Emily Sheetz
MATH-401, Capstone
Project Proposal

Evaluating Text Generated by Probabilistic Language Models

I. Abstract

Research in natural language processing has improved the ability of language models to generate text. Probabilistic language models predict the next word in a sequence based on probabilities learned from training data. As text generation methods progress, these models are able to generate higher quality text and perform a task that is considered to be inherently human. While machine learning approaches to generating text produce high quality results, simpler and less computationally expensive probabilistic language models are capable of producing quality text as well. One challenge in text generation is evaluating the model's representation of the language and the quality of the generated text. This is a complex task because evaluation depends on the application of the model and the desired criteria to be evaluated. The purpose of this project is to investigate how text generated from probabilistic language models—specifically n-grams and Markov models—can be quantitatively evaluated with respect to the content of the generated text samples. Evaluation methods such as the perplexity of the model, context free grammar (CFG), and probabilistic context free grammar (P-CFG) will be investigated using horror fiction texts. Variations of n-gram and Markov models will be built and trained to generate sentences of horror text on the word level. These models will be evaluated using different evaluation methods. The goal of the project is to understand what evaluation methods can tell researchers about the quality

of different characteristics of a text sample and how text generated by probabilistic language models compares quantitatively to that from human authors.

This project proposal serves to outline socio-technical issues surrounding the project, describe the project specifications, discuss foundational material, and document the project plan that will be followed during the spring semester. The social issues related to probabilistic language models generating text force us to consider the significance of storytelling in our lives, specifically when eliciting emotion. The project description specifies the details of the project, focusing on the end product. Foundational material includes a description of probabilistic language models, estimation techniques that allow these models to generate higher quality text, and evaluation criteria such as perplexity and context free grammars. The project plan includes checkpoints to be met throughout the semester, alternative plans in the event that the project needs to be scaled back, opportunities for project extensions, and tools to be used during implementation. The project proposal is the culmination of all preparations that took place during this semester and is a starting point for implementation next semester.

II. Introduction

As researchers explore methods for generating text, even from a machine learning approach, it is important to investigate the task of evaluating the quality of the generated text. These methods for evaluation will provide a way to gauge the progress of the models and guide these approaches in a way that is most beneficial to particular applications.

The purpose of this project is to explore methods for evaluating text generated by probabilistic language models. Perplexity, context free grammars, and probabilistic context free grammars will be the evaluation methods investigated. Baseline measurements for each of these

methods will be established on sentences of horror text from human authors. These baselines will be compared to text generated by probabilistic models—specifically n-grams and Markov models—which will be trained to generate sentences of horror text one word at a time. The goal is to understand what different evaluation methods measure, what these metrics can collectively tell researchers about the quality of a text sample, and how text from probabilistic language models compares to text from human authors.

Evaluation methods for text generation relate to several areas of mathematics. Text generation falls under the field of natural language processing, also called computational linguistics. The probabilistic language models that will generate text and the context free grammars that will evaluate the generated text are specific to the field of statistical natural language processing. Knowledge of probability is necessary to understand how probabilistic language models are built and how they use learned probabilities from training data to generate text. Probability theory can also be used to estimate the likelihood that a particular sample would be generated, which is one way to quantify the quality of a text sample. This project also relates to information theory. Perplexity is a measure used to evaluate how well a language model represents the language by quantifying the uncertainty in the model. The measure of perplexity is very similar to that of entropy, but is scaled differently and used specifically in the field of computational linguistics. All of these areas of mathematics will guide this project.

There is not just one defined method that can be used to evaluate any text from any language model. Evaluation depends on the application of the model and the desired criteria of the text to be measured (Kawthekar et al. 1; Theis et al. 1). Previous research projects have used many different methods for evaluating text. One method of evaluation used in the field of computational linguistics is the perplexity of the language model. Perplexity quantifies the

uncertainty in a model (Manning and Schütze 73-78). Manning and Schütze offer a way to interpret perplexity: "a perplexity of *k* means that you are as surprised on average as you would have been if you had had to guess between *k* equiprobable choices at each step" (78). Perplexity measures the quality of a model. One could infer that a model with lower perplexity produces higher quality text samples because the language model better represents the language. One research team used a support vector machine (SVM) to identify the authors of emails by looking at structural and linguistic characteristics—such as organization and word choice—of the emails. They found that preselecting features of the text eliminated the evaluation of features that do not contribute to categorizing the emails according to author (De Vel). Neural networks have been used to evaluate the quality of short stories by looking at separate regions of the text interdependently and taking a holistic view of the text ("A Good Read"). A research team led by Sai Rajeswar evaluated the quality of their machine learning approach to text generation using a context free grammar (CFG) and a probabilistic context free grammar (P-CFG). CFGs determine how well a generated text sample follows certain grammar rules. If a text sample follows common grammar rules, the sample is considered to have a good quality. P-CFGs differ from CFGs in that they can identify more complex grammatical structure and evaluate for greater abstraction of meaning. CFGs and P-CFGs allowed the researchers to evaluate the quality of the generated text samples with respect to grammar (Rajeswar et al.). The variety of research projects and evaluation methods demonstrate that evaluation depends on the application and the criteria of the text to be evaluated.

This project will take a closer look at the evaluation methods used in several research projects. Because of its use within the field of computational linguistics, the measure of perplexity will be used to evaluate text generation models. As grammar is an intuitive way for humans to

evaluate text, a context free grammar and a probabilistic context free grammar will be built and trained for evaluation of generated text. This project will serve as an opportunity to understand how several different evaluation methods work within the context of a single application.

While this mathematics capstone project is being implemented, a related project in computer science will also be carried out. The computer science project will explore how text can be generated from a machine learning approach. A generative adversarial network (GAN) using long short-term memory networks (LSTMs) will be trained on horror texts to generate sentences of text at the character level. The quality of the text will be evaluated using a survey asking participants to rank the quality of text samples from the GAN and from human authors. The results of this survey will provide a qualitative evaluation of the generated text and serve as a way to compare the quality to that of existing horror fiction. The goal of the project is to investigate how well machine learning can participate in the human task of storytelling.

The task of evaluating text generated by language models is not only of technical interest within the field of natural language processing, but it is also significant in a broader social context. The focus of this project is to apply methods for evaluating text on probabilistic language models. However, there are other approaches to the task of text generation—such as machine learning approaches—that will need to be evaluated. Evaluation methods will help gauge the success of a particular approach to text generation while helping guide the development of these approaches to a particular application. In order to understand the value in investigating methods for evaluating text quality, it is important to consider not just probabilistic language models, but all approaches to text generation, including the more rigorous and computationally expensive machine learning approaches.

Human beings use stories to communicate with one another. As artificial intelligence becomes more powerful, it is more common for people to interact with artificially intelligent agents. However, these agents are typically considered to be an "alien sort of intelligence" (Riedl). If artificial intelligence can learn how to tell stories, it will become a more familiar form of intelligence and therefore become more useful to us. Furthermore, works of fiction have "sociocultural knowledge encoded" into them from "different cultures and societies" (Riedl). Teaching artificial intelligence through fiction not only allows it to learn narrative intelligence, but also about the values and norms of a culture. Artificial intelligence learning to tell stories will provide the powerful capacity for more human-like intelligence. Methods for quantitatively evaluating the quality of these models will allow them to be developed in a way that is more useful to us.

Horror fiction in particular is personal and human. The goal of horror fiction is to "elicit an emotional reaction that includes some aspect of fear or dread." As author Douglas Winter said, "Horror is an emotion" (Horror Writers Association). Horror texts are compelling because they create powerful emotions in their readers. What one reader finds scary is dependent on who they are. A horror story "forces us to confront who we are, to examine what we are afraid of, and to wonder what lies ahead down the road of life." Horror "reminds us of how little we actually know and understand" (Horror Writers Association). Horror fiction is closely related to our humanity and our emotions. This makes horror a particularly interesting genre through which to explore the storytelling capabilities of language models. This project will provide an opportunity to investigate techniques for evaluating the quality of text generated by language models. As language models progress and produce higher quality texts, these evaluation methods will be

important in guiding improvements and determining how well these models demonstrate human-like intelligence through storytelling.

III. Project Description

The purpose of this project is to apply several methods for evaluating generated texts that have been used in previous research projects. Evaluation through perplexity, context free grammars, and probabilistic context free grammars will be investigated using horror fiction from human authors. Probabilistic language models—specifically n-gram models and Markov models—will be built and trained to generate sentences of horror text one word at a time. If time allows, different versions of these models will be built in order to compare performance of the models using different probability estimators. The goal of the project is to understand what can be measured using different evaluation methods, what these metrics show about a text sample, and how text generated by language models compares to text from human authors.

For the language models to learn meaningful probabilities that can generalize well, sufficient training data must be obtained. Horror texts for this project will be gathered from Project Gutenberg. Texts that will be used for training include:

- *Twenty-Five Ghost Stories*, an anthology edited by W. Bob Holland
- *Metamorphosis* by Franz Kafka
- *The Trial* by Franz Kafka
- *The Shunned House* by H. P. Lovecraft
- *The Works of Edgar Allen Poe*, Volumes 1 – 5 by Edgar Allen Poe
- *Varney the Vampire: Or the Feast of Blood* by Thomas Preskett Prest
- *The Strange Case of Dr. Jekyll and Mr. Hyde* by Robert Louis Stevenson

- *Dracula* by Bram Stoker
- *Dracula's Guest* by Bram Stoker
- *The Jewel of Seven Stars* by Bram Stoker
- *The Lady of the Shroud* by Bram Stoker
- *Lair of the White Worm* by Bram Stoker
- *The Man* by Bram Stoker

These particular texts were selected for author name recognition as well as the variety of horror elements. This will create language models that are trained on multiple types of horror characters and trained on texts that are known to elicit strong emotions from readers.

The goal is that by the end of the project, samples of text from human authors and all probabilistic models built during the semester can be compared quantitatively using several different evaluation metrics. This collection of metrics will show what features of a text can be quantified, what the combination of measurements can explain about the quality of a text sample, how samples generated by different language models compare, and how text from probabilistic language models compares to text from human authors. The project will serve as an exploration of the information that can be gathered by evaluating text using different methods, which will help researchers gauge how well text generation models perform in the human task of storytelling.

IV. Foundations

The foundations of this project lie in the field of natural language processing. Natural language processing (NLP) is the field of study that allows us to "characterize and explain" how humans and computers "acquire, produce, and understand language" (Manning and Schütze 3). For

computers to generate text, they must have a way of learning the rules, structure, and meaning that govern language.

NLP is a challenging task because it must account for shades of meaning and ambiguity inherent in language. There are several different levels of NLP:

- *Phonology*: deals with spoken language and how sounds can be interpreted within and across words.
- *Morphology*: examines "the componential nature of words" by breaking words down into their smallest units of meaning, called morphemes.
- *Lexical*: interprets the meaning of individual words.
- *Syntactic*: focuses on the words and grammatical structure in a sentence.
- *Semantic*: looks at interactions between words to determine possible meanings of a sentence.
- *Discourse*: interprets meaning across sections of text that are longer than a sentence.
- *Pragmatic*: utilizes the context and reads into a text to understand the meaning that is not explicitly encoded into the text.

Different levels of meaning and understanding of a text are available on each level of NLP (Liddy). Each level of NLP provides its own implementation challenges and its own opportunities for understanding meaning within language.

One of the probabilistic language models that will be investigated is the n-gram model. These models predict the nth word in a sequence given the previous $n - 1$ words. The n-gram model makes the Markov assumption that the local context affects the next word in the sequence (Manning and Schütze 138-139). To make a prediction on the next word in a text, it is not necessary to consider all of the prior words, just the previous few words. In practice, n-gram

models are typically bigrams or trigrams and are trained on relatively small vocabularies in order to minimize the number of parameters the model has to estimate (193). The n-gram model makes predictions on the next word in the sequence by sampling out of a predicted probability distribution over all possible words in the vocabulary.

The probability distributions output by the n-gram model are generated by probability estimators. The probability estimator the n-gram model uses affects how well the model represents the language and generalizes to data outside of the training dataset. Probability estimators that could be implemented in the models used in this project are:

- *Maximum Likelihood Estimation (MLE)*: estimates probabilities based on the relative frequencies of sequences of words in the training data. MLE does not generalize to sequences not seen in the training dataset.
- *Expected Likelihood Estimation (ELE)*: when learning from sequences of words in the training dataset, some probability is left over for unseen events. However, ELE has the problem of determining how much probability should be used to account for unseen events.
- *Held Out Estimation*: better estimates how much probability space is left for unseen events by comparing the n-grams in the training dataset to the same n-grams in a held out dataset. This technique requires training, validation (held out), and testing datasets, meaning the model will be trained using fewer data samples.
- *Deleted Estimation*: provides a method so that data samples can be used as a part of both the training and held out datasets. This allows the model to be trained on more data while using both a training and a validation dataset to improve the probability estimates of the model.

- *Good-Turing Estimation*: assumes that probability estimates follow a binomial distribution. Though n-grams do not have a binomial distribution, this assumption works well for large datasets and large vocabularies, and therefore performs well as a probability estimator for n-gram models.

The different probability estimators give the n-gram model different text generation capabilities. Better probability estimators allow the model to generalize to sequences of words that were not seen in the training dataset (Manning and Schütze 196-217). Throughout the project, n-gram models will be built and trained using as many of these probability estimators as can be implemented. A variety of n-gram models will allow for a quantitative comparison of the models.

The other probabilistic language model to be investigated during the project is the Markov model. Whereas n-grams work on the word level of text generation, Markov models allow for higher levels of abstraction, consider the grammar of a sentence, and work on the syntax level of meaning (Manning and Schütze 317). Markov models have the properties of limited horizon—the Markov assumption, as describe previously—and time invariance—the model will make the same prediction given the same sequence of words (318). The two types of Markov model are Visible Markov Models (VMMs) and Hidden Markov Models (HMMs). Both models work by predicting the next word in a sequence given the previous few words. In VMMs, the state of the model—the words that are being used to predict the next word in the sequence—is known. This means that n-gram models are VMMs (319). For HMMs, the user does not necessarily know what state the model is currently in, but they do know the probabilistic function of the state transitions of the model. This allows for higher abstraction in terms of the meaning of the text (320). In this project, a Hidden Markov Model will be built and trained for text generation.

The focus of this project is to evaluate the quality of text generated by these probabilistic language models. The evaluation methods that will be applied are perplexity, context free grammars, and probabilistic context free grammars. The measure of perplexity is commonly used in the field of natural language processing. Perplexity quantifies the uncertainty in the language model or how well the model represents the language. Though perplexity evaluates the language model and not the text samples directly, it can be inferred that a model that better represents the language—a model with a lower perplexity—will produce higher quality text samples. The other evaluation methods that will be applied are context free grammars (CFGs) and probabilistic context free grammars (P-CFGs). These techniques evaluate the content of the text sample by determining how well the sample adheres to certain grammar rules. P-CFGs evaluate more complex grammar rules and therefore allow for greater abstraction in meaning. Text samples that follow grammar rules are likely to be quality sentences. Perplexity, context free grammars, and probabilistic context free grammars will provide a starting point for evaluating generated text.

V. Implementation Plan and Timeline

The timeline for project implementation during the spring semester is based on biweekly checkpoints. The checkpoints for the semester are described below.

- *Checkpoint 1*: By the end of this checkpoint, the script that will handle text preprocessing will be created and the evaluation method of perplexity will be explored. The goal of text preprocessing is to edit the raw text files from Project Gutenberg so that the language models can effectively learn from the data. Punctuation and capitalization will have to be removed from the text because it may affect how words are distinguished from one another. For example, in Edgar Allen Poe's *The Masque of the Red Death*, if punctuation and

capitalization were not ignored, “death,” “Death,” and “death.” would all be considered different words. Text preprocessing will also include breaking texts into sentences, which will serve as individual samples in the training dataset. While the script for text preprocessing will handle much of the preprocessing automatically, there will likely be some formatting that will need to be done manually before the dataset is ready to train the probabilistic language models. By the end of this checkpoint, it will be understood how much preprocessing will need to be done manually. The second goal of this checkpoint is to learn about how to measure the perplexity of language models and obtain a baseline measurement based on existing horror fiction by human authors.

- *Checkpoint 2*: This checkpoint will focus on working with CFGs as a means of evaluating text. The CFG will be built and trained using a dataset comprised of existing horror fiction from human authors. A separate testing dataset will be used to obtain a baseline measurement for the quality of text from human authors.
- *Checkpoint 3*: By this checkpoint, the P-CFG will be built, trained, and used to obtain a baseline evaluation on horror text from human authors. By the end of this checkpoint, the three methods for evaluation will have been investigated and implemented. A baseline for horror text generated by human authors will be established to understand how the three measurements differ and what these measurements show about the quality of a text sample. This baseline will later be used for comparing the evaluation measurements on text generated by probabilistic language models.
- *Checkpoint 4*: This checkpoint will focus on building and training n-gram models. The goal is to have several n-gram models with different probability estimators. In order of priority, the probability estimators that will be implemented in each n-gram version are:

maximum likelihood estimation, expected likelihood estimation, held out estimation, deleted estimation, and Good-Turing estimation. The goal is to have at least two variations of n-grams, using the maximum likelihood estimation and expected likelihood estimation as probability estimators. However, as many of these will be implemented as can be during this checkpoint to give more points of comparison for evaluation methods. As the models are implemented, some text samples will be generated and evaluated using the investigated methods. The final poster presentation, which will be given on Scholar's Day in the spring, will also be started.

- *Checkpoint 5*: Building and training a Markov model—specifically a Hidden Markov Model—will be the focus of this checkpoint. Once the HMM is built, text samples will be generated and evaluation of these samples will begin. Because only one language model will be built during this checkpoint, a significant amount of time will be dedicated to putting the final poster presentation together.
- *Checkpoint 6*: During the final checkpoint, text samples from the language models will be generated and evaluated using the investigated evaluation methods. By the end of the checkpoint, there will be a comparison between different versions of n-gram models, all of the probabilistic models built, and probabilistic models and human authors. The poster presentation will be finished, including the results and the conclusions of the project.

Checkpoints are structured to maximize success in the event of an unforeseen obstacle. For example, the early checkpoints focus on the evaluation methods and using these methods to evaluate horror text from human authors. This means that by Checkpoint 3, the project will have sufficient information to show how text can be quantitatively evaluated and what the metrics collectively show about the quality of a text sample. Therefore, any model built and trained during

Checkpoint 4 or Checkpoint 5 can be used to determine how text from probabilistic language models can be quantitatively evaluated and how text samples generated by language models compares to text from human authors.

If time allows, there are a few possible extensions of this project. Another evaluation method could be investigated, implemented, and tested if there is extra time in the first three checkpoints. The quantitative evaluation methods investigated in this project could also be applied to the GAN developed in the related computer science project, as described in the introduction. Furthermore, text samples from the probabilistic language models could be included in the qualitative survey that will be used to evaluate the success of the computer science project. Though the focus of this project is quantitative evaluation of text, it would be interesting and informative to compare the qualitative and quantitative metrics. This could give a better understanding of what each metric measures and whether the metrics cannot accurately represent some characteristic of the text.

The primary tool required for this project is Python. The Natural Language Toolkit (NLTK) in Python includes functionality for text preprocessing and building n-gram models. Documentation and tutorials will be used as a reference and guide throughout implementation. Part of the researcher's time this semester was spent putting together a technical presentation to begin working with the foundational material for the project. This presentation explored how n-gram models can be built and coded in Python and served as a brief introduction to the NLTK. The experience of the technical presentation gave the opportunity to start working with the tools for building the probabilistic language models that will be used to investigate quantitative evaluation of generated text.

VI. Conclusion

The goal of this mathematics capstone project is to investigate methods for quantitatively evaluating generated text and testing these methods on text from both human authors and probabilistic language models. The language models—specifically n-gram models with different probability estimators and Markov models—will be built and trained to generate one sentence of horror text one word at a time. The project will provide an understanding of what characteristics of a text can be evaluated, what these metrics collectively tell us about the quality of a text, and how text generated by language models compares to that written by human authors. The investigated evaluation methods will provide a way of determining how well the language models—from either a natural language processing or a machine learning approach—participate in the human task of storytelling.

The work to implement this project during the spring semester will be challenging yet rewarding. While there are sure to be obstacles given that the researcher has never worked with natural language processing techniques, the opportunity to experience research in text generation will ultimately be the driving force of the project. Planning biweekly checkpoints, possible ways to scale back the project, and opportunities to extend the project has forced the researcher to carefully consider and prioritize realistic goals. No matter which goals are met, the final product of this capstone project will serve as an exploration of the methods to quantify the quality of generated text. The results of this project will carry implications about how well language models can participate in human tasks.

Bibliography

"A Good Read: AI Evaluates Quality of Short Stories." *PhysOrg*. 2017.

Allen, James F. "Natural Language Processing." *Encyclopedia of Computer Science*. 2003, 1218-1222.

Bird, Steven, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media. 2009.

Clark, Alexander, Chris Fox, and Shalom Lappin, eds. *The Handbook of Computational Linguistics and Natural Language Processing*. John Wiley & Sons. 2013.

De Vel, Olivier. "Mining E-mail Authorship." *Proc. Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining*. 2000.

Horror Writers Association. "What is Horror Fiction?" *Horror Writers Association*, http://horror.org/horror-is.htm. 2009. Accessed Nov. 2017.

Kawthekar, Prasad, Raunaq Rewari, and Suvrat Bhooshan. "Evaluating Generative Models for Text Generation." *Stanford University*. 2017.

Liddy, Elizabeth D. "Natural Language Processing." *Encyclopedia of Library and Information Science*. 2001.

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

McKeown, Kathleen. *Text Generation*. Cambridge University Press, 1992.

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

Press, Ofir, et al. "Language Generation with Recurrent Generative Adversarial Networks without Pre-training." *arXiv preprint arXiv:1706.01399*. 2017.

Project Gutenberg. “Free ebooks – Project Gutenberg.” *Project Gutenberg*, www.gutenberg.org/. 2017. Accessed Oct. 2017.

Rajeswar, Sai, et al. “Adversarial Generation of Natural Language.” *arXiv preprint arXiv:1705.10929*. 2017.

Riedl, Mark. “Why Artificial Intelligence Should Read and Write Stories.” *Huffington Post*, www.huffingtonpost.com/mark-riedl/why-artificial-intelligen_b_8287478.html. 2015. Accessed Oct. 2017.

Riffaterre, Michael. “Criteria for Style Analysis.” *Word* 15.1. 1959, 154-174.

Theis, Lucas, Aäron van den Oord, and Matthias Bethge. “A Note on the Evaluation of Generative Models.” *International Conference on Learning Representations*. 2016.