マルコフ連鎖モデルによるかな文と英語文の誤り訂正

1Q－8

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

１．はじめに

ＯＣＲや音声認識装置を通して入力された文は、通常、誤字、脱落あるいは誤挿入の誤りを含んでおり、これらの誤りを自動的に検出し訂正するために、自然言語処理技術が期待されている。しかし、現在の自然言語解析の技術は正しい文に対して開発されているため、上記の問題に直接適用することが出来ない。

これまでに、誤りの種別とマルコフ連鎖確率が小さな値を取り続ける回数の関係に着目して、これらの誤りを検出し訂正する方法が提案され、２重マルコフ連鎖モデルを用いて、日本語文における誤字、脱落及び誤挿入の誤りを検出し訂正する方法の有効性が知られている[1]。

本論文では［１］の方法を、日本語の音節文字及び英字の３重マルコフモデルに適用し、その効果を新聞記事の統計データを用いた実験により評価する。

２．誤り検出と訂正方法

［仮定］誤りの音節列または英字列に対する各マルコフ連鎖確率の値は、正しい音節列または英字列に対するマルコフ連鎖確率の値に比べて小さい。

この仮定に従うと、誤りの種別と位置を決定する手順が次のように決まる。

（１）誤りの検出法

文節内に誤字、脱落および誤挿入の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値が減少する。このときの減少回数を調べることにより誤り種別および誤り位置を識別する。（表１）

（２）誤り文字の訂正法

(1)の方法により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、その中で最も高いマルコフ連鎖確率値をもつ時の文字候補を挿入した文を正しい文とする。また、(ii)誤挿入および誤字誤りならば、最初に誤挿入の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が見られれば、そのとき取り除いた文字を誤挿入文字の候補とする。次に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を調べる。このとき、連鎖確率値の最も高いときに置き換えた文字を誤字の候補とする。最後に、誤挿入文字の候補と誤字の候補を比較して連鎖確率値の高い方をとる。

３．実験

３．１　実験条件

①日本語新聞記事７７日分の文節数：283,963
②英字新聞４日分の単語数：155,459
③誤り種別と文節数：誤字、脱落あるいは誤挿入の誤りを１文節（または単語）当たり１箇所設定したものを1200文節（単語）使用
④マルコフ連鎖モデル：音節または英字の３重マルコフ連鎖モデル

３．２　実験結果

（１）誤り検出／訂正の適合率と再現率の関係

A Method of Error Correcting for Japanese and English Sentences Using Markov Chain Models
Tetsuo Araki[+] Satoru Ikehara[++]
Nobuyuki Tsukahara[+] Yasunori Komatsu[+]
[+] Faculty of Engineering Fukui University
[++]NTT Network Information Systems Laboratories

2.の方法により誤字、脱落および誤挿入について誤り検出と訂正を行った結果、誤り訂正後の文字列の適合率、再現率は図１（音節文節）、図２（英単語）の通りである。同図より、最大値を比較すると、英単語と音節文節はほぼ同程度であることがわかる。

（２）スペルチェッカーを用いた英単語の誤り検出／訂正

誤字、脱落および誤挿入の誤りを含む英単語に対して、スペルチェッカーを用いて誤り検出／訂正を行った結果を表２に示す。その結果、既存のスペルチェッカーは、誤った単語のほとんど全てを検出するが、訂正能力についてみれば、１文字の誤りでは正解候補が得られるのは80%程度であるのに対して、連続２文字の誤りでは正解候補は10%程度に低下するため誤り単語の検出についてはスペルチェッカーが、また誤り訂正についてはマルコフモデルによる方法が有利であると考えられる。

４．おわりに

本論文では３重のマルコフ連鎖モデルを用いて、日本語の音節文節ならびに英単語の、誤字、脱落または誤挿入を検出し訂正する方法を日英の新聞記事を用いた実験により評価した結果、２重マルコフモデルに比べて5-20%程度適合率及び再現率が改善する効果があることがわかった。また、スペルチェッカーによる誤り検出／訂正実験により、誤り検出についてはスペルチェッカーが、また訂正についてはマルコフモデルによる方法が有利であることがわかった。

今後の課題は、単語辞書引き法と組み合わせた効果を評価していくことがあげられる。

参考文献

（１）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,pp29,35(1993)

表１　誤り種別とマルコフ連鎖確率値の関係

	３重マルコフ	m重マルコフ
１文字脱落	３回落ち込み	m回落ち込み
２文字脱落	３回落ち込み	m回落ち込み
n文字脱落	３回落ち込み	m回落ち込み
１文字挿入	４回落ち込み	m+1回落ち込み
２文字挿入	５回落ち込み	m+2回落ち込み
n文字挿入	n+3回落ち込み	m+n回落ち込み
１文字誤字	４回落ち込み	m+1回落ち込み
２文字誤字	５回落ち込み	m+2回落ち込み
n文字誤字	n+3回落ち込み	m+n回落ち込み

表２　スペルチェッカの誤り検出と訂正能力

	検出可		検出不可
	正解候補有り	正解候補なし	
１文字脱落	７６．０	１７．５	６．５
２文字脱落	０	７９．５	２０．５
１文字挿入	８２．０	１８．０	０
２文字挿入	０	１００．０	０
１文字誤字	８０．５	１８．５	１．０
２文字誤字	４．０	９６．０	０

単位：　%

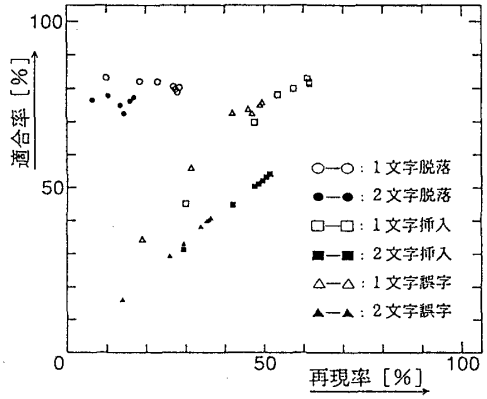

図１　誤り文字の訂正結果（音節文節：３重）

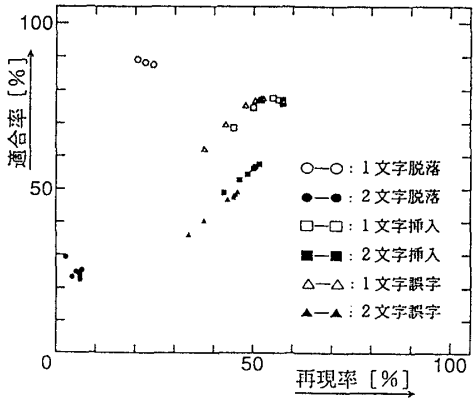

図２　誤り文字の訂正結果（英文：３重）