NTMI - Project Exercises - Part A3

Alex Khawalid (10634207)
Wessel Klijnsma (10172432)
Winand Renkema (10643478)

February 16, 2016

Introduction

The third step of Part A of the assignments for the natural language and interfaces course is to implement different smoothing methods and compare these methods to each other. This assignment is about smoothing methods. Smoothing is used to determine frequencies of unseen events, which are 0 in maximum likelihood estimator.
Using the bi-gram language model and various smoothing methods, the probability of a sentence occurring is calculated. The methods to calculate the sentence probabilities are: no smoothing, add one smoothing, and Good-Turing smoothing. In the sections below the exercises will be discussed.

Method

Commands to run:

```
1: python a1-step3.py -training-corpus austen.txt -n 2 -test-corpus ja-pers-clean.txt
2: python a1-step3.py -training-corpus austen.txt -n 2 -test-corpus ja-pers-clean.txt -s add1
3: python a1-step3.py -training-corpus austen.txt -n 2 -test-corpus ja-pers-clean.txt -s gt
```

Smoothing

Add-one smoothing

Add-one smoothing is a simple method for smoothing n-gram frequencies. The process is to add one to all the frequencies. So an unseen will have a frequency of 1. We can directly apply this method for calculating n-gram probabilities. The following formula is used (for bi-grams in this case):

$$P^*(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)+1}{C(w_{n-1})+V}$$

Here C is the n-gram frequency. V is the amount for n-grams (vocabulary size). It is necessary to add V to the denominator to make sure all the probabilities add up to 1.

Good Turing smoothing

The intuition of Good-Turing smoothing is that the mass of probability is based on the frequency of things that occur once. Good-Turing smoothing is only applied to n-grams with occurrences $\leq k$. Good-Turing smoothing is applied using the following formula where $1 \leq occurrences \leq k$:

$$r^* = \frac{(r+1)\frac{n_r+1}{n_r} - r\frac{(k+1)n_{k+1}}{n_1}}{1-\frac{(k+1)n_{k+1}}{n_1}}$$

For n-grams with occurrences$=0$, mass is distributed uniformly.

Results

The percentage of sentences that are assigned a zero probability is zero for the smoothed models. This is because there are no events of which the frequency is zero. The percentage of zero probability sentences for the unsmoothed model is 96.75%
These are the first five sentences that are assigned a zero probability by the unsmoothed model:

```
[START] Persuasion by Jane Austen 1818 [END]
[START] Sir Walter Elliot of Kellynch Hall in Somersetshire was a man who [END]
[START] for his own amusement never took up any book but the Baronetage [END]
[START] there he found occupation for an idle hour and consolation in a [END]
[START] distressed one there his faculties were roused into admiration and [END]
```

Conclusion

Using smoothing methods enables a more meaningful calculation of sentence probability. The occurrence of an n-gram that has not been seen before will, without smoothing , make the probability of a sentence 0. With the use of add-one smoothing or Good-Turing smoothing, one unknown n-gram does drop the probability to 0 percent. The difference between add-1 smoothing and Good-Turing smoothing is that Good-Turing distributes mass uniformly. Add-one smoothing just gives unknown n-grams the probability of an n-gram occurring once in the corpus. However add-one is faster as it requires less computation, both methods have pros and cons