

Are All Languages Equally Hard to Language Model?

Cotterelle et al.

JungsooPark

Data Mining & Information Systems Lab.

Department of Computer Science and Engineering,
College of Informatics, Korea University

Introduction



Problem with Bits Per Character

$$p(\text{the cat sat on the mat}) = 0.0000000000341 \\ p(\text{the mat sat on the cat}) = 0.0000000000239 \\ p(\text{the cat the on mat mat}) = 0.0000000000001 \\ ppl_{word} = \exp\frac{-\log 0.0000000000341}{6+1} = 22.5 \\ ppl_{word} = \exp\frac{-\log 0.000000000341}{22+1} = 2.7 \\ p(\text{the}|\varepsilon) = 0.01 \\ p(\text{cat}|\text{the}) = 0.001 \\ p(\text{sat}|\text{the cat}) = 0.008 \\ p(\text{EOS}|\text{the cat sat}) = 0.04 \\ \Rightarrow p(\text{the cat sat}) = 0.1 \cdot 0.01 \cdot 0.008 \cdot 0.04 = 0.00000032$$

Introduction



Open Vocab vs Clsed Vocab

```
p_{word}(	exttt{the wolpertinger sat}) < p_{char}(	exttt{the wolpertinger sat})
```

"wolpertinger" is an oov

Therefore, the word–level language model will not assign probability to the word "wolpertinger"

$$p_{word}(exttt{the cat sat}) > p_{char}(exttt{the cat sat})$$

However when the in-domain word comes in, word-level will likely assign high probability

In other words, character-level LM's support doesn't match that of word-level LM's

Experiment



Fair evaluation metric across different languages?

Standard Perplexity(Bits Per Character)

$$\frac{1}{|c|+1} \sum_{i=1}^{|c|+1} \log p(c_i \mid c_{< i})$$



Czech puč Serman Putsch.

Proposed Metric(Bits Per English Character)

$$\frac{1}{|\boldsymbol{c}_{English}|+1} \sum_{i=1}^{|\boldsymbol{c}|+1} \log p(c_i \mid \boldsymbol{c}_{< i})$$

Experiment



MCC and Evaluation Table

bg 0.71/4.3 96 1.13/ 4 1.03/ 1 0.95/ 3 0.80/ cs 0.65/3.9 195 1.20/ -8 1.05/-12 0.97/ -6 0.83/ da 0.70/4.1 15 1.10/ -1 1.06/ -4 0.85/ -1 0.82/ de 0.74/4.8 38 1.25/ 17 1.18/ 13 1.04/ 14 0.90/ e1 0.75/4.6 50 1.18/ 13 1.08/ 5 0.90/ 10 0.82/ en 0.75/4.1 6 1.10/ 0 1.08/ -3 0.85/ 0 0.83/ es 0.81/4.6 71 1.15/ 12 1.07/ 7 0.87/ 9 0.80/ et* 0.55/3.9 110 1.20/ -8 1.11/-15 0.97/ -6 0.89/-fi* 0.52/4.2 198 1.18/ 2 1.02/-11 1.05/ 1 0.79/ fr 0.88/4.9 30 1.13/ 17 1.06/ 13 0.92/ 14 0.78/ hu* 0.63/4.3 94 1.25/ 5 1.12/ -9 1.09/ 5 0.89/ it 0.85/4.8 52 1.15/ 16 1.08/ 14 0.96/ 14 0.79/ lt 0.59/3.9 152 1.17/ -6 1.12/ -7 0.93/ -5 0.88/ lv 0.61/3.9 81 1.15/ -6 1.04/ -9 0.91/ -5 0.81/ nl 0.75/4.5 26 1.20/ 11 1.16/ 4 0.92/ 8 0.91/ p1 0.65/4.3 112 1.21/ 6 1.09/ -1 0.97/ 5 0.84/ pt 0.89/4.8 77 1.17/ 16 1.09/ 9 0.88/ 12 0.82/ ro 0.74/4.4 60 1.17/ 8 1.09/ 0 0.90/ 6 0.84/ sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/		BPEC / ΔBPC (e-2)					
bg 0.71/4.3 96 1.13/ 4 1.03/ 1 0.95/ 3 0.80/ cs 0.65/3.9 195 1.20/ -8 1.05/-12 0.97/ -6 0.83/ da 0.70/4.1 15 1.10/ -1 1.06/ -4 0.85/ -1 0.82/ de 0.74/4.8 38 1.25/ 17 1.18/ 13 1.04/ 14 0.90/ e1 0.75/4.6 50 1.18/ 13 1.08/ 5 0.90/ 10 0.82/ en 0.75/4.1 6 1.10/ 0 1.08/ -3 0.85/ 0 0.83/ es 0.81/4.6 71 1.15/ 12 1.07/ 7 0.87/ 9 0.80/ et* 0.55/3.9 110 1.20/ -8 1.11/-15 0.97/ -6 0.89/- fi* 0.52/4.2 198 1.18/ 2 1.02/-11 1.05/ 1 0.79/ fr 0.88/4.9 30 1.13/ 17 1.06/ 13 0.92/ 14 0.78/ hu* 0.63/4.3 94 1.25/ 5 1.12/ -9 1.09/ 5 0.89/ it 0.85/4.8 52 1.15/ 16 1.08/ 14 0.96/ 14 0.79/ lt 0.59/3.9 152 1.17/ -6 1.12/ -7 0.93/ -5 0.88/ lv 0.61/3.9 81 1.15/ -6 1.04/ -9 0.91/ -5 0.81/ p1 0.65/4.3 112 1.21/ 6 1.09/ -1 0.97/ 5 0.84/ pt 0.89/4.8 77 1.17/ 16 1.09/ 9 0.88/ 12 0.82/ ro 0.74/4.4 60 1.17/ 8 1.09/ 0 0.90/ 6 0.84/ sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/		data (M)		hybrid n-gram		LSTM	
cs 0.65/3.9 195 1.20/ -8 1.05/-12 0.97/ -6 0.83/ 0.70/4.1 15 1.10/ -1 1.06/ -4 0.85/ -1 0.82/ de 0.74/4.8 38 1.25/ 17 1.18/ 13 1.04/ 14 0.90/ e1 0.75/4.6 50 1.18/ 13 1.08/ 5 0.90/ 10 0.82/ en 0.75/4.1 6 1.10/ 0 1.08/ -3 0.85/ 0 0.83/ es 0.81/4.6 71 1.15/ 12 1.07/ 7 0.87/ 9 0.80/ et* 0.55/3.9 110 1.20/ -8 1.11/-15 0.97/ -6 0.89/- fi* 0.52/4.2 198 1.18/ 2 1.02/-11 1.05/ 1 0.79/ fr 0.88/4.9 30 1.13/ 17 1.06/ 13 0.92/ 14 0.78/ hu* 0.63/4.3 94 1.25/ 5 1.12/ -9 1.09/ 5 0.89/ it 0.85/4.8 52 1.15/ 16 1.08/ 14 0.96/ 14 0.79/ lt 0.59/3.9 152 1.17/ -6 1.12/ -7 0.93/ -5 0.88/ lv 0.61/3.9 81 1.15/ -6 1.04/ -9 0.91/ -5 0.81/ pl 0.65/4.3 112 1.21/ 6 1.09/ -1 0.97/ 5 0.84/ pt 0.89/4.8 77 1.17/ 16 1.09/ 9 0.88/ 12 0.82/ ro 0.74/4.4 60 1.17/ 8 1.09/ 0 0.90/ 6 0.84/ sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/	lang	wds / ch	MCC	form	lemma	form	lemma
da 0.70/4.1 15 1.10/ -1 1.06/ -4 0.85/ -1 0.82/ de 0.74/4.8 38 1.25/ 17 1.18/ 13 1.04/ 14 0.90/ el 0.75/4.6 50 1.18/ 13 1.08/ 5 0.90/ 10 0.82/ en 0.75/4.1 6 1.10/ 0 1.08/ -3 0.85/ 0 0.83/ es 0.81/4.6 71 1.15/ 12 1.07/ 7 0.87/ 9 0.80/ et* 0.55/3.9 110 1.20/ -8 1.11/-15 0.97/ -6 0.89/- fi* 0.52/4.2 198 1.18/ 2 1.02/-11 1.05/ 1 0.79/ fr 0.88/4.9 30 1.13/ 17 1.06/ 13 0.92/ 14 0.78/ hu* 0.63/4.3 94 1.25/ 5 1.12/ -9 1.09/ 5 0.89/ it 0.85/4.8 52 1.15/ 16 1.08/ 14 0.96/ 14 0.79/ lt 0.59/3.9 152 1.17/ -6 1.12/ -7 0.93/ -5 0.88/	bg	0.71/4.3	96	1.13/ 4	1.03/ 1	0.95/ 3	0.80/ 1
de 0.74/4.8 38 1.25/17 1.18/13 1.04/14 0.90/10 el 0.75/4.6 50 1.18/13 1.08/5 0.90/10 0.82/ en 0.75/4.1 6 1.10/0 1.08/-3 0.85/0 0.83/ es 0.81/4.6 71 1.15/12 1.07/7 0.87/9 0.80/ et* 0.55/3.9 110 1.20/-8 1.11/-15 0.97/-6 0.89/- fi* 0.52/4.2 198 1.18/2 1.02/-11 1.05/1 0.79/ fr 0.88/4.9 30 1.13/17 1.06/13 0.92/14 0.78/ hu* 0.63/4.3 94 1.25/5 5 1.12/-9 1.09/5 0.89/ it 0.85/4.8 52 1.15/16 1.08/14 0.96/14 0.79/ lt 0.59/3.9 152 1.17/-6 1.12/-7 0.93/-5 0.88/ lv 0.61/3.9 81 1.15/-6 1.04/-9 0.91/-5 0.81/	cs	0.65/3.9	195	1.20/ -8	1.05/-12	0.97/ -6	0.83/ -9
el 0.75/4.6 en 0.75/4.1 6 1.10/ 0 1.08/ -3 0.85/ 0 0.83/ es 0.81/4.6 71 1.15/ 12 1.07/ 7 0.87/ 9 0.80/ et* 0.55/3.9 110 1.20/ -8 1.11/-15 0.97/ -6 0.89/- fr 0.88/4.9 30 1.13/ 17 1.06/ 13 0.92/ 14 0.78/ hu* 0.63/4.3 94 1.25/ 5 1.12/ -9 1.09/ 5 0.89/ it 0.85/4.8 52 1.15/ 16 1.08/ 14 0.96/ 14 0.79/ lt 0.59/3.9 152 1.17/ -6 1.12/ -7 0.93/ -5 0.88/ nl 0.75/4.5 26 1.20/ 11 1.16/ 4 0.92/ 8 0.91/ pl 0.65/4.3 112 1.21/ 6 1.09/ -1 0.97/ 5 0.84/ pt 0.89/4.8 77 1.17/ 16 1.09/ 9 0.88/ 12 0.82/ ro 0.74/4.4 60 1.17/ 8 1.09/ 0 0.90/ 6 0.84/ sk 0.64/3.9	da	0.70/4.1	15	1.10/ -1	1.06/ -4	0.85/ -1	0.82/ -3
en 0.75/4.1 6 1.10/ 0 1.08/ -3 0.85/ 0 0.83/ es 0.81/4.6 71 1.15/ 12 1.07/ 7 0.87/ 9 0.80/ et* 0.55/3.9 110 1.20/ -8 1.11/-15 0.97/ -6 0.89/- fi* 0.52/4.2 198 1.18/ 2 1.02/-11 1.05/ 1 0.79/ fr 0.88/4.9 30 1.13/ 17 1.06/ 13 0.92/ 14 0.78/ hu* 0.63/4.3 94 1.25/ 5 1.12/ -9 1.09/ 5 0.89/ it 0.85/4.8 52 1.15/ 16 1.08/ 14 0.96/ 14 0.79/ lt 0.59/3.9 152 1.17/ -6 1.12/ -7 0.93/ -5 0.88/ lv 0.61/3.9 81 1.15/ -6 1.04/ -9 0.91/ -5 0.81/ nl 0.75/4.5 26 1.20/ 11 1.16/ 4 0.92/ 8 0.91/ pl 0.65/4.3 112 1.21/ 6 1.09/ -1 0.97/ 5 0.84/ pt 0.89/4.8 77 1.17/ 16 1.09/ 9 0.88/ 12 0.82/ ro 0.74/4.4 60 1.17/ 8 1.09/ 0 0.90/ 6 0.84/ sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/	de	0.74/4.8	38	1.25/ 17	1.18/ 13	1.04/ 14	0.90/ 10
es 0.81/4.6 71 1.15/12 1.07/ 7 0.87/ 9 0.80/ et* 0.55/3.9 110 1.20/ -8 1.11/-15 0.97/ -6 0.89/- fi* 0.52/4.2 198 1.18/ 2 1.02/-11 1.05/ 1 0.79/ fr 0.88/4.9 30 1.13/ 17 1.06/ 13 0.92/ 14 0.78/ hu* 0.63/4.3 94 1.25/ 5 1.12/ -9 1.09/ 5 0.89/ it 0.85/4.8 52 1.15/ 16 1.08/ 14 0.96/ 14 0.79/ lt 0.59/3.9 152 1.17/ -6 1.12/ -7 0.93/ -5 0.88/ lv 0.61/3.9 81 1.15/ -6 1.04/ -9 0.91/ -5 0.81/ n1 0.75/4.5 26 1.20/ 11 1.16/ 4 0.92/ 8 0.91/ pl 0.65/4.3 112 1.21/ 6 1.09/ -1 0.97/ 5 0.84/ pt 0.89/4.8 77 1.17/ 16 1.09/ 9 0.88/ 12 0.82/ ro 0.74/4.4 60 1.17/ 8 1.09/ 0 0.90/ 6 0.84/ sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/	el	0.75/4.6	50	1.18/ 13	1.08/ 5	0.90/ 10	0.82/ 4
et* 0.55/3.9 110 1.20/ -8 1.11/-15 0.97/ -6 0.89/- fi* 0.52/4.2 198 1.18/ 2 1.02/-11 1.05/ 1 0.79/ fr 0.88/4.9 30 1.13/ 17 1.06/ 13 0.92/ 14 0.78/ hu* 0.63/4.3 94 1.25/ 5 1.12/ -9 1.09/ 5 0.89/ it 0.85/4.8 52 1.15/ 16 1.08/ 14 0.96/ 14 0.79/ lt 0.59/3.9 152 1.17/ -6 1.12/ -7 0.93/ -5 0.88/ lv 0.61/3.9 81 1.15/ -6 1.04/ -9 0.91/ -5 0.81/ nl 0.75/4.5 26 1.20/ 11 1.16/ 4 0.92/ 8 0.91/ pl 0.65/4.3 112 1.21/ 6 1.09/ -1 0.97/ 5 0.84/ pt 0.89/4.8 77 1.17/ 16 1.09/ 9 0.88/ 12 0.82/ ro 0.74/4.4 60 1.17/ 8 1.09/ 0 0.90/ 6 0.84/ sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/	en	0.75/4.1	6	1.10/ 0	1.08/ -3	0.85/ 0	0.83/ -3
fi* 0.52/4.2 198 1.18/ 2 1.02/-11 1.05/ 1 0.79/ fr 0.88/4.9 30 1.13/ 17 1.06/ 13 0.92/ 14 0.78/ hu* 0.63/4.3 94 1.25/ 5 1.12/ -9 1.09/ 5 0.89/ it 0.85/4.8 52 1.15/ 16 1.08/ 14 0.96/ 14 0.79/ lt 0.59/3.9 152 1.17/ -6 1.12/ -7 0.93/ -5 0.88/ lv 0.61/3.9 81 1.15/ -6 1.04/ -9 0.91/ -5 0.81/ nl 0.75/4.5 26 1.20/ 11 1.16/ 4 0.92/ 8 0.91/ pl 0.65/4.3 112 1.21/ 6 1.09/ -1 0.97/ 5 0.84/ pt 0.89/4.8 77 1.17/ 16 1.09/ 9 0.88/ 12 0.82/ ro 0.74/4.4 60 1.17/ 8 1.09/ 0 0.90/ 6 0.84/ sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/	es	0.81/4.6	71	1.15/ 12	1.07/ 7	0.87/ 9	0.80/ 5
fr 0.88/4.9 30 1.13/17 1.06/13 0.92/14 0.78/ hu* 0.63/4.3 94 1.25/5 1.12/9 1.09/5 0.89/ it 0.85/4.8 52 1.15/16 1.08/14 0.96/14 0.79/ lt 0.59/3.9 152 1.17/6 1.12/7 0.93/5 0.88/ lv 0.61/3.9 81 1.15/6 1.04/9 0.91/5 0.81/ n1 0.75/4.5 26 1.20/11 1.16/4 0.92/8 0.91/ pl 0.65/4.3 112 1.21/6 1.09/9 0.88/12 0.82/ ro 0.74/4.4 60 1.17/8 1.09/9 0.90/6 0.84/ sk 0.64/3.9 40 1.16/6 1.06/-11 0.92/5 0.87/	et*	0.55/3.9	110	1.20/ -8	1.11/-15	0.97/ -6	0.89/-12
hu* 0.63/4.3 94 1.25/5 1.12/-9 1.09/5 0.89/ it 0.85/4.8 52 1.15/16 1.08/14 0.96/14 0.79/ lt 0.59/3.9 152 1.17/-6 1.12/-7 0.93/-5 0.88/ lv 0.61/3.9 81 1.15/-6 1.04/-9 0.91/-5 0.81/ nl 0.75/4.5 26 1.20/11 1.16/-4 0.92/-8 0.91/- pl 0.65/4.3 112 1.21/-6 1.09/-1 0.97/-5 0.84/- pt 0.89/4.8 77 1.17/-16 1.09/-9 0.88/-12 0.82/- ro 0.74/4.4 60 1.17/-8 1.09/-0 0.90/-6 0.84/- sk 0.64/3.9 40 1.16/-6 1.06/-11 0.92/-5 0.87/-	fi^*	0.52/4.2	198	1.18/ 2	1.02/-11	1.05/ 1	0.79/ -9
it 0.85/4.8 52 1.15/ 16 1.08/ 14 0.96/ 14 0.79/ 1t 0.59/3.9 152 1.17/ -6 1.12/ -7 0.93/ -5 0.88/ 1v 0.61/3.9 81 1.15/ -6 1.04/ -9 0.91/ -5 0.81/ nl 0.75/4.5 26 1.20/ 11 1.16/ 4 0.92/ 8 0.91/ pl 0.65/4.3 112 1.21/ 6 1.09/ -1 0.97/ 5 0.84/ pt 0.89/4.8 77 1.17/ 16 1.09/ 9 0.88/ 12 0.82/ ro 0.74/4.4 60 1.17/ 8 1.09/ 0 0.90/ 6 0.84/ sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/	fr	0.88/4.9	30	1.13/ 17	1.06/ 13	0.92/ 14	0.78/ 10
lt 0.59/3.9 152 1.17/ -6 1.12/ -7 0.93/ -5 0.88/ lv 0.61/3.9 81 1.15/ -6 1.04/ -9 0.91/ -5 0.81/ nl 0.75/4.5 26 1.20/ 11 1.16/ 4 0.92/ 8 0.91/ pl 0.65/4.3 112 1.21/ 6 1.09/ -1 0.97/ 5 0.84/ pt 0.89/4.8 77 1.17/ 16 1.09/ 9 0.88/ 12 0.82/ ro 0.74/4.4 60 1.17/ 8 1.09/ 0 0.90/ 6 0.84/ sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/	hu*	0.63/4.3	94	1.25/ 5	1.12/ -9	1.09/ 5	0.89/ -7
1v 0.61/3.9 81 1.15/ -6 1.04/ -9 0.91/ -5 0.81/ n1 0.75/4.5 26 1.20/ 11 1.16/ 4 0.92/ 8 0.91/ pl 0.65/4.3 112 1.21/ 6 1.09/ -1 0.97/ 5 0.84/ pt 0.89/4.8 77 1.17/ 16 1.09/ 9 0.88/ 12 0.82/ ro 0.74/4.4 60 1.17/ 8 1.09/ 0 0.90/ 6 0.84/ sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/	it	0.85/4.8	52	1.15/ 16	1.08/ 14	0.96/ 14	0.79/ 10
nl 0.75/4.5 26 1.20/ 11 1.16/ 4 0.92/ 8 0.91/ pl 0.65/4.3 112 1.21/ 6 1.09/ -1 0.97/ 5 0.84/ pt 0.89/4.8 77 1.17/ 16 1.09/ 9 0.88/ 12 0.82/ ro 0.74/4.4 60 1.17/ 8 1.09/ 0 0.90/ 6 0.84/ sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/	lt	0.59/3.9	152	1.17/ -6	1.12/ -7	0.93/ -5	0.88/ -6
pl 0.65/4.3 pt 0.89/4.8 ro 0.74/4.4 60 1.16/ -6 1.06/-11 0.92/ -5 0.84/ sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/	lv	0.61/3.9	81	1.15/ -6	1.04/ -9	0.91/ -5	0.81/ -7
pt 0.89/4.8 77 1.17/ 16 1.09/ 9 0.88/ 12 0.82/ ro 0.74/4.4 60 1.17/ 8 1.09/ 0 0.90/ 6 0.84/ sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/	nl	0.75/4.5	26	1.20/ 11	1.16/ 4	0.92/ 8	0.91/ 4
ro 0.74/4.4 60 1.17/ 8 1.09/ 0 0.90/ 6 0.84/ sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/	pl	0.65/4.3	112	1.21/ 6	1.09/ -1	0.97/ 5	0.84/ -1
sk 0.64/3.9 40 1.16/ -6 1.06/-11 0.92/ -5 0.87/	pt	0.89/4.8	77	1.17/ 16	1.09/ 9	0.88/ 12	0.82/ 7
	ro	0.74/4.4	60	1.17/ 8	1.09/ 0	0.90/ 6	0.84/ 0
	sk	0.64/3.9	40	1.16/ -6	1.06/-11	0.92/ -5	0.87/ -9
sl 0.64/3.8 100 1.15/-10 1.02/-10 0.90/ -8 0.80/	sl	0.64/3.8	100	1.15/-10	1.02/-10	0.90/ -8	0.80/ -7
sv 0.66/4.1 35 1.11/ -2 1.06/ -8 0.86/ -2 0.83/	sv	0.66/4.1	35	1.11/ -2	1.06/ -8	0.86/ -2	0.83/ -7

```
n.lemmatize('dies', 'v')

'die'

n.lemmatize('watched', 'v')

'watch'

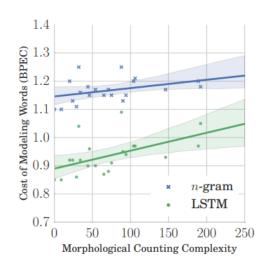
n.lemmatize('has', 'v')

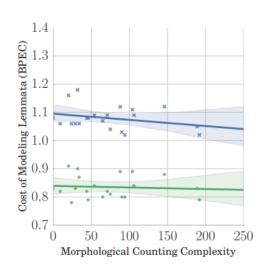
'have'
```

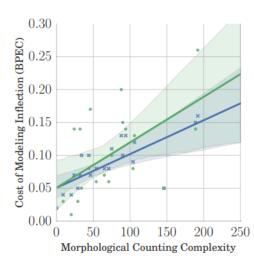
Experiment



Experiment Result







- High Counting Complexity results in high BPEC
- If lemmatized, no correlation between counting complexity and BPEC
- Inflectional morphology is the main culprit