Are Some Words Worth more than Others?*

Shiran Dudy, Steven Bedrick
Oregon Health & Science University

*Accepted to Eval4NLP workshop EMNLP 2020

Research shows: Text generation is dull

Serban et al, 2015: However, the majority of the predictions are generic, such as "I don't know" or "I'm sorry."

Holtzman et al 2019: producing dull and repetitive text that is not aligned with human generated text

- Common evaluation metrics for language models (LM) are accuracy and perplexity
- They tend to overlook linguistic properties of words
- Neural LMs are biased towards frequently occurring words-types, creating dull and repetitive text
- We demonstrate that a model's performance depends greatly upon word frequency

The Goal: Promote Diversity through evaluation metrics

Term accuracy

Term Diversity

Term Perplexity

model	top_1 (top_{10})	T_1 (T_{10})	ppx
GPT-2	35.63 (67.76)	26.60 (47.27)	34.8
GPT	29.37 (60.89)	15.96 (30.80)	37.9
RoBerta	28.18 (59.55)	24.73 (42.63)	42.2
Bert	22.11 (50.98)	15.59 (29.61)	50.7

Table 1: Experimental results on wiki-103 corpus

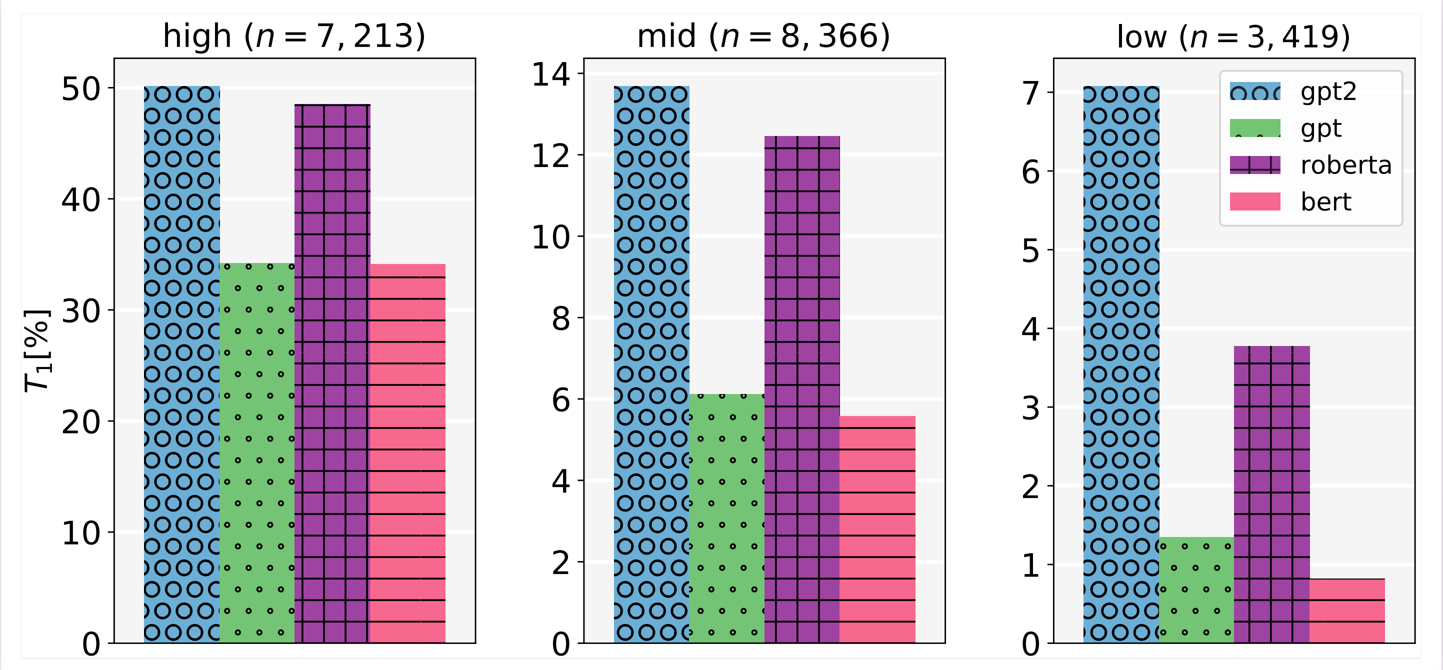

Fig 1: T_1 distribution by frequency

Take away 1: T_1 Diversity is not correlated with accuracy (top_1) or ppx

Take away 2: State-of-the-art models perform poorly on infrequent words

What about the effect of infrequent words on a downstream task?

A toy experiment on a paraphrasing task

Why paraphrasing? Easy way to measure the effects of single word substitutions

1. Which `dog` has longer hair
2. Which `cat` has longer hair
3. Which `poodle` has longer hair

We hypothesize that (`1`,`3`) are more similar than (`1`,`2`)

$$Bertscore(s(\texttt{dog}), s(\texttt{poodle})) > Bertscore(s(\texttt{dog}), s(\texttt{cat}))$$

model	*hits*	*misses*	*total*
Bert$_{rare}$	14	36	50
RoBerta$_{rare}$	11	39	50
Bert$_{common}$	40	10	50
RoBerta$_{common}$	39	11	50

Table 2: Paraphrasing sentences with wiki-103 words

Take away 3: a possible link between bad performance on rare words in downstream tasks to low prediction of infrequent types

(1) Hierarchical neural network generative models for movie dialogue, Serban et al 2015
(2) Diversity-Promoting Objective Function for Neural Conversation Models, Li et al 2016
(3) Neural Text Generation with Unlikelihood Training, Welleck et al, 2019
(4) The curious case of neural text degeneration, Holtzman et al, 2019