# Statistical Methods in NLP 1
# Assignment 1: Exploring Entropy and Language Modeling

Andrew McIsaac

January 31, 2022

## Contents

# 1 Entropy of a Text

## 1.1 English

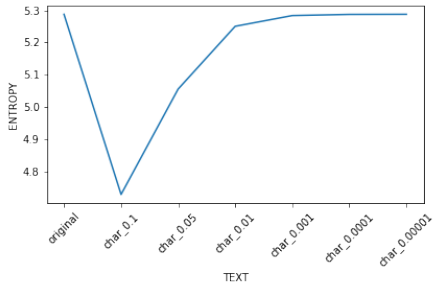
Table 1: Conditional Entropy of English

| Text | Entropy | | | Avg. Perplexity |
|---|---|---|---|---|
| original | 5.2874 | | | 39.0553 |
| | Min | Max | Average | |
| char 10% | 4.7262 | 4.7367 | 4.7308 | 26.5533 |
| char 5% | 5.0522 | 5.0622 | 5.0564 | 33.2764 |
| char 1% | 5.2473 | 5.2529 | 5.2504 | 38.0654 |
| char 0.1% | 5.2825 | 5.2842 | 5.2835 | 38.9484 |
| char 0.01% | 5.2868 | 5.2873 | 5.2871 | 39.0454 |
| char 0.001% | 5.2874 | 5.2875 | 5.2874 | 39.0549 |
| word 10% | 5.4508 | 5.4602 | 5.4572 | 43.9313 |
| word 5% | 5.3769 | 5.3836 | 5.3800 | 41.6440 |
| word 1% | 5.3053 | 5.3087 | 5.3072 | 39.5928 |
| word 0.1% | 5.2890 | 5.2899 | 5.2894 | 39.1089 |
| word 0.01% | 5.2875 | 5.2878 | 5.2877 | 39.0609 |
| word 0.001% | 5.2874 | 5.2876 | 5.2875 | 39.0560 |

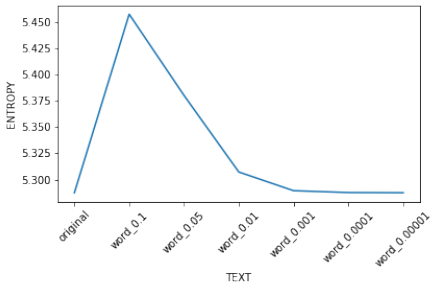
## 1.2 Czech

Table 2: Conditional Entropy of Czech

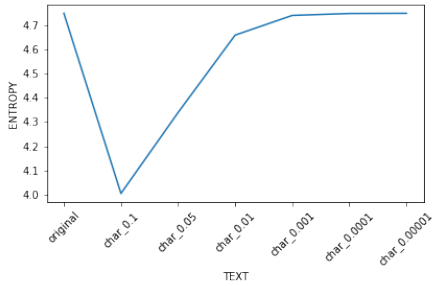
| Text | Entropy | | | Avg. Perplexity |
|---|---|---|---|---|
| Original | 4.7478 | | | 26.8685 |
| | Min | Max | Average | |
| char 10% | 3.9976 | 4.0117 | 4.0048 | 16.0534 |
| char 5% | 4.3338 | 4.3443 | 4.3383 | 20.2283 |
| char 1% | 4.6547 | 4.6610 | 4.6578 | 25.2430 |
| char 0.1% | 4.7384 | 4.7399 | 4.7390 | 26.7039 |
| char 0.01% | 4.7467 | 4.7471 | 4.7469 | 26.8516 |
| char 0.001% | 4.7477 | 4.7478 | 4.7478 | 26.8671 |
| word 10% | 4.6335 | 4.6444 | 4.6378 | 24.8945 |
| word 5% | 4.6965 | 4.7022 | 4.6992 | 25.9768 |
| word 1% | 4.7383 | 4.7417 | 4.7394 | 26.7126 |
| word 0.1% | 4.7464 | 4.7474 | 4.7468 | 26.8508 |
| word 0.01% | 4.7475 | 4.7478 | 4.7477 | 26.8662 |
| word 0.001% | 4.7478 | 4.7479 | 4.7478 | 26.8686 |



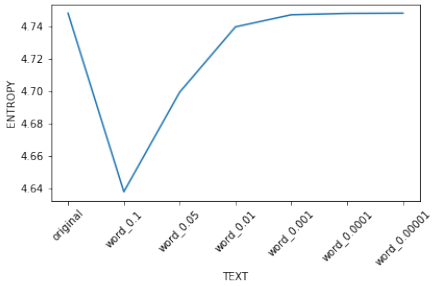

(a) English with messed up characters




(b) English with messed up words




(c) Czech with messed up characters




(d) Czech with messed up words

Figure 1: Conditional entropy of text for English and Czech with varying probabilities of messing up the text at the character and word level.

## 1.3 Discussion

Table 3: Text statistics at word, bigram, and character level

| | English | Czech |
|---|---|---|
| Text Size | 221098 | 222412 |
| Vocabulary Size | 9607 | 42826 |
| Top 20 Word Frequency | 109798 | 78095 |
| Unique Words | 3811 | 26315 |
| Different Bigrams | 73246 | 147136 |
| Unique Bigrams | 49600 | 125007 |
| Total Characters | 972917 | 1030631 |
| Average Characters | 4.400 | 4.633 |
| Different Characters | 74 | 117 |

The conditional entropy of the original Czech data is lower than the original English data (4.7478 compared to 5.2874). This suggests that the ability to predict a word given another word is slightly better for Czech. From Table 3, comparing vocabulary size shows that the number of different words in Czech is about 4.5x more than in English, meaning that there are more different bigrams for Czech (about 2x). Further, the number of unique bigrams in Czech represents a larger proportion (85.0%) of the total different bigrams when compared to English (67.7%). This means that 85% of the bigrams have perfect conditional predictive ability, and do not increase the entropy at all. So the total conditional

entropy of Czech is lower than English. The reason for the difference in the distribution between languages may be because of the morphological richness of Czech, with many inflectional forms that English does not have contributing to the larger vocabulary.

For both English and Czech data, when messing up characters within words, the entropy changes in a similar pattern (Figs. 1a, c). For very small probabilities the entropy is approximately equal to the original text, but as the probability of messing up a character increases, the entropy decreases. Again, this may be because the number of hapax legomena (words occurring only once in the text) increases and thus the number of unique bigrams, meaning fewer bigrams have any uncertainty at all.

When messing up words, however, the conditional entropy of English increases at larger probabilities, while for Czech it decreases (Figs. 1b, d). A large reason for this may be because of the respective vocabulary sizes, and specifically the hapax legomena. The number of hapax legomena reduce quite significantly in English because of the smaller vocabulary, so that when changing a word with some probability in two similar sized texts, the hapax legomena are chosen more often and are no longer unique. Hence the log conditional probability of a bigram is 0 less often, and this results in an increase in conditional entropy. For Czech, this effect isn't seen to the same extent because there are far more hapax legomena, (26315 vs 3811), and so there are still many 0 terms which keep the conditional entropy down.

Finally, the range in which the entropy changes differs when messing up characters or words. The conditional entropy has a range of 0.56 for English and 0.74 for Czech for characters, while for words it is just 0.17 and 0.11 for English and Czech respectively. This emphasises the importance of hapax legomena in determining the entropy, with changing characters making many more unique words.

### 1.4 Independent Languages $L_1$ and $L_2$

Assuming that texts $T_1$ and $T_2$ have start and end tags, so that the bigram count of the final word of $T_1$ is not affected by appending $T_2$, the conditional entropy of the joined text will be greater than the individual entropies $E$ of $T_1$ and $T_2$. This is because the conditional probabilities of the entropy equation do not change, and the joint probabilities, as computed by MLE, only change in their denominator, which is equal to the sum of the number of bigrams in the text minus one (accounting for the reduced number of bigrams including start/end tags). So with a smaller denominator and the same numerator over all pairs of words, the total conditional entropy is larger than $E$.

## 2 Cross-Entropy and Language Modeling

### 2.1 Smoothing Parameters

Trained smoothing parameters using both heldout and training data for both languages are detailed in Table 4.

Table 4: Smoothing parameters computed with the EM algorithm with terminating condition $\epsilon = 0.00001$

| Data | l_0 | l_1 | l_2 | l_3 |
|---|---|---|---|---|
| EN Heldout | 0.0700 | 0.2539 | 0.4922 | 0.1836 |
| EN Train | 7.504e-29 | 1.365e-14 | 5.181e-06 | 0.9999 |
| CZ Heldout | 0.1402 | 0.4289 | 0.2446 | 0.1862 |
| CZ Train | 1.717e-42 | 1.110e-21 | 3.524e-06 | 0.9999 |

## 2.2 Cross-Entropy

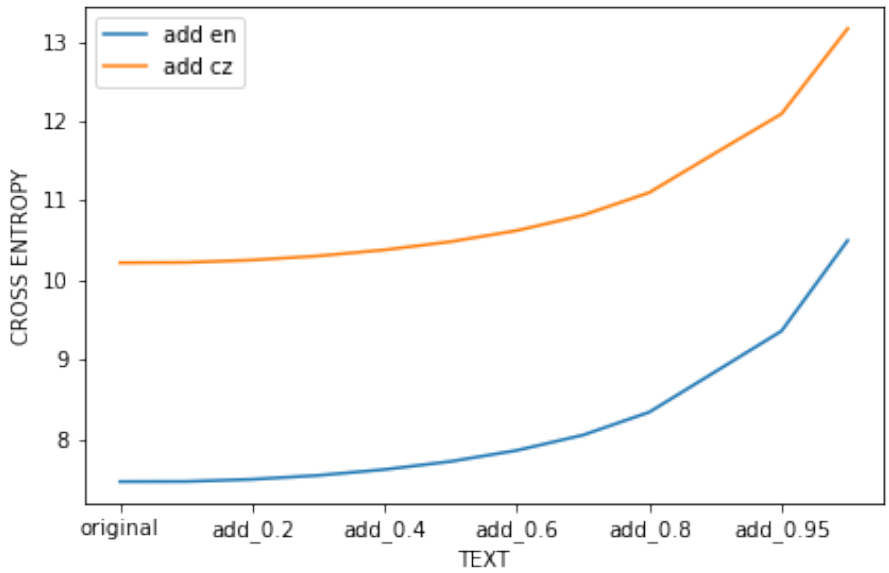
Cross-entropy for all permutations, along with the associated smoothing lambda parameters, are found for English in Table 5, and for Czech in Table 6.

Table 5: Cross-entropy on English test data with different smoothing parameters

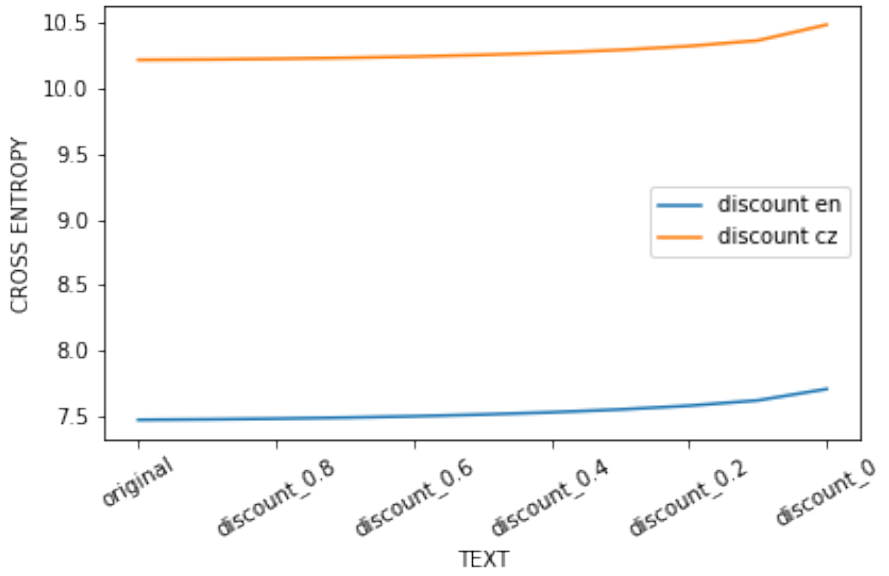
| Text | Lambdas | | | | Cross-Entropy |
|---|---|---|---|---|---|
| | l_0 | l_1 | l_2 | l_3 | |
| original | 0.0700 | 0.2539 | 0.4922 | 0.1836 | 7.4677 |
| add 10% | 0.0630 | 0.2285 | 0.4430 | 0.2652 | 7.4693 |
| add 20% | 0.0560 | 0.2031 | 0.3938 | 0.3469 | 7.4962 |
| add 30% | 0.0490 | 0.1777 | 0.3445 | 0.4285 | 7.546 |
| add 40% | 0.0420 | 0.1523 | 0.2953 | 0.5101 | 7.6197 |
| add 50% | 0.0350 | 0.1269 | 0.2461 | 0.5918 | 7.7215 |
| add 60% | 0.0280 | 0.1015 | 0.1969 | 0.6734 | 7.8602 |
| add 70% | 0.0210 | 0.0761 | 0.1476 | 0.7550 | 8.0533 |
| add 80% | 0.0140 | 0.0507 | 0.0984 | 0.8367 | 8.3409 |
| add 90% | 0.0070 | 0.0253 | 0.0492 | 0.9183 | 8.8501 |
| add 95% | 0.0035 | 0.0127 | 0.0246 | 0.9592 | 9.3614 |
| add 99% | 0.0007 | 0.0025 | 0.0049 | 0.9918 | 10.5003 |
| discount 90% | 0.0716 | 0.2597 | 0.5033 | 0.1652 | 7.4716 |
| discount 80% | 0.0732 | 0.2654 | 0.5144 | 0.1469 | 7.4773 |
| discount 70% | 0.0748 | 0.2711 | 0.5254 | 0.1285 | 7.4851 |
| discount 60% | 0.0763 | 0.2768 | 0.5365 | 0.1101 | 7.4954 |
| discount 50% | 0.0779 | 0.2825 | 0.5476 | 0.0918 | 7.5086 |
| discount 40% | 0.0795 | 0.2882 | 0.5587 | 0.0734 | 7.5255 |
| discount 30% | 0.0811 | 0.2939 | 0.5697 | 0.0550 | 7.5471 |
| discount 20% | 0.0826 | 0.2997 | 0.5808 | 0.0367 | 7.5758 |
| discount 10% | 0.0842 | 0.3054 | 0.5919 | 0.0183 | 7.6166 |
| discount 0% | 0.0858 | 0.3111 | 0.6030 | 0 | 7.7035 |

Table 6: Cross-entropy on Czech test data with different smoothing parameters

| Text | Lambdas | | | | Cross-Entropy |
|---|---|---|---|---|---|
| | l_0 | l_1 | l_2 | l_3 | |
| original | 0.1402 | 0.4289 | 0.2446 | 0.1862 | 10.2198 |
| add 10% | 0.1262 | 0.3860 | 0.2201 | 0.2675 | 10.2243 |
| add 20% | 0.1121 | 0.3431 | 0.1956 | 0.3489 | 10.2538 |
| add 30% | 0.0981 | 0.3002 | 0.1712 | 0.4303 | 10.3059 |
| add 40% | 0.0841 | 0.2573 | 0.1467 | 0.5117 | 10.3817 |
| add 50% | 0.0701 | 0.2144 | 0.1223 | 0.5930 | 10.4852 |
| add 60% | 0.0560 | 0.1715 | 0.0978 | 0.6744 | 10.6250 |
| add 70% | 0.0420 | 0.1286 | 0.0733 | 0.7558 | 10.8180 |
| add 80% | 0.0280 | 0.0857 | 0.0489 | 0.8372 | 11.1027 |
| add 90% | 0.0140 | 0.0428 | 0.0244 | 0.9186 | 11.6002 |
| add 95% | 0.0070 | 0.0214 | 0.0122 | 0.9593 | 12.0925 |
| add 99% | 0.0014 | 0.0042 | 0.0024 | 0.9918 | 13.1646 |
| discount 90% | 0.1434 | 0.4387 | 0.2502 | 0.1675 | 10.2230 |
| discount 80% | 0.1466 | 0.4485 | 0.2558 | 0.1489 | 10.2281 |
| discount 70% | 0.1498 | 0.4583 | 0.2614 | 0.1303 | 10.2353 |
| discount 60% | 0.1530 | 0.4681 | 0.2669 | 0.1117 | 10.2451 |
| discount 50% | 0.1562 | 0.4780 | 0.2725 | 0.0930 | 10.2578 |
| discount 40% | 0.1595 | 0.4878 | 0.2781 | 0.0744 | 10.2743 |
| discount 30% | 0.1627 | 0.4976 | 0.2837 | 0.0558 | 10.2959 |
| discount 20% | 0.1659 | 0.5074 | 0.2893 | 0.0372 | 10.3250 |
| discount 10% | 0.1691 | 0.5172 | 0.2949 | 0.0186 | 10.3679 |
| discount 0% | 0.1723 | 0.5270 | 0.3005 | 0 | 10.4885 |




(a) l_3 boosted




(b) l_3 discounted

Figure 2: Cross-entropy of English and Czech data with a modified l_3 parameter

Figure 2 shows how the cross-entropy changes as the smoothing parameters change. It can be seen that the cross-entropy is lowest for the original trained lambdas for both languages, and in both adding and discounting the l_3 parameter. This would be expected to be the case, as the trained lambdas

represent the optimal values for the heldout data, so assuming a similar distribution between heldout and test data means that the original lambdas are close to optimal for the test data.

The cross-entropy for Czech is higher than it is for English. This suggests that the distribution learnt by the training algorithm is closer to that found in the test data. It can be seen that for English the coverage of words in the test data that have been seen in the training data is 75.8%, while for Czech it is just 65.2%. As seen in Table 3, the reason for this may again be because of the higher number of unique words in Czech arising from its rich morphology.

# A Code and Results Files

Note that `numpy` is required to run the code.

The code for Section 1 is in the file `entropy.py`. Running the file with `python entropy.py` will run the entire code for two text files called `TEXTEN1.txt` and `TEXTCZ1.txt`. The results of all experiments will be saved in files called `entropy_resultsEN.txt` and `entropy_resultsCZ.txt` respectively.

The code for Section 2 is in the file `lm.py`. Running the file with `python lm.py` will run the entire code for two text files called `TEXTEN1.txt` and `TEXTCZ1.txt`. The results of all experiments will be saved in files called `lm_resultsEN.txt` and `lm_resultsCZ.txt` respectively.