News Article Generator

Miguel A. Mascorro, Victor Ortega, Mario Velasquez
miguel.mascorro01@utrgv.edu, victor.ortega01@utrgv.edu, mario.velasquez01@utrgv.edu

October 2, 2019

Summary

For this project, we're looking to design a recurrent neural network which is shown to be the type of neural network that is best suited for this task. Ideally after training on data sets such as the CNN News Story data set, will produce news articles that are virtually indistinguishable from one made by an author. Of course, there already exists bots that conduct similar tasks. However for the purposes of this course, ours will be a stepping stone in exploring just how accurate it is at producing such articles and perhaps how quickly it can be trained to do so.

Background

Currently, there exist several programs that can create news articles based on stocks, sports scores, and similar data. These use certain preset sentences to describe changes such as the description of a price dropping or rising. These types of programs are very useful to newspaper companies, both digital and traditional, since about a third of their articles are composed of these types news, where the only things to interpret are given tables and scores. What we want to do is to generate articles more akin to ones a person would write about a certain hot topic, complete with buzzwords and eye-catching sentences.

As of right now, the software that exists for generating these sorts of stories are far and few in between. For example, a student's small time project that they have to continue presented in a news article is a neural network that creates presidential speeches based on the hundreds of already existing recorded ones. A lot of the generated speeches share very similar structure such as starting with a "Thank you" and being very polite in general, but looking more closely at these reveals that the bot has typed nothing but gibberish only resembling a speech in passing. This is due to the fact that these are generated character by character, rather than word by word or phrase by phrase. One of our goals then is to, first, make sure the program we create is spelling words correctly and to make sure it is writing coherent sentences or at the very least semi-coherent based on keywords or seeds the user feeds it.

Measuring the accuracy of the spelling is no problem as long as we have access to a dictionary. This is of course very limiting since checking for spelling alone requires the program to go character by character which could potentially take a very long time. In addition to that, we would like to be able to check its coherence. This turns out to be more of a challenge since coherence has a somewhat subjective definition. We will have to define what being coherent means in a way that will be relatively simple for the program to check since comparing a sentence in our generated article with hundreds of other sentences to see if they match somewhat could take an even longer time than checking for spelling.

Goal and Objectives

The goal of this research is to be able to generate articles that at a short glance seem indistinguishable from a news article. The first objective will be to collect the data for our algorithm to mimic. The second objective is to produce sentences in a grammatically correct manner through machine learning techniques.

Data and Methods

We need a data set that has multiple records to allow us to generate text from, specifically from news as that is part of our goal in this text generating research. Therefore, we will be using the CNN News Story Data set for our data. It holds 93,000 news articles. CNN News will hold 93,000 news or "stories" as they will be extracted, from a tar file, as a file extension of ".story". The data will need to be cleaned out as we are only interested in the main body and title of the data. There is extra information such as a "highlight" and "CNN information" that must be removed for our project. The data will allow us to train using only the ASCII information in our machine learning algorithm.

References

- https://www.nytimes.com/2019/02/05/business/media/artificial-intelligence-journalism-robots.html
- https://www.wired.com/2015/10/this-news-writing-bot-is-now-free-for-everyone/
- https://machinelearningmastery.com/inspirational-applications-deep-learning/