も異なる A と E との類似度は，わずか 0.02458 である．また，A と B，C，D，E の類似度を見ると，内容が離れていくにつれて，類似度が小さくなる．この結果から，お互いに内容の近いテキストのテキスト辞書ほど類似度が大きいことが推測できる．

3 テキスト辞書の選択機能の実現

選択機能は類似度を算出する機能 (類似度算出機能) と，類似度が近いテキスト辞書に対するリンク情報を付加する機能 (リンク情報付加機能) から構成される．選択機能の構成図を図 1に示す．

選択機能の実行はユーザがシェルで指令する．ユーザが選択機能を指令すると，類似度算出機能が各々のテキスト辞書と現在入力中のテキストに対応したテキスト辞書の間の類似度を計算し，ユーザの指定した閾値以上の類似度を持ったテキスト辞書ファイル名を select ファイルに登録する．次に，リンク情報付加機能は類似度が select ファイルの中の情報を用いて，現テキスト辞書と類似したテキスト辞書のファイル名を現テキスト辞書に付加する．

自動短縮登録機能は，テキスト辞書内の語を短縮登録辞書に登録するだけでなく，テキスト辞書内のリンク情報で指定されたテキスト辞書ファイル内の語も登録する．同じ表記で同じ短縮読みの語が既に短縮登録辞書に存在する場合は，その語は登録しない．また，短縮登録辞書の最大登録語数は 10000 語とし，それ以上の単語の登録に関しては LRU で管理する．

4 考察

テキスト辞書の選択を行う上で問題になるのは，これから書こうとしているテキストの内容にあったテキスト辞書を選ばなくてはいけないことである．これから出現する語の集合が分からないのに，その集合の中の単語が多く出現するテキスト辞書を選ばなくてはいけない．この研究では，テキストを途中まで書き終えた時点でそれまでの文章に出現した語が多く登録されているテキスト辞書を選択する．早い段階で適切な選択ができれば，その辞書を用いて入力する残りの文章が多いためテキスト辞書の選択の効果が大きい．しかし，早すぎる段階ではテキスト内の語の数が少ないため，適切な選択ができない．どの程度まで入力した時点で選択機能を起動すると最適なテキスト辞書が選択できるかは，選択の基準となる類似度の閾値やテキスト辞書の大きさ，これから入力しようとしている文章の内容などに依存するため，容易には把握できない．現段階ではユーザにその判断を委ねている．選択をどのタイミングで行うと有効かを特定することがこれからの課題である．

参考文献

[1] 蓮井洋志，西野順二，小高知宏，小倉久和：頻出する自立語の動的な推定による入力補助，第 54 回全国大会講演論文集，pp. 4/249 − 250，情報処理学会，(1996)