電子英和辞書検索ツールにおける見出し語の曖昧検索

静岡大学 光電機械工学科
若林研究室 50413028 永谷 真一郎

1 序論

電子辞書を用いて単語を調べる際，既存のツールの出力方法では，使いづらい時もある．そこで，英和電子辞書ツールの検索方法の改良を行う．具体的には綴りの違いを距離で表し，入力と比べて辞書中の単語から距離の近いものを見つけるといった検索方法を提案する．距離は，全く同じ綴りを 0 とし，違いがある程距離の値も多きくなるとする．その距離が近い順に出力することで，曖昧検索を実現する．単語間の距離を出力する距離関数を作成して，それを用いた，出力ツールを同時に作成する．

2 既存の距離関数の使用

2.1 既存の距離関数

既存の距離関数を辞書検索に使えるか検討する．この距離関数が対象とする誤りを表 1 に示す。

表 1: 誤りの種類

要因	記号
入力文字数が比較文字数より多い	G
入力文字数が比較文字数より少ない	L
比較文字列にない文字が入力文字列に表れている	N
比較文字列と入力文字列の文字の位置のずれ	R

上記の誤りを計数して，式 1 により距離 $dist$ を計算する．

$$dist = \frac{W_g \times [G] + W_l \times [L] + W_n \times [N] + W_r \times [R]}{M} \tag{1}$$

ここで，添字のある W は添字の誤りの重み．[ ] で囲まれているものは，その誤りの個数である．

2.2 結果

結果は，あまり似ていない単語が出力されたり，同じ点数の単語が出力として多く表れてしまったりして，好ましくないものであった．よって，新しく距離関数を作成することにし，そのための実験を行う．

3 距離関数の修正

3.1 実験

実験は，10 人の被験者に辞書中の日本語の意味 50 個を見てもらい，その意味に該当する単語を想像して答えてもらった．この実験で得られたことは，日本人が，外来語をカタカナで表すことが多いためであろうか r や l ，m や n などのカタカナ的に似ているものの誤りが多いということである．実験を基に，距離関数で扱うべき誤りを決定した．

3.2 修正した距離関数

新しく作成した距離関数は，文字列の長さ，位置のずれと誤った文字でも誤りやすい文字で間違った場合を対象にする．ここで，m_{err} は文字列の長さによって決まるずれても良い範囲である．

文字列 1 は，入力文字と比較文字列のうち文字数の少ない方である．比較する文字は文字列 1 を先頭から順に移動する．文字列 2 のその場所に対応する場所から，± m_{err} の近傍を参照して，誤りやすい文字，または，比較する文字を探して減点する．

距離関数の計算の考え方を図 1 に示す．

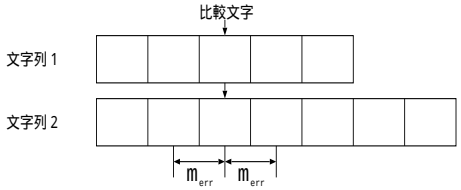

図 1: 距離関数の考え方

4 出力ツール

辞書ツールを作るにあたって，処理を簡略化するために自作のツールによって見出し語のみをファイルにした index.txt を作っておく．

辞書ツールは，距離の近い順に 6 番目まで出力する．使用者は，出力された単語のリストの中から数字を追加入力すると，意味を出力できる．

5 結果

結果を図 2 に示す．

```
   input = berner
1:          vernal: dist = 0.100000
2:           brine: dist = 0.160000
3:          berber: dist = 0.166667
4:         blender: dist = 0.166667
5:          burner: dist = 0.166667
6:          bearer: dist = 0.183333
7:          bender: dist = 0.183333
```

図 2: 出力結果

6 まとめ

英単語を対象とし，曖昧な見出し語検索を可能にするため，2 つの単語の距離を計算する方法を提案し，実験によりパラメータの調整を行った．距離関数を用いて，辞書ツールを作成した．距離関数は，日本人的な誤りを対象に選んだためカタカナ表記の単語を想像して調べる時にも訳に立つのではないかと思われる．今後の課題として，検索の高速化や，パターンによる検索等がある．