Evaluating Horror Text Generated by Probabilistic Language Models

Emily Sheetz

Supervised by Professor Sostarecz, Professor Richards, Professor Eary, Professor Bond

Monmouth COLLEGE®

Department of Mathematics and Computer Science

Abstract

Probabilistic language models predict the next word in a sequence based on probabilities learned from training data and generate text by sampling out of the learned probability distribution. Evaluating the model's representation of the language is a challenging task because evaluation depends on the application of the model and the desired criteria to be evaluated. A variety of language models trained to generate sentences of horror text on the word level are evaluated using perplexity, a measure commonly used in Natural Language Processing. As models more closely imitate the way human authors write, the model more accurately represents the language and generates better text samples.

N-Gram Models

N-gram models are a particular type of probabilistic language model. N-grams predict the n^{th} word in a sequence given the previous n-1 words by estimating the conditional probability in Equation 1. Different probability estimators can be used to derive this probability. This project uses the Maximum Likelihood Estimation (MLE), which depends on the relative frequencies of sequences of words, as seen in Equation 2. The n-gram models used in this project are bigram, trigram, 4-gram, and 5-gram models. For example, we can look at the trigrams in the phrase "mathematics capstone project".

[(None, None, "mathematics"), (None, "mathematics", "capstone"), ("mathematics", "capstone", "project"), ("capstone", "project", None), ("project", None, None)]

Perplexity

Cross entropy measures the average uncertainty based on the probability the language model assigns to a text sample from the testing data. Equation 3 shows the computation for cross entropy. Perplexity—commonly used in the field of Natural Language Processing (NLP)—uses the cross entropy to determine the accuracy of the model's learned probabilities, as seen in Equation 4. Models with lower perplexity better represent the language, which implies the model generates better text.

Equations

Equation 1: Probability estimated by n-gram models

$$P(w_n|w_1 \dots w_{n-1})$$

Equation 2: Maximum Likelihood Estimation (MLE)

$$P_{MLE}(w_n|w_1 \dots w_{n-1}) = \frac{C(w_1 \dots w_n)}{C(w_1 \dots w_{n-1})}$$

Equation 3: Cross Entropy for sentence S with N(S) n-grams

$$H = -\frac{1}{N(S)} \sum_{w_{1\dots n} \in S} \log_2 p(w_n|w_1 \dots w_{n-1})$$

Equation 4: Perplexity

$$PPL = 2^H$$

Computer or Human?

Try to determine whether each text sample below was generated by a computer or human.

"it may be so"

"tuesday the th instant in the expectation of an atmosphere of sorrow"

"startled her mood that youve either loosen his countrymen"

"left alone sank on her ringing"

"burn out the vampire"

"graceful acceptance of good things came to her naturally as it does to one who is born to be a great lady"

"dont speak such things must begin at once she broke out again"

"surprisingly beautiful"

"such individuals"

"the sea very smooth all day with little or no wind"

"dvelf"

"kepler admitted that it be disproved be kept in ignorance there are ninety nine persons out of your wives my brothers i am restless and uneasy animation to the buried"

"amidst the vast primeval forces there were new sources of doubt"

Varney the Vampire by Thomas Preskett Prest; 4-gram model; bigram model; trigram model; 5-gram model; *The Man* by Bram Stoker; trigram model; 4-gram model; bigram model; *Narrative of Gordon A. Pym* by Edgar Allan Poe; bigram model; trigram model; 4-gram model

Evaluation of N-Gram Models

Bigram, trigram, and 4-gram models trained on horror text from Project Gutenberg [3] were evaluated using perplexity. To quantify the difference in quality between human authors and probabilistic models, 5-gram models trained on text from specific horror authors were also evaluated. The 5-gram models most closely represent how human authors write as word choices are made conditioned on more information. As shown in the table below, the average cross entropy and perplexity were computed for each model over 100 randomly selected text samples from *The Turn of the Screw* by Henry James. As the model makes better predictions, the perplexity goes down. This means the model better represents the language because it makes decisions based on more local context. However, the probabilistic models do not achieve as low a perplexity as human authors.

Model	Cross Entropy	Perplexity
Bigram Model	5.11	46.28
Trigram Model	2.18	6.18
4-gram Model	1.10	2.49
Franz Kafka	0.57	1.58
Edgar Allen Poe	0.70	1.77
Thomas Preskett Prest	0.72	1.81
Bram Stoker	0.81	1.92

Conclusions

Using perplexity to evaluate language models, the difference in quality between n-gram models and human authors can be quantified. Though the text samples from probabilistic language models are of lower quality than text from human authors, these models are capable of generating interesting text and learning to represent the language.

References

[1] Manning, Christopher D., and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Vol. 999. Cambridge: MIT Press. 1999.

[2] NLTK Project. "Natural Language Toolkit." *NLTK Project*, http://www.nltk.org/. 2017. Accessed Oct. 2017.

[3] Project Gutenberg. "Free ebooks – Project Gutenberg." *Project Gutenberg*, www.gutenberg.org/. 2017. Accessed Oct. 2017.