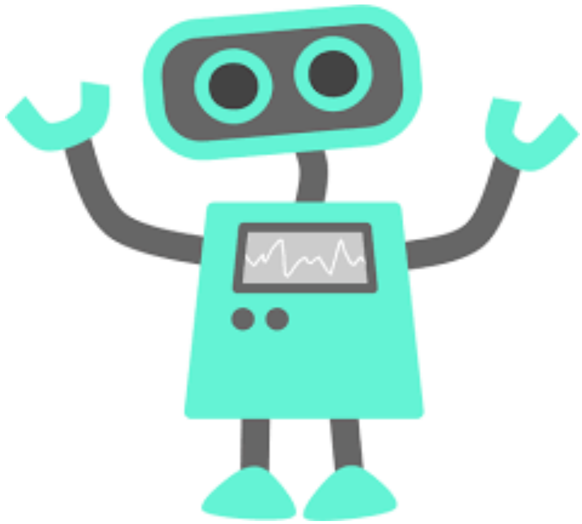
Interpretability and Robustness for Multi-Hop QA

Mohit Bansal

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

(MRQA-EMNLP 2019 Workshop)

Multihop-QA's Diverse Requirements



- Interpretability and Modularity
- Multiple Reasoning Chains Assembling
- Adversarial Shortcut Robustness
- Scalability and Data Augmentation
- Commonsense/External Knowledge

Outline

- **Interpretability & Modularity for MultihopQA:**
    - Neural Modular Networks for MultihopQA
    - Reasoning Tree Prediction for MultihopQA

- **Robustness to Adversaries and Unseen Scenarios for QA/Dialogue:**
    - Adversarial Evaluation and Training to avoid Reasoning Shortcuts in MultihopQA
    - Robustness to Over-Sensitivity and Over-Stability Perturbations
    - Auto-Augment Adversary Generation
    - Robustness to Question Diversity via Question Generation based QA-Augmentation
    - Robustness to Missing Commonsense/External Knowledge

- Thoughts/Challenges/Future Work

Interpretability and Modularity

Single-Hop QA

[Rajpurkar et al., 2016]

Question

"Which NFL team represented the AFC at Super Bowl 50?"

Answer

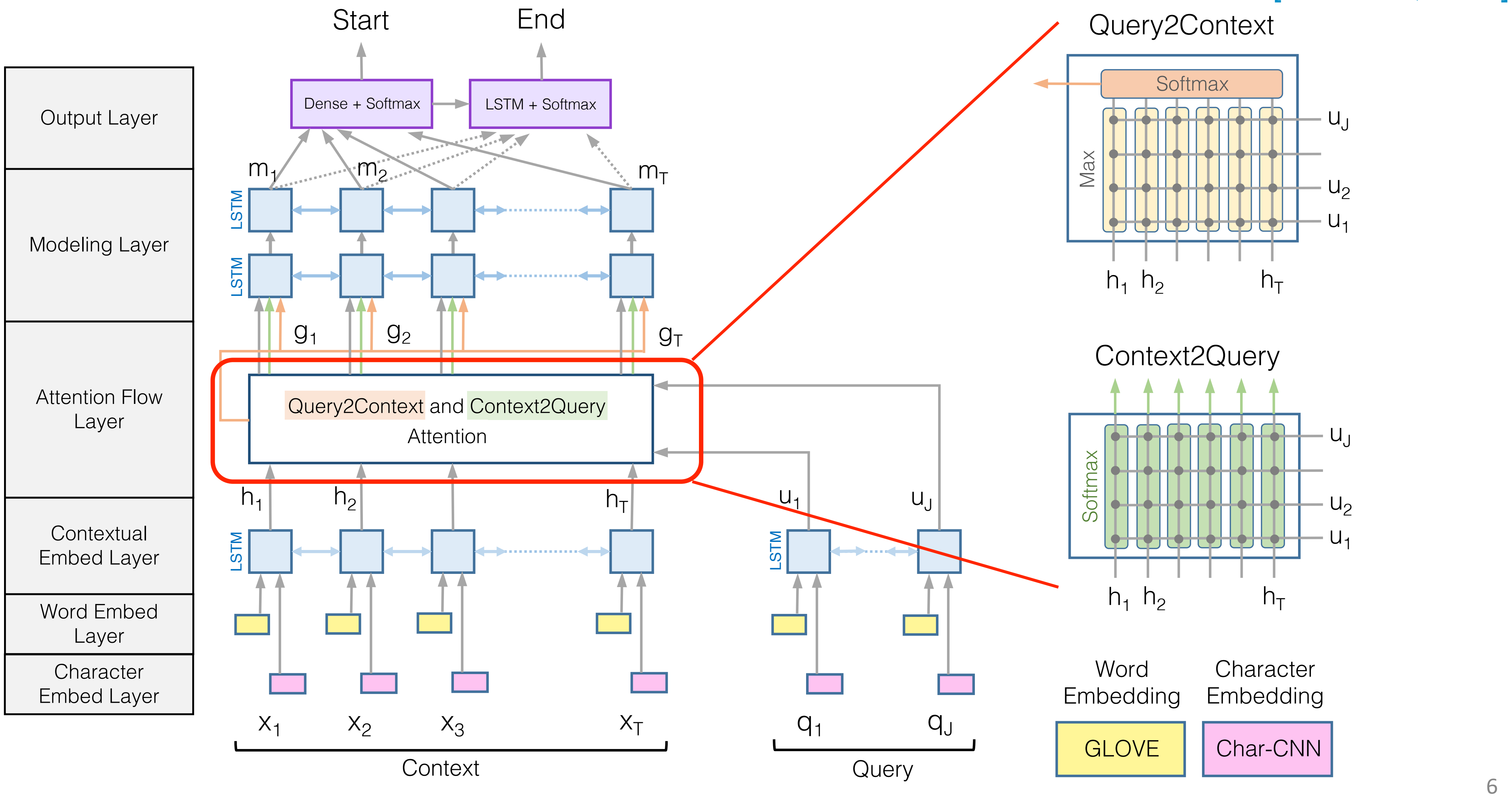
"Denver Broncos"

Context

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion <u>Denver Broncos</u> defeated the National Football Conference (NFC) champion Carolina Panthers …

Bi-directional Attention Flow Model (BiDAF)

[Seo et al., 2017]



Multi-Hop QA: Bridge-Type

[Yang et al., 2018]

Question

"What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?"

Context

Kasper Schmeichel is a Danish professional footballer ... He is the son of former Manchester United and Danish international goalkeeper **Peter Schmeichel**.

Peter Bolesław Schmeichel is a Danish former professional footballer … was voted the IFFHS World's Best Goalkeeper in 1992 …

Kasper Schmeichel $\xrightarrow{son\_of}$ Peter Schmeichel $\xrightarrow{voted\_as}$ $\boxed{\text{World's Best Goalkeeper}}$

Bridge Entity

Multi-Hop QA: Comparison-Type

[Yang et al., 2018]

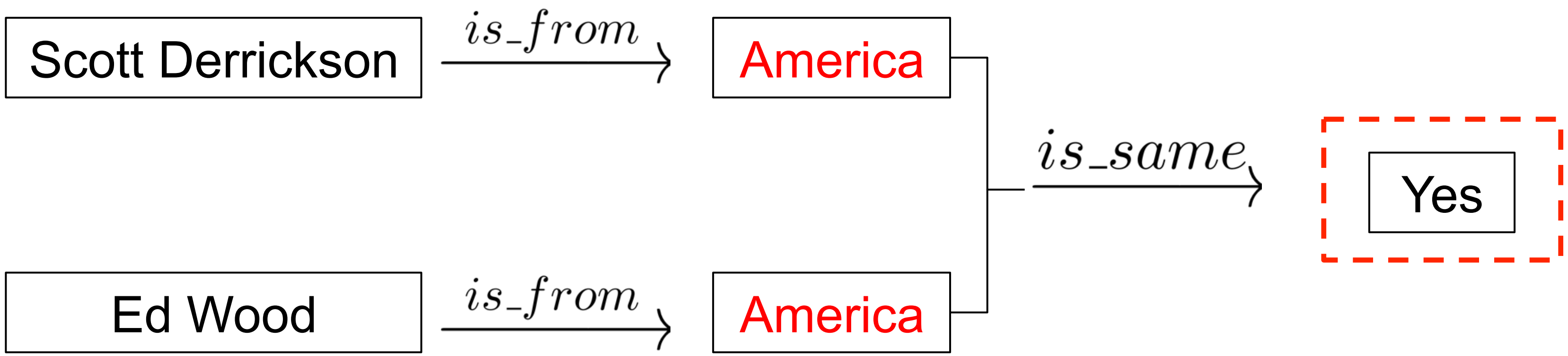
Question

"Were Scott Derrickson and Ed Wood of the same nationality?"

Context

Scott Derrickson is an **American** director ...

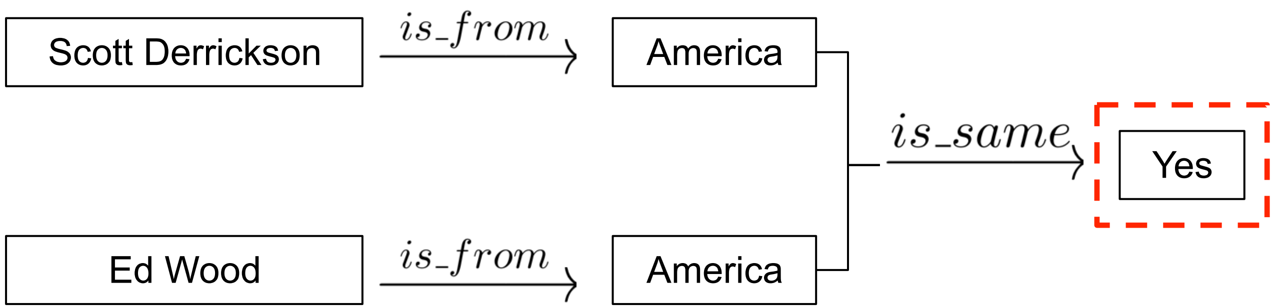
Edward Wood Jr. was an **American** filmmaker ...



Challenges: Different Reasoning Chains in Multi-Hop QA

"What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?"

"Were Scott Derrickson and Ed Wood of the same nationality?"



(1) Self-Assembling Neural Modular Networks

What we want:

A modular network dynamically constructed according to different question types.

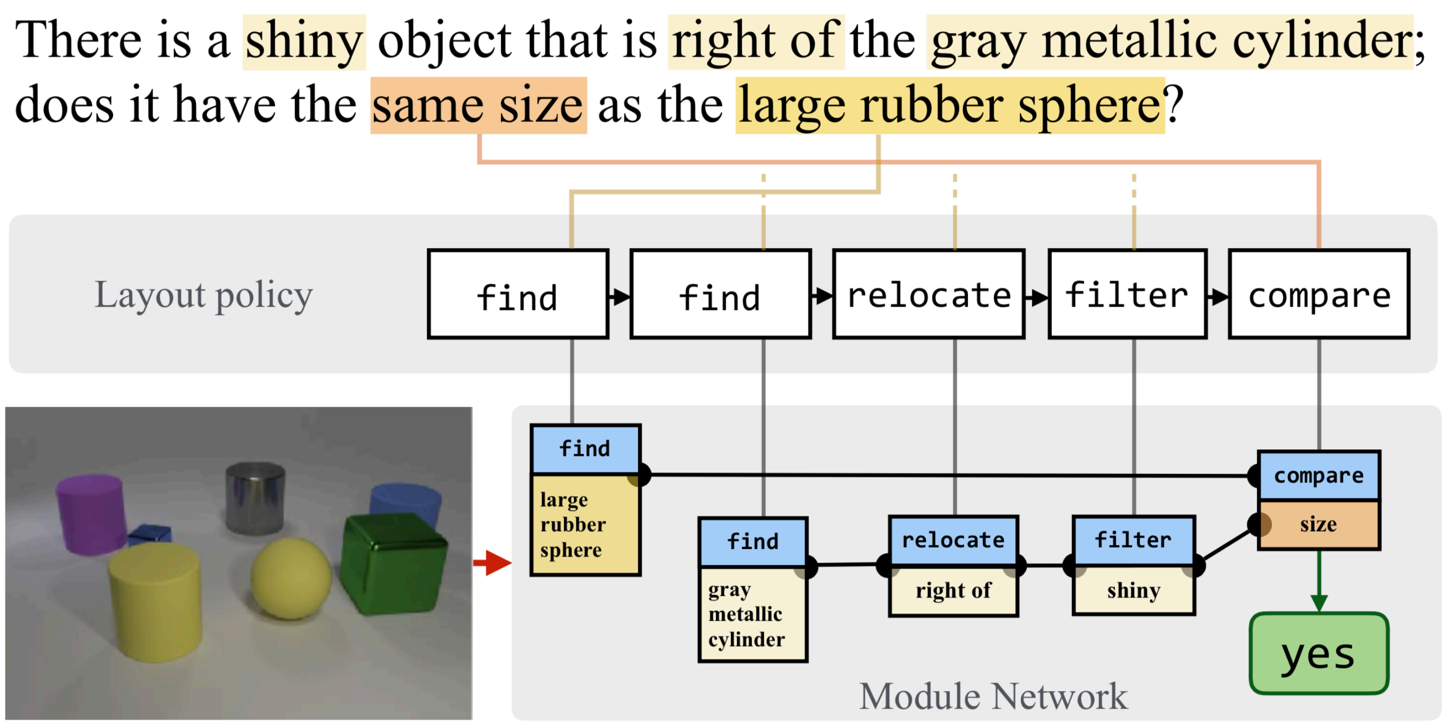
To achieve this, we need:

- A number of modules, each designed for a unique type of single-hop reasoning.
- A controller to
  - decompose the multi-hop question to multiple single-hop sub-questions,
  - design the network layout **based on the question** (decides which module to use for each sub-question).

[Jiang and Bansal, EMNLP 2019]

Neural Modular Networks

Neural Modular Network was originally proposed to solve Visual Question Answering (VQA), including VQA dataset and CLEVR dataset (Andreas et al. 2016, Hu et al. 2017).



[Jiang and Bansal, EMNLP 2019]

Controller RNN

The original NMN controllers are usually trained with RL. Hu et al. (2018) proposed stack-based NMN w/ soft module execution to avoid indifferentiability in optimization

-Average over the outputs of all modules at every step instead of sample a single module at every step.

-Modules at different timestep communicate by popping/pushing the averaged attention output from/onto a stack.

- **Inputs:**
  - Question emb: u
  - Decoding timestep: t
- **Intermediate:**
  - Distribution over question words: cv_t (softly decompose the question)
- **Outputs:**
  - Module probability: p (Which module should be used at step t)
  - Sub-question vector: c_t (What sub-question to solve at step t)

[Jiang and Bansal, EMNLP 2019]

Reasoning Modules

Inputs: Question emb: u, Sub-question vector: c_t, Context emb: h

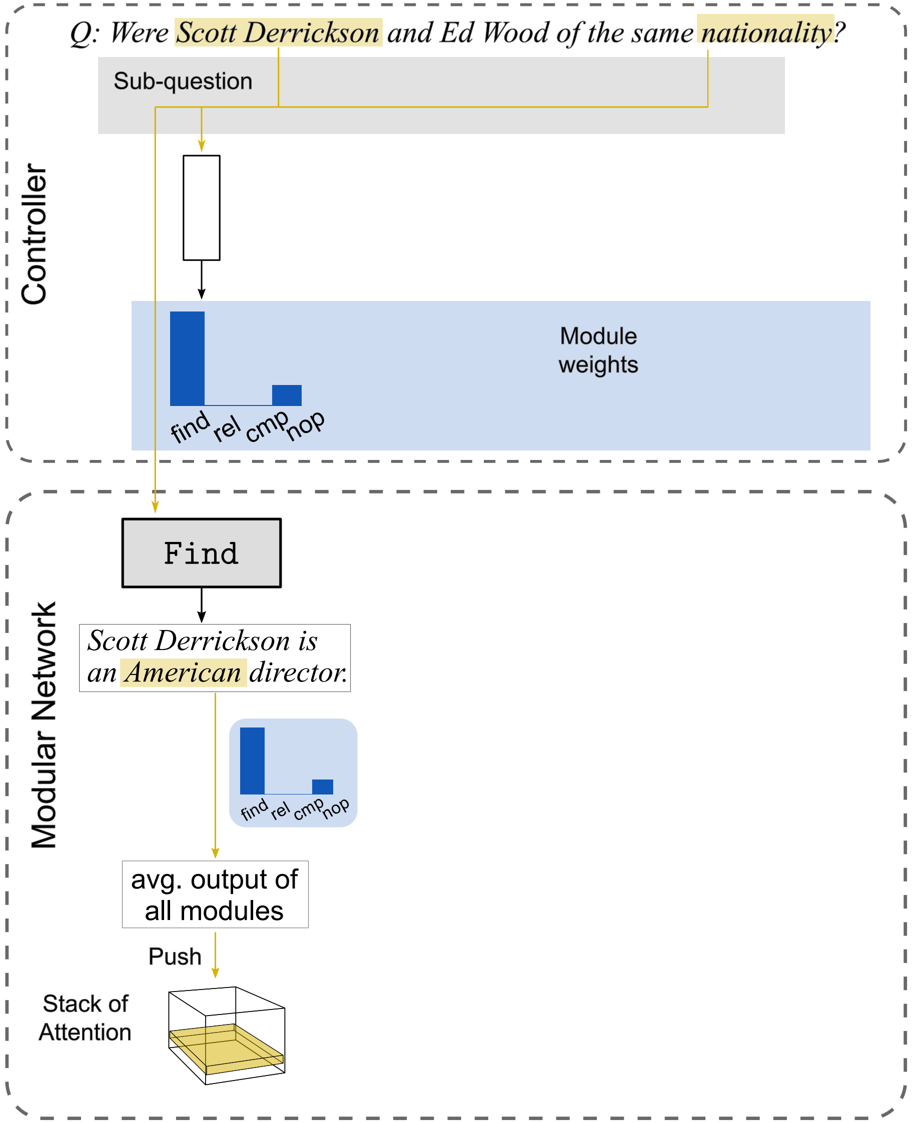
Module Name	Input Attention	Output Types	Implementation Details
`Find`(u, c, h)	(None)	Attention	$\mathrm{BiAttn}(h \odot c_t, u)$
`Relocate`(u, c, h)	a1	Attention	$\mathtt{Find}(u, c_t, h \odot (a_1 \cdot h))$
`Compare`(u, c, h)	a1, a2	Yes/No	$\sigma(\mathrm{MLP}([c_t, a_1 \cdot h, a_2 \cdot h, c_t \cdot (a_1 - a_2) \cdot h]))$
`NoOp`(u, c, h)	(None)	(None)	(None)

[Jiang and Bansal, EMNLP 2019]

Putting an NMN together...

Controller:

Modules:

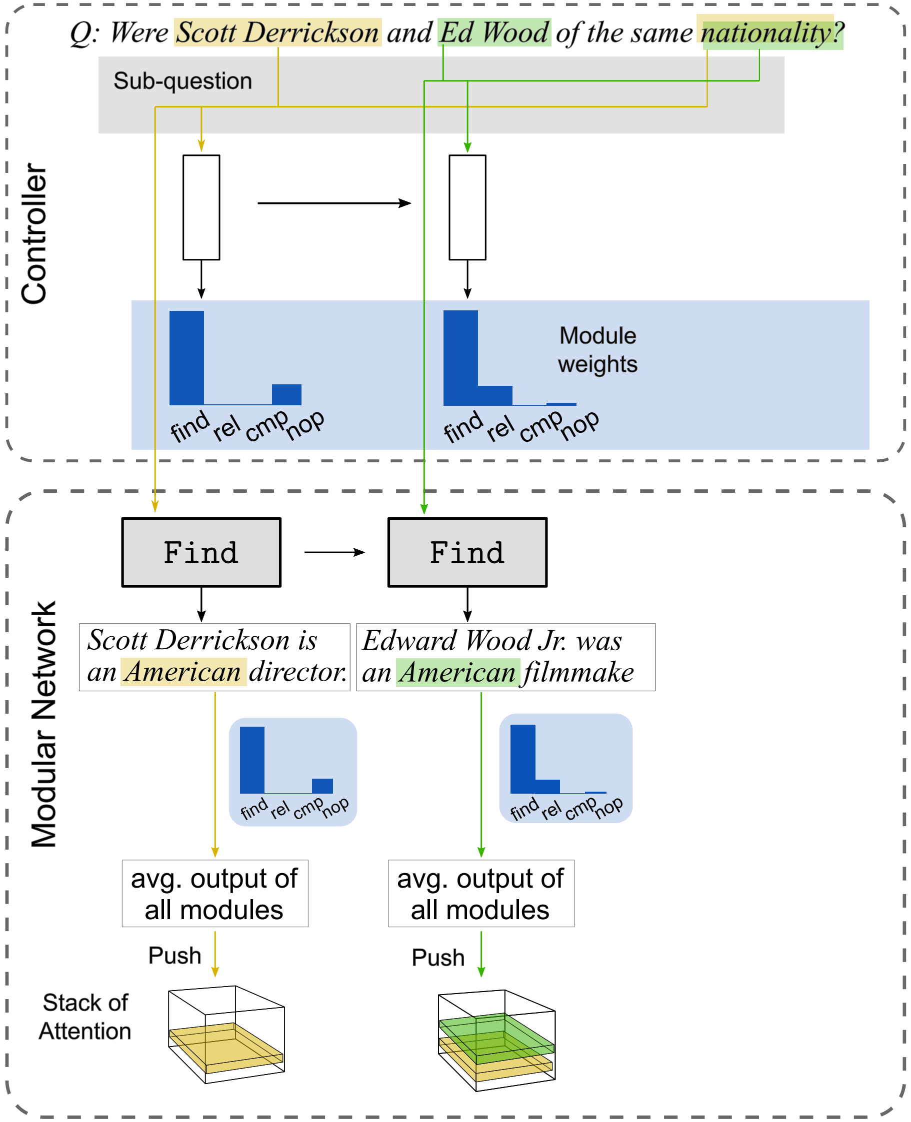


[Jiang and Bansal, EMNLP 2019]

Putting an NMN together...

Controller:

Modules:

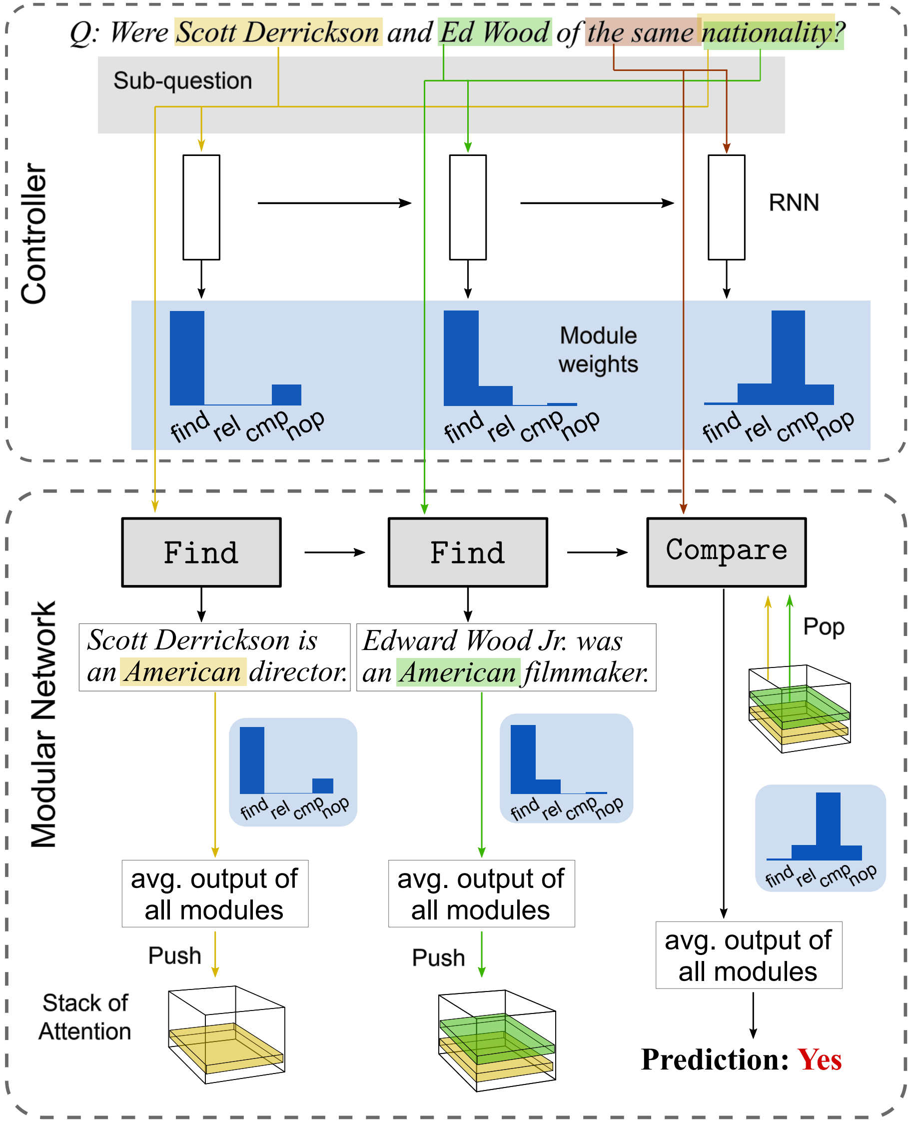


[Jiang and Bansal, EMNLP 2019]

Putting an NMN together...

Controller:

Modules:



[Jiang and Bansal, EMNLP 2019]

Main Results on HotpotQA

	Dev	Test
	F1	F1
BiDAF Baseline	57.19	55.81
Original NMN	40.28	39.90
Our NMN	63.35	62.71

[Jiang and Bansal, EMNLP 2019]

Ablation Studies

	Bridge F1	Comparison F1
Our NMN	64.49	57.20
-Relocate	60.13	58.10
-Compare	64.46	56.00

*All models are evaluated on our dev set.

[Jiang and Bansal, EMNLP 2019]

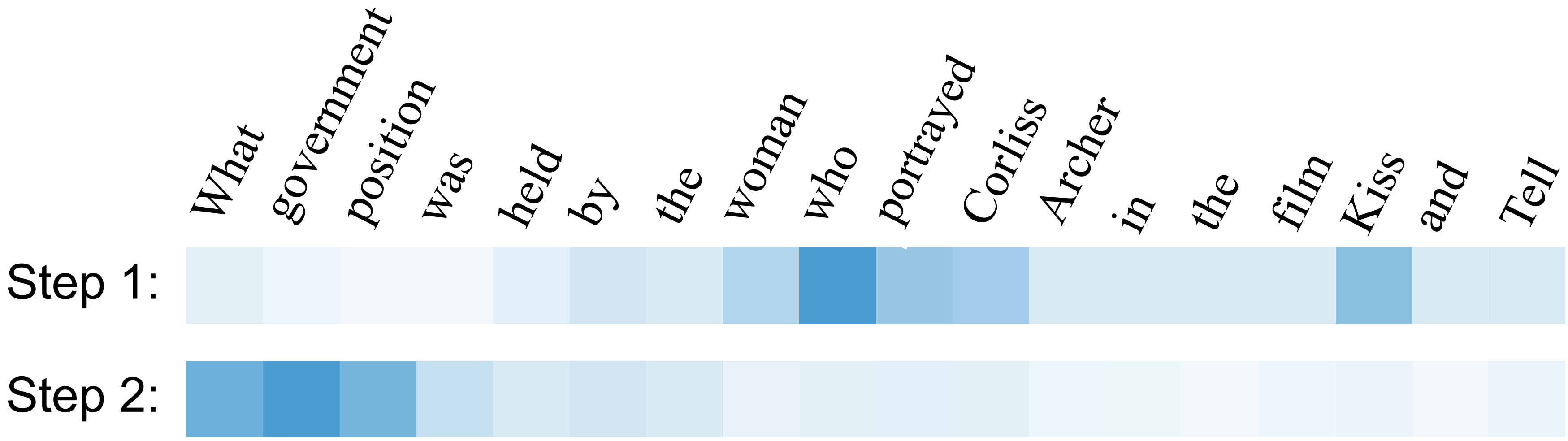
Adversarial Evaluation

Train	Reg	Reg	Adv	Adv
Eval	Reg	Adv	Reg	Adv
BiDAF Baseline	43.12	34.00	45.12	44.65
Our NMN	**50.13**	**44.70**	**49.33**	**49.25**

Table 4: EM scores after training on the regular data or on the adversarial data from Jiang and Bansal (2019), and evaluation on the regular dev set or the adv-dev set.

[Jiang and Bansal, EMNLP 2019]

Analysis: Controller Attention Visualization



Step 1: Kiss and Tell is a 1945 American comedy film starring then 17-year-old Shirley Temple as Corliss Archer. ...

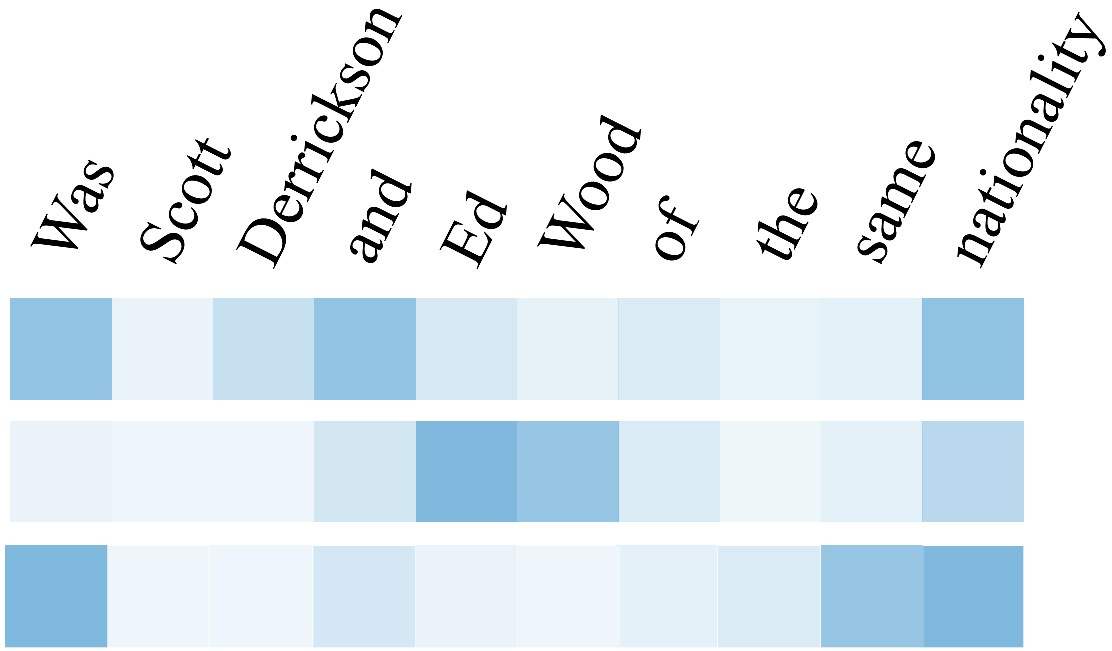
Step 2: Shirley Temple Black was an American actress, ..., and also served as Chief of Protocol of the United States.

- We also have initial human evaluation results on controller's sub-question soft decomposition/attention.

[Jiang and Bansal, EMNLP 2019]

Analysis: Controller Attention for Comparison Questions

UNC NLP



Ctrl Step 1:

Ctrl Step 2:

Ctrl Step 3:

Mod. Step 1: Scott Derrickson is an American director. ...

Mod. Step 2: Edward Wood Jr. was an American filmmaker. ...

Mod. Step 3: Yes

[Jiang and Bansal, EMNLP 2019]

Analysis: Evaluating Module Layout Prediction

Bridge:

"What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?"

`Find -> Relocate`: 99.9%

Comparison Yes/No:

"Were Scott Derrickson and Ed Wood of the same nationality?"

`Find -> Find -> Compare`: 4.8 %

`Find -> Relocate -> Compare`: 63.8%

[Jiang and Bansal, EMNLP 2019]

Recent Results with BERT

- BERT+NMN achieves >= results as Fine-tuned BERT-base (71.26 vs 70.66 F1).
- Module Layout Prediction results improved (compared to the non-BERT NMN):
- Hence, BERT+NMN model allows for stronger interpretability than non-modular BERT models (& non-BERT NMNs), but while maintaining BERT-style numbers.

Bridge-Type:	"What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?"	**`Find -> Relocate:`** 99.9%
Comparison Yes/No:	"Were Scott Derrickson and Ed Wood of the same nationality?"	**`Find -> Find -> Compare:`** ~~4.8 %~~ **96.9%** **`Find -> Relocate -> Compare:`** ~~63.8%~~ **0%**

[Jiang and Bansal, EMNLP 2019]

Recent Results with BERT

- BERT+NMN achieves >= results as Fine-tuned BERT-base (71.26 vs 70.66 F1).
- Module Layout Prediction results improved (compared to the non-BERT NMN):
- Hence, BERT+NMN model allows for stronger interpretability than non-modular BERT models (& non-BERT NMNs), but while maintaining BERT-style numbers.

Bridge-Type: "What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?" **`Find -> Relocate`**: 99.9%

Still several challenges/ long way to go, e.g., more complex MultihopQA datasets with more hops, more types of reasoning behaviors, etc.!

Comparison Yes/No: "Were Scott D and Ed Wood of the same nationality?" **`Find -> Relocate -> Compare`**:

See Yichen's full talk on Nov7 10.30am!

[Jiang and Bansal, EMNLP 2019]

(2) Divergent Reasoning Chains

[Welbl et al. 2018]

The *Polsterberg Pumphouse* (German : Polsterberger Hubhaus) is a pumping station above **the Dyke Ditch** in the **Upper Harz** in central Germany ...

The Dyke Ditch is the longest artificial ditch in the **Upper Harz** in central Germany.

The **Upper Harz** refers to ... the term Upper Harz covers the area of the seven historical mining towns (\"Bergst\u00e4dte\") - Clausthal, Zellerfeld, Andreasberg, Altenau, Lautenthal, Wildemann and Grund - in the present-day German federal state of **Lower Saxony**.

Query subject: *Polsterberg Pumphouse*
Query body: located_in_the_administrative_territorial_entity
Answer: Lower Saxony

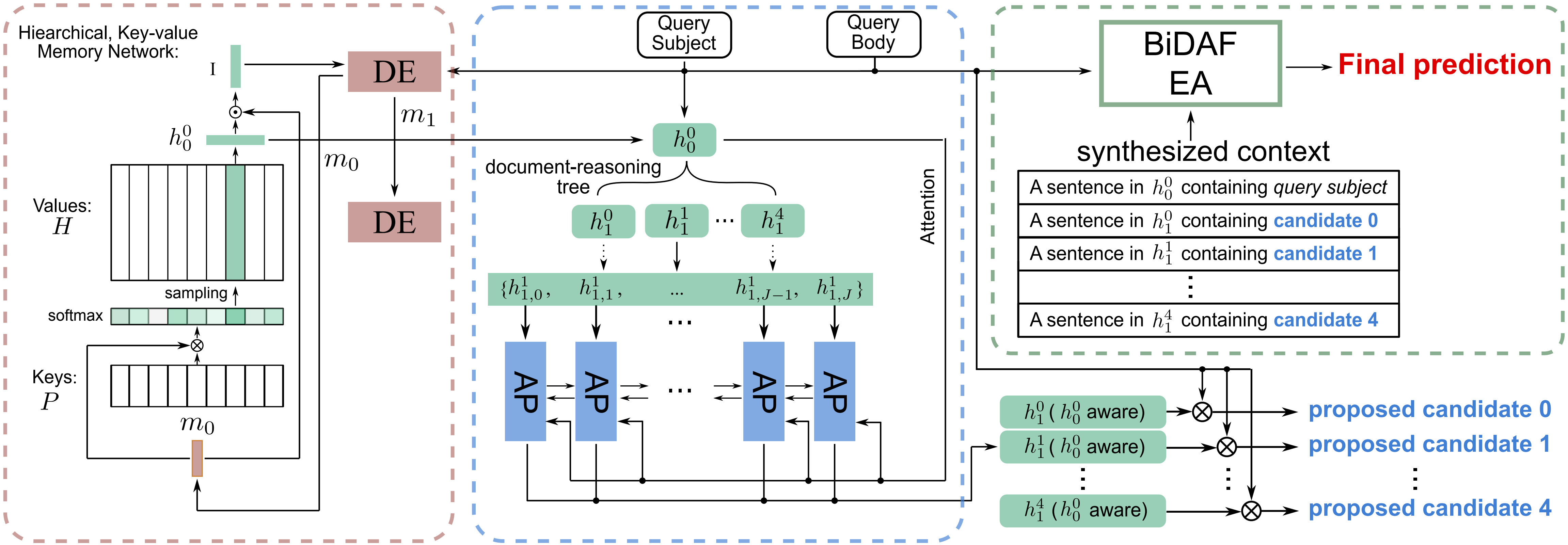
[Jiang, Joshi, Chen, Bansal, ACL 2019a]

Multi-Hop QA Requirements

- Success on Multi-Hop Reasoning QA requires a model to:
  - Locate a reasoning chain of important/relevant documents from a large pool of documents
  - Consider evidence loosely distributed in all documents from a reasoning chain to predict the answer
  - Weigh and merge evidence from **MULTIPLE** reasoning chains to predict the answer

[Jiang, Joshi, Chen, Bansal, ACL 2019a]

EPAr: Explore-Propose-Assemble reader



Document Explorer (DE):
Iteratively selects relevant documents and represents multiple reasoning chains in a tree structure

Answer Proposer (AP):
Proposes a candidate answer from every ancestor-aware root-to-leaf chain in the reasoning tree

Evidence Assembler (EA):
Extracts key sentences from every reasoning chain and combines them to make a unified prediction

[Jiang, Joshi, Chen, Bansal, ACL 2019a]

Results - WikiHop and MedHop

	Dev	Test
BiDAF Welbl et al., 2017*	-	42.9
Coref-GRU (Dhingra et al., 2018)	56.0	59.3
WEAVER (Raison et al., 2018)	64.1	65.3
MHQA-GRN (Song et al., 2018)	62.8	65.4
Entity-GCN (De Cao et al., 2018)	64.8	67.6
BAG (Cao et al., 2019)	66.5	69.0
CFC (Zhong et al., 2019)	66.4	70.6
EPAr (Ours)	**67.2**	69.1

WikiHop

	Test (Masked)	Test
FastQA (Weissenborn et al., 2017)	23.1	31.3
BiDAF (Seo et al., 2017)	33.7	47.8
CoAttention	-	58.1
Most Frequent Candidate	10.4	58.4
EPAr (Ours)	**41.6**	**60.3**

MedHop

[Jiang, Joshi, Chen, Bansal, ACL 2019a]

Human Evaluation: Quality of Reasoning Tree

UNC NLP

- Recall-k score is the % of examples where one of the human-annotated reasoning chains is recovered in the top-k root-to-leaf paths in the reasoning tree

	R@1	R@2	R@3	R@4	R@5
Random	11.2	17.3	27.6	40.8	50.0
1-hop TFIDF	32.7	48.0	56.1	63.3	70.4
2-hop TFIDF	42.9	56.1	70.4	78.6	82.7
DE	38.8	50.0	65.3	73.5	83.7
TFIDF+DE	**44.9**	**64.3**	**77.6**	**82.7**	**90.8**

- 2-hop TF-IDF performs much better than simple 1-hop TF-IDF retrieval
- DE without any TF-IDF retrieval pre-processing performs worse than 2-hop TF-IDF
- Combination of TF-IDF retrieval and DE performs better than each one of them alone

[Jiang, Joshi, Chen, Bansal, ACL 2019a]

Human Evaluation: Quality of Reasoning Tree

- Recall-k score is the % of examples where one of the human-annotated reasoning chains is recovered in the top-k root-to-leaf paths in the reasoning tree

	R@1	R@2	R@3	R@4	R@5
Random	11.2	17.3	27.6	40.8	50.0
1-hop TFIDF	32.7	48.0	56.1	63.3	70.4
2-hop TFIDF	42.9	56.1	70.4	78.6	82.7
DE					
TFIDF+DE					

Still several challenges/ long way to go, e.g., more complex MultihopQA datasets with more hops, longer and more #reasoning chains, etc.!

- 2-hop TF-IDF performs much better than simple 1-hop TF-IDF retrieval
- DE without any TF-IDF retrieval pre-processing performs worse than 2-hop TF-IDF
- Combination of TF-IDF retrieval and DE performs better than each one of them alone

[Jiang, Joshi, Chen, Bansal, ACL 2019a]

Adversarial Robustness

Is *compositional reasoning* necessary to answer these multi-hop questions?

Not always!

[Jiang and Bansal, ACL 2019]

Reasoning Shortcut

Question

"What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?"

Reasoning Chain:

Kasper Schmeichel	$\xrightarrow{son\_of}$	Peter Schmeichel	$\xrightarrow{voted\_as}$	World's Best Goalkeeper
Question Entity		*Bridge Entity*		*Answer*

Reasoning Shortcut:

[Placeholder]	$\xrightarrow{voted\_as}$	World's Best Goalkeeper
		Answer

[Jiang and Bansal, ACL 2019]

Reasoning Shortcut

Question

"What was the father of Kasper Schmeichel **voted to be by the IFFHS in 1992**?"

The answer can be directly inferred by word-matching the documents to maximum of the question !!!

Context

Peter Bolesław Schmeichel is a Danish former professional footballer .., and was **voted** the **IFFHS** World's Best Goalkeeper **in 1992** and 1993.

Edson Arantes do Nascimento is a retired Brazilian professional footballer. In 1999, he was **voted** World Player of the Century by **IFFHS**. [Missing: 1992]

Kasper Hvidt is a Danish retired handball goalkeeper, .. also **voted** as Goalkeeper of the Year March 20, 2009, [Missing: 1992, IFFHS]

[Jiang and Bansal, ACL 2019]

Min et al., 2019; Chen & Durrett, 2019

How to eliminate this reasoning shortcut from the data to **ENFORCE** compositional reasoning?

Building **adversarial documents** as better distractors

[Jiang and Bansal, ACL 2019]

Adversarial Document

Question

"What was the father of Kasper Schmeichel **voted to be by the IFFHS in 1992**?"

Context

Peter Bolesław Schmeichel is a Danish former professional footballer .., and was **voted** the **IFFHS** World's Best Goalkeeper **in 1992** and 1993.

Adversarial Document

R. Kelly Schmeichel is a Danish former professional footballer .., and was **voted** the **IFFHS** World's Best Defender **in 1992** and 1993.

A model exploiting the reasoning shortcut will now find two plausible answers!

[Jiang and Bansal, ACL 2019]

BERT (Document Retrieval Results)

* Exact-Match scores between 2 golden documents and 2 retrieved documents

Train \ Eval	Eval = Regular	Eval = Adv
Train = Regular	89.44	44.67
Train = Adv	89.03	80.14

- The performance of the BERT retrieval model trained on the regular training set **dropped** a lot when evaluated on the adversarial data.
- BERT is actually exploiting the reasoning shortcut instead of performing multi-hop reasoning.

[Jiang and Bansal, ACL 2019]

BERT (Document Retrieval Results)

* Exact-Match scores between 2 golden documents and 2 retrieved documents

Train \ Eval	Eval = Regular	Eval = Adv
Train = Regular	89.44	44.67
Train = Adv	89.03	80.14

- After being trained on the adversarial data, BERT achieves significantly higher EM score in adversarial evaluation.
- Adversarial training is able to teach the model to be aware of distractors and force it not to take the reasoning shortcut, but there is still a remaining drop in performance.

[Jiang and Bansal, ACL 2019]

Bi-attention + Self-attention Baseline

* Exact-Match scores

Train \ Eval	Eval = Regular	Eval = Adv
Train = Regular	43.12	34.00
Train = Adv	45.12	44.65

- The performance of the baseline trained on the regular training set **dropped** a lot when evaluated on the adversarial data.
- The model that performs well in the original data is actually exploiting the reasoning shortcut instead of performing multi-hop reasoning.

[Jiang and Bansal, ACL 2019]

Bi-attention + Self-attention Baseline

* Exact-Match scores

Train \ Eval	Eval = Regular	Eval = Adv
Train = Regular	43.12	34.00
Train = Adv	45.12	44.65

- After being trained on the adversarial data, the baseline achieves significantly higher EM score in adversarial evaluation.
- Adversarial training is able to teach the model a bit to be aware of distractors and force it not to take the reasoning shortcut, but still big room for improvement.

[Jiang and Bansal, ACL 2019]

Analysis

- Manual Verification of Adversaries
  - 0 out of 50 examples had contradictory answers

- Model Error (Adversary Success) Analysis
  - In 96.3% of the failures, the model's prediction spans at least one of the adversarial documents

- Adversary Failure Analysis
  - Sometimes the reasoning shortcut still exists after adversarial documents are added

- **Next Steps/Questions:**
  - We might have made the model robust to one kind of attack but there might be others?
  - How do we ensure robustness to other adversaries we haven't thought of?

[Jiang and Bansal, ACL 2019]

Auto-Augment Adversary Generation

Ribeiro et al., 2018; Zhao et al., 2018

How do we automatically generate the best adversaries without manual design? Our AutoAugment model consists of a controller and a target model. The controller first samples a policy that transforms the original data to augmented data, on which the target model retrains. After training, the target model is evaluated to obtain the performance on the validation set. This performance is then fed back to the controller as the reward signal.

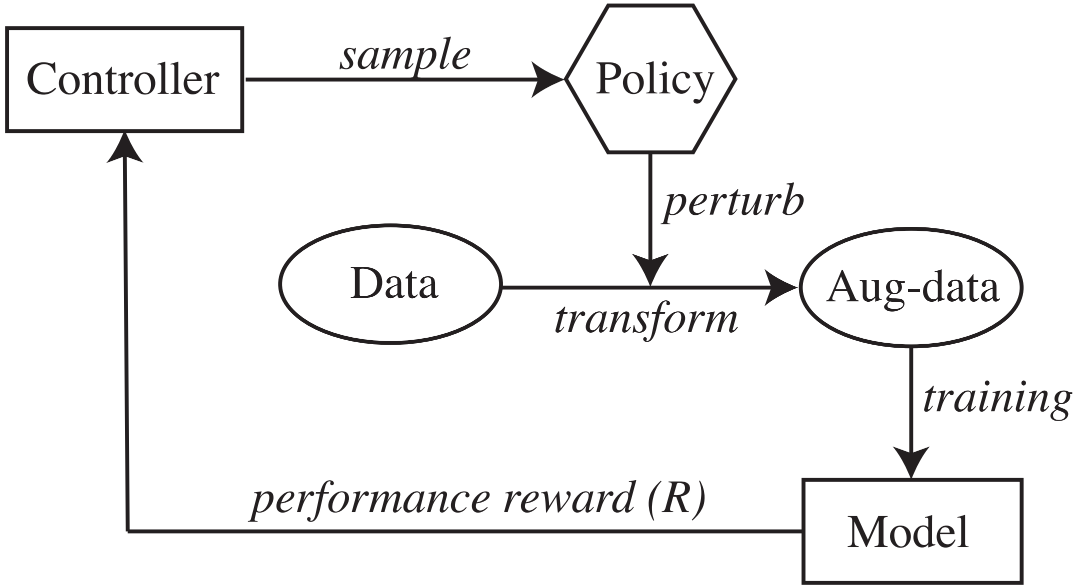


Figure 1: The controller samples a policy to perturb the training data. After training on the augmented inputs, the model feeds the performance back as reward.

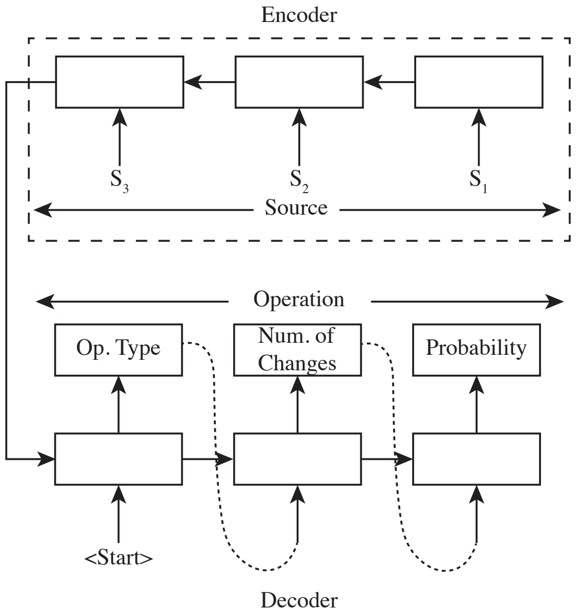


Figure 3: AutoAugment controller. An input-agnostic controller corresponds to the lower part of the figure. It samples a list of operations in sequence. An input-aware controller additionally has an encoder (upper part) that takes in the source inputs of the data.

[Cubuk et al., 2018] [Niu and Bansal, EMNLP 2019]

Auto-Augment Adversary Generation

Policy Hierarchy and Search Space:

- A policy consists of <u>4 sub-policies</u>;
- Each sub-policy consists of <u>2 operations</u> applied in sequence;
- Each operation is defined by <u>3 parameters</u>: **Operation Type**, **Number of Changes** (the maximum # of times allowed to perform operation, and **Probability** of applying that operation.
- Our pool of operations contains **Random Swap**, **Stopword Dropout**, **Paraphrase**, **Grammar Errors**, and **Stammer**.

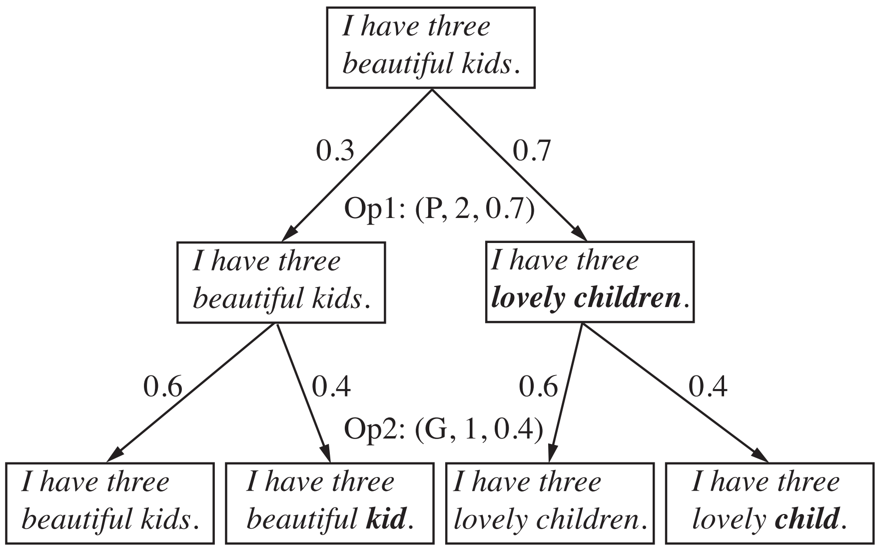


Figure 2: Example of a sub-policy applied to a source input. E.g., the first operation (Paraphrase, 2, 0.7) paraphrases the input twice with probability 0.7.

Subdivision of Operations:

- **Stopword Dropout:** To allow the controller to learn more nuanced combinations of operations, divide Stopword Dropout into 7 categories: Noun, Adposition, Pronoun, Adverb, Verb, Determiner, and Other.
- **Grammar Errors:** Noun (plural/singular confusion) and Verb (verb inflected/base form confusion).

[Niu and Bansal, EMNLP 2019]

Auto-Augment Adversary Generation

- **Setup:** Variational Hierarchical Encoder-Decoder (VHRED) (Serban et al., 2017b) on troubleshooting Ubuntu Dialogue task (Lowe et al., 2015); REINFORCE (Williams, 1992; Sutton et al., 2000) to train the controller.
- **Evaluation**: Serban et al. (2017a), evaluate on F1s for both activities (technical verbs) and entities (technical nouns). We also conducted human studies on Mturk, comparing each of the input-agnostic/aware models with the VHRED baseline and All-operations from Niu and Bansal (2018).

	Activity F1	Entity F1
LSTM	1.18	0.87
HRED	4.34	2.22
VHRED	4.63	2.53
VHRED (w/ attn.)	5.94	3.52
All-operations	6.53	3.79
Input-aware	**7.04**	3.90
Input-agnostic	7.02	**4.00**

Table 1: Activity, Entity F1 results reported by previous work, the All-operations and AutoAugment models.

	W	T	L	W - L
Input-agnostic vs. baseline	48	23	29	19
Input-aware vs. baseline	45	27	28	17
Input-agnostic vs. All-ops	43	27	30	13
Input-aware vs. All-ops	50	13	37	13

Table 4: Top 3 policies on the validation set and their test performances. Operations: R=Random Swap, D=Stopword Dropout, P=Paraphrase, G=Grammar Errors, S=Stammer. Universal tags: n=noun, v=verb, p=pronoun, adv=adverb, adp=adposition.

Sub-policy1	Sub-policy2	Sub-policy3	Sub-policy4
P, 1, 0.5	D_v, 3, 0.2	R, 3, 0.9	D_p, 2, 0.3
D_{adv}, 4, 0.4	R, 1, 0.5	D_{adp}, 1, 0.5	D_{adp}, 2, 0.1
D_n, 1, 0.8	D_o, 3, 1.0	P, 4, 0.4	G_n, 3, 0.3
G_v, 1, 0.9	D_o, 3, 0.1	S, 3, 0.4	R, 1, 0.2
D_v, 2, 0.5	D_v, 2, 0.7	S, 3, 0.5	P, 1, 1.0
R, 2, 0.2	G_v, 1, 0.9	D_o, 1, 0.5	G_n, 2, 0.6

Table 2: Human evaluation results on comparisons among the baseline, All-operations, and the two AutoAugment models. W: Win, T: Tie, L: Loss.

[Niu and Bansal, EMNLP 2019]

Auto-Augment Adversary Generation

- **Setup:** Variational Hierarchical Encoder-Decoder (VHRED) (Serban et al., 2017b) on troubleshooting Ubuntu Dialogue task (Lowe et al., 2015); REINFORCE (Williams, 1992; Sutton et al., 2000) to train the controller.
- **Evaluation**: Serban et al. (2017a), evaluate on F1s for both activities (technical verbs) and entities (technical nouns). We also conducted human studies on Mturk, comparing each of the input-agnostic/aware models with the VHRED baseline and All-operations from Niu and Bansal (2018).

	Activity F1	Entity F1
LSTM	1.18	0.87
HRED	4.34	2.22
VHRED	4.63	2.5[illegible]
VHRED (w/ attn.)	5.94	3.[illegible]
All-operations	6.53	3.[illegible]
Input-aware	**7.04**	3.[illegible]
Input-agnostic	7.02	**4.[illegible]**

Table 1: Activity, Entity F1 results reported by pre[illegible] work, the All-operations and AutoAugment model[illegible]

	W	T	L	W - L
Input-agnostic vs. baseline	48	23	29	19
Input-aware vs. baseline	45	27	28	17
Input-agnostic vs. All-ops	43	27	30	13
[illegible]	[illegible]	[illegible]	[illegible]7	13

Table 2: Human evaluation results on comparisons among the baseline, All-operations, and the two AutoAugment models. W: Win, T: Tie, L: Loss.

D_{adv}, 4, 0.4	R, 1, 0.5	D_{adp}, 1, 0.5	D_{adp}, 2, 0.1
D_n, 1, 0.8	D_o, 3, 1.0	P, 4, 0.4	G_n, 3, 0.3
G_v, 1, 0.9	D_o, 3, 0.1	S, 3, 0.4	R, 1, 0.2
D_v, 2, 0.5	D_v, 2, 0.7	S, 3, 0.5	P, 1, 1.0
R, 2, 0.2	G_v, 1, 0.9	D_o, 1, 0.5	G_n, 2, 0.6

[illegible]ormances. Operations: [illegible]rammar Errors, [illegible]=adverb, adp=adposition.

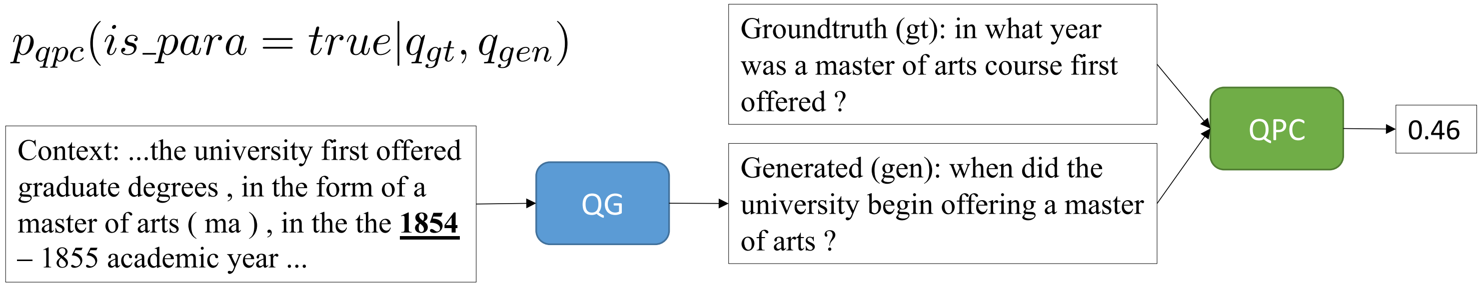
Still several challenges: better AutoAugm algorithms for RL speed, reward sparsity, other NLU/NLG tasks? Visit Tong's poster Nov5 3.30pm for more details!

[Niu and Bansal, EMNLP 2019]

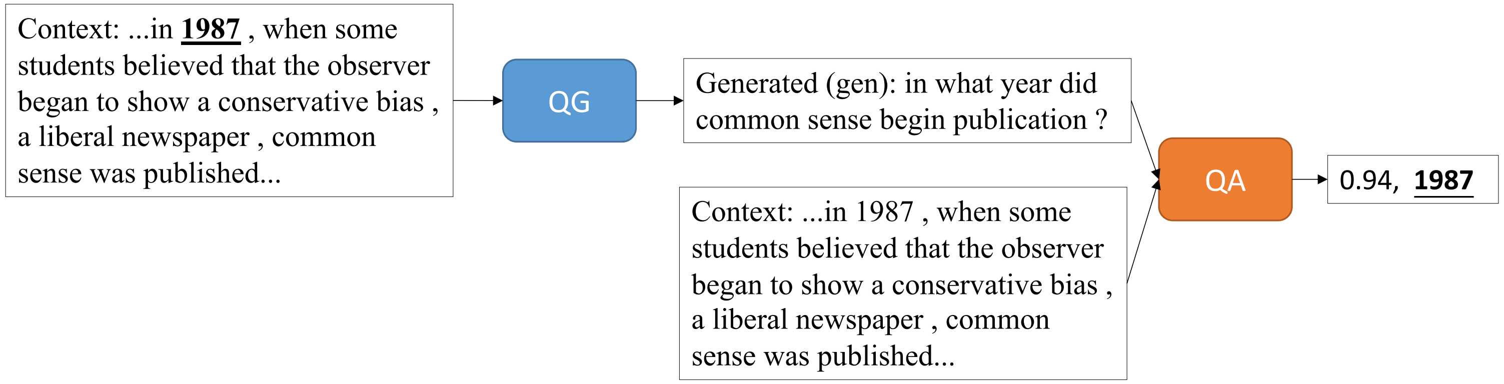
Robustness to New Questions via Semi-Supervised QG-for-QA

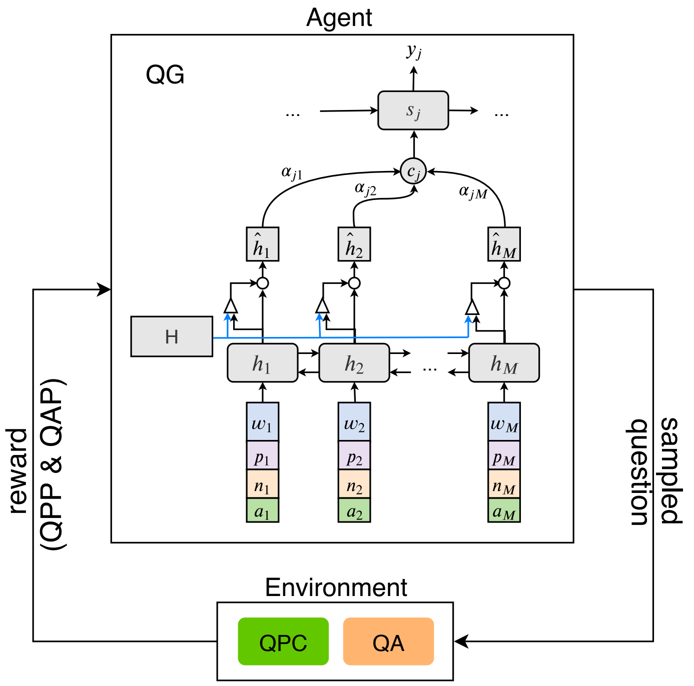
- Can also address Auto-Augment Robustness for QA by making it robust to new types of questions it has not seen before (via automatic question generation)!
- **Semantics-reinforced QG:** We first improve QG by addressing a "semantic drift" problem with two semantics-enhanced rewards (QPP = Question Paraphrasing Probability & QAP = Question Answering Probability) and introduce a QA-based QG evaluation method.

$$p_{qpc}(is\_para = true|q_{gt}, q_{gen})$$



$$p_{qa}(a|q_{gen}, context); q_{gen} \sim p_{qg}(q|a, context)$$





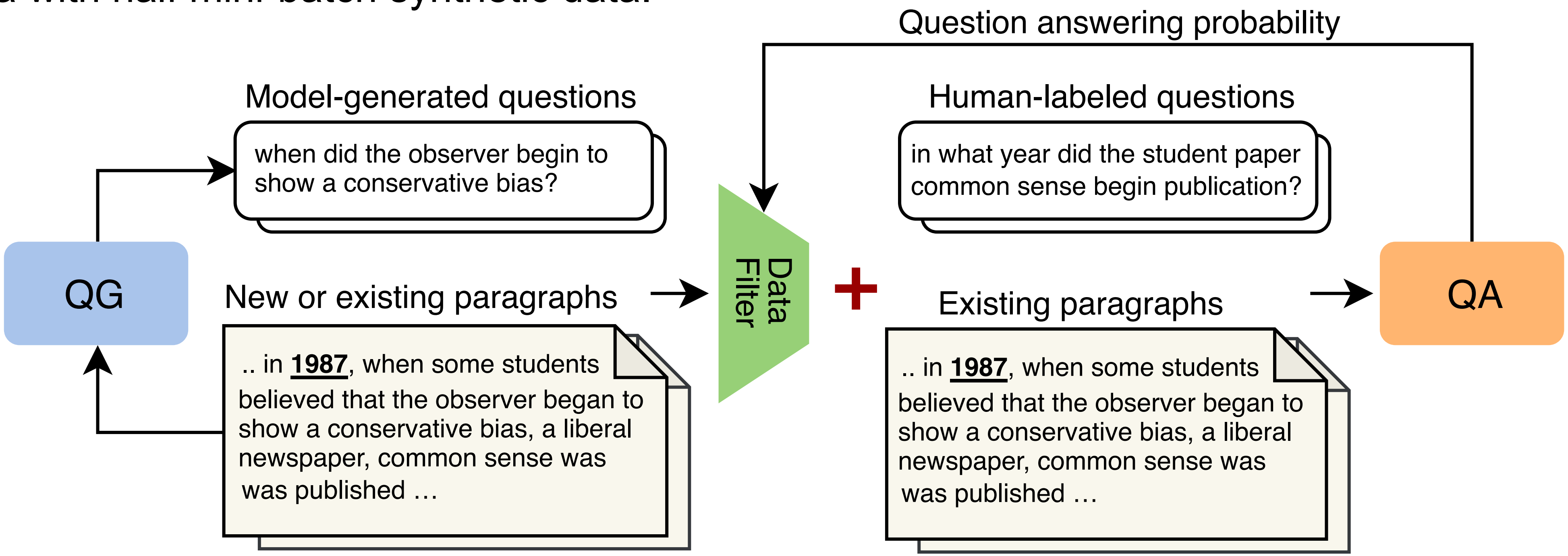
[Zhang and Bansal, EMNLP 2019]

Semi-Supervised QA with QG-Augmentation

Augment QA dataset with QG-generated examples (Generate from Existing Articles, and Generate from New Articles)

(1) QAP filter: To filter out poorly-generated examples; Filter synthetic examples with QAP < ε.

(2) Mixing mini-batch training: To make sure that the gradients from ground-truth data are not overwhelmed by synthetic data, for each mini-batch, we combine half mini-batch ground-truth data with half mini-batch synthetic data.



[Zhang and Bansal, EMNLP 2019]

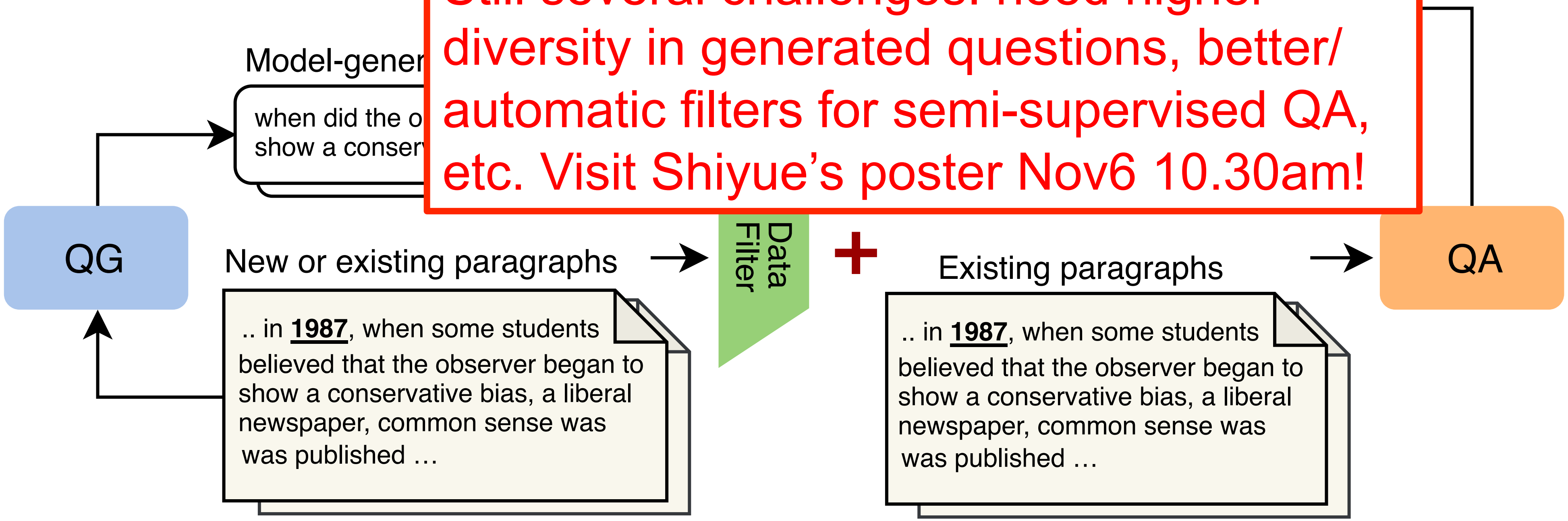
Semi-Supervised QA with QG-Augmentation

Augment QA dataset with QG-generated examples (Generate from Existing Articles, and Generate from New Articles)

(1) QAP filter: To filter out poorly-generated examples; Filter synthetic examples with QAP < ε.

(2) Mixing mini-batch training: To make sure that the gradients from ground-truth data are not overwhelmed by synthetic data, for each mini-batch, we combine half mini-batch ground-truth data with half mini-batch sy

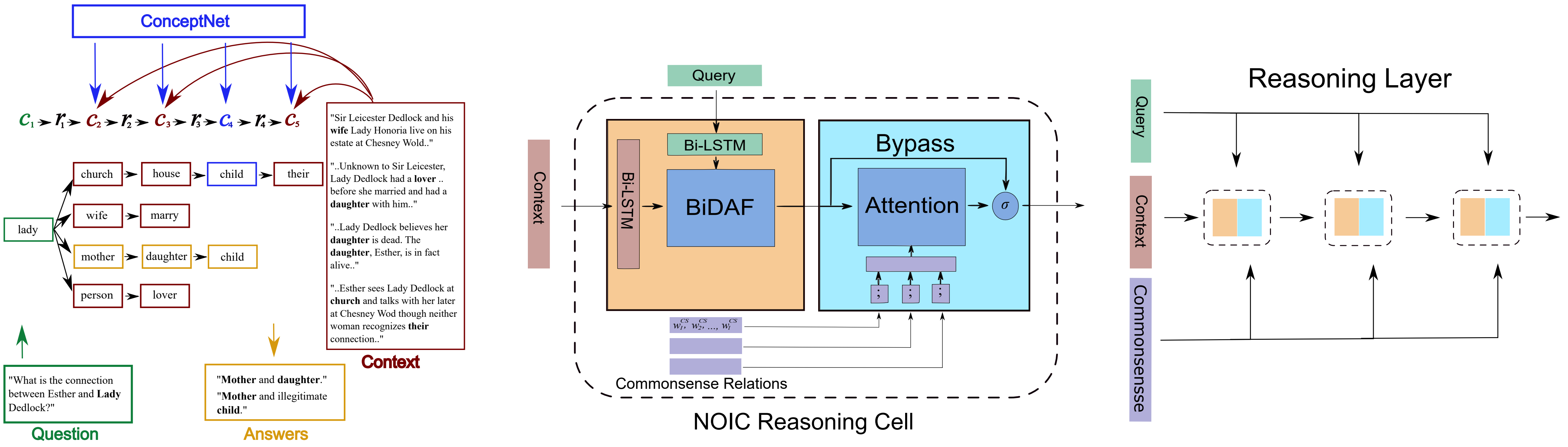
Still several challenges: need higher diversity in generated questions, better/ automatic filters for semi-supervised QA, etc. Visit Shiyue's poster Nov6 10.30am!



[Zhang and Bansal, EMNLP 2019]

Commonsense/Missing Knowledge Robustness in QA

- We use 'bypass-attention' mechanism to reason jointly on both internal context and external commonsense, and essentially learn when to fill 'gaps' of reasoning and with what information



[Bauer, Wang, and Bansal, EMNLP 2018]

Thoughts/Challenges/Current+Future Work

- BERT vs modularity?
- Evaluating NMN's interpretability when using contextualized input embeddings (BERT).
- New reasoning behaviors in more complex tasks?
- Structured knowledge as commonsense for QA and other NLU/NLG tasks
- Ongoing: Question generation for Multihop QA
- Ongoing: Auto-Augment for MultihopQA and addressing RL slowness, reward sparsity, etc.
- Ongoing: Multilingual extensions of QA/MultihopQA
- Our Multimodal QA work: TVQA and TVQA+

PhD Students


Lisa Bauer
PhD at UNC


Darryl Hannan
PhD at UNC


Peter Hase
PhD at UNC


Yichen Jiang
PhD at UNC


Hyounghun Kim
PhD at UNC
(co-advised w/ H. Fuchs)


Jie Lei
PhD at UNC
(co-advised w/ T. Berg)


Adyasha Maharana
PhD at UNC


Yixin Nie
PhD at UNC


Ramakanth Pasunuru
PhD at UNC


Swarnadeep Saha
PhD at UNC


Hao Tan
PhD at UNC


Shiyue Zhang
PhD at UNC


Yubo Zhang
PhD at UNC
(co-advised w/ A. Tropsha)


Xiang Zhou
PhD at UNC

Undergraduate Students


Tsion Coulter
UG at UNC


Han Guo
UG at UNC


Akshay Jain
UG at UNC


Sweta Karlekar
UG at UNC


Antonio Mendoza
UG at UNC


Yicheng Wang
UG at UNC


Songhe Wang
UG at UNC

THE UNIVERSITY *of* NORTH CAROLINA *at* CHAPEL HILL

Thank you!

Webpage: http://www.cs.unc.edu/~mbansal/

Email: mbansal@cs.unc.edu

UNC-NLP Lab: http://nlp.cs.unc.edu/

Postdoc Openings!!: ~mbansal/postdoc-advt-unc-nlp.pdf