2J-3

日本文校正支援システムにおける評価方法の考察

村上仁一　荒木哲郎　池原悟

NTT情報通信処理研究所

1　はじめに

新聞記事などの作成において大量の漢字かな混じり文中に含まれる誤字、脱字などの誤りを検出する校正支援システムの研究、実用化が進められている（文献1）。このシステムの誤り検出の性能を評価する方法として実際の誤りデータに基づいて評価する方法が文献2で報告されている。しかしこのような方法は評価のための標本データが大量に必要なため人にかかる負荷が非常に大きい。ここではシステムの誤り検出の性能の概略を評価する新しい方法を提案する。この方法はかな漢字変換において生じる膨大な漢字かな混じり文を利用するもので、ほぼ自動化が可能であり、そのため人に対する負荷を必要としない。

この方法を用いて実際の校正支援システムで実験をおこなったところ、文献1で報告されている誤り検出の能力の機能評価値とほぼ等しい値が得られ、この評価方法の有効性が推定できた。

2　新しい評価方法の提案

2．1　コンセプト

文献2で報告されている評価方法は日本文の校正中において実在する誤りを校正支援システムに入力して、誤りが検出した数で評価している。ここで提案する方法は、かな漢字変換において生じる、莫大な数の誤りの持つ漢字かな混じり文をシステムに入力して、誤りが検出できた漢字かな混じり文の数で評価する。

この方法は「校正中における実際の誤りはかなから漢字への変換誤りで良く近似できる」と仮定している。この仮定については最後の考察において記す。

2．2　評価方法のフローチャート

校正支援システムの性能評価のための具体的なデータの流れを次にしめす。図1はこの流れを図で示したものである。

1)漢字かな混じり文を評価文として用意する。

2)評価文をひらがな文にする。

3)ひらがな文を漢字変換する。その結果、大量の候補文(漢字かな混じり文)が出力する。

4)出力された大量の候補文を校正支援システムに入力する。

5)校正支援システムにおいて誤りが検出された候補文の数を数える。

6)誤りが検出された候補文の数を、校正支援システムに入力した候補文の数で割って、その値を校正支援システムの誤りの検出の性能評価値とする。

$$\text{校正支援システムの誤りの検出の性能評価値} = \frac{\text{誤りが検出された候補文の数}}{\text{校正支援システムに入力した候補文の数}}$$

ひらがな文を漢字変換したとき大量の候補文が出力される。この中には元の評価文も含まれているが、その他の候補文は全て誤りであると推定される。したがって、6)で与えられる性能評価値が校正支援システムの性能評価の概略をしめす値と考えられる。

この方法は2)のブロックのひらがな文に変換するところを除いてすべて全自動化が可能である。そのため人にかかる負荷がほとんどない。

3　評価実験

ここで提案した評価方法を実際の校正支援システム(REVISE 文献1)に適用して性能の評価を行った。

3．1　実験の条件

The Evaluation Method for Revision Support System for Japanese Text

Jin'ichi Murakami　Teturou Araki　Satoru Ikehara

NTT communications and Information Processing Laboratories

この実験における入出力の条件は次の通り。

1)評価文

評価文は文節ごとに区切った新聞記事とした。実験は50文節、行なった。

2)ひらがな文

ひらがな文に変換するのは人間が行った。

3)かな漢字変換

かな漢字変換は文節数最小法を使用した。ただし、分割数は最小分割数+1まで分割した。また、頻度情報や、文法による絞り込みは行わなかった。使用した単語辞書は約１６万語である。

３．２　実験結果

入力50文節における平均の性能評価値は次の通りであった。

平均の性能評価値

誤りが検出できたもの　70%

誤りが検出できなかった残りの30% の内訳は次の通りである。

1)名詞連続複合語として解釈されたもの　25%

2)名詞＋動詞もしくは動詞＋名詞として解釈されたもの　5%

これらの値はすでに文献１で報告されている校正支援システムにおける誤り検出機能性能値７割とほぼ同じ値であり、未機能チェックの1)固有名詞、実在性、知識ベースによる照合チェック20%, 2)助詞同音異義チェック10% とほぼ等しい。

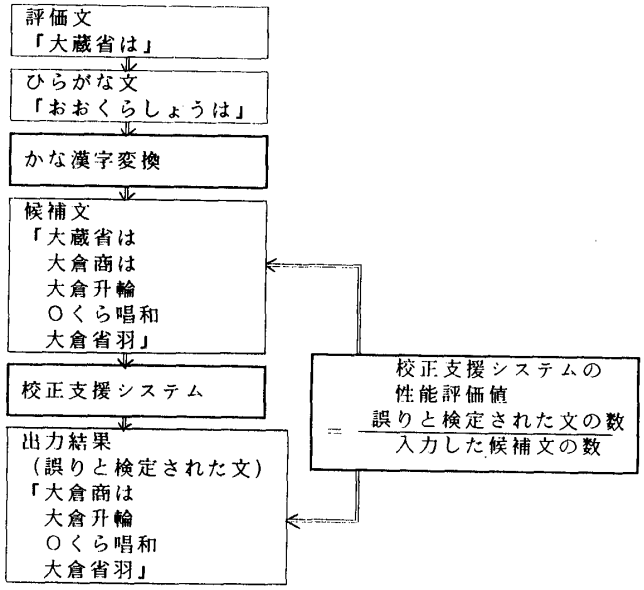

図　１　校正支援システムの評価方法

４　まとめ　ならびに　考察

ここでは新しい校正支援システムの評価方法として、かな漢字変換において生成される漢字かな混じり文を誤りのデータとして用いる方法を提案した。そして、この方法によって校正支援システムの評価を簡単に推定できる見込みがあることを示した。この方法は、自動化が可能であり、従って人に大きな負荷がかからない利点を持っている。

しかし、この方法は２つの大きな仮定を含んでいる。

「かな漢字変換で生成された大量の候補文の中で正しい候補文は１つである。」

「校正中における実際の誤りは　かなから漢字への変換誤り　で良く近似できる。」

これらの仮定は日本語における同音異義語や固有名詞の存在などの点から必ずしも正しくない。しかし、実験結果が実際の校正支援システムの性能値とほぼ一致したことからこれらの仮定がそれほど大きく外れたものでもないことが推定される。今後この仮定の妥当性を確かめるために別の校正支援システムでこの実験をすることなどが必要である。

参考文献

1)高木伸一郎、安田恒雄、島崎勝美、池原悟
「日本文訂正支援システム(REVISE)における誤り検定方式の検討」情報処理第34回全国大会6x-4

2)島崎勝美、安田恒雄、高木伸一郎、池原悟
「日本文訂正支援システムにおける評価法の検討」情報処理第36回全国大会5U-3