Jan aděra
janmadera97@gmail.com

Assignment 1
Exploring Entropy and Language Modeling

2021/01/31

1 Entropy of a Text

1.1 Basic properties of texts

As can be seen in the table 1 both texts have similar length and also average length of words is similar with czech words being longer by 0.234 character.

	Czech text	English text
Words count	222412	221098
Characters count	1030631	972917
Avg chars per word	4.634	4.400
Number of different words	42826	9607
Words with freq. 1	26315	3811
Words with freq. higher than 1	16511	5796
Ratio of words with freq. 1 to higher freq. w.	1.594	0.658

Table 1: Basic properties of Czech and English texts.

"Word"	Count
,	13788
.	12931
a	4486
v	4043
:	3434
se	3378

Table 2: Most frequent "words" in Czech text.

"Word"	Count
,	14721
the	13299
of	9368
.	5645
and	5537
in	4761

Table 3: Most frequent "words" in English text.

1.2 Text differences

There is a significant difference between the Czech and English text in terms of their conditional entropy. As can be seen in the figure 1 the Czech language has lower entropy than the English text even when we mess it up a little (messing up is explained in the assignment text).

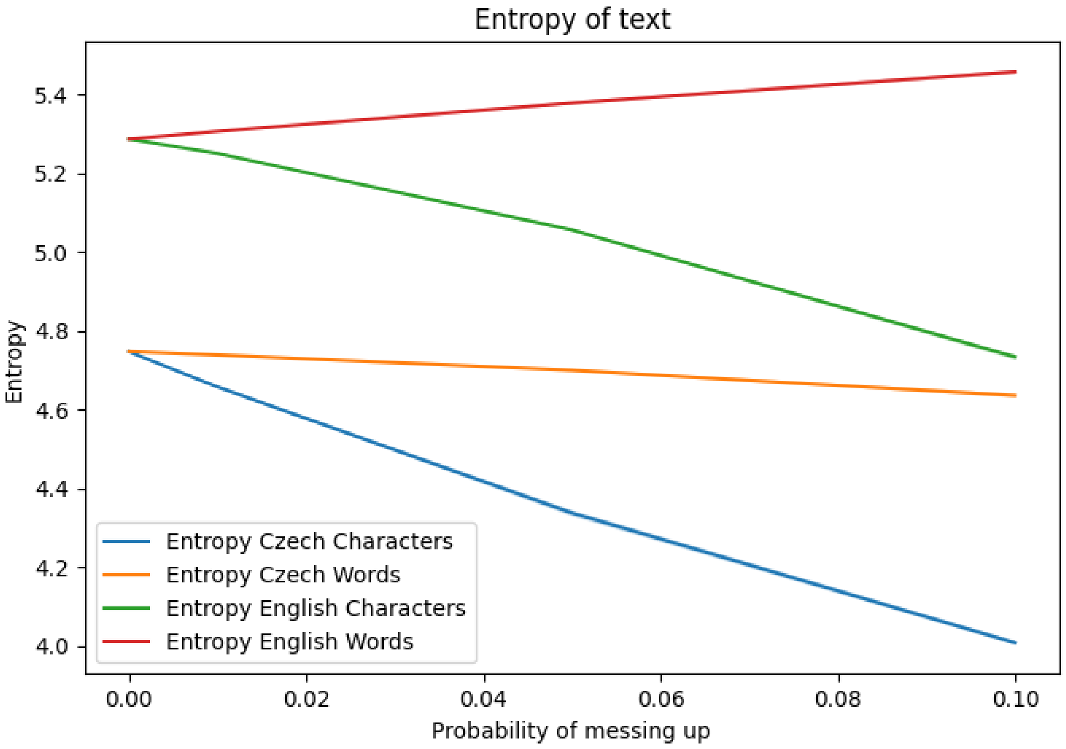

Figure 1: Entropy of Czech and English text with dependence on probability of words (or characters) being messed up.

1.2.1 Role of number of different words and single appearance words

The difference in entropy can be in my opinion partly explained by the fact that the Czech language has significantly more words with frequency of 1 and also significantly more different words as is shown in table 1. The Czech language has 26315 words with single appearance which is almost 7 times more than in the case of English language, which has only 3811 single appearance words. Also the Czech text has approximately 4.5 times more different words than English text.

Text with significantly more single appearance words (or so called tail words) and also with more different words will in many cases have much lower conditional entropy than similar text with less single appearances and less different words. This is caused by the fact that we are 100% sure what word will follow after the tail word and therefore it will not increase entropy of the text.

If we create a text where no word is repeated, then the conditional entropy of such text would be zero. If we create text where there is only one word repeated in it, then the conditional entropy would be also zero as we would always know which word would be next. To get as small entropy of text as possible, we want to have as many different words in the text as possible and at the same time the words that are repeated more than once to be repeated as many times as possible. There are many more factors that play a significant role in the size of entropy of text (for example order of words in text) and role of the number of words with one appearance and number of different words is based only on my observations.

1.2.2 Messing up English text makes the entropy grow

The second difference that I discovered is the more we mess up whole words in the English text, the more the Entropy grows. In the Czech text it is other way around. As the mess up grows the conditional entropy gets a little smaller. This can be seen in the figure 1.

I think that it can be explained by the ratio of words with single appearance and words with frequency higher than 1. This ratio of words is in case of Czech text 1.594 and in case of English text 0.658 as is shown in the table 1. This means that in English text, there are less tail words than words with frequency of appearance higher than 1. Therefore, when we randomly choose from these words a word, we have higher probability (than in the Czech text) that it will be the more frequent word. If we are messing up the tail word which does not increase entropy, we therefore have high chance, that we will increase entropy. This is in my opinion one of possible reasons why the entropy in case of English text rises when we mess the text up.

	Entropy		
Mess up likelihood	Min	Max	Average
0%	5.2874	5.2874	5.2874
0.001%	5.2873	5.2875	5.2874
0.01%	5.2868	5.2875	5.2871
0.1%	5.283	5.284	5.284
1%	5.248	5.253	5.250
5%	5.053	5.064	5.059
10%	4.725	4.737	4.731
	Perplexity		
Mess up likelihood	Min	Max	Average
0%	39.0548	39.0548	39.0548
0.001%	39.0510	39.0558	39.0546
0.01%	39.038	39.056	39.046
0.1%	38.946	38.969	38.955
1%	37.994	38.126	38.065
5%	33.191	33.456	33.328
10%	26.453	26.676	26.557

Table 4: Entropy and perplexity of **English** text with given percent of **CHARACTERS** messed up.

	Entropy		
Mess up likelihood	Min	Max	Average
0%	5.2874	5.2874	5.2874
0.001%	5.2874	5.2875	5.2875
0.01%	5.2875	5.2878	5.2876
0.1%	5.2891	5.2902	5.2895
1%	5.305	5.308	5.307
5%	5.376	5.384	5.380
10%	5.451	5.463	5.458
	Perplexity		
Mess up likelihood	Min	Max	Average
0%	39.0548	39.0548	39.0548
0.001%	39.0546	39.0566	39.0556
0.01%	39.058	39.064	39.060
0.1%	39.100	39.130	39.110
1%	39.544	39.613	39.583
5%	41.521	41.762	41.648
10%	43.749	44.099	43.959

Table 5: Entropy and perplexity of **English** text with given percent of **WORDS** messed up.

1.3 Entropy of two concatenated texts

The two languages L_1 and L_2 do not share any vocabulary items. The texts T_1 and T_2 from these two languages have same entropy E. If the last word of the T_1 does not repeat in the T_1 then the entropy of concatenated text will be still E as by concatenation we did not get any new information which would decrease entropy of whole text and also we did not make any bigrams which would increase entropy, because last word in the T_1 does not repeat more than once.

	Entropy		
Mess up likelihood	Min	Max	Average
0%	4.7478	4.7478	4.7478
0.001%	4.7477	4.7478	4.7477
0.01%	4.7467	4.7471	4.7470
0.1%	4.7378	4.7400	4.7386
1%	4.655	4.660	4.658
5%	4.331	4.347	4.337
10%	4.001	4.012	4.006
	Perplexity		
Mess up likelihood	Min	Max	Average
0%	26.868	26.868	26.868
0.001%	26.865	26.868	26.867
0.01%	26.848	26.855	26.852
0.1%	26.68	26.72	26.70
1%	25.19	25.29	25.25
5%	20.13	20.35	20.21
10%	16.01	16.13	16.07

Table 6: Entropy and perplexity of **Czech** text with given percent of **CHARACTERS** messed up.

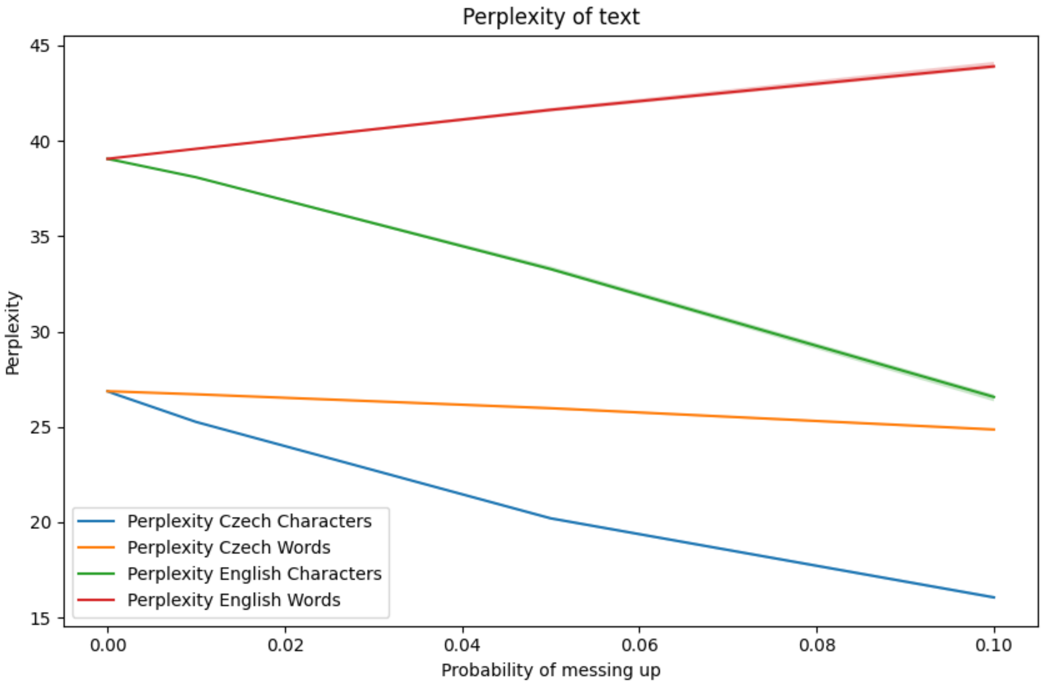

Figure 2: Perplexity of Czech and English text with dependence on probability of words (or characters) being messed up.

However, the entropy will increase when the last word of the T_1 is repeated in the T_1 more than once (let us name this repeated word w_r). We can show it by first appending first word of T_2 to T_1. In this case, the conditional probability $P(j|w_r)$ for all bigrams that begin with the word w_r will decrease, because $P(j|w_r) = P(w_r, j)/P(w_r)$ and $P(w_r, j)$ decreases. As conditional entropy is computed as $H(J|I) = -\sum_{i \in I, j \in J} P(i,j) log_2 P(j|i)$ where the decreased conditional probability is inside the logarithm, this sum will increase. By concatenating the T_2 we increased entropy of T_1. Therefore, in this case, the final entropy of concatenated T_1 with T_2 will increase.

	Entropy		
Mess up likelihood	Min	Max	Average
0%	4.7478	4.7478	4.7478
0.001%	4.7477	4.7479	4.7478
0.01%	4.7476	4.7479	4.7478
0.1%	4.7467	4.7475	4.7470
1%	4.7377	4.7410	4.7394
5%	4.6976	4.7026	4.7004
10%	4.631	4.643	4.636
	Perplexity		
Mess up likelihood	Min	Max	Average
0%	26.868	26.868	26.868
0.001%	26.867	26.869	26.868
0.01%	26.863	26.870	26.867
0.1%	26.847	26.862	26.853
1%	26.680	26.742	26.711
5%	25.948	26.039	26.000
10%	24.782	24.978	24.872

Table 7: Entropy and perplexity of **Czech** text with given percent of **WORDS** messed up.

2 Cross-Entropy and Language Modeling

I computed lambdas as was said in the assignment. I also used the hints from lecture slides and hints said in lectures which are:

- Set vocabulary size $|V|$ to number of all words in training data. It is used to compute uniform distribution probability value.
- Instead of prepending two words in front of all data, I started by computing trigram on three words w_{i-2}, w_{i-1}, w_i, then I computed bigram on w_{i-1}, w_i, and unigram with word w_i. This was said in the lecture and also avoids the beginning-of-data problems with cost of not using the two words in computation of probability distributions approximations.
- I assigned 0 probability to $p_n(w|h)$ whenever $c_{n-1}(h) > 0$, and $1/|V|$ otherwise.

2.1 Computing smoothing parameters

As was said on lecture, when we compute lambdas on the same data on which we computed the conditional probabilities, we will get in theoretical case whole weight on the lambda at highest n-gram (which is in our case the trigram). In our practical case, we get these values of lambdas l_i at given i-gram:

- Czech train text:
  - $l_0 = 8.59726111 \cdot 10^{-30}$
  - $l_1 = 1.51276074 \cdot 10^{-12}$
  - $l_2 = 5.04154935 \cdot 10^{-4}$
  - $l_3 = 0.999495845$
- English train text:
  - $l_0 = 1.64263705 \cdot 10^{-26}$
  - $l_1 = 5.33458275 \cdot 10^{-9}$
  - $l_2 = 4.95124323 \cdot 10^{-4}$
  - $l_3 = 0.999504870$

On heldout data, we get:

- Czech heldout data:
  - $l_0 = 1.35701832 \cdot 10^{-11}$

  - $l_1 = 5.84292545 \cdot 10^{-1}$
  - $l_2 = 3.33073535 \cdot 10^{-1}$
  - $l_3 = 8.26339198 \cdot 10^{-2}$
- English heldout data:
  - $l_0 = 3.72759995 \cdot 10^{-22}$
  - $l_1 = 3.07029408 \cdot 10^{-1}$
  - $l_2 = 5.52072714 \cdot 10^{-1}$
  - $l_3 = 1.40897878 \cdot 10^{-1}$

2.2 Computing cross-entropy of test data

Cross-entropy of the test parts of texts using the trained ngram probability distributions and smoothing parameters is:

- Czech text: 10.184347
- English text: 7.538001

The cross-entropies of test parts with tweaked smoothing parameters are in table 8.

Additions to trigram smoothing parameter

percent added	Czech text	English text
10%	10.189	7.545
20%	10.230	7.578
30%	10.297	7.634
40%	10.388	7.714
50%	10.507	7.821
60%	10.660	7.963
70%	10.865	8.157
80%	11.157	8.443
90%	11.650	8.940
95%	12.123	9.433
99%	13.127	10.525

Reduction of trigram smoothing parameter

percent of t.s.p. value	Czech text	English text
90%	10.187	7.540
80%	10.190	7.544
70%	10.195	7.549
60%	10.200	7.555
50%	10.207	7.564
40%	10.216	7.575
30%	10.228	7.589
20%	10.243	7.609
10%	10.267	7.637
0%	10.336	7.701

Table 8: Cross-entropies of Czech and English texts with tweaked trigram smoothing parameter.

Coverage graph is:

- Czech text: 86.46%
- English text: 95.56%

One of the main reasons why the cross-entropy of English text is lower than cross-entropy of Czech text is that coverage graph of English text is almost 96% which is much higher than the coverage graph of Czech text which is approximately 86%. Our trained model was therefore better suited for the test data in the case of

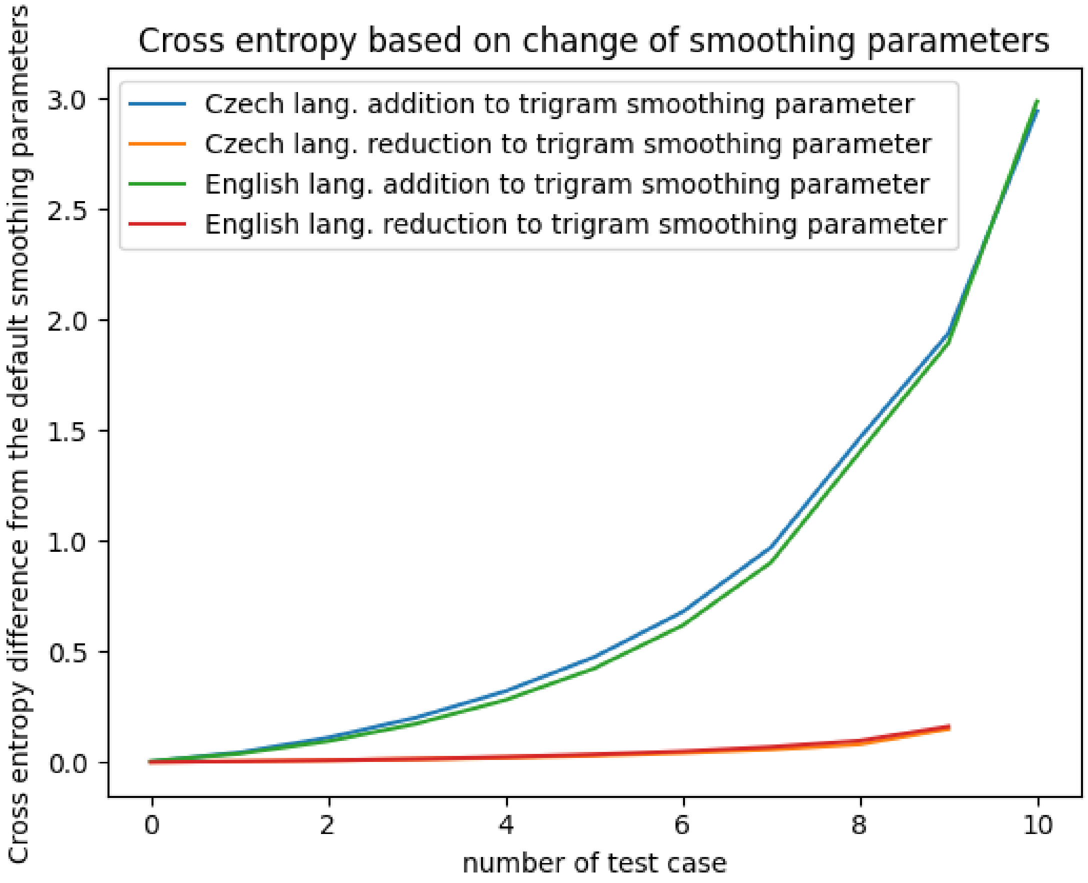

Figure 3: Plotted data from table 8.

English text compared to the model on Czech text. We therefore did not have to use uniform probability for that much words in English text case.
The smaller coverage graph of Czech text could be caused by the high number of different words and high number of words with frequency 1 in the Czech text compared to the English text.
If we for example switch the test text of Czech text and English text, than we get these coverage graphs:

- Czech train text, English test text: 26.00%
- English train text, Czech test text: 27.08%

And the cross-entropies are:

- Czech test text: 15.143
- English test text: 15.428

We can see on this example, that with smaller coverage graph, we can get higher cross-entropies. (side note: the similar "words" are mostly punctuations and other non-alphabetical symbols.)