Assignment 1: Kneser Ney Implementation

Natasa Farmaki - DS3517018

Dimitris Georgiou - DS3517004

Stratos Gounidellis - DS3517005

Course: Text Engineering & Analytics

Professor: Ion Androutsopoulos

Athens, Feb 2018

INTRODUCTION

The aim of this assignment is to implement a natural language model(NLP) based on a corpus from the European Parliament in order to extract probabilities for a sentence based on bigrams and trigrams. Then given a sentence we will predict the next word and based on these predictions and on efficiency measures we will compare the predictive ability of bigram and trigram models. Also, in case a word is not correctly spelled the model will suggest the most probable word based on the probabilities extracted and the edit distance. For this assignment the NLP algorithm implemented is Kneser Ney as it is much more efficient than Laplace.

You can find the python code in the following github link:

https://thinkingtea.github.io/repoTEA/

PREPROCESSING OF THE CORPUS

As mentioned before the corpus used was a text from the European parliament. In order to test the algorithm we first split the text into a training and a test set with a proportion of 77%-33%.

The first issue to tackle in the training corpus was the words that do not appear often. In our case the words appearing less than ten times within the training corpus are replaced with “UNK”.In order to do that we tokenized the training set, calculaed the frequency of each word using thefdist function provided by nltk and replaced the words for which the freuquency was less than ten.

To extract the probabilities of the bigrams and trigrams we need to introduce a “start1”, ”end12” word at the start and the end of each sentence for the bigrams and a “start1”, “start2”, ”end12” for the trigrams. To achieve this, we created two copies of the initial training set. The original will be used as is for the unigram probabilities and the copies will be processed for the bigram and the trigram probabilities.

It is important to note that the process of replacing the infrequent words is time consuming and thus we have already ran the code and stored the results in the files “unigram.txt”, “bigram.txt,”trigram.txt”. Sosince the files are already created, we can tokenize the text and implement the Kneser Ney algorithm to calculate the smoothed probabilities for both the bigram and the trigram model.

KNESER NEY ALGORITHM

Kneser–Ney smoothing is a method primarily used to calculate the probability distribution of n-grams in a document based on their histories. ReinhardKneser and Hermann Ney proposed the method on 1995. More specifically, it uses absolute discounting by subtracting a fixed value fromthe probability's lower order terms to omit n-grams with lower frequencies. For this reason, it is considered one of the most effective smoothing techniques both for lower and higher order n-grams.

An example that could illustrate the concept behind this method is the frequency of the bigram "European Union" in our corpus. Suppose this phrase is abundant in a given training corpus. Then the unigram probability of the unigram "Union" will also be high. If we unwisely use something like absolute discounting interpolation in a context where our bigram model is weak, the unigram model portion may take over and lead to some strange results. In other words, relying on only the unigram frequency to predict the frequencies of n-grams leads to skewed results. In contrast, Kneser–Ney smoothing aims to fix this problem by considering the frequency of the unigram in relation to the possible tokens preceding it.

The interpolated recursive Kneser-Ney smoothed bigram model is as follows:

$$PKN(w2|w1)=\frac{\max\{c(w1w2)-D,0\}}{c(w1)}+D*\frac{N(w1\cdot)}{c(w1)}*\frac{c(w1w2)}{N(\cdot\cdot)})$$

The interpolated recursive Kneser-Ney smoothed trigram model is as follows:

$$PKN(w3|w1w2)=\frac{\max\{c(w1w2w3)-D,0\}}{c(w1w2)}+D*\frac{N(w1w2\cdot)}{c(w1w2)}*(\frac{\max\{N(\cdot w2w3)-D,0\}}{N(\cdot w2\cdot)}+D*\frac{N(w2\cdot)}{N(\cdot w2\cdot)}*\frac{N(\cdot w3)}{N(\cdot\cdot)})$$

PREDICTING USING KNESER NEY

Predictive Keyboard

Predictive keyboard is a keyboard with the power to predict the next word. Given a sentence we focus mainly on the last part of it (i.e mostly the last four words) and we predict the next word. If the last token does not exist in the vocabulary of the trained model (ex. a word that is not yet completed), we utilize the edit distance and among the closest words, the most probable bigrams or trigrams are chosen. Obviously, in that case the outcome is not an extra word or a prediction but a word that could possibly replace the last word of the given sentence.

Although we implemented edit distance using dynamic programming, we used the implementation of nltk for efficiency reasons and we just set the substitution cost to two. That way, we penalize more words that do not contain the letters existing in the last token of the given sentence, while with that approach we simulate better real case scenarios. The implementation of the edit distance algorithm can also be found in the github repository.

Finally, it should be mentioned that both approaches utilize the bigram and the trigram model. Both models calculate the smoothed probabilities and then the top three words associated with the highest probabilities are returned.

	correct_sentence	logProb_cs	wrong_sentence	logProb_ws
0	Yes, I totally agree: let us set challenging t...	-71.970115	let I confuse with compliance but let totally ...	-74.484186
1	The allocation of the budget to the Member Sta...	-161.105469	every country cofinancing The Member different...	-167.246107
2	In Greece, the dangers come from the exploitat...	-83.185418	in from catchment the dangers exploitation Bul...	-84.050367
3	Has that been checked, before an emergency inc...	-41.121926	been incident that Has before checked emergenc...	-25.796982
4	That was, I think, one of the most important s...	-233.773617	most talking about important to has that a mad...	-295.163526
5	In order to ensure that there is no misunderst...	-135.855826	order use like fossil to for In is and I our e...	-181.456244
6	We are delighted that we will be welcoming a S...	-116.676106	will once a South be Joint we Sudanese has Ass...	-118.811617
7	We have to revert back to peace mediation with...	-48.881705	We mediation without or revert peace to winner...	-41.737315
8	(HU) Ladies and gentlemen, in the course of it...	-142.581948	Central wound of crisis 2008 as its into in 20...	-153.908244
9	It is also worth emphasising the role played b...	-114.506859	by development is areas also role by and in em...	-147.419297
10	Hence we all - MEPs and ministers in the regio...	-87.093408	and federal all are we the feel and behind reg...	-108.355207
11	Rather, they bring clear, measurable benefits ...	-33.085304	citizens benefits clear measurable bring Rathe...	-27.681346
12	in writing. - (NL) The Dutch People's Party fo...	-95.421112	2012 for The NL an is People and Democracy s a...	-108.621463
13	It is left to the Member States and peer revie...	-42.381679	It peer Member the monitoring is States suppor...	-80.838691
14	However, of what use are the euro and the Euro...	-102.494285	promote the if use and and they are of do resp...	-94.669496
15	In the current economic context, we could desc...	-212.477361	the of describe Greece country in context of o...	-162.078378
16	The German Government is currently conducting ...	-139.892821	can its we and multiannual Government so end d...	-128.057071
17	I would also like to thank the President of th...	-29.519224	to I the European would the of thank President...	-60.740351
18	The vote will take place today at 11:30.\n	-32.925108	30 11 vote today The will at place take	-29.891449
19	This issue cannot be tackled appropriately by ...	-83.973544	be a be discussed appropriately cannot by alon...	-81.203198
20	(SL) I am in favour of Croatia's membership of...	-90.167675	favour Union of but am SL I European of Croati...	-128.724221
21	Finally, it is important for us to bear in min...	-94.718968	mind the safety other to the of coal branches ...	-96.385654
22	Moreover, I very much appreciate Turkey's posi...	-45.145911	Turkey very positive appreciate the in role Ca...	-45.756501
23	However, to quote a popular Hungarian saying, ...	-105.913375	unless horseshoes as quote a Hungarian is dead...	-130.420393
24	The President of the Republic of Lithuania too...	-63.718232	by amendments immediately tabling Lithuania Re...	-47.424333

Figure1 Predictions for bigram model

	correct_sentence	logProb_cs	wrong_sentence	logProb_ws
0	Yes, I totally agree: let us set challenging t...	-61.003979	targets confuse agree not let but let set comp...	-85.747118
1	The allocation of the budget to the Member Sta...	-126.407836	into of every budget cohesion take the capacit...	-143.480522
2	In Greece, the dangers come from the exploitat...	-58.314793	dangers come basin the catchment from exploita...	-73.170688
3	Has that been checked, before an emergency inc...	-39.554890	Has checked incident been occurs an before eme...	-48.573466
4	That was, I think, one of the most important s...	-188.840725	important been statement the that the politica...	-216.429279
5	In order to ensure that there is no misunderst...	-103.991229	our to many environment I use on fossil impact...	-147.863137
6	We are delighted that we will be welcoming a S...	-93.932756	South the has signed we are a We parliamentari...	-136.304019
7	We have to revert back to peace mediation with...	-48.864177	losers to to back winners have without revert ...	-52.359325
8	(HU) Ladies and gentlemen, in the course of it...	-108.391608	gentlemen September with in wound as in was it...	-150.077836
9	It is also worth emphasising the role played b...	-101.593907	worth economic sustaining areas the promoting ...	-138.408540
10	Hence we all - MEPs and ministers in the regio...	-83.708616	the and facts MEPs we federal feel in Hence we...	-93.226107
11	Rather, they bring clear, measurable benefits ...	-35.871112	clear for bring benefits citizens Rather measu...	-38.712548
12	in writing. - (NL) The Dutch People's Party fo...	-58.366926	opposed s VVD in 2012 NL for the Democracy wri...	-107.855548
13	It is left to the Member States and peer revie...	-46.282331	Commission peer supported by by the States rev...	-67.480427
14	However, of what use are the euro and the Euro...	-67.516882	do of Eurogroup not responsibility and what if...	-87.115933
15	In the current economic context, we could desc...	-190.976426	exaggeration in the without current largest ec...	-198.986092
16	The German Government is currently conducting ...	-115.153562	this have can until wait so will section begin...	-156.372036
17	I would also like to thank the President of th...	-14.758040	European I the also of President Commission li...	-24.868229
18	The vote will take place today at 11:30.\n	-16.166811	11 The today will place 30 take vote at	-33.292053
19	This issue cannot be tackled appropriately by ...	-67.280393	cannot This be level tackled appropriately nee...	-81.799098
20	(SL) I am in favour of Croatia's membership of...	-61.462528	interests SL Croatia but European membership U...	-104.813586
21	Finally, it is important for us to bear in min...	-62.997376	workers safety is branches in it and of other ...	-103.723818
22	Moreover, I very much appreciate Turkey's posi...	-39.778100	very Turkey role s positive Caucasus Moreover ...	-52.643484
23	However, to quote a popular Hungarian saying, ...	-86.541864	a much will be unless as it on is quote worth ...	-106.245592
24	The President of the Republic of Lithuania too...	-60.922060	The of amendments tabling Republic took of Pre...	-60.798148

Figure2 Predictions for trigram model

EFFICIENCY MEASURES

The most common metric for evaluating the performance of a language model is the probability that the model assigns to test data, or the derivative measures of cross-entropy (or just entropy) and perplexity. As the cross-entropy of a model on test data gives the number of bits required to encode that data, cross-entropy is a direct measure of application performance for the task of text compression.

Entropy isalso a measure of information. Given a random variable X ranging over whatever we are predicting and with a particular probability function. The entropy is useful when we don't know the actual probability distribution p that generated some data. It is generally assumed that lower entropy correlates with better performance.

There are many parameters that could be optimized. The way to test the model is testing the perplexity on the validation corpus. Perplexity is a measurement of how well a probability distribution or probability model predicts a sample. We wrote a routine to calculate the perplexity and we tested it on the validation corpus.

As mentioned earlier, we evaluate smoothing methods through their cross-entropy on test data over a variety of training set sizes for both bigram and trigram models. We see that cross-entropy decreases steadily as the training set used grows in size, this decrease is somewhat slower than linear in the logarithm of the training set size. Furthermore, we see that the entropies of different corpora can be very different, and that trigram models perform substantially better than bigram models only for larger training sets.

The entropy of a bigram model is a little bigger, we are weighting the number of times that a word x occurs first in a bigram out of all the words in the text, but then we have the weighted probabilities of

each of the ‘next words.’ We can easily observe from the result that for bigram the entropy and perplexity is 4.32 and 75.193 respectively and as far as is concern trigram is 3.83 and 46.097 respectively.

SUMMARY

The objective of the assignment was to implement Kneser Ney in order to predict the next word using the previous two words or one (trigram, bigram). Results are positive; we conclude that we have better predictions using the last two words than the one. In case having specific letters of a word, it give us the words, which have the smallest distance from that letters. Kneser-Ney smoothing makes use of the probability of a word being a novel continuation. Kneser-Ney’s determines how likely a word is to appear in an unfamiliar bigram or trigram context. The code can be found here: https://thinkingtea.github.io/repoTEA/.

FUTURE IMPROVEMENTS

There are a lot of things that could be tried:

- It could be fruitful to correct mangled or misspelled word before to try the prediction.
- Calculate the optimal value of D through test various values and selecting that values which results to the lowest perplexity.
- What is needed is a way to check faster different combination of parameters or just perform repeated simulation.