analysis-entropy-of-text

November 1, 2020

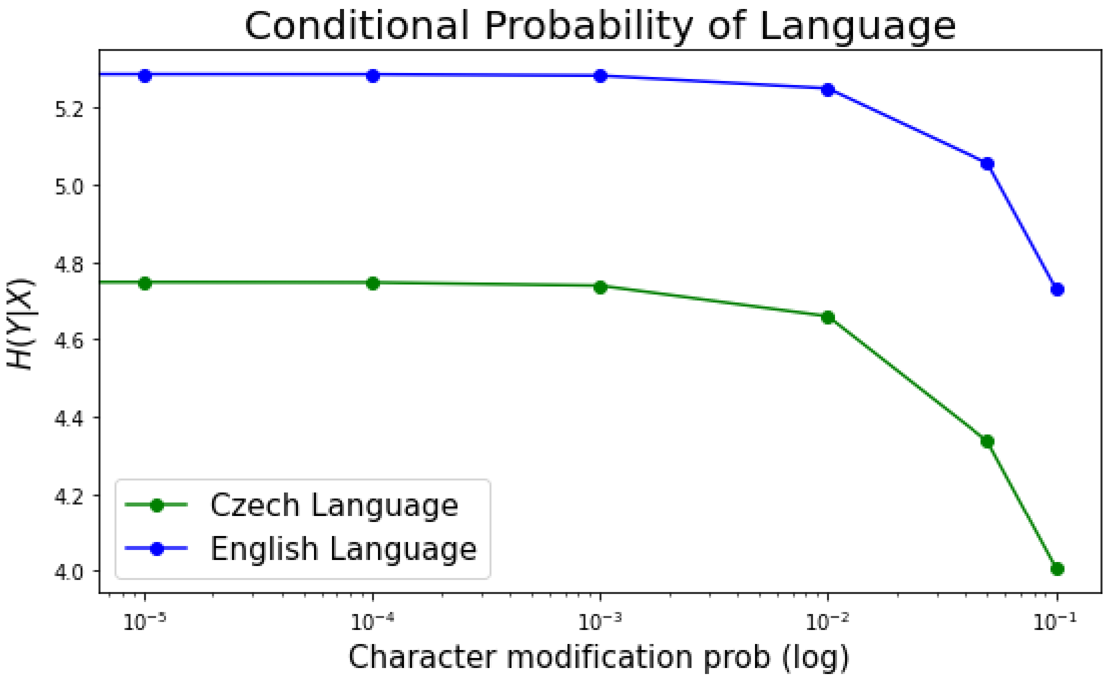
1 Conditional Entropy H(Y|X)

- All results are computed using the 10 repeat average.
- Note that `mess_prob=0.0` repesents original (unmodified) dataset…

1.1 Character modification

- In the table below (and also in the plot) we can see, that modifying the characters leads to a linear decrease of the conditional entropy.
- When we modify characters of words, we are creating new words and thus making the vocabulary much larger.
- Moreover we destroy the relationship between the words, because for example if in the original text we had "can of coke" 100 times, now we will probably end up each time with a different word, thus each combination of words w_{i-1}, w_i will be more and more unique, because each word w_i will start becoming more unique.
- This means that $p(w_{i-1}w_i)$ will go towards $1/|V|$ and $p(w_i|w_{i-1})$ will go to 1, thus $\log_2(p(w_i|w_{i-1})$ will more often be 0 and the conditional entropy will be lower.

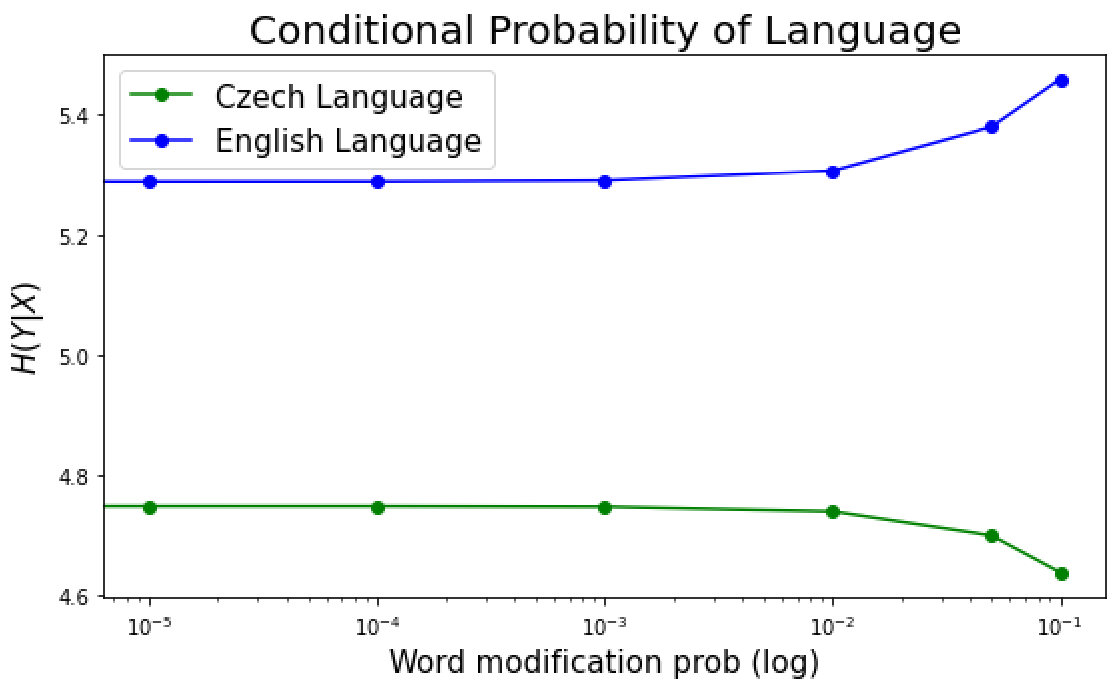
dataset mess_prob	cz	en	cz_perplexity	en_perplexity
0.0	4.747801	5.287398	26.867694	39.053984
1e-05	4.747718	5.287392	26.866157	39.053815
0.0001	4.746771	5.287000	26.848525	39.043208
0.001	4.738600	5.283650	26.696888	38.952674
0.01	4.659225	5.250462	25.267749	38.066804
0.05	4.336144	5.057127	20.198049	33.292550
0.1	4.007994	4.731907	16.088900	26.573320



1.2 Word modification

- Contrary to character modification, when we are substituting the words, we are not making the vocabulary larger. But we are changing the distribution of words. Each substitution leads to the fact that this word is in this position with uniform probability.
- Let's see the results first

dataset	cz	en
mess_prob		
0.0	4.747801	5.287398
1e-05	4.747787	5.287406
0.0001	4.747751	5.287613
0.001	4.746902	5.289483
0.01	4.738953	5.306460
0.05	4.698919	5.380170
0.1	4.637736	5.459387



- I think that $p(w_{i-1}w_i)$ will decrease for the most common word combinations.. because there is a high probability of modification of one of those two words (and result will uniformly random word.
- I though $p(w_i|w_{i-1})$ that will decrease, because now we will start having more and more random history for this word... and i=that it will tend to go to $1/|V|$ ... so that the Entropy will be increasing ... and the distribution of the words will go towards the maximal entropy...
- But we can see, that this is the case only for English. For the Czech, it quite surprisingly decreases the entropy. My explanation is that this is happening because of the very large Czech dictionary. Let us see the stats about the languages... In the table, we have the size of the dataset, vocabulary size, number of different characters... and quantiles with counts of most common words...

	dataset	T	V	C	max	q25	q50	q75	q95
0	cz	222412	42826	117	13788	1	1	2	11
1	en	221098	9607	74	14721	1	2	7	56

- We can see that Czech vocabulary has 42826 different words. Also half of the words are only single time in the dataset and even 75% of words from the vocabulary is at most two times in the dataset.
- So if we randomly change 10% of the words, we are changing about $22,241$ of the words. Most probably we will choose again and again some common words.
- These words will be changed uniformly to some of $42,826$ different words. So there is a high chance that often we sample a very common word and substitute it with a word that was in the

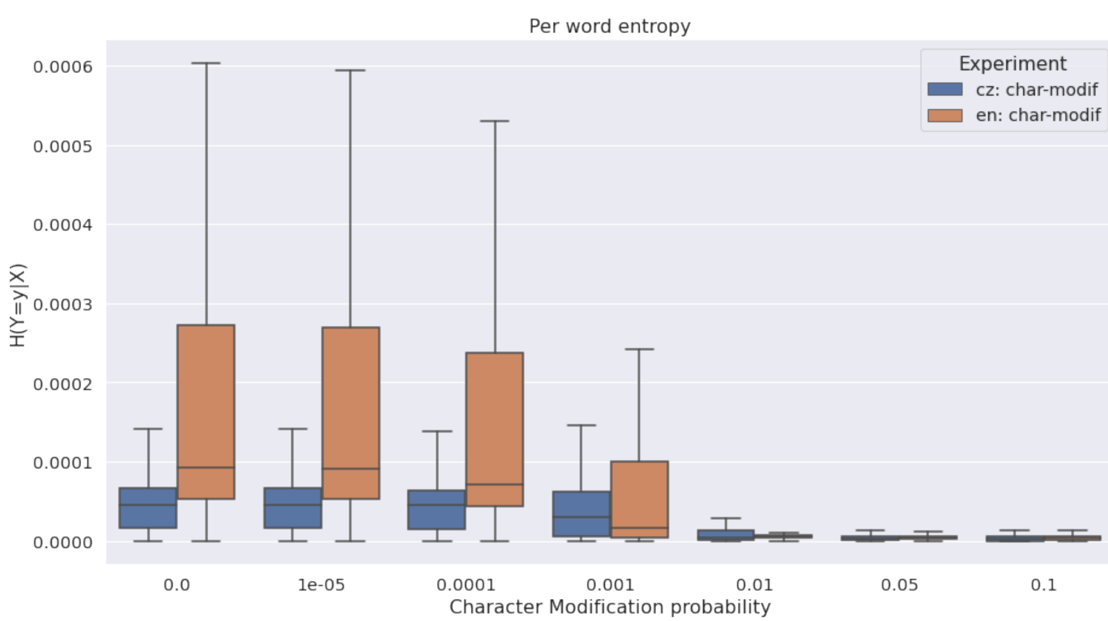
dataset only once or twice ... If this new changed 'rare' word r is in the history e.g. $p(w_i|r_{i-1})$ then the probability will be very high 1 or 0.5 .. thus this will lower the entropy.

- This is not the case for English, where is still a higher chance that a new word was at least multiple times in the dataset.

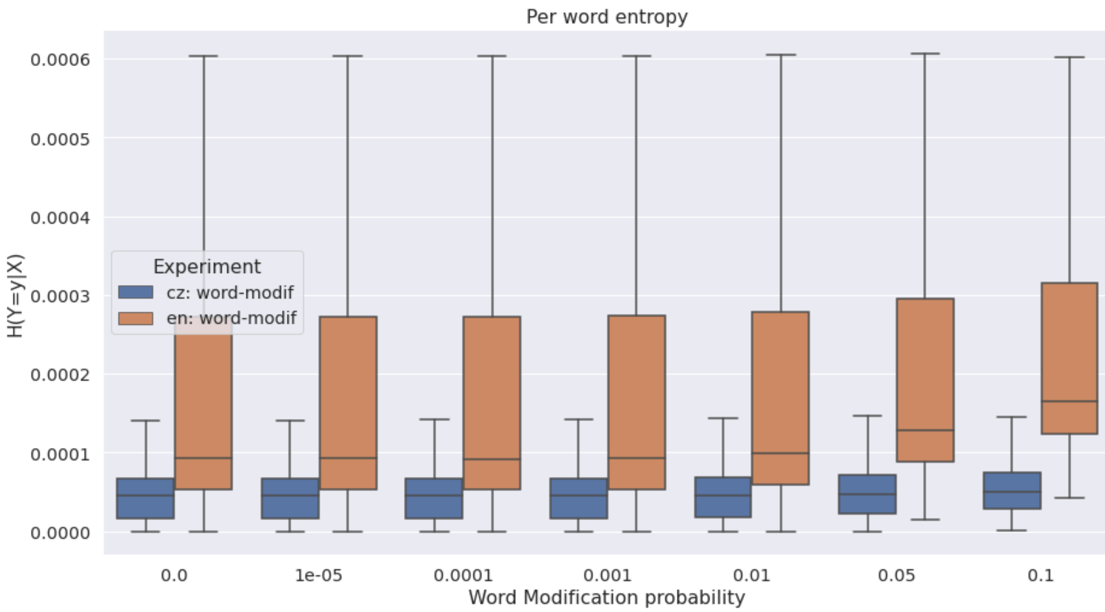
2 Per-word contribution to Conditional Entropy $H(Y = y|X), y \in \mathcal{Y}$

- I thought that it would also be interesting to see how individual words contribute to the conditional entropy.
- In the following table, we will find maximum, minimum, average, and quantiles for $H(Y = y|X)$ for different modification probabilities.
- First table and plot are for character modification. The second table and plot are for word modification.
- For the character modification we can see, that all words reduce their contribution to the conditional entropy quite linearly. We can also see that some words contribute a lot (common words) but most words contribute a little.
- For the word modification, we can see, that for English the skew of the distribution is changing and that the more we change words uniformly randomly to other words, the mean per-word contribution to the conditional entropy is getting closer to 0.5 quantile. For Czech, this is happening also, but too little.

```
<pandas.io.formats.style.Styler at 0x7fcdcde8f280>
```




```
<pandas.io.formats.style.Styler at 0x7fcdcde5b940>
```



We can also show those words, which contribute to conditional entropy the highest. Result is not suprsing ...

English

	word	avg_1
1713947	,	0.146136
1713956	the	0.143909
1713958	of	0.083818
1713955	in	0.080007
1713964	and	0.073680
1713941	.	0.068793
1714039	a	0.061798
1713968	to	0.050796

Czech

	word	avg_1
15	,	0.107800
1	.	0.101425
120	v	0.056312
53	a	0.047288
142	"	0.040737
46	se	0.040549
80	na	0.035754
18	-	0.025175

3 Paper and Pen excersise

- Assume languages,$ L_1$ and L_2 which do not share any vocabulary items
  - $H(J|I) = E$ for T_1 in laguage L_1 and also for T_2 in L_2
  - now we create text T_1T_2.
  - What will be the $H(J|I)$?
- **Lets split it by all possible cases...**
- $H(J|I)$ where $J, I \in T_1$ will be E (by definition)
- $H(J|I)$ where $J, I \in T_2$ will be E also (by definition)
  - NOTE: There is single change for first word in T_2
  - It was $P(S, t_1) \log_2 P(t_1|S)$ where S is starting symbol
  - but $P(t_1|S)$ was equal to 1 ... so $\log(P(t_1|S) = 0$
- $H(J|I)$ where $I \in T_2$, $J \in T_1$ is 0 because $P(i, j) = 0$ for all such words
- $H(J|I)$ where $I \in T_1$, $J \in T_2$ will be 0 except...
  - $P(t_n^1 t_1^2)P(t_1^2|t_n^1)$ where t_n^1 is last word in text $ T_1 $ and t_1^2 if first word from text T_2.
  - In this case we get $P(i, j) = 1$ and $P(j|i) \geq 0$
- So we can conclude that conditional entropy of T_1T_2 will be $2E + P(t_1^2|t_n^1)$.