

m重マルコフモデルによる日本語の誤字、脱落及び挿入誤りの検出法

2L-8

荒木 哲郎⁺ 池原 悟⁺⁺ 塚原信幸⁺⁺ 福井大学工学部⁺⁺ NTT情報通信網研究所

1. はじめに

漢字OCRやWP(ワードプロセッサ)さらには音声認識装置などの入力装置を用いて計算機入力を行った日本語文には、一般に誤字、脱落及び挿入誤りが含まれるために、これらの誤りを自動的に検出し正しい日本語文に訂正する技術が必要となる。

これまでに、日本語の誤字を対象に単語解析プログラムを用いた誤字検出法並びに1重マルコフモデルによる訂正方法[1]がありまた、日本語文節内の連続した脱落、挿入誤りに対して、m重マルコフ連鎖確率を用いて誤り位置の検出並びに正しい日本語文に訂正するアルゴリズムが提案されている[2]。

本論文では、[2]について更に、誤字に対しても誤り位置の検出並びに訂正が行えるように、そのアルゴリズムを拡張する方法を示す。さらにその有効性を確認するために、1文字並びに2文字の置換誤りを埋め込んだ新聞記事400文節を用いて、文節内の誤り位置を検出し、訂正する実験を行う。

2. 誤り位置の検出と訂正方法

漢字かな交じり文節内の文字間の結合力は、一般に誤字、誤挿入または脱落の文字列がある場合には、正解文字列の場合に比べて弱くなる性質があるので、以下の仮説を設ける。

[仮説] 漢字かな交じり文節(一般には文)内に誤字、脱落または誤挿入の文字列が存在するときには、m重マルコフ連鎖確率が一定区間だけ連続してあるしきい値以下の値をとる。(仮説終)

A method for detecting characters wrongly substituted, deleted and inserted in Japanese "bunsetu"

Tetuo Araki⁺ Satoru Ikehara⁺⁺ Nobuyuki Tukahara⁺

⁺ Faculty of Engineering Fukui University

⁺⁺ NTT Network information Systems Laboratories

この仮説が成り立てば、誤りのタイプ(脱落、誤挿入または誤字)とその誤りの文字列長が存在する位置を決定する手順が次のように決まる。

(1) 誤り位置の検出法

文節内に誤字、脱落及び挿入誤りが存在する場合、誤り位置の前後において一定回数だけマルコフ連鎖確率が減少する。このときのマルコフ連鎖確率の減少回数を調べることにより誤りタイプ並びに誤り位置を識別する。(表1)

(2) 誤り文字の訂正法

(1)において誤り位置が検出された後、(i)脱落誤りならば、誤り位置に任意な文字を挿入しマルコフ連鎖確率値の改善(しきい値より高くなる)が見られれば、その中で最も高い文字候補を挿入した文を正しい文とする。また、(ii)挿入及び誤字誤りならば、最初に挿入誤りと仮定して誤り位置の文字を取り除きマルコフ連鎖確率値の改善性を調べ、連鎖確率が最も高くなる時の文字を挿入誤りの文字候補とする。次に、(iii)誤字と仮定して誤り位置の文字を任意な文字と置き換え、マルコフ連鎖確率値の改善性を調べる。このとき連鎖確率値の最も高くなる文字を誤字候補とする。最後に、(iv)挿入文字候補と誤字候補のうち連鎖確率値の高い方をとる。すなわち、もし置換文字候補の方が連鎖確率が高ければ置換誤りとし誤り位置の文字を置換文字候補と置き換えることで正しい文に訂正する。(表2)

3. 実験

(1) 実験の実施条件

- ①日本語文の種類 : 新聞記事のべた書き漢字かな交じり文節
- ②誤りの個数 : 1文節当たり1箇所、1~2まで連続した誤字、脱落あるいは挿入誤りを設定
- ③日本語文の個数 : 新聞記事データから任意に200文節を使用

④マルコフ連鎖確率辞書 : 新聞記事77日分の2重マルコフ連鎖確率辞書

(2) 実験結果

①誤字誤りの検出結果

2. に示された誤り検出法に基づいて誤りタイプが既知の場合の誤り位置を検出した結果は図1の通りである。同図より、1文字誤字で再現率93%、適合率100%、2文字誤字で再現率88.5%適合率100%となり、挿入と比べて若干悪くなるが脱落に比べて非常に高い精度で検出可能であることがわかる。

②誤りタイプが未知(誤字、脱落及び挿入誤りが混在する)の場合の検出結果

2. の方法により誤りタイプが未知の場合の検出結果は図2の通りである。同図より、1文字誤字で再現率93%、適合率93.5%、2文字誤字で再現率88.5%、適合率88.5%となり、誤りタイプが既知の場合に比べて適合率が0~15%程度低くなるものの高い精度で検出可能であることがわかる。

4. おわりに

本論文では、m重マルコフモデルによる誤字、脱落及び挿入誤りの検出並びに訂正手順を示し、その有効性を実験により確認した。その結果、誤りタイプが既知の場合、1文字置換では再現率93%、適合率100%、2文字置換では再現率88.5%、適合率100%であり、誤りタイプが未知の場合、1文字置換で再現率93%、適合率93.5%、2文字置換で再現率88.5%、適合率88.5%という良好な結果を得た。

今後は更に、文節境界、単語境界など誤りの位置による誤りの検出、訂正能力等について詳細に研究していく予定である。

[文献]

- [1]池原、臼井：「単語解析プログラムによる日本語誤字の自動検出と二次マルコフモデルによる訂正候補の抽出」、情報処理論文誌、Vol. 25、No2 pp298-305(1984)
- [2]塚原、荒木、池原：「べた書き漢字かな文における脱落・挿入誤り訂正」、電子情報通信学会春季大会講演論文集、H5. 3、No6 pp105

表1 べた書き日本語文における脱落/挿入/置換誤りの検出方法

	1重マルコフ	2重マルコフ	
脱落	1文字脱落	連続1回落ち込み	連続2回落ち込み
	2文字脱落	連続1回落ち込み	連続2回落ち込み
	n文字脱落	連続1回落ち込み	連続2回落ち込み
挿入	1文字挿入	連続2回落ち込み	連続3回落ち込み
	2文字挿入	連続3回落ち込み	連続4回落ち込み
	n文字挿入	連続(n+1)回落ち込み	連続(n+2)回落ち込み
置換	1文字置換	連続2回落ち込み	連続3回落ち込み
	2文字置換	連続3回落ち込み	連続4回落ち込み
	n文字置換	連続(n+1)回落ち込み	連続(n+2)回落ち込み

表2 べた書き日本語文における脱落/挿入/置換誤りの訂正方法

	1重マルコフ	2重マルコフ	
脱落	1文字脱落	連続2回改修	連続3回改修
	2文字脱落	連続3回改修	連続4回改修
	n文字脱落	連続(n+1)回改修	連続(n+2)回改修
挿入	1文字挿入	連続1回改修	連続2回改修
	2文字挿入	連続1回改修	連続2回改修
	n文字挿入	連続1回改修	連続2回改修
置換	1文字置換	連続2回改修	連続3回改修
	2文字置換	連続3回改修	連続4回改修
	n文字置換	連続(n+1)回改修	連続(n+2)回改修

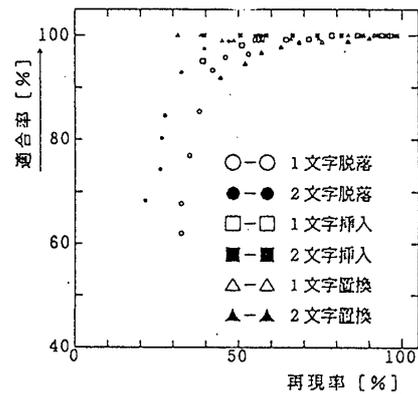


図1 誤り位置の検出結果(誤りタイプ既知)

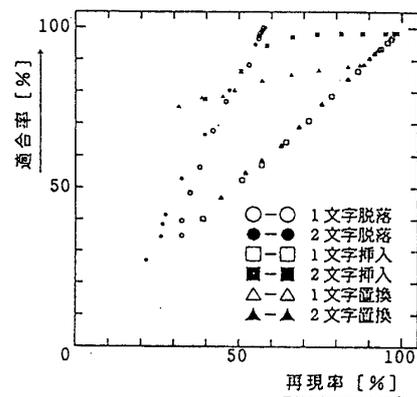


図2 誤り位置の検出結果(誤りタイプ未知)