Add-1 and Kneser-Ney Smoothing in Bigram Models

Daniela Verratti Souto (5692183)

January 28th, 2022

1 Perplexity

After completing the project, we registered the following perplexities and OOV rates for the test data (see table 1).

	ja-sas-test1.txt	ja-sas-test2.txt	ja-pap-test.txt
Add-1	359.345	331.660	313.815
Kneser-Ney	89.326	83.960	79.561
OOV Rate	8.408 %	8.488 %	9.064 %

Table 1: Perplexity values for each test file with each model and OOV rate for each file.

With regard to perplexity, most of our expectations were met. Firstly, we were aware that the perplexity values obtained with a model using Laplace smoothing should be considerably high. This is due to the fact that Add-1 discounts too much probability from seen events to account for unseen events, and does so blindly. This is a very rudimentary approach which leads to underestimating the probabilities of the perfectly valid sentences which occur in the test data. From this, an increase in perplexity is to be expected.

In comparison, the Kneser-Ney model produced much better perplexity values for the test data since it approaches the problem in a more sophisticated manner: discounting by a certain amount (d), distributing the discounted probability mass and accounting for how likely it is for the second word of our bigram to appear in a new context. This results in higher probabilities for more likely sentences, like the ones we have in our test sets.

2 Sentence generation

In terms of the sentences generated, the difference between both smoothing approaches is apparent: whereas Laplace smoothing yields very random words and extremely long sentences (whose length we have decided to limit), Kneser-Ney produces relatively coherent phrases (*e.g.* "She said", or "He replied"). However, as was to be expected from a bigram model, complete sentences rarely have a full meaning.

Perhaps what is more difficult to understand is why the perplexity of *Pride and Prejudice* is lower for both models. We would have expected it to be the other way around on account of it coming from a different book. In order to answer this, it would be safer to compare the test sets and perhaps also carry out some significance testing to see whether this difference is significant or a product of chance.

3 OOV rate

It is also unsurprising to find that the OOV rate is higher for *Pride and Prejudice* (see table 1). As it is a different novel, some of the difference could be accounted for by the difference in character names, places and topics the writer focused on. However, the difference is not dramatic, because after all, *Pride and Prejudice* and *Sense and Sensibility* are both novels written by Jane Austen which deal with similar topics, take place during the late 18^{th} and early 19^{th} centuries and feature characters from the English middle class of the time. As such, the vocabulary used is quite uniform. Finally, the fact that the test data is much shorter than the training set also explains the low rate.

4 Final remarks

Prior to working on the project, we had some expectations based on mostly theoretical sources. Having finished, we have more accurate impressions of what the practical counterpart entails. For example, we know what values to expect from the perplexity of a corpus and what they imply. We also have a better sense of what would happen should we test on a corpus that is less representative of our data, such as a different genre. This would have resulted in much higher values for perplexity and OOV.