

3L-5

日本語校正支援システム F1eCS
—新聞校正における実力診断—

協田早紀子・奥村薫・金子宏
日本アイ・ビー・エム(株) 東京基礎研究所

日本語校正支援システム F1eCS は 1992 年 12 月 7 日から産経新聞社で実稼働を始めた。実務で入力される記事修正に F1eCS がどの程度の力を発揮しているか、開発者としても興味津々で見守っているところである。今回は、現状での精度、残された課題と今後の方針について報告する。

F1eCS の警告には二通りの使用方法がある。

- (1) パソコンの入力画面上で色をつけて表示
- (2) 記事のプリントアウト(モニター)に傍線を表示

校閲者は(2)の紙に直接「赤字(校閲者による修正)」を入れる。この「モニター」を大量に入手できるので、F1eCS の精度を調べたり校正用の辞書・ルールを修正したりするのに便利である。

1. 精度

まず、モニターをもとに F1eCS の精度を調べてみよう。

精度というと、普通は検出率(警告を出せた割合)のことを指すことが多いだろう。しかし誤り個所の 9 割以上に警告が出ていたとしても、そのほかに不要な警告がたくさん出ていたら、その中から必要な警告を捜すのは煩わしい。つまり、検出率・過検出率(余計な警告の割合)の両方が重要となる。

それでは、検出率・過検出率というのは、どう計算するのが正しいのだろうか。この数字は、書かれている話題の種類によっても含む誤りの量によっても、また数え方によっても激しく変化するので実態をつかむのは難しい。具体例で述べよう。

データ : 1993 年 1 月 14 日分 約 420 字詰め×57 枚の本文(見出し・タグなど含まず)

記事の種類 : 書評欄・投書欄その他

記事の特徴 : 記者が直接ワープロで入力したものと異なり、書き原稿をもとにオペレーターが入力したもので、タイミス・脱落・口語・方言・原稿誤りなどが多い。署名入り原稿については産経用字用語規則に合わなくても原稿を尊重する(赤字をいれない)。

赤字部分 115 個所のうち

- | | |
|-------------------------|----|
| a. 誤り・警告が出ていた | 52 |
| b. 誤り・警告が出ていなかった | 25 |
| c. 日本語としての誤りではない(注) | 38 |
| (注) 組み版指定・文体の好み・数字の誤りなど | |

警告があり赤字がなかった部分 113 個所のうち

- | | |
|---------------------|----|
| d. 警告は正しいがあえて直さなかった | 51 |
| e. 不要な警告だった | 62 |

a. と b. の内容例(カッコ内が訂正後)

a. 誤り・警告が出ていた

タイプミス

つかれたような(つかれたような), それにしもなわなければ(それにともなわなければ)
実感をものではない。(実感をともなつたものではない。), 密柑畑(蜜柑)

送り仮名など表記規則違反

お話しができました(お話ができました)
ジャカルタ(ジャカルタ), 葉書(はがき)

変換ミス

それと平行して(並行して)
私が理事を努めている…(務めて)

その他

感じ。。(。), 珠江-マカオ(珠江-マカオ; 長音と単柱の間違い)

b. 誤り・警告が出ていなかった

表記規則登録漏れ

と博場(とばく場), しょうがない(しょうがない)

変換ミス

病因追及の(追究の), 留学性たちの(留学生)
詳細名計画報告書を正式に提出した。(詳細な)

タイプミス

扱おうとして時代と定義し(扱おうとした時代), 数学上の八十年代(数字上の)
異論さえめつたにない(移動さえめつたにない)
自分の心に生じた疑問の心に生じた疑問を(自分の心に生じた疑問を)

[検出率]

検出率の定義はいくつか考えられる。

- (1) 赤字(実際の直し)のうちいくつ警告を出していたか
 $a/(a+b+c) = 45\%$
- (2) 誤りのうちいくつ警告を出していたか
 $(a+d)/(a+b+d+警告も赤字も出なかった誤り) = 8割弱$
- (3) 赤字のうち文自体が誤りである中でいくつ警告を出していたか
 $a/(a+b) = 74\%$

これを見てもわかるように、計算方法によってずいぶん印象がかわるものである。開発する側としては(2)を指標として向上させる努力をするが、使用する側の実感としては(1)に近くなってしまう。ただし、逆に機械の助けを借りて、現在不徹底な表記規則を社内に広めていこうという動きもあり、dは今後減る傾向にあるかもしれない。すると見た目の検出率が上がる。変な話だが。

[過検出率]

過検出率といっても何を何で割るべきかは明確でない。開発者としては、単位長さ文章あたりの不要な警告数というのが、文章の質によらないひとつの指標だろう。また、過検出率というのは見た目の問題だという立場にたつて誤りの実数はとりえず無視し、警告のうちどれだけがウソだったか

$$e/(a+d+e) = 38\%$$

というのもある意味で妥当な指標である。けれども使用者にしてみれば、結果として不要だった率

$$(d+e)/(a+d+e) = 68\%$$

の方が実感かもしれない。開発者とのずれは大きい。

また、これらの数字に出ない精度として、警告の出る誤りの傾向がある。「○○に関してなら機械が信用できる」となれば心理的な負担はずっと減るものだ。人間が見過しがちなところに機械が強ければなおよい。例えば校正辞書・ルールのカズや不足はほぼなくなって用字用語規則については間違えないようになれば、校正者は文脈上の誤りなどに専念できる。

以上、まとめると

- ・この記事について現在のF1eCSは誤りの8割近くを検出するが、警告のうち4割近くが余計なもの(未登録語)である。(政治面などでは未登録語が少なくなってくると印象が変わる)
- ・ただし使用者側の実感はこの数字とだいぶひらきがあるかもしれない。

2. 課題と今後の方針

- b. 誤り・警告が出ていなかった
- e. 不要な警告だった

の二項目をつぶしていく。両項目とも重要だが、現在のところ使用者側からはe.についての要求の方がやや強い。

b. について

表記規則の登録漏れについては発見し次第登録する。これはしばらくすると落ち着くと思う。

変換ミスは難しい問題である。ヒューリスティクスにより発見するルール³⁾を今後も充実させていくが、それだけではうまくいかないものもある。

タイプミスはある意味でさらに難しい問題を含んでいる。予想もしない間違い方がいくらでもありうるし、ミスの結果が文法的におかしい場合などは検出漏れになりがちである。人が読めばおかしいとわかる場合でも、機械が形態素解析をするつもりやり解釈できてしまう場合は数多くある。形態素列の特徴や単語の長さ、字種などを利用して、その先に踏み込む方法はいくつか考えられるが、副作用として不必要な警告が出ることもある。この問題については後日報告する予定である。

e. について

現在のところこの数はかなり多い。校正用辞書・ルールのバグまたは副作用という場合もあるが、それはその都度修正していけるし数も少ない。圧倒的に多いのは未登録単語である。新聞記事に用いられる単語の多さはあらためて言うまでもない。特に外国人名・地名・スポーツ用語などの登録単語がまだまだ不足しているので、これから充実していきたい。「未登録単語または接続不良」として警告が出た単語は、自動的に収集しておく仕組みを作った。そのリストをチェックしたのち追加していくことで、早い時期に大量に登録単語を増やせるだろう。しばらくすると未登録単語は減ってくるが、なくなるということはない。人名・会社名・本や映画の題名など、日々新しい単語が生まれる。くだけた表現や方言、かな書きカナ書きなどの表記をすべて登録するというわけにもなかなかいかない。「未登録単語または接続不良」の警告は他の警告と区別がつくように表示しているが、さらに見せ方の工夫などをしていく必要がある。

3. まとめ

当面の問題である「未登録語」大量登録が済めば、スペルチェッカーとしてある程度のレベルになる見通しがたってきた。細かな修正によって達成できる精度の限界をなるべく早い時期に見極め、さらなる飛躍を目指したい。

謝辞

F1eCSを育てる地道で膨大な作業に協力して下さっている産経新聞社の方々に感謝の意を表します。

参考文献

- 1) 奥村ほか: 日本語校正支援システム「F1eCS」: 92-NL-87, 情処 自然言語処理研究会(1992)
- 2) 脇田ほか: 文中における語句の『近さ』について: 92-NL-90, 情処 自然言語処理研究会(1992)
- 3) 脇田ほか: 日本語校正支援システムF1eCS-新聞用ルールの獲得と表現: 情処第45回全国大会3F-4(1992)
- 4) 奥村ほか: 日本語校正支援システムF1eCS-新聞社における実用化報告: 情処第45回全国大会3F-5(1992)