A Structured Variational Autoencoder for Contextual Morphological Inflection

Parameter Estimation:
Wake Sleep Algorithm

- Wake Step (like E step)
- Sleep Step (like M step)

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

Morphological Inflection

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

A Generative Model of Inflected-Form Sequences

Parameter Estimation:
SOTA methods are generally trained on type-level
Morphological Inflection

POS/morph.
tag LM

Results

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

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