

Lawrence Wolf-Sonkin*
Johns Hopkins University

Jason Naradowsky*
Johns Hopkins University

Sebastian J. Mielke*
Johns Hopkins University

Ryan Cotterell*
University of Cambridge

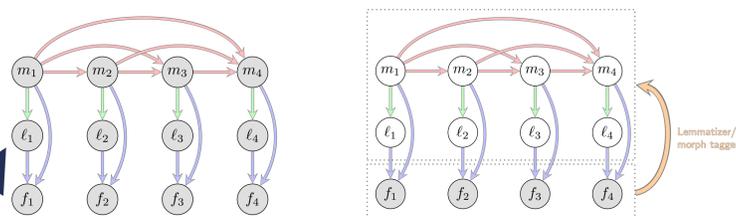
Morphological Inflection

hablar + V;IND;PST;3;SG;PF → habló

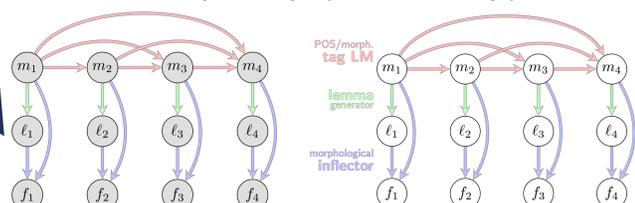
- Typically trained on type-level lexicons
- SOTA methods are generally neural and extremely data-hungry
- what to do for low-resource languages?

Parameter Estimation: Wake Sleep Algorithm

Wake Step (like E step)



Sleep Step (like M step)

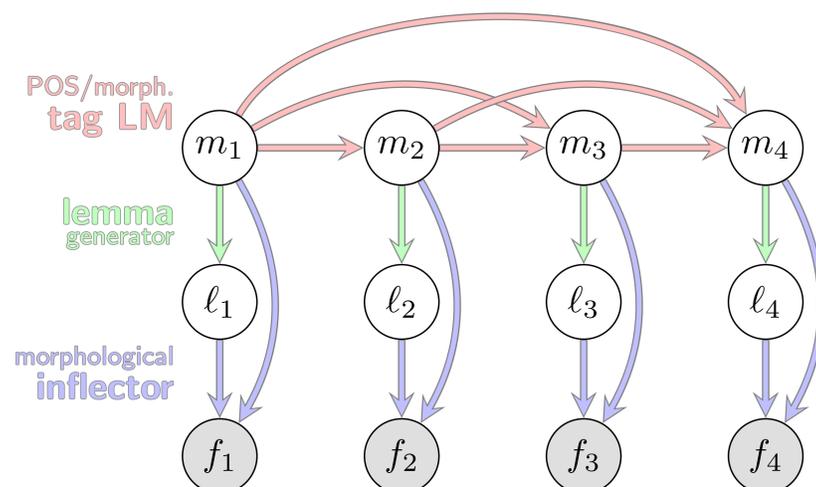


Alternate between 2 steps

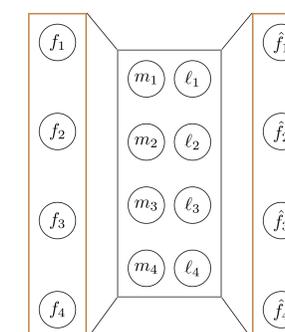
A Generative Model of Inflected-Form Sequences

Tag Sequence: PRN;GEN N;PL ADV V;V.PTCP;PST
 Lemma Sequence: I wug gently weep
 Form Sequence: my wugs gently wept

$$p_{\theta}(\mathbf{m}, \mathbf{l}, \mathbf{f}) = \underbrace{p_{\theta}(\mathbf{m})}_{\text{m-tag LM (1)}} \prod_{i=1}^{|\mathbf{f}|} \underbrace{p_{\theta}(l_i | m_i)}_{\text{lemma generator (2)}} \cdot \underbrace{p_{\theta}(f_i | l_i, m_i)}_{\text{morphological inflector (3)}}$$



View as an Autoencoder



- Encoder is morphological tag joint tagger/lemmatizer
- Latent space prior is morphological tag LM and lemma generator
- Decoder is morphological inflector

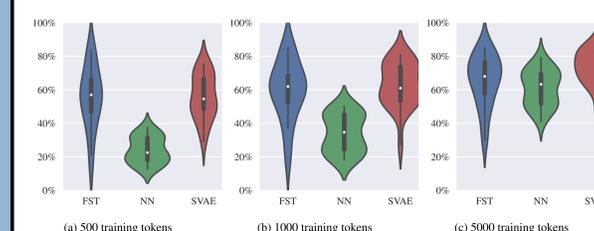
Data Provenance

- Model requires token-level data
 - Universal Dependencies (UD)
- 23 typologically diverse languages
 - Romance
 - Slavic
 - Semitic
 - Germanic

Evaluation

- Evaluate using morphological inflection accuracy
- Consider annotated dataset sizes consisting of 500, 1000, 5000 tokens
- Baselines:**
 - high-resource SOTA NN seq2seq model for inflection
 - FST baseline from CoNNL-SIGMORPHON 2017

Results



- In 500 token case, FST wins, and as we ramp up to 5000, SVAE wins