NPFL067 - Assignment 1

Thuong-Hai Pham

February 2018

1 Data overview

The provided input data are two files TEXTCZ1.txt (TEXTCZ1) and TEXTEN1.txt (TEXTEN1), which some basic properties are presented in Table 1.

Properties		TEXTCZ1	TEXTEN1
Text size (#tokens)		222412	221098
Vocabulary size (#words)		42826	9607
Text size / vocabulary size		5.193387	23.014260
Characters count		1030631	972917
Word length (#characters)	min	1	1
	avg	4.633882	4.400388
	max	28	18
Most frequent word	word	,	,
	freq. (#tokens)	13788	14721
Word(s) with freq. = 1	count (#words)	26315	3811
	%	61.446318%	39.668991%

Table 1: Data statistics

2 Entropy of a Text

2.1 Original texts

We first compute the conditional entropy (CE) and perplexity of the original texts. Table 2 shows the results, in which CE of TEXTCZ1 is smaller than TEXTEN1.

Data	Entropy	Perplexity
TEXTCZ1	4.747841	26.868452
TEXTEN1	5.287446	39.055279

Table 2: Conditional entropy and perplexity of the original texts

While $H(J|I) = -\sum_{i \in I, j \in J} P(i,j) \log_2 P(j|i)$, TEXTCZ1 has a higher percentage of words which have been seen only once (Table 1). Hence, TEXTCZ1 has higher conditional probabilities in its formula, which leads to smaller negative logarithms i.e. smaller conditional entropy.

2.2 Messing things up

Then, words and characters of the original texts are messed up with a list of given probabilities. Because of the random factor, each was repeated 10 times. Table 3 and Figure 1 present the results from the experiments.

	%	TEXTCZ1			TEXTEN1		
		min	avg.	max	min	avg.	max
Words	0.100000	4.633735	4.637704	4.641148	5.452914	5.457731	5.461215
	0.050000	4.694249	4.698542	4.703365	5.376085	5.380456	5.383470
	0.010000	4.738789	4.739497	4.740605	5.305614	5.306357	5.307111
	0.001000	4.746599	4.746873	4.747409	5.288634	5.289587	5.290370
	0.000100	4.747659	4.747765	4.747849	5.287408	5.287578	5.287801
	0.000010	4.747825	4.747863	4.747900	5.287444	5.287483	5.287566
Chars	0.100000	4.000039	4.014189	4.006891	4.728963	4.736937	4.731700
	0.050000	4.333974	4.342041	4.337972	5.049540	5.060010	5.056234
	0.010000	4.656389	4.660191	4.657792	5.247956	5.250353	5.249333
	0.001000	4.737460	4.739415	4.738394	5.283175	5.285097	5.283937
	0.000100	4.746881	4.747159	4.747035	5.286797	5.287380	5.287142
	0.000010	4.747700	4.747820	4.747767	5.287350	5.287449	5.287410

Table 3: Conditional entropy of messed texts (min, avg. and max of 10 times)

In Figure 1, it is obvious that when messing up characters more frequently, the CEs of both texts decrease because we have created more rare words, as explained in previous section.

However, messing up words have a strange behavior that while CE of TEXTCZ1 decreases, CE of TEXTEN1 increases. In TEXTCZ1, messing up a word is more likely to turn it into a rare word (61.4% of vocabulary), hence decreases the CE. However, the rate of decreasing is smaller than messing characters because there is also a chance to turn the word into non-rare word. In contrary, TEXTEN1 has a smaller percent of rare words (39.7%) in the vocabulary, so messing up a word is more likely to turn it into non-rare word, which increases the CE.

2.3 Concatenation

Two texts T_1, T_2 of two imaginary languages L_1, L_2 that do not share any vocabulary item are given, and that CEs of both are E. We create a new text T by appending T_2 to T_1.

Consider the notation of $C(i,j)$: count of bigram i followed by j.

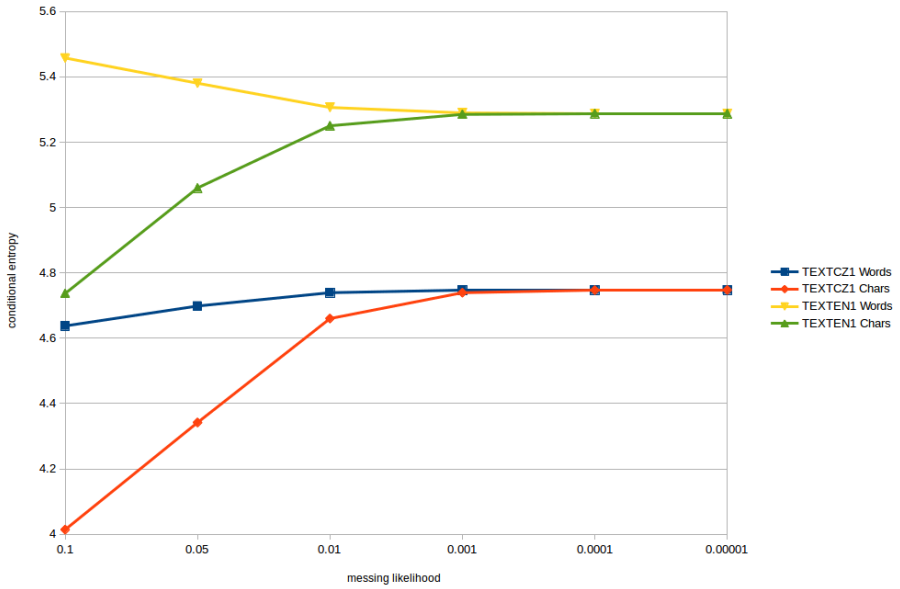


Figure 1: Conditional entropy of messed texts (avg. of 10 times)

$H(J|I) = -\sum_{i\in I, j\in J} P(i,j)\log_2 P(j|i)$, I, J are from the new text T. Because L_1, L_2 do not share vocabulary items,

$$H(J|I) = H(J_1|I_1) + H(J_2|I_2) + H(J_1|I_2) + H(J_2|I_1)$$

where J_1, I_1 are from T_1, J_2, I_2 are from T_2. Hence,

$$\begin{aligned}
H(J|I) &= -\sum_{i\in I_1, j\in J_1} P(i,j)\log_2 P(j|i) - \sum_{i\in I_2, j\in J_2} P(i,j)\log_2 P(j|i) + H(J_1|I_2) + H(J_2|I_1) \\
&= -\sum_{i\in I_1, j\in J_1} P(i,j)\log_2 P(j|i) - \sum_{i\in I_2, j\in J_2} P(i,j)\log_2 P(j|i) + H(J_1|I_2) + H(J_2|I_1) \\
&= -\sum_{i\in I_1, j\in J_1} \frac{C(i,j)}{|T|}\log_2 P(j|i) - \sum_{i\in I_2, j\in J_2} \frac{C(i,j)}{|T|}\log_2 P(j|i) + H(J_1|I_2) + H(J_2|I_1)
\end{aligned}$$

assume that T_1, T_2 have equal length

$$\begin{aligned}
&= -\sum_{i\in I_1, j\in J_1} \frac{C(i,j)}{2|T_1|}\log_2 P(j|i) - \sum_{i\in I_2, j\in J_2} \frac{C(i,j)}{2|T_2|}\log_2 P(j|i) + H(J_1|I_2) + H(J_2|I_1) \\
&= \frac{1}{2}E + \frac{1}{2}E + H(J_1|I_2) + H(J_2|I_1)
\end{aligned}$$

Because $H(J_1|I_2) + H(J_2|I_1) \geq 0$, the new conditional entropy $H(J|I) \geq E$.

3 Cross-Entropy and Language Modeling

Table 4 below presents the lambda values learned by EM algorithm using train set and held-out set, and the corresponded cross-entropy (XE) values when being tested against the test set for both texts.

		l0	l1	l2	l3	XE (test)
TEXTCZ1	train set	0.000000	0.000000	0.000000	1.000000	32.268907
	held-out set	0.140276	0.428926	0.244601	0.186197	10.220183
TEXTEN1	train set	0.000000	0.000000	0.000000	1.000000	23.701589
	held-out set	0.070058	0.253975	0.492288	0.183679	7.468154

Table 4: Lambdas tuned by train/held-out set, cross-entropy (XE) on test set

When using train set to tuned lambda values, we get into over-fitting, hence, leads to a higher cross-entropy, in which the algorithm decides to choose trigram only.

To observe this effect more closely, we tweak the values of l3 (tuned with held-out set). It is intuitive that changing l3 will increase the cross-entropy as we are getting away from the optimum learned by EM algorithm. However, because the values of l3 are closer to 0 (0.186197 in TEXTCZ1, 0.183679 in TEXTEN1), increasing l3 gives more 'error' than shrinking it (Table 5, Figure 2).

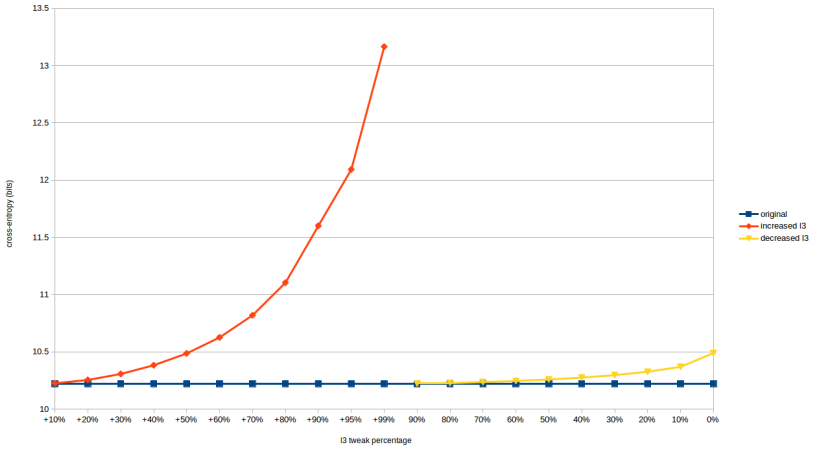
In Figure 2, both TEXTCZ1 and TEXTEN1 have similar rate of increasing cross-entropy when tweaking l3. Nonetheless, the absolute values from TEXTCZ1 are always higher than TEXTEN1 (even without tweaking l3 in Table 4), which can be explained by the lower coverage of unigrams, bigrams and trigrams in the test sets over train sets (Table 6).

		TEXTCZ1	TEXTEN1
original		10.220183	7.468154
increase	+10%	10.224665	7.469738
	+20%	10.254126	7.496635
	+30%	10.306235	7.546413
	+40%	10.382028	7.620124
	+50%	10.485595	7.721969
	+60%	10.625398	7.860674
	+70%	10.818398	8.053792
	+80%	11.103174	8.341477
	+90%	11.600713	8.850705
	+95%	12.093014	9.362038
	+99%	13.165277	10.501043
shrink	90%	10.223369	7.471996
	80%	10.228451	7.477708
	70%	10.235684	7.485535
	60%	10.245417	7.495815
	50%	10.258154	7.509026
	40%	10.274664	7.525886
	30%	10.296231	7.54756
	20%	10.325334	7.576209
	10%	10.36831	7.617011
	0%	10.4889	7.703941

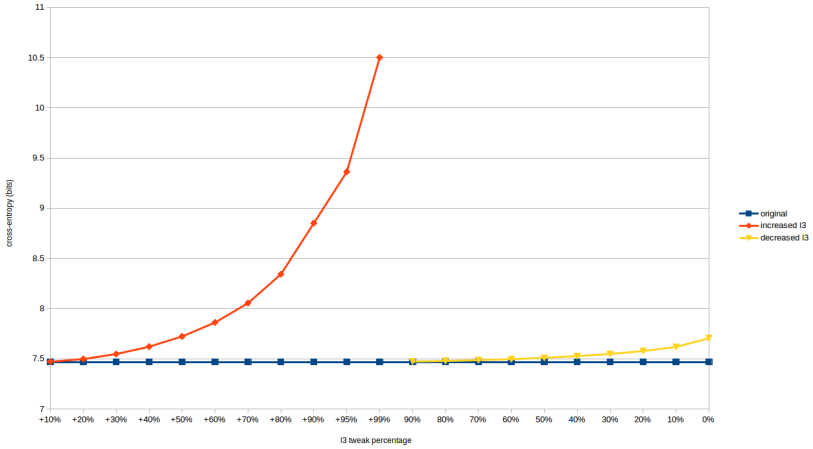
Table 5: Lambdas tuned by train/held-out set and cross-entropy computed on test set

	unigram	bigram	trigram
TEXTCZ1	65.172%	29.618%	12.292%
TEXTEN1	75.815%	52.592%	27.433%

Table 6: Test set coverage



(a) TEXTCZ1



(b) TEXTEN1

Figure 2: Cross-entropy on test set when tweaking l3