CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)

Houjun (Jack) Liu

Project Number

38372

Project Title

Comparing the Effects of Various Corpora's Qualities on NLG/NLP Systems

Abstract

Objectives/Goals

The experiment is designed to analyze the relationship between the input quality of a dataset and the output quality of a dataset after it is processed a Natural Language Generation (NLG) system - a system that would generate text based on an input. During the experiment, the focus is to identify whether or not the input/output qualities affect each other in a linear pattern (hence, having an equal ratio in growth.)

Methods/Materials

Four corpora of incrementing quality, two generation systems, Python and Wolfram Mathematica -- where scripts written for this experiment computation is executed, Gold Standard benchmark corpus "1200 Graded Sentence for Analysis" for evaluation.

The four corpora are all separately fed in small batches into the two generation systems, which generates five batches of 100 sentences based on the corpora. These sentences are evaluated according to the benchmarks of the Gold Standard corpus. An average quality score between 1-5 is found for each batch. A best fit line between the input quality of each corpus against the average quality score across all five batches for that corpus is found, and finally, the error rate of that fit line is identified to test the linearity of the dataset.

Results

According to the data, both model's best fit lines all have near-zero error values. Furthermore, the ratios between input and output qualities on each of these models stayed approximately the same around 1.

Conclusions/Discussion

Since both sets of numerical data collected above identified a linear pattern between the input quality and the output quality of the NLG system, it is shown that the data observed had a linear relationship with each other. This result would help model the performance of an NLG system for the future.

Summary Statement

NLG systems' input and output corpora's qualities are compared, and it is found that, when data of various quality is fed through and NLG system, the input and output data had a linear relationship in quality.

Help Received

My science teachers helped proofread the display board and check for semantic and grammatical errors.