

E-041

携帯電話での文字入力における誤り自動訂正手法の性能評価

Performance Evaluation of Automatic Correction of Erroneous Text Input on Mobile Phone

菊地 直樹[†] 松原 雅文[†] Goutam Chakraborty[†] 馬淵 浩司[†]
 Naoki Kikuchi Masafumi Matsuhara Goutam Chakraborty Hiroshi Mabuchi

1. はじめに

近年、携帯電話をはじめとした携帯端末の普及とメール機能、ブラウザ性能の向上によって、携帯端末での文字入力の機会と必要性は増大している。携帯端末は小型であるため、常に持ち歩いての使用が可能であり、文字入力の状況も歩きながらやテレビを見ながらなど、他の動作と並行して行われることが想定される。そのような場合、ユーザは入力するために画面を終始注視しているわけではない。そのため誤入力が増え、訂正のために打鍵数が増加し、迅速な入力が困難になるという問題が発生する。

そこで本研究では、この問題を解決するために、携帯電話での文字入力における誤り自動訂正手法を提案する。この手法では、誤りを含む入力文字列に対して、かな漢字変換を行った際に、システムにより自動的に誤り訂正を含む変換候補が生成される。そして、出力された変換候補の中から、正しい変換結果をユーザが選択することにより処理が完了する。これにより、誤入力の訂正にかかる手間を軽減することができる。

本稿では、本手法の概要を示し、実際の文字入力データから得られた、誤りを含む入力文字列に対してかな漢字変換を行った結果から、本手法の有効性を述べる。

2. 誤入力について

2.1 携帯電話での日本語入力

現在、携帯電話で一般的に利用されている日本語入力方式として、文字循環指定方式が挙げられる。この方式では、携帯電話に実装されているキー数の少なさを補うため、1つのキーに複数のかな文字が割り当てられている。従って、1文字を入力するために同じキーを複数回打鍵する必要がある。そのため、入力のために必要な打鍵数が多くなり、誤入力が起こりやすくなってしまふ。

2.2 対象とする誤入力

一般的な誤入力には、誤字、脱字、文字区切り誤りなどが挙げられるが [1]、携帯電話での誤入力を考えた場合には、文字循環指定方式特有の誤りに着目することが必要であると考えられる。そこで、この方式で最も多く発生する誤りである、同一キー打鍵数誤差±1の誤りを本手法での訂正対象とする。

同一キー打鍵数誤差±1の誤りとは、例えば「そうじ」と入力しようとして「せうじ」や「そえじ」というように、キーを打鍵する回数を1回多く、または少なく打鍵してしまった場合の誤りである。なお、本手法では、入力文字列中の1文字を誤った場合の誤入力を対象としている。

[†]岩手県立大学大学院 ソフトウェア情報学研究所

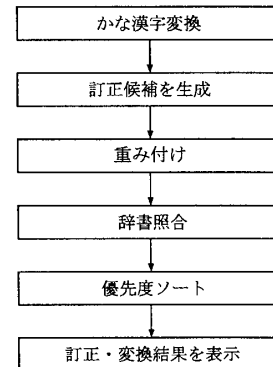


図 1: 提案手法の流れ

3. 提案手法

3.1 概要

本手法における処理の流れを図 1 に示す。ユーザが入力した文字列へのかな漢字変換の際に、システムにより入力文字列に対しての訂正候補かな文字列が自動的に生成される。次に、生成された訂正候補に対し、入力傾向に基づいた数値を利用して重み付けが行われる。その後、訂正候補かな文字列に対して辞書照合を行い、候補を漢字に変換する。そして、変換が行われた候補を、先ほど与えられた重み付けを用いて、ユーザが意図していた可能性の高い順にソートし、変換結果として表示する。最終的に、表示された変換候補一覧の中からユーザが候補を選択することで誤入力の訂正を完了する。

3.2 誤り訂正候補生成

入力された文字列から、まず、同一キー打鍵誤差±1の誤りに基づき訂正候補の生成を行う。文字数が n 文字であった場合、生成される候補数は基本的に $2n + 1$ 通りである。例えば「すうじ」という3文字の文字列が入力された場合、「しうじ」、「せうじ」、「すいじ」、「すえじ」、「すうざ」、「すうず」、「すうじ」という7個の候補文字列が生成される。

3.3 重み付け

生成された候補に対して、携帯電話での文字入力傾向に基づいて重み付けを行う。重み付けには誤入力傾向と正入力傾向を用いる。正入力傾向を c 、誤入力傾向を e とすると、重み w の値は $w = \alpha c - \beta e$ の式で表される。ここで α 、 β は定数である。

3.4 辞書照合

生成された候補に対して辞書照合を行い、候補を漢字へと変換する。照合の際には SKK 辞書を用いる [2]。この辞書は、あらかじめ言語的に現れやすい候補が上位にくるように語が収録されている。さらに、活用する語の送り仮名の開始位置の情報を持っている。そのため、生

表 1: か行:誤入力の種類例

誤りの種類	1→2	2→1	2→3	3→2	3→4
重み	1	1	0	7	8
誤りの種類	4→3	4→5	5→4	5→6	-
重み	2	5	2	14	-

表 2: か行:正入力の種類例

打鍵数	1	2	3	4	5
重み	193	86	144	65	70

成された候補に対してそのまま辞書照合を行うことが可能である。

3.5 優先度ソート

辞書照合によって漢字へと変換された候補に対して、与えられた重み付けの値を用いて優先度を決定し、優先度の高い順にソートを行い、変換候補を表示する。

その後、ユーザによって候補から正解である単語が選択され、処理が完了となる。

4. 評価実験

4.1 実験方法

本手法の有効性を確認するため、実際に誤りを含む入力文字列に対してかな漢字変換を行った。実験には、日本語約 6,000 文字の論文データ [3] より得られた誤り傾向の数値と、正しい入力傾向の数値を利用し、入力データ中に出現した誤入力単語 100 個、正入力単語 100 個の合計 200 個に対して処理を行った。なお、入力傾向取得のためのデータは、実際に第一著者によって入力されたものである。

評価には、システムにより生成された変換候補の中で、正解の文字列が何位に出現しているかという情報を使用する。正解文字列が候補の先頭に出現したものを 1 位と分類し、正解文字列が先頭以外に出現したものをまとめて 2 位以下と分類する。

4.2 重み付け

誤り訂正候補の重み付けには、実際の携帯電話での文字入力データから得られた入力傾向を用いる。誤入力傾向は、同一キー打鍵の際に何打鍵から何打鍵へ誤ったかという情報に基づき、誤りの種類を 10 通りに分類し、さらにその分類を「あ行」から「わ行」までの各行ごとに分けた件数であり、今回は実験データより得られた 227 件の誤り件数を用いる。表 1 に、か行の誤り分類の例を示す。例えば、「こ」を「け」と誤って入力した場合、「か行:5→4」という誤りに分類される。

正入力傾向は、入力データ中で正しく入力された文字を分類した件数である。ここでは、入力データ約 6,000 文字から、誤入力傾向の 227 件を除いたものを正入力傾向として利用する。表 2 に、か行の正入力の分類例を示す。上記の例の「こ」に対して判定を行う場合、70 という数値が与えられる。

上記の誤入力傾向と正入力傾向を基に重みの値を決定する。重み w の値は $w = \alpha c - \beta e$ で示される。今回の実験では、 α , β の値を $\alpha = 1.8$, $\beta = 0.8$ と設定した。

表 3: 評価実験結果

順位	誤入力		正入力		合計	
	件数	精度%	件数	精度%	件数	精度%
1 位	94	94.0	92	92.0	186	93.0
2 位以下	6	6.0	8	8.0	14	7.0

4.3 実験結果および考察

実験結果を表 3 に示す。処理を行った結果、誤入力 94 件、正入力 92 件の単語で、正解単語が訂正候補の 1 位に出力された。また、訂正後の正解単語が候補の 2 位以下に出力された単語が 6 個確認され、正入力でも、順位が 2 位以下で出力された例が 8 件確認された。これら 2 位以下の候補の平均順位は 2.9 位であった。このように、正解の候補が 1 位ではない場合でも、比較的上位に候補が出現している。そのため、実用時を想定した場合、順位の低下による入力の手間の極端な増加はないものと考えられる。しかし、今後研究を進める上で、システムの精度に大きな影響を与えることも予想される。その場合には、重み付けを与える際に使用している α , β の数値を検討し、最適な値を使用する必要がある。

以上のように、合計で 93% の精度が得られ、本手法の有効性を示すことができた。

5. おわりに

本稿では、携帯電話での文字入力における誤り自動訂正手法を提案し、実験によってその性能評価を行った。本手法では、文字循環指定方式での入力傾向を利用して、誤り訂正を含む変換候補を生成し、ソートを行うことでユーザの意図していた順に出力する。これにより誤入力訂正にかかる手間を軽減することが可能である。また、重み付けの際に誤入力傾向と正入力傾向を利用することで、入力文字列が正しい場合、誤っている場合の両方に対して、区別することなく処理を行うことが可能となっている。

評価実験として、誤りを含む入力文字列に対してかな漢字変換処理を行い、全体で 93.0% の精度が得られ、本手法の有効性が示された。

今後は、より多くの文字入力データによって実験を行い、さらなる精度向上を目指す。また、今回実験で使用したデータはクローズドデータであるため、今後オープンデータでの実験も行う予定である。さらに、候補の出現順位による評価のみではなく、本手法を使用した際の打鍵数による評価も行っていく予定である。

参考文献

- [1] 田中 穂積, “自然言語処理 - 基礎と応用 -”, 電子情報通信学会, 1999.
- [2] 佐藤 雅彦, “Simple Kana to Kanji conversion program”, <http://openlab.jp/skk/index-j.html>.
- [3] 市村 由実, 齋藤 佳美, 木村 和広, 平川 秀樹, “入力予測機能を組み込んだ仮名漢字変換システム” 電子情報通信学会論文誌 D-II, Vol.J85-D-II, No.12, pp.1853-1863, December 2002.