2J-6　曖昧変換システムと従来型の日本語入力システムの比較評価実験*

蓮井洋志†
室蘭工業大学情報工学科‡

1　はじめに

日本語の入力には仮名漢字変換システムを用いる。仮名漢字変換システムによる入力は単語の読みと文を文節に切り分ける知識を必要とする。日本語の入力は英語とちがって文字の種類が多いため、入力に手間がかかる。

我々は、これまでに日本語の入力の効率を上げるために入力補助システムを開発した。日本語入力補助手法の1つとして、曖昧変換システムがある。このシステムはユーザが入力する短縮形と正しい読みの類似度を定義し、類似度が大きい漢字表記を変換結果とする。類似度を定めるために、大学の学生に短縮形を書くアンケートを実施し、類似度の閾値を統計的に定めた。また、単語毎に閾値を定め、その値と類似した曖昧度を持つ漢字表記を優先的に順番を定める。

本研究では、曖昧変換システム **uum.5a** と従来の仮名漢字変換システム **uum** と比較評価実験を行なった。その結果、入力に必要な所用時間や打鍵数などが1割程、少なくて済むことがわかった。このシステムは、繰り返し現れる複合語を入力するときに打鍵数が少なくて済む。話者の喋った内容を即座にメモする速記に向いたシステムであることがわかった。

本稿では、2節では従来型の仮名漢字変換システムと曖昧変換システムの入力実験をおこなう。3節では実験結果をふまえて考察する。

2　比較評価実験

比較入力実験の結果を表1に示す。実験の目的は、仮名漢字変換システムと曖昧変換システムのどちらが速く、打鍵数を少なく、正確に入力できるのかを見積もることである。

この実験は3日にわたって4つのテキストを対象に行なった。まず、1日目には、社説を2編、2日目にはコンピュータ雑誌、3日目に情報処理に関する論文を入力した。4つのテキスト各々に対して、まず、被験者が仮名漢字変換システムでそのテキストを入力し、実験用の文章の内容を把握しておく。30分以上リラックスした後に仮名漢字変換システムで入力し、その後、30分後に曖昧変換システムで入力した。2回目と3回目のデータを比較する。1回目の入力は、仮名漢字変換システムと曖昧変換システムの条件を整えるために必要である。

計測するパラメータは、曖昧変換による入力文字数、曖昧変換成功回数、曖昧変換失敗回数、入力誤り、仮名入力の打鍵数、総打鍵数、所用時間の7つである。この中で、曖昧変換成功回数は曖昧変換によって変換結果が得られた回数のことである。曖昧変換失敗回数は変換結果がなかった回数である。入力誤りとは入力結果の中と本物のテキストの間で食い違った箇所の数である。仮名入力の打鍵数とは読みの入力に要する打鍵数のことである。仮名入力はローマ字入力で行なった。総打鍵数はエディタ内での制御コードを混ぜた入力実験すべてにおいて打鍵された数を表す。

すべてのテキストの入力において、仮名入力の打鍵数は曖昧変換システムの方が従来の仮名漢字変換システムより約1/8少ない。所用時間は明らかに曖昧変換システムの方が約1/8ほど短い。これは曖昧変換システムがワープロで速記をする時に役立つことを示している。

社説にはこの仮名漢字変換システムのシステム辞書に登録されていない語が繰り返し使われていた。その結果、従来の変換では、単漢字変換を利用してしか入力できないために選択の打鍵数が多くなる。本システムは2度目以降の入力では未登録の単語を曖昧変換できるために、選択回数が少なくて済む。このため、社説は他のテキストと比較して選択、伸ばし、縮みなどの制御コードの打鍵数が多い。

曖昧変換によって入力された文字数は社説のテキスト全体の約1/18であるのに対して、コンピュータ雑誌、論文では約1/7である。社説はコンピュータ雑誌、論文より文字数が少ないことと、多くの長い外来語、複合語が少ないことが原因である。

入力実験すべてを合計すると、曖昧変換成功回数は失敗回数と比較して、約1/17である。以前の文脈にある語しか変換の対象とならないために覚えておかな

*Comparation of Approximate Translation System with Conventional IME
†Hiroshi Hasui
‡Department of Computer Science and Systems Engineering in Muroran Institute of Technology

表 1: 入力実験

(1) 比較入力実験 (社説 I) 1256 文字

	仮名入力キー打鍵数	総打鍵数	入力誤り	所要時間	曖昧変換成功回数	曖昧変換失敗回数	曖昧変換による入力文字数
既存のシステム	2730	3937	7	1194	-	-	-
曖昧変換システム	2643	3488	8	1008	8	2	28

(2) 比較入力実験 (社説 II) 1151 文字

	仮名入力キー打鍵数	総打鍵数	入力誤り	所要時間	曖昧変換成功回数	曖昧変換失敗回数	曖昧変換による入力文字数
既存のシステム	2723	3692	9	1146	-	-	-
曖昧変換システム	2609	3409	7	984	26	4	105

(3) 比較入力実験 (コンピュータ雑誌) 2298 文字

	仮名入力キー打鍵数	総打鍵数	入力誤り	所要時間	曖昧変換成功回数	曖昧変換失敗回数	曖昧変換による入力文字数
¡ 既存のシステム	5032	6355	7	2070	-	-	-
曖昧変換システム	4638	5964	5	1825	61	3	300

(4) 比較入力実験 (論文) 2282 文字

	仮名入力キー打鍵数	総打鍵数	入力誤り	所要時間	曖昧変換成功回数	曖昧変換失敗回数	曖昧変換による入力文字数
既存のシステム	5718	6918	7	2025	-	-	-
曖昧変換システム	5086	6529	9	1816	90	2	349

いと、曖昧変換は失敗する。しかし、実験結果では失敗回数は決して大きい数ではない。

入力誤りの数は、5 つのテキストそれぞれに 7〜8 個ある。これはすべてささいな誤りで入力時にテキストを見間違えたためである。被験者の打鍵が 1 秒間に 3 回程度で、速い入力であることも入力誤りが生じた原因である。入力誤りの数は両方のシステムともに大差がない。

3 考察

朝日新聞社説 100 編を対象に調べた結果、4 文字以上の読みを持つ自立語連鎖の中で、繰り返し使われる語の割合は全体の 54% である。つまり、1000 文字程度の長い語の存在しない社説においても、約 5 割の語に対して、曖昧変換を利用できる計算になる。

入力履歴情報ファイルとテキスト辞書ファイルを文書ごとに持てば、文書中に頻繁に現れる語を登録するために、変換結果の競合が少ない。人間によって、ちがう文書を書いていても、良く使う表現が一致することが多い。頻繁に曖昧変換を行なう単語については、共通辞書を作るべきである。

入力補助システムには、曖昧変換システムの他に予測システム、自動短縮登録システム [1] などがある。予測システムは、前の文脈の情報から次入力の単語の読みを予測し、補完するシステムである。前の文脈の語と同じ表記が予測されるために、予測結果には揺れが入らない。しかし、読みを予測するために予測結果の選択とそれを変換した結果の選択の 2 度手間がかかる。曖昧変換システムはユーザの入力した短縮形を仮名漢字交じり表記に変換することでこの問題を解決した。

自動短縮登録システムは、前の文脈に現われた語を短縮形で、辞書にすべて登録する。再度入力をする時にはその短縮形で逆変換する。この変換を短縮変換と呼ぶ。前の文脈と同じ表記に変換する。また、短縮形で変換するために、繰り返される長い語の入力の手間も減る。このシステムは規則で短縮形を決定する。そのため、仮名入力の前にユーザは短縮形を作り記憶する必要がある。これは、ユーザにとって負担である。

曖昧変換を応用した自動短縮登録システムでは短縮形の規則にこだわらない設計であるが、システム表記が短縮形であるために変換結果の競合が多過ぎるという問題点があった。本システムは短縮形ではなく読みを登録する自動登録システムに、曖昧変換を持ち込むことでこれら 2 つの問題点を解消している。

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. 情報処理学会第 54 回全国大会講演論文集, pp. 4/249–4/250, 1997.