Are All Languages Equally Hard to Language-Model?

Ryan Cotterell[1] and **Sebastian J. Mielke**[1] and **Jason Eisner**[1] and **Brian Roark**[2]
[1] Department of Computer Science, Johns Hopkins University [2] Google
{ryan.cotterell@,sjmielke@,jason@cs.}jhu.edu roark@google.com

1 Overview

How cross-linguistically applicable are NLP models, specifically *language models*? A fair comparison between languages is tricky: not only do training corpora in different languages have different sizes and topics, some of which may be harder to predict than others, but standard metrics for language modeling depend on the orthography of a language. We argue for a fairer metric based on the bits per utterance using utterance-aligned multi-text. We conduct a study on 21 languages, training and testing both n-gram and LSTM language models on "the same" set of utterances in each language (modulo translation), demonstrating that in some languages, especially those with complex inflectional morphology, the textual expression of the information is harder to predict.

2 Open-Vocab BPEC on Multi-Text

Multi-Text. To avoid the problem of incomparable corpora, we use *multi-text*: k-way translations of the same semantic content, so each document, in principle, contains the same information.

Open-Vocabulary Language models. Most language models operate on a *word level*, employing a distinguished symbol UNK that represents all word types not present in their training vocabulary. This makes the task easier (because the model only predicts the presence of a rare word, but not the word itself), making the comparison unfair, especially in morphologically rich languages, which simply have more word forms and would thus more often only have to predict UNK. We thus require our language models to be "open-vocabulary": they predict every *character* in an utterance, rather than skipping some characters because they appear in words that were (arbitrarily) replaced by UNK in that language.

Bits per English Character. Open-vocabulary LMs are most commonly evaluated under *bits per character* (BPC). Even with multi-text, however, comparing BPC is not fair, as it relies on the vagaries of individual writing systems: consider the Czech word *puč* and its German equivalent *Putsch*. Even if these words are both predicted with the *same* probability in a given context, German will end up with a lower BPC, because the phoneme /tʃ/ is expressed with *tsch* instead of *č*, spreading the information over more characters.

Luckily, multi-text allows us to compute a fair metric that is invariant to the orthographic (or phonological) changes discussed above: the *bits per utterance*. To control for length, we divide this number for every language by the *same* factor, *arbitrarily* chosen to be average English characters per utterance, yielding *bits per English character* (BPEC). Any other choice of language would simply scale the values by a constant factor.

3 Inflectional Morphology

Inflectional morphology increases the number of word types in a language. The English lexeme BOOK for example only has the singular *book* and the plural *books*. The Turkish lexeme KITAP, in contrast, distinguishes at least 12 forms.[1]

To compare the degree of morphological inflection in our evaluation languages, we use *counting complexity* (Sagot, 2013), a simple metric that counts the number of inflectional categories distinguished by a language.[2]

To crudely "control" for the inflection, we also perform experiments on *lemmatized* text, where we replace every word with its lemma,[3] stripping away its inflectional morphology.[4]

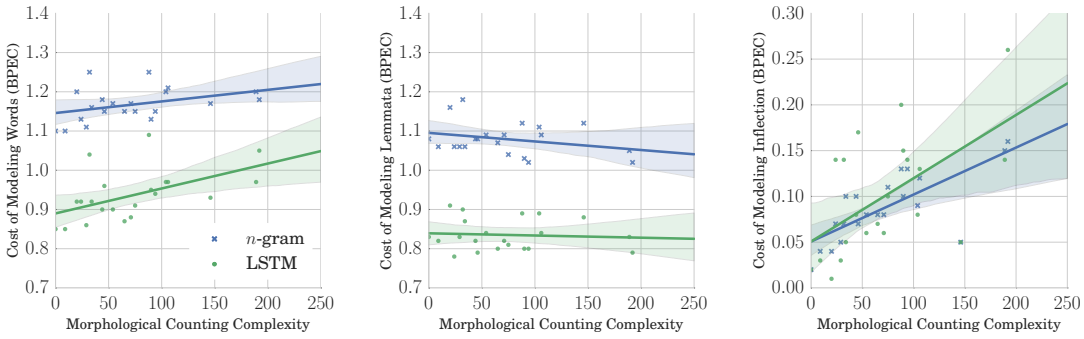
[1] The exact number depends on what forms are considered part of the nominal paradigm (Underhill, 1976).

[2] We count the categories annotated in the language's UniMorph (Kirov et al., 2018) lexicon.

[3] We use UDPipe (Straka et al., 2016) to obtain lemmata.

[4] Our BPEC measure always normalizes by the length of

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 361-362.
New York City, New York, January 3-6, 2019



(a) BPEC for word sequences (b) BPEC for lemma sequences (c) Difference: (a) − (b)

Figure 1: BPEC performance of n-gram (blue) and LSTM (green) LMs; each point is a language. The correlation between LM performance and counting complexity disappears after lemmatization of the corpus (from (a) to (b)), indicating that inflectional morphology is the origin for the lower BPEC.

4 Experimental setup

For each language, we train a "flat" hybrid word/character open-vocabulary n-gram[5] model (Bisani and Ney, 2005) and a long short-term memory (LSTM)[6] language model (Sundermeyer et al., 2012, but at the character-level), for the original and lemmatized text each.

Our experiments are conducted on the 21 languages of the Europarl corpus (Koehn, 2005), which consists of cross-linguistically aligned utterances made in the European parliament. With the exceptions of Finnish, Hungarian and Estonian, which are Uralic, the languages are Indo-European.

5 Discussion and Analysis

The results are shown in Fig. 1. While it is not surprising to see LSTM models outperform the baseline n-gram models across the board, it is interesting to see that rich inflectional morphology is a difficulty for both n-gram and LSTM LMs.[7] Studying Fig. 1a, we find that Spearman's rank correlation between a language's BPEC and its counting complexity (§3) is quite high.[8] This clear correlation between the level of inflectional morphology and the LSTM performance indicates that *character-level models do not automatically fix the problem of morphological richness*. If we lemmatize the words, however (Fig. 1b), the correlation disappears.[9] The difference of the two previous graphs (Fig. 1c) shows more clearly that the LM penalty for modeling inflectional endings is greater for languages with higher counting complexity.[10]

Why is this? (1) Text in highly inflected languages may be *inherently harder to predict* (higher entropy per utterance) if its extra morphemes carry additional, unpredictable information. (2) Alternatively, perhaps the extra morphemes are *predictable in principle*—for example, redundant marking of grammatical number on both subjects and verbs, or marking of object case even when it is predictable from semantics or word order—but our current language modeling technology fails to predict them.

the original, not lemmatized, English.

[5]We use 5-grams of words and 7-grams for hybrid strings.

[6]We use 1024-dim. character embeddings, 2 1024-dim. hidden layers, and 100 iterations (with early stopping) of SGD with gradients clipped to 5.

[7]In this section we give numbers for the LSTMs.

[8]$\rho = 0.59$, significant at $p < 0.005$

[9]$\rho = -0.13$, $p \approx 0.56$

[10]Indeed, this penalty is arguably a more appropriate measure of the complexity of the inflectional system.

References

M. Bisani and H. Ney. 2005. Open vocabulary speech recognition with flat hybrid models. In *INTERSPEECH*.

C. Kirov, R. Cotterell, J. Sylak-Glassman, G. Walther, E. Vylomova, P. Xia, M. Faruqui, A. McCarthy, S. J. Mielke, S. Kübler, D. Yarowsky, J. Eisner, and M. Hulden. 2018. Unimorph 2.0: Universal morphology. In *LREC*.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

B. Sagot. 2013. Comparing complexity measures. In *Comp. Approaches to Morph. Complexity*.

M. Straka, J. Hajič, and J. Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *LREC*.

M. Sundermeyer, R. Schlüter, and H. Ney. 2012. LSTM neural networks for language modeling. In *INTERSPEECH*.

R. Underhill. 1976. *Turkish Grammar*. MIT Press.