Better Character Language Modeling Through Morphology

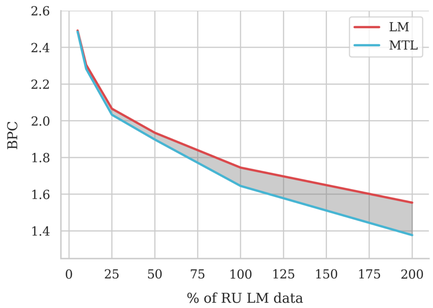
Terra Blevins[1] and Luke Zettlemoyer[1,2]

[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Facebook AI Research, Seattle

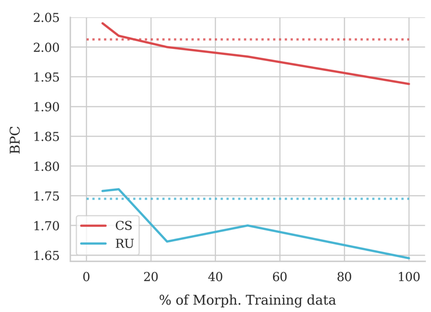
Morphology Improves CLMs

- Character language models (CLMs) have the capacity to share subword infromation across morphological forms.
- **Hypothesis**: accurately modeling morphology improves CLM performance, but it is difficult for CLMs to learn this from the language modeling objective alone.
- We incorporate morphology annotations into a CLM using a multi-task objective.
- Adding morphology into CLMs improves bits-per-character (BPC) across 24 languages, even when the LM and morphology data is disjoint.

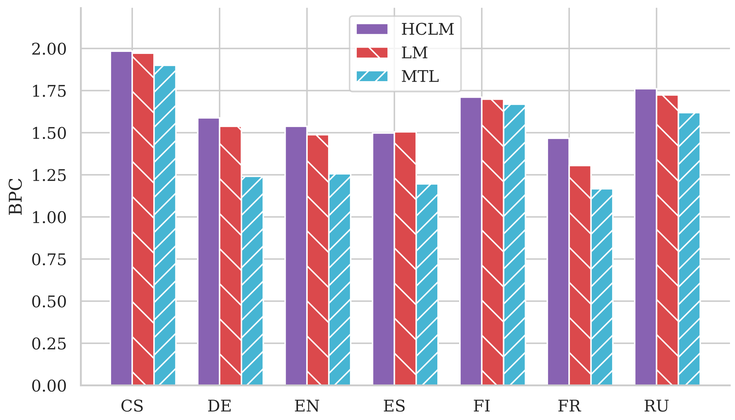
Effect of Training Data Quantity



BPC performance on MWC dev set when varying the amount of LM training data

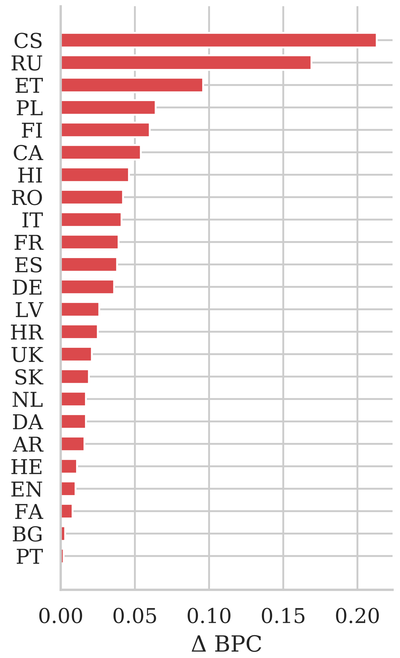


BPC performance on MWC dev set when varying the amount of morphology training data (dashed line is LM baseline)

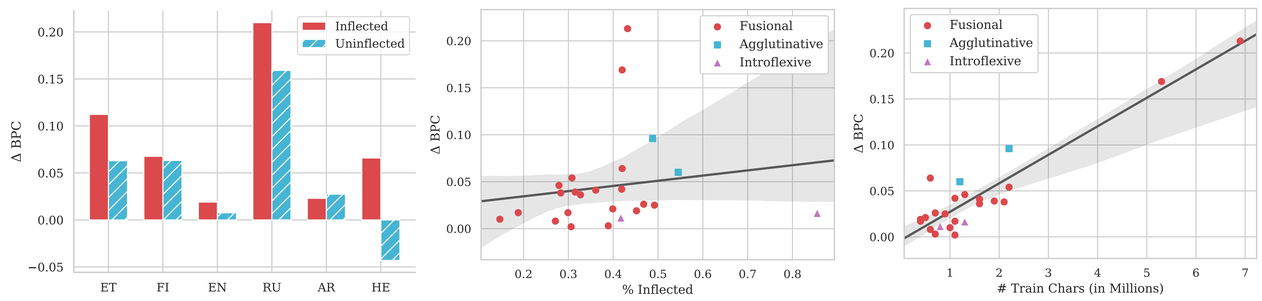


Language Modeling Results

- **Above**: BPC performance on MWC test set of **HCLM** (best model from *Kawakami et al., 2017*), our baseline **LM**, and the **MTL** model.
- **Right**: Improvement in BPC of the **MTL** model over the **LM** baseline on the UD test set for 24 languages.
- Find that across all languages, BPC improves when morphology supervision is added to vanilla CLM.



What drives this improvement in CLM performance?



(A) **(B)** **(C)**

(A) Improvement in BPC of MTL model over the LM baseline on inflected and uninflected forms in the UD development set.

(B) Improvement in BPC of MTL model over the LM baseline over the percentage of inflected forms for each language. (r = 0.15)

(C) Improvement in BPC of MTL model over the LM baseline over the amount of training data available for each language in the UD dataset. (r = 0.93)

Methodology + Models

- Incorporate morphology supervison into model via multi-task learning:

$$\mathcal{L}(\mathbf{c}, \mathbf{m}) = \mathcal{L}_{\text{LM}}(\mathbf{c}) + \delta \sum_{i=1}^{n} \mathcal{L}_i(\mathbf{m})$$

- The LM architectures consist of a stacked LSTM model with the layer at which we multi-task morphology selected as a hyperparameter.
- We train both baseline LMs (LM) and mulitasked LMs (MTL) on the text of Univeral Dependencies (UD) for 24 langauges, as well as on the Multilingual Wikipedia Corpus (MWC).
- UD morphological features are used as supervision for all MTL models.

Cross-Lingual Transfer

- Can morphology from a high-resouce language improve LMs on a low-resource, typologically similar language?
- We incorporate additional data from a high-resource language into CLMs for a related, low-resource one.
- Find that adding both LM data and morphological features helps model the low-resource language.

LM data	Morph. data	BPC
SK	None	2.806
	SK	2.779
	CS	2.752
	CS+SK	2.777
CS+SK	None	2.668
	CS+SK	**2.446**
UK	None	2.369
	UK	2.348
	RU	2.348
	RU+UK	2.351
RU+UK	None	2.495
	RU+UK	**2.316**