Commonsense for Generative Multi-Hop Question Answering Tasks

EMNLP2018

UNC Chapel Hill（北卡罗来纳大学教堂山分校）

Lisa Bauer* Yicheng Wang* Mohit Bansal

Xiachong Feng

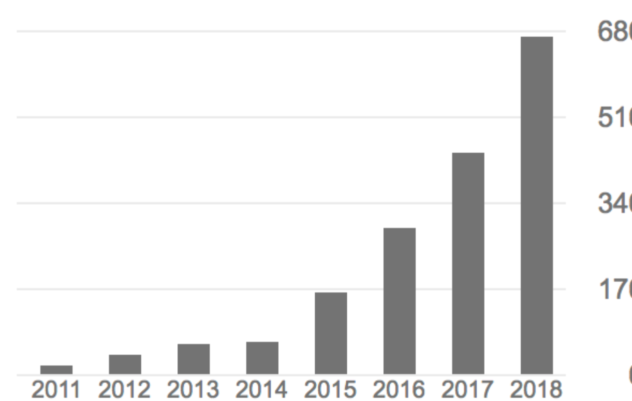
Author

Lisa Bauer

- Second year Ph.D. UNC Chapel Hill
- B.A. Johns Hopkins University
- natural language generation、 QA
- Dialogue、 deep reasoning
- knowledge-based inference

Mohit Bansal

- Director of the UNC-NLP Lab
- Assistant Professor
- Ph.D. from the
  UC Berkeley



Commonsense for Generative Multi-Hop Question Answering Tasks

QA Dataset

- **Task**
  - **Machine reading comprehension (MRC) based QA**, asking it to answer a question based on a passage of relevant content.
- **Dataset**
  - **bAbI：**smaller lexicons and simpler passage structures
  - **CNN/DM、SQuAD：**fact-based、answer extraction、select a context span
  - **Qangaroo(WikiHop):** extractive dataset、multi-hop reasoning

Mary moved to the bathroom. John went to the hallway. Daniel went back to the hallway. Sandra moved to the garden. John moved to the office. Sandra journeyed to the bathroom. Mary moved to the hallway. Daniel travelled to the office. John went back to the garden. John moved to the bedroom.,

Question → Where is Sandra?, Answer → bathroom|⟩

bAbI

QA Dataset

- **Dataset**
  - **NarrativeQA** generative dataset
  - includes fictional stories, which are **1,567 complete stories** from **books and movie scripts**, with human written questions and answers based solely on human-generated abstract summaries.
  - There are **46,765 pairs of answers to questions** written by humans and includes mostly the more complicated variety of questions such as “when / where / who / why”.
  - Requiring **multi-hop reasoning** for long, complex stories
- **Experiment**
  - Qangaroo: extractive dataset、multi-hop reasoning
  - NarrativeQA: generative dataset、multi-hop reasoning

Commonsense Dataset

- **ConceptNet**
  - Large-scale graphical commonsense databases

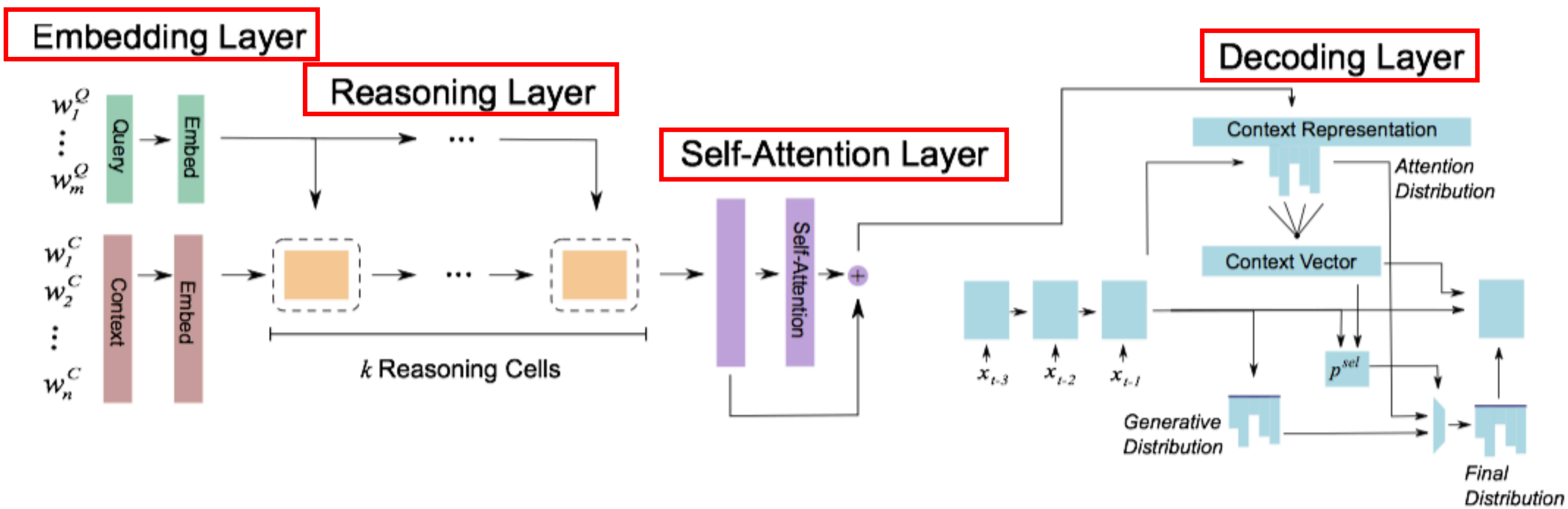
Task

- **generative QA**
  - Input:
    - Context $X^C = \{w_1^C, w_2^C, \ldots, w_n^C\}$
    - Query $X^Q = \{w_1^Q, w_2^Q, \ldots, w_m^Q\}$
  - Output：
    - series of answer tokens：$X^a = \{w_1^a, w_2^a, \ldots, w_p^a\}$

Model overview

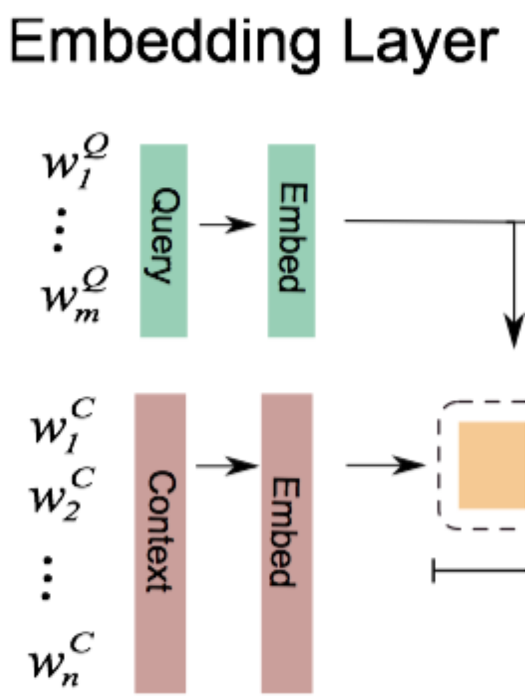
- **Multi-Hop Pointer-Generator Model (MHPGM)**
  - baseline model
  - Baseline reasoning cell
  - multiple hops of bidirectional attention
  - self-attention
  - pointer-generator decoder
- **Necessary and Optional Information Cell (NOIC)**
  - NOIC Reasoning Cell
  - **Choose knowledge**
    - pointwise mutual information (PMI)
    - term-frequency-based scoring function
  - **Insert knowledge**
    - Selectively gated attention mechanism

Multi-Hop Pointer-Generator Model



Embedding Layer

- learned embedding space of dimension d
- pretrained embedding from language models (ELMo)
- The embedded representation for each word in the context or question：

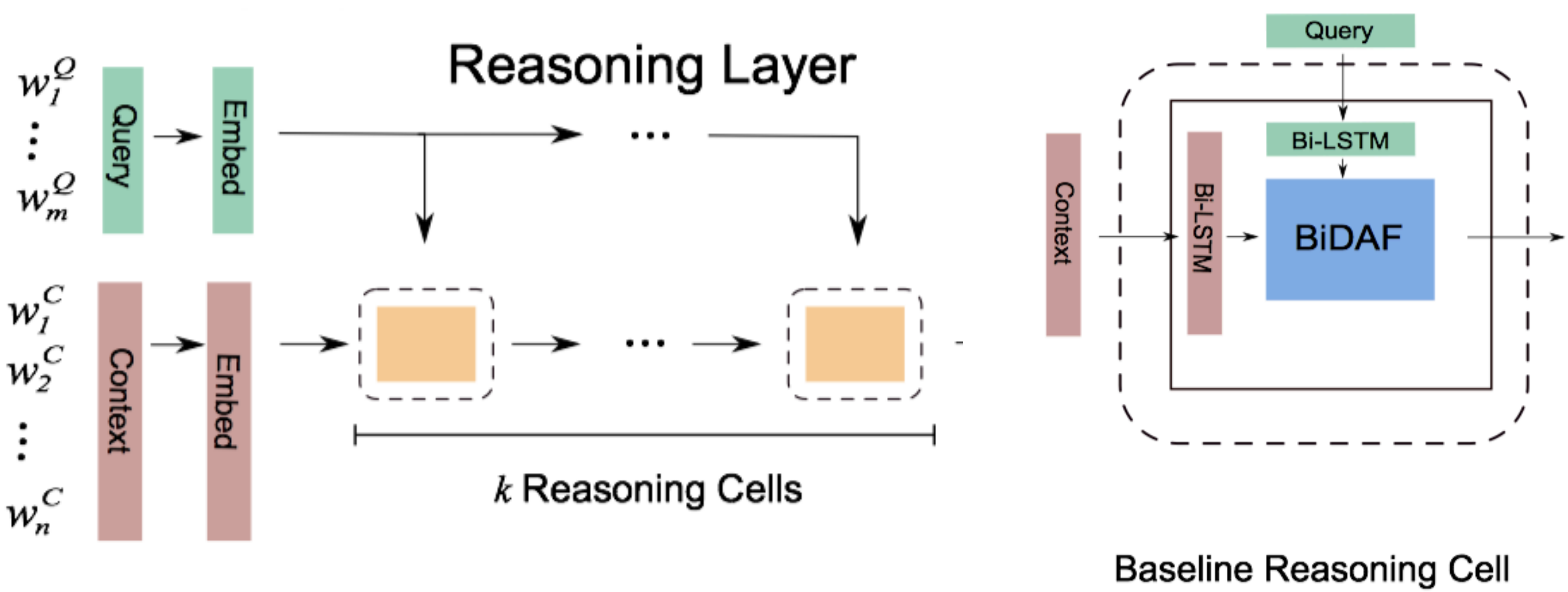


$$\mathbf{e}_i^C \text{ or } \mathbf{e}_i^Q \in \mathbb{R}^{d+1024}$$

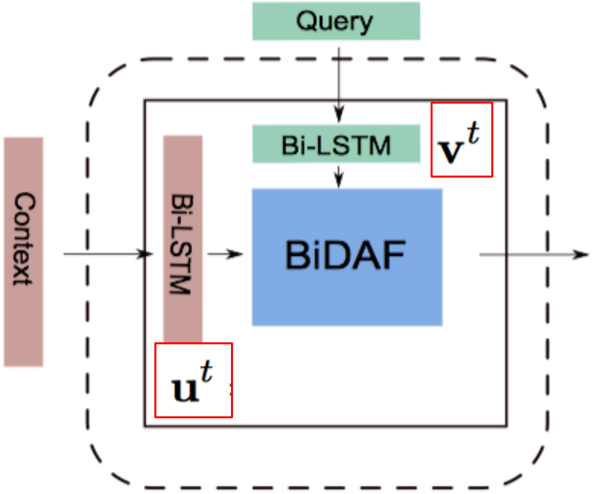
Reasoning layer

- k reasoning cells
- The t^{th} reasoning cell's inputs are the previous step's output $(\{\mathbf{c}_i^{t-1}\}_{i=1}^n)$ and the embedded question $(\{\mathbf{e}_i^Q\}_{i=1}^m)$
- First creates step-specific context and query encodings via cell-specific bidirectional LSTMs:

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



Reasoning layer



- Use **bidirectional attention** to emulate a hop of resoning by focusing on relevant aspects of the context.
- **Context-to-query attention**

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

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

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

About Query

- **Query-to-context attention**

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

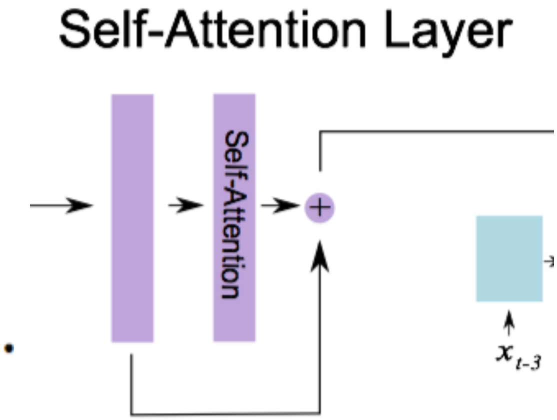
$$p_i^t = \frac{\exp(m_i^t)}{\sum_{j=1}^n \exp(m_j^t)}$$

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

About Context

- **Final**

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



Self-Attention Layer

- Residual static self-attention mechanism
- Input： output of the last reasoning cell $\mathbf{c}^k$.
  1. fully-connected layer
  2. a bi-directional LSTM $\mathbf{c}^{SA}$.
- Self attention representation

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

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

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

- Output of the self-attention layer is generated by another layer of bidirectional LSTM.

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

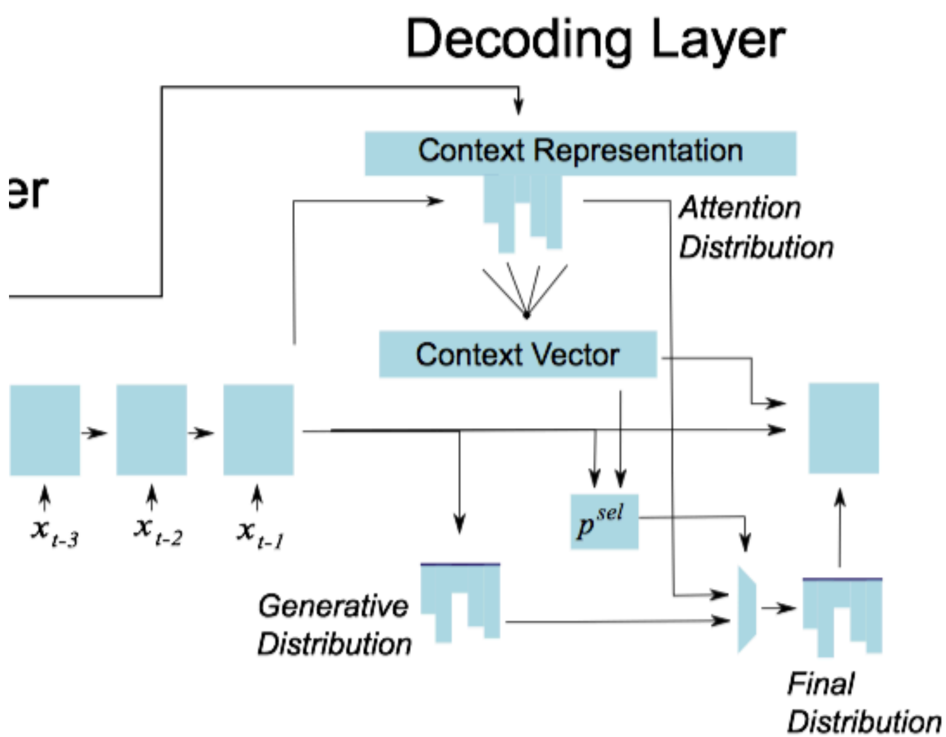
- Final encoded context:

$$\mathbf{c} = \mathbf{c}^k + \mathbf{c}''.$$

Pointer-Generator Decoding Layer

- embedded representation of last timestep's output $\mathbf{x}_t$
- the last time step's hidden state $\mathbf{s}_{t-1}$
- context vector $\mathbf{a}_{t-1}$

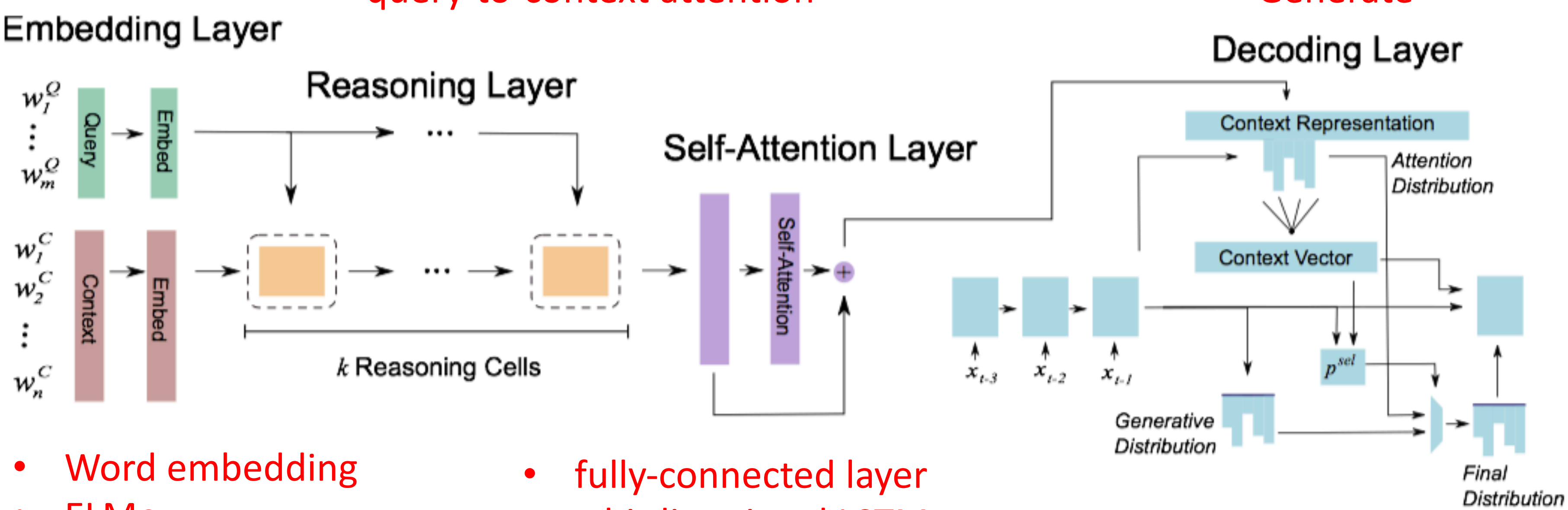
$$\mathbf{s}_t = \mathrm{LSTM}([\mathbf{x}_t; \mathbf{a}_{t-1}], \mathbf{s}_{t-1})$$



Multi-Hop Pointer-Generator Model

- BiDAF
- cell-specific bidirectional LSTMs
- context-to-query attention
- query-to-context attention

- Attention
- Copy
- Generate



- Word embedding
- ELMo

- fully-connected layer
- a bi-directional LSTM
- Self attention
- a bi-directional LSTM
- residually

Commonsense Selection Representation

- QA tasks often needs knowledge of relations not directly stated in the context

Dataset	Outside Knowledge Required
WikiHop	11%
NarrativeQA	42%

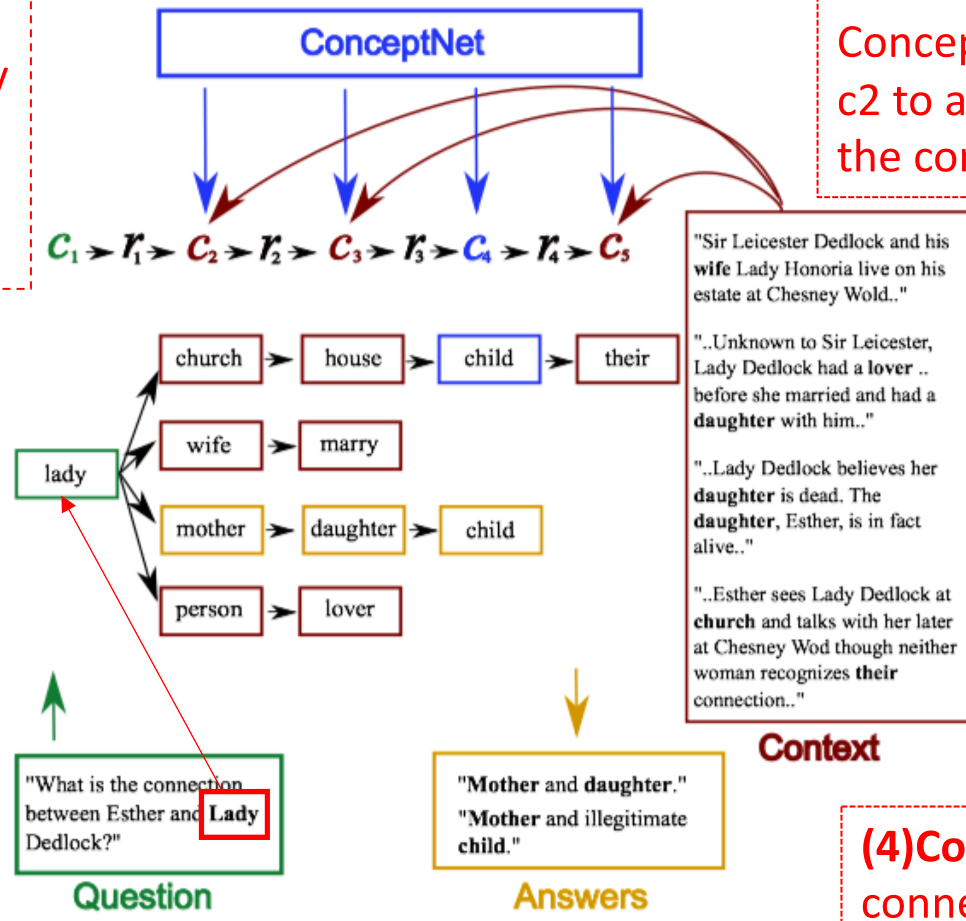
- **Key idea**
  - Introducing useful connections between **concepts** in the **context** and **question** via **ConceptNet**
  1. collect potentially relevant concepts via a **tree construction method**
  2. rank and filter these paths to ensure both the quality and variety of added via a **3-step scoring strategy**

Tree Construction

(1)Direct Interaction
select relations r1 from ConceptNet that directly link c1 to a concept within the context c2 ∈ C

(2)Multi-Hop
select relations in ConceptNet r2 that link c2 to another concept in the context, c3 ∈ C.

For each concept c1 in the question



(3)Outside Knowledge
an unconstrained hop into c3 's neighbors in ConceptNet

(4)Context-Grounding
connecting c4 to c5 ∈ C

Example

Question	What shore does Michael's **corpse** wash **up** on?
Context	"..as the play begins nora and cathleen receive word from the priest that a **body**, that may be their brother michael, has washed up on shore in donegal, the island farthest **north** of their home island of inishmaan.."
Answers	the shore of donegal / donegal
	up → RelatedTo → wind → Antonym → her → RelatedTo → person up → RelatedTo → north → RelatedTo → up wash → RelatedTo → up up → Antonym → down wash → RelatedTo → water → PartOf → sea → RelatedTo → fish up → RelatedTo → wind wash → RelatedTo → water → PartOf → sea shore → RelatedTo → sea wash → RelatedTo → body wash → Antonym → making up → Antonym → down → Antonym → up wash → RelatedTo → water → PartOf → sea → MadeOf → water up → RelatedTo → wind → Antonym → her wash → RelatedTo → water up → RelatedTo → south

Rank and Filter(3-step scoring method)

- **Initial Node Scoring**
  - **For c2、c3、c5**
    - Term frequency
    - Heuristic: important concepts occur more frequently
      $$\text{score}(c) = \text{count}(c)/|C|$$
    - |C| is the **context length** and count() is the **number of times a concept appears in the con**text.
  - **For c4**
    - want c4 to be a logically consistent next step in reasoning following the path of c1 to c3
    - Heuristic: logically consistent paths occur more frequently
    - Pointwise Mutual Information (PMI)

Rank and Filter(3-step scoring method)

- **Initial Node Scoring**
  - **For c4**
    - Pointwise Mutual Information (PMI)

$$\mathrm{PMI}(c_4, c_{1-3}) = \log(\mathbb{P}(c_4, c_{1-3})/\mathbb{P}(c_4)\mathbb{P}(c_{1-3}))$$

$$\mathbb{P}(c_4, c_{1-3}) = \frac{\text{\# of paths connecting } c_1, c_2, c_3, c_4}{\text{\# of distinct paths of length 4}}$$

$$\mathbb{P}(c_4) = \frac{\text{\# of nodes that can reach } c_4}{|\text{ConceptNet}|}$$

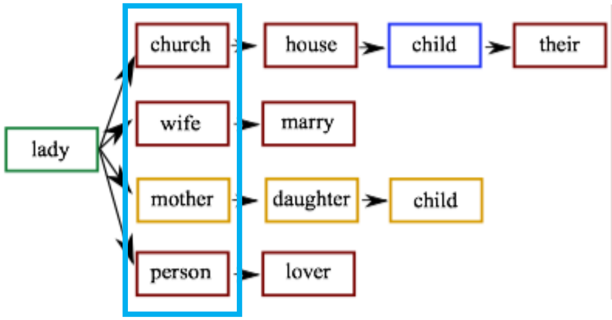
$$\mathbb{P}(c_{1-3}) = \frac{\text{\# of paths connecting } c_1, c_2, c_3}{\text{\# of paths of length 3}}$$

    - normalized PMI (NPMI)

$$\mathrm{score}(c_4) = \mathrm{PMI}(c_4, c_{1-3})/(-\log \mathbb{P}(c_4, c_{1-3})).$$

- **Normalize each node's score against its siblings**

$$\text{n-score}(c) = \mathrm{softmax}_{\mathrm{siblings}(c)}(\mathrm{score}(c)).$$



Rank and Filter(3-step scoring method)

- **Cumulative Node Scoring**
  - **re-score** each node based not only on its relevance and saliency but also that **of its tree descendants**.
  - **When at the leaf nodes**
    - c-score = n-score
  - **for cl not a leaf node**
    - c-score(cl) = n-score(cl) + f(cl)
    - f of a node is the average of the c-scores of its top 2 highest scoring children

lady → **mother** → daughter(high)
→ married(high)
→ book(low)

example

Rank and Filter(3-step scoring method)

1. Starting at the root
2. recursively take two of its children with the highest cumulative scores
3. until reach a leaf

Final: directly give these paths to the model as **sequences of tokens.**

Commonsense Model Incorporation

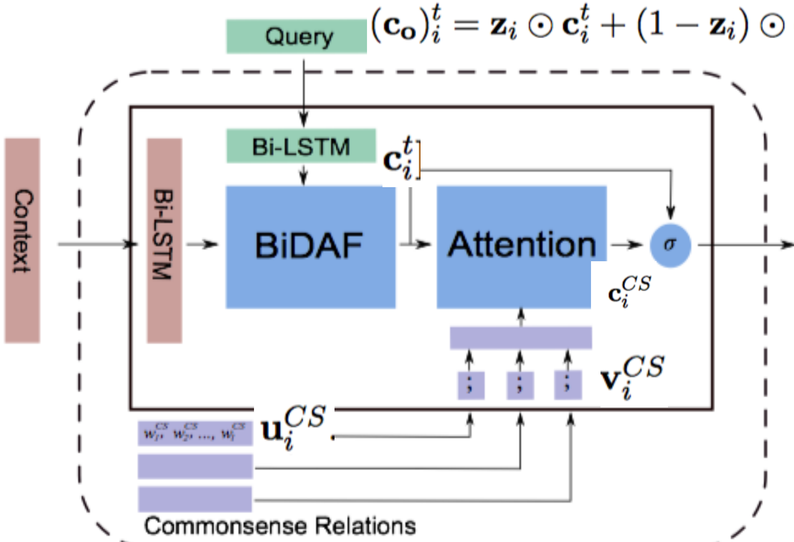
- **Given:**
  - list of commonsense logic paths as sequences of words

$$X^{CS} = \{w_1^{CS}, w_2^{CS}, \ldots, w_l^{CS}\}$$

  - **Example**: <lady, AtLocation, church, RelatedTo, house, RelatedTo, child, RelatedTo, their>
- **Necessary and Optional Information Cell (NOIC)**

$$\mathbf{z}_i = \sigma(W_z[\mathbf{c}_i^{CS}; \mathbf{c}_i^t] + b_z)$$

$$(\mathbf{c_o})_i^t = \mathbf{z}_i \odot \mathbf{c}_i^t + (1 - \mathbf{z}_i) \odot \mathbf{c}_i^{CS}$$



NOIC Reasoning Cell

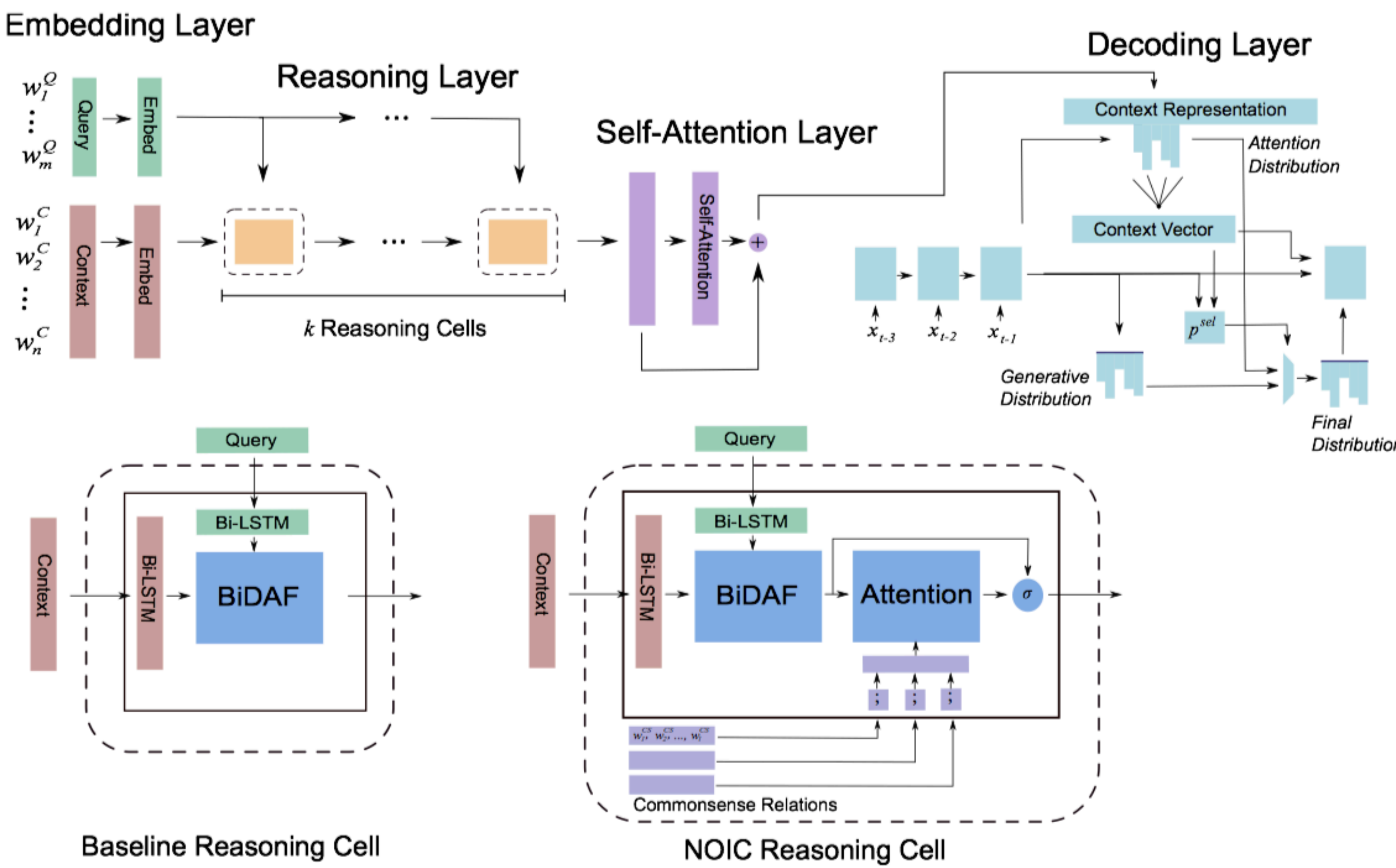
- concatenating the embedded commonsense $\mathbf{u}_i^{CS}$.
- project it to the same dimension as $\mathbf{v}_i^{CS}$
- attention between commonsense and the context

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

$$p_{ij}^{CS} = \frac{\exp(S_{ij}^{CS})}{\sum_{k=1}^{l} \exp(S_{ij}^{CS})}$$

$$\mathbf{c}_i^{CS} = \sum_{j=1}^{l} p_{ij}^{CS}\mathbf{v}_j^{CS}$$

Total Model



Embedding Layer
Reasoning Layer
Self-Attention Layer
Decoding Layer
Query
Embed
Context
Embed
k Reasoning Cells
Self-Attention
Context Representation
Attention Distribution
Context Vector
Generative Distribution
Final Distribution
Query
Bi-LSTM
Context
Bi-LSTM
BiDAF
Query
Bi-LSTM
Context
Bi-LSTM
BiDAF
Attention
Commonsense Relations
Baseline Reasoning Cell
NOIC Reasoning Cell

Experiment

- **Dataset**
  - generative NarrativeQA
  - extractive QAngaroo WikiHop
    - For multiple-choice WikiHop, we rank candidate responses by their generation probability.
- **Metric**
  - **NarrativeQA**
    - Bleu-1、Bleu-4 、METEOR 、RougeL 、CIDEr
  - **WikiHop**
    - Accuracy

Result

- **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**

- **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**

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**

Table 4: Model ablations on NarrativeQA val-set.

Commonsense Ablations

- **NumberBatch** :naively add ConceptNet information by initializing the word embeddings with the ConceptNet-trained embeddings
- **In-domain noise** :giving each context-query pair a set of random relations grounded in other context-query pairs
- Using a **single hop** from the query to the context.

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**

Table 5: Commonsense ablations on NarrativeQA val-set.

Human Evaluation Analysis

- **Commonsense Selection**

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

Table 6: NarrativeQA's commonsense requirements and effectiveness of commonsense selection algorithm.

- **Model Performance**

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

Table 7: Human evaluation on the output quality of the MHPGM+NOIC vs. MHPGM in terms of correctness.

Conclusion

- **Effective reasoning-generative QA architecture**
  1. multiple hops of bidirectional attention and a pointer-generator decoder
  2. select grounded, useful paths of commonsense knowledge
  3. Necessary and Optional Information Cell (NOIC)
- **New state-of-the-art on NarrativeQA.**

Thank you!