

複合語の同音異義語誤り検出手法

4M-6

鈴木 努

東京大学工学系研究科電子工学専攻

浜田 喬

学術情報センター

1 はじめに

ワープロで作成された文書には、仮名漢字変換ミスによる同音異義語誤りが生じやすい。したがって、文書に含まれる同音異義語の誤りを自動的に検出することは有用である。本稿では、意味的な情報を用いて複合語の同音異義語誤りを検出する手法を提案する。

2 同音異義語の誤り検出

文書中の誤りを検出する基本となる言語処理技術は、形態素解析である。入力文を辞書に登録された単語と照合し、辞書に登録されていない語が見つかった場合は、その箇所に誤りがあると判断する。

しかしこの方法では、同音異義語の誤りを検出することができない。同音異義語の誤り語は単語辞書に登録されているからである。同音異義語の誤りを検出するためには、前後の単語との間の関係を利用する必要が出てくる。

3 複合語の同音異義語誤り検出手法

複合語を構成する単語間の関係に着目して、複合語に含まれる同音異義語の誤りを検出することを考える。問題とする複合語を x_1w_i とする。ただし単語 x_1, x_2 が同音異義語の対であり、単語 w_i は正しいことがわかっているとする。

1つの方法として、 x_1 に隣接しうる単語をすべてあらかじめ辞書に記述しておき、 w_i がその辞書に登録されていなければ誤りとする、というものがある。しかし、 x_1 に隣接しうる単語をすべて網羅するのは困難である。

そこで、似た意味を持つ単語を1つのグループ（意味分類）にまとめ、そのグループごとに x_1 に隣接しうるかどうかを記述しておく方法を考える。

表 1: 接続表 (S_i : 意味分類、 x_1, x_2 : 同音異義語の対、○: 接続可能、×: 接続不可)

	S_1	S_2	...	S_i	...
x_1	○	×	...	○	...
x_2	×	○	...	×	...

単語を、その意味によって有限個の集合に分類したものを、意味分類と呼ぶことにする。本手法では、分類語彙表 [2] を意味分類として利用する。分類語彙表では、似た意味を持つ単語が同じ分類に属し、分類どうしが階層構造を成している。

単語 w_1, w_2 が同音異義語であるとする。EDR 日本語コーパス [3] から w_1 を含む複合語を抽出し、 w_1 と接続している単語の意味属性を記述しておく。これにより、 w_1 がそれぞれの意味属性の単語と接続して複合語をつくることがありうるかどうかを表す接続表 (表 1) を作成する。文書の校正にはこの接続表を使う。

文書の校正にはこの接続表を使う。複合語 x_1w_i が入力されたとき、 w_i の意味分類を S_i とすると、 x_1 と S_i が接続不可であれば x_1 が誤りであると判断できる。

4 まとめ

複合語に含まれる同音異義語の誤りを検出する手法を提案した。今後は、接続表作成に用いるコーパスデータの充実をはかる予定である。また、同音異義語以外の誤り（文字脱落、置換）への応用も考えている。

参考文献

- [1] 奥 雅博: 日本文推敲支援システム REVISE における複合語同音異義語誤りの検出および訂正支援手法, 電子情報通信学会論文誌, Vol. J79-D-II, No. 11, pp. 1836-1846 (1996)
- [2] 国立国語研究所: 分類語彙表, 秀英出版 (1964)
- [3] 日本電子化辞書研究所: EDR 電子化辞書 1.5 版 (1996)