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
“Which NFL team represented the AFC at Super Bowl 50?”

“Denver Broncos”

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 Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers …
Our Bi-directional Attention Flow Model (BiDAF) outperforms all previous approaches on the highly-competitive Stanford Question Answering Dataset (SQuAD) test set leaderboard at the time of submission. With a modification to only the output layer, BiDAF achieves the state-of-the-art results on the CNN/DailyMail cloze test. We also provide an in-depth ablation study of our model on the SQuAD development set, visualize the intermediate feature spaces in our model, and analyse its performance as compared to a more traditional language model for machine comprehension (Rajpurkar et al., 2016).

Our machine comprehension model is a hierarchical multi-stage process and consists of six layers (Figure 1):

1. Character Embedding Layer
   - Maps each word to a vector space using character-level CNNs.
2. Word Embedding Layer
   - Maps each word to a vector space using a pre-trained word embedding model.
3. Contextual Embedding Layer
   - Utilizes contextual cues from surrounding words to refine the embedding of the words. These first three layers are applied to both the query and context.
4. Attention Flow Layer
   - Couples the query and context vectors and produces a set of query-aware feature vectors for each word in the context.
5. Modeling Layer
   - Employs a Recurrent Neural Network to scan the context.
6. Output Layer
   - Provides an answer to the query.

Our code and interactive demo are available at: allenai.github.io/bi-att-flow/
Multi-Hop QA: Bridge-Type

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 $\xrightarrow{voted\_as}$ World's Best Goalkeeper

Bridge Entity
Multi-Hop QA: Comparison-Type

**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 ...

Scott Derrickson $\xrightarrow{is\_from} \text{America}$

Ed Wood $\xrightarrow{is\_from} \text{America}$

$\xrightarrow{is\_same}$

Yes
Challenges: Different Reasoning Chains in Multi-Hop QA

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

```
Kasper Schmeichel \(\text{son-of}\) Peter Schmeichel \(\text{voted-as}\) World’s Best Goalkeeper
```

“Were Scott Derrickson and Ed Wood of the same nationality?”

```
Scott Derrickson \(\text{is-from}\) America
Ed Wood \(\text{is-from}\) America
```

```
\text{is_same}, \text{Yes}
```
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).
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).

There is a shiny object that is right of the gray metallic cylinder; does it have the same size as the large rubber sphere?
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: $c_u t$ (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]
Inputs: Question emb: $u$, Sub-question vector: $c_b$ Context emb: $h$

<table>
<thead>
<tr>
<th>Module Name</th>
<th>Input Attention</th>
<th>Output Types</th>
<th>Implementation Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find($u, c, h$)</td>
<td>(None)</td>
<td>Attention</td>
<td>$\text{BiAttn}(h \odot c_t, u)$</td>
</tr>
<tr>
<td>Relocate($u, c, h$)</td>
<td>$a_1$</td>
<td>Attention</td>
<td>$\text{Find}(u, c_t, h \odot (a_1 \cdot h))$</td>
</tr>
<tr>
<td>Compare($u, c, h$)</td>
<td>$a_1, a_2$</td>
<td>Yes/No</td>
<td>$\sigma(\text{MLP}([c_t, a_1 \cdot h, a_2 \cdot h, c_t \cdot (a_1 - a_2) \cdot h]))$</td>
</tr>
<tr>
<td>NoOp($u, c, h$)</td>
<td>(None)</td>
<td>(None)</td>
<td>(None)</td>
</tr>
</tbody>
</table>

[Jiang and Bansal, EMNLP 2019]
Putting an NMN together...

Controller:

Module:

Q: Were Scott Derrickson and Ed Wood of the same nationality?

Sub-question:

Find:

Scott Derrickson is an American director.

avg. output of all modules

Push:

Stack of Attention

[73x471] Putting an NMN together...

Controller:

Module:

Q: Were Scott Derrickson and Ed Wood of the same nationality?

Sub-question:

Find:

Scott Derrickson is an American director.

avg. output of all modules

Push:

Stack of Attention

[Jiang and Bansal, EMNLP 2019]
Putting an NMN together...

**Controller:**

**Modules:**

*Q: Were Scott Derrickson and Ed Wood of the same nationality?*

Controller:

- Sub-question
- Find
- Find

Find

- Scott Derrickson is an American director
- Edward Wood Jr. was an American filmmake

Find

- avg. output of all modules
- Push

- avg. output of all modules
- Push

Stack of Attention

[Jiang and Bansal, EMNLP 2019]
Putting an NMN together...

Controller:

Q: Were Scott Derrickson and Ed Wood of the same nationality?

Modules:

Sub-question

Find

Find

Compare

Scott Derrickson is an American director. Edward Wood Jr. was an American filmmaker.

avg. output of all modules

avg. output of all modules

Prediction: Yes

avg. output of all modules

Jiang and Bansal, EMNLP 2019
## Main Results on HotpotQA

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev F1</th>
<th>Test F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiDAF Baseline</td>
<td>57.19</td>
<td>55.81</td>
</tr>
<tr>
<td>Original NMN</td>
<td>40.28</td>
<td>39.90</td>
</tr>
<tr>
<td>Our NMN</td>
<td>63.35</td>
<td>62.71</td>
</tr>
</tbody>
</table>

[Jiang and Bansal, EMNLP 2019]
## Ablation Studies

*All models are evaluated on our dev set.*

<table>
<thead>
<tr>
<th></th>
<th>Bridge</th>
<th></th>
<th>Comparison</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td></td>
<td>F1</td>
<td></td>
</tr>
<tr>
<td>Our NMN</td>
<td>64.49</td>
<td></td>
<td>57.20</td>
<td></td>
</tr>
<tr>
<td>-Relocate</td>
<td>60.13</td>
<td></td>
<td>58.10</td>
<td></td>
</tr>
<tr>
<td>-Compare</td>
<td>64.46</td>
<td></td>
<td>56.00</td>
<td></td>
</tr>
</tbody>
</table>

[Chiang and Bansal, EMNLP 2019]
Adversarial Evaluation

<table>
<thead>
<tr>
<th>Train Eval</th>
<th>BiDAF Baseline</th>
<th></th>
<th>Our NMN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reg</td>
<td>43.12</td>
<td>50.13</td>
<td></td>
</tr>
<tr>
<td>Adv</td>
<td>45.12</td>
<td>49.33</td>
<td></td>
</tr>
<tr>
<td>Adv Reg</td>
<td>44.65</td>
<td>49.25</td>
<td></td>
</tr>
<tr>
<td>Adv Adv</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.
## Analysis: Controller Attention Visualization

<table>
<thead>
<tr>
<th>Step 1:</th>
<th>Step 2:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What government position was held by the woman portrayed Corliss Archer in the film Kiss and Tell</strong></td>
<td><strong>Kiss and Tell is a 1945 American comedy film starring then 17-year-old Shirley Temple as Corliss Archer.</strong>...</td>
</tr>
<tr>
<td><strong>Shirley Temple Black was an American actress, ..., and also served as Chief of Protocol of the United States.</strong></td>
<td><strong>We also have initial human evaluation results on controller’s sub-question soft decomposition/attention.</strong></td>
</tr>
</tbody>
</table>

- We also have initial human evaluation results on controller’s sub-question soft decomposition/attention.
Analysis: Controller Attention for Comparison Questions

<table>
<thead>
<tr>
<th></th>
<th>Was</th>
<th>Scott</th>
<th>Derrickson</th>
<th>and</th>
<th>Ed</th>
<th>Wood</th>
<th>of</th>
<th>the</th>
<th>same</th>
<th>nationality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ctrl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ctrl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ctrl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 $\geq$ 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 $\geq$ 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:

“Were Scott Derrickson and Ed Wood of the same nationality?”

Find -> Relocate: 99.9%

Comparison Yes/No:

“Find -> Relocate: 99.9%

Find -> Relocate -> Compare:

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

See Yichen’s full talk on Nov7 10.30am!
(2) Divergent Reasoning Chains

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 ("Bergstädtle") - 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*
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
### Results - WikiHop and MedHop

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

<table>
<thead>
<tr>
<th></th>
<th>Test (Masked)</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>FastQA (Weissenborn et al., 2017)</td>
<td>23.1</td>
<td>31.3</td>
</tr>
<tr>
<td>BiDAF (Seo et al., 2017)</td>
<td>33.7</td>
<td>47.8</td>
</tr>
<tr>
<td>CoAttention</td>
<td>-</td>
<td>58.1</td>
</tr>
<tr>
<td>Most Frequent Candidate</td>
<td>10.4</td>
<td>58.4</td>
</tr>
<tr>
<td><strong>EPAr (Ours)</strong></td>
<td><strong>41.6</strong></td>
<td><strong>60.3</strong></td>
</tr>
</tbody>
</table>

WikiHop                  MedHop

[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

<table>
<thead>
<tr>
<th>Method</th>
<th>R@1</th>
<th>R@2</th>
<th>R@3</th>
<th>R@4</th>
<th>R@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>11.2</td>
<td>17.3</td>
<td>27.6</td>
<td>40.8</td>
<td>50.0</td>
</tr>
<tr>
<td>1-hop TFIDF</td>
<td>32.7</td>
<td>48.0</td>
<td>56.1</td>
<td>63.3</td>
<td>70.4</td>
</tr>
<tr>
<td>2-hop TFIDF</td>
<td>42.9</td>
<td>56.1</td>
<td>70.4</td>
<td>78.6</td>
<td>82.7</td>
</tr>
<tr>
<td>DE</td>
<td>38.8</td>
<td>50.0</td>
<td>65.3</td>
<td>73.5</td>
<td>83.7</td>
</tr>
<tr>
<td>TFIDF+DE</td>
<td><strong>44.9</strong></td>
<td><strong>64.3</strong></td>
<td><strong>77.6</strong></td>
<td><strong>82.7</strong></td>
<td><strong>90.8</strong></td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Method</th>
<th>R@1</th>
<th>R@2</th>
<th>R@3</th>
<th>R@4</th>
<th>R@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>11.2</td>
<td>17.3</td>
<td>27.6</td>
<td>40.8</td>
<td>50.0</td>
</tr>
<tr>
<td>1-hop TFIDF</td>
<td>32.7</td>
<td>48.0</td>
<td>56.1</td>
<td>63.3</td>
<td>70.4</td>
</tr>
<tr>
<td>2-hop TFIDF</td>
<td>42.9</td>
<td>56.1</td>
<td>70.4</td>
<td>78.6</td>
<td>82.7</td>
</tr>
<tr>
<td>DE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFIDF+DE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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!
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

**Reasoning Shortcut:**

[Placeholder] \(\xrightarrow{voted\_as}\) World’s Best Goalkeeper

[33]

[Jiang and Bansal, ACL 2019]
<table>
<thead>
<tr>
<th>Question</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>“What was the father of Kasper Schmeichel <strong>voted to be by the IFFHS in 1992</strong>?”</td>
<td>Peter Bolesław Schmeichel is a Danish former professional footballer .., and was <strong>voted</strong> the IFFHS World's Best Goalkeeper <strong>in 1992</strong> and 1993.</td>
</tr>
<tr>
<td>The answer can be directly inferred by word-matching the documents to maximum of the question !!!</td>
<td>Edson Arantes do Nascimento is a retired Brazilian professional footballer. In 1999, he was <strong>voted</strong> World Player of the Century by IFFHS. [Missing: 1992]</td>
</tr>
<tr>
<td></td>
<td>Kasper Hvidt is a Danish retired handball goalkeeper, .. also <strong>voted</strong> as Goalkeeper of the Year March 20, 2009, [Missing: 1992, IFFHS]</td>
</tr>
</tbody>
</table>
How to eliminate this reasoning shortcut from the data to **ENFORCE** compositional reasoning?

Building adversarial documents as better distractors

Min et al., 2019; Chen & Durrett, 2019
“What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?”

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

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!
**BERT (Document Retrieval Results)**

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

<table>
<thead>
<tr>
<th>Train \ Eval</th>
<th>Eval = Regular</th>
<th>Eval = Adv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train = Regular</td>
<td>89.44</td>
<td>44.67</td>
</tr>
<tr>
<td>Train = Adv</td>
<td>89.03</td>
<td>80.14</td>
</tr>
</tbody>
</table>

- 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)

<table>
<thead>
<tr>
<th>Train \ Eval</th>
<th>Eval = Regular</th>
<th>Eval = Adv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train = Regular</td>
<td>89.44</td>
<td>44.67</td>
</tr>
<tr>
<td>Train = Adv</td>
<td>89.03</td>
<td>80.14</td>
</tr>
</tbody>
</table>

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

• 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

<table>
<thead>
<tr>
<th>Train \ Eval</th>
<th>Eval = Regular</th>
<th>Eval = Adv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train = Regular</td>
<td>43.12</td>
<td>34.00</td>
</tr>
<tr>
<td>Train = Adv</td>
<td>45.12</td>
<td>44.65</td>
</tr>
</tbody>
</table>

* Exact-Match scores

- 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.
Bi-attention + Self-attention Baseline

<table>
<thead>
<tr>
<th>Train \ Eval</th>
<th>Eval = Regular</th>
<th>Eval = Adv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train = Regular</td>
<td>43.12</td>
<td>34.00</td>
</tr>
<tr>
<td>Train = Adv</td>
<td>45.12</td>
<td>44.65</td>
</tr>
</tbody>
</table>

* Exact-Match scores

- 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.
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]
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 retrain. 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.
Auto-Augment Adversary Generation

Policy Hierarchy and Search Space:

- A policy consists of 4 sub-policies;
- Each sub-policy consists of 2 operations applied in sequence;
- Each operation is defined by 3 parameters: 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.

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).

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.
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).

<table>
<thead>
<tr>
<th>Activity F1</th>
<th>Entity F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>1.18</td>
</tr>
<tr>
<td>HRED</td>
<td>4.34</td>
</tr>
<tr>
<td>VHRED</td>
<td>4.63</td>
</tr>
<tr>
<td>VHRED (w/ attn.)</td>
<td>5.94</td>
</tr>
<tr>
<td>All-operations</td>
<td>6.53</td>
</tr>
<tr>
<td>Input-aware</td>
<td>7.04</td>
</tr>
<tr>
<td>Input-agnostic</td>
<td>7.02</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>W</th>
<th>T</th>
<th>L</th>
<th>W - L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input-agnostic vs. baseline</td>
<td>48</td>
<td>23</td>
<td>29</td>
</tr>
<tr>
<td>Input-aware vs. baseline</td>
<td>45</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td>Input-agnostic vs. All-ops</td>
<td>43</td>
<td>27</td>
<td>30</td>
</tr>
<tr>
<td>Input-aware vs. All-ops</td>
<td>50</td>
<td>13</td>
<td>37</td>
</tr>
</tbody>
</table>


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

<table>
<thead>
<tr>
<th>Activity</th>
<th>Activity F1</th>
<th>Entity F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>1.18</td>
<td>0.87</td>
</tr>
<tr>
<td>HRED</td>
<td>4.34</td>
<td>2.22</td>
</tr>
<tr>
<td>VHRED</td>
<td>4.63</td>
<td>2.53</td>
</tr>
<tr>
<td>VHRED (w/ attn.)</td>
<td>5.94</td>
<td>3.59</td>
</tr>
<tr>
<td>All-operations</td>
<td>6.53</td>
<td>3.73</td>
</tr>
<tr>
<td>Input-aware</td>
<td><strong>7.04</strong></td>
<td>3.95</td>
</tr>
<tr>
<td>Input-agnostic</td>
<td>7.02</td>
<td>4.06</td>
</tr>
</tbody>
</table>

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

Still several challenges: better AutoAugment algorithms for RL speed, reward sparsity, other NLU/NLG tasks? Visit Tong’s poster Nov 5 3.30pm for more details!
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.

\[
p_{qpc}(\text{is \_ para} = \text{true}|q_{gt}, q_{gen})
\]

- **Groundtruth (gt):** in what year was a master of arts course first offered?
- **Generated (gen):** when did the university begin offering a master of arts?

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

- **Generated (gen):** in what year did common sense begin publication?
- **Context:** ...in 1987, when some students believed that the observer began to show a conservative bias, a liberal newspaper, common sense was published...

[Zhang and Bansal, EMNLP 2019]
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 < \( \varepsilon \).

(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.
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 $\text{QAP} < \varepsilon$.
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.

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]
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.
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

Yichun Jiang
PhD at UNC

Hyoungchun 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

Shiyou Zhang
PhD at UNC

Yubo Zheng
PhD at UNC
(co-advised w/ A. Trappshah)

Xiang Zhou
PhD at UNC

Undergraduate Students

Tision 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
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