Are All Languages Equally Hard to Language-Model?

Ryan Cotterell[1], **Sabrina J. Mielke**[1], **Jason Eisner**[1], **Brian Roark**[2]

[1]*Department of Computer Science, Johns Hopkins University, Baltimore, USA* [2]*Google*

Q1: Are all languages equally hard to model?

Not with current models, all models we test have very different performance on different languages.

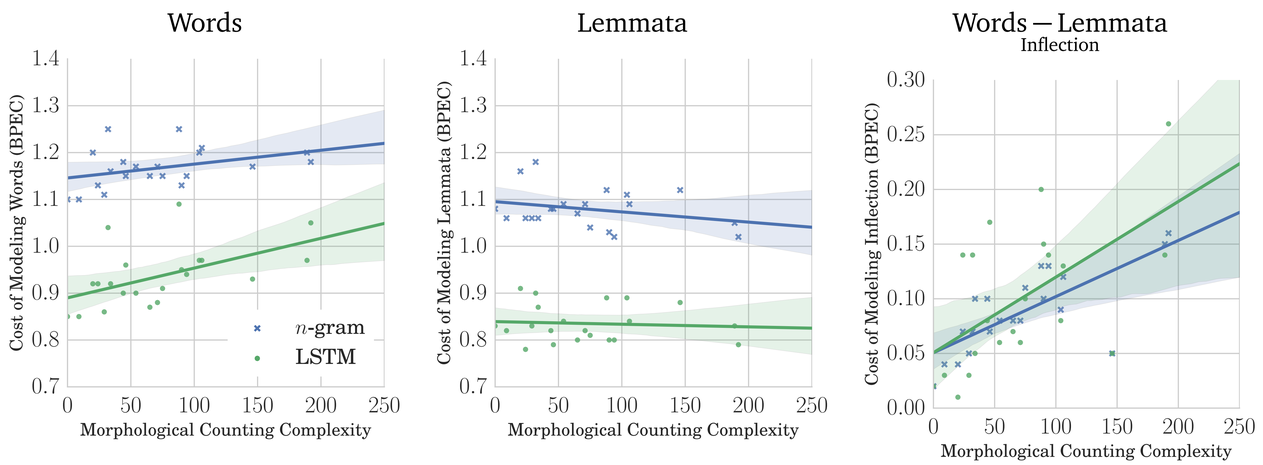
Q2: So what makes a language hard to model?

Hypothesis: inflectional morphology.

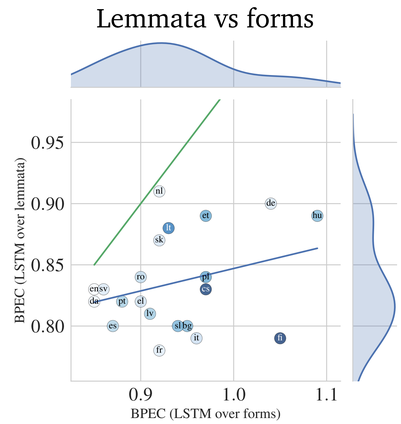
→ LM performance negatively correlated with *morphological counting complexity* (MCC; Sagot, 2013)

→ Correlation disappears when modeling lemmata (obtained using UDPipe (Straka et al., 2016)) instead of forms

Information contained in...



Each point is a language, the cost of modeling is plotted against the MCC of a language.



All languages are below $y = x$. Richer morphology (darker circles) ⇝ farther from it.

Differing corpora are unfair → Multi-text

Domain differences impact estimates, so use *aligned parallel text*, 21 usable languages in Europarl (Koehn, 2005):

bg/cs/da/de/el/en/es/fr/it lt/lv/nl/pl/pt/ro/sk/sl/sv (Indo-European) et/fi/hu (Uralic)

Translationese (Baker, 1993, translations is stylistically different from "native" text) only *underestimates* the difficulty of non-English languages—and we find the opposite!

Closed-vocab is unfair → Open-vocab LMs

Replacing rare words with UNK leads to unfairly good scores for languages with many word types (e.g., morphologically rich languages).
So instead use *open-vocabulary* LMs like:

1. Kneser-Ney smoothed **n-gram** LM over "flat" hybrid representations (Bisani and Ney, 2005):

   the G A N EOW trains EOS (3-gram)

2. A character-level **LSTM** LM (Sundermeyer et al., 2012; Zaremba et al., 2014):

$$p(c_t \mid \mathbf{c}_{<t}) = \text{softmax}(W h_i + b)$$
$$h_i = \text{LSTM}(h_{i-1}, c_{i-1})$$

BPC is unfair → Bits per English character

BPC (i.e., the information contained in one character) values depend on a language and cannot be compared. Example: these three strings that contain equal information (total cross-entropy is around 6 bits for each of them), but different BPC:

Sentence/word/lemma	Σ bits	BPC	BPEC
EN c o u p	≈ 6	$6/4 = 1.5$	$6/4 = 1.5$
DE P u t s c h	≈ 6	$6/6 = 1.0$	$6/4 = 1.5$
CZ p u č	≈ 6	$6/3 = 2.0$	$6/4 = 1.5$

We normalize the total number of bits (i.e., information) for length, arbitrarily choosing the number of English characters in the utterance, obtaining *bits per **English** character* (BPEC).