Statistical NLP 2013 - Assignment 1

due September 17 - at 5pm

In this assignment you will build a language model and evaluate its performance on a speech recognition task.

1 Update

The code release included the solution. It is not necessary to hand in a write-up, but please run the experiments and submit your result to Git.

2 Setup

Please send me an email so that I can point you to the Java source code for the course and the data sets needed for this assignment. This is necessary due to licensing restrictions.

Unzip the source files to your local working directory. Some of the classes and packages won't be relevant until later assignments, but feel free to poke around. Make sure you can compile the entirety of the course code without errors. If you get warnings about unchecked casts, ignore them. If you cannot get the code to compile, email me or post to the newsgroup. If you are at the source root (i.e. your current directory contains only the directory 'nlp'), create a directory called 'classes'. You can then compile the provided code with

```
javac -d classes */*/*.java
```

While you can certainly do everything from the command-line and using your favorite text editor, I would recommend using an IDE. Eclipse is easy to setup and makes writing code and debugging it a lot easier. Next, unzip the data into a directory of your choice. For this assignment, the first Java file to inspect is:

```
src/nlp/assignments/LanguageModelTester.java
```

Try running it with:

```
java nlp.assignments.LanguageModelTester -path DATA -model baseline
```

where DATA is the directory containing the contents of the data zip. If everythings working, you will get some output about the performance of a baseline language model being tested. The code is reading in some newswire and building a basic unigram language model that I have provided. This is phenomenally bad language model, as you can see from the strings it generates - you will improve on it.

3 Git and Leaderboard

On the course webpage there is a leaderboard for tracking how well everybody is doing. We will be using public GitHub repositories to manage the submissions. If you are not familiar with *git*, there are plenty of good tutorials online, for example: `http://www.vogella.com/articles/Git/article.html`.

If you don't already have a GitHub account yet, please create one here: `https://github.com/signup/free`. You are welcome to reuse an existing account for this course.

For the master repository `https://github.com/slavpetrov/stat-nlp-nyu`. We will be mainly using the repository for submitting system output, so you will find a skeleton structure for `hw0` and `hw1`.

Please edit `hw0/output.txt` so it contains your name, email address, and a screen name for the course leaderboard and submit it via: `git add output.txt; git commit -m 'my first check-in'; git push;`

Finally, post your repository to Piazza, so that I can make sure it is included in the leader board. Your repository path should be of the form: `https://github.com/username/stat-nlp-nyu`.

4 Description

In this assignment, you will construct several language models and test them with the provided code framework. Take a look at the main method of LanguageModelTester.java, and its output.

Training: Several data objects are loaded by the harness. First, it loads about 1M words of WSJ text (from the Penn treebank, which we will use again later). These sentences have been ”speechified”, for example translating ”$” to ”dollars”, and tokenized for you. The WSJ data is split into training data (80%), validation (held-out) data (10%), and test data (10%). In addition to the WSJ text, the harness loads a set of speech recognition problems (from the HUB data set). Each HUB problem consists of a set of candidate transcriptions of a given spoken sentence. For this assignment, the candidate list always includes the correct transcription and never includes words unseen in the WSJ training data. Each candidate transcription is accompanied by a pre-computed acoustic score, which represents the degree to which an acoustic model matched that transcription. These lists are stored in SpeechNBestList objects. Once all the WSJ data and HUB lists are loaded, a language model is built from the WSJ training sentences (the validation sentences are ignored entirely by the provided baseline language model, but may be used by your implementations for tuning). Then, several tests are run using the resulting language model.

Evaluation: Each language model is tested in two ways. First, the harness calculates the perplexity on two different held-out sets. This number can be useful for evaluation, but can be misleading. If you have a bug (and your probabilities do not sum to 1), you will get wonderful perplexities (but potentially terrible word error rates). If you don’t have any bugs, higher order n-gram models and more sophisticated smoothing techniques will result in lower perplexities.

The harness will first calculate the perplexity of the WSJ test sentences. In the WSJ test data, there will be unknown words. Your language models should treat all entirely unseen words as if they were a single UNK token. This means that, for example, a good unigram model will actually assign a larger probability to each unknown word than to a known but rare word - this is because the aggregate probability of the UNK event is large, even though each specific unknown word itself may be rare. To emphasize, your model’s WSJ perplexity score will not strictly speaking be the perplexity of the exact test sentences, but the UNKed test sentences (a lower number). The harness will then calculate the perplexity of the correct HUB transcriptions. This number will, in general, be worse than the WSJ perplexity, since these sentences are drawn from a different source. Language models predict less well on distributions which do not match their training data. The HUB sentences, however, will not contain any unseen words.

More importantly, the harness will also compute a word error rate (WER) on the HUB recognition task. The code takes the candidate transcriptions, scores each one with the language model, and combines those scores with the pre-computed acoustic scores. The best-scoring candidates are compared against the correct answers, and WER is computed. The testing code also provides information on the range of WER scores which are possible: note that the candidates are only so bad to begin with (the lists are pre-pruned n-best lists). You should inspect the errors the system is making on the speech re-ranking task, by running the harness with the -verbose flag.

Finally, the harness will generate sentences by randomly sampling from you language models. The provided language models outputs arent even vaguely like well-formed English, though yours will hopefully be a little better. Note that improved fluency of generation does not mean improved modeling of unseen sentences.

Experiments: You will implement several language models, though you can choose which specific ones to try out. Along the way you must build the following:

- Start by building a bigram language model using an interpolation method of your choice
- Extend you bigram model to a trigram model
- Implement a Kneser-Ney estimator (bigram ok)

As you increase the complexity of your language model, it may be that lower perplexity, especially on the HUB sentences, will translate into a better WER, but don't be surprised if it doesnt. The actual performance of your systems does not directly impact your grade on this assignment, though I will announce students who do particularly interesting or effective things.

To submit results to the leaderboard, edit `hw1/output.txt` and replace the GUESS and GOLD lines with what the harness prints when you set the -verbose flag. You can then push your changes to the repository and the leaderboard will automatically update shortly after. Feel free to submit results as often as you want.

With an interpolated bigram model ($\lambda = 0.6$) you should be able get the word error rate down to 7.3% with a perplexity of around 480. An interpolated trigram model reduces the word error rate further to 6.9%. In both cases I simply reserved 1/totalCount for words not in the vocabulary.

Random Advice: In nlp.util there are some classes that might be of use - particularly the Counter and CounterMap classes. These make dealing with word to count and history to word to count maps much easier.

5 Write-ups

What will impact your grade is the degree to which you can present what you did clearly and make sense of what is going on in your experiments using thoughtful error analysis. When you do see improvements in WER, where are they coming from, specifically? Try to localize the improvements as much as possible. Some example questions you might consider: Do the errors that are corrected by a given change to the language model make any sense? Are there changes to the models which substantially improve perplexity without improving WER? Do certain models generate better text? Why? Similarly, you should do some data analysis on the speech errors that you cannot correct. Are there cases where the language model is not selecting a candidate which seems clearly superior to a human reader? What would you have to do to your language model to fix these cases? For these kinds of questions, it is actually more important to sift through the data and find some good ideas than to implement those ideas. The bottom line is that your write-up should include concrete examples of errors or error-fixes, along with commentary.

For this assignment, you should turn in a write-up of the work you have done, but not the code (it is sometimes useful to mention code choices or even snippets in write-ups, and this is fine). The write-up should specify what models you implemented and what significant choices you made. It should include tables or graphs of the perplexities, accuracies, etc., of your systems. It should also include some error analysis - enough to convince me that you looked at the specific behavior of your systems and thought about what it is doing wrong and how you would fix it. There is no set length for write-ups, but a ballpark length might be 3-4 pages, including your evaluation results, some graphs, and some interesting examples. I am more interested in knowing what observations you made about the models or data than having a reiteration of the formal definitions of the various models.

While typesetting your write-up in Latex is not a requirement, it can be an useful exercise and will come in handy for later assignments (and the project) that might involve writing down some formulas.

You can email me your write-up in PDF format before the class on the due date or bring a print-out to class.