B4ゼミ2014/2/28(第2週)

スペル訂正と編集距離

長岡技術科学大学　電気系
自然言語処理研究室　高橋寛治

はじめに

- 真嘉比さんからスペル訂正について教えてもらい興味を持ったので、自分で動かしてみた
- 入門自然言語処理のNLTKに含まれるコーパスを用いる練習
- 編集距離を考えやすい英語を用いた
  - 日本語だとどうなる？

編集距離

- **編集距離**
  - 一方の語から他方の語を得るのに必要な修正の回数
  - 編集（削除、転位、置換、挿入）
- **長さnの語に対し**
  - 削除：n
  - 転位：n-1
  - 置換：26n
  - 挿入：26(n+1)
  - トータル：54n+25
- **例　”on” (n=2)**
  - [“o”, “n”]
  - [“no”]
  - [“an”, “bn”,・・・”oz”]
  - [“aon”, “obn”, “onc”]
  - 54×2+25 = 133個

スペルミス

- **英語におけるスペルミス**
  - speling、runing、plint・・・
- **スペルミスにも種類がある**
  - タイポ
  - 覚え間違い
    - rとl、mとn

実際に正解候補を出すところまで
プログラムを組んだ

手順

- 入力した単語に対して、編集距離1の単語パターンを取得
- ↑の単語で、辞書に載っているものを候補として絞る
- ↑の候補がコーパス中に何回使用されるか数え、頻度を返す
- ↑頻度が最大のものを正解として返す

入力について

- Wikipedia（Commonly misspelled English words）より
  - http://en.wikipedia.org/wiki/Commonly_misspelled_English_words
- 例
  - 誤り例
    - alcohol　　alchohol
    - Amateur　amatuer　amature
  - 正解例
    - affect　　affect

結果

	間違えた単語のみ入力	正解、間違えを混ぜた単語を入力
入力	271	542
意図した単語に修正	180	323
意図していない単語に修正	40	167
判断できなかった	51	52
適合率	0.82	0.66
再現率	0.66	0.60
F値	0.73	0.63