自然言語処理 48－3
（1985. 3. 27）

ビジネス文のかな漢字変換における文節数最小法適用の評価

吉田 正行，三ッ矢 裕一，宮本 義昭
（日本ユニバック）

1 まえがき

日本語文の入力には音声入力，オンライン手書き入力，カナ漢字変換による入力などがある。この中でカナ漢字変換による入力が最もポピュラーな方法として普及している。これはカナ漢字変換では現在使用中のＪＩＳキーボードの端末装置がそのまま日本語文入力の端末として使用出来ること，入力に対する訓練を必要としないことなどにある。カナ漢字変換方式の中でもカナ文字の入力形態や変換アルゴリズムによって種々の方式が提唱されている。我々はＵＮＩＶＡＣシリーズ１１００のカナ漢字変換方式として九州大学の吉田研究室が提唱するべた書き（自由形式）のカナ文字列を対象とした「文節数最小法」を取り入れ「日本語文書システム／入力」（略称ＪＡＳＴＹ）を商品化した。文節数最小法とは一つの入力文に対して形態素解析を行ない，その解析結果が複数個存在する場合，文節数が少ない解析結果のほうが多い解析結果よりも日本文として正しい可能性が大きいと考えて文節数の少ない解析結果を優先する手法である。

福岡大学の吉村講師らによって入力文として「人生論」を使用し最長一致法と文節数最小法の比較実験が行なわれ，文節数最小法が有効であることが報告された[5]。ＪＡＳＴＹの使われ方は小説だけでなく論文や報告書など会社で書かれる全ての文書（以降ビジネス文と呼ぶ）が対象となる。そこで，文節数最小法がビジネス文に対してどの程度の変換率が得られるかを調査するため，大量のビジネス文を入力し変換実験を行なった。また，最小文節数を持つ候補の中のどれを最初に出力するかの優先順位を決めたがこれの妥当性をも評価した。

ビジネス文に対しては，第１候補文が正しく変換されていた割合が７４％，次候補文の中に正しい変換が含まれていた割合が９５％，また，第１候補文中正しく変換された文字数は９５．８％，次候補文の中に正しく変換された文字数は９９．３％であった。

2 ビジネス文

2．1 ビジネス文の種類

会社で使われる文書全てをビジネス文と呼ぶことにすると，ビジネス文の分類としては以下のものが考えられる。

a．本　：解説書，手引き書などマニュアル類
b．論文　：研究発表の論文
c．提案書　：新しい機械やシステムの導入を提案する書き物

d．議事録　：会議の議事録
e．案内書　：展示会や発表会の案内状など
f．報告書　：出張報告，各種催し物への参加報告など
g．法例集　：「不法行為による損害賠償請求」や「無権代理行為の追認」など法律に関する文
h．社則集　：就業規則，給与規定など

2．2　ビジネス文の特徴

ビジネス文も日本語であることには変わり無く，国文法に基づいた文法規則で書かれている。しかし，小説とビジネス文を比較してみるとその用途，目的の違いにより文の作り方に幾つか相違点がある。

〔1〕 敬語が使われる
提案書や法例集のように読み手に顧客を想定している文書では「お考え」や「御説明」の様に敬語や謙譲語が多く使われる。

〔2〕 専門用語が使われる
マニュアルや論文などでは，その内容または記述する分野によって特有の専門用語が使われる。

〔3〕 外来語が使われる
コンピュータ関連の文書のように最近の文書ではカタカナ書きされる外来語が多く使われる。

〔4〕 固有名詞が使われる
特に人名，会社名が多く使われる。

3　実　験

3．1　実験環境

カナ漢字変換の実験は，ＵＮＩＶＡＣシリーズ１１００を使用して行なった（図３－１）。

・一般用語辞書としては約８６０００語の自立語が登録されている。この辞書は九州芸工大の稲永講師が作成し，研究機関に対し公開しているデータ・ベースと同じものである。
・付属語辞書には約３１０語の付属語，形式名詞，補助用言が登録されている。
・接辞辞書には約５２０語の接頭語，接尾語および助数詞などが登録されている。派生語は使用者指向辞書へ登録したため，今回は助数詞のみを使用した。
・使用者指向辞書としてはビジネス文の種類によって各々別の辞書を作成した。使用者指向辞書には専門用語，派生語など一般用語辞書には存在しない用語を登録した。
・高頻度用語辞書は会話型で使用する場合に次候補選択で選んだ用語を登録する辞書であり，今回の実験をバッチ処理で行なったのでほとんど使用していない。
・ＪＡＳＴＹでは派生語や外来語は変換種別を指定することによって変換することが出来るが，本実験では指定入力は行なわずにべた書き入力とした。派生語や外来語など未登録語となる用語は新しく作成し使用者指向辞書へ登録した。

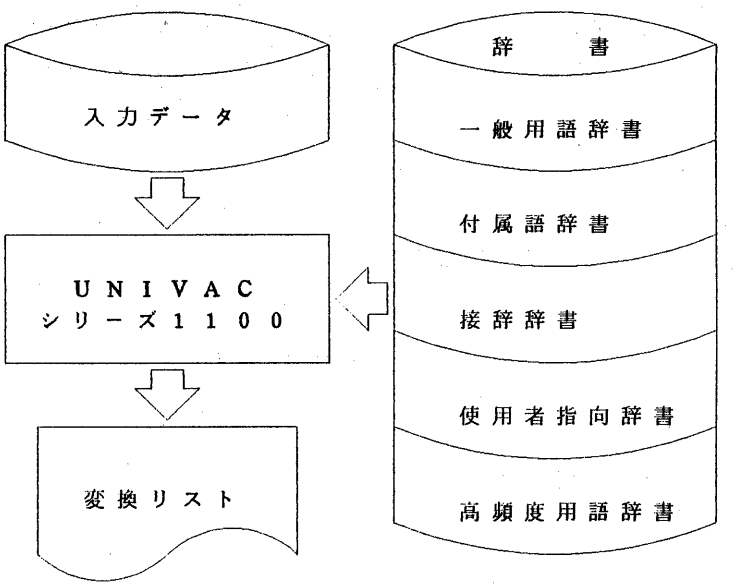

図３－１　実験環境

ただ，無変換（英数記号をそのまま漢字コードに変換する）と固有名詞（姓，名，企業名および地名）は指定入力を行なった。使用者指向辞書へ登録した用語の内訳は表３－１のとおりである

３．２　実験データ

実験には入力データとしてビジネス文の中から以下のデータを用意した。

・議事録：１０１３文
・本　　：１０５８文
・提案書：　７４５文
・法例集：　４４９文
・社則集：　６２５文

以上合計３８９０文を入力データとした。入力は１文当たり約２５字（読みの長さ）から構成されている。

３．３　最小文節候補内での順序付け

文節数最小法によって得られた変換結果が複数個存在する場合は次の優先順位で出力した。

	議事録	本	提案書	法例集	社則集	合 計
派生語	73	56	63	59	41	292
外来語	1	41	72	0	0	114
専門用語	12	18	8	7	3	48
固有名詞	9	0	0	2	0	11
その他	27	9	9	9	8	62
合計	122	124	152	77	52	527

表３－１　登録用語分類表

〔1〕 自立語を検索した辞書の順序

a．高頻度用語辞書

b．使用者指向辞書

c．一般用語辞書

一般的な辞書の用語よりも対象分野に依存した使用者辞書の用語を優先する。

〔2〕 自立語の読みの長さの長い順序

例．　データの型　＞　データのか田

ビジネス文において「指定入力」，「補助用言」など複合語が多く使われる。この様な複合語を優先して出力するためにこの順序付けを行なっている。

〔3〕 辞書中の同音語の優先順位

例．　今日　＞　経　＞　興　＞　供

この優先順位は全ての分野での使用頻度を考慮している。

３．４　実験結果

実験の結果を次の４種類に分けて集計した一覧表を表３－２に示す。

a．第一候補に正しい変換結果が出力された文の数。

b．次候補中に正しい変換結果が含まれる文の数。

c．文節数は同じであるが領域不足のため正しい変換結果が出力されない文の数。

d．次候補中にも正しい変換結果が含まれていない文の数。

（注）b＃，c＃，d＃はb，c，dの第一候補の中で正しくない単語の漢字文字数，＃＃は出力漢字文字総数

	議事録	本	提案書	法例集	社則集	合 計
a	625	817	632	308	495	2877
b	316	203	95	122	107	843
c	64	30	16	19	21	150
d	8	8	2	0	2	20
b#	962	576	242	314	269	2363
c#	190	84	39	55	53	421
d#	28	33	8	0	10	79
##	17324	18950	13162	9338	9806	68580

表３－２　変換結果集計表

変換結果の一部を次に示す。

〔1〕 議事録

入力　：ﾜﾉｾｲｼﾞﾆﾂｷﾏｼﾃﾊｺﾝｺﾞﾄﾓｺﾚﾖｹﾝｼﾞｼﾃﾏｲﾘﾀｲ，

第一候補：輪の政治につきましては今後ともこれを堅持して参りたい、

次候補　：和の政治につきましては今後ともこれを堅持して参りたい、
羽の政治につきましては今後ともこれを堅持して参りたい、
把の政治につきましては今後ともこれを堅持して参りたい、
輪の正字につきましては今後ともこれを堅持して参りたい、
・・・

入力　：ｿﾉﾖｳﾅｼｾｲﾃﾞｾｲｼﾞﾆﾄﾘｸﾝﾃﾞﾏｲﾘﾀｲ，

第一候補：其の様な姿勢で政治に取り組んで参りたい、

次候補　：其の様な市政で政治に取り組んで参りたい、
其の様な至誠で政治に取り組んで参りたい、
其の様な氏姓で政治に取り組んで参りたい、
其の様な市井で政治に取り組んで参りたい、
・・・

入力　：ｺﾉﾖｳﾆｶﾝｶﾞｴﾃｵﾙﾜｹﾃﾞｺﾞｻﾞｲﾏｽ．

第一候補：この様に考えて居る訳で御座います。

次候補　：この様に考えて折る訳で御座います。
此の様に考えて折る訳で御座います。
此の様に考えて居る訳で御座います。
・・・

〔2〕 本（第一候補のみ）

Cは汎用のプログラミング言語である。此の言語は従来，
ＵＮＩＸシステムと密接に関係するものとされてきているが，
此れはCが，ＵＮＩＸシステムの上で開発され，ＵＮＩＸと
其のソフトウェアがCで書かれているためである。
然し，此の言語は特定のオペレーティング・システム（ＯＳ）や
計算機と結び付いているものではない。
・・・

〔3〕 提案書（第一候補のみ）

貴行益々御清栄の段お慶び申し上げます。
この度は、
貴行新本部システムのご検討に当たり、
弊社に提案の機械を給わりましたこと誠に有り難く熱く御礼申し上げます。
早速貴行より御提示頂きました諸条件を、
永年に渡り積み重ねて参りました実績と経験を元に検討・
分析させて頂きましたが、
・・・

〔4〕 法例集（第一候補のみ）

売買契約の取り消しについて
・・・
ところが其の後調査の結果、其の様な都営地下鉄線の計画は現在無く、
貴社の御説明は全くの偽りであったことが判明致しました。
したがって、私はここに上記売買契約における
申し込みの意思表示を取り消し致します。
・・・

〔5〕 社則集（第一候補のみ）

第１１条
従業員は会社の許可を受けないで、
会社内で集会、
演説会を開き、
又は印刷物其の他の文書を掲示し、
若しくは配布し、
あるいは宣伝、
放送などの行為を行なってはならない。
・・・

正しい変換結果が得られなかった２０文を調べると以下の４つの原因に分類できた。

〔1〕 原文より文節数の少ない変換結果があった。（１６件）

原文	変換結果
やや｜異なる	稚子となる
できるだけ｜早く｜役に	できるだけは｜役々に
例は｜意図したほど	礼拝としたほど
引用符が｜無いと｜言う｜つれない	引用符がな｜厭い｜移れない
よく｜似ている	欲にている
実行｜可能｜文	実行家｜能文
長い｜使用	長居しよう
プログラムの｜破壊｜防止	プログラムのは｜解剖し

（注）文節の区切りは縦棒（｜）で示した。

〔2〕 自立語が省略されていた。（２件）

原文	変換結果
（そう）かといって	過渡一手
（それ）じゃ	蛇

〔3〕 自立語と接尾語の間に他の用語が挿入されていた。（１件）

原文：インプリメンテーション（実現）上の

変換結果：インプリメンテーション（実現）情の

〔4〕 付属語が接続しない。（１件）

それはどうなりましたですか

接続不可

4　評　価

カナ漢字変換の変換精度を入力文全体を一つの単位として評価する場合，評価式として次の式を使用する。

$$\text{正変換率（文）} = \frac{\text{第１候補が正しい変換の文数}}{\text{入力した総文数}} \times 100$$

$$\text{多変換率（文）} = \frac{\text{次候補中に正しい変換のある文数}}{\text{入力した総文数}} \times 100$$

$$\text{誤変換率（文）} = \frac{\text{次候補中にも正しい変換のない文数}}{\text{入力した総文数}} \times 100$$

実験で使用した３８９０文について評価すると次の通りである。

正変換率（文）＝２８７７／３８９０×１００＝７３．９６％

多変換率（文）＝８４３／３８９０×１００＝２１．６７％

誤変換率（文）＝１７０／３８９０×１００＝４．３７％

ただし，領域不足で正しい変換が次候補中にも得られなかった文も領域を拡張すれば変換されることは明白であるが，あえて誤り文数に含めた。文全体の評価としては，次候補中も含めて正しい変換結果が得られたのは，７３．９６＋２１．６７＝９５．６３％である。現在，世の中一般に使用されているカナ漢字変換の精度の評価では文全体の変換率ではなく文字の変換率である。この方式に従うと評価式として次の式を使用する。

$$\text{正変換率} = \frac{\text{正しく変換された字数}}{\text{入力文の総字数}} \times 100$$

$$\text{多変換率} = \frac{\text{同音異議語となった字数}}{\text{入力文の総字数}} \times 100$$

$$\text{誤変換率} = \frac{\text{誤って変換された字数}}{\text{入力文の総字数}} \times 100$$

この評価式で実験データを評価すると

正変換率＝６５７１７／６８５８０×１００＝９５．８２％

多変換率＝２３６３／６８５８０×１００＝３．４５％

誤変換率＝（４２１＋７９）／６８５８０×１００＝０．７３％

第１候補と次候補の中に正しい変換結果が含まれる文字の割合は，９５．８２＋３．４５＝９９．２７％と非常に高い精度が得られた。

5　おわりに

「文節数最小法」を使用したビジネス文の評価を4章に示した。第1候補で正解が得られる正変換率と次候補に正解が得られる多変換率を合わせると，９５．６３％（文字数を基準にすると９９．２７％）の高い変換率が得られ，文節数最小法のアルゴリズムがビジネス文に対しても有効であることが実証できた。また，最小文節候補内での順序付けは多変換率２１．６７％で分かるとおりまだ検討の余地が残されている。我々が行なった順序付けの中で「自立語を検索した辞書の順序」は良い結果が得られた。しかし「読みの長さの長い順」に関しては良い順序付けとは言い難い。「辞書中の同音語の優先順位」に関しては文書の分野によって順序が変わってくる。

今後の課題として以下の点が残されている。

〔1〕 正変換率の向上

カナ漢字変換の操作をする使用者にとって次候補の中から正しく変換された結果を選択する操作は煩わしいことである。現在，２１．６７％を占めている多変換をできる限り正変換へ持って行くことで次候補選択の余分な操作を減らすこと。

この作業として次の事項が考えられる。

１．係り受け構造解析

２．最小文節数候補内での順序付けの検討

〔2〕 操作性の向上

現在，派生語は入力時に使用者が指示を行なうか新しく用語を登録しておく必要がある。今後は自動分かち書き処理へ接辞処理を組み込み余分な指示は無くしていくこと。

最後に，「日本語文書システム／入力」の実現のために，ご指導下さった九州大学吉田将教授，日高達助教授，福岡大学吉村賢治講師，および当システムで使用した一般用語辞書を作成してこられた九州芸工大稲永紘之講師に深く感謝します。

［参考文献］

〔1〕 吉村賢治，日高達，吉田将，「表方式を用いた文節構造分析アルゴリズムとその能率について」，情報処理，計算言語学研資，２５－６，１９８０．
〔2〕 日高達，吉田将，「効率的日本語機械辞書」，情報処理学会第２４回全国大会論文集，５Ｇ－７，１９８２．
〔3〕 森健一，河田勉，「かな漢字変換」，情報処理，ＶＯＬ．２０，ＮＯ．１０，１９７９．
〔4〕 稲永紘之，吉田将，「日本語処理のための機械辞書」，情報処理，ＶＯＬ．２３，ＮＯ．２，１９８２．
〔5〕 吉村賢治，日高達，吉田将，「日本語文の形態素解析における最長一致法と文節数最小法について」，自然言語処理，３０－７，１９８２．