Auto grader for Short Answer Questions

Pranjal Patil (ppatil@stanford.edu) Ashwin Agrawal (ashwin15@stanford.edu)

Background and Motivation

Short-answer questions can target learning goals more effectively than multiple choice as they eliminate test-taking shortcuts like eliminating improbable answers. However, staff grading of textual answers simply doesn't scale to massive classes. Grading answers has always been time consuming and costs a lot of Public dollars in the US. We start in this project by tackling the simplest problem where we attempt to make an machine learning based system which would automatically grade one line answers based on the given reference answers.

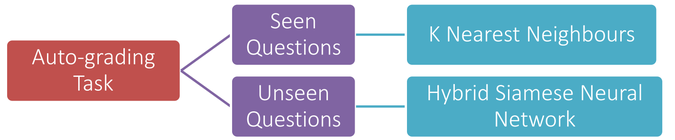
Dataset and Features

- We chose the publicly available Student Response Analysis (SRA) dataset. Within the dataset we used the SciEntsBank part.
- This dataset consists of 135 questions from various physical sciences domain. It has a reference short answer and 36 student responses per question.
- Total size of dataset is 4860 data points.
- Ground truth labels are available in the dataset whether each student response is correct or incorrect.

Data pre-processing including tokenization, stemming and spell checking each of the student responses.

We used the Pre-trained Glove embedding trained on Wikipedia and Gigaword 5 with 400K vocabulary and 300 features.

Models and Results

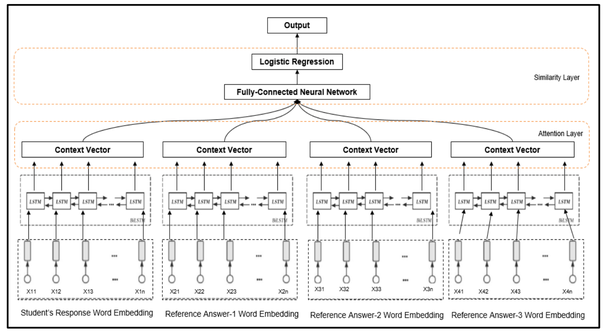


K-nearest neighbours

Accuracy = 79%

Weight = *idf * (w_pos - w_neg);* W_pos = Correct answers with given word/Total number of correct answers
W_neg = Wrong answers with given word/Total number of wrong answers

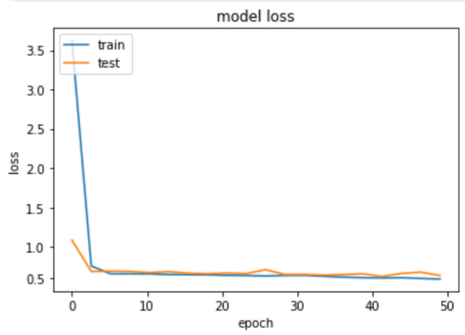
Hybrid Siamese Neural Network



Model	Accuracy (%)	MSE
LSTM + Manhattan Distance [1]	62%	0.25
LSTM + Attention + FNN [2]	73%	0.18
CNN + Bi - LSTM + Manhattan	69%	0.20
Our Model	76%	0.16

- We created an expansion of the Siamese neural network to employ bidirectional LSTM with attention layer and mixed it with KNN's intuition to achieve better results.
- The branches of the network learn sentence embedding for each of the student answer and reference answer. After merging, a fully connected layer measured the similarity between the two answers to score the answer as correct or incorrect.
- In the initial models we used *Manhattan distance, cosine similarity* as the similarity metric.

Data split : 80% train, 10% validation, 10% test data ; ***Loss***: Binary Cross Entropy; ***Optimizer***: Adam
Epochs: 50; ***Attention Layer***: Softmax



Discussion

- Evaluation of sentence similarity can be improved by providing a sentence similarity index in addition to 0/1 labels.
- We found in k-NN approach that correct responses are unexpectedly very similar and hence we inserted more reference answers to cover all writing styles and reinforce the algorithm's similarity detection.
- The hybrid model tend to misclassify long sentences which probably can be improved by using a different attention layer.
- The hybrid model also especially misclassifies the sentences which have the keyword missing in them or written in some other form.

Q. *What is the relation between tree rings and time?*

Ref: As time increases, number of tree rings also increases.
Ans: They are both increasing

Original Label: Correct
Model Result: Misclassified due to missing keywords

Future Work

- Trying out different attention layer to smooth out key word issue
- We would like to improve this model and run it on a larger unseen and out of domain dataset to gauge its robustness.
- Try adding better reference answers or better similarity detection mechanisms.

References

1. Jonas Muller and Aditya Thygarajan, "Siamese Recurrent Architecture for learning sentence similarity", AAAI-16
2. Ziming Chi and Bingyan Zhang, "A sentence similarity estimation method based on improved Siamese Network", JILSA-2018
3. Tianqi Wang et.al, "Identifying Current Issues in Short Answer Grading", ANLP-2018