

かん字漢字変換の変換率について

建石由佳・金子宏・島原信一

日本アイ・ビー・エム(株)東京基礎研究所

1. はじめに

記入 助のこめ変力率 一
表ブ2) 换そるた誤入率を定
字イ1) の ゼるた れ漢を
漢タしたムえ文れっは 汉を
チレテ考をわが稿さ換の率
てッラスを宇失な本変中換
しタメシと漢がら 動力变
し 始力こに報こ5) 入て
力はが入るに報こ5) 入て
入」究語す入るはつ正し
を換研本用すにかがめに成
列変て日利が聞合わ入たを
字字しるを(4)に場が(るト
文漢とよ換3)みると率ベス
む手段に変た読すこ換調キル
含ん手続きの換る変をテア
をか力書漢篆字變ずてかたで
字「入手字構文ら生つるし告
漢る的はんをてかによれ戻報
にす助れかムなたに変にの
的力補わてテよか新記うみ果
分出のれしスに、が表ど説結
部を力わとシドに換のが部た

2. かん字漢字変換プログラム

は含てて約は変沢ストムをる。象記る字選コラ字あ持對表あ漢捕るグ漢でての力でな候すロをのし索い入)か。対ブ書もと索語る倍節るに換書もと考るの書き1文い語を変のしキ辞書で、連てるで字ム張索辞と5。いあ定漢ラ拵接な。象約は用一。字グにをうい対のムまですん口う法よな索数ズまのらかブよ記るは検語りのもよた換る表い限。(ゴそるにし変きのて制りリルをよ記現字でれるあおアムに表実漢索べわすとのズ法力のな検す行対語千外り小入れかではに千4以ゴ最のわ節一書2)記7万索ルトそれ文キ辞、表万9検アス、わ連だのりる73書のコは

3. 変換率測定の方法

(れ確、用ある
文かのし驗で
2書)成實の
2まゝ作。も
1まーをたな
3の1のせう
たそ(もさ
つが)をた換の
取字し交次
ら漢漢房節は
か。に文の
どた中み連た
ない文説して
事用原なし意
記を、然括用
聞)て自一て
新文しにし
はノとムとと
に字。ダニト
定3をン文ス
測2率ラをキ

A. $p = 0$, すなわち記事の漢字部をすべて（その読みに対応する）ひらがなに直したものを

B. $p = 1/4, 1/2, 3/4$ のテキストをそれぞれ5つずつ

C. $p = 1$. すなわち記事の表記そのままのもの

手かグロは、かロブの
はくブ換る
の書識変す
た接認。測
し直字で予
し漢が文のを
に一。るさ
ムザ度す
ダ一頻存や
ンユ。依れ。
ラを数に現る
に字画のある
き漢。も字で
どるはな漢ら
すあかまのか
戻。るざ中た
にはすま力え
みで換さ入考
読力変どがと
を入らぬまだ
字きかムラ難
漢書なラグ困

4. 結果

文を単表1位のに変換率（正しく表記の換換の数／3122）して示す。なお、たれられた文は正解としている。

変換率の高さはよくなるようになる。この表2に示すとおり、漢字が多くなればなるほど、漢字の関係をプロットすると、必ず原点を通る直線となる。

表4でこの記入は、そのもしさりて味か通い意事除をる記をとな聞れこに新搖るうの出よは記がの率表り次換誤と變も換るのて變すきしの類とと度分るのた程を1しトらでをシレ数ロカセニ

a. かな書きされた語を正しくない漢字に変換
(63)

例：国道をふさいでいる
→ 国道を夫婦がいた
灯油をまいだをいた
→ 灯油を巻き始めた

b. 語が辞書にない
(38)

例：エジプトと復交した
→ エジプトと[未知語]下
それがびたつと相手の体に届いて
→ それが[未知語]と相手の体に

(この区間で [未知語] とはシステムが未知語と推定したを示す)

c. 漢字書きされた語を前後の語
かなか漢字交じりで表記されたもの

d. 文語表現または口語表現
(11)
例：私はタケウチでーす
→私はタケウチデー

こ以語の推す換起分知上とに変も部未、語とくで】、る知し換語個わ未る正変知6か、す上字未3がし力け漢【が果も出かなはの結、をみかでも換たまはは、な变つま個漢【あうのあの6e、よ分でそ3d、だ例の2表すd、たのが力よb、下ほの入のa、あいめな分のうり正るよた、の誤はあのしばこる外が例定れ

されているようになる。

c. のような誤変換の数とそのあいだには強い相間がある。対して異なるものでは、漢字書籍がそのまま読み取られることが多い。しかし、漢字書籍がそのまま読み取れない場合は、漢字書籍がそのまま読み取れない場合がある。これは、漢字書籍がそのまま読み取れない場合がある。これは、漢字書籍がそのまま読み取れない場合がある。

部分的に読み戻した($0 < p < 1$)テキストでは、不完全な漢字書籍を変換するために、別の音の、もともと対立しない語に変換されてしまうケースがある。下の例のようないふたつの数を数えると表2のようになる。

アル・オマヤかい上き地
アル・オマヤ買上生地
(正: アル・オマヤ海上基地)

喪主は妻し保さん
喪主は妻子保さん
(正: 喪主は妻志保さん)

かい物をして
怪物をして
(正: 買い物)

p とこのような誤変換の数とのあいだには強い相間はみられなかった。

5.まとめ

入力記号の中には漢字書籍がそのまま読み取れない場合がある。これは、漢字書籍がそのまま読み取れない場合がある。これは、漢字書籍がそのまま読み取れない場合がある。

今回のはたたかれた割合は、漢字書籍がそのまま読み取れない場合がある。これは、漢字書籍がそのまま読み取れない場合がある。

- 1) 塩見・喜多・河合・大岩: 「2ストローク入力のための仮名漢字変換」, 情報処理学会論文誌 Vol. 33, No. 7, pp. 920-928 (1992).
- 2) 小野: 「Tコードの補助入力: 字形組み合わせ法と文書書き換え法」, 情報処理学会論文誌 Vol. 31, No. 3, pp. 405-413 (1990).
- 3) 金子・鳥原・達石: 「表記変換つきの形態素解析プログラムとその応用」, 情報処理学会第45回全国大会4C-4(論文集3, pp. 255-256).
- 4) 日経BP社: 「日本IBMが技術展を開催--ベンチマークなど展示」, 日経エレクトロニクス No. 557 (1992).
- 5) 達石・金子・鳥原: 「連文節かん字漢字変換」, 情報処理学会第45回全国大会4C-1(論文集3, pp. 249-250).

p	0	1/4	1/2	3/4	1
		0.826	0.882	0.930	
		0.827	0.866	0.892	
	0.802	0.826	0.892	0.928	0.958
		0.837	0.890	0.923	
		0.820	0.878	0.927	
平均	0.802	0.827	0.881	0.927	0.958

表1 変換率測定結果

p	0	1/4	1/2	3/4	1
		10	24	15	
		9	14	16	
	0	18	21	16	16
		8	11	18	
		11	15	19	
平均	0	11.2	17.0	16.8	16

表2 異音語への変換の起きた数

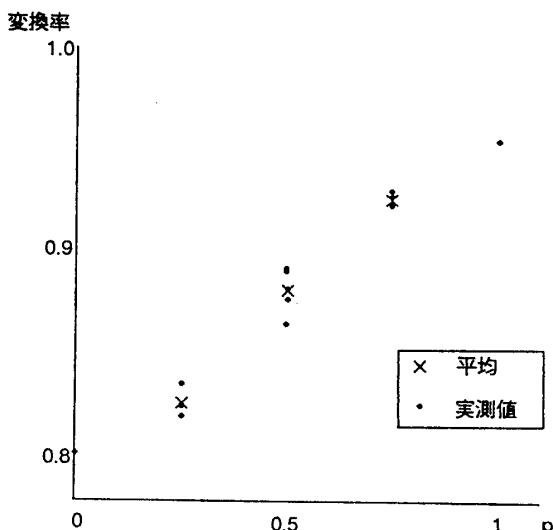


図1 変換率測定結果