Developing A Real-Word Spelling Corrector

L445/L515

Autumn 2010

1. We’ve talked about $n$-grams for various language processing techniques before, and now I want you to think about how you would use trigrams in order to develop a real-word spelling corrector. Some issues to think about include:

- What will the probability model look like? That is, which probabilities will we compare?
- How will candidate correction sentences be generated?
  - How many changes per sentence will you allow? (Think about efficiency.)
  - Do you want to use pre-defined confusion sets, sets of commonly confused words (e.g., \{their, there, they’re\})?
- How will you handle issues of data sparseness?

Sketch out a design in very broad terms.\(^1\)

2. Now, let’s not use trigrams, but instead base our system on these confusion sets. What other kinds of information would help us disambiguate such content-based confusions sets like \{weather, whether\}; \{principal, principle\}; etc.?\(^2\)

References


\(^1\)For more on a trigram model, see: Mays et al. (1991); Wilcox-O’Hearn et al. (2006)

\(^2\)See, e.g., Golding and Roth (1999); Hirst and Budanitsky (2005)