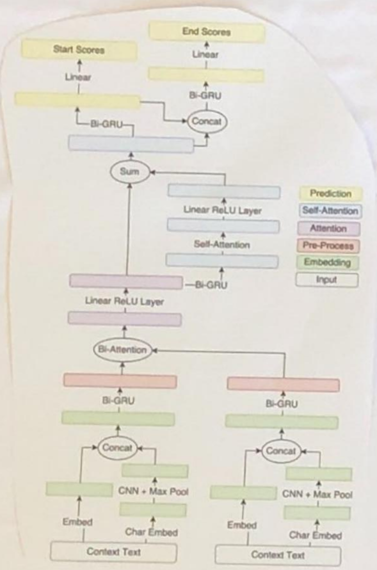
Implementation and Analysis of Character-level Embeddings, Self-attention, and R-NET on BiDAF for QA on SQuAD 2.0

By: Rohin Manvi and Avash Shrestha

Question answering (QA) is one of the most important tasks in the field of natural language processing (NLP). The overarching goal of QA is to make models that are able to extract meaning from texts in order to accurately answer questions on the model. This is a crucial task for a variety of applications, from chat bots to social media to medical texts.

We propose implementing character embeddings, self-attention, and a combination of embeddings and encoder layers on top of the starter model. We ran various combinations of the features we implemented on top of the starter model. We found that adding these features generally helped **improve the EM/F1 scores** of our model, likely because the information the model had access to was of higher quality.



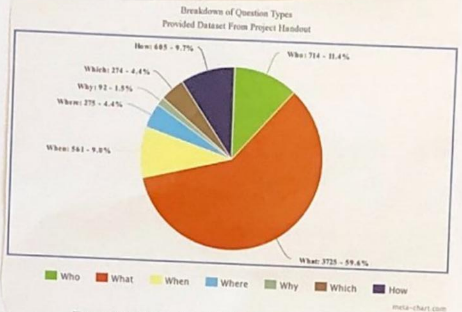


Figure 4: Breakdown of QUestion Types from provided dataset [1]

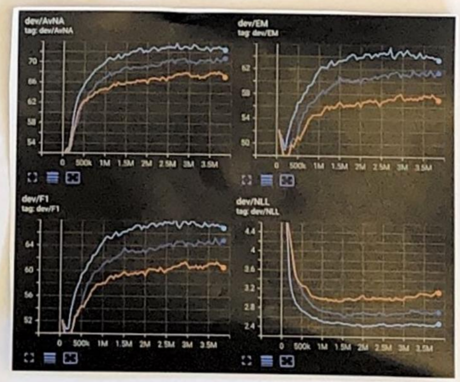


Table 1: Results

Model	Leaderboard	LR	Dropout	Hidden	EM	F1
BiDAF (baseline)	dev	0.5	0.2	100	58.58	61.90
BiDAF + Char Embed	dev	0.5	0.2	100	62.02	65.34
BiDAF + Char Embed	dev	0.6	0.15	100	62.81	66.21
BiDAF + R-NET (concat)	dev	0.5	0.2	100	57.50	60.90
BiDAF + R-NET (add)	dev	0.5	0.2	100	57.60	60.97
BiDAF + R-NET (concat) + Self-Att.	dev	0.5	0.2	100	61.91	64.89
BiDAF + R-NET (add) + Self-Att.	dev	0.5	0.2	100	63.20	66.32
BiDAF + Char Embed + Self-Att.	dev	0.5	0.2	100	63.91	67.26
BiDAF + Char Embed + Self-Att.	dev	0.6	0.15	100	63.23	66.16
BiDAF + Char Embed + Self-Att.	dev	0.5	0.2	125	65.48	68.40
BiDAF + Char Embed + Self-Att.	test	0.5	0.2	125	62.06	65.14

Table 2: Best Model by Question Type

Question Type	Who	What	When	Where	Why	Which	How
Size	714	3726	561	275	92	274	605
EM	65.28	65.26	68.10	65.44	59.09	68.69	62.64
F1	67.39	68.17	69.97	69.29	65.87	70.95	67.13
Mean Answer Length (characters)	17.6	21.6	11.6	18.9	45.3	17.2	18.5
Mean Prediction Length (characters)	17.8	20.3	11.3	17.8	48.9	16.9	16.1
Mean Answer Rate	0.44	0.45	0.52	0.48	0.50	0.53	0.45
Mean Prediction Rate	0.55	0.49	0.59	0.56	0.35	0.52	0.55

To conclude, we implemented 3 distinct feature: **character embeddings, self-attention, and R-NET's embeddings/encoding layers.** We found our best model to be a combination of the baseline, character embeddings, and self-attention, which performed slightly worse on the test set, likely due to overfitting, but still performed well. We showed an **overall increase in performance** as compared to the given baseline.