Assignment 1 - Exploring Entropy and Language Modeling

Adrián Pauer

1 First part - conditional entropy

In the first part we calculated conditional entropy for text files using the formula

$$H(J|I) = -\sum_{i \in I,\, j \in J} P(i,j) \log P(j|i).$$

To begin, we counted the occurrences of word pairs (i, j) and individual unigram frequencies within the text. The joint probability $P(i, j)$ was determined as the count of (i, j) divided by number of possible positions for word i in text, including those corresponding to the symbol '.'. For the conditional probability$P(j|i)$ we computed $\frac{P(i,j)}{word\ counts(i)}$. Finally, we conducted an experiment by perturbing the text files with a given probability, as specified in the assignment..

For both text files we observed that the entropy of the original text was approximately 5. Specifically, the entropy for `English` text file it was 5.280207, while for the `Czech` text file, it was 4.676353. This suggests that both languages encode a similar amount of information. However, the Czech language allows for greater certainty in predicting the next word given the previous one. In terms of perplexity, which reflects the uncertainty corresponding to a $2^{H(J|I)}$-sided die, the Czech text had a perplexity of 25.569513, much lower than the English text, which had a perplexity of 38.859809.

From the experiment, we recorded the maximum, minimum, and mean entropy and perplexity values for both text files under each configuration of the text perturbation process.

- messing up characters

data	mess up prob	min entropy	mean	max entropy	perplexity
	0	5.280207	5.280207	5.280207	38.859809
	0.00001	5.280096	5.280163	5.280240	38.858634
	0.0001	5.279586	5.279801	5.280164	38.848887
ENG	0.001	5.275291	5.275991	5.276331	38.746419
	0.01	5.234007	5.236100	5.239298	37.689746
	0.05	5.016802	5.022485	5.025618	32.502646
	0.1	4.679645	4.684037	4.689995	25.706067
	0.	4.676353	4.676353	4.676353	25.569513
	0.00001	4.676179	4.676262	4.676346	25.567898
	0.0001	4.675005	4.675362	4.675639	25.551954
	0.001	4.666025	4.666794	4.667814	25.400653
CZK	0.01	4.582814	4.585807	4.587945	24.014053
	0.05	4.260054	4.262299	4.265119	19.190210
	0.1	3.928082	3.934035	3.937836	15.284894

As shown in the table, both entropy and perplexity decrease as the probability of character exchange increases. Intuitively, character exchanges introduce more unique words and bigrams, which reduces the overall conditional entropy.

- messing up words

data	mess up prob	min entropy	mean	max entropy	perplexity
	0	5.280207	5.280207	5.280207	38.859809
	0.00001	5.280130	5.280236	5.280309	38.860590
	0.0001	5.280294	5.280406	5.280591	38.865171
ENG	0.001	5.281787	5.282115	5.282886	38.911249
	0.01	5.297370	5.299360	5.301445	39.379160
	0.05	5.366603	5.370550	5.373823	41.371052
	0.1	5.444380	5.447712	5.451418	43.644012
	0.	4.676353	4.676353	4.676353	25.569513
	0.00001	4.676291	4.676345	4.676385	25.569381
	0.0001	4.676127	4.676276	4.676458	25.568160
CZK	0.001	4.674899	4.675529	4.676040	25.554920
	0.01	4.667936	4.668813	4.670392	25.436237
	0.05	4.630099	4.631945	4.634214	24.794453
	0.1	4.573507	4.576144	4.580605	23.853753

When words are exchanged, entropy increases for the English data but decreases slightly for the Czech data. This suggests that disrupting the natural structure of language impacts the two languages differently. The Czech language exhibits a high level of grammatical predictability, so word exchange has less effect on its entropy. This could be attributed to the frequent use of common word endings in Czech, which preserve some linguistic structure even when words are rearranged. In contrast, for English, the increase in entropy indicates that its grammatical structure and word similarity are less robust in terms of predictability.

We also visualize the entropy results in Figure 1, which corresponds to the outcomes of our experiment and the data presented in the tables.

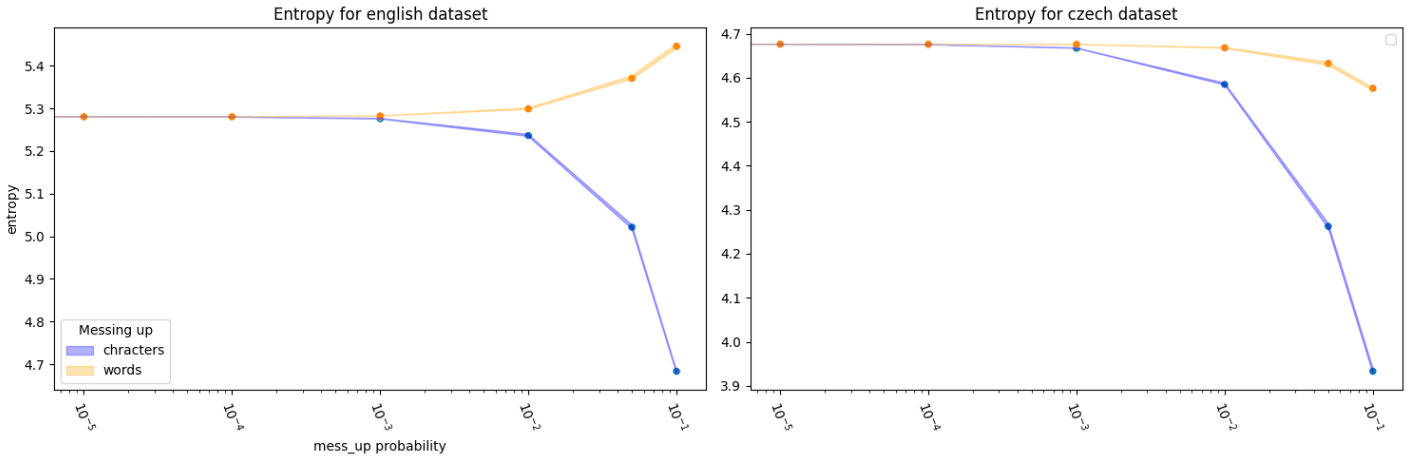

Figure 1: Conditional entropy for text files

We also present basic characteristics of the two languages, such as the most frequent words, single-word counts, and vocabulary sizes, in Table 1. Additionally, the visualization of the most frequent words is shown in Figure 2.

Table 1: Basic language characteristics

Data	Words	Vocabulary	Total Char	Unique char	Avg Word Length	Freq1 Words
ENG	221098	9607	972917	74	4.400388	3811
CZK	222412	42826	1030631	117	4.633882	26315

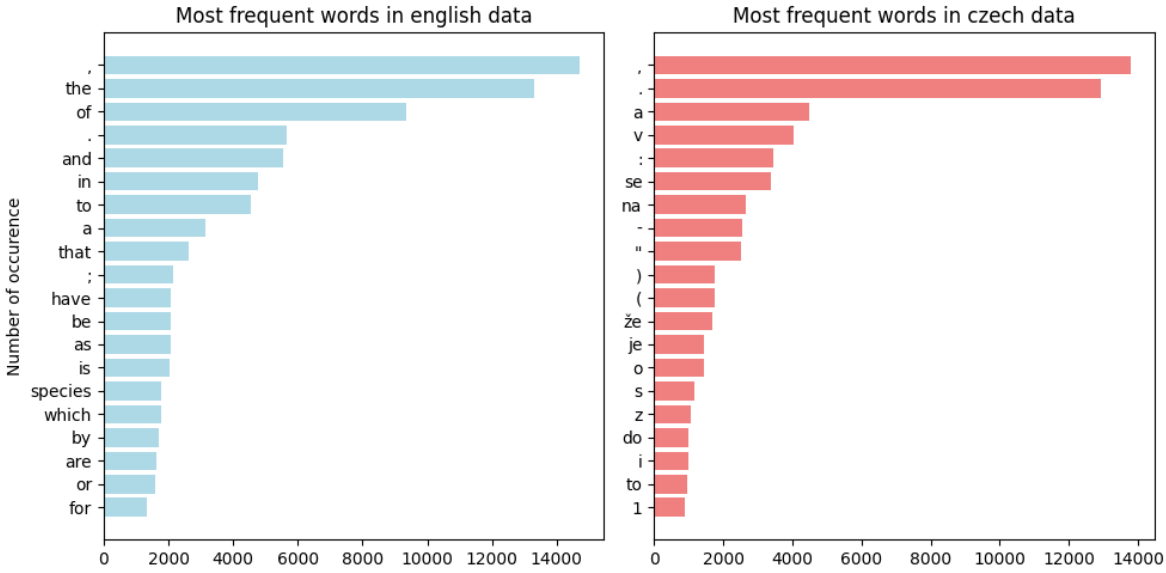

Figure 2: Most frequent words

In Figure 2, we observe that the most frequent words in the Czech language are primarily prepositions (e.g., a, v, se, na) and punctuation marks (, . :). These words are generally short, with lengths not exceeding two characters. While grammatically essential, they carry less semantic information compared to content words, which likely contributes to the lower entropy observed in Czech data. In contrast, the English dataset includes some content words among the most frequent words, reflecting differences in the grammatical and lexical structures of the two languages.

From Table 1, we observe that both datasets have a similar number of words and average word lengths. However, the Czech dataset contains significantly more words with a frequency of 1, which contributes to its lower entropy. When replacing words, it is more likely that a word in Czech is replaced with another low-frequency word, which could explain the observed decrease in entropy for Czech. In contrast, for English, the replacement process likely disrupts its structure more, leading to an increase in entropy.

1.1 Conditional entropy of L_1, L_2

As 2 languages don't share any word, new text $T_3 = T_1T_2$ contains only 1 new bigram (k, l) at the end of T1 an beginning of T2. Counts of bigrams $c(i, j)$ and unigrams $c(i)$ in new text remain the same as

for T_1 and T_2 separately. If we denote d_x as length of text T_x we can compute entropy as

$$H(J|I) = -\sum_{i\in I, j\in J} P_3(i,j)\log P_3(j|i) = -\sum_{i\in I_1,\, j\in J_1} P_3(i,j)\log P_3(j|i) - \sum_{i\in I_2,\, j\in J_2} P_3(i,j)\log P_3(j|i) - P_3(k,l)\log P_3(l|k),$$

where I_1 ,J_1 and I_2 ,J_2 corresponds to decomposition of I,J according to datasets T_1 and T_2. As (k,l) appears only once, $P_3(k,l) = \frac{1}{d_3}$ and $P_3(l|k) = \frac{1}{c_3(k)}$, which gives us

$$-P_3(k,l)\log P_3(k|l) \geq 0.$$

For pairs from I_1, J_1 we have

$$P_3(i,j) = \frac{c_3(i,j)}{d_3} \leq \frac{c_2(i,j)}{d_2} = P_2(i,j)$$

as $d_2 \leq d_3$. For conditional probability

$$P_3(j|i) = \log(\frac{c_3(i,j)}{c_3(i)}) = \log(\frac{c_2(i,j)}{c_2(i)}) = P_2(j|i) \tag{1}$$

And

$$\sum_{i\in I_2,\, j\in J_2} P_3(i,j)\log P_3(j|i) \geq \sum_{i\in I_2,\, j\in J_2} P_2(i,j)\log P_2(j|i)$$

$$-\sum_{i\in I_2,\, j\in J_2} P_3(i,j)\log P_3(j|i) \leq -\sum_{i\in I_2,\, j\in J_2} P_2(i,j)\log P_2(j|i)$$

For P_1 we obtain same upper boundary. In the result we get upper boundary on entropy:

$$H(J|I) \leq E + E + e_{k,l} = 2E + e_{k,l}.$$

Intuitively, we can get entropy E for combined text T_1T_2. For instance, if unigrams and bigrams occur with the same frequency in both texts. In this case $d_3 = 2d_1$ and $d_1 = d_2$. Under these conditions entropy is

$$H(J|I) = -\frac{1}{d_3}\sum_{i\in I_1,\, j\in J_1} c_3(i,j)\log P_3(j|i) \; - \frac{1}{d_3}\sum_{i\in I_2,\, j\in J_2} c_3(i,j)\log P_3(j|i) - P_3(k,l)\log P_3(l|k) =$$

$$-\frac{2}{2*d_1}\sum_{i\in I_1,\, j\in J_1} c_3(i,j)\log P_3(j|i) - P_3(k,l)\log P_3(l|k) = E - P_3(k,l)\log P_3(l|k)$$

For $c_1(k) = 1$ we get $log(1/1) = 0$ and $H(X|Y)_{T_1T_2} = E$. If word k occur in text more than once entropy gets higher than E. From this observation we get intuition, that T_2 cannot reduce the uncertainty of predicting words given the previous ones, since the two texts are independent and T_2 can only bring more uncertainty to combined text. But from (1) we see, that joint probabilities decrease, what is caused by higher length of combined text T_1T_2. As the conditional probabilities stay the same, one would expect the resulting entropy to be lower than E. According to our ideas, we conclude that the entropy of T_1T_2 should be greater or equal to E.

2 Language model and smoothing

First, we computed the trigram counts from the training data, followed by the bigram counts, and then the unigram counts. To ensure that every word was included in the unigram count, we extended the text by adding two virtual words at the beginning and at the end. Using the Expectation-Maximization algorithm with $\epsilon = 10^{-6}$ we obtained the following coefficients for the `training data`:

eng data : $\lambda_0 = 1.88773400 * 10^{-6}$, $\lambda_1 = 4.19734724 * 10^{-5}$, $\lambda_2 = 2.23883826 * 10^{-3}$, $\lambda_3 = 0.997717301$
czk data : $\lambda_0 = 1.49496220 * 10^{-7}$, $\lambda_1 = 9.64676621 * 10^{-6}$, $\lambda_2 = 2.28809454 * 10^{-3}$, $\lambda_3 = 0.997702109$

From the EM algorithm, we observe that the coefficient λ_3 gradually increases, approaching 1. As expected, the model favours the trigram distribution as the best fit, with the weight converging close to one. For `heldout` data we obtained the following coefficients:

$$\text{eng data : } \lambda_0 = 0.000115,\ \lambda_1 = 0.16646,\ \lambda_2 = 0.650400,\ \lambda_3 = 0.183019$$
$$\text{czk data : } \lambda_0 = 0.001133,\ \lambda_1 = 0.446049,\ \lambda_2 = 0.439173,\ \lambda_3 = 0.113646$$

Highest coefficient for English model is λ_2 and for Czech model λ_1. This suggests that in English, trigrams capture more specific word order dependencies, whereas in Czech, bigrams play a more significant role.

As shown in Table 1, the Czech dataset contains a significant proportion of unique words, which explains the highest λ_1 coefficient for the Czech model.

For both models, lowest coefficients are λ_0, suggesting that the models are quite complex and capable of making accurate predictions based on the data. For English, highest coefficient is λ_2. As shown in table 1 smaller vocabulary size indicates that English relies on a more limited set of words, leading to more frequent and predictable word pairings.

We obtained cross entropy of 7.0213 for English dataset and 9.6937 for Czech dataset. From this we conclude, that English model is a closer approximation of the true distribution of English data. It may results from limited English morphological complexity, making it easier for the model to predict sequences based on trigrams, bigrams, and unigrams.

Experiment with increasing λ_3 coefficient is displayed on figure 3.

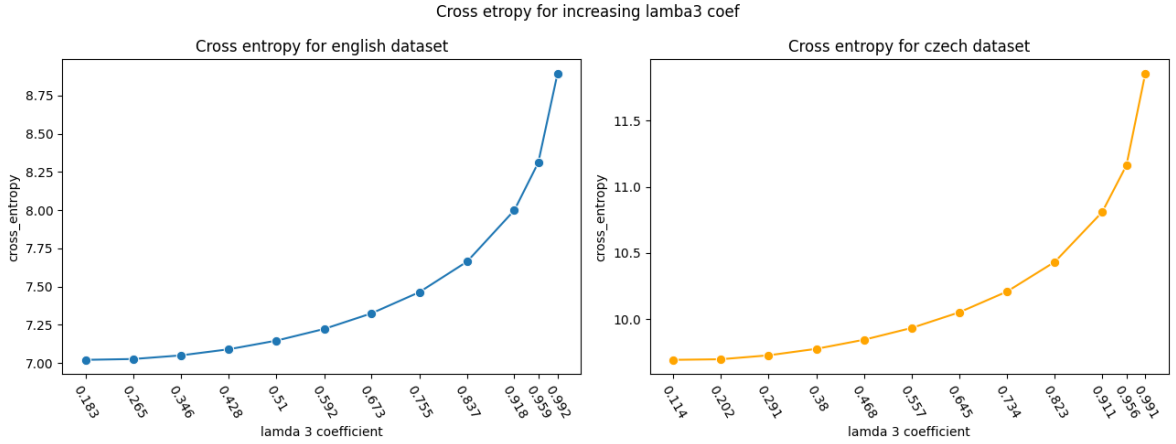

Figure 3: Cross entropy change with increasing lambda 3 coefficient

As shown in figure 3, an increasing λ_3 coefficient results in an increase in entropy for both models. The magnitude of the increase is similar, around 10^6, and the plots follow a similar trend. By increasing λ_3, we assign more weight on trigrams. When an exact trigram match is missing in the data, the model's predictive capacity drops, leading to higher uncertainty and cross entropy for words in new contexts.

For decreasing λ_3 we obtained same effect displayed on figure 4.

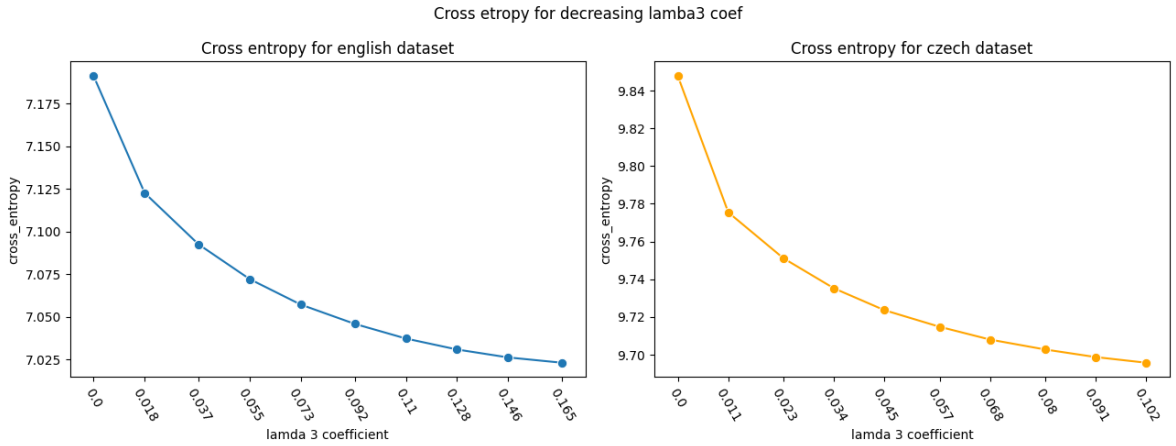

Figure 4: Cross entropy change with decreasing lambda 3 coefficient

For decrease of λ_3 coefficient we obtained cross entropy increase of same magnitude and similar graphs for both datasets. As expected cross entropy gets higher because original coefficients were optimal.

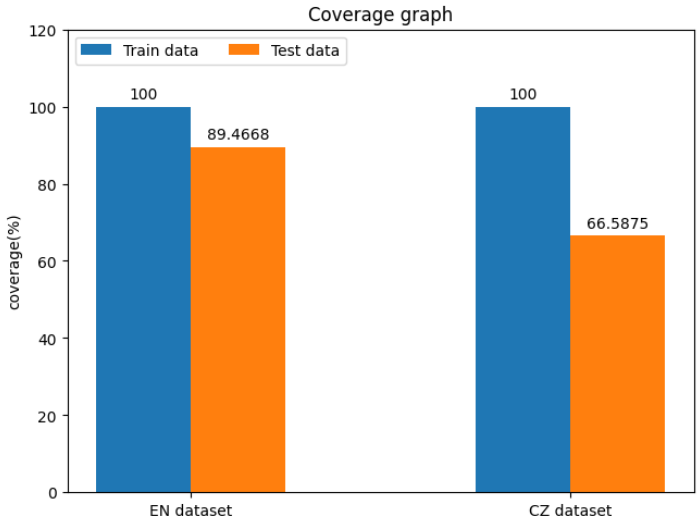

Figure 5: Coverage graph

As shown in figure 5, English training data covers approximately 90% of the test data, which contributes to the lower cross-entropy. In contrast, the Czech training data covers only 66.6% of the test data, making it intuitive that the cross-entropy for the Czech model is higher.