ローマ字入力における誤りの修復

5B－1

今村 剛　中西 正和
慶応義塾大学理工学部数理科学科

1. はじめに

我々は、ワープロなどの道具を利用して文書を作成する際に、ミスタイプのために再入力を強いられることがかなりある。本研究ではこの点に注目し、実際にどの様なタイプのミスが生じているのかを調査すると共に、その原因や傾向（誤り特性）を分析する。また、その結果を誤りの修復に利用することで、どの程度処理が効率化されるのかを実験する。

2. 誤りデータの収集

2.1 収集方法

文献 [1] から文章を抜粋し、句読点などの区切り記号を全て取り除いたものを入力テキストとした。このテキストを被験者に提示し、QWERTY 配列のキーボードからローマ字で入力してもらう。この際、1 度入力した文字は修復できないようにした。また、ミスタイプを犯してもごく自然な入力となるように、自分がタイプした文字は出力しないことにした。こうして入力されたテキストから、目視にてタイプミスを抽出した。

2.2 収集結果

タイプミスを、次の代表的な誤り操作とその他に分類した [2]。

- 1 文字の置換（letter → litter）
- 1 文字の挿入（letter → lettear）
- 1 文字の削除（letter → leter）
- 隣接文字の互換（letter → eltter）

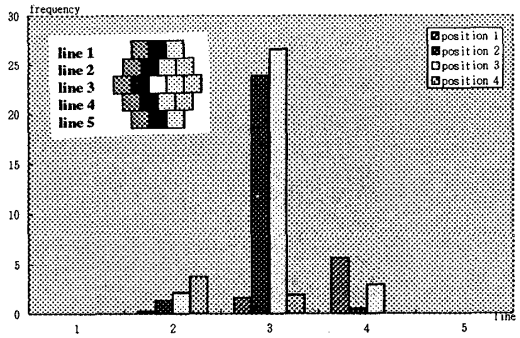

図 1: 相対位置ごとの頻度（置換）

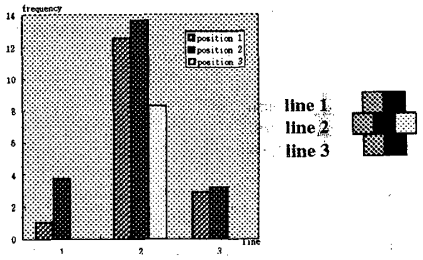

図 2: 相対位置ごとの頻度（挿入）

被験者 10 人によって入力された合計文字数は 37228 で、この中に合計 656 の誤りがあった。その内訳は、一文字の置換が 313、一文字の挿入が 169、一文字の削除が 127、隣接文字の互換が 20、その他が 27 であった。

2.3 誤りデータの分析

一文字の置換 各キーに対してスコープおよび相対位置を決め、その位置別に文字 10000 字当たりに生じる誤りの頻度を計算したものが図 1 である。この図から、置換の対象が左右に隣接したキーに集中していることがわかる。

一文字の挿入 挿入されたキーに対して前後のキーがスコープ内にどの様に分布しているのかを、文字 10000 字当たりに生じる誤りについて計算したものが図 2 である。この図から、挿入されたキーと同じキーや左右に隣接したキーが、前後によく現れていることがわかる。

A Study for Automatically Correcting Romanized Japanese Words
Takeshi IMAMURA, Masakazu NAKANISHI
Department of Mathematics, Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa Pref., 223, Japan

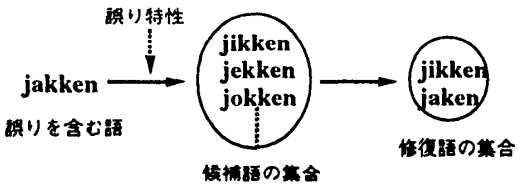

図 3: 処理の流れ

一文字の削除 'n' の削除される確率は 1.968×10^{-2} で、他に比べて非常に高くなっている。'n' が削除された時にはその前後に極端に 'n' が出現していることから、'n' を複数回押す必要がある際に誤りが生じ易いことがわかる。

隣接文字の互換 20 の誤りのうち、左手と右手の組み合わせによる誤りが 16 であることから、この誤りはキーの位置に強く依存していることがわかる。

3. データに基づく修復実験

3.1 実験方法

誤りを修復する方法の 1 つとして、一致法がある。これは、誤りを含む語から可能な語を推定することによって、修復語（基本的には辞書に存在する語）を得る方法である。推定は、誤りが先の 4 種の誤り操作のいずれかによることに着目し、誤り操作の逆操作を施すことによって行う。

単純な一致法では常に一定数の候補語を作成するため、辞書の規模が大きくなるにつれて探索時間の増加が予想される。また、修復率の低下も予想される。そこで誤り特性を利用し、無駄な候補語の作成を抑えることにする（図 3)。

誤りを修復する上で辞書が必要となってくるが、ここでは Wnn 4 で利用されている基本語 22709 語を採用することにした。

3.2 実験結果

誤り特性をもとに実験的に誤りを含む語（未知語）を 50000 語作成し、それを修復する実験を行った。

候補語数 誤り特性を利用することで、作成される候補語は語長ごとで 64.5 〜 73.0 %、全体として 67.6 % に絞り込まれている（図 4)。

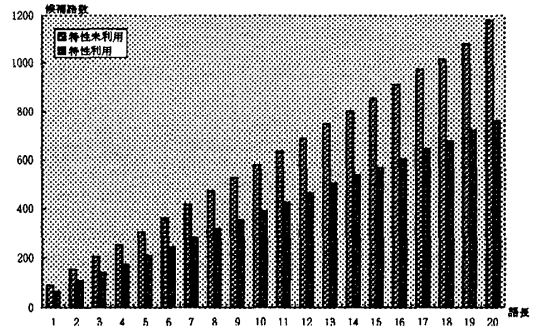

図 4: 語長ごとの候補語数の比較

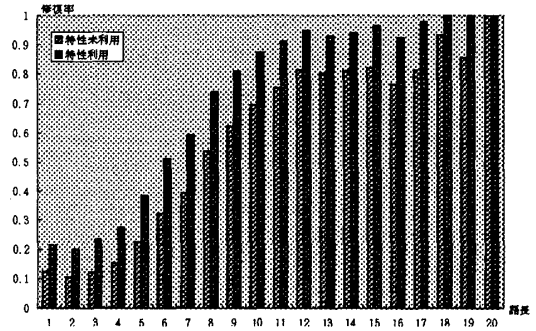

図 5: 語長ごとの修復率の比較

修復率 語長によってかなり変化しており、語長が長いものほど値が高くなっている。両者を比較すると、語長の短いものほど誤り特性の効果が顕著に現れており、全体としては 0.373 から 0.570 へと約 50 % の向上につながっている（図 5)。

4. おわりに

誤り特性の利用は、候補語の絞り込みおよび修復率の向上に対して効果的であることを確認した。今後は、誤り特性の利用を前提として、タイプミスを自動的に修復するシステムを実装する予定である。

参考文献

[1] 松本人志. 遺書. 朝日新聞社, 1994.

[2] 川合慧. 英文綴り検査法. 情報処理学会論文誌, Vol. 24, No. 4, pp. 507–513, 1983.

[3] 野田雄三. 誤打鍵特性の調査と分析. 情報処理学会第 43 回全国大会 1W-3, pp. 217–218, 1993.

[4] 野田雄三. 未知語の復元. 自然言語処理研究報告, Vol. 99, No. 9, pp. 63–70, 1994.