LING439/539 - Statistical NLP
Assignment #2

Due Tuesday, September 27 2016 at 11:00AM

Simple linear interpolation

function interpolation(*corpus*) **return** $\lambda_1, \lambda_2, \lambda_3$

$\lambda_1 \leftarrow 0$

$\lambda_2 \leftarrow 0$

$\lambda_3 \leftarrow 0$

foreach trimgram t_1, t_2, t_3

if $\frac{C(t_1,t_2,t_3)}{C(t_1,t_2)} > 0$: increase λ_3

else if $\frac{C(t_2,t_3)}{C(t_2)} > 0$: increase λ_2

else if $\frac{C(t_3)}{N} > 0$: increase λ_1

end

end

normalize $\lambda_1, \lambda_2, \lambda_3$

return $\lambda_1, \lambda_2, \lambda_3$

1. Download the training and held-out corpora from `https://www.dropbox.com/s/132x4k222wvkf8z/corpus-brown.tar.gz?dl=0` (Brown corpus 50-50)
2. Insert `<s>` and `</s>` at the beginning and end of the sentences for your language model (1pt)
3. Using the training corpus find uni, bi, and tri-grams (3pts)
4. Using the held-out corpus calculate λs (5pts)
5. What is the trigram probabilities of *i want English food* ($P($`<s> i want English food </s>`$)$) (original and interpolated probabilities) ? (1pt)
6. Propose the better algorithm for the interpolation and calculate their λs and the probabilities of the above sentence. (optional 5pts).
7. Describe your work (the number of N-grams, the values of λs, how to execute your programs for steps 3 and 4, etc.) in README.txt (plain text format) within ONE page MAX and send to `jungyeul@email.arizona.edu` before 11:00AM on Tuesday, September 27. DO NOT SEND N-GRAM FILES. Use "LING439/539 Assignment #2" as a subject of the mail.