Motivation:

To understand the power of open-text generation models, estimate how similar are the style of generated text to the reference text.

Introduction

Theory

Components of Style:

* Word Choices
* Sentence Structure
* Imagery

Models

* GPT-2 (pretrained) hugging face
* GPT-Neo (pretrained) hugging face
* RNN
* N gram as baseline

Sampling

* Temperature sampling
* Top-k
* Top – p

Evaluation

Semantic Preservation:

* Corpus- Bleu
* M-Bleu (named entities replaced with <M>)

Style Strength:

Accuracy by separately trained style classifier:

* Model trained on trump tweets (take trained one)
* Model trained on speeches (to train) How to? Should I remove named entities from data? trained only on POS?

Fluency – as it doesn’t affect style, no need to measure

Data

* Trump rally speeches – 370480 words
* Other speeches (Biden, Obama, Kenye, Bernie, Mike, Warren) - 109709

Method

* Train models on trump speeches, use different sampling to generate text
* Train style classifier to classify how Trumpish is the generated text (POS, data without named entities, should I use word embedding (embeddings give meaning of text not the style?)?)
* Evaluate with BLEU and classifiers.

Some more questions:

Can I use same data for generator training and classifier?

For BLEU estimation reference should be Validation data or can be all data, Should model be conditioned?

Should I generate few sentences or whole speech?