Introduction to NLP (CS7.401)

Spring 2022, IIIT Hyderabad
Assignment 1 Report

Abhinav S Menon (2020114001)

Perplexities

The average perplexities of the models on the different sets are as follows.

Model	Trainset	Testset
LM 1	80.44468448137053	162429158.7184405
LM 2	5.133071978651511	835376.8607449322
LM 3	114.60145540433574	195.9678043633212
LM 4	3.8572394985083975	261.2306081115929

Analysis

Perplexities

The basic expectation we have is that the perplexities of the testset ought to be higher than those of the training set. This is borne out by the observations. However, the actual magnitudes of the differences vary considerably. We notice two main patterns.

First, the models trained using Kneser-Ney smoothing (LMs 1 and 3) return a much higher perplexity on the training set than those trained using Witten-Bell smoothing (LMs 2 and 4). This correlates with the empirical evidence of Witten-Bell performing better than Kneser-Ney.

Second, the models using the Europarl corpus (LMs 1 and 2) return much, much higher perplexities on their test sets. This is possibly because the medical corpus has a large number of low-frequency tokens, which means that many of them are treated as `<UNK>` (see below). The Europarl corpus, however, returns high perplexities due to low-frequency n-grams, which are treated normally. The high perplexities in these cases tend to be caused by a small number of sentences (*e.g.* in LM1 and LM2, only three sentences have perplexities exceeding 10^5).

Computation

The code replaces the count of unseen n-grams with a uniform frequency of 2. Words in the training set occurring less than or equal to five times are treated as `<UNK>`, and all unseen words in the test set are considered to be `<UNK>`.