

Assignment 1: Unscrambling Sentences

January 27, 2009

In section 1.4.5 of the reading, we covered the “noisy channel” model used in speech recognition and machine translation. In this assignment, you will build a language model and use it to solve a simulated machine translation task.

Our simulated machine translation system creates a garbled version of the input sentence, permuting, deleting or copying words at random. For instance, the sentence:

‘Response to Civil War in Sierra Leone’

might be altered into:

‘Response to Civil War in **in** Sierra Leone’

(Duplicate words often appear in situations where the language being translated uses two words where only one English word is needed— for instance, French ‘*ne ... pas*’ might become *not ... not.*)

The files `/data/lang-mod/validate.ex` and `/data/lang-mod/test.ex` contain simulated translations from the Canadian Hansards corpus in the following format:

```
0 [tab] ORIGINAL SENTENCE 1
ERROR [tab] TRANSLATION 1-1
ERROR [tab] TRANSLATION 1-2
...
[blank line]
0 [tab] ORIGINAL SENTENCE 2
ERROR [tab] TRANSLATION 2-1
...
```

For each English sentence, we provide up to 100 proposed translations, each with an error score measuring how bad a translation it is. (The score is

a rough count of the number of errors we inserted.) The first proposed translation for each sentence is the original itself, which has an error score of 0.

Your task is to build a language model that can select translations with low error. To select the best translation according to the language model, rank each proposed translation T_i by:

$$\frac{\log(P(T_i))}{\text{len}(T_i)}$$

(Why must we divide by the length?) To evaluate the different language models you build, you will report their average error (the mean error of the best translation of each sentence).

1 Trivia Assignment: Due Feb 3

Your first task is to compute a baseline score— the score you could obtain without using any language model at all. Report the expected error of a system that picks a translation at random for every sentence. (We plan to give out some support code for reading in the data files in the next day or two.) Email the result to *cs146tas@cs.brown.edu*.

2 Main Assignment: Due Feb 10

Develop your systems using the `validate.ex` file; when you are confident your assignment is working, run it on `test.ex` and report the final scores. Email a written report to *cs146tas@cs.brown.edu*

- The `/data/hansards` directory in the materials distributed with this book contains the data we will use to train and test the language models. (This data comes from the Canadian Hansards, which are parliamentary proceedings and appear in both English and French). These files have one sentence per line, and have been tokenized, i.e., split into words. To start with, we will train language models from the Senate Hansards, i.e., train the English language model using `english-senate-0.txt` as the main training data, `english-senate-1.txt` as heldout training data.
- Use the training set to produce a smoothed unigram model for each language with the smoothing parameter $\alpha = 1$. Compute and report

the log probability of the correct (0 error) translations. What is the model's average error? It's probably not very good— why not?

- Now set the unigram smoothing parameter α to optimize the likelihood of the heldout data as described in the text. What values of α do you find? Repeat the evaluation described in the previous step using your new unigram models. The log probability of the language-specific test data should increase.
- Now construct smoothed bigram models as described in the text, setting $\beta = 1$, and repeat the evaluation. Do these models do better than the unigram models you constructed in the previous step?
- Finally, set the bigram smoothing parameter β as described in the text, and repeat the evaluation. What values of β maximize the likelihood of the heldout data? How do these models compare to the other models you have constructed?
- What kind of errors does your best system make? Report two or three of the worst selected translations, and describe the kind of information that might help to get them right.