Video link:
https://youtu.be/OlA0VxDsQ_A

Automated Essay Scoring

Junaid Syed, Sai Shanbhag, Vamsi Ravilla

Video link: https://youtu.be/OlA0VxDsQ_A

Georgia Tech

Agenda

1. Introduction
2. Dataset Description
3. Methods
4. Results Summary
5. Future Steps

Georgia Tech

Introduction

Motivation:

- Writing is a foundational skill that only a few students can hone, often because writing tasks are infrequently assigned in school.
- Automated Essay scoring makes it easier for teachers to assign more writing tasks and provide feedback.
- However, current tools lack in their scope because providing a simple overall score provides little to no feedback to the student and does not help the students in their progression

Objective:

- The goal of this project is to evaluate the essays on granular factors such as cohesion, grammar, syntax rather than just a single score
- We have used the ELLIPSE and PERSUADE corpus datasets available on Kaggle to train our automated essay scoring models

Evaluation: Mean Column Root Mean Square Error (MCRMSE)

$$MCRMSE = \frac{1}{N_t}\Sigma_{j=1}^{N_t}\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(p_{ij} - y_{ij})^2}$$

N_t: *number of ground truth score columns*
p_{ij}: *the predicted score*
y_{ij}: *the ground truth score*
n: *number of training samples*

Georgia Tech

Data Description

ELLIPSE corpus available on Kaggle; contains essays written by students in grades 8-12 annotated by human raters for language proficiency.

ELLIPSE Exploration:

- 3911 essay samples with scores for six analytical measures
  - Cohesion
  - Syntax
  - Vocabulary
  - Phraseology
  - Grammar
  - Conventions

- Scores range from 1.0 to 5.0 with an increment of 0.5

- Average length of essays was ~500 tokens with max length of 1453 tokens

percentile	essay_length
mean	496.985170
std	218.322784
min	16.000000
50%	464.000000
90%	775.000000
91.1%	795.000000
92.2%	817.975000
93.4%	841.962500
94.5%	884.000000
95.6%	936.875000
96.8%	1007.925000
97.9%	1104.737500
99%	1239.700000
max	1453.000000

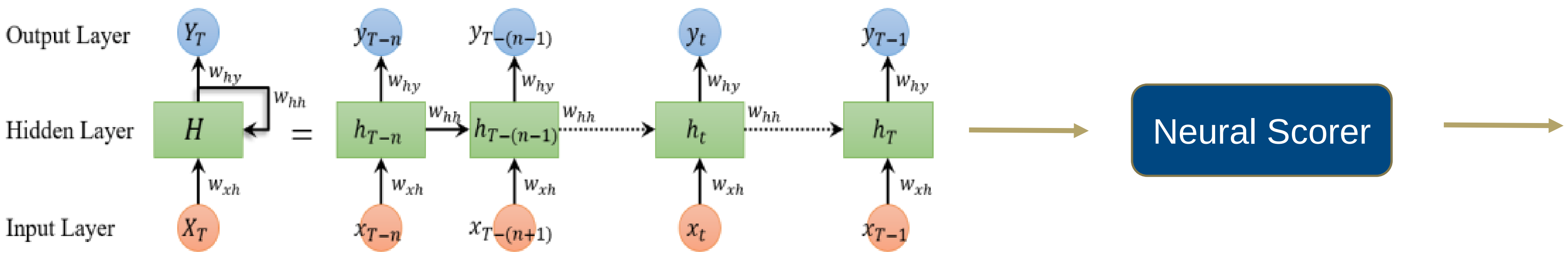
Georgia Tech

Text Encodings

- Inputs to Regression model
- Baseline: Bidirectional LSTM with Glove embeddings
- Pre-trained Language Models:
  - DistilBERT
  - Longformer
  - RoBERTa-base
  - T5-base

Method I: LSTM with GloVe

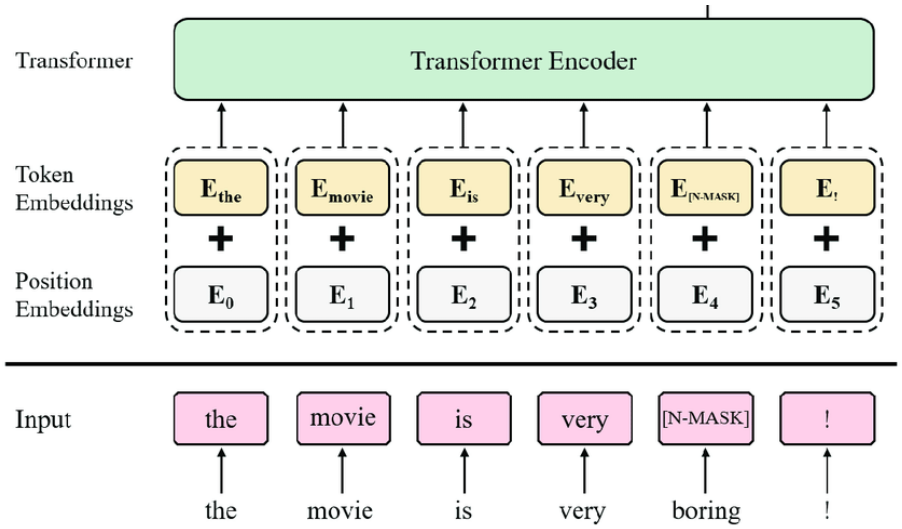
- Performed data cleaning to remove white spaces, punctuations and any special html characters
- Used NLTK's tokenizer to tokenize the processed essays
- Used Glove embeddings to obtain vector representation of the tokens
- Trained a bidirectional LSTM network with hidden size = 400 and obtained the final hidden state
- Finally, a two-layer neural network converts this into a 6-dimensional output vector representing the scores for each of the six writing attributes described earlier



- Cohesion score
- Syntax score
- Vocabulary score
- Phraseology score
- Grammar score
- Conventions score

Method II: DistilBERT

- BERT which uses self-attention provides context dependent embeddings as opposed to Glove
- This improves model’s ability to capture contextual information and provide a more accurate score
- Used Huggingface's AutoTokenizer class to tokenize the essays before passing them to the pre-trained distilBERT model
- A two-layer neural network described earlier was used to obtain the essay scores from distilBERT embeddings

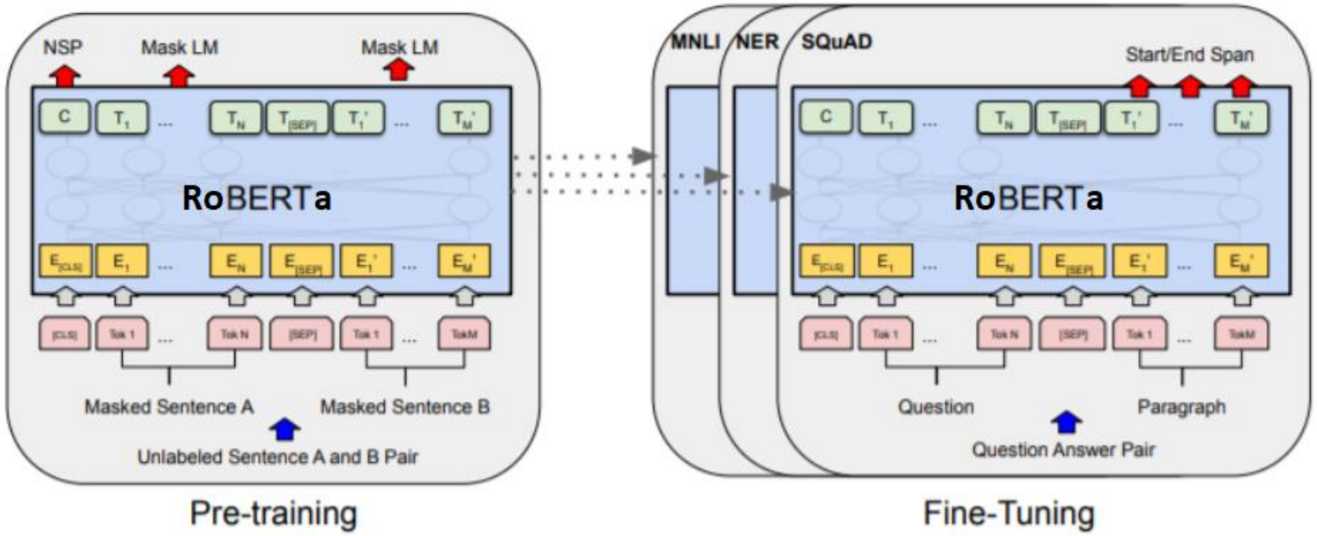


first hidden state → **Neural Scorer** →

- Cohesion score
- Syntax score
- Vocabulary score
- Phraseology score
- Grammar score
- Conventions score

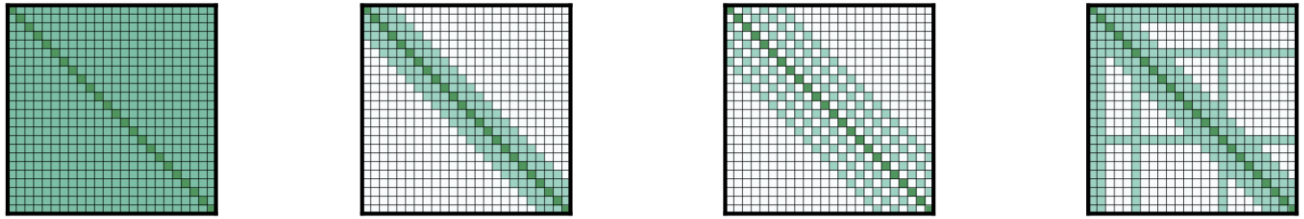
Method III: RoBERTA

- RoBERTa is a BERT like masked language model developed by Facebook - outperforms BERT on most GLUE and SQuAD tasks
  - Differs from BERT with regard to the masking process - uses dynamic masking.
  - Trained on a much larger corpus of data compared to BERT (10x) and a larger vocabulary set.
- Used Huggingface's AutoTokenizer class to tokenize the essays before passing them to the pre-trained RoBERTa-base model
- A two-layer neural network to obtain the essay scores from RoBERTa embeddings



Method IV: Longformer

- DistilBERT supports a max sequence length of only 512, but 40% of training essays have a length > 512
- Longformer model supports sequences upto length 4096
- Instead of self-attention, it uses a sliding-window and dilated sliding-window mechanism to capture the local as well as global context
- Like distilBERT, used Huggingface's AutoTokenizer class to tokenize the essays before passing them to the pre-trained Longformer-base model
- A two-layer neural network described earlier was used to obtain the essay scores from the longformer embeddings

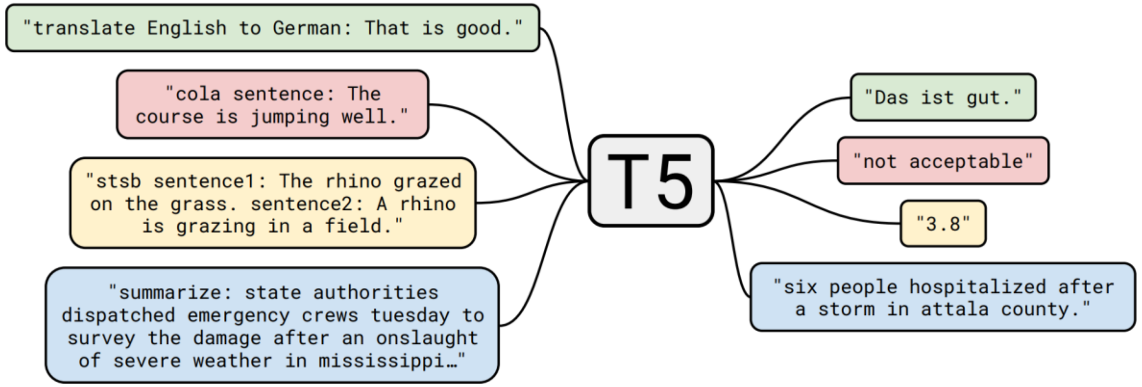


(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

Figure 2: Comparing the full self-attention pattern and the configuration of attention patterns in our Longformer.

Method V: T5-base

- T5 or Text-To-Text Transfer Transformer is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks
- This pre-training framework provides the model with general-purpose “knowledge” that might improve its performance on downstream tasks like sequence classification
- Used Huggingface's AutoTokenizer class to tokenize the essays before passing them to the pre-trained T5-base model
- A two-layer neural network described earlier was used to obtain the essay scores from the T5 encoder output



Results: Baseline + Pretrained Language Models

Model	MCRMSE
Baseline (LSTM + GloVe)	1.36
distilBERT	0.4934
T5-base	0.5320
RoBERTa	0.4746
Longformer	0.4899

- The bidirectional LSTM with glove embeddings has the poorest performance
- Masked language models (DistilBERT, RoBERTa and Longformer) are seen to perform better than the generative model T5
  - Cause masked models are more tuned towards discriminative tasks with numeric outputs
- RoBERTa architecture produced the best results with a MCRMSE score of 0.4746
  - Plausibly due to its much larger training corpus and superior masking

Georgia Tech

Improvements to Regression Modeling

- Output Quantization
  - constrain output between 1 and 5, with increments of 0.5

- Weighted RMSE (WRMSE)
  - Account for imbalance in score distribution.

- Multi Head Architecture
  - Use 6 single-task models instead of one multi-task model

- Autoencoder
  - Use bottleneck layer or denoised output from decoder. Also perform semi-supervised learning using other essays in ELLIPSE + PERSUADE corpus.

Results: Improvements to Regression Modeling

Experiment	MCRMSE
distilBERT + output quantization	0.5294
distilBERT + WRMSE	0.5628
distilBERT + Multi-Head Architecture	0.508
distilBERT + Autoencoder	0.575

- Unfortunately, none of these variations to training the regression model result in a significant improvement
- Further study with a larger dataset is essential to verify that this reduction in performance is not an artifact of the current dataset

Results: Individual analytic measure MCRMSE

Model (or) Experiment	Cohesion	Syntax	Vocabulary	Phraseology	Grammar	Conventions
Baseline	1.37	1.35	1.32	1.34	1.44	1.36
distilBERT	0.54	0.51	0.46	0.52	0.57	0.49
T5-Base	0.55	0.52	0.48	0.54	0.58	0.53
RoBERTa	0.51	0.47	0.42	0.47	0.51	0.46
Longformer	0.54	0.48	0.46	0.49	0.53	0.47
distilBERT + output quantization	0.55	0.53	0.48	0.53	0.57	0.51
distilBERT + WRMSE	0.56	0.56	0.55	0.56	0.61	0.53
distilBERT + Multi-Head Architecture	0.53	0.5	0.45	0.51	0.56	0.49
Autoencoder + distilBERT	0.59	0.56	0.52	0.56	0.61	0.55

- Cohesion and grammar seem to be the toughest to predict across all models
- Future works should focus on improving language models to better capture the grammatical aspects of the language