キーワード方式べた書き文かな漢字変換システムにおける付属語情報を用いた単語のあてはめ

2K-6

荒木健治、内田幸司、山田洋志、栃内香次、永田邦一

北海道大学工学部

1．はじめに

我々は、べた書き文を確実性の高い部分から段階的に分割して変換する方式（キーワード方式）によるべた書き文のかな漢字変換システムを開発している[1]。このシステムでは、最初べた書き文字列中から一意に識別可能なキーワードによってべた書き文をいくつかに分割し、ついでその各部分ごとに一般の単語をあてはめている。これまで得られた実験結果によると、本システムにおける変換エラーの約７割は、一般単語のあてはめの際に、あてはめるべき語を決定するための連接情報（単語の前後に連接している文字の組）が辞書中に存在しないことによる未変換である[2]。しかし、これに対して、辞書中の連接情報を増やすと誤変換も増大する。そこで、付属語を手掛りとしてあらかじめ単語を分類しておき、この情報を併用して一般単語のあてはめを行うことを検討している。本稿では、そのアルゴリズムおよびそれを用いて文章中より関連のある語を取り出した実験結果について述べる。

2．付属語情報を用いた単語のあてはめ

2．1　付属語情報の取り出し方法

単語への分割処理を終了し、字種指定された文字列より付属語情報を取り出す。この例を図１に示す。今、あらかじめ図１（a）の①～⑤のようなルールが与えられているとする。このルールを用いて図１（b）の文を解析すると、図１（c）の(i)～(v)のような関連語（修飾、並列などの関係にある語）の組が得られ、これを各語ごとに、その関係を示すフラグと共に辞書に蓄えておく。

① AはBてある → AとBは主語と述語
② AなB → AはBの修飾語
③ AやB → AとBは並列
④ AのB → AはBの修飾語
⑤ AについてBてある → AはBの対象

(a)　ルール

本章｜は、｜アセンブラ｜言語｜の｜起源｜や｜基本的｜な｜性格｜について｜述べ｜てある。

注）「｜」：単語の切れ目
「＿」：ルールとして着目する付属語

(b)　解析結果

(i)　①より「本章」は「述べ」の主語
(ii)　②より「基本的」は「性格」の修飾語
(iii)　②③より「起源」と「性格」は並列
(iv)　④より「言語」は「起源」の修飾語
(v)　⑤より「性格」は「述べ」の対象

(c)　得られた関連語の組

図１　情報の取り出し方法の例

2．2　単語のあてはめ方法

上述のようにして取り出された情報を利用して一般単語のあてはめを行う。この方法を図２の例文で説明する。例文を本システムで処理すると助詞候補の検出が終った時点では図2（a)のようにあてはめられている。ここで、例えば「言語」の後ろの助詞「の」を付属語として辞書中の「言語」の項の被修飾語のグループを見ると「起源」が図１の解析により蓄えられているので、これをあてはめる。また、同様にして「性格」、「について」、「てある」より「性格」の被対象語である「述べ」をあてはめる。この結果を図2(b)に示す。なお、２つの語の間の関連性をもとに単語をあてはめるので、誤ってあてはまった語の関連語を用いてあてはめを行うと、エラーが拡大する。そこで、当

Use of Annexed Word in the "Key-word" Method Kana-Kanji Translation System
Kenzi ARAKI, Kouji UTIDA, Hirosi YAMADA, Koji TOCHINAI, Kuniiti NAGATA
HOKKAIDOU University

面エラー発生頻度の少ないキーワードについてのみ関連語情報を用いることにしている。

３．実　　験

本手法をシステムに組み込む前に、実際に関連語を正確に取り出せることを確かめる実験を行った。

３．１　実験方法

資料として、「データベース要論」（穂鷹良介著）の第１章（2458文字）を用いた。まず、最初に、資料の初めより８文(455文字)の中から24のルールを取り出し、次にそのルールを用いて、資料中から関連語の組を取り出し、その情報の正当性を評価した。

(a) [言語] のきげんとして (の) [様々] な [性格] @について@のべてある。

(b) [言語] (の) きげん として (の) [様々] な [性格] @について@のべてある。

@@：１種キーワード　[] ：２種キーワード
() ：文節端助詞候補

図２　関連語による単語のあてはめの例

３．２　実験結果

図３に、実験結果を示す。ここで、検出されるべき関連語組数とは、24のルールにより当然取り出されるべき関連語の組の個数であり、検出されなかった関連語組数とは、その中で取り出されなかった関連語の組の個数である。

４．考　　察

図３からわかるように、未検出率は0%であるから本ルールを用いた場合取り出されるべきものはすべて取り出している。しかし、誤検出率も9.6%に達し、誤ったものを１割程度取り出している。これは、ルールがないために修飾関係や係り受け関係が検出できず結果として誤った語の組を関連語の組として判断してしまったためである。したがって、ルールを整備すれば、ある程度解決できるものと思われる。

正しく検出された関連語組数(A)	141
誤って検出された関連語組数(B)	15
検出総数(A+B)	156
検出されなかった関連語組数(C)	0
検出されるべき関連語組数(A+C)	141
検出率=A/(A+C)	100.0%
未検出率=C/(A+C)	0.0%
正検出率=A/(A+B)	90.4%
誤検出率=B/(A+B)	9.6%

図３　　実　験　結　果

この実験では、文章中に存在する関連語の組のうち、ルールに適合するものはすべて取り出せることがわかった。さらに、ルールを整備して誤検出率を下げることにより、本手法を用いて単語のあてはめの精度を高めることが可能であると考えられる。

５．おわりに

本稿では、キーワード方式べた書き文かな漢字変換システムにおける一般単語のあてはめ段階での連接情報の欠如による未変換と連接情報の増加による誤変換を防ぐために、付属語を用いて単語をあらかじめ分類しておき、この情報をもとにして、一般単語のあてはめを行うことの可能性を示した。今後は、本手法を実際にシステムに組み込み、どの程度変換性能が向上されるかの実験を行う予定である。

＊参考文献

1) 荒木、鈴木、栃内、永田　　情報処理学会第29回(昭和59年後期)全国大会５Ｊ－５
2) 荒木、鈴木、中山、栃内、永田 昭和60年度電子通信学会総合全国大会　１４５５