5Q－6

日本語校正支援システムにおける校正知識 -同音異義語について-

奥村薫　脇田早紀子　金子宏·
日本アイ・ビー・エム㈱　東京基礎研究所

１．はじめに

著者らは新聞記事を対象として日本語校正支援システムＦｌｅＣＳを開発してきた。当システムは、１９９２年１２月より産経新聞社にて使用されており、誤りの約９割を検出している[1]。本論では校正対象文書に現れる同音語誤りおよび、かな漢字変換プログラム結果の同音語誤りを収集し、比較分析を行った。

２．問題提起

ＦｌｅＣＳ開発では実用性を重視して『よく起こる誤りから対処する』という方針を取ってきた。すなわち誤り単語を（もしくは単語列を一単語と取り扱って）校正辞書に登録し、また辞書では難しいが誤りやすい語に付いては、それぞれヒューリスティックスを校正ルールとして記述した[2]。これにより典型的誤りは発見するものの、同音語一般に対処しているわけではなく、他の種類の誤りより検出率も低いので、さらなる向上を目指している。

同音語の選択に付いては、かな漢字変換プログラムですでに多くの研究がなされている。そこで今回、かな漢字変換プログラムと校正支援対象での同音語誤りを収集し、比較検討を行なった。

３．調査対象

同音語の誤りで、次の種類を除いたもの。
(1)新聞社では使わないことが決まっている単語。
常用漢字以外の字、正書以外の送り仮名など。
(2)同一単語の漢字・かな表記の相違。

その理由：(1)は校正辞書で網羅的に検出できるし、(2)も校正辞書・ルールで既にその多くに対処できているため、今回の調査から省いた。

【かな漢】かな漢字変換プログラム出力の第一候補と期待された漢字書きとの差。
・新聞記事、手紙文から
・約3万6千文字分
・収集誤り数　227件

【校正】新聞社において校閲者が赤字訂正した誤り。
・新聞記事。各種の面から。
・約36万文字分。
・収集誤り数　77件

４．分析

同音語誤りを以下のカテゴリーに分類する。例示の括弧内が正解である。

(1)文節・単語の境界や、品詞が異なる
(1A)自立語⇔自立語＋付属語
例：大阪市来たく[北区]長柄西１丁目
(1B)単語境界が変化する
例：今期での高大切[交替説]もささやかれて
延長洗浄[延長線上]のものだから
(1C)単語の品詞が異なる
例：寺山自信[自身]のためにも
質問を内臓[内蔵]していた

Knowledge on Homonim Error for Japanese Critiquing System,
Kaoru Okumura, Sakiko Wakita, Hiroshi Kaneko,
Tokyo Research Laboratory, IBM Japan.

	【かな漢】			【校正】		
カテゴリー	件数	%	発生率	件数	%	発生率
1A	32	14%	8.9	1	1%	0.03
1B	22	10%	6.1	2	3%	0.07
1C	16	7%	4.4	3	4%	0.08
2A	122	53%	34	11	14%	0.31
2B	35	16%	9.7	60	78%	1.7
計	227	100%	63	77	100%	2.1

※発生率は1万文字あたりの件数をあらわす。

表1:誤りの分類

(2)同一品詞の単語

(2A)かなり異なる単語同士

例：神鍋広言[高原]マラソンの時に
　　単勝４番人気に指示[支持]された

(2B)類義語、または形が似ている。

例：事情を聞[聴]いてみなければ
　　ビール戦争幕明[開]けとなる気配
　　見る側に選択肢[枝]があふれている

表１にかな漢字変換結果および校正対象文書中の上記分類による同音語誤りを示す。直観的には分類(1)は見た目がかなり違う場合が多く、逆に(2B)は似ているといえるだろう。

校閲がチェックする文はかな漢字変換結果を人間が修正したものであり、それでも見逃してしまった、あるいは思い違いから残ったものが対象となる。かな漢字変換プログラム結果の誤りに比べると、(1)の類い、ついで(2A)の誤りの比率が少ないのはもっともなことであろう。校正で発見すべき誤りの８割近くが同音類義語であることも特徴的である。

対象文書が異なっているが、人間の変換ミス見逃し率を概算してみたものが表2である。

また現在のＦｌｅＣＳによる検出率も表3に付加する。これは14,000語程度の誤り語辞書と400種程度の校正ルールで検出される、かなり再現性のある同音語誤りの割合と見ることが出来よう。

カテゴリー	見逃し率
1A	0.3%
1B	0.9%
1C	1.9%
2A	0.9%
2B	17%

表2:人間の見逃し率の目安

カテゴリー	件数	検出数（内辞書/ルール）		
1A	1	0		
1B	2	0		
1C	3	3	0	3
2A	11	5	5	0
2B	60	36	30	6
計	77	44(57%)	35	9

表3:ＦｌｅＣＳによる検出件数

5．おわりに

校正支援では同音語のうち特に同音類義語に対する誤り検出が重要であることがいえる。かな漢字変換プログラムで同音語対策とされている共起情報を用いるにしても、一般的な共起頻度よりは類義語にしぼり込み、類義語を識別し得る情報を主に収集すると有効であると考える。

現在、著者らは同音語のチェックのため、単語間の共起、助詞や接辞との親和性などのデータを産経新聞社校閲センターと共同で整備しつつある。

謝辞：　本研究に多大な協力をいただいている産経新聞社校閲センターの方々に深謝いたします。

参考文献:

[1] 奥村ほか：日本語校正支援システムＦｌｅＣＳによる新聞記事校正，情報処理学会第47回全国大会4W-7，(1993).

[2]奥村ほか：日本語校正システムＦｌｅＣＳの新聞社における実用化、情処研報92-NL-91,(1992).