5R－3

日本語校正支援における同音語誤り検出 －警告レベル分けの判定基準

脇田早紀子・奥村薫・金子宏
日本アイ・ビー・エム東京基礎研究所

１．はじめに[1]

日本語校正支援システムにおいて、変換ミスの単語にもれなく警告を出してほしいという要求が以前から強かった。だからといって同音異義語があるものにすべて警告しては、警告の数が多くなりすぎてとても実用にならない。

そこで前回の発表では、「同音異義語あり」の警告を３段階にレベル分けすることにより、間違いの可能性が高いものを強調し、間違いの可能性の低いものを消して見やすくする方法について提案した。

今回は、レベル分けの際の判定基準についてより詳しく検討する。

２．警告レベル分けのしくみ

まず、校正支援システムで通常の解析に用いていた辞書のほかに、同音語組の単語を登録した辞書を用いる。この辞書の内容例を表１に示す。

表１　同音異義語辞書の例

[よういん]
*要員　名詞
直後：の　を　数　枠　面
近く：安全　医療　介護　援助　必要　不足　任務
　　派遣　参加　装備　展開　撤回　内訳　配置
　　代替　保安　足り(る)　送(る)
*要因　名詞
直前：主　諸　一
直後：が　と　に
近く：悪化　圧迫　安定　円安　円高　価格　回復
　　外部　外的　考慮　構造　減少　分析　複雑
　　輸出　抑制　気象　阻害　一つ　除(く)
　　増(す)　多(い)　大き(い)

表１の例では、一つの読み「よういん」を持つ単語「要員」「要因」（以下、見出し語と呼ぶ）があり、それぞれに対して直前／直後／近くに出てきやすい語（以下、手がかり語と呼ぶ）を列挙している。

このような辞書を基にして、文章中の手がかり語により警告のレベルを決定する。

- レベルA［危険］
  「円高の要員」など（「要因」の方の手がかり語「円高」があるので）
- レベルB［注意］
  有効な手がかり語が文中にないとき
- レベルC［O.K.］
  「要員が不足している」など（「要因」の方の手がかり語「不足」があるので）

３．レベル分けにおける問題点

以上が警告レベル分けの基本的な考え方だが、考慮しておかなければならない点がいくつかあるので以下に述べる。

同じ手がかり語の重複

---

[1] Homonym Error Detection in Japanese Critiquing
-How to define the Levels of Notices,
Sakiko Wakita, Kaoru Okumura, Hiroshi Kaneko,
Tokyo Research Laboratory, IBM Japan.

意味が似ているか分野が同じ同音語の場合は同じ手がかり語を持っていることがよくある。同音語組のすべてが同じ手がかり語を持っているときは手がかりとしての意味がないので辞書から削除するが、一部の見出し語のみに重複しているときは候補をしぼる役に立つ。

例文１：「新しい政治を打ち出すべきだ」などと奇声を上げた。

「きせい」の同音語組「奇声」「気勢」「規制」「既成」「寄生」のうち、「奇声」「気勢」の両方が「上げ(る)」を手がかり語として持っているのでどちらかわからない。そこでレベルＢ［注意］のままになるとしても、「規制」「寄生」「既成」の手がかり語はないので置き換え候補から削除して「気勢」のみを示すことができる。

弱い手がかり語

たいていの手がかり語は、ある見出し語をレベルＣ［O.K.］にし、他の見出し語をレベルＡ［危険］にする働きをするが、そこまではっきりしない手がかり語を、前者の働きのみをする語として登録する。

例文１の場合、「気勢」は特に「～と気勢を上げた」と使われやすいので「気勢」の前に「と」があればとりあえずレベルＣ［O.K.］にする。ただし「奇声」「既成」「規制」の前に「と」があってもレベルＡ［危険］にはしない。

遠い手がかり語

一つの文中に出てきても、あまり遠くの手がかり語は見ない方が安全である。

例文２：代表者会議の論議を見極めたうえで十八日以降に開く中央執行委員会に誤る意向を表明した。

「表明」は「意向」の手がかり語であるが遠すぎる。それより「以降」の直前に「日」があることからレベルＡ［危険］と判定したい。

便宜上、判定したい語の前後 15 文字を見ることにしているが、この「遠さ」判定にはなお検討が必要である。

複数手がかり語の競合

レベルＡ［危険］と主張する手がかり語とレベルＣ［O.K.］と主張する手がかり語が競合してしまうことがある。

例文３：子供を取り巻く状況は改善されていることを協調し、こうした前進を維持するために先進国の支援を求めている。

正しくは「強調」であるが「強調」の手がかり語「改善」「（直前の）を」とともに協調の手がかり語「維持」も出てくる。

競合する場合、直前・直後の手がかり語があればそれを優先し、手がかり語の数が違えば多い方にしておくなどの工夫をしているが、決め手がない場合はレベルＢ［注意］のままとする。

４．まとめ

以上、「同音異義語あり」警告をレベル分けする際の問題点として、

- 同じ手がかり語の重複
- 弱い手がかり語
- 遠い手がかり語
- 複数手がかり語の競合

について検討した。現在は、これらの点を考慮したシステムを試験的に使用している。

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

参考文献

奥村ほか: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,情処 48 全国大会 5Q-6,(1994)

奥村ほか:日本語校正支援における同音語誤り検出－警告レベル分けの提案,情処 49 全国大会 3K-6,(1994)