

# CIS 530 Spring 2008 Assignment 2

Due: Thursday, February 21th, 2008, 4:30pm

Please check our website <http://www.seas.upenn.edu/~cis530/> for general information and updated announcements on Assignments.

## Experiments on Smoothing Techniques

### Background

Consider a two-stage statistical machine translation system for individual sentences, such as the 'Spanish-English' one introduced in the lecture. Based on statistical analysis on foreign/native bilingual corpus, we compute a bag of 'broken English'(in our context, the native language) from a given Spanish sentence(the foreign language); then based on statistical analysis on English corpus, we generate a English sentence constituted with the bag of English words/segments.

Specifically, at the second stage, we search for a English sentence  $S$  that maximizes  $P(S|\lambda)$ , where  $\lambda$  is the language model we built on the English Corpus. Recall that, according to a language model  $\lambda$ ,

$$P(S) = P(w_1) * P(w_2|w_1) * P(w_3|w_1w_2) * \dots * P(w_n|w_1\dots w_n), \text{ where } S = w_1w_2\dots w_n$$

In this assignment, we expect you to implement a language model,  $\lambda$ , to be used in the second stage of this kind of machine translation system as introduced above; then select the most probable translated sentence constituted with output of the first stage. To avoid the use of and advanced search algorithm, a pool of grammatical/ungrammatical native sentences will be provided, instead of a bag of English words/segments.

(Note: the machine translation system we play with here is not an architecture widely-used nowadays.)

### Task

From the pool of English sentences we provide, pick out the proper translation of the German sentence

- (1) *Fuer meinen Geschmack sprechen und lachen sie viel zuviel*  
for my taste talk and laugh they much too.much  
"They talk and laugh too much for me."

Please download the pool of sentences, i.e. the test corpus, from our website. The pool of sentences contain the following four sentences:

Sample1: They talk and laugh a great deal too much for me .

Sample2: For my taste speak and laugh they much too much .

Sample3: For me talk they and laugh a great deal too much .

Sample4: They speak and laugh much too much for my taste .

Use the test corpus, i.e. the pool of sentences, the same way as for corpus imported from `nlk`.

```
from hw_data import sentpool

print sentpool.__doc__
```

## 1 Bigram based language model (30 points)

\*Based on the conditional frequency distribution of Austen's *Emma*, `bigram_dist`, prepared in homework I, compute a bigram based language model estimated by the *Maximum Likelihood Estimator*.

Suppose you store the bigram model in `bigram_MLE`; add the following line to your script as a test code:

```
print bigram_MLE['I'].prob('hear')
```

\*Turning off backoff, compute  $P(S|bigram\_MLE)$  for each sentence in the test corpus.

\*Do you get satisfying figures? If not, comment in a few words.

Note that, compute the sum of log-probabilities instead of the product of probabilities to avoid underflow. For example, given a list of small probabilities stored in `probs`, DON'T compute `reduce(lambda x,y: x*y, probs)`, instead, compute `reduce(lambda x,y: x+y, [log(p) for p in probs])`

## 2 Smoothing techniques (50 points)

`nlk` provides implementation of various smoothing techniques in `nlk.probability`. Among them, play with `LaplaceProbDist`, `WittenBellProbDist`, `CrossValidationProbDist`, and `GoodTuringProbDist`. References for NLTK can be found on our website.

\*Making use of the four different techniques, smooth the conditional frequency distribution `bigram_dist`; recompute  $P(S|\lambda)$  for each sentence in the test corpus, where  $\lambda$  is a smoothed bigram model.

Add the following lines, or counterparts, to your script as test code:

```
#in the loop body for each sentences in the pool
print reduce(lambda x, y: x+' '+y, sent)

#for each smoothing techniques
#sprob contains the probabilities of the corresponding sentence computed by smoothed models
print tech, ':', sprob[tech]
```

### 3 Pick out the one (20 points)

\*Now, pick up the sentence that maximizes  $P(S|\lambda)$  according to some language model.

\*Have you got an acceptable translation of the German sentence? Is every smoothed language model able to differentiate the sentences in our pool? If not, explain.

### 4 Debriefing

1. Approximately how many hours did you spend on this assignment?
2. Would you rate it as easy, moderate, or difficult?
3. How deeply do you feel you understand the material it covers (0%-100%)?
4. Any other comments?

This question is intended to help us calibrate the homework assignments. Your answers will not affect your grade.