Hear Ye Hear Ye: Machine Translation and Sentiment Analysis of Modern Literature into Early English Language

Maggie Brodbeck, MSI & Marilu Duque, MSI

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Introduction

The English language has evolved over the ages as civilizations advanced and societies expanded. As language has changed, many of the sentiments and subjectivity expressed have shifted meaning as well. Having a clearer understanding of sentiment analysis throughout history and helps us acknowledge public perceptions of certain topics throughout time and how they have evolved. This project trains models on existing popular literature to understand this evolution.

Objective

Our project aims to translate existing modern literature into early English using machine learning (ML) methods. Additionally, we aim to analyze the sentiment of the modern literature before and after translation into early English. This will help us better understand how sentiment perception has evolved as the English language has evolved.

Data

Harry Potter and the Sorcerer's Stone - GitHub - Text File
Original translation of *Romeo & Juliet* - GitHub - Text File
Modern translation of *Romeo & Juliet* - GitHub - Text File

Methodology

We employed an experimental research design methodology in order to both explore the texts and best ascertain the models that would best suit our datasets. Two models were created, and each dataset was preprocessed, trained on the models, tuned as needed, and tested.

Related Work

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. (2014). Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14). MIT Press, Cambridge, MA, USA, 3104–3112.

Xu W., Ritter A., Dolan W.B., Grishman R., Cherry C. (2012). Paraphrasing for style. 24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers, pp. 2899-2914.

Sentiment Analysis Model

The sentiment analysis models utilized a preliminary Bag of Word (BoW) model that was then applied to Logistic Regression (LR) and Naive Bayes (NB) models. The BoW model was chosen due to its specialized ability to process textual data for machine learning. The LR model was chosen due to its unique ability to both classify and predict outcomes of assigned data. The NB model was chosen due to its focus on independence between features and its usually higher accuracy results. Additionally, extra analysis was conducted by applying the TextBlob NLP tool to better evaluate sentiment.

Machine Translation Model

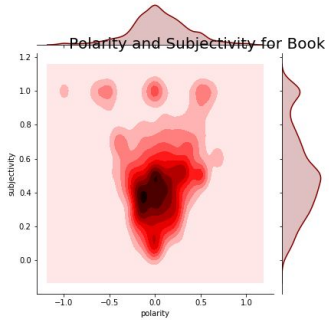
The machine translation model utilized deep learning algorithms on the Shakespearean datasets. It uses a Bidirectional Long Short-Term Memory (LTSM) algorithm with additional layers.. This recurrent neural network (RNN) was chosen as it is typically used for machine translation models due to its effectiveness in learning sequence-to-sequence order, especially with the Bidirectional LSTM since it processes data in both directions. The model was trained with the original translation as the input data, and the modern translation as the output data.

Results

Despite producing higher accuracy scores (~80%), low loss (~1.00) and undergoing several rounds of hypertuning, the experimental machine translation model could not produce a coherent translation and struggled to map the input sequences to the target sequences.

Input text	Target text	Predicted text
as mine on hers , so hers is set on mine , and all combined , save what thou must combine by holy marriage .	we re bound to each other in every possible way , except we need you to marry us	you me the <end> you , , , , <start> you you you sucked
then plainly know my heart s dear love is set on the fair daughter of rich capulet .	i love rich capulet s daughter .	will to i . to to i . the i , i . .
riddling confession finds but riddling shrift .	a jumbled confession can only receive a jumbled absolution .	<start> <start> <start> . don don don

Through the plot below, we can see that average subjectivity expressed in the text, is .475 which is slightly above neutral subjectivity. The polarity is much more neutral score at .050. The subjectivity was expected due to the emotional topic of this text, but we hypothesized the polarity to be higher for the same reason. In looking at the joint plot, polarity seems more stable and neutral while subjectivity seems fluid.



Discussion

The machine translation model produced high accuracy scores during training, but it performed poorly in producing direct translation results. It had difficulty in matching target sequences beyond the first few parts of the input index, which may be improved with more data, layers, and epochs. The sentiment analysis models produced high, but identical accuracies which begs the question of potential embedded errors. More time to work on these would help ascertain these errors.

Future Work

With more time, we hope to experiment with a wider range of models including Latent Semantic Analysis (LSA) to help maximize model accuracy. We would also aim to complete the training and translation of the Shakespearean text given more time and computational resources. Lastly, adding additional text of early English literature for the model to train from would help ensure better results.

Contact Us: Brodbeck@umich.edu & Marilud@umich.edu