文の先頭・末尾位置を考慮したべた書き日本語文の検出・訂正効果

2Ｈ－5　荒木哲郎+　池原　悟++　小松康則+
+福井大学工学部　++ＮＴＴコミュニケーション科学研究所

１．はじめに

コンピュータのマン・マシン・インターフェイスの改善のため、光学式文字読み取り装置（ＯＣＲ）や音声認識装置などの入力装置が開発されてきた。しかしながら、日本語文は非常に多くの文字、とりわけ数千種に上る漢字によって記述されるため、これらの装置を用いて入力することは容易ではなく、一般に誤った文字列が含まれる。

これらの誤りを見つけ、訂正する自然言語処理技術が期待されているが、現在の自然言語解析技術は正しい文に対して開発されているため、このような問題に直接適用することができない。

今日までに、この問題に対し、マルコフモデルを用いた統計的なアプローチがなされている。日本語漢字かな混じり文節における置換誤り、脱落誤りおよび挿入誤りの３タイプの誤りを判別し、これらの誤りをm重マルコフ連鎖モデルを用いて訂正する、選択可能な誤り訂正方法が提案され、ランダムに設定された誤りの検出・訂正に有効であることが示されている[1][2]。しかし、マルコフモデルを用いた日本語文節誤りの検出手法は、ＦＡＸ－ＯＣＲ複合誤りに対して、その先頭・末尾文字誤りにおける検出率が低いことが問題となっていた。

本研究では、これらの先頭・末尾文字誤りに対し、マルコフモデルのタイプ（順方向、逆方向）を組み合わせた手法を提案し、誤り文字検出実験による結果を評価し、考察する。

Effect of Detecting and Correcting
the Erroneous Strings located
at the top and tail positions
of Non-Segmented Japanese Sentences
Tetsuo ARAKI+　Satoru IKEHARA++
Yasunori KOMATSU+
+Fukui University
++NTT Communication Science Laboratories

２．２重マルコフモデルを用いた誤り文字検出方法

文節内に置換、脱落および挿入誤りが存在する場合、マルコフ連鎖確率値が一定区間だけ減少する。この時、マルコフ連鎖確率値が「しきい値Ｔ」を下回る回数を調べることにより、誤りタイプの判別が可能である[2]。

２重マルコフ連鎖モデルによる誤り文字の検出方法の説明図を図１に示す。位置X_2およびX_3に関わるマルコフ連鎖確率値が「しきい値Ｔ」より下回ることから、誤り文字を検出する。

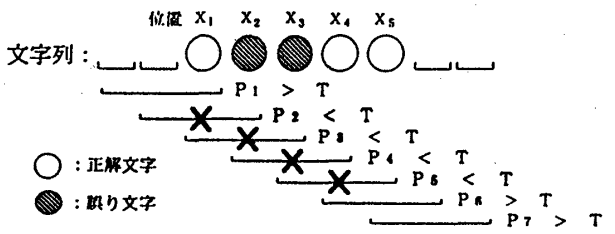

図１　マルコフ連鎖モデルによる誤り文字の検出方法

３．ＦＡＸ－ＯＣＲ複合誤りを含む日本語文節に対する実験結果

３．１　実験条件

１．統計情報を得るために使用した文節数：
283,963（日本語新聞記事70日分）
２．マルコフ連鎖確率辞書（２重）の学習方向：
順方向、逆方向、中間方向
３．使用した文節数：1000文節
（ａ）平均文節長（漢字かな混じり文節）：６
（ｂ）文字フォントの大きさ：１０ポイント
（ｃ）ＯＣＲへの入力方法：ファックス

３．２　マルコフモデルのタイプの組み合わせに関する実験結果および考察

誤り検出の精密さは、マルコフ連鎖確率の足切り値Ｔに依存している。「適合率Ｐ」および「再現率Ｒ」は、Ｔの値を変化させることで得られる。

各マルコフモデルと、ＦＡＸ－ＯＣＲ複合誤りの誤り位置ごとのデータの組み合わせによる、適合率Ｐと再現率Ｒの値を表１に示す。

［１］単独のマルコフモデルでの誤り検出

順方向に学習されたマルコフモデルでは、先頭・内部文字に対する検出率が高く、末尾文字に対する検出率は低い。逆方向に学習されたマルコフモデルでは、先頭文字に対する検出率は低いが、内部・末尾文字に対する検出率が高い。

［２］複数のタイプのマルコフモデルの組み合わせによる誤り検出

順方向と逆方向を組み合わせたものでは、順方向の先頭文字に対する特徴と、逆方向の内部・末尾文字に対する特徴とが顕著に現れ、全体の検出率を向上させている。

各しきい値に対する、順方向マルコフモデル単独の場合と、順方向と逆方向を組み合わせた場合の適合率・再現率の変化を図２に示す。

図２より、適合率および再現率がしきい値１１（対数値で表す）でピークをもっていることに注意する。

実験結果から、順方向と逆方向のマルコフモデルの積の条件で誤り検出を行う手法が、従来の順方向単独での場合より、適合率で１２．８％、再現率で１２．２％それぞれ向上した。

表１.マルコフ辞書の違いによる誤り検出結果

	順	逆	中	順＋逆	逆＋中	中＋順	順逆中
全体	64.2 60.4	64.0 60.4	51.0 48.1	77.0 72.6	60.0 56.6	59.8 57.5	65.4 65.1
先頭	72.4 58.2	26.5 21.3	——	69.1 56.6	26.5 21.3	70.8 61.5	60.4 57.4
内部	72.3 71.6	81.2 80.4	72.3 71.6	81.2 80.4	77.2 76.5	72.3 71.6	75.2 74.5
末尾	21.6 17.4	67.6 54.3	——	67.6 54.3	67.6 56.5	——	70.3 56.5

上段：適合率Ｐ　下段：再現率Ｒ　単位：％

順：順方向２重漢字かな文節マルコフ辞書
逆：逆方向２重漢字かな文節マルコフ辞書
中：中間方向２重漢字かな文節マルコフ辞書

４．結論

本研究では、マルコフモデルを用いた日本語文節の誤り文字検出実験を行い、種々のマルコフモデルを組み合わせ、より有効な手法を提案した。

順方向に学習されたマルコフモデルと逆方向のそれとを組み合わせた積の条件で誤り検出を行うことで、より高い検出率が得られることが分かった。

今後は、これらのマルコフ連鎖確率辞書の組み合わせだけでは検出されない誤りについて、効果的な検出手法の研究を進める。

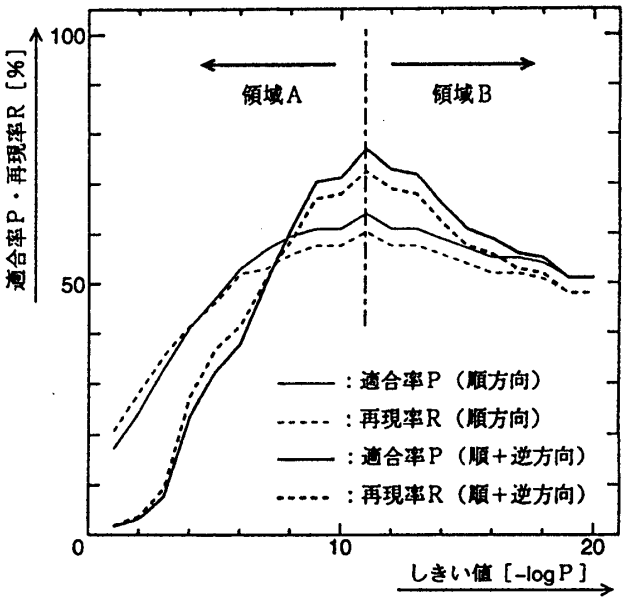

図２　しきい値による適合率・再現率の変化

領域Ａで再現率が低下する要因について

しきい値を甘く設定すると、誤りらしい文字列は多く見つけられるが、その範囲がぼやけ、元々独立していた誤り文字列が互いに干渉しあうため、一つの誤り文字列とみなされることで、再現率が低下する。

領域Ｂで適合率が低下する要因について

しきい値を厳しく設定すると、誤りの文字列の一部が正しいと判定され、誤り文字列の長さ（範囲）を正しく判定できなくなるため、適合率が低下する。

参考文献

［１］荒木、池原、塚原:"べた書き日本語文の脱落・挿入誤りの検出方法",情報処理学会自然言語処理研究会,93-NL-94-7,pp.49-54(1993)

［２］荒木、池原、塚原:"２重マルコフモデルによる日本語文の誤り検出並びに訂正法",情報処理学会自然言語処理研究会,93-NL-97-5,pp.29-35(1993)