Exercise 3

Dejan Porjazovski

November 18, 2018

1 Question 1

1.1 Fetch the 1-gram counts again and compute the maximum likelihood estimates for the following 1-gram probabilities by hand (you can use "octave" or "matlab" as a calculator)

The number of $< s >$ is 3 but it's being ignored because it is always. assumed in the beginning of the sentence.
Likelihood estimates $P(in) = 0.136$
Likelihood estimates $P(a) = 0.045$
Likelihood estimates $P(< /s >) = 0.136$

2 Question 2

2.1 Use ngram-count to get necessary counts and compute the following 2-gram and 3-gram estimates (maximum likelihood) below by hand. Note that the notation $P(bag|in\, the)$ means the probability that word "bag" appears after "in the" (for example in the sentence "in the bag")

bigrams:
Likelihood estimates $P(the|is) = 0.333$
Likelihood estimates $P(box|is) = 0$
Likelihood estimates $P(is|is) = 0$
Likelihood estimates $P(in|is) = 0.666$
Likelihood estimates $P(bag|is) = 0$
Likelihood estimates $P(< /s > |is) = 0$
Likelihood estimates $P(it|is) = 0$
Likelihood estimates $P(a|is) = 0$
Likelihood estimates $P(on|is) = 0$

trigrams:
Likelihood estimates $P(the|in\, the) = 0$
Likelihood estimates $P(box|in\, the) = 0$
Likelihood estimates $P(is|in\, the) = 0$
Likelihood estimates $P(in|in\, the) = 0$
Likelihood estimates $P(bag|in\, the) = 0.5$
Likelihood estimates $P(< /s > |in\, the) = 0$
Likelihood estimates $P(it|in\, the) = 0$
Likelihood estimates $P(a|in\, the) = 0$
Likelihood estimates $P(on|in\, the) = 0$

3 Question 3

3.1 Using interpolated absolute discounting (D=0.5) compute $P(in|is)$ and $P(</s>|is)$ by hand

Calculating the probabilities by hand, I got the following results:
$P(in|is) = 0.545$
$P(</s>|is) = 0.045$

Using the command **'ngram -lm 2gram.lm -ppl test.txt -debug 2'** we can see that the values computed by hand match the probabilities computed with the above command.

3.2 Compare to results you got in Question 2

For $P(in|is)$, in question 2, I got probability of 0.666, compared to 0.545 that I got using the interpolated absolute discounting.
For $P(</s>|is)$, in question 2, I got probability of 0, compared to 0.045 that I got using the interpolated absolute discounting.

4 Question 4

4.1 What are the log-probabilities of the above sentences?

The log probability for the first sentence is: -3.35094
The log probability for the second sentence is: -2.83334
The log probability for the third sentence is: -2.51316

4.2 Which sentence is the most probable one according to the model?

According to the model, the third sentence is the most probable with log probability of -2.51316

4.3 Give an example of a sentence (non-empty, no out-of-vocabulary words) whose probability is even higher than any of the above.

An example sentence: "the box", with probability of: -1.39936

5 Question 5

5.1 Which of the models gave the best probability for the test data?

1-gram model has probability of -264042
2-gram model has probability of -237930
3-gram model has probability of -235817

The 3-gram model gave the best probability for the data.

5.2 What is the proportion of out-of-vocabulary (OOV) words in the test data (the ngram tool prints the relevant information for this)?

The proportion of out-of-vocabulary data is: 27097.

6 Question 6

6.1 Which of the models is the best one?

1-gram morph model has probability of: -662489
2-gram morph model has probability of: -503229
3-gram morph model has probability of: -465729

The 3-gram morph model gave the best probability for the morph data.

6.2 What was now the number of OOV morphs (the tool talks about words since it knows nothing about morphs)?

The proportion of out-of-vocabulary morph data is: 0