Developing A Real-Word Spelling Corrector

Based on Dickinson, Brew, & Meurers (2013)

Fall 2013

Warning: this activity may cause you to think numerically.

1. We’ve talked about n-grams for language processing techniques, and I now 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:

   • Where will you get your probabilities? That is, what type of data are you interested in correcting?
     – Also: how will you handle issues of data sparsity?
   • How will candidate correction sentences be generated?
     – How many changes per sentence will you allow? (Think about efficiency.) Will you allow every word to be changed?
     – Will you make changes in the same order as the original sentence (e.g., changes to the first word first, then the second word, and so on)?
     – Do you want to use pre-defined confusion sets, sets of commonly confused words (e.g., {their, there, they’re})? How many sets would you need?
   • What will the probability model look like? That is, which probabilities will you compare?
     – Which trigrams will you use to calculate your probabilities? i.e., which words will you use in your trigrams? For example, if you’re trying to correct word C, do you look at \( p(C|AB_1) \), \( p(C|B_2D_3) \), \( p(C|D_4E_5) \), or some other probability?
     – How will you combine/compare the probabilities? Do you multiply probabilities? Do you want to focus only on the probability of a correction vs. the original, or take the whole sentence into account?

   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)