IBM Watson

On Machine Reading Comprehension and Question Answering

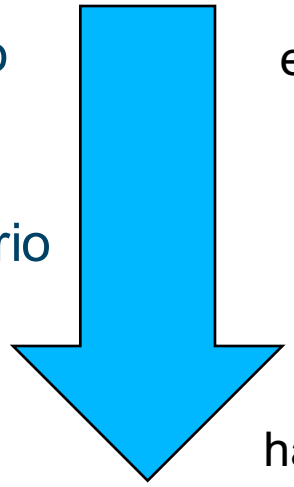
Wei Zhang, Yang Yu, Bowen Zhou
RCQA Squad, IBM Watson

Nov. 21, 2016 @Harvard NLP

IBM

Overview

- RCQA as inference task in an ideal scenario
- RCQA in a real world single passage scenario
- RCQA in a multi-passage scenario



- Remaining Challenges

Overview

- **RCQA as inference task in an ideal scenario**
- RCQA in a real world single passage scenario
- RCQA in a multi-passage scenario
- Remaining Challenges

RCQA as inference task in an ideal scenario

Tasks:

1）Associative Recall

Input string	Target
c9k8j3f1??c	9
j0a5s5z2??a	5

2）bAbI dataset (Weston et al. 2015)

task (1): Single supporting fact

F: Mary went to the Kitchen.
Q: Where is marry?
A: Kitchen.

task (2): two supporting facts

F: John got the football there.
F: John went to the hallway.
Q: Where is the football?
A: Hallway

task (3): three supporting facts

F: Mary journeyed to the office.
F: Mary journeyed to the bathroom.
F: Mary dropped the football.
Q: Where was the football before the bathroom?
A: office

An excerpt from bAbI dataset

1 John travelled to the hallway.
2 Mary journeyed to the bathroom.
3 Where is John? hallway 1 4
Daniel went back to the bathroom.
5 John moved to the bedroom.
6 Where is Mary? bathroom 2

RCQA as inference task in an ideal scenario

Model:

Structured Memory NTM (Zhang et al., 2015, NIPS RAM workshop)

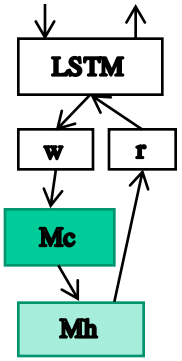
next page

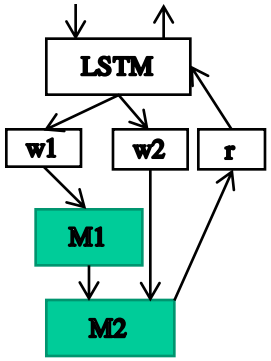
Neural Machine Translation (Yu et al., 2015)

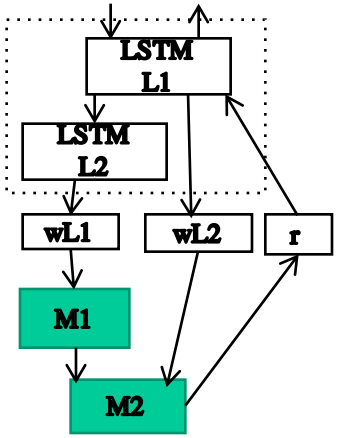
Concatenate passage and question, feed into words into encoder. Decode single token answer. (87% acc. on bAbI 10K, v.s. ~100% so far)

Structured Memory for Neural Turing Machines

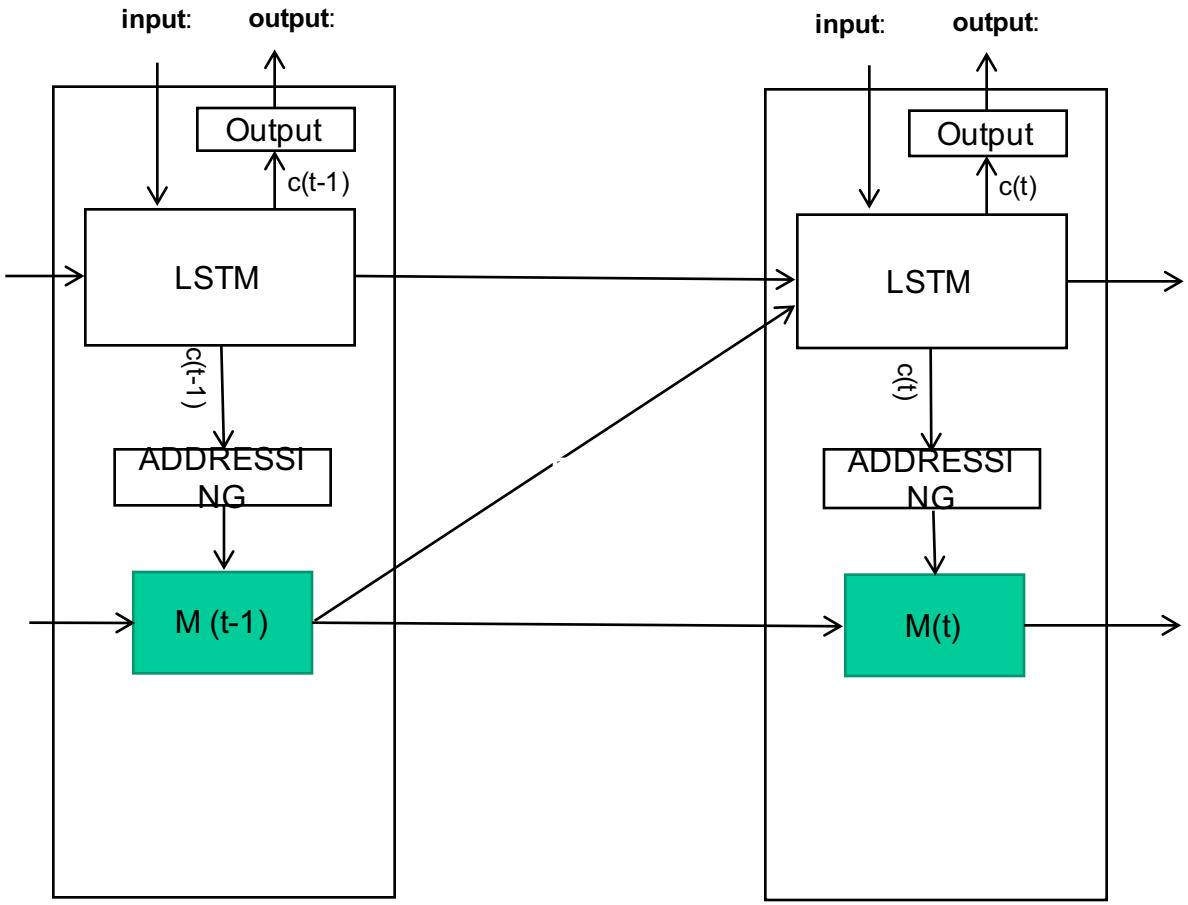
Memory visability (w.r.t. write heads): *memory is 'Controlled' if it is modified by controller outputs through write heads directly, or 'Hidden' if not.*





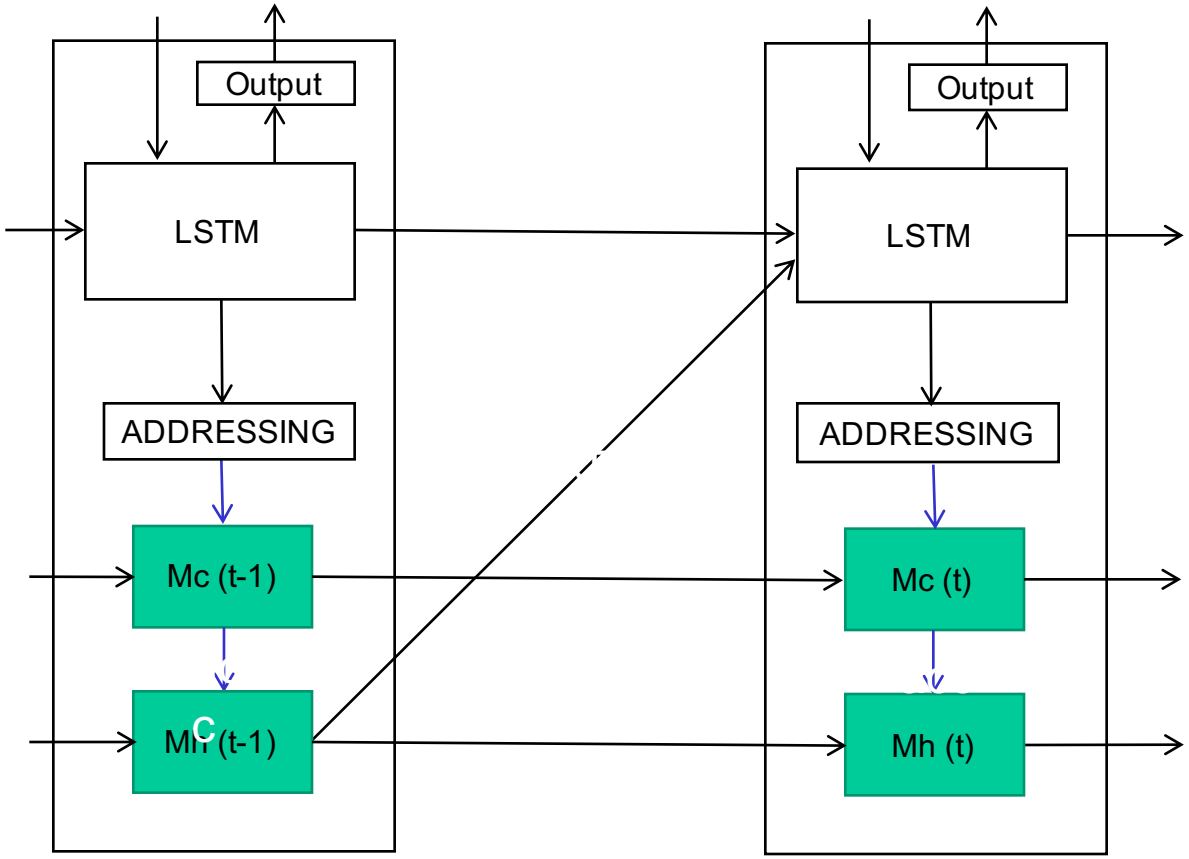


Vanilla NTM

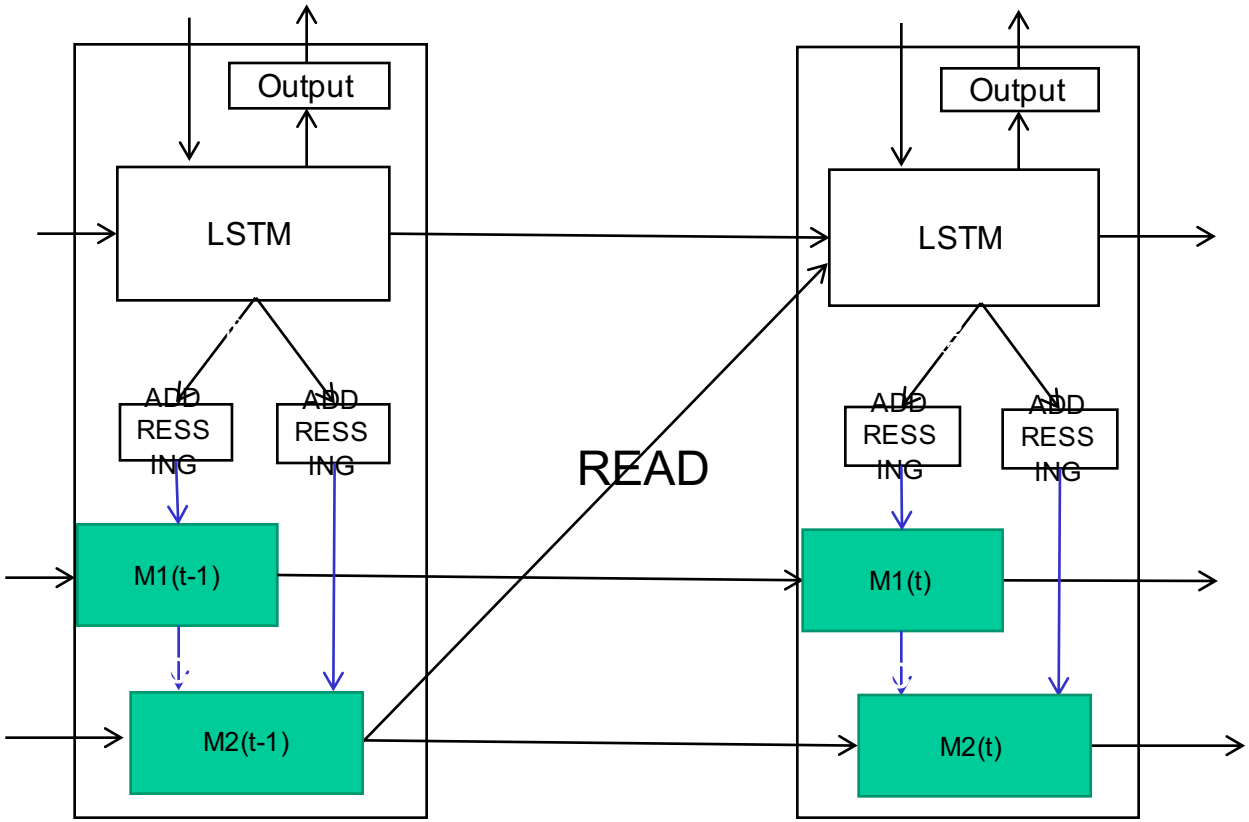

input:
output:
Output
c(t-1)
LSTM
c(t-1)
ADDRESSI
NG
M (t-1)
input:
output:
Output
c(t)
LSTM
c(t)
ADDRESSI
NG
M(t)

#1: *Hidden-Memory NTM*

$$write: \quad \mathbf{M}_c(t) = h\big(\mathbf{M}_c(t-1), \mathbf{w}(t-1), \mathbf{c}(t)\big)$$
$$update: \quad \mathbf{M}_h(t) = a\mathbf{M}_h(t-1) + b\mathbf{M}_c(t)$$
$$read: \quad \mathbf{r}(t) = \mathbf{w_r}(t)\mathbf{M}_h(t)$$



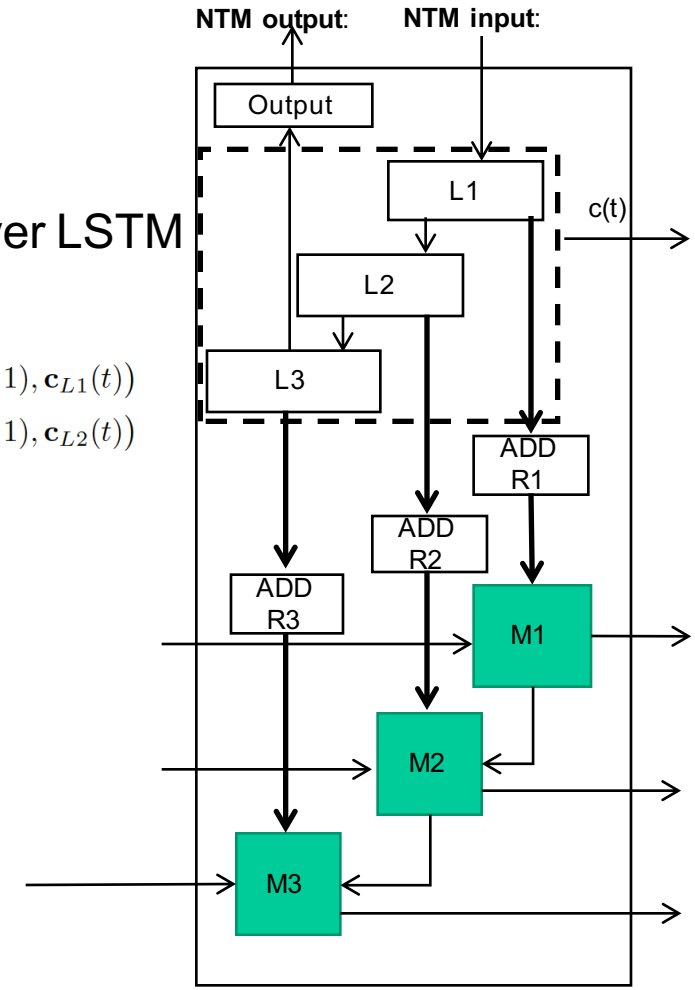
#2: *Double-Controlled NTM*



#3 Tightly-Coupled NTM

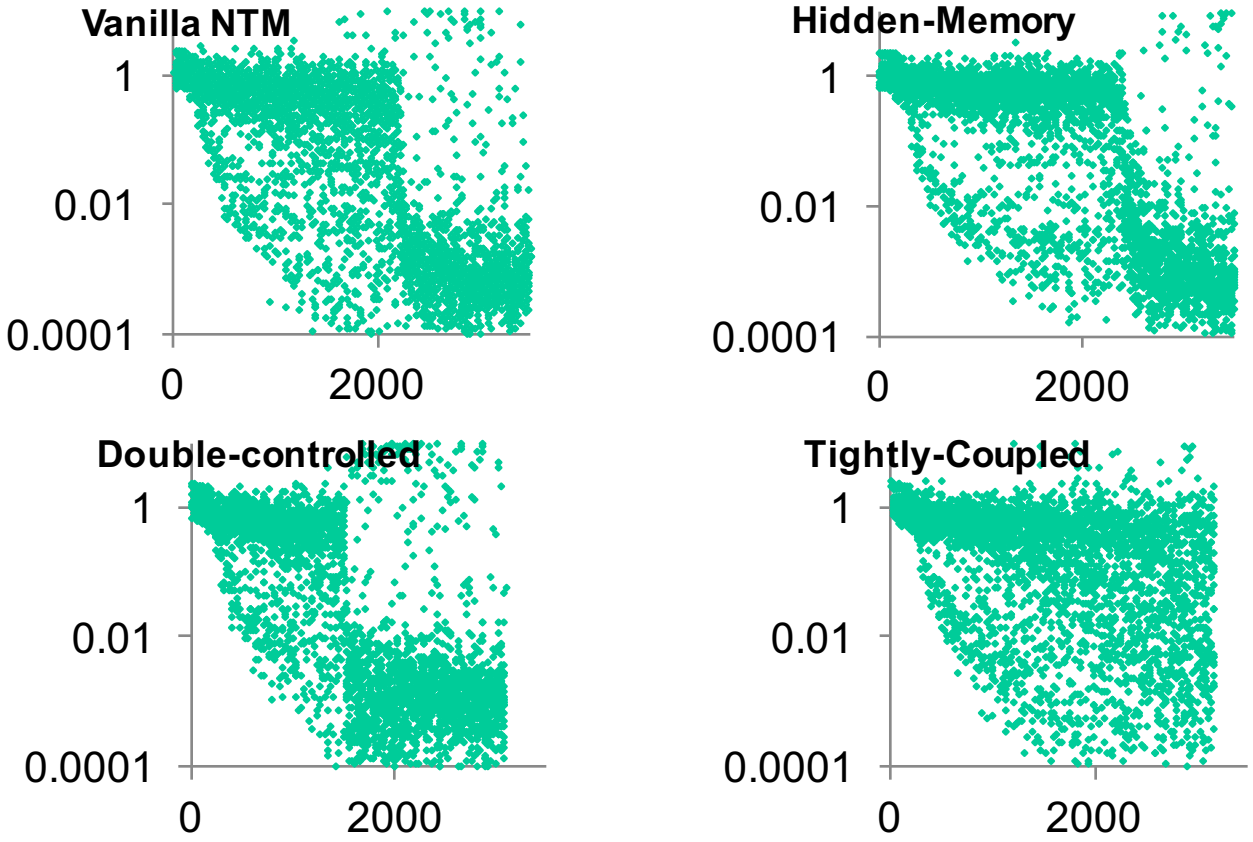
Multi-layer LSTM

$$
\begin{aligned}
write: \quad & \mathbf{M}_1(t), \mathbf{w}_{L1}(t) = h\big(\mathbf{M}_1(t-1), \mathbf{w}_{L1}(t-1), \mathbf{c}_{L1}(t)\big) \\
& \widetilde{\mathbf{M}}_2(t), \mathbf{w}_{L2}(t) = h\big(\mathbf{M}_2(t-1), \mathbf{w}_{L2}(t-1), \mathbf{c}_{L2}(t)\big) \\
update: \quad & \mathbf{M}_2(t) = a\widetilde{\mathbf{M}}_2(t) + b\mathbf{M}_1(t) \\
read: \quad & \mathbf{r}(t) = \mathbf{w}_{\mathbf{r}}(t)\mathbf{M}_2(t)
\end{aligned}
$$



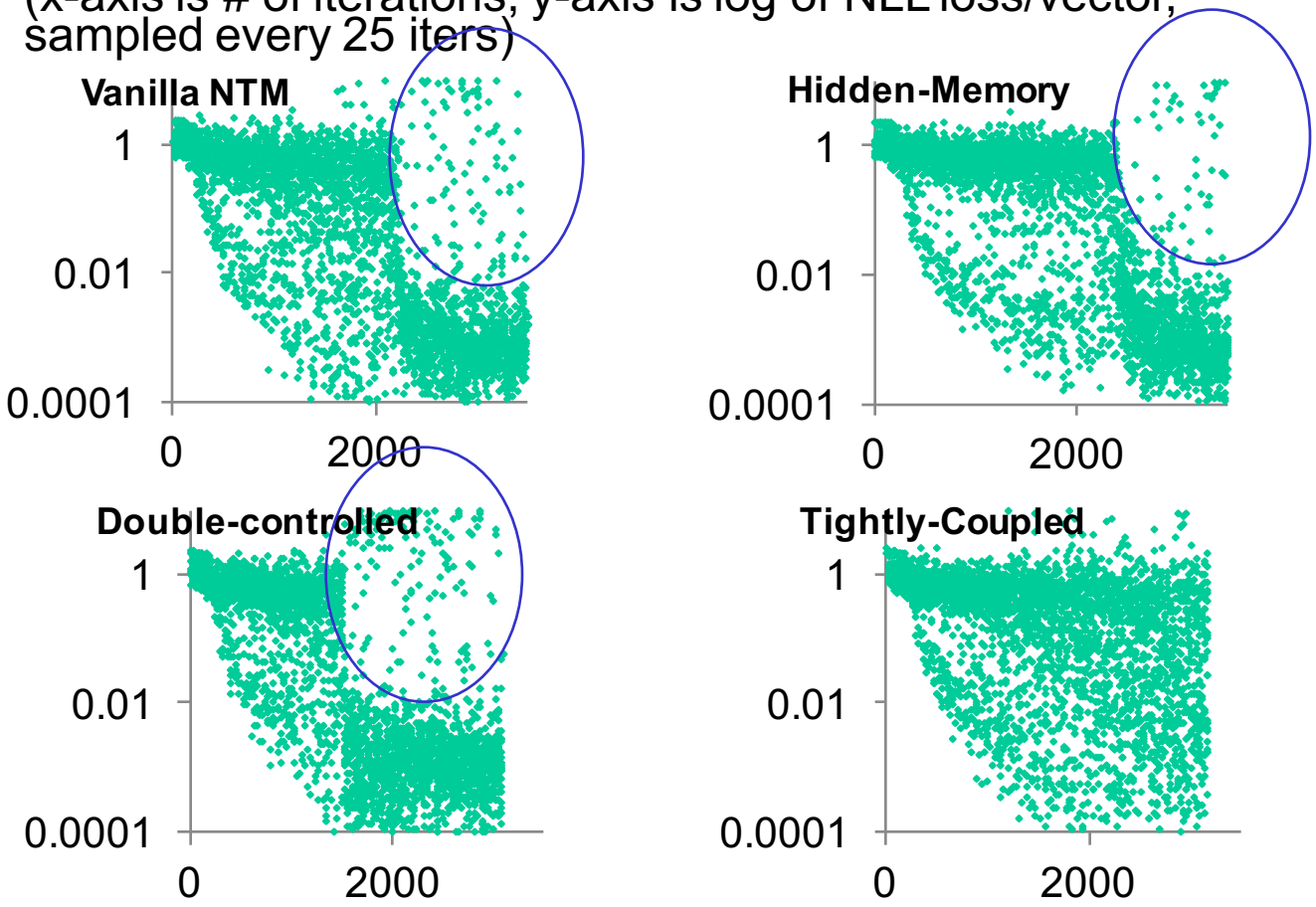
Associative Recall

(x-axis is # of iterations, y-axis is log of NLL loss/vector, sampled every 25 iters)

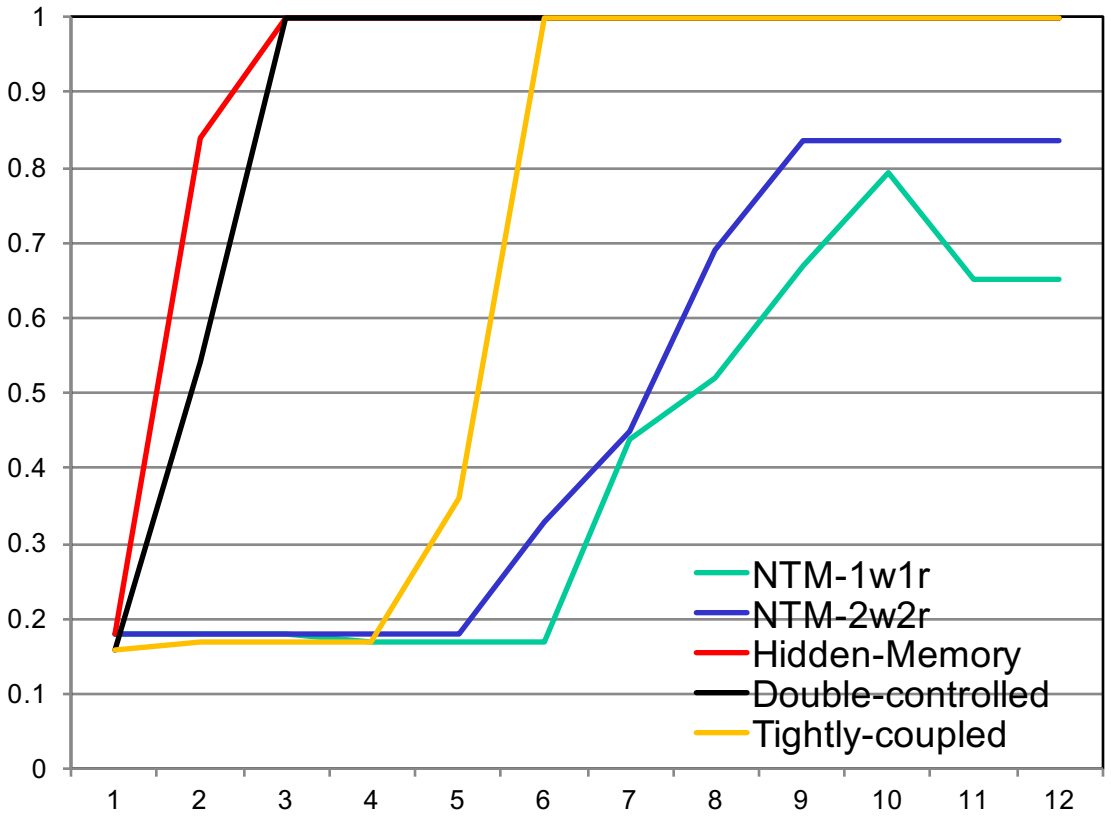


Associative Recall

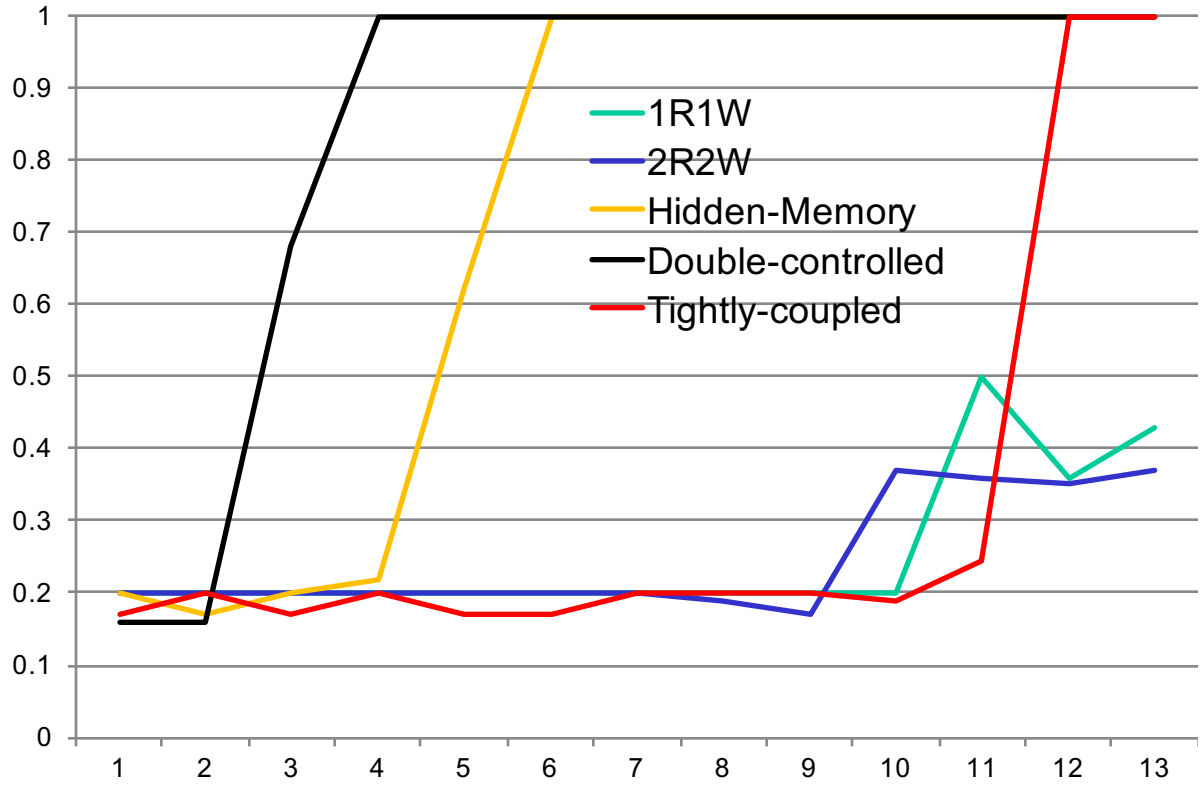
(x-axis is # of iterations, y-axis is log of NLL loss/vector, sampled every 25 iters)



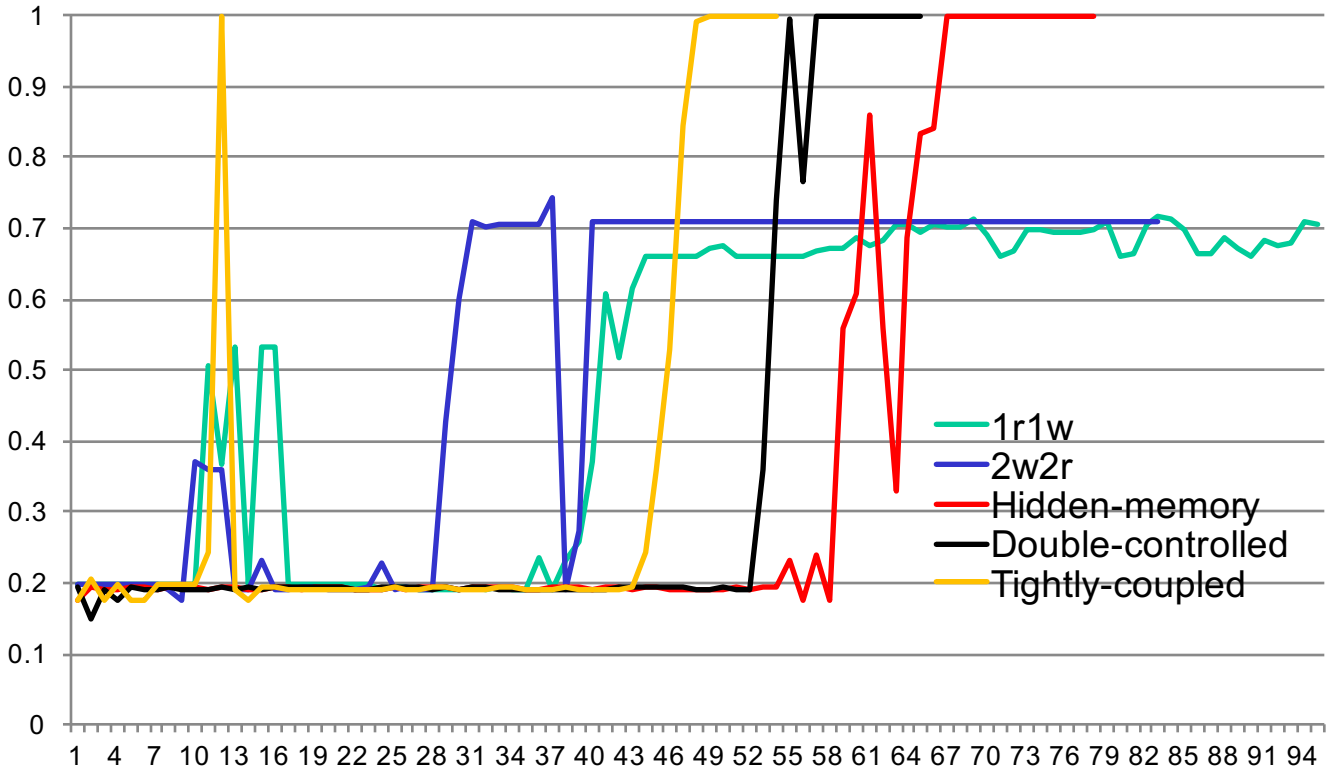
bAbI task 1 (1K)
supporting fact only



bAbI task 2 (1k train/test)
supporting fact only



Task 3(1K) Accuracy



Overview

- RCQA as inference task in an ideal scenario
- **RCQA in a real world single passage scenario**
- RCQA in a multi-passage scenario
- Remaining Challenges

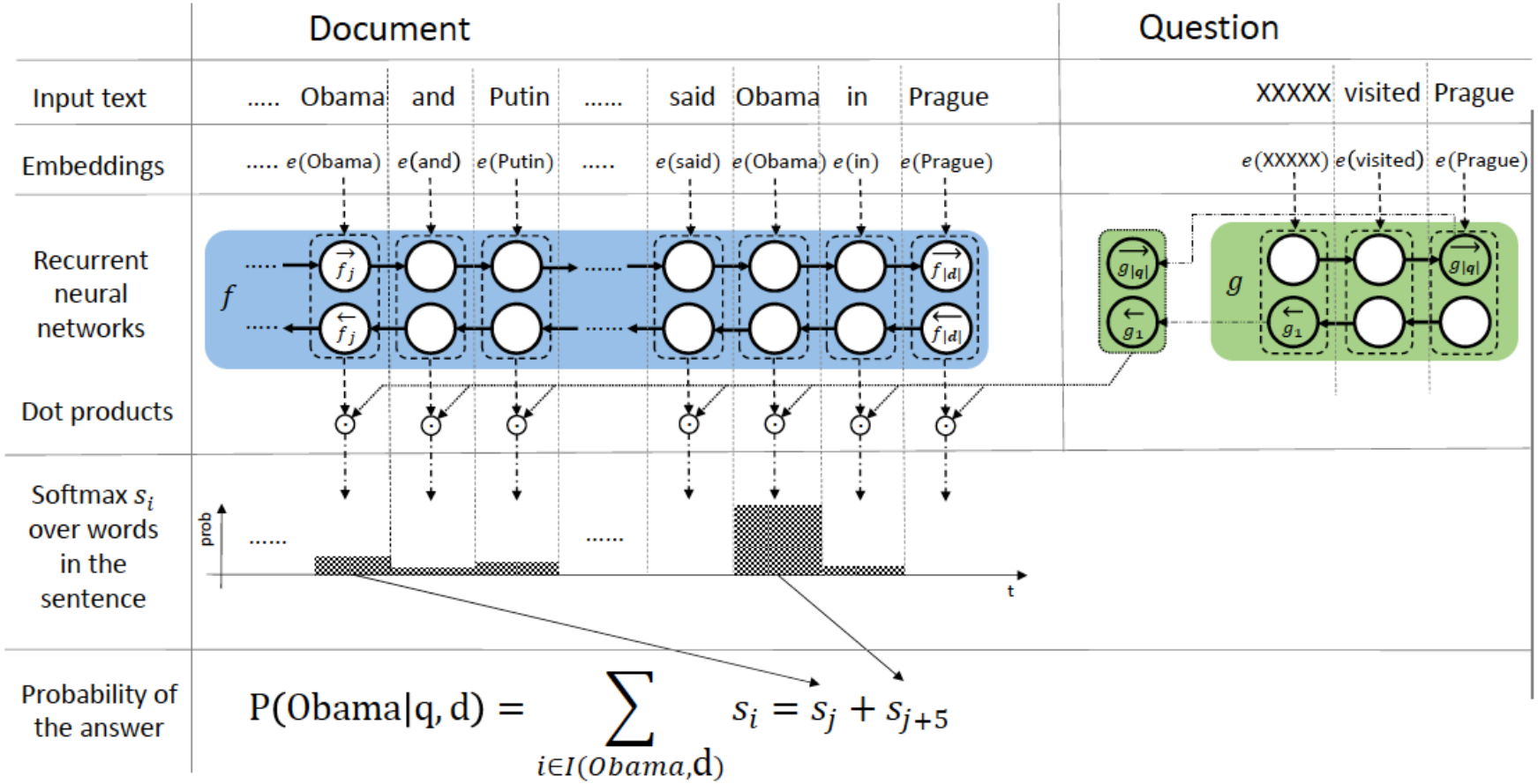
Quiz and Cloze Style Question Answering

- MC Test / Movie QA
  - Complexity is 4-5
- CNN/DailyMail/CBT
  - Complexity is 30 to 40


```
{
  "qid": "train:0",
  "question": "Why does Hook leave Eamon's apartment?",
  "correct_index": 4,
  "imdb_key": "tt2614684",
  "answers": [
    "Because he is warned it's not safe.",
    "Because he wants to get some fresh air.",
    "Because he is called away.",
    "Because he has to get some stuff.",
    "Because he senses danger."
  ],
  "video_clips": []
},
```


Original Version	Anonymised Version
Context	
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " ...
Query	
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	Producer **X** will not press charges against *ent212*, his lawyer says.
Answer	
Oisin Tymon	*ent193*

Attention Sum Reader (Kadlec et. al 2016)



Attention Sum Reader (Kadlec el. al 2016)

	CNN		Daily Mail	
	valid	test	valid	test
Deep LSTM Reader †	55.0	57.0	63.3	62.2
Attentive Reader †	61.6	63.0	70.5	69.0
Impatient Reader †	61.8	63.8	69.0	68.0
MemNNs (single model) ‡	63.4	66.8	NA	NA
MemNNs (ensemble) ‡	66.2	69.4	NA	NA
AS Reader (single model)	68.6	69.5	75.0	73.9
AS Reader (avg for top 20%)	68.4	69.9	74.5	73.5
AS Reader (avg ensemble)	73.9	**75.4**	78.1	77.1
AS Reader (greedy ensemble)	74.5	74.8	78.7	**77.7**

CNN/Daily Mail Results

CBT Results

	Named entity		Common noun	
	valid	test	valid	test
Humans (query) (*)	NA	52.0	NA	64.4
Humans (context+query) (*)	NA	***81.6***	NA	***81.6***
LSTMs (context+query) ‡	51.2	41.8	62.6	56.0
MemNNs (window memory + self-sup.) ‡	70.4	66.6	64.2	63.0
AS Reader (single model)	73.8	68.6	68.8	63.4
AS Reader (avg for top 20%)	73.3	68.4	67.7	63.2
AS Reader (avg ensemble)	74.5	70.6	71.1	**68.9**
AS Reader (greedy ensemble)	76.2	**71.0**	72.4	67.5

Stanford Question Answering Dataset (SQuAD)

- Answer is a span of text in any length
  - Complexity of choosing candidates is O(n^2), n is passage length which is often over 1000
  - 86k Train / 10k Dev

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

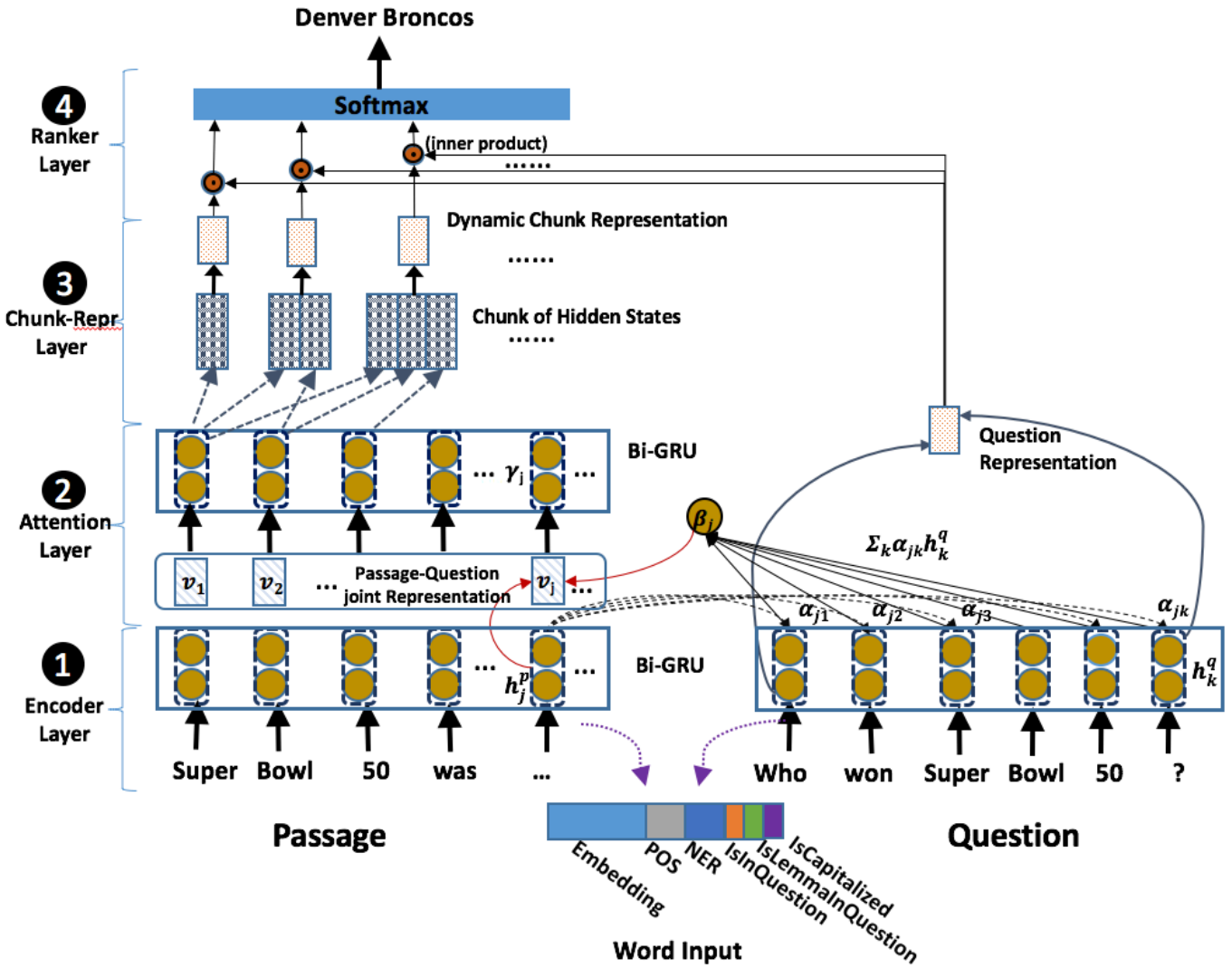
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

SQuAD

Reasoning	Description	Example	Percentage
Lexical variation (synonymy)	Major correspondences between the question and the answer sentence are synonyms.	Q: What is the Rankine cycle sometimes **called**? Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle.	33.3%
Lexical variation (world knowledge)	Major correspondences between a question and the answer sentence require world knowledge to resolve.	Q: Which **governing bodies** have veto power? Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process.	9.1%
Syntactic variation	After a question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence after light pruning.	Q: What Shakespeare scholar **is currently on the faculty**? Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.	64.1%
Multiple sentence reasoning	There is a coreference or answering requires higher-level fusion of multiple sentences.	Q: What collection does **the V&A Theatre & Performance galleries** hold? Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance.	13.6%
Ambiguous	We don't agree with the crowdworkers' answer or the question does not have a unique answer.	Q: What is the main goal of criminal punishment? Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment.	6.1%

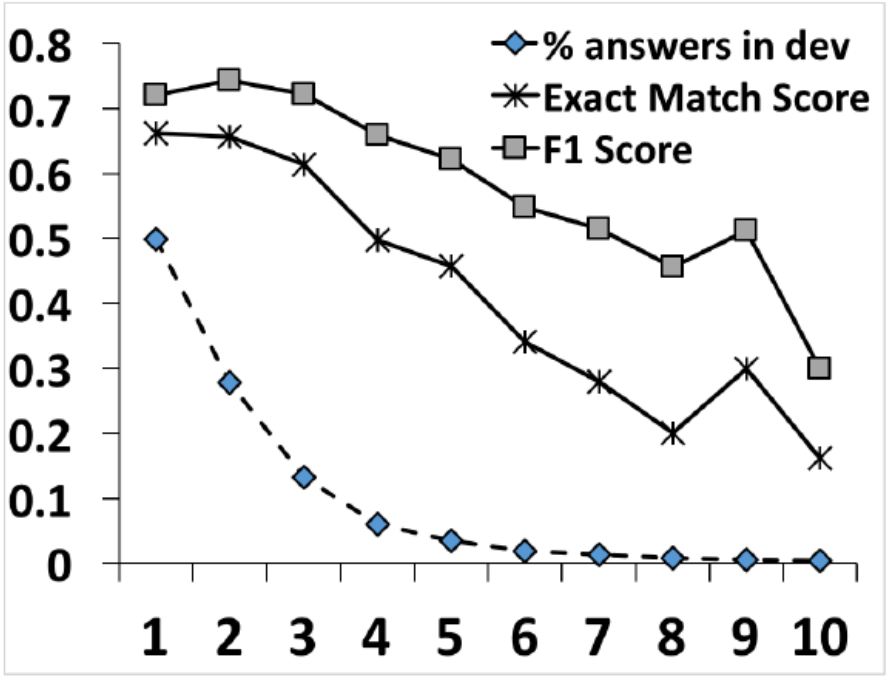
Dynamic Chunk Reader



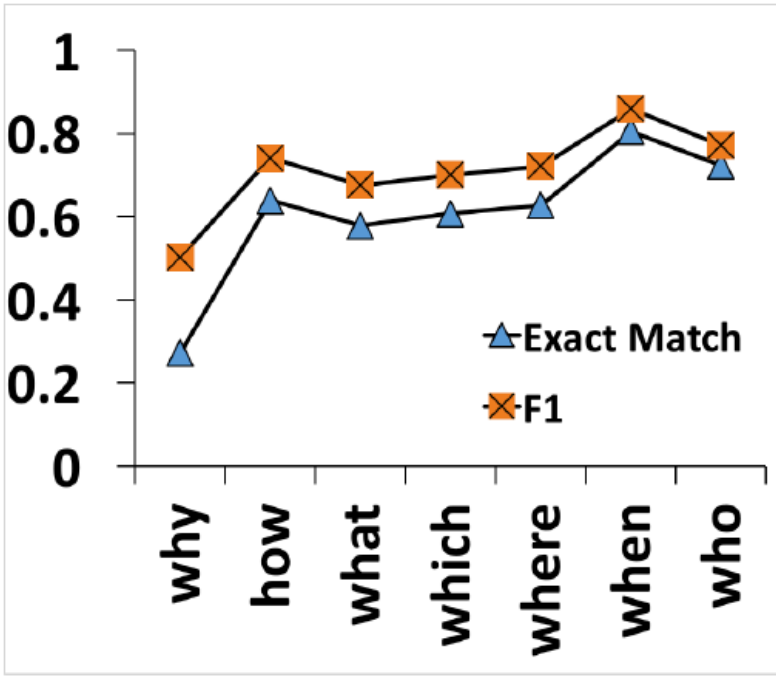
Leaderboard (as of Sept 15)

	Dev		Test	
	EM	F1	EM	F1
Logistic Regression (Stanford)	40.0	50.0		
MatchLSTM (SMU)	59.1	70.0	59.5	70.3
DCR (IBM)	62.5	71.2	62.5	71.0

Result Analysis

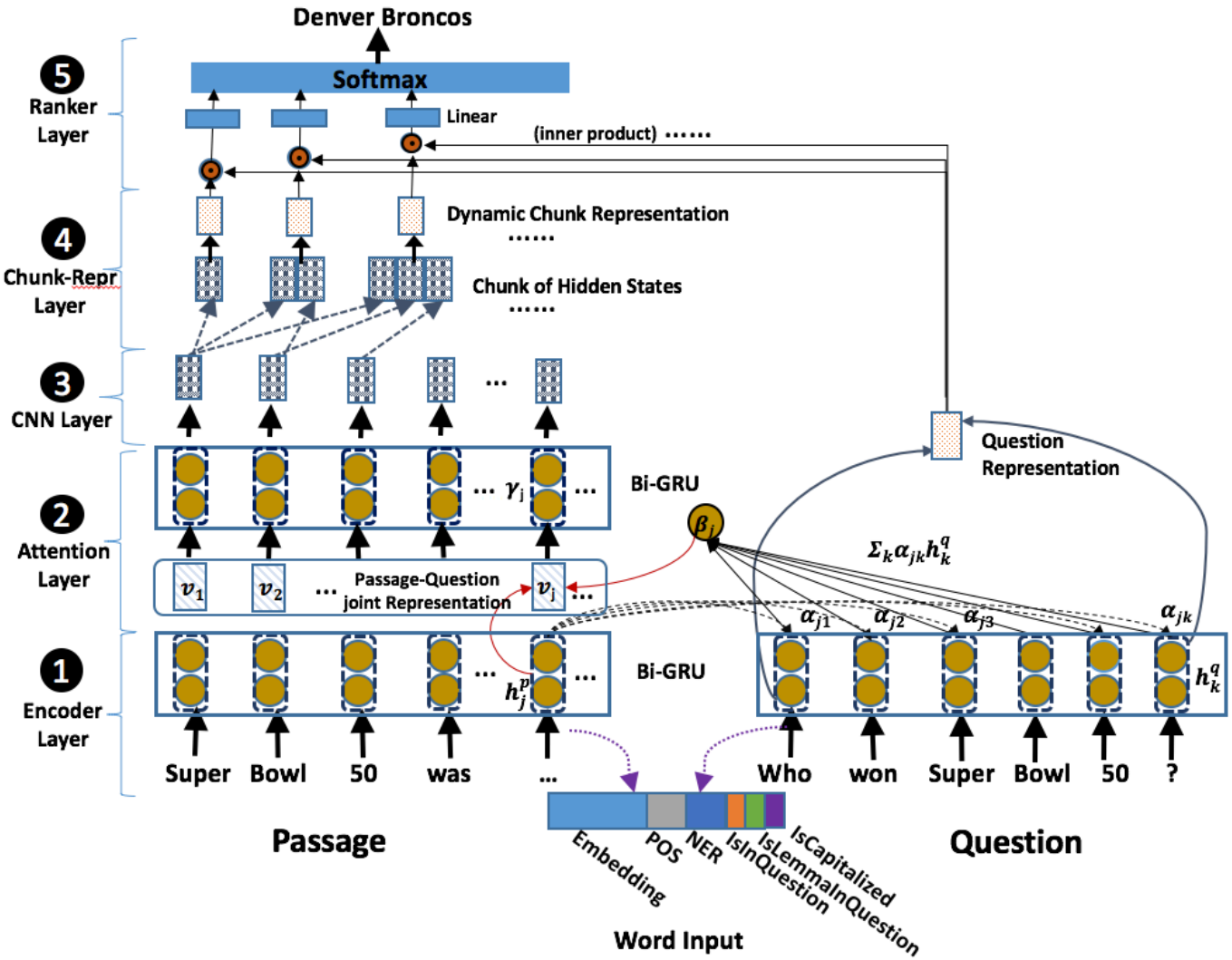


Variations of DCR performance on ground truth answer length (up to 10) in the development set.



Performance of the DCR across question types.

Dynamic Chunk Reader v2



Leaderboard (as of Nov. 14)

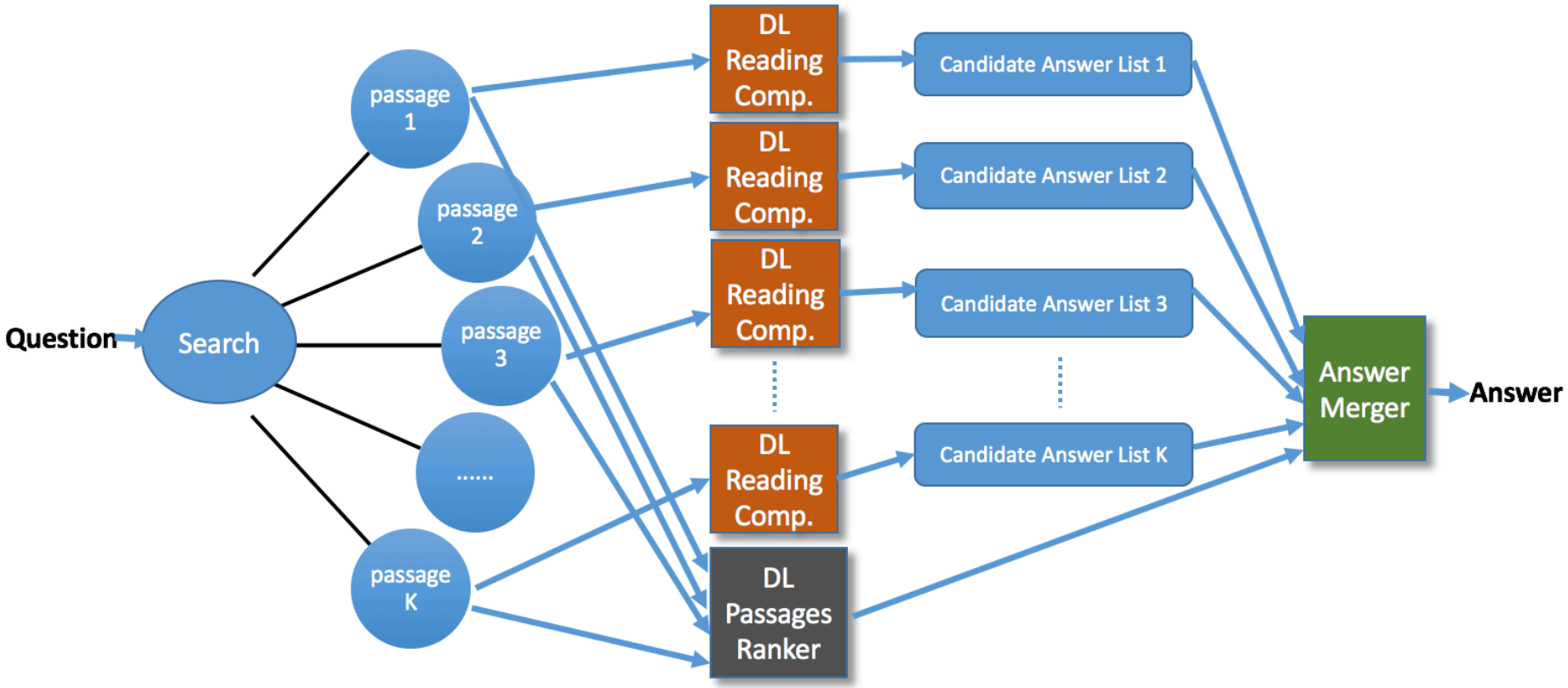
	Dev		Test	
	EM	F1	EM	F1
MatchLSTM (SMU)	64.1	73.9		
MatchLSTM ensemble (SMU)	67.6	76.8	67.9	77.0
Span Classifier (Google)	66.4	74.9		
Span Classifier ensemble (Google)	68.2	76.7		
Bi-attention (AI2)	64.0	74.5		
Bi-attention ensemble (AI2)	69.2	77.8	69.9	78.1
Co-attention (Salesforce)	65.4	75.6		
Co-attention ensemble (Salesforce)	70.3	79.4	71.2	80.4
DCR v2 (IBM)	63.6	72.5		
DCR v2 ensemble (IBM)	66.3	74.7		

Overview

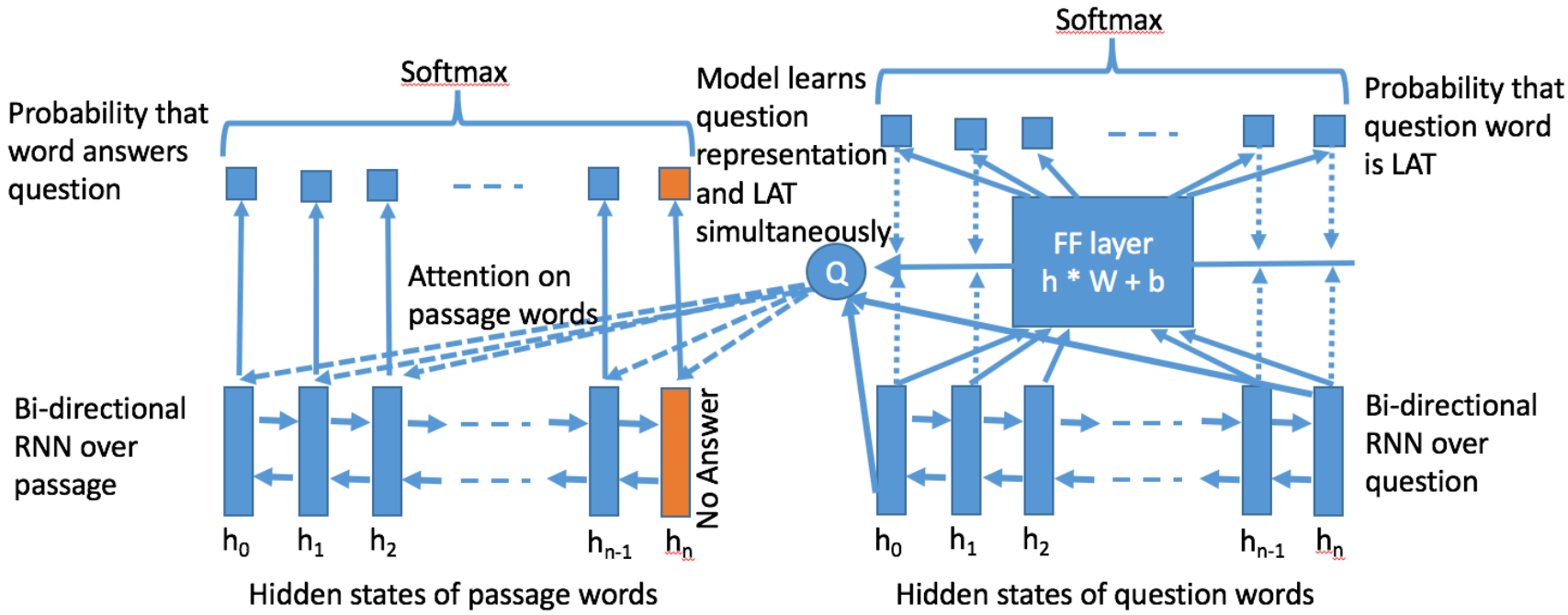
- RCQA as inference task in an ideal scenario
- RCQA in a real world single passage scenario
- **RCQA in a multi-passage scenario**
- Remaining Challenges

Multi-passage Factoid QA

In real world scenario, only question is provided.



RC on single document



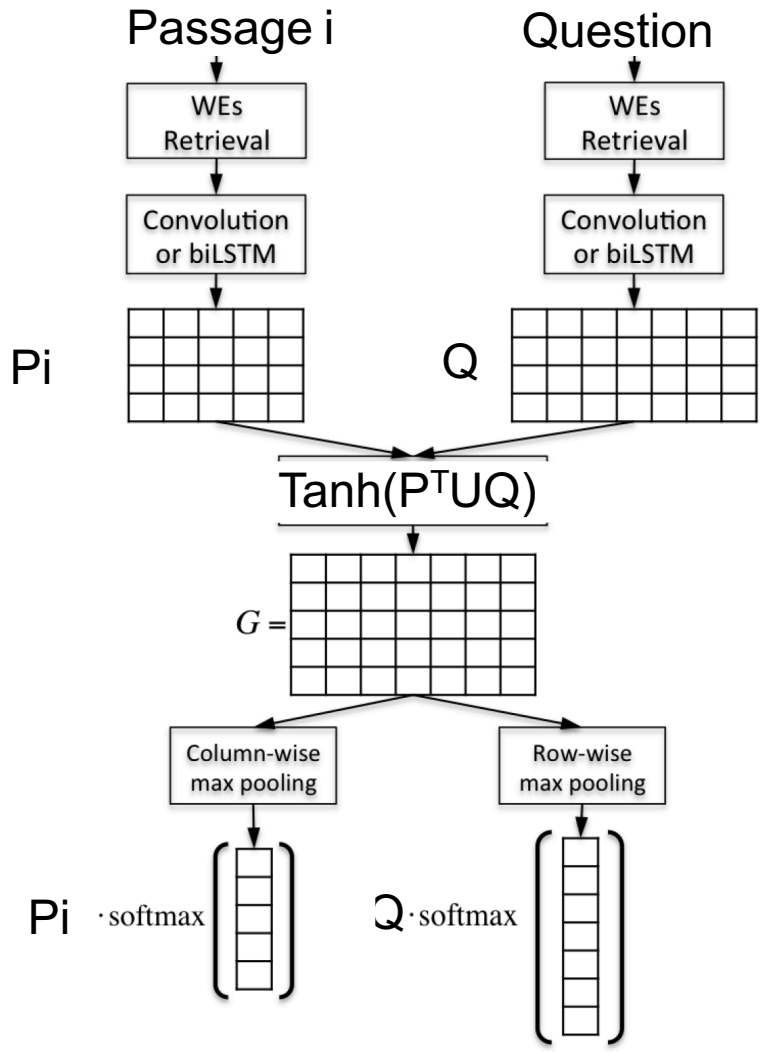
Retrieved Pasasge ranking with Attentive Pooling-Convolutional Neural Networks

Assumption: Each of the retrieved passage i is evaluated by its similarity to the question, using AP-CNN. The more similar the question is to the passage, the more relevant the passage is to the question, and in turn the more trust we put to the passage for generating the correct answer.

APCNN uses the two-way attention mechanism between passage words and question words for evaluating similarities. For passage i for the question Q, we first generate similarity score score(P_i). After all scores are generated, the scores are then normalized to a distribution with sharpening parameter .

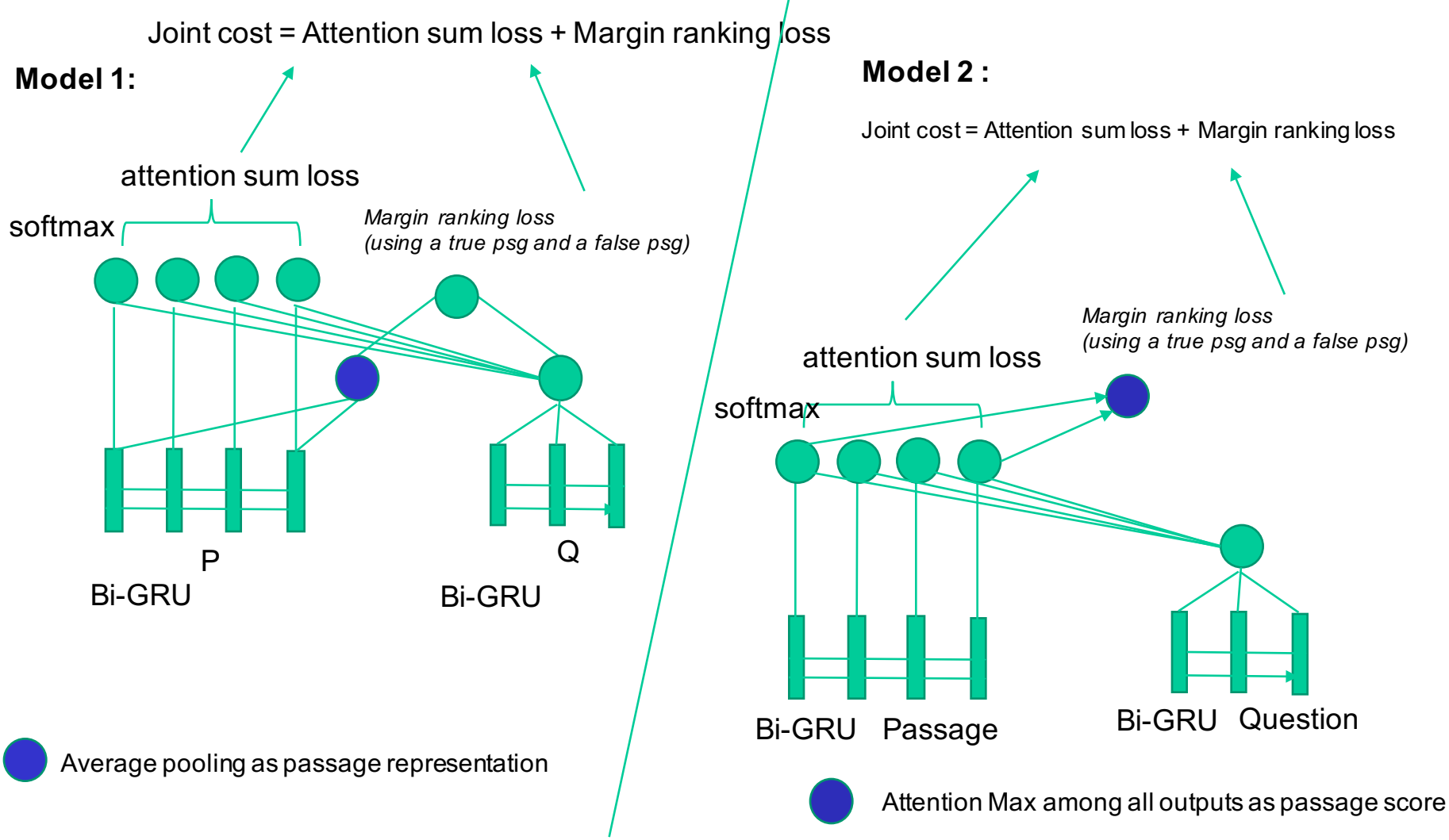
$$score(p_i) = similarity_{AP}(p_i, Q)$$

$$P(p_i|Q) = \frac{score(p_i)^\gamma}{\sum_{i=1}^{N} score(p_i)^\gamma}$$

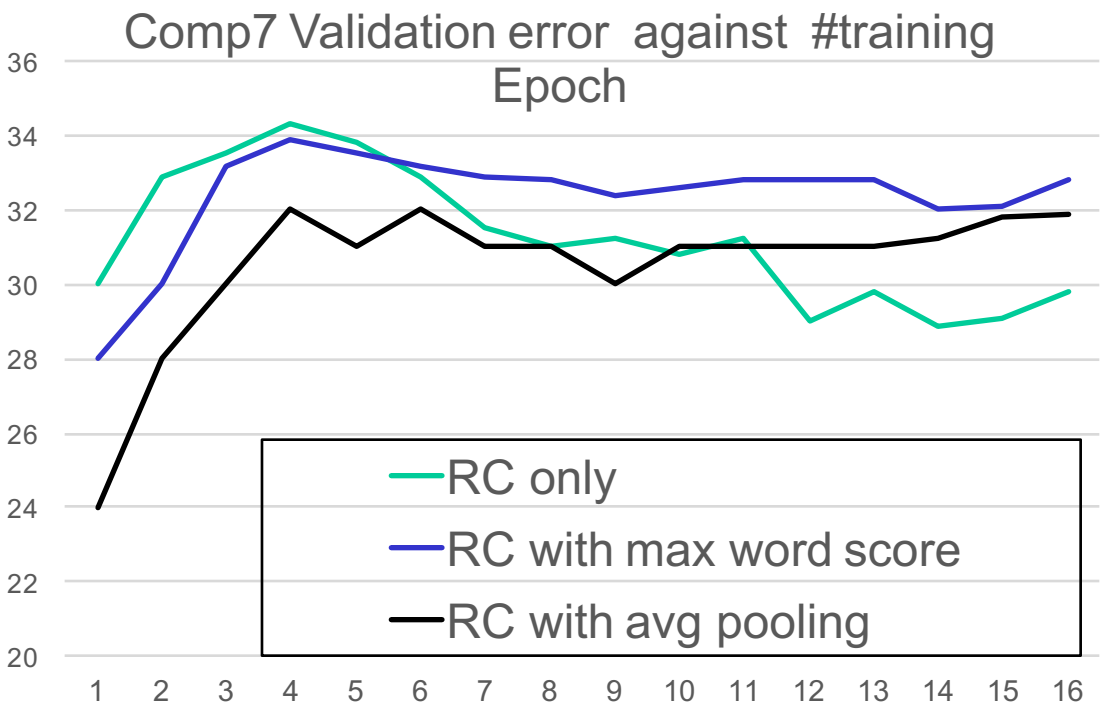


IBM Internal Real User Question Answering

Module combinations	Accuracy on validation
RC	31.75%
RC+ heuristics	33.09%
RC+APCNN	**35.85%**
RC+APCNN+entail	**36.32%**
RC+APCNN+DDQA passage scores	36.91%


Joint cost = Attention sum loss + Margin ranking loss
Model 1:
attention sum loss
softmax
Margin ranking loss
(using a true psg and a false psg)
P
Q
Bi-GRU
Bi-GRU
Average pooling as passage representation
Model 2 :
Joint cost = Attention sum loss + Margin ranking loss
Margin ranking loss
(using a true psg and a false psg)
attention sum loss
softmax
Bi-GRU Passage
Bi-GRU Question
Attention Max among all outputs as passage score

Joint Learning RC and Passage ranking



two strategies for passage score using only scores from bi-encoder:

1) max token score in passage
2) Avg pooling of scores from bi-encoder
3) Attentive Pooling scores from bi-encoder

The final scores are only evaluating RC part.

Max word score or avg. pooling scores can be used as passage scores.

Overview

- RCQA as inference task in an ideal scenario
- RCQA in a real world single passage scenario
- RCQA in a multi-passage scenario
- **Remaining Challenges**

Remaining Challenges

Learning representation

- new RNN/CNN models
- more labeled/unlabeled data
- dataset creation

Effective alignment between representations

At last,

Collaboration is Welcome,

And, we are hiring

internship/full-time candidates

Contact Sasha or us, if interested.

Topics we care about:

Innovative NN models, reinforcement Learning for NLP, machine reading comprehension, answer generation, knowledge graph based QA, semantic relatedness

Questions

Reference

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. ICLR , 2015.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems , pp. 1693–1701, 2015.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children's books with explicit memory representations. arXiv preprint arXiv:1511.02301 , 2015.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. Text understanding with the attention sum reader network. ACL , 2016.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 , 2016.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In EMNLP , volume 3, pp. 4, 2013.

Dynamic Chunk Reader v1