Arizona State University
CSE 576 - Natural Language Processing (Fall 2020)
Submission of NLP Project - Word Math Problems
Training Phase-1

Group Members: Shivam Raval, Parthav Patel, Kalp Patel,
Sharadhi Jagnath, Rushita Thakkar

November 10, 2020

- **Introduction:**

In this project phase we trained different models that have been already proposed for solving word math problems. Additionally, we also performed testing using the original test dataset as well the synthesized dataset generated in earlier phase. The training and testing summary for all the models that we implemented can be found below. The ipynb notebook that consists of how the training and testing was performed can be found here on Github.

- **Summary of Training Models:**

Model	F1-Score	Exact-Match Accuracy (EM)	Loss
NumNet +v2	0.73935	0.695	0.79775
NumNet	0.5224	0.4782	2.441
GenBert	0.4913	0.4743	-

- **Summary of Testing Models:**

Model	F1-Score	Exact-Match Accuracy (EM)
NumNet +v2	61.95	58.38
GenBert (bert + drop)	43.12	41.18
GenBert (bert + drop + ND + TD)	66.65	64.71

NumNet +v2 - This model ran for 5 epochs and 14.3 K iterations. The dataset consists of 11.5K questions randomly sampled from the original 95 K questions. The complete dataset wasn't used to train the model as this resulted in memory issues.

GenBert - This model ran for 4 epochs with 200 iterations per epoch for 1.5 k samples. The smaller size of the training samples is due to the large time taken to train the dataset. In parallel, a pre-trained model provided by gen-bert github repo which is trained on original 95k

questions was used to validate 500 samples.

NumNet - This model ran for 8 epochs. The dataset consists of 11.5 K questions randomly sampled from the original 95 K questions. The complete dataset wasn't used to train the model as this resulted in memory issues.

Testing on Synthesized Dataset:

After training and testing the model on the original dataset, we also tested the model on a synthesized version of the DROP dataset prepared using the text spinner technique proposed in the data creation phase. The results of testing the synthesized dataset on NumNet +v2 is shown below:
Synthesized Testing Samples: 499
Exact-match accuracy: 36.07
F1 score: 39.65