誤文中の形態素出現確率を用いる校正支援

5Q－5

金子　宏　奥村　薫　脇田　早紀子
日本アイ・ビー・エム　東京基礎研究所

1．はじめに

われわれは，日本語校正支援システムＦｌｅＣＳを構築し，新聞社における実用化を行い，機能の向上を続けてきた[1][2]．そこでは，主に習慣性のある誤りを対象にして，校閲者の知識を効率的に取り込むことを目指してきた．新聞社の場合には表記上の基準が規定されているため，この方式が有効であった．

このシステムを習慣性のない誤りにも適応させるために，偶発的に発生する形態素列を，あたかも習慣性のある誤りであるかのように取り扱って一応の成果を得たが[3]，習慣性のない誤りへの対応としては確率的手法が妥当と考えられる．

確率的手法を用いると，第１種の過誤（検出漏れ），第２種の過誤（過剰検出）が発生する．何らかのしきい値を用いて検出限界を定めるならば，第１種の過誤と第２種の過誤はトレード・オフの関係になる．われわれは，しきい値に，正文・誤文双方における形態素列出現確率を用いて，システムの最適化を試みている．

本稿では，この方式の概要を報告する．

2．過誤のトレード・オフ

例えば形態素列の出現確率が低いものを警告するような校正支援システムを考える．ここで警告するかしないかのしきい値となる確率の値は，それを小さくすれば第１種の過誤が増え，大きくすれば第２種の過誤が増える．このようなトレード・オフは，形態素列の出現確率を用いる手法に特有のものではなく，校正支援システムに必然的に伴う問題と考えられる．

このトレード・オフに対して，以下のアプローチが報告されている．牛島らの「推敲」では完全な文書を作成するためのツールとの立場で，第１種の過誤を０とする前提で，第２種の過誤を最小化している[4]．下村らは，正しい文中における確率に基づいて，しきい値を変化させて実験し，実用的と考えられるしきい値を選択している[5]．

われわれは，第２種の過誤をあらかじめ定めた値以下にするという制約のもとで，第１種の過誤を最小化するというアプローチを採用した．これは，以下の理由による．

新聞社の現場では，第１種の過誤は「省力化に役立つかどうか」，第２種の過誤は「オペレータが不満をもたずに使えるかどうか」を示す指標となる．われわれのシステムは，習慣性のある誤りについての性能のみでも，省力化に役立つと評価されている．今回の研究は，検出範囲をさらに拡張するものであるから省力化の点は実証済みであり，使い勝手に悪影響を与えることのないよう，第２種の過誤をコントロールすることが重要である．

A Critiquing Support by Probabilities of Morphemes in Erroneous Texts
Hiroshi Kaneko, Kaoru Okumura and Sakiko Wakita
IBM Research, Tokyo Research Laboratory, IBM Japan
1423-16, Shimotsuruma, Yamato, Kanagawa, 242 Japan

3．われわれの方式の定式化

$a_1, a_2, \ldots, a_n$を互いに素な事象とし，a_iの正文，誤文中の出現確率をp_i，q_iとする．全事象 $N=\{1,2,\ldots,n\}$の部分集合としての警告対象事象 Wを定めたとき，第k種の過誤率 E_kは，

$$E_1 = e\sum_{i\in N-W} q_i$$

$$E_2 = (1-e)\sum_{i\in W} p_i$$

となる．ここで，eは校正対象文章中の誤り率である（池原ら[6]もeを考慮する必要性を指摘している）．また検出率 Rは，

$$R = \sum_{i\in W} q_i$$

となる．

われわれは，e，p_i，q_iを実測して値を求め，それに基づき，$E_2<E$を満たして Rを最大にするように Wを最適化することを提案する．（この最適化は「ナップザック問題」になり，既存プログラムによって解くことができる．）

4．実用のための工夫

われわれの方式を実現する上で，最も困難な点は，q_iの測定である．統計的に意味のあるq_iの値を得るには，最低$10/eq_i$程度のデータ（誤りを含む文書）が必要である．われわれは，限られたデータからq_iを測定するため，「互いに素な事象」を以下の手順により実験的に定義した．

(1) 全ての形態素３連鎖を「互いに素な事象」とみなして，誤文中の出現数を数える．
(2) 出現数が10未満のものは，適宜，事象を統合する．例えば，名詞－（行にかかわらず）五段動詞語幹－（連用形以外の）活用語尾の連鎖は，「名詞と動詞の間の助詞の脱落」と考えられるので，これを一つの事象とする．
(3) 出現数が10未満で，統合できないものは出現数０とする（検出をあきらめる）．
(4) 統合された事象を「互いに素な事象」としてq_iを測定する．

このような手順では，(2)の統合過程において校正についての知識を利用することになる．

5．実験

新聞の校正前原稿約130万字を用いて，上に述べた方式を実験した．Eは，経験的に0.0004とした．また，E_2の計算に現れるeの値については，$1-e\fallingdotseq 1$として近似した（統計的誤差の方が大きいので問題ない）．

この結果，以下の３種のものが確率的手法による検出対象となった．

(1)連体形，連体詞の直後の用言
(2)孤立した１漢字固有名詞（「勝ことが」）
(3)文節頭の形式名詞，補助用言

検出数は，上記の３種について32件であった．習慣性のある誤りに加えて，この分だけ習慣性のない誤りを検出できるようになった．

謝辞

産経新聞社校閲センターの方々に感謝いたします．

文献

[1] 奥村他：情報処理学会ＮＬ研，92-NL-87，pp.83-90（1992）
[2] 脇田他：情報処理学会第45回全国大会，3F-4（1992）
[3] 脇田他：情報処理学会ＮＬ研，93-NL-97，pp.19-26
[4] 下園他：情報処理学会第46回全国大会，3L-2（1993）
[5] 下村他：情報処理学会論文誌，Vol.33，No.4，pp.457-464（1992）
[6] 池原他：情報処理，Vol.34，No.10，pp.1249-1258（1993）