Every Answer begins with a Question: Ask me Another!

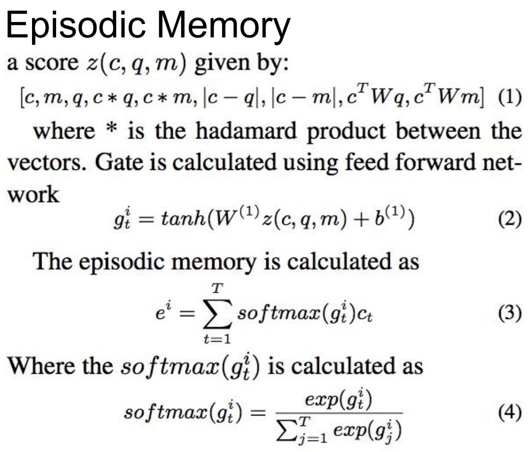
W266 Section 3

Anup Jha, Chitra Agastya

Dynamic Memory Network for SQuAD 2.0 QA

DMN Worked well with bAbi tasks.

Is it effective on SQuAD 2.0?



Episodic Memory

a score $z(c,q,m)$ given by:

$$[c, m, q, c * q, c * m, |c-q|, |c-m|, c^T W q, c^T W m] \quad (1)$$

where * is the hadamard product between the vectors. Gate is calculated using feed forward network

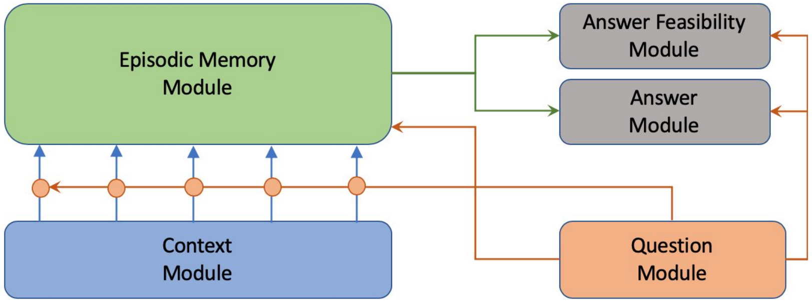
$$g_t^i = tanh(W^{(1)} z(c,q,m) + b^{(1)}) \quad (2)$$

The episodic memory is calculated as

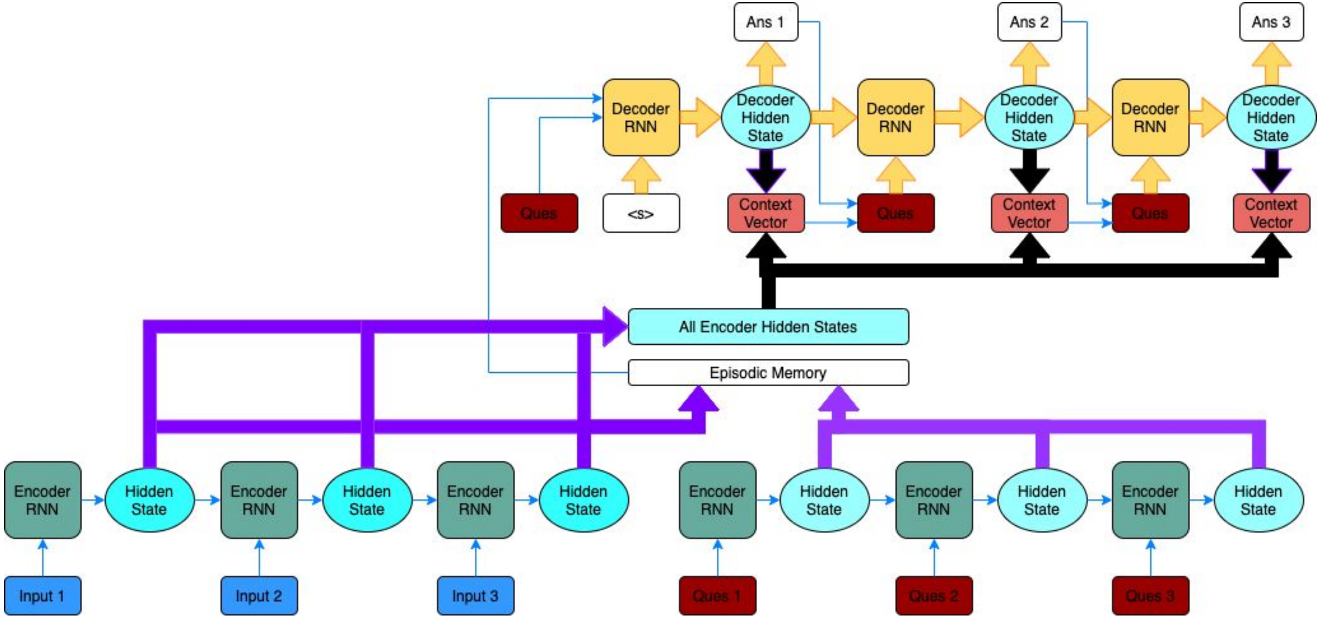
$$e^i = \sum_{t=1}^{T} softmax(g_t^i) c_t \quad (3)$$

Where the $softmax(g_t^i)$ is calculated as

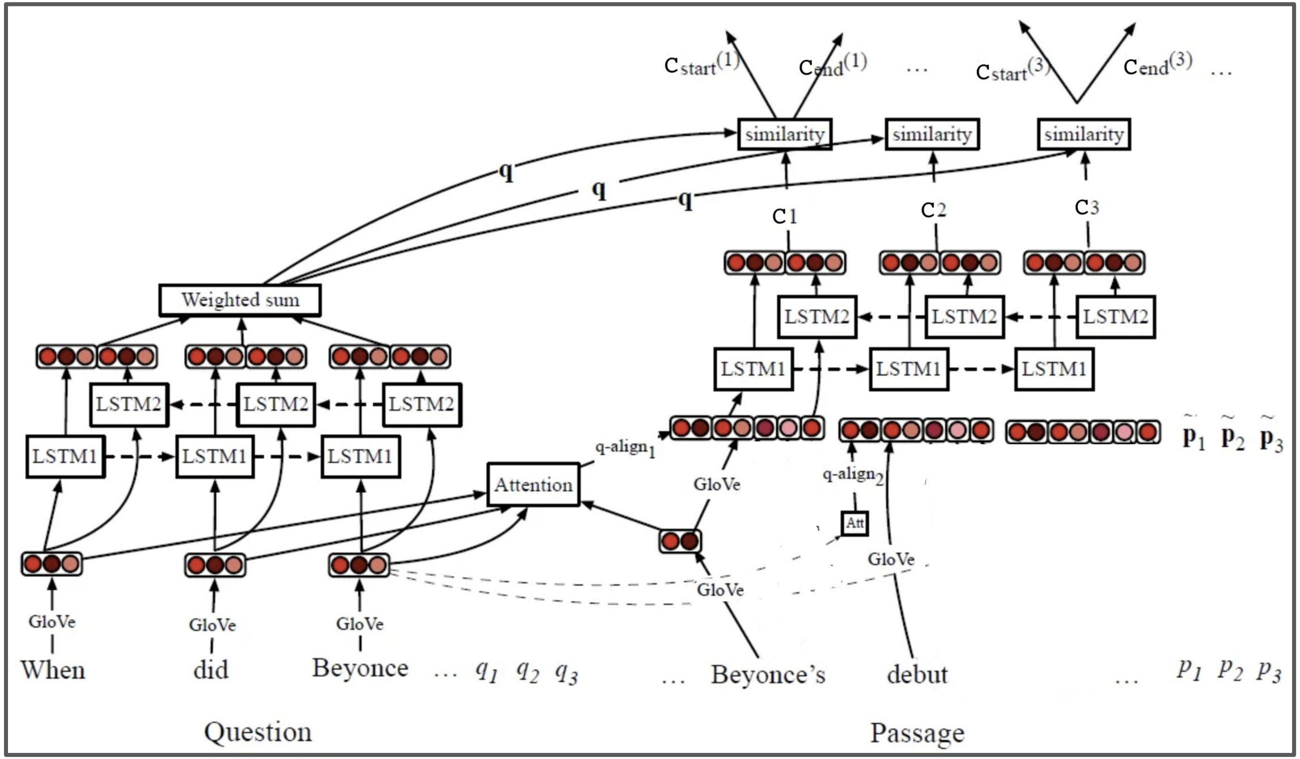
$$softmax(g_t^i) = \frac{exp(g_t^i)}{\sum_{j=1}^{T} exp(g_j^i)} \quad (4)$$



Attention based Model



Span Based Model



$$f_{align}(c_i) = \sum_j a_{i,j} E(q_j)$$

$$a_{i,j} = \frac{exp(\alpha(E(p_i)).\alpha(E(q_j)))}{\sum_k exp(\alpha(E(p_i)).\alpha(E(q_k)))}$$

$$C_{start(i)} \propto exp(c_i W_s q)$$

$$C_{end(i)} \propto exp(c_i W_e q)$$

Experiments

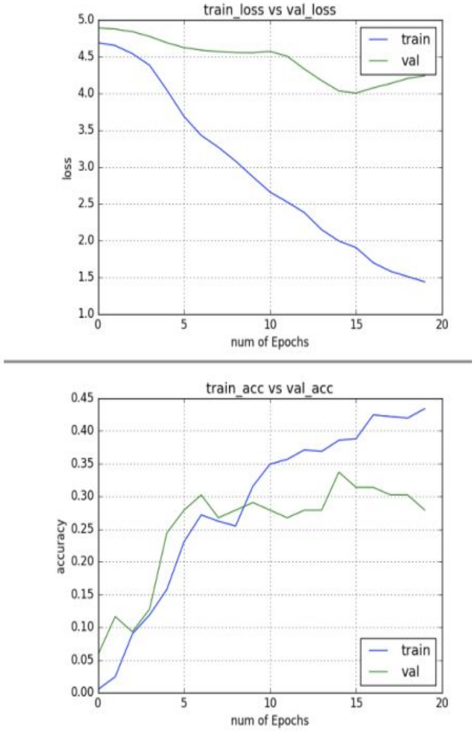
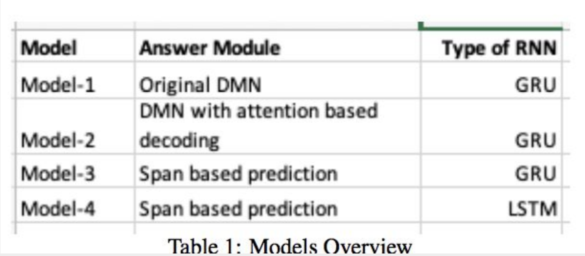
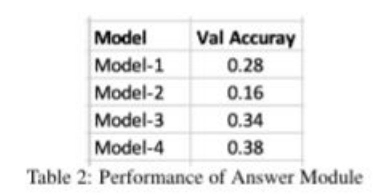


Figure 3: Loss and Accuracy: Model-1



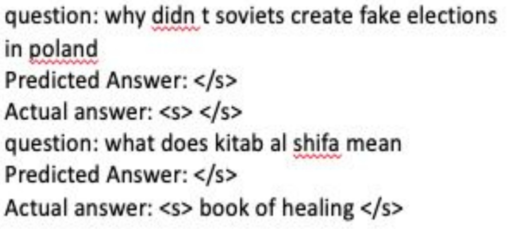
Model	Answer Module	Type of RNN
Model-1	Original DMN	GRU
Model-2	DMN with attention based decoding	GRU
Model-3	Span based prediction	GRU
Model-4	Span based prediction	LSTM

Table 1: Models Overview



Model	Val Accuray
Model-1	0.28
Model-2	0.16
Model-3	0.34
Model-4	0.38

Table 2: Performance of Answer Module



question: why didn t soviets create fake elections in poland
Predicted Answer: </s>
Actual answer: <s> </s>
question: what does kitab al shifa mean
Predicted Answer: </s>
Actual answer: <s> book of healing </s>

Sample Predictions: Model-1

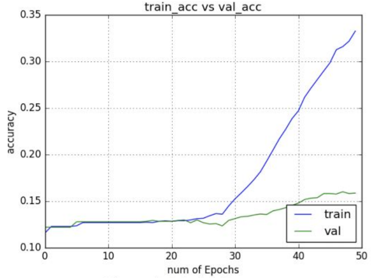
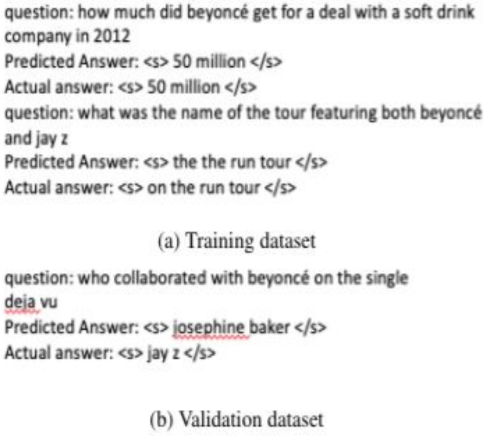


Figure 4: Model-2 accuracy



question: how much did beyoncé get for a deal with a soft drink company in 2012
Predicted Answer: <s> 50 million </s>
Actual answer: <s> 50 million </s>
question: what was the name of the tour featuring both beyoncé and jay z
Predicted Answer: <s> the the run tour </s>
Actual answer: <s> on the run tour </s>

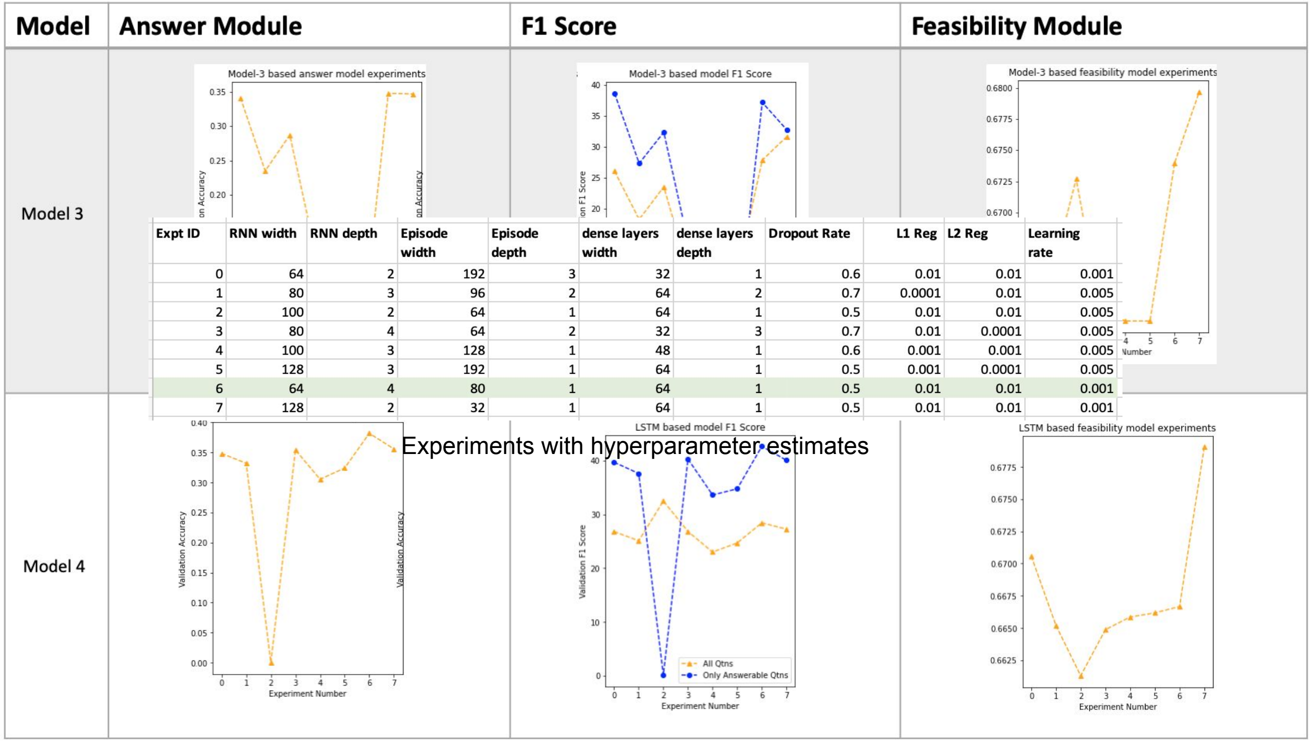
(a) Training dataset

question: who collaborated with beyoncé on the single deja vu
Predicted Answer: <s> josephine baker </s>
Actual answer: <s> jay z </s>

(b) Validation dataset

Figure 9: Sample Predictions: Model-2

Experiments

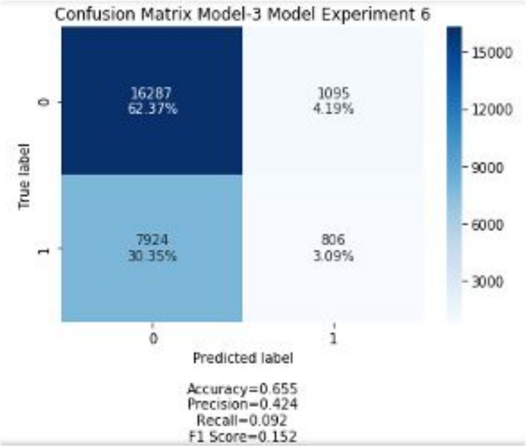


Model	Answer Module	F1 Score	Feasibility Module
Model 3	Model-3 based answer model experiments	Model-3 based model F1 Score	Model-3 based feasibility model experiments
Model 4		LSTM based model F1 Score	LSTM based feasibility model experiments

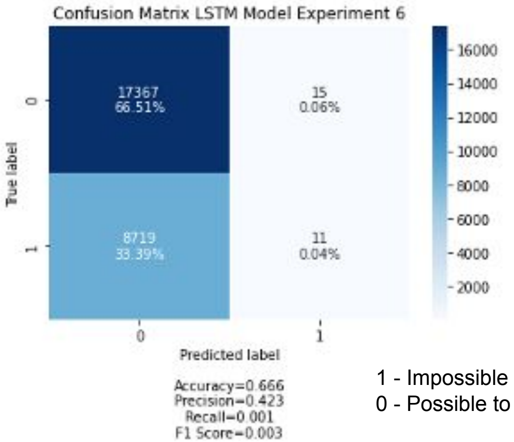
Expt ID	RNN width	RNN depth	Episode width	Episode depth	dense layers width	dense layers depth	Dropout Rate	L1 Reg	L2 Reg	Learning rate
0	64	2	192	3	32	1	0.6	0.01	0.01	0.001
1	80	3	96	2	64	2	0.7	0.0001	0.01	0.005
2	100	2	64	1	64	1	0.5	0.01	0.01	0.005
3	80	4	64	2	32	3	0.7	0.01	0.0001	0.005
4	100	3	128	1	48	1	0.6	0.001	0.001	0.005
5	128	3	192	1	64	1	0.5	0.001	0.0001	0.005
6	64	4	80	1	64	1	0.5	0.01	0.01	0.001
7	128	2	32	1	64	1	0.5	0.01	0.01	0.001

Experiments with hyperparameter estimates

Results

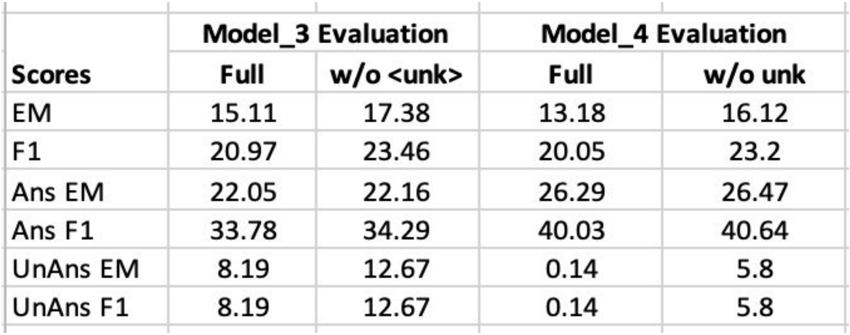

Confusion Matrix Model-3 Model Experiment 6

Accuracy=0.655
Precision=0.424
Recall=0.092
F1 Score=0.152


Confusion Matrix LSTM Model Experiment 6

Accuracy=0.666
Precision=0.423
Recall=0.001
F1 Score=0.003

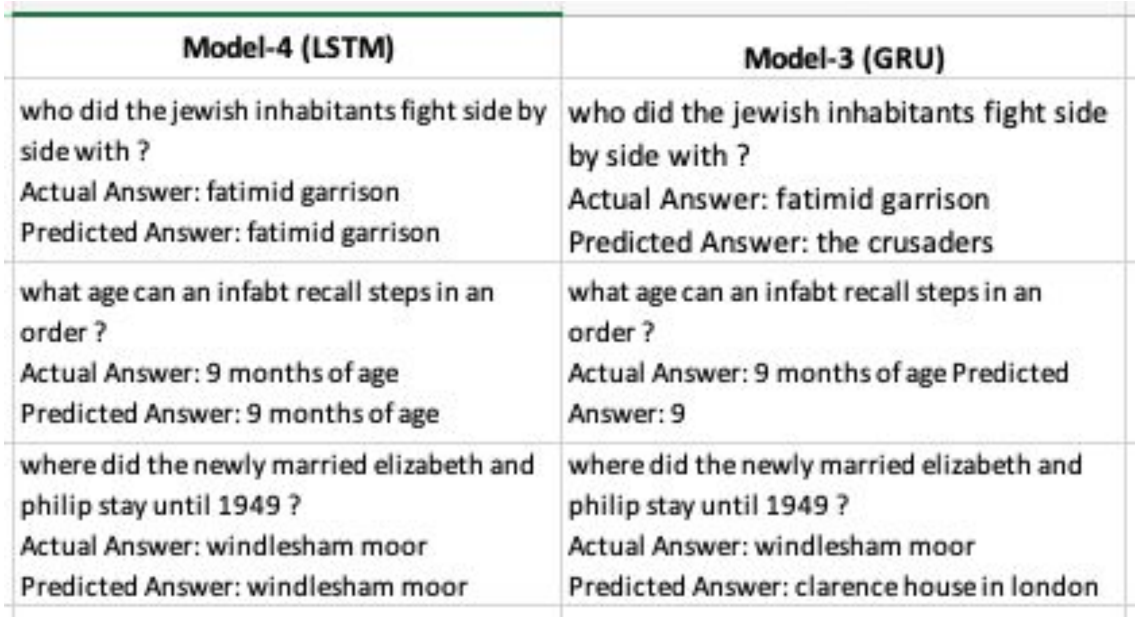
1 - Impossible to answer
0 - Possible to answer



	Model_3 Evaluation		Model_4 Evaluation	
Scores	Full	w/o <unk>	Full	w/o unk
EM	15.11	17.38	13.18	16.12
F1	20.97	23.46	20.05	23.2
Ans EM	22.05	22.16	26.29	26.47
Ans F1	33.78	34.29	40.03	40.64
UnAns EM	8.19	12.67	0.14	5.8
UnAns F1	8.19	12.67	0.14	5.8

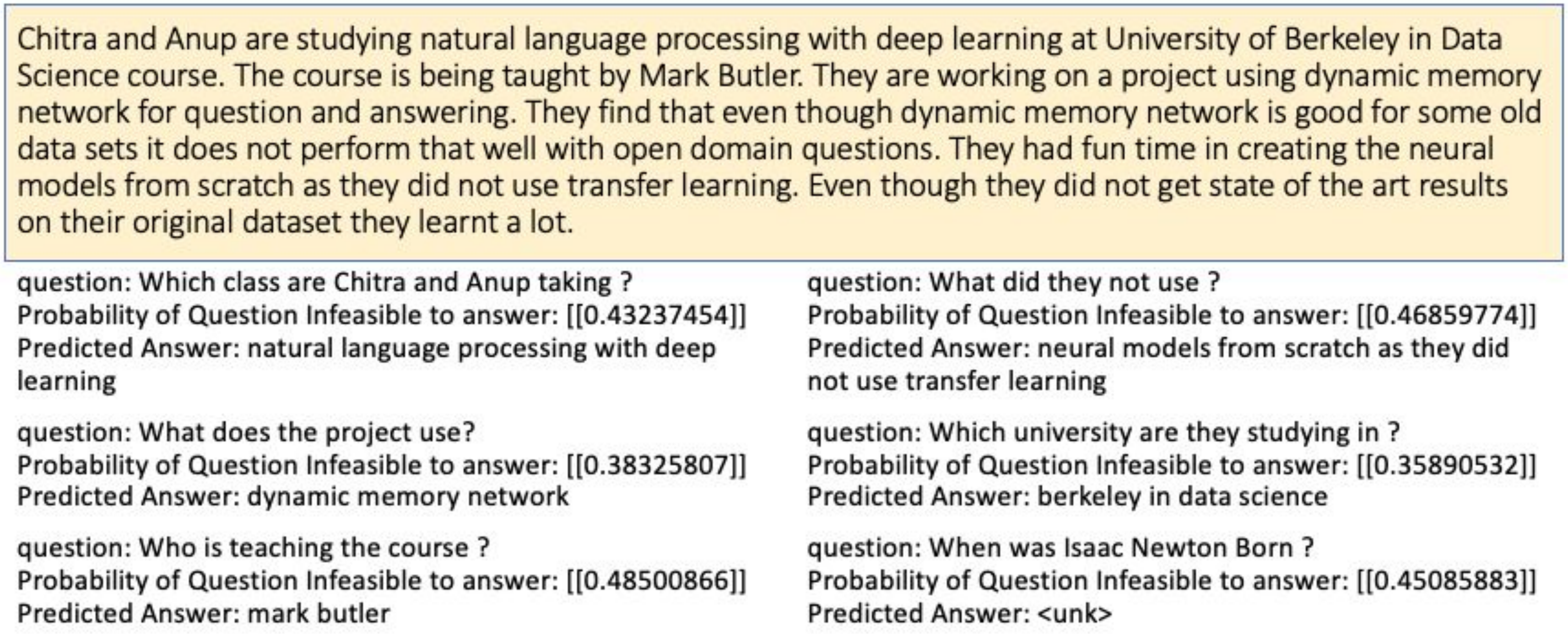
Evaluation scores for Test Data

LSTM vs GRU Predictions



Model-4 (LSTM)	Model-3 (GRU)
who did the jewish inhabitants fight side by side with ? Actual Answer: fatimid garrison Predicted Answer: fatimid garrison	who did the jewish inhabitants fight side by side with ? Actual Answer: fatimid garrison Predicted Answer: the crusaders
what age can an infabt recall steps in an order ? Actual Answer: 9 months of age Predicted Answer: 9 months of age	what age can an infabt recall steps in an order ? Actual Answer: 9 months of age Predicted Answer: 9
where did the newly married elizabeth and philip stay until 1949 ? Actual Answer: windlesham moor Predicted Answer: windlesham moor	where did the newly married elizabeth and philip stay until 1949 ? Actual Answer: windlesham moor Predicted Answer: clarence house in london

Sample - non Wiki passage



Chitra and Anup are studying natural language processing with deep learning at University of Berkeley in Data Science course. The course is being taught by Mark Butler. They are working on a project using dynamic memory network for question and answering. They find that even though dynamic memory network is good for some old data sets it does not perform that well with open domain questions. They had fun time in creating the neural models from scratch as they did not use transfer learning. Even though they did not get state of the art results on their original dataset they learnt a lot.

```
question: Which class are Chitra and Anup taking ?
Probability of Question Infeasible to answer: [[0.43237454]]
Predicted Answer: natural language processing with deep
learning

question: What does the project use?
Probability of Question Infeasible to answer: [[0.38325807]]
Predicted Answer: dynamic memory network

question: Who is teaching the course ?
Probability of Question Infeasible to answer: [[0.48500866]]
Predicted Answer: mark butler
```

```
question: What did they not use ?
Probability of Question Infeasible to answer: [[0.46859774]]
Predicted Answer: neural models from scratch as they did
not use transfer learning

question: Which university are they studying in ?
Probability of Question Infeasible to answer: [[0.35890532]]
Predicted Answer: berkeley in data science

question: When was Isaac Newton Born ?
Probability of Question Infeasible to answer: [[0.45085883]]
Predicted Answer: <unk>
```

Conclusion

- DMN by itself performed poorly on SQuAD 2.0 dataset
  - SQuAD 2.0 has longer spans of answer
  - SQuAD 2.0 has a semi open nature where not all questions are answerable by the context
- Use of attentions and spans helped improved the model’s prediction of answers
- LSTM gave better answer prediction than GRU
- Deeper RNN gave better performance than wider RNN

Backup

Modifications to original DMN

Original DMN	**Our Version of DMN**
Input module emits vectors for every sentence in context	Input module emits vectors for every word in context
RNN in forward direction	Bidirectional RNN
4 modules: question, input, episodic memory and answer	5 modules: question, input, episodic memory, feasibility and answer
No use of explicit attentions.	Some of our models have attentions in answer modules