Assignment 1: Language Models

Anand Mohan - SR No: 14417 - M Tech(SE)
moghan.anand@gmail.com

1 Methodology

1.1 Language Models

Models that assign probabilities to sequences of words are called language models or LMs. An interpolated model of Unigrams and Bigrams are used for as the base language model for finding the perplexity in different settings. Trigrams are also included in generating sentences model.

1.2 Data Splits and Tokenization

Data in each setting is divided into train set, held-out set and test set in the ratio 8:1:1 based on the number of sentences. Tokenization is done on the data sets and some unnecessary characters and digits were removed. Also tokens <*s*> and </*s*> are added at the start and end of sentences. <*UNK*> is used to represent any tokens that are not present in the vocabulary.

1.3 Smoothing

In order to accommodate the n-grams not in your language model, well have to take off a bit of probability mass from some more frequent events and give it to the n-grams not in occurring in the training set. I have used *absolute discounting* (subtracting a fixed (absolute) discount d from each count) in my language model.

$$d = \begin{cases} 0.5 & \text{if count } = 1 \\ 0.75 & \text{if count } > 1 \end{cases}$$

1.4 Metric

Perplexity: The perplexity (PP) of a language model on a test set is the inverse probability of the test set, normalized by the number of words.
For a test set $W = w_1 w_2 ... w_N$:

$$PP(W) = P(w_1 w_2 ... w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_N)}}$$

1.5 Sentence Generation

Sentences are generated using a combination of Trigram, Bigram and Unigram model. The sentence starts from the START token <*s*> and the next word is generated by the model. The model takes the top five probable words and assigns it randomly, so that there a variety of sentences will be generated each time you run.

2 Results

D1 - Brown Corpus
D2 - Gutenberg Corpus

- S1: Train: D1-Train, Test: D1-Test
  Perplexity = 289.48
- S2: Train: D2-Train, Test: D2-Test
  Perplexity = 155.13
- S3: Train: D1-Train + D2-Train, Test: D1-Test - Perplexity = 152.78
- S4: Train: D1-Train + D2-Train, Test: D2-Test - Perplexity = 103.91

Sample Generated Sentence: He is into a standing position in the looks.