Preliminary Study on Automatic Grammaticality Evaluation

Yang

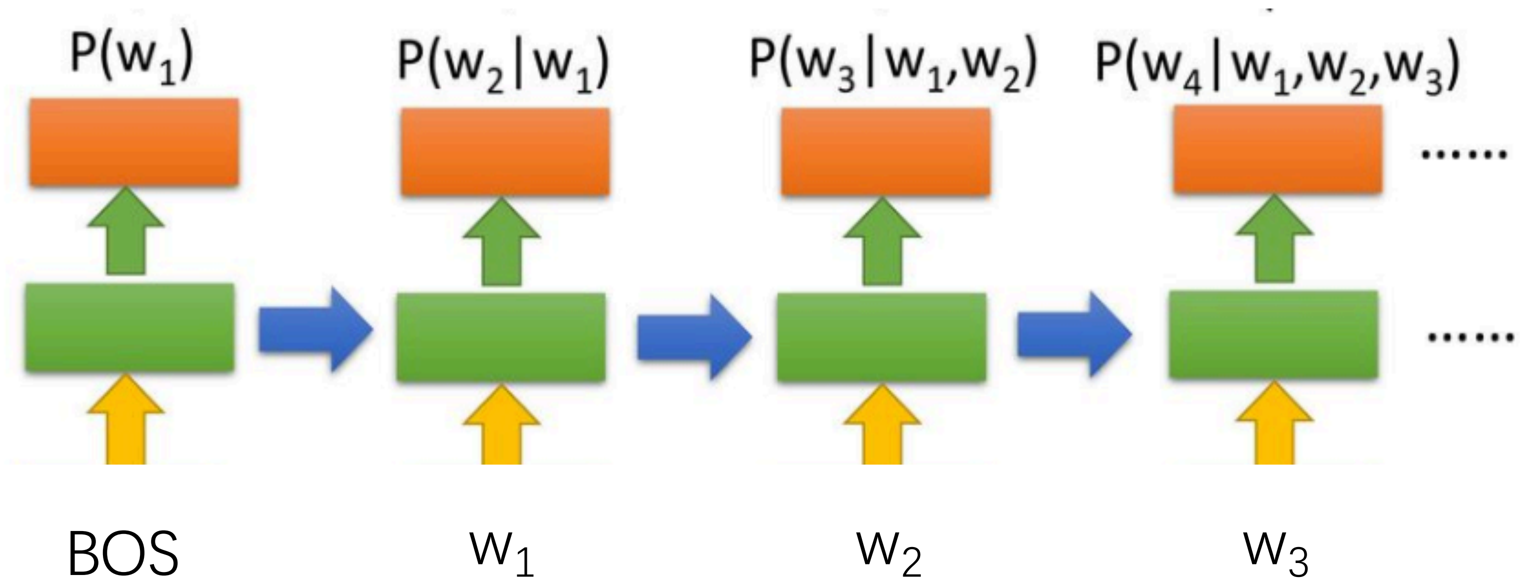
20190913

Background

- Given a grammatical or ungrammatical sentence, grammaticality Evaluation is to access the grammaticality (or fluency) by scoring
- Most common way is to use language model

Background – Neural Language Model

- The sentence to be evaluated $w_1 \ w_2 \ w_3 \ w_4 \ldots$



Sentence Perplexity = $e^{-1/N(\Sigma \log(Pi))}$

Background - Perplexity

- The lower the Perplexity is, the more likely this sentence occur
- Word frequency has great impact on Perplexity
- Using larger training data usually leads to lower Perplexity

Sample sentence	Perplexity
sees he i often mary ?	7555.2
it seems that it is likely that john will win .	48.9

Perplexity Good for Grammaticality Judgment?

- No

Sample sentence	Perplexity
I travel to London	800.0
I travel to Tuvalu	233.9

GPT2 language model used

- Perplexity favors frequent words, although sentences are equally grammatical

Goals

- An idealized grammaticality evaluator should

    1. Avoid or alleviate the impact of low frequency word (this is especially important for the text contains a number of low-frequency entity)

    2. Grammatical sentence > ungrammatical sentence

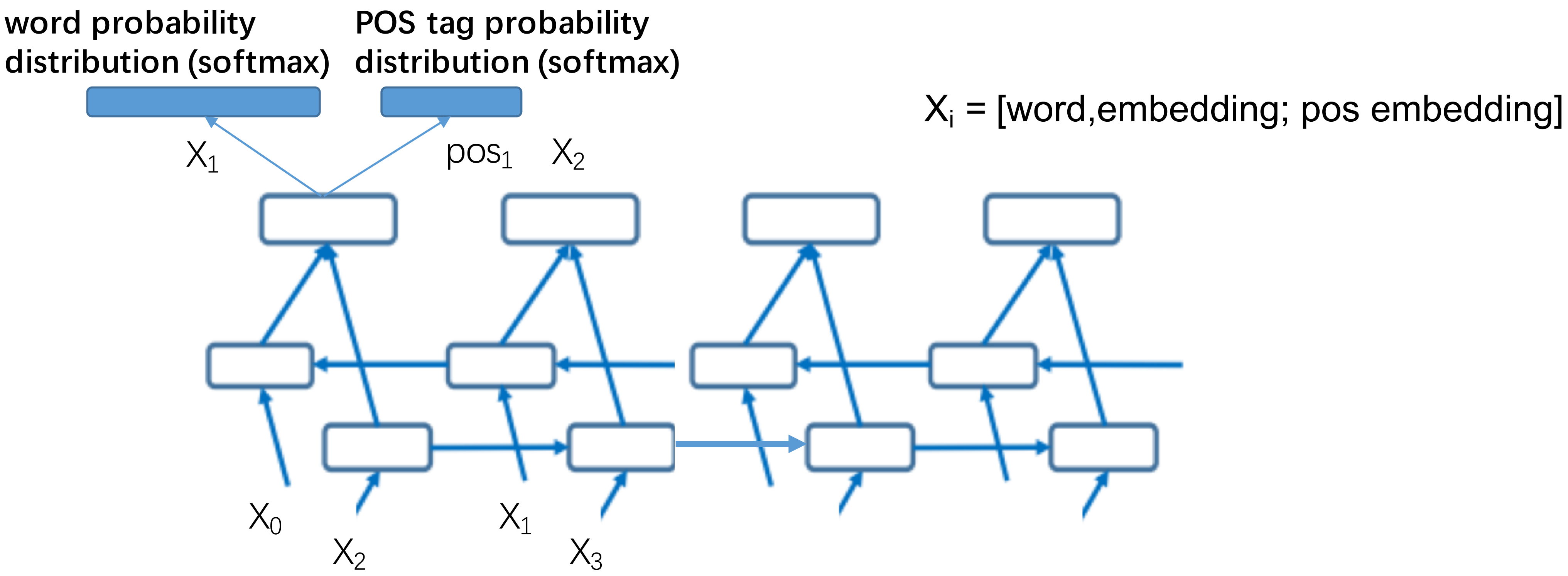
    3. ungrammatical sentence with 1 grammar errors > ungrammatical sentence with 4 grammar errors

Eliminating impact of low frequency word by considering POS tags

Sample sentence	Perplexity
I travel to London [('I', PRON'), ('travel', VERB'), ('to', ADP'), ('London', PROPN')]	800.0
I travel to Tuvalu [('I', PRON'), ('travel', VERB'), ('to', ADP'), (Tuvalu', PROPN')]	233.9

POS based Neural Language Model

“two ppl loss terms”



Training

- Training LM with Gigawords (Agence France-Presse, English Service (afp_eng) scoure, etc.)

- Vocabulary size: ~30,000 subwords

- One-layer unidirectional LSTM language model

- Test Dataset - Corpus of Linguistic Acceptability (CoLA; Warstadt et al., 2018): 1,000 sentences with human judgments (1: grammatical or 0: ungrammatical)

Results

- take (ppl<100) as grammatical prediction and (ppl>100) as ungrammatical prediction

	f(x)
LM	0.62
POS-guided LM	0.66

$$f(\boldsymbol{x}) = \frac{1}{1 + H(\boldsymbol{x})}$$ D ∈ **(0, 1]**

$$H(\boldsymbol{x}) = -\frac{\sum_{i=1}^{|\boldsymbol{x}|} \log P(x_i | \boldsymbol{x}_{<i})}{|\boldsymbol{x}|}$$ D ∈ **[0, +∞)**

Expectation

- The number of POS tags is less than 50, making pos tag embedding well learned during training
- Model might learn correct collocation of POS tags
- Limitation of this approach:
  - Need pos tagging beforehand
  - Ungrammatical sentence might have wrong pos tagging

Goals

- An idealized grammaticality evaluator should

1. Avoid or alleviate the impact of low frequency word (this is especially important for the text contains a number of low-frequency entity)

2. Grammatical sentence > ungrammatical sentence

   ? Large room for improvement

3. ungrammatical sentence with 1 grammar errors > ungrammatical sentence with 4 grammar errors

   Need category errors

Next

- Testify whether LM is able to detect the grammar errors like
  - Subject-Verb