Assignment 1 Write Up

Kit Lao

January 26th, 2020

Part 1

n-gram	Training	Testing	Development
Uni-gram	976.544	896.499	892.247
Bi-gram	77.073	∞	∞
Tri-gram	6.444	∞	∞

The perplexity numbers are the average branching factor for predicting the next word in a sentence. For uni-grams, the perplexities are generally high because for every sentence in the data set, the probability of it is calculated as a product of chances for each word occurring, based on the training data. For any n-grams where n is greater than 1, the perplexity decreases at an exponential level because the probability of the next word is based on it's previous $n-1$ words. The perplexities for uni-grams for testing and development is a bit lower than training because there are more unknown-words in them, due to how the data is pre-processed. Bi-gram and tri-gram for testing and development are infinity because there are certain grams that doesn't exist in the training data, resulting in a probability of 0.

Part 2

1. This part answers 1 and 2.
   Row 1: For training, the perplexity is closest to tri-gram perplexity. This is due to the fact that the weight for the probabilities for the tri-grams are the highest, and non-zero. For testing and development, the perplexities aren't infinity due the fact that there is always a non-zero probability added together from the uni-grams. Since there are zero-probabilities in bi-grams, and especially tri-grams, there will be a lot of probabilities used in the smoothed perplexity that will have no weight from the bi-gram and tri-gram probabilities, which causes the perplexity to somewhere between all three, rather than closer whichever gram has more weight. The probabilities are zero due to the fact that grams don't appear in the training data.
   Row 2: For training, the perplexity is a bit higher than the one on row 1,

due to the weights being more distributed than row 1, but is still closer to the tri-gram perplexity due to the fact that most of the probabilities are higher in bi-gram and especially tri-gram. For testing and development, the perplexities are a bit lower due to less weight on the tri-grams. Explanation is similar to row 1.
Row 3, 4, 5: For training, the perplexity decreases as less weight is applied to the n-gram with the lowest probabilities. For testing and development, explanation is similar to row 1.

$(\lambda_1, \lambda_2, \lambda_3)$	Training	Testing	Development
$(0.1, 0.3, 0.6)$	11.104	349.264	350.468
$(0.33, 0.33, 0.34)$	16.777	273.206	273.553
$(0.8, 0.1, 0.1)$	44.412	352.923	352.222
$(0.1, 0.8, 0.1)$	27.960	290.907	291.279
$(0.1, 0.1, 0.8)$	9.291	477.938	479.794

3. For training, using half of the training data resulted in increased perplexity of the uni-grams and decreased perplexity of bi-gram and tri-gram. For uni-grams, since the data is cut in half, the number of words in the corpus is about half of what it's supposed to be, but the frequency for each most of the words would not decrease at the same ratio, and the unique word count is about two thirds. Since the decrease in word count is proportionally greater than the decrease in unique word count, the frequencies for each word is generally smaller, causing the probabilities to be lower, and perplexity to be higher. For bi-gram, the frequency for both a uni-gram and bi-gram decreased due to word count decrease, but the frequency for uni-gram decreased a bit more from the unique word count decreasing proportionally less than the word count. Similar reason for tri-gram. For testing and development, since unique word count is smaller, your training data is smaller, resulting in more words in your testing and development data represented as unknown-words, which causes the probability of the unknown-word word to be higher, which results to overall probability to be higher, when then causes perplexity to be lower. The bi-gram and tri-gram for testing and development is still infinity due to some grams not appearing in training data.

	Training	Testing	Development
Uni-gram	1020.970	698.109	695.773
Bi-gram	74.246	∞	∞
Tri-gram	5.797	∞	∞

4. When the unknown-word threshold is increased, either more noise is blocked out, or certain valuable data is being removed from the training. Since the

threshold is only increased from 3 to 5, it is likely that more noise is blocked out. This outcomes will cause the uni-gram perplexity for training to decrease because unknown-words will have higher frequency, and more low-frequency words will be replaced by the frequency of the unknown-word. When more are replaced with an unknown-word, this causes the unique token count to decrease, which causes more tokens from the testing and development to be replaced with unknown-words. Since the testing and development data has more unknown-words, and frequency for unknown-words in the training data is increased, their will be higher probabilities used in calculating the their perplexities. More unknown-words in training data will result to a slightly higher chance of a bi-gram with an unknown word in one or the other position, or in both. This results certain probabilities of a bi-gram involved in the calculation of the perplexity to be slightly higher, resulting in slightly lower perplexity. But for the specific case where the unknown-word is in the first position of the bi-gram, the probability of it could be lower, due to the increased amount of the particular unknown-word uni-gram. For tri-grams in training, there will be a slight increase in occurrences of a tri-grams where an unk-word appears. But certain tri-grams where the unknown-word appears in the first 2 positions, there probabilities are lowered due to the increased occurrences of bi-grams with at least an unknown-word. This will cause the perplexity to decrease slightly. The infinities appearing because of reasons mentioned in this write-up.

	Training	Testing	Development
Uni-gram	803.485	756.689	754.300
Bi-gram	75.633	∞	∞
Tri-gram	7.112	∞	∞