

Developing A Real-Word Spelling Corrector

Linguistics 445/515

Autumn 2008

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 confusion sets like {*weather, whether*}; {*principal, principle*}; etc.?²

References

- Golding, Andrew R. and Dan Roth (1999). A Winnow-Based Approach to Context-Sensitive Spelling Correction. *Machine Learning* 34(1-3), 107–130.
- Hirst, Graeme and Alexander Budanitsky (2005). Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering* 11(1), 87–111.
- Mays, Eric, Fred J. Damerau and Robert L. Mercer (1991). Context based spelling correction. *Information Processing and Management* 23(5), 517–522.
- Wilcox-O'Hearn, L. Amber, Graeme Hirst and Alexander Budanitsky (2006). Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer model. <http://www.cs.toronto.edu/compling/Publications/Abstracts/Papers/Wilcox%0Hearn-etal-2006-abs.html>.

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

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