NLP – EX1

Theoretical:

1. Let’s mark all sentences as S,

From law of total probability:

Observe

Which according the Bigram model, is equal to:

We can assume that the total number words in the corpus are finite, therefore, the total distinct pairs of sequenced words is finite as well.

On that basis, there exists some s.t.

Hence:

And

Which means that

* 1. Unigram model:

Where

and

Spelling corrector model:

Given a sentence, generate all possible combinations of the sentence with ‘where’ and ‘were’, and return the generated sentence with max probability in unigram model.

If the word ‘where’ appeared more times in the corpus than the word ‘were’

The sentence with the highest probability will be:

He went where there where more opportunities.

(And vice versa)

A mixture of both options won’t receive a high probability in this model, and thus it will never give the correct sentence the highest probability (unless both appear the same number of times, under which the behavior in undefined).

* 1. Bigram model:

Spelling corrector model:

Given a sentence, generate all possible combinations of the sentence with ‘where’ and ‘were’, and return the generated sentence with max probability in Bigram model.

This model can give the correct sentence the highest probability, because it is likely that the combination ‘there were’ appears more times in a corpus than ‘there where’. The Same goes for combinations ‘went where’ and ‘went were’ respectively.

We risk the possibility that a given two-word sequence in the sentence will not appear in the corpus, and in such a scenario the sentence probability will be 0, and the model will perform badly.

1. 1. We need to show that:

Moving

* 1. According to Add One Laplace smoothing, the smoothed frequency of a word that appears in the corpus times is:

where is the number of word types in corpus.

Assuming that the frequency MLE is

Let’s find a threshold , s.t. , else :

Hence

* 1. Let’s examine when the Good Turing frequency smoothing estimator is smaller than MLE:

This is satisfied iff:

Thus, if the number of all instances of words that appear c times Is larger than the number of all words instances that appear c+1 times the smoothed frequency is lower, else its higher or equal to the MLE, and since this can change intermittently between consecutive pairs, the threshold does not necessarily exist.

1. 1. Trigram model:

The made assumption is that each word’s appearance is only dependet on the two previous words, and in independent from the rest.

* 1. In case of emergency, the child hastily fastens the safety belt.

הילדה החכמה הרצתה.

* 1. In case of emergency, children should hastily fasten the safety belts.

הכלבה שראינו אתמול נובחת.

We would need a 4-Gram model for capturing subject-verb agreement in both sentences.

1. Pairs:

ראיתי איש הלך.

Triplets:

ראיתי איש הולך ראיתי.

4-Tuples:

ראיתי איש הולך למכולת ראיתי.

As we use a higher degree Markov model, we are more likely to generate correct sentences using the model, or to give a higher probability to a correct sentence. However, there is a tradeoff with corpus size and computation time.

Practical:

2. most probable word in sentence “I have a house in”, as predicted by bigram model is:

“the”

3. Bigram:

a. Prob for “Brad Pitt was born in Oklahoma”:

Prob for “The actor was born in USA”:

b. Perplexity:

4. Interpolation of Unigram and Bigram:

a. Prob for “Brad Pitt was born in Oklahoma”:

Prob for “The actor was born in USA”:

b. Perplexity: