Commonsense for Generative Multi-Hop Question Answering Tasks

Lisa Bauer, Yicheng Wang, Mohit Bansal

EMNLP 2018

Mengdi Huang (mengdih@seas.upenn.edu)

03/27/2019

Introduction

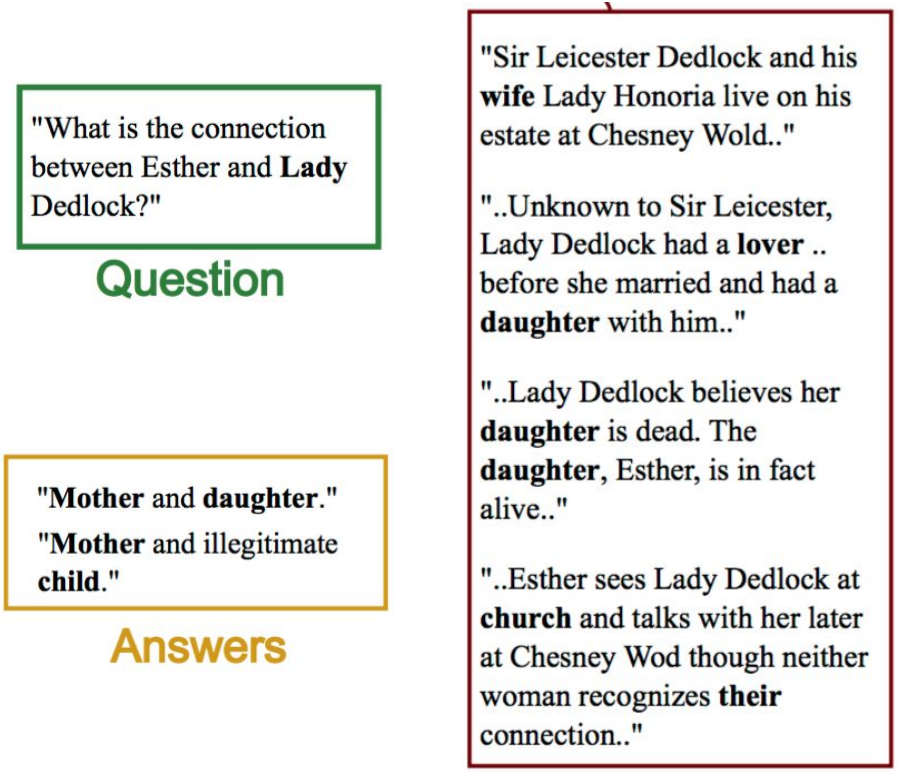
- Machine Reading Comprehension-based QA

  This task tests a model's natural language understanding capabilities by asking it to answer a question

- Types of MRC-QA
  - **Fact-finding extractive QA**: answers are guaranteed to be **spans** within the input text (CNN/DM (Hermann et al.,2015), SQuAD (Rajpurkar et al., 2016), QAngaroo (Welbl et al., 2018))
  - **Generative QA**: answers require multi-hop reasoning for long, complex stories and other narratives, which requires the model to go beyond fact linking and to synthesize **non-span** answers (**NarrativeQA generative dataset** (Kocisky et al., 2018)).

Problem

- Fact-finding extractive QA does not work on multi-hop generative tasks.
  - e.g.: Question: “What is the connection between A and B?”
  - Content: “A believes her daughter is dead. The daughter, B, is in fact alive.”
- Although generative model can gather and synthesize disjoint pieces of information within the context, it cannot understand implicit relations and fill in gaps of reasoning without external, background commonsense knowledge.



"What is the connection between Esther and **Lady** Dedlock?"

Question

"**Mother** and **daughter**."

"**Mother** and illegitimate **child**."

Answers

"Sir Leicester Dedlock and his **wife** Lady Honoria live on his estate at Chesney Wold.."

"..Unknown to Sir Leicester, Lady Dedlock had a **lover** .. before she married and had a **daughter** with him.."

"..Lady Dedlock believes her **daughter** is dead. The **daughter**, Esther, is in fact alive.."

"..Esther sees Lady Dedlock at **church** and talks with her later at Chesney Wod though neither woman recognizes **their** connection.."

Motivation

- We want the model to be able to answer questions that require multi-hop reasoning for long, complex stories and other narratives, which requires the model to go beyond fact linking and to synthesize non-span answers.

- Can we build a new generative model that can, on one hand, **gather and synthesize disjoint pieces of information within the context**, while exploiting, on the other hand, the background commonsense knowledge from external knowledge bases?

Dataset: NarrativeQA

- Description
  - Documents: 1,572 stories (books, movie scripts) & human generated summaries
  - Questions: 46,765 human generated, based on summaries
  - Answers: human generated, based on summaries
- Motivation:
  - Answer Spans: 44.05%
  - Outside Knowledge Required: 42%
- Challenges
  - Intricate Event Timelines

    e.g., Who leads Mickey back to boxing after the HBO documentary is released?
  - Large Number of Characters

    e.g., Why did Sophia go to Russia with Alexei, instead of John?
  - Complex Structure

    e.g. Why did Mickey have reservations about his flight in Atlantic City?

Contents:

- Problem & Motivation
- Previous Approaches (SoTA)
- Contributions
- Model & Methods
- Results & Analysis
- Conclusions
- Shortcomings & Future Work

Previous approaches (SoTA)

- This model performs substantially better than previous generative models, and is competitive with current state-of-the-art span prediction models.
- **Machine Reading Comprehension:**
  - Models designed for previous tasks (Seo et al., 2017; Kadlec et al., 2016) have limited success on multi-paragraph, multi-hop inference QA datasets such as QAngaroo (Welbl et al., 2018) and NarrativeQA (Kocisky et al., 2018)
- **Commonsense/Background Knowledge:**
  - SoTA techniques such as Knowledge path extraction (Bordes et al., 2014; Bao et al., 2016) is applied in MRC-QA in this paper to extract useful commonsense knowledge paths.
- **Incorporation of External Knowledge:**
  - Using contextually-refined word embeddings which integrated information from ConceptNet via a single layer bidirectional LSTM (Weissenborn et al. (2017)) .
  - Using context-to-commonsense attention, where commonsense relations were extracted as triples (Mihaylov and Frank(2018))

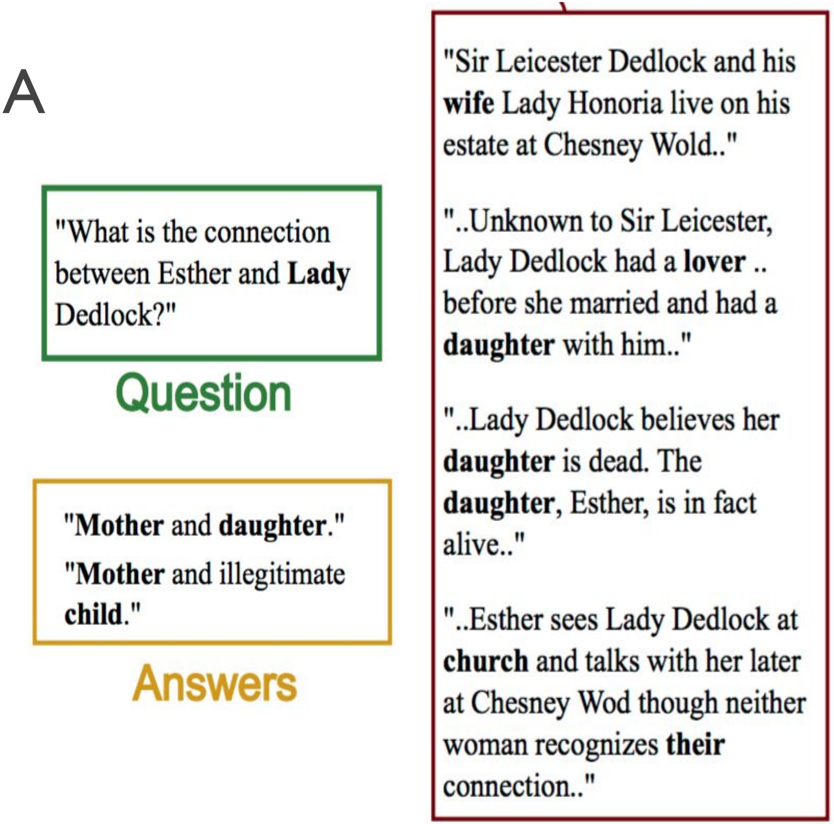
Contributions

- Baseline: Multi-Hop Pointer-Generator Model (MHPGM)
- SoTA extended model: MHPGM via the Necessary and Optional Information Cell (MHPGM-NOIC)

- Key contributions:
  - Proposed a strong baseline for generative QA task
  - Used a novel filtering algorithm to effectively find relevant subgraphs in large commonsense knowledge graphs.
  - Effectively incorporated commonsense paths into the multi-hop baseline.

Baseline Multi-Hop Pointer-Generator

- Success on Multi-Hop Reasoning QA datasets require a model to have:
  - Strong NLU capabilities
  - Abilities to extract disjoint pieces of information
  - Abilities to process long/interconnected context (**> 1000 tokens**)
  - Strong generative modeling capabilities (**non-span answers, rare words**)



"What is the connection between Esther and **Lady** Dedlock?"

Question

"**Mother** and **daughter**."

"**Mother** and illegitimate **child**."

Answers

"Sir Leicester Dedlock and his **wife** Lady Honoria live on his estate at Chesney Wold.."

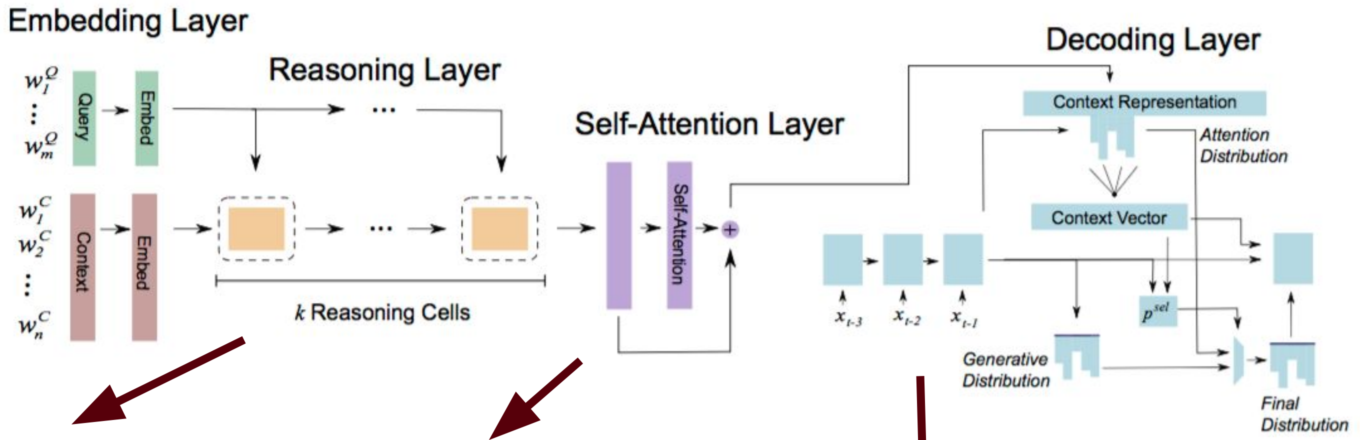
"..Unknown to Sir Leicester, Lady Dedlock had a **lover** .. before she married and had a **daughter** with him.."

"..Lady Dedlock believes her **daughter** is dead. The **daughter**, Esther, is in fact alive.."

"..Esther sees Lady Dedlock at **church** and talks with her later at Chesney Wod though neither woman recognizes **their** connection.."

Model Structures: Baseline Multi-Hop Pointer-Generator

Given context and query tokens, embed them in both task-specific learned word embedding space and Elmo pretrained context-aware embedding space.

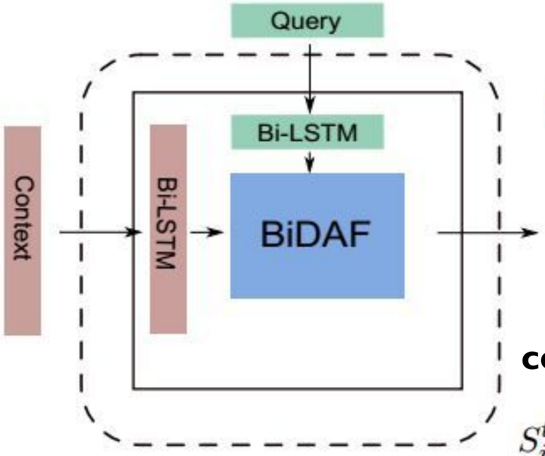


The context and query representation are fed through multiple reasoning layers, each of which represents one "hop" (reasoning step) of inference between the query and the context.

The context representation is passed through a residual self-attention layer to deal with long dependencies within the context.

A attention-pointer-generator decoder that attends on and potentially **copies** from the context is used to create the answer.

Baseline Reasoning Cell



$$\mathbf{u}^t = \text{BiLSTM}(\mathbf{c}^{t-1}); \qquad \mathbf{v}^t = \text{BiLSTM}(\mathbf{e}^Q)$$

updated context

$$\mathbf{c}_i^t = [\mathbf{u}_i^t; (\mathbf{c_q})_i^t; \mathbf{u}_i^t \odot (\mathbf{c_q})_i^t; \mathbf{q_c}^t \odot (\mathbf{c_q})_i^t]$$

context-to-query attention:

$$S_{ij}^t = W_1^t \mathbf{u}_i^t + W_2^t \mathbf{v}_j^t + W_3^t(\mathbf{u}_i^t \odot \mathbf{v}_j^t)$$

$$p_{ij}^t = \frac{\exp(S_{ij}^t)}{\sum_{k=1}^m \exp(S_{ik}^t)}$$

$$(\mathbf{c_q})_i^t = \sum_{j=1}^m p_{ij}^t \mathbf{v}_j^t$$

query-to-context attention:

$$m_i^t = \max_{1 \leq j \leq m} S_{ij}^t$$

$$p_i^t = \frac{\exp(m_i^t)}{\sum_{j=1}^n \exp(m_j^t)}$$

$$\mathbf{q_c}^t = \sum_{i=1}^n p_i^t \mathbf{u}_i^t$$

Self-Attention Layer

Obtain self attention representation c’:

$$S_{ij}^{SA} = W_4\mathbf{c}_i^{SA} + W_5\mathbf{c}_j^{SA} + W_6(\mathbf{c}_i^{SA} \odot \mathbf{c}_j^{SA})$$

$$p_{ij}^{SA} = \frac{\exp(S_{ij}^{SA})}{\sum_{k=1}^{n}\exp(S_{ik}^{SA})}$$

$$\mathbf{c'}_i = \sum_{j=1}^{n} p_{ij}^{SA}\mathbf{c}_j^{SA}$$

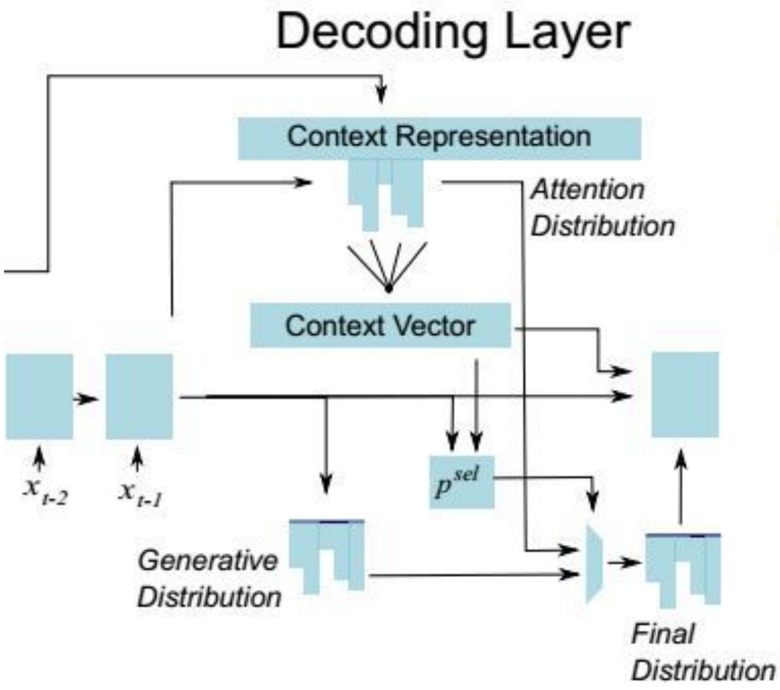
c^SA is obtained by passing the representation c^k through a fully-connected layer and then a bi-directional LSTM.

obtain residual c”:

$$\mathbf{c''} = \mathrm{BiLSTM}([\mathbf{c'}; \mathbf{c}^{SA}; \mathbf{c'} \odot \mathbf{c}^{SA}$$

obtain the encoded context c = ck + c”

Pointer-Generator Decoding Layer



hidden state st:

$$\mathbf{s}_t = \mathrm{LSTM}([\mathbf{x}_t; \mathbf{a}_{t-1}], \mathbf{s}_{t-1})$$

$$P_{gen} = \mathrm{softmax}(W_{gen}\mathbf{s}_t + \mathbf{b}_{gen})$$

$$\alpha_i = \mathbf{v}^\intercal \tanh(W_c\mathbf{c}_i + W_s\mathbf{s}_t + b_{attn})$$

$$\hat{\alpha}_i = \frac{\exp(\alpha_i)}{\sum_{j=1}^n \exp(\alpha_j)}$$

$$\mathbf{a}_t = \sum_{i=1}^n \hat{\alpha}_i \mathbf{c}_i$$

calculate select distribution:

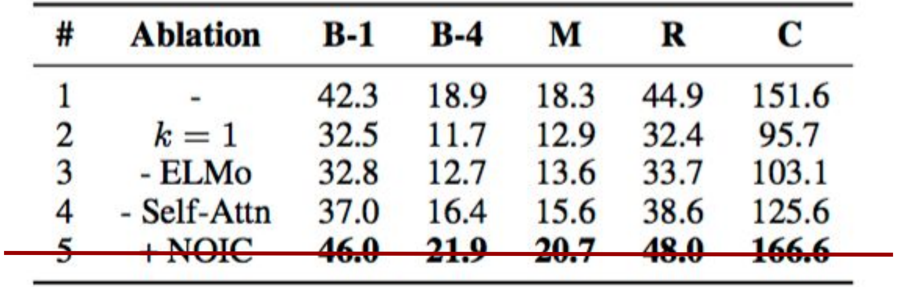
$$\mathbf{o} = \sigma(W_a\mathbf{a}_t + W_x\mathbf{x}_t + W_s\mathbf{s}_t + b_{ptr})$$

$$\mathbf{p}^{sel} = \mathrm{softmax}(\mathbf{o})$$

final output distribution:

$$P_t(w) = p_1^{sel} P_{gen}(w) + p_2^{sel} \sum_{i:w_i^C = w} \hat{\alpha}_i$$

Results: Model Ablations



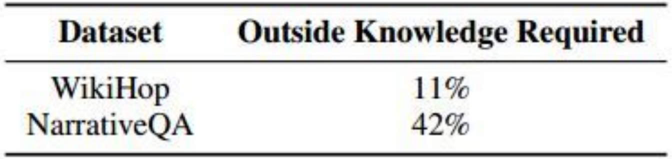
#	Ablation	B-1	B-4	M	R	C
1	-	42.3	18.9	18.3	44.9	151.6
2	$k = 1$	32.5	11.7	12.9	32.4	95.7
3	- ELMo	32.8	12.7	13.6	33.7	103.1
4	- Self-Attn	37.0	16.4	15.6	38.6	125.6
~~5~~	~~+ NOIC~~	~~**46.0**~~	~~**21.9**~~	~~**20.7**~~	~~**48.0**~~	~~**166.6**~~

According to the baseline ablation results:

-- **multi-hop architecture** is most important to the baseline model performance.

-- Other components such as Elmo, residual self-attention are important as well.

How can we make it better?



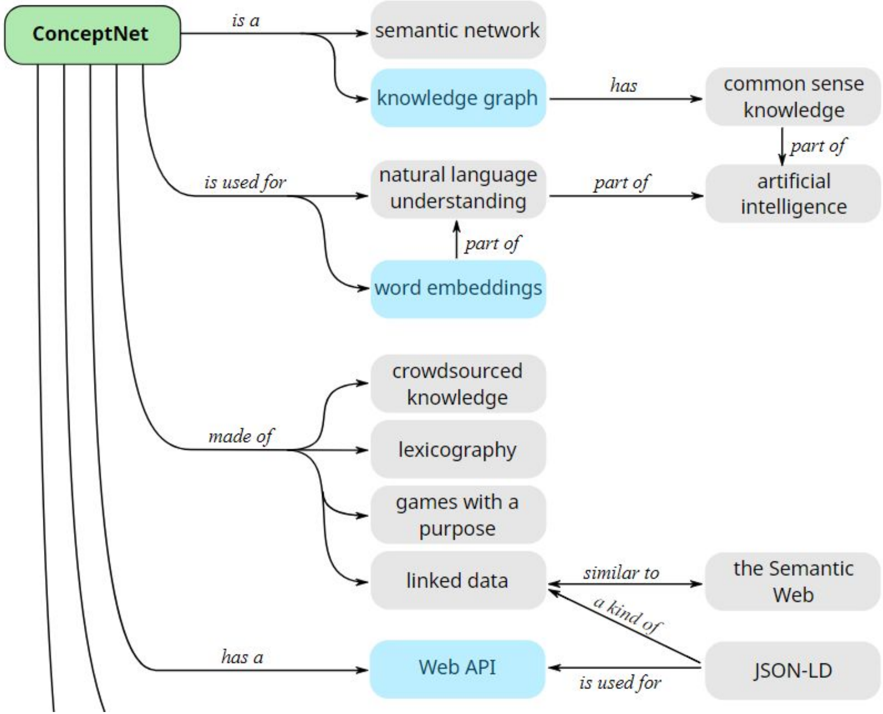
Dataset	Outside Knowledge Required
WikiHop	11%
NarrativeQA	42%

Commonsense Requirements

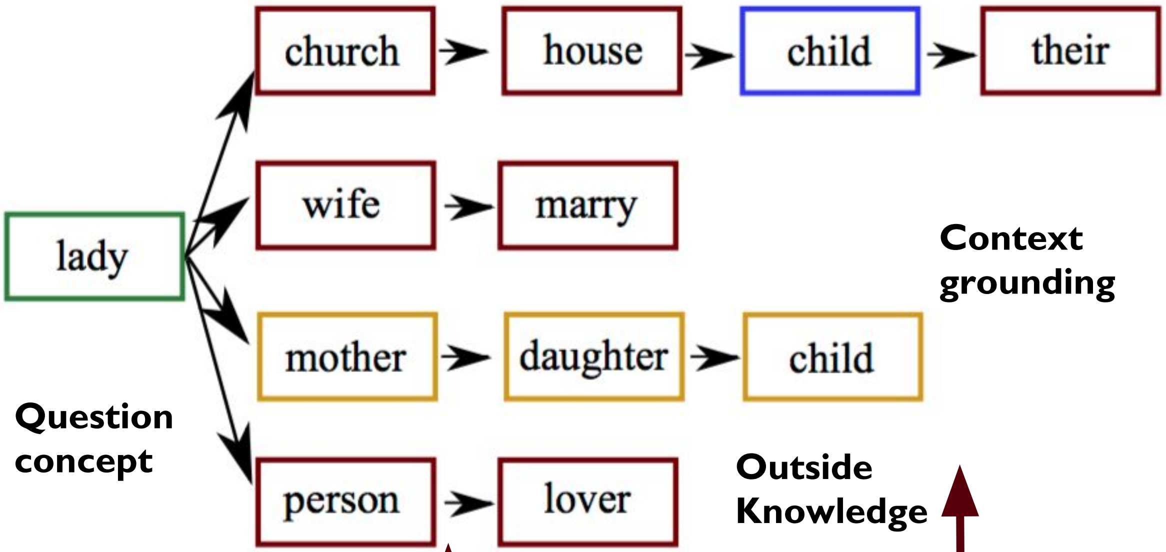
- Success on Multi-Hop Reasoning QA datasets require a model to have:
  - Strong NLU capailities
  - Abilities to extract disjoint pieces of information
  - Abilities to process long/interconnected context (**> 1000 tokens**)
  - Strong generative modeling capabilities (**non-span answers, rare words**)
  - Reason with implicit relations not mentioned in the context

ConceptNet

- A knowledge graph of semantic relations between concepts
- Has 28 million edges
- Each edge represents one of 37 types of semantic relationship, e.g. UsedFor, FormOf, CapableOf, etc.



Tree Construction



Question concept

Context grounding

Outside Knowledge

Pointwise mutual information

Direct interaction

Multi-hop

term frequency → softmax normalization

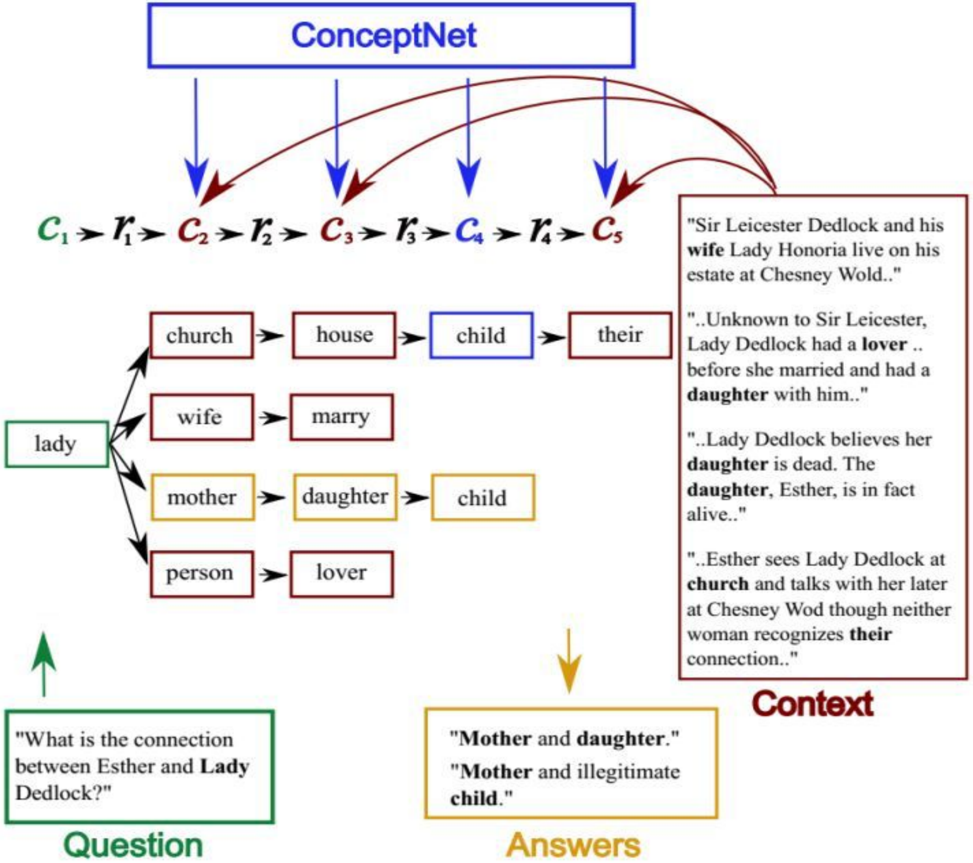
Initial Scoring

Cumulative Scoring

Path Selection

Output: Optimal Path!

Commonsense Selection Approach



"Sir Leicester Dedlock and his **wife** Lady Honoria live on his estate at Chesney Wold.."

"..Unknown to Sir Leicester, Lady Dedlock had a **lover** .. before she married and had a **daughter** with him.."

"..Lady Dedlock believes her **daughter** is dead. The **daughter**, Esther, is in fact alive.."

"..Esther sees Lady Dedlock at **church** and talks with her later at Chesney Wod though neither woman recognizes **their** connection.."

Context

"What is the connection between Esther and **Lady** Dedlock?"

Question

"**Mother** and **daughter**."

"**Mother** and illegitimate **child**."

Answers

Penn Engineering

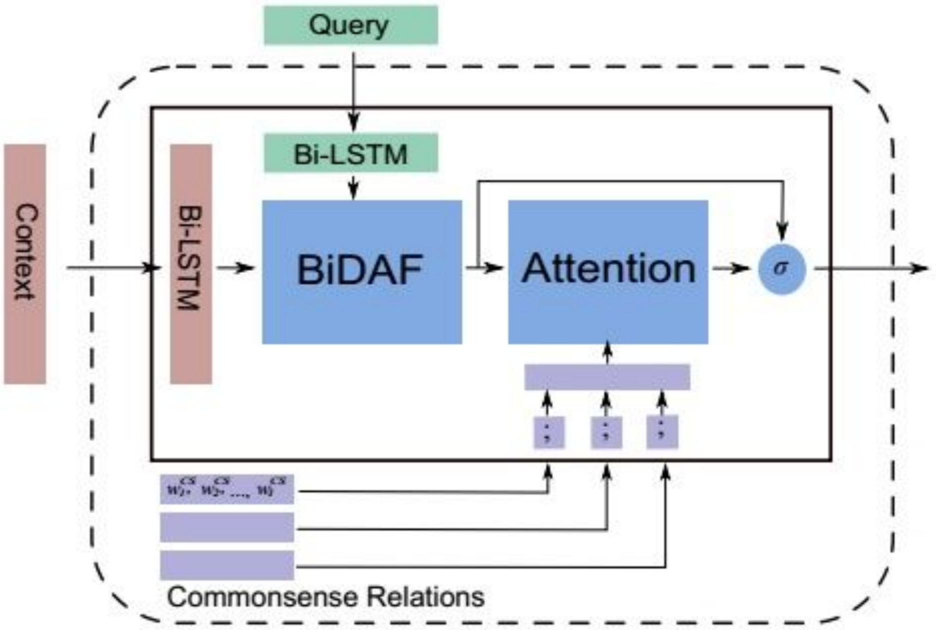
Commonsense Incorporation

Effective incorporation of commonsense information requires:

- Multi-hop, selective commonsense incorporation
- Ability to “denoise” and ignore unnecessary commonsense

Necessary and Optional Information Cell (NOIC) incorporates optional commonsense information via a gated-attention layer.

NOI Cell



NOIC: baseline reasoning cell + extracted commonsense path

1. Create Layer-specific representation for each of the extracted path via projection layer.
2. Use attention mechanism to model the interaction between context representation and extracted commonsense information.
3. Add a sigmoid gate that allows the model to select which commonsense info to include.

Summary on Experiments

Datasets:

generative NarrativeQA and extractive QAngaroo WikiHop.

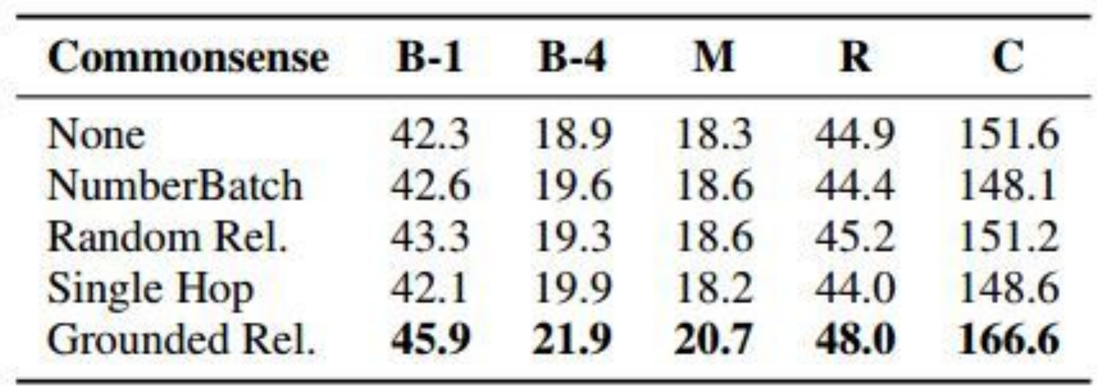
Evaluation Metrics:

NarrativeQA: Bleu-1, Bleu-4, METEOR, Rouge- L, and CIDEr which emphasizes annotator consensus
WikiHop: accuracy

Experiments:

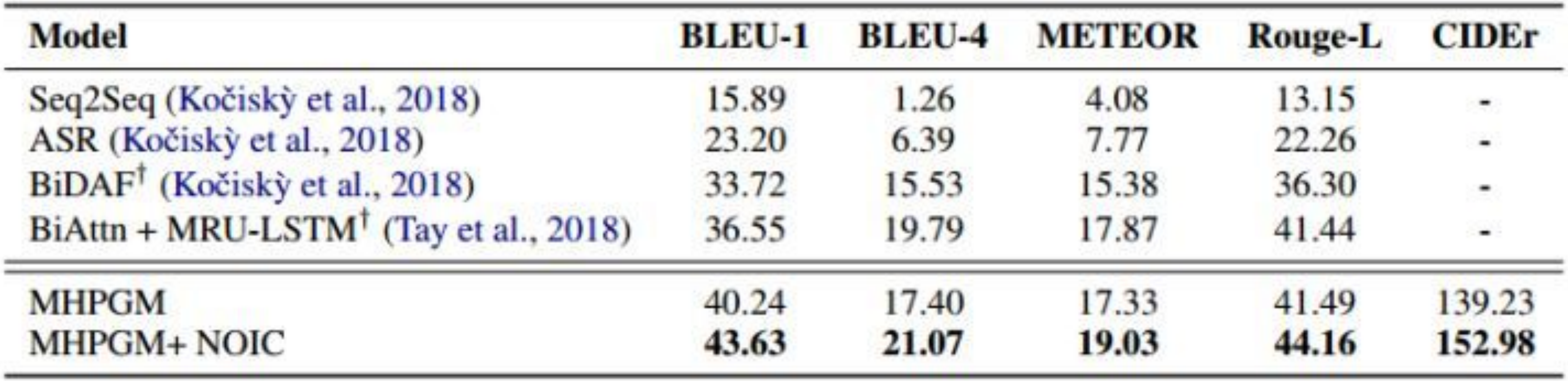
Main Experiment: Testing model performance on both NarrativeQA and WikiHop with and without commonsense incorporation
Model Ablations: Testing effectiveness of each component of the architecture
Commonsense Ablations: Testing effectiveness of commonsense selection and incorporation techniques.
Human Evaluations on model performance
Human Evaluations on commonsense selection

Results: Commonsense Ablations



Commonsense	**B-1**	**B-4**	**M**	**R**	**C**
None	42.3	18.9	18.3	44.9	151.6
NumberBatch	42.6	19.6	18.6	44.4	148.1
Random Rel.	43.3	19.3	18.6	45.2	151.2
Single Hop	42.1	19.9	18.2	44.0	148.6
Grounded Rel.	**45.9**	**21.9**	**20.7**	**48.0**	**166.6**

Results: on NarrativeQA

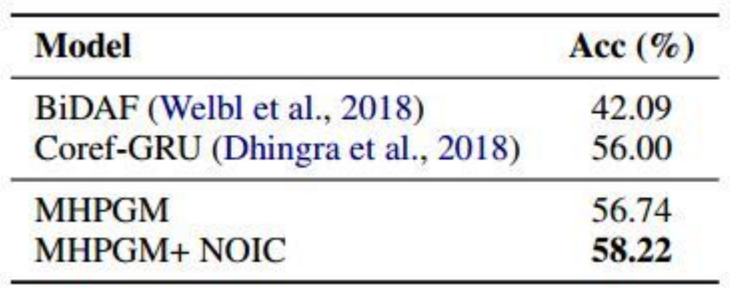


Model	BLEU-1	BLEU-4	METEOR	Rouge-L	CIDEr
Seq2Seq (Kočiskỳ et al., 2018)	15.89	1.26	4.08	13.15	-
ASR (Kočiskỳ et al., 2018)	23.20	6.39	7.77	22.26	-
BiDAF† (Kočiskỳ et al., 2018)	33.72	15.53	15.38	36.30	-
BiAttn + MRU-LSTM† (Tay et al., 2018)	36.55	19.79	17.87	41.44	-
MHPGM	40.24	17.40	17.33	41.49	139.23
MHPGM+ NOIC	**43.63**	**21.07**	**19.03**	**44.16**	**152.98**

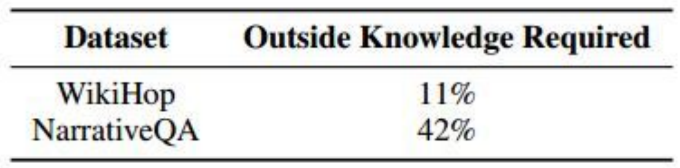
† indicates span prediction models trained on the Rouge-L retrieval oracle.

$p < 0.001$ on all metrics: Stat. significance computed using bootstrap test with 100K iterations (Noreen, 1989; Efron and Tibshirani, 1994).

Results: on WikiHop



Model	Acc (%)
BiDAF (Welbl et al., 2018)	42.09
Coref-GRU (Dhingra et al., 2018)	56.00
MHPGM	56.74
MHPGM+ NOIC	**58.22**



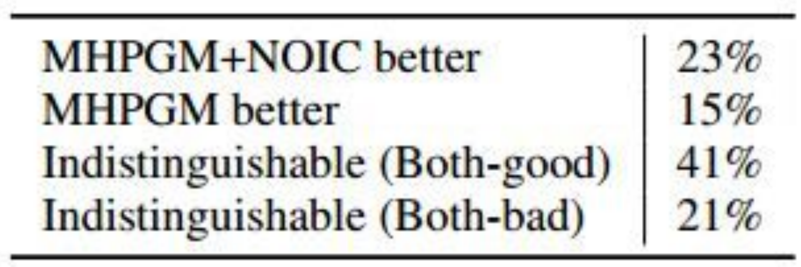
Dataset	Outside Knowledge Required
WikiHop	11%
NarrativeQA	42%

WikiHop is multi-hop QA dataset which diverge from NarrativeQA in that:

-- Only 11% of examples need outside knowledge as opposed to 42% on NQA

-- Needs more fact-based commonsense (Freebase) instead of semantic-based ones (ConceptNet)

Results: Human Evaluation (Model)

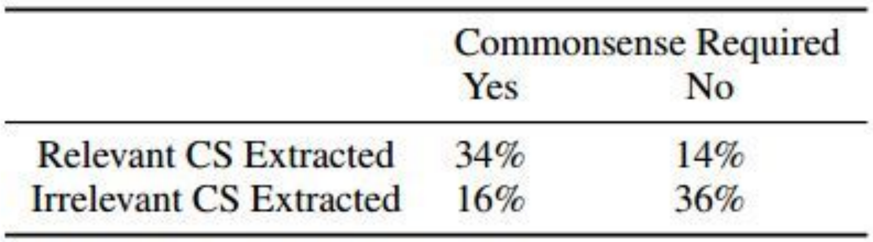


MHPGM+NOIC better	23%
MHPGM better	15%
Indistinguishable (Both-good)	41%
Indistinguishable (Both-bad)	21%

They randomly selected 100 examples from the NarrativeQA test set, along with both models' predicted answers, and for each datapoint, they asked 3 external human evaluators to decide if one is strictly better than the other, or that they were similar in quality (both-good or both-bad).

Fleiss κ = 0.831, indicating 'almost-perfect' agreement between the annotators (Landis and Koch, 1977).

Results: Human Evaluation (CS Extraction)



	Commonsense Required	
	Yes	No
Relevant CS Extracted	34%	14%
Irrelevant CS Extracted	16%	36%

Goal: To check the effectiveness of the commonsense selection algorithm.

Experiment Data: 50 sample subset of the NarrativeQA test set

How: given a context-query pair, and the commonsense selected by the algorithm, two independent evaluations were conducted: (1) was any external commonsense knowledge necessary for answering the question?; (2) were the commonsense relations provided by our algorithm relevant to the question?

Conclusions

In this work, the authors:

- Proposed a strong multi-hop baseline for generative QA task
- Used PMI/TF based filtering algorithm to effectively query large knowledge graphs for relevant subgraphs.
- Effectively incorporated commonsense paths into the multi-hop baseline via multiple hops of selectively gated attention.

Shortcomings & Future Work

- Other datasets may need more fact-based commonsense instead of semantic-based ones.
- Information loss in the modules

- Explore adding different types of commonsense to other domains.
- Explore the possibility of adding graph-based attention to more directly incorporate semantic networks.

Thank you!

Model Structures: Baseline Multi-Hop Pointer-Generator

- **Embedding layer:**
  Given context and query tokens, embed them in both task-specific learned word embedding space and Elmo pretrained context-aware embedding space.

- **Reasoning Layer:**
  The context and query representation are then fed through multiple reasoning layers, each of which represents one “hop” (reasoning step) of inference between the query and the context.

- **Self-Attention Layer:**
  After this, the context representation is passed through a residual self-attention layer to deal with long dependencies within the context

- **Pointer-Generator Decoding Layer:**
  This is finally passed into a pointer-generator decoder that not only allows the model to generate non-span answers, but also allows models to pull context-specific words directly if necessary.

Type of Commonsense

- Taxonomy

  e.g. Physical disorders include insomnia

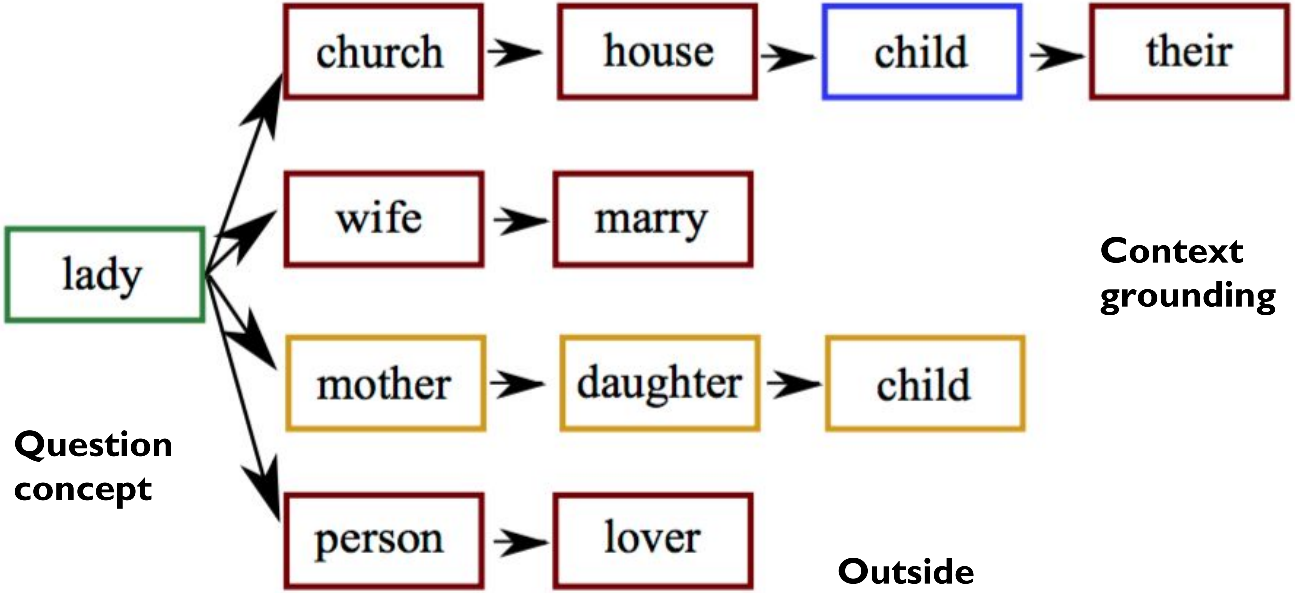
- Cause and Effect

  e.g. Take an offer → take a position in

- Colloquialisms

  e.g. make ends meet = pay for necessities

Tree Construction



Context grounding

Question concept

Outside Knowledge

Direct interaction

Multi-hop