Homework #2: Autocorrect

Deadline: 11:59pm Saturday 22 March 2014

Please present a report with all your answers, explanations, and sample images or plots. Submit also a soft copy of the source code and binaries used to generate these results. Please note that copying of any results or source code will result in ZERO credit for the whole homework.

Acknowledgment: This homework is adapted from Chris Manning and Dan Jurafsky's Coursera NLP class from 2012.

In this assignment you will be training a language model to build a spell checker. Specifically, you will be implementing part of a noisy-channel model for spelling correction. We will give the likelihood term, or **edit model**, and your job is to make a **language model**, the prior distribution in the noisy channel model. At test time you will be given a sentence with exactly one typing error. We then select the correction which gets highest likelihood under the noisy-channel model, using your language model as the prior. Your language models will be evaluated for accuracy, the number of valid corrections, divided by the number of test sentences.

Data

We will be using the writings of secondary-school children, collected by David Holbrook. The training data is located in the data/ directory. A summary of the contents:

- **holbrook-tagged-train.dat:** the corpus to train your language models
- **holbrook-tagged-dev.dat:** a corpus of spelling errors for development
- **count_1edit.txt:** a table listing counts of edits $x|w$, taken from Wikipedia. You don't need to modify any code which uses this.

Note that the data files do not contain <s> and </s> markers, but the code which reads in the data adds them.

Your Assignment

Implement the following language models:

- **Laplace Unigram Language Model:** a unigram model with add-one smoothing. Treat out-of-vocabulary items as a word which was seen zero times in training.
- **Laplace Bigram Language Model:** a bigram model with add-one smoothing.
- **Stupid Backoff Language Model:** use an unsmoothed bigram model combined with backoff to an add-one smoothed unigram model

•**Custom Language Model:** implement a language model of your choice. Ideas include interpolated Kneser-Ney, Good-Turing, linear interpolated, trigram, or any other language model you can come up with. You should not train your models on different training data than supplied. Your goal is for your custom language model to perform better than any of the other three language models we ask you to implement.

We have provided you with a uniform language model so you can see the basic layout, located in UniformLanguageModel.py.

To implement a language model you need to implement two functions:

•**train(HolbrookCorpus corpus):** takes a corpus and trains your language model. Compute any counts or other corpus statistics in this function. See the example UniformLanguageModel for how to access sentences in the corpus and the words in those sentences.

•**score(List words):** takes a list of strings as argument and returns the numerical score, which should be the log-probability of the sentence using your language model. Use whatever data you computed in train() here.

Evaluation

Your language models will be evaluated on the development data set. To help with your implementation, we give you the expected performance of each language model on the development set:

•**Laplace Unigram Language Model:** 0.11

•**Laplace Bigram Language Model:** 0.13

•**Stupid Backoff Language Model:** 0.18

•**Custom Language Model:** at least as good as Stupid Backoff, so 0.18.

Note that the performance we expect from your language model is not that great! We have provided you with a very simple edit model, not a lot of training data, and the task is rather difficult. You will receive full credit for implementations which meet the stated thresholds, and a linearly decaying score for accuracy less than the reference.

For more information on the methods discussed in the lectures, or other methods, you can check the paper "An Empirical study of smoothing Techniques for Language Modeling", or check SRILM's man page http://www.speech.sri.com/projects/srilm/manpages/ngram-discount.7.html, or search Google!

Given Code

The rest of the scaffolding has been provided (reading corpora, computed edit probabilities, computing the argmax correction). A short summary of the remaining files:

•**SpellCorrect.py**: Computes the most likely correction given a language model and edit model. The main() function here will load all of your language model and print performance on the

development data, useful for debugging. It may be useful to comment out some of the tests in main() when developing.

•**EditModel.py**: Reads the count_1edit.txt file and computes the probability of corrections. The candidate corrections are all strings within Damerau-Levenshtein edit distance 1. The probability of no correction is set at .9 (Pxx.9). Note that the EditModel isn't great, but your language models will be evaluated using this model, so it won't effect your grade.
•**HolbrookCorpus.py**: Reads in the corpus and generates test cases from misspellings.

•**Sentence.py**: Holds the data for a given sentence, which is a list of Datums. Contains helper functions for generating the correct sentence and the sentence with the spelling error.

•**Datum.py**: Contains two strings, word and error. The word is the corrected word, and error contains the spelling error. For tokens which are spelled correctly in the corpus, error = ''.

Running the code
Execute:

```
$ cd python

$ python SpellCorrect.py
```

This will train all of the language models and output their performance.

Requirements

You are required to implement the code to train your language models and all supporting functions/code. You are required to obtain performance better than the Supidbackoff i.e. 0.18.
Please submit your code and report in one zip file, named CMP462.HW02.bench#.firstname.lastname.zip. For example, if your name is Mohamed Aly and your bench number is 26, your file should be named CMP462.HW02.26.Mohamed.Aly.zip

Grading

4 pts: working code
2 pts: report and submission file name
4 pts: performance better than 0.18.
-1 pt: for any 0.02 loss in performance.