3K－6

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

奥村薫・脇田早紀子・金子宏
日本アイ・ビー・エム株式会社

１．はじめに

日本語において、変換ミスはやっかいな問題としてよくとりあげられる。これまで我々が開発してきた校正支援システムでは、変換ミスに「誤用語辞書」と「校正ルール」で対処してきた[1)]が、まだ使い分けを記述していない同音異義語組が大量にある[2)]。そこで、変換ミスの可能性のある単語にはもれなく「同音異義語あり」の印をつけてほしい、とユーザーから要求されることがある。しかし、文章が印だらけになって結局何もつけないのと変わらないことになっては役に立たない。そこで今回は、テキストから自動抽出したデータを用いて、「同音異義語あり」の警告を危険度でレベル分けすることを提案する。

２．警告レベル分けの提案

「間違いやすい同音異義語がある」ものを全て警告すれば「検出率」は上がるが、警告が多過ぎて見る気がしなくなってしまう。そこで、「同音異義語あり」の警告を危険度で３段階にレベル分けして見やすくすることを考える。

レベルA[危険]
　同音異義語があり間違いの可能性大→目立つ警告
レベルB[注意]
　同音異義語があるので注意→やや控えめに警告
レベルC[O.K.]
　同音異義語があるがまず大丈夫→警告しない

このようにすれば警告の総数が減らせるし、急いでいるときはレベルAだけを見ることもできる。全部見るとしてもメリハリがついて見やすい。

今回は、以下の基準でレベルA～Cの分類を行った。

同音異義語組X0～Xnがあり、その中の単語(X0)が出現したとする。
レベルA[危険]　X1～Xnのどれかの使用法と似ていて、X0の使用法としては登録されていないもの。
レベルC[O.K.]　X1～Xnの使用法としては登録されていず、X0の使用法と似ているもの。
レベルB[注意]　レベルAでもレベルCでもないもの

本研究で言う「使用法」とは、データを自動作成する都合を考慮して、直前直後または近くに使われている語（助詞・名詞・接頭語・接尾語・動詞・括弧など）を指している。

例えば、「かせつ」には「仮設」と「仮説」があるが、「仮設」のそばには「住宅」「劇場」、「仮説」のそばには「検証」「限界」などが登録されているので、「仮設を検証する」はレベルAとして強く警告し、逆に「仮設の劇場」はレベルCとする。

３．レベル分けデータの自動作成

前節で述べた基準によりレベル分けを行うためのデータは、人力で作成するには膨大過ぎる。そこで今回は、手間をかけずに自動作成したデータでどのくらい役に立つレベル分けができそうかを見積もることにした。既存のテキストをもとに、手がかりとなる語のデータを生成する。

Homonym Error Detection in Japanese Critiquing - Three Levels of Notices,
Kaoru Okumura, Sakiko Wakita, Hiroshi Kaneko,
Tokyo Research Laboratory, IBM Japan.

材料テキスト：
　日経新聞の記事データ２カ月分（1400万文字分）
対象単語：
　材料テキスト中に10回以上出てくる一般名詞二文字漢字単語同士の同音異義語組

材料テキスト中に対象単語が出てきたとき、その単語の
●直前・直後の助詞、接頭語、接尾語、括弧
●近くの名詞・動詞など
を数え、用例の５％以上かつ複数回出てきたものを登録した。

自動生成したデータ例「いこう」

以降　直前:は　の　も
　　近く:四月
意向　直前:を　が　で
　　直後:だ
　　近く:表明　示(す)

見ると、余計なもの、あやしいものもあるがそれなりの手がかりになりそうなことがわかる。

４．レベル分け精度の見通し

産経新聞社で、実際の記事原稿に対し人間の校閲者が赤字を出した「変換ミス」を130例収集し、今回作成したデータをもとにレベル分けを行った。

例　新聞記事原稿の変換ミス例(括弧内は正表記)

●法的な面を調べ、もし必要があれば適性(適正)な対応をとりたいと考えている。
●ベトナムへの融資・技術援助の再会(再開)についても、
●長男の喜美氏が再(最)有力視されているが、
●短大教授を勤(務)める母から十八歳になる息子へ送る
●東京佐川急便の渡辺弘康(広康)元社長から金丸氏への五億円献金は、

無作為に収集した130文のうち、二文字漢字語同士の変換ミスは42個だった(人名の間違いを除く)。前節で作成したデータをもとにレベル分けを行うと、このうち、

　レベルA[危険]…18
　レベルB[注意]…17
　レベルC[O.K.]… 1

であり、残り6個は今回のレベル分け対象単語には含まれていなかった。

一方、不要な警告をどれだけ減らせるだろうか。レベル分け対象単語であるが、変換ミスでない例102個に対して、

　レベルA[危険]… 5
　レベルB[注意]…49
　レベルC[O.K.]…48

であった。

おおまかに言うと、誤りの半分近くにレベルA[危険]をつけ、正しいものについてしまう警告の半分近くをレベルC[O.K.]として取り除くことができる。そして、誤りをレベルC[O.K.]として除いてしまう危険はかなり少ない。

５．まとめ

同音異義語が存在し、変換ミスの可能性がある単語に注意を促す際、既存のテキストから自動抽出したデータを用いて３段階にレベル分けすることを試みた。その結果、間違いの可能性が高いものを強調し、間違いの可能性が低いものを消すことによって大幅に見やすくすることができる見通しが立った。今後、
●より大量のデータから抽出する
●抽出方法を工夫する
●自動抽出したデータの一部を人手により修正する
などによりさらに精度を向上させ、「使う気がする変換ミスチェッカー」を提供したい。

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致します。

参考文献:

[1]脇田ほか:日本語校正支援システムFleCS-新聞用ルールの獲得と表現,情処45全国大会3F-4,(1992)
[2]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)