Avoiding Reasoning Shortcuts: Adversarial Evaluation, Training, and Model Development for Multi-Hop QA

Yichen Jiang
www.jiang-yichen.io

Mohit Bansal
www.cs.unc.edu/~mbansal/

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

UNC NLP

Data/code available at https://github.com/jiangycTarheel/Adversarial-MultiHopQA

Single-Hop QA

[Rajpurkar et al., 2016]

Question

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

Single-Hop QA

[Rajpurkar et al., 2016]

Question

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

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 Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers …

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 Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers …

Multi-Hop QA

[Yang et al., 2018]

Question

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

Multi-Hop QA

[Yang et al., 2018]

Question

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

$$\boxed{\text{Kasper Schmeichel}} \xrightarrow{son\_of} \text{???} \xrightarrow{voted\_as} \text{???}$$

Multi-Hop QA

[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.

$$\boxed{\text{Kasper Schmeichel}} \xrightarrow{son\_of} ??? \xrightarrow{voted\_as} ???$$

Multi-Hop QA

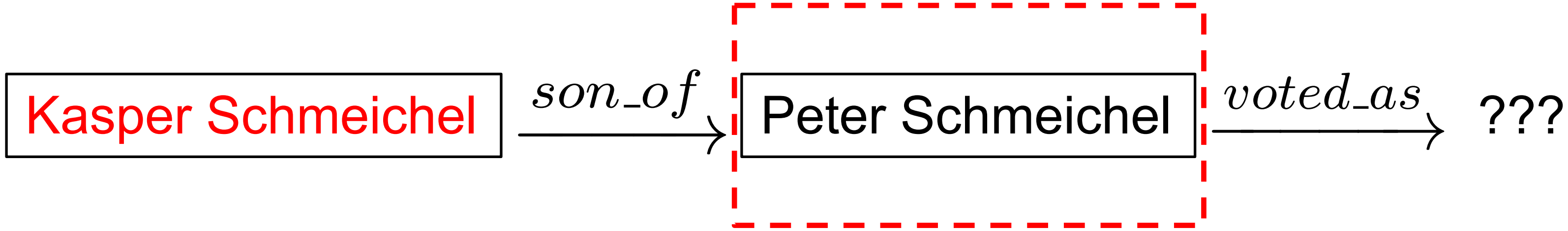
[Yang et al., 2018]

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.



Kasper Schmeichel

son_of

Peter Schmeichel

voted_as

???

Multi-Hop QA

[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 …

$$\boxed{\text{Kasper Schmeichel}} \xrightarrow{son\_of} \boxed{\text{Peter Schmeichel}} \xrightarrow{voted\_as} \text{???}$$

Multi-Hop QA

[Yang et al., 2018]

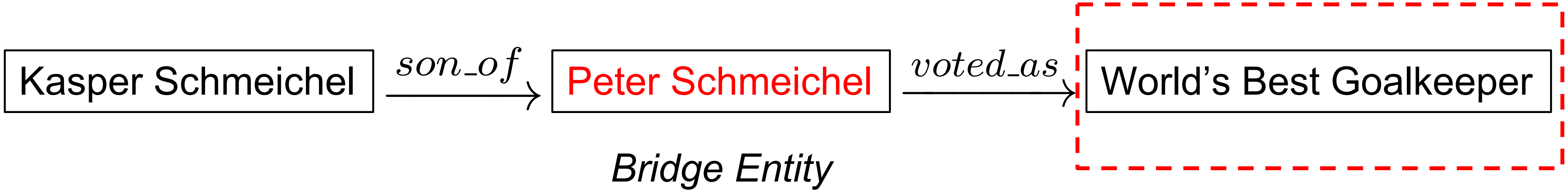
Question

"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 …



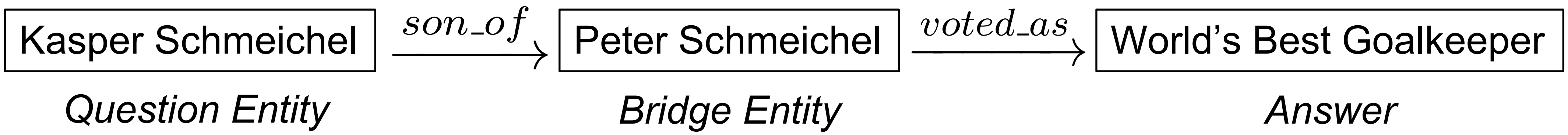
Kasper Schmeichel $\xrightarrow{son\_of}$ Peter Schmeichel $\xrightarrow{voted\_as}$ World's Best Goalkeeper

Bridge Entity

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

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

Reasoning Chain:



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

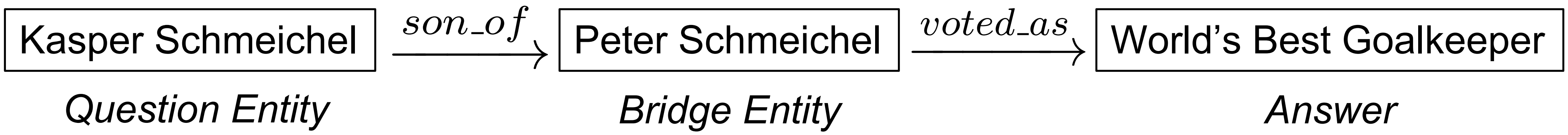
Not always!

Reasoning Shortcut

Question

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

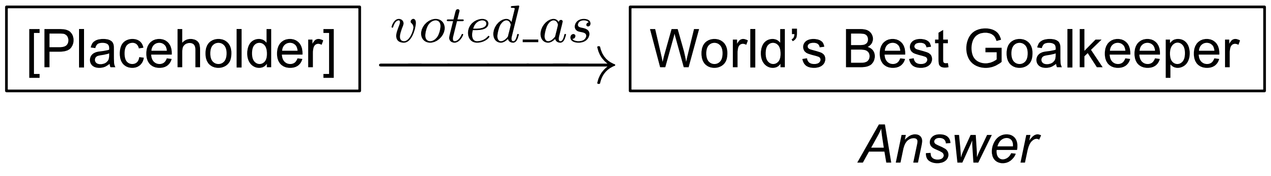
Reasoning Chain:



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

Question Entity *Bridge Entity* *Answer*

Reasoning Shortcut:



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

Answer

Reasoning Shortcut

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.

Reasoning Shortcut

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.

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

Reasoning Shortcut

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.

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]

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 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]

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

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

Building **adversarial documents**

as better distractors

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

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

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. Bolesław Kelly 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 would found two plausible answers !

Title-Balancing

Question

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

Problem: The title of the adversarial document ("R. Bolesław Kelly") is never mentioned in other documents in the context.

Context

Kasper Schmeichel is a Danish professional footballer ... former Manchester United and Danish international goalkeeper Peter Schmeichel.

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

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

Title-Balancing

Question

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

Solution: Add another document from Wikipedia that mentions the adversarial title.

Context

Kasper Schmeichel is a Danish professional footballer ... former Manchester United and Danish international goalkeeper Peter Schmeichel.

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

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

… R. Kelly also participated in the …

Related Works (Adversarial Eval/Train)

- Jia and Liang, 2017: **Adversarial Examples for Evaluating Reading Comprehension Systems**
  - Created adversaries to attack existing QA models on SQuAD dataset
  - Adversarial training had bias

- Wang and Bansal, 2018: **Robust Machine Comprehension Models via Adversarial Training**
  - Improved adversarial training on SQuAD with diverse fake answers and random adversary placement

- Niu and Bansal, 2018: **Adversarial Over-Sensitivity and Over-Stability Strategies for Dialogue Models**
  - Adversarially-trained model performs significantly better in both adversarial inputs and original inputs.

Related Works (Multi-Hop QA)

- Chen & Durrett, NAACL 2019: **Understanding Dataset Design Choices for Multi-hop Reasoning**
  - Models that cannot do multi-hop reasoning by design are still able to solve a large number of examples in HotpotQA and WikiHop

- Min et al., ACL 2019: **Compositional Questions Do Not Necessitate Multi-hop Reasoning**
  - A single-hop BERT-based RC model can achieve near-state-of-the-art performance
  - Humans, who are not shown all of the necessary paragraphs for the intended multi-hop reasoning, can still answer over 80% of questions
  - Suggested generating better distractors using adversarial examples.

Related Works (Multi-Hop QA)

- Chen & Durrett, NAACL 2019: **Understanding Dataset Design Choices for Multi-hop Reasoning**
- Min et al., ACL 2019: **Compositional Questions Do Not Necessitate Multi-hop Reasoning**

- These two concurrent works identified reasoning shortcuts by **building single-hop-only models** that achieve good performance in HotpotQA.
- We instead create adversaries to **eliminate reasoning shortcuts**, and show that models achieving strong performance in the original HotpotQA cannot solve our adversarial examples.

HotpotQA

[Yang et al., 2018]

- 2 settings
  - Distractor setting: 2 documents with evidence + 8 distractor documents
  - Full-wiki setting: entire Wikipedia as the context

HotpotQA

[Yang et al., 2018]

- 2 settings
  - <u>Distractor setting: 2 documents with evidence + 8 distractor documents</u>
  - Full-wiki setting: entire Wikipedia as the context

- 4 different reasoning types:
  - Bridge-type I

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

HotpotQA

[Yang et al., 2018]

- 2 settings
  - Distractor setting: 2 documents with evidence + 8 distractor documents
  - Full-wiki setting: entire Wikipedia as the context

- 4 different reasoning types:
  - Bridge-type I

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

    "What city is the Marine Air Control Group 28 located in?"

HotpotQA

[Yang et al., 2018]

- 2 settings
  - Distractor setting: 2 documents with evidence + 8 distractor documents
  - Full-wiki setting: entire Wikipedia as the context

- 4 different reasoning types:
  - Bridge-type I
    *"What was **the father of Kasper Schmeichel** voted to be by the IFFHS in 1992?"*
  - Bridge-type II
    "What city is the Marine Air Control Group 28 located in?"
  - Checking Multiple Entities
    *"Which **French ace pilot and adventurer fly L'Oiseau Blanc**?"*

HotpotQA

[Yang et al., 2018]

- 2 settings
  - Distractor setting: 2 documents with evidence + 8 distractor documents
  - Full-wiki setting: entire Wikipedia as the context

- 4 different reasoning types:
  - Bridge-type I

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

    "What city is the Marine Air Control Group 28 located in?"
  - Checking Multiple Entities

    *"Which **French ace pilot and adventurer fly L'Oiseau Blanc**?"*
  - Comparison

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

Baselines

- BERT Base (Retrieval) [Devlin et al., 2018]
- Bi-attention + Self-attention [Yang et al., 2018]

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.

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.

Bi-attention + Self-attention Baseline

UNC NLP

* 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.

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 to be aware of distractors and force it not to take the reasoning shortcut.

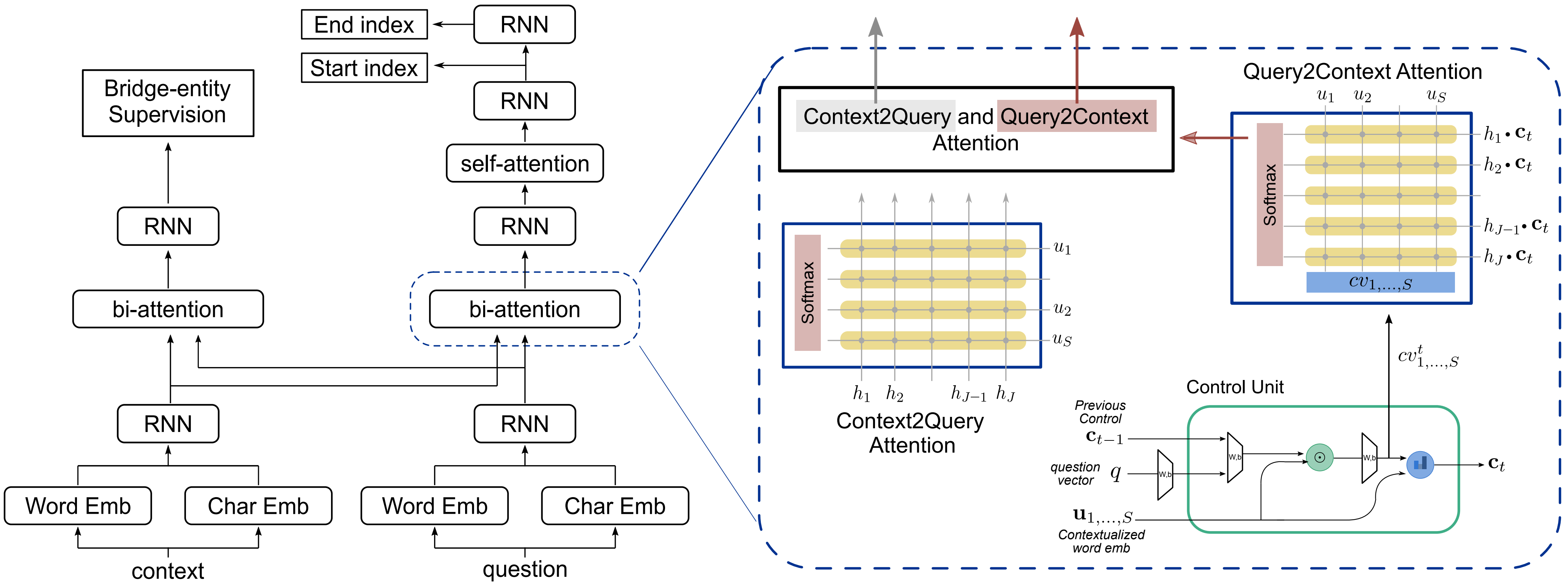
Bi-attention + Self-attention Baseline

* 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 also obtains better performance in the regular evaluation.
- The multi-hop reasoning skills learnt from the adversarial data is also beneficial to the regular evaluation.

An Initial 2-Hop Architecture



2-Hop Model

Train \ Eval	Eval = Regular	Eval = Adv
Train = Regular	46.41	32.30
Train = Adv	47.08	46.87

- The performance of the 2-hop model 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.

2-Hop Model

Train \ Eval	Eval = Regular	Eval = Adv
Train = Regular	46.41	32.30
Train = Adv	47.08	46.87

- After being trained on the adversarial data, the 2-hop model 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.

2-Hop Model

Train \ Eval	Eval = Regular	Eval = Adv
Train = Regular	46.41	32.30
Train = Adv	47.08	46.87

- After being trained on the adversarial data, the 2-hop model also obtains better performance in the regular evaluation.
- The multi-hop reasoning skills learnt from the adversarial data is also beneficial to the regular evaluation.

Analysis

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

Analysis

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

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

Analysis – Adversary Success

Question

"Where is the company that Sachin Warrier worked as a software engineer headquartered?"

Context

Sachin Warrier, …, he was working as a software engineer in Tata Consultancy Services in Kochi. …

Tata Consultancy Services is … company headquartered in Mumbai, …

Answer Before Adversary

Mumbai

Analysis – Adversary Success

Question

"Where is the company that Sachin Warrier worked as a software engineer headquartered?"

Answer Before Adversary

Mumbai

Context

Sachin Warrier, …, he was working as a software engineer in Tata Consultancy Services in Kochi. …

Tata Consultancy Services is … company headquartered in Mumbai, …

Adversarial Document

Valencia Street Circuit is … company headquartered in Delhi, …

Analysis – Adversary Success

Question

“Where is the company that Sachin Warrier worked as a software engineer headquartered?”

Answer Before Adversary

Mumbai

Answer After Adversary

Delhi

Context

Sachin Warrier, …, he was working as a software engineer in Tata Consultancy Services in Kochi. …

Tata Consultancy Services is … company headquartered in Mumbai, …

Adversarial Document

Valencia Street Circuit is … company headquartered in Delhi, …

Analysis

- Manual Verification of Adversaries
  - 0 out of 50 examples has 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
  - Model is still able to predict the correct answer with the adversary added

Analysis – Adversary Failure

Question

“Who produced the film that was Jennifer Kent’s directorial debut?”

Context

The Badabook is a 2014 Australian film … directed by Jennifer Kent in her directorial debut, and produced by Kristina Ceyton and Kristian Moliere. …

Answer

Kristina Ceyton and Kristian Moliere

Analysis – Adversary Failure

Question

“Who produced the film that was Jennifer Kent’s directorial debut?”

Context

The Badabook is a 2014 Australian film … directed by Jennifer Kent in her directorial debut, and produced by Kristina Ceyton and Kristian Moliere. …

Answer

Kristina Ceyton and Kristian Moliere

Adversarial Document

The Aphra Behn is a 2014 Australian film … directed by Scott Hahn in her directorial debut, and produced by Kristina Mutrux and Kristian Ionesco. …

Conclusions

- In this work, we…
  - identified reasoning shortcuts in the HotpotQA.

Conclusions

- In this work, we…
  - identified reasoning shortcuts in the HotpotQA.
  - constructed adversaries that can fool the models exploiting the shortcut, and the strong models' scores drop significantly.

Conclusions

- In this work, we…
  - identified reasoning shortcuts in the HotpotQA.
  - constructed adversaries that can fool the models exploiting the shortcut, and the strong models' scores drop significantly.
  - showed that models can improve on the adversarial evaluation after being trained on the adversarial data.

Conclusions

- In this work, we…
    - identified reasoning shortcuts in the HotpotQA.
    - constructed adversaries that can fool the models exploiting the shortcut, and the strong models' scores drop significantly.
    - showed that models can improve on the adversarial evaluation after being trained on the adversarial data.
    - proposed to use a control unit to guide the bi-attention in multi-hop reasoning, which achieved some initial improvements in both regular and adversarial settings.

Thank you for listening!
Questions?

Data/code available at https://github.com/jiangycTarheel/Adversarial-MultiHopQA

Acknowledgement: DARPA (YFA17-D17AP00022), Google, Bloomberg, NVidia, AWS