多跳阅读

汇报人：陶思雨

Multi-hop

Figure 1: An example of the multi-hop questions in HOTPOTQA. We also highlight the supporting facts in *blue italics*, which are also part of the dataset.

Two Benchmark Settings:

distractor: challenge the model to find the true supporting facts in the presence of noise

fullwiki: fully test the model's ability to locate relevant facts as well as reasoning about them

Paragraph A, Return to Olympus:

[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:

[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

Multi-hop

Leaderboard (Distractor Setting)

In the *distractor* setting, a question-answering system reads 10 paragraphs to provide an answer (Ans) to a question. They must also justify these answers with supporting facts (Sup).

	Model	Code	Ans		Sup		Joint	
			EM	F_1	EM	F_1	EM	F_1
1 Dec 1, 2019	HGN-large (single model) *Anonymous*		**69.22**	**82.19**	**62.76**	**88.47**	**47.11**	**74.21**
2 Oct 18, 2019	C2F Reader (single model) *Joint Laboratory of HIT and iFLYTEK Research*		67.98	81.24	60.81	87.63	44.67	72.73
3 Nov 19, 2019	SAE-large (single model) *JD AI Research* Tu, Huang et al., AAAI 2020		66.92	79.62	61.53	86.86	45.36	71.45
4 Sep 27, 2019	HGN (single model) *Microsoft Dynamics 365 AI Research* Fang et al., 2019		66.07	79.36	60.33	87.33	43.57	71.03

$$P^{(\text{joint})} = P^{(\text{ans})} P^{(\text{sup})}, \quad R^{(\text{joint})} = R^{(\text{ans})} R^{(\text{sup})}$$

$$\text{Joint } F_1 = \frac{2P^{(\text{joint})} R^{(\text{joint})}}{P^{(\text{joint})} + R^{(\text{joint})}}.$$

Bridge-IR

EMNLP2019 **Simple yet Effective Bridge Reasoning for Open-Domain Multi-Hop Question Answering**

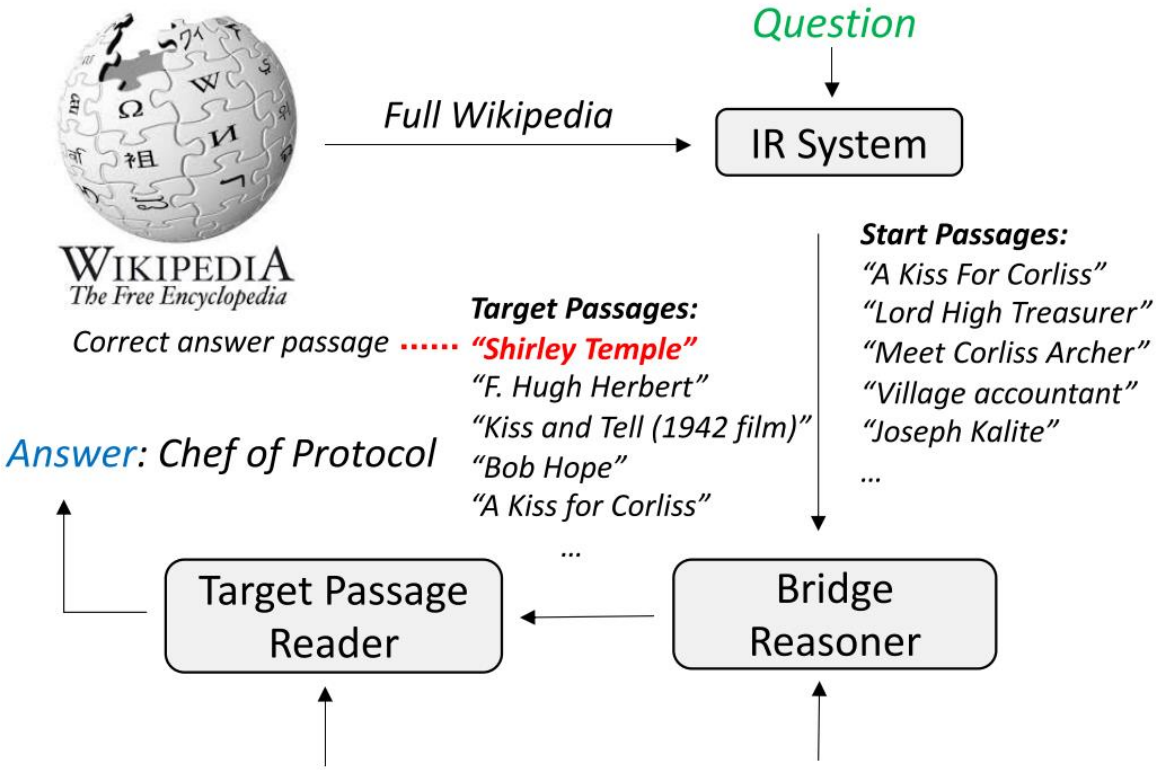
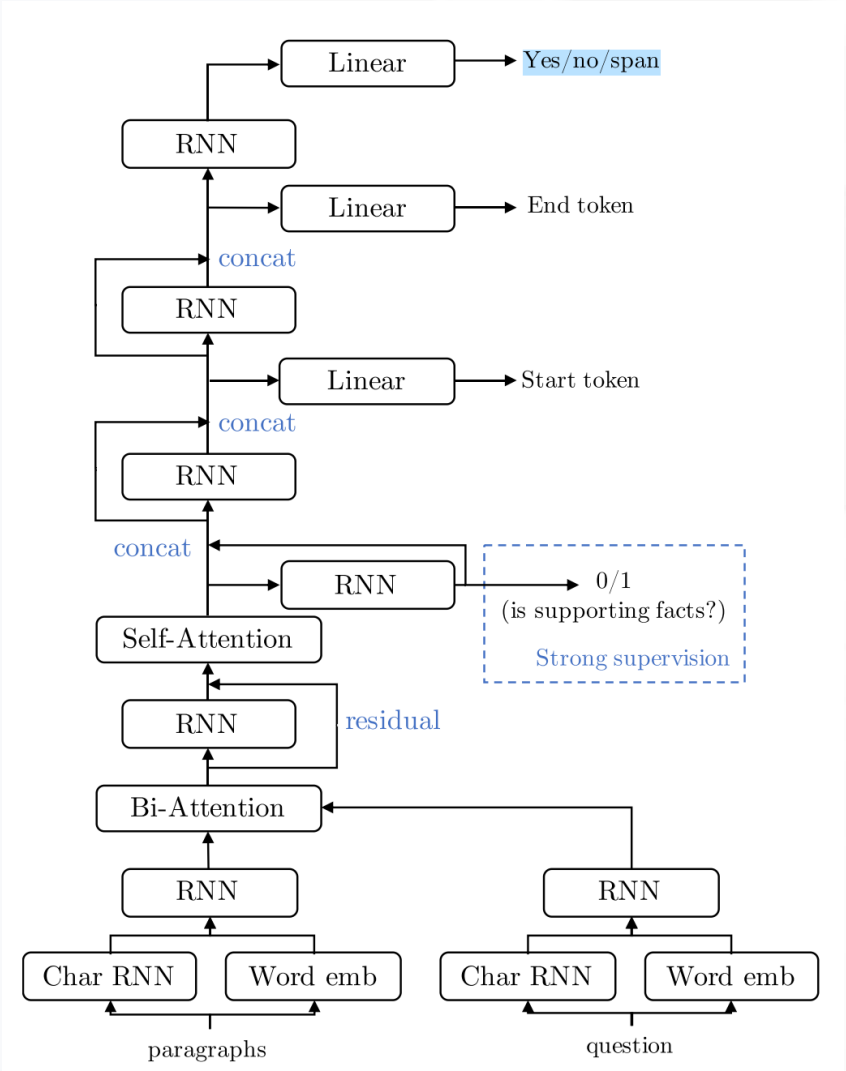


Figure 1: The overview of our QA system. The **bridge reasoner** reads the start passages retrieved by an IR system and predicts a set of candidate bridges (anchor links) that lead to the answer passages, which is further processed by the **passage reader** to return the answer.

Bridge-IR

EMNLP2019 **Simple yet Effective Bridge Reasoning for Open-Domain Multi-Hop Question Answering**



Local Context Evidence：
use each anchor's start token representation hc
as to represent the anchor's local context evidence

Passage Content Evidence：
use a bi-LSTM to encode the abstract passages and use max-pooling on the output states to get the passage content representation

Both the local context evidence and passage content evidence are integrated into our final bridge reasoner by a linear layer

Bridge-IR

EMNLP2019 **Simple yet Effective Bridge Reasoning for Open-Domain Multi-Hop Question Answering**

Approach	Hits@10
HotpotQA IR	48.4
Our Methods	
Bridge Reasoner	76.6
w/o local context evidence	75.4
w/o passage content evidence	65.7
Bridge Reasoner + entity linking	**80.6**

Table 1: Answer passage prediction performance, measured by Hits@10 on dev bridge questions.

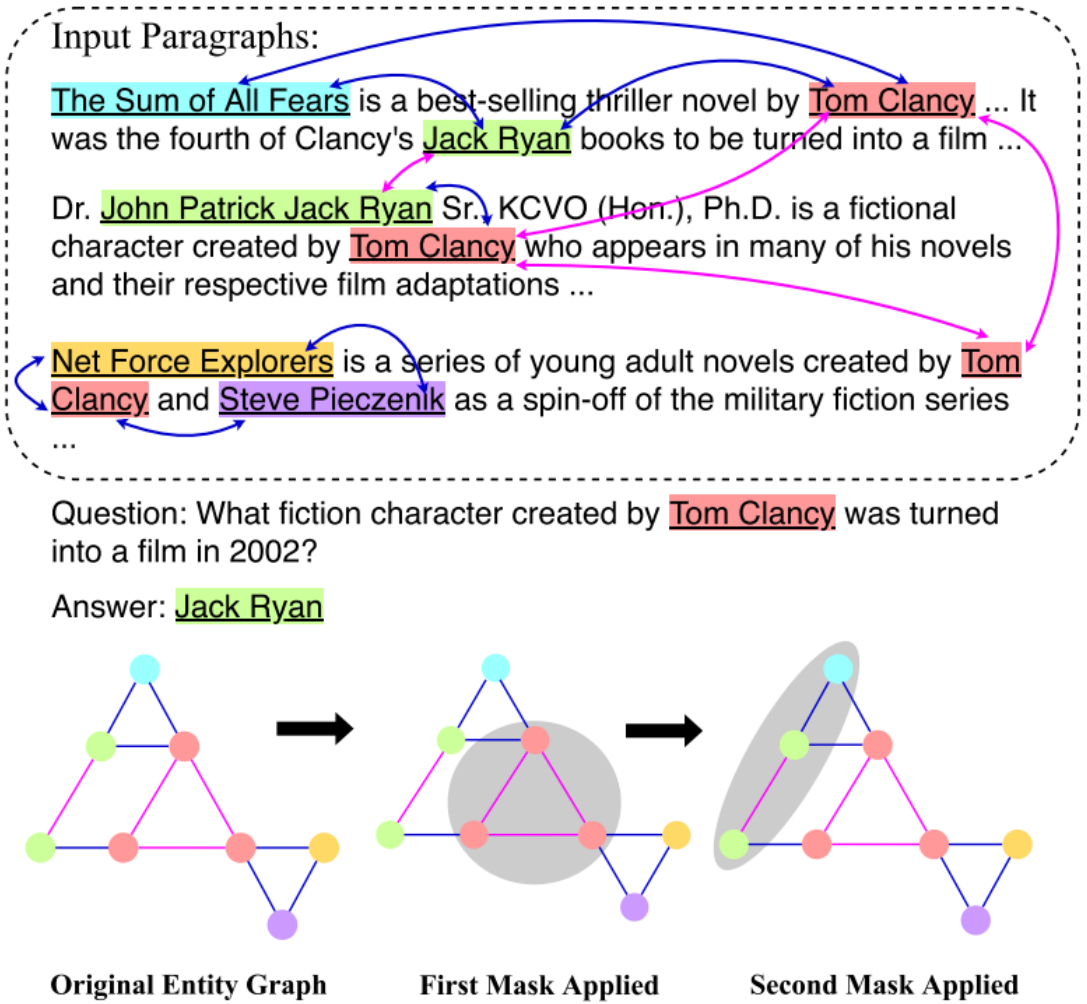
Model	Dev EM	Dev F1	Test EM	Test F1
Methods w/o BERT				
HotpotQA Baseline	24.68	34.36	23.95	32.89
GRN	-	-	27.34	36.48
Ours	36.81	48.48	36.04	47.43
w/o EL	35.00	46.16	-	-
Methods with BERT				
GRN + BERT	-	-	29.87	39.14
CogQA	37.6	49.4	37.12	48.87
w/o EL	34.6	46.2	-	-
w/o re-scoring	33.6	45.0	-	-
Methods with Unknown Usage of BERT				
DecompRC	-	-	30.00	40.65
MUPPET	-	-	30.61	40.26

Table 2: QA performance on HotpotQA. The underline methods use the same resource, but our method does not use any pre-trained contextual embeddings like BERT.

DFGN

ACL2019

Dynamically Fused Graph Network for Multi-hop Reasoning



Example of multihop text-based QA. One question and three document paragraphs are given. Our proposed DFGN conducts multi-step reasoning over the facts by constructing an entity graph from multiple paragraphs, predicting a dynamic mask to select a subgraph, propagating information along the graph, and finally transfer the information from the graph back to the text in order to localize the answer. Nodes are entity occurrences, with the color denoting the underlying entity. Edges are constructed from cooccurrences. The gray circles are selected by DFGN in each step.

DFGN

ACL2019

Dynamically Fused Graph Network for Multi-hop Reasoning

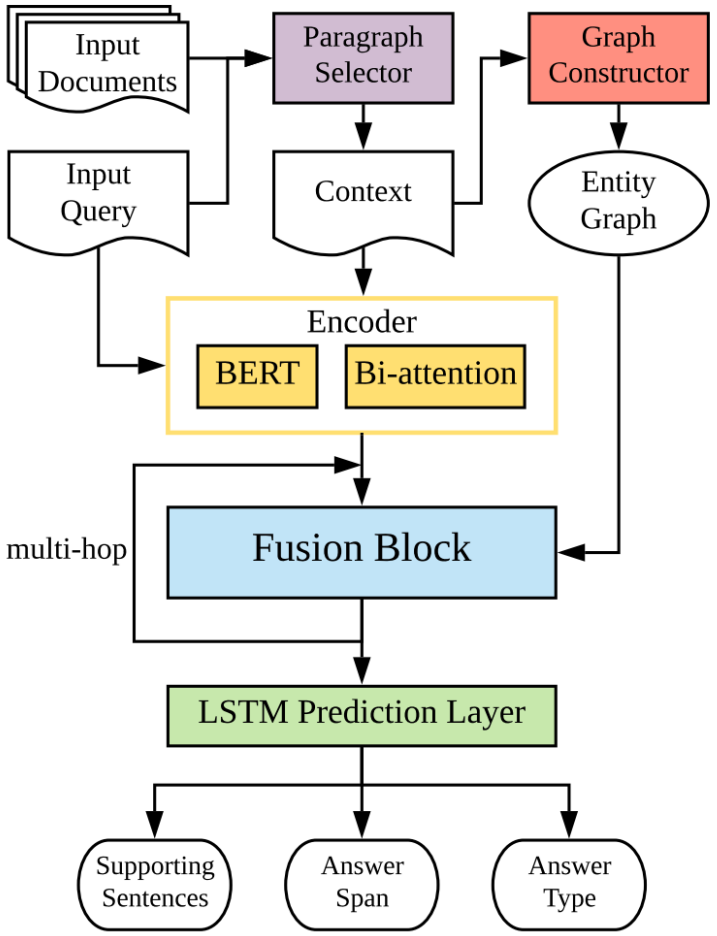


Figure 3: Overview of DFGN.

Paragraph Selector:
The selector network takes a query Q and a paragraph as input and outputs a relevance score between 0 and 1.

Graph Constructor Rules:
sentence-level links: for every pair of entities appear in the same sentence

context-level links:
for every pair of entities with the same mention text

paragraph-level links:
between a central entity node and other entities within the same paragraph

DFGN

ACL2019

Dynamically Fused Graph Network for Multi-hop Reasoning

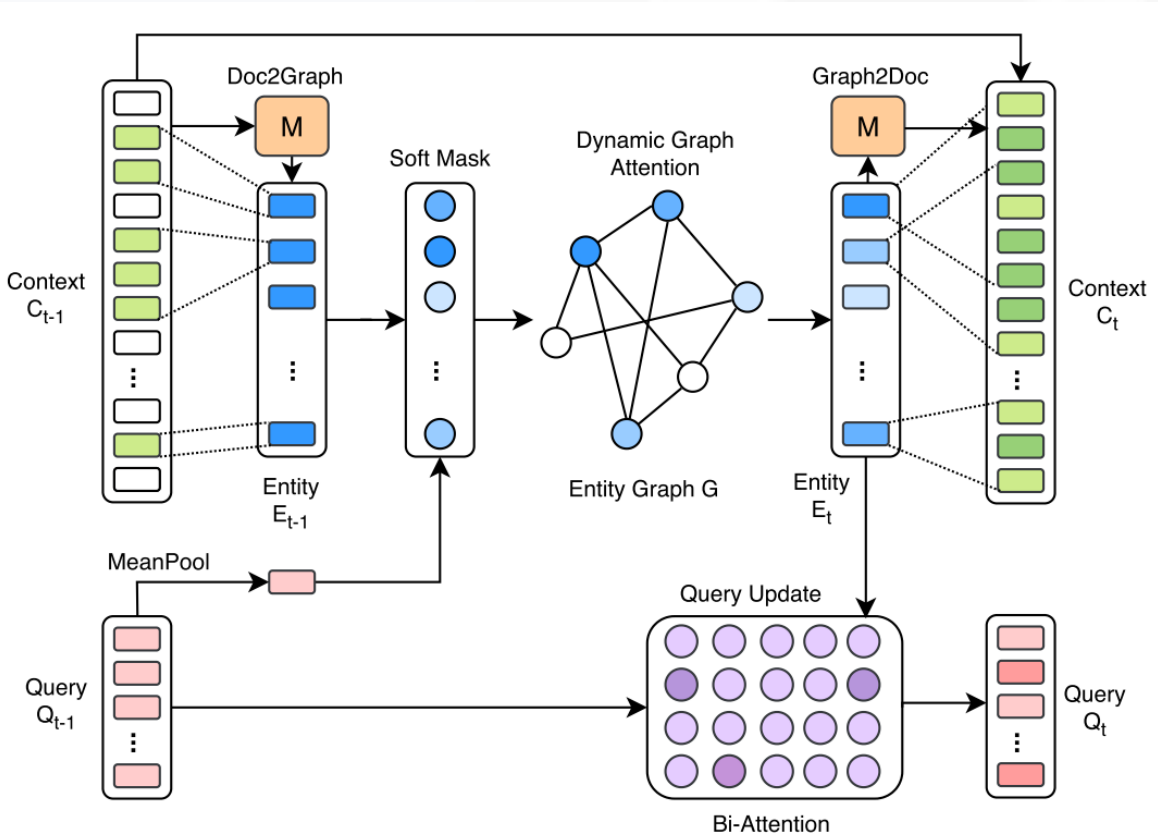


Figure 4: Reasoning with the fusion block in DFGN

Document to Graph Flow:

$$\mathbf{E}_{t-1} = [\mathbf{e}_{t-1,1}, \ldots, \mathbf{e}_{t-1,N}]$$

Dynamic Graph Attention:

$$\tilde{\mathbf{q}}^{(t-1)} = \text{MeanPooling}(\mathbf{Q}^{(t-1)})$$

$$\gamma_i^{(t)} = \tilde{\mathbf{q}}^{(t-1)} \mathbf{V}^{(t)} \mathbf{e}_i^{(t-1)} / \sqrt{d_2}$$

$$\mathbf{m}^{(t)} = \sigma([\gamma_1^{(t)}, \cdots, \gamma_N^{(t)}])$$

$$\tilde{\mathbf{E}}^{(t-1)} = [m_1^{(t)} \mathbf{e}_1^{(t-1)}, \ldots, m_N^{(t)} \mathbf{e}_N^{(t-1)}]$$

DFGN

ACL2019

Dynamically Fused Graph Network for Multi-hop Reasoning

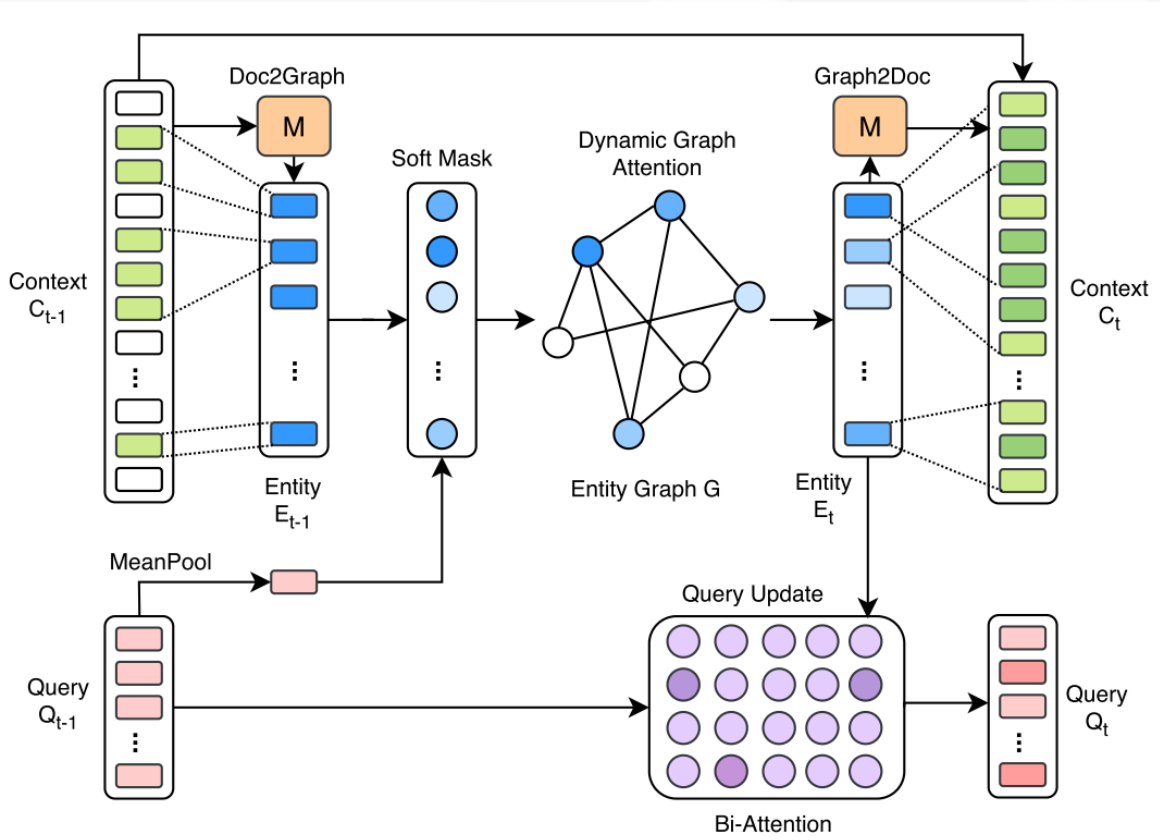


Figure 4: Reasoning with the fusion block in DFGN

$$\mathbf{h}_i^{(t)} = \mathbf{U}_t \tilde{\mathbf{e}}_i^{(t-1)} + \mathbf{b}_t$$

$$\beta_{i,j}^{(t)} = \text{LeakyReLU}(\mathbf{W}_t^\top [\mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t)}])$$

$$\alpha_{i,j}^{(t)} = \frac{\exp(\beta_{i,j}^{(t)})}{\sum_k \exp(\beta_{i,k}^{(t)})}$$

$$\mathbf{e}_i^{(t)} = \text{ReLU}(\sum_{j \in B_i} \alpha_{j,i}^{(t)} \mathbf{h}_j^{(t)})$$

Update Query：

$$\mathbf{Q}^{(t)} = \text{Bi-Attention}(\mathbf{Q}^{(t-1)}, \mathbf{E}^{(t)})$$

Graph to Document Flow:

$$\mathbf{C}^{(t)} = \text{LSTM}([\mathbf{C}^{(t-1)}, \mathbf{M}\mathbf{E}^{(t)\top}])$$

Model	Answer		Sup Fact		Joint	
	EM	F1	EM	F1	EM	F1
Baseline Model	45.60	59.02	20.32	64.49	10.83	40.16
GRN*	52.92	66.71	52.37	84.11	31.77	58.47
DFGN(Ours)	55.17	68.49	49.85	81.06	31.87	58.23
QFE*	53.86	68.06	**57.75**	**84.49**	**34.63**	59.61
DFGN(Ours)†	**56.31**	**69.69**	51.50	81.62	33.62	**59.82**

Table 1: Performance comparison on the private test set of HotpotQA in the distractor setting. Our DFGN is the second best result on the leaderboard before submission (on March 1st). The baseline model is from Yang et al. (2018) and the results with * is unpublished. DFGN(Ours)† refers to the same model with a revised entity graph, whose entities are recognized by a BERT NER model. Note that the result of DFGN(Ours)† is submitted to the leaderboard during the review process of our paper.

DFGN

ACL2019

Dynamically Fused Graph Network for Multi-hop Reasoning

0.67	0.01	0.01	Barrack
0	0.8	0.67	Provisional Irish Republican Army
0.01	0.69	1.09	IRA
0.02	0	0	British Royal Navy
0.01	0	0	British Army Gazelle
0	0	0	Falkland Islands
0.74	0	0.01	British Army Lynx
0	0.82	0.41	Provisional Irish Republican Army
0	0.81	0.73	IRA
0	0.73	0.13	Northern Ireland
0.01	0.33	0.61	IRA

Q1: Who used a Barrack buster to shoot down a British Army Lynx helicopter?
Answer: IRA ***Prediction***: IRA
Top 1 Reasoning Chain: British Army Lynx, Provisional Irish Republican Army, IRA

Supporting Fact 1:
"Barrack buster is the colloquial name given to several improvised mortars, developed in the 1990s by the engineering group of the Provisional Irish Republican Army (IRA)."
Supporting Fact 2:
" On 20 March 1994, a British Army Lynx helicopter was shot down by the Provisional Irish Republican Army (IRA) in Northern Ireland."

0.01	0.07	0	Sasanid
0	0.07	0	Iran
0	0.02	0	Islam
0	0.02	0	House of Sasan
0	0.02	0	Roman-Byzantine Empire
0	0.11	0	Samo
0.04	0.03	0	King
0	0.03	0.01	Samo
0	0	0	Moravia

Q2: From March 631 to April 631, Farrukhzad Khosrau V was the king of an empire that succeeded which empire?
Answer: the Parthian Empire **Prediction:** Parthian Empire **Top 1 Reasoning Chain:** n/a

Supporting Fact 1:
"Farrukhzad Khosrau V was briefly king of the Sasanian Empire from March 631 to ..."
Supporting Fact 2:
"The Sasanian Empire, which succeeded the Parthian Empire, was recognised as ... the Roman-Byzantine Empire, for a period of more than 400 years."

SAE

AAAI2020

Select, Answer and Explain: Interpretable Multi-hop Reading Comprehension over Multiple Documents

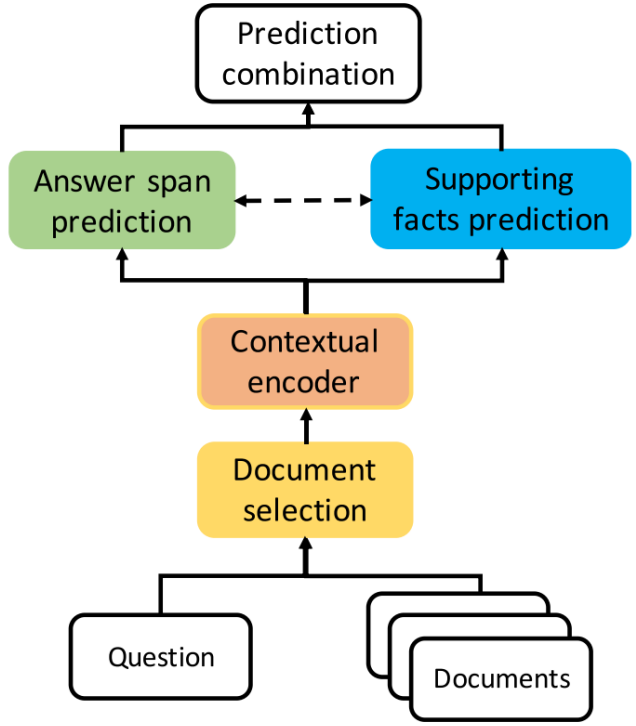


Figure 2: Diagram of the proposed SAE system. The dashed arrow line indicates the mixed attention based interaction between the two tasks

Select gold documents

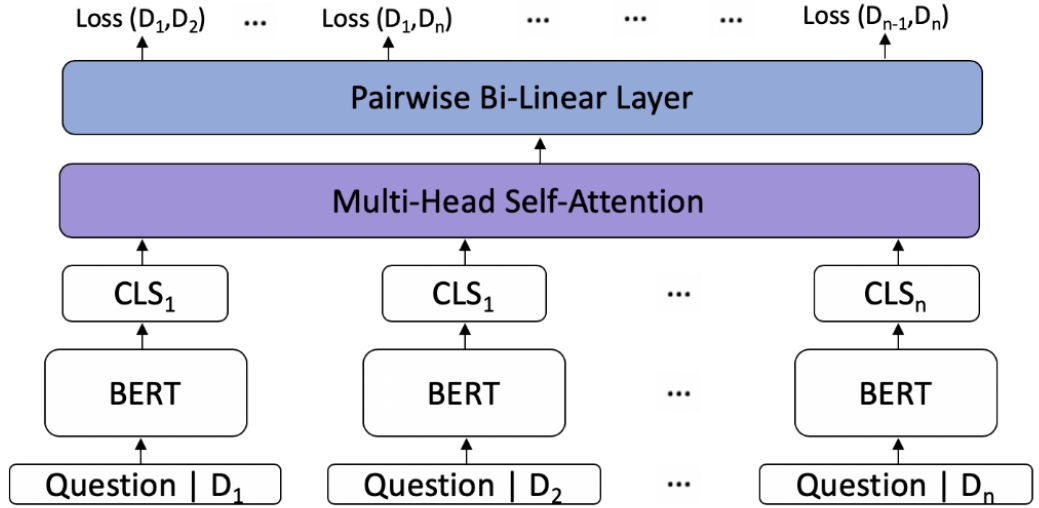


Figure 3: Diagram of document selection module. N indicates the total number of documents.

SAE

AAAI2020

Select, Answer and Explain: Interpretable Multi-hop Reading Comprehension over Multiple Documents

Select gold documents

$$L = -\sum_{i=0}^{n} t_i log P(D_i) + (1 - t_i) log(1 - P(D_i))$$

binary cross entropy loss

if Di is gold document: $S(D_i) = 1$

is a golden document containing the answer span: $S(D_i) = 2$

else: $S(D_i) = 0$

$$l_{i,j} = \begin{cases} 1 & \text{if } S(D_i) > S(D_j) \\ 0 & \text{if } S(D_i) <= S(D_j) \end{cases}$$

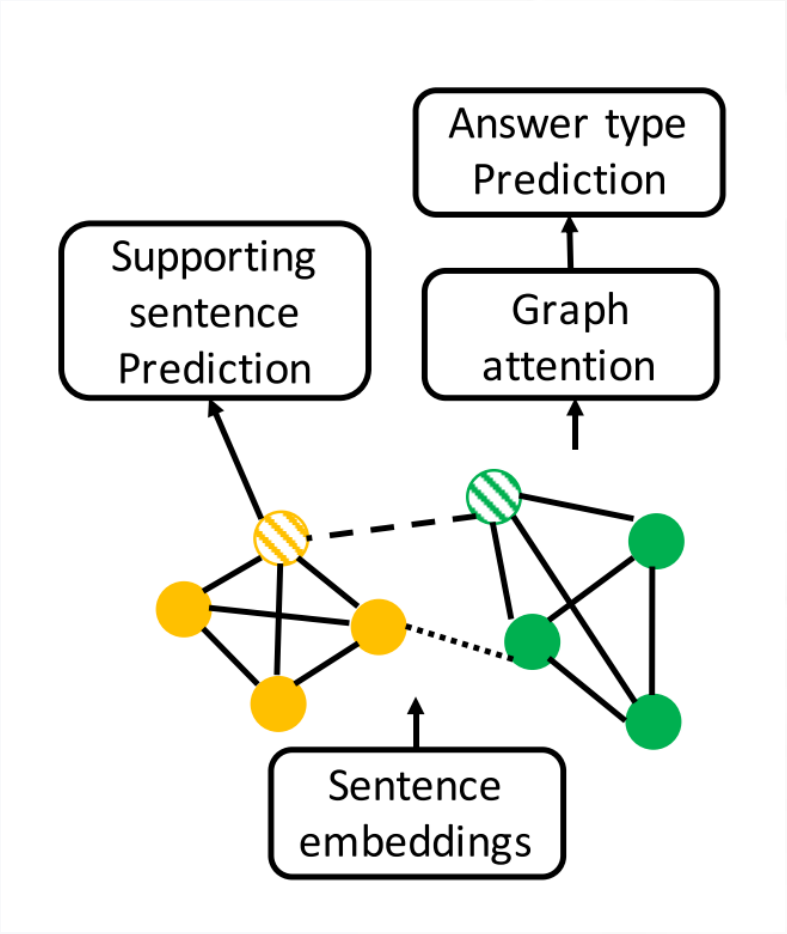
$$L = -\sum_{i=0}^{n} \sum_{j=0, j \neq i}^{i} l_{i,j} log P(D_i, D_j) + (1 - l_{i,j}) log(1 - P(D_i, D_j))$$

$$R_i = \sum_j^n \mathbb{1}(P(D_i, D_j) > 0.5)$$

SAE

AAAI2020

Select, Answer and Explain: Interpretable Multi-hop Reading Comprehension over Multiple Documents



Graph Constructor Rules:

1. Add an edge between two nodes if they are originally from the same document

2. Add an edge between two nodes from different documents if the sentences representing the two nodes both have named entities or noun phrases (can be different) in the question

3. Add an edge between two nodes from different documents if the sentences representing the two nodes have the same named entities or noun phrases

SAE

AAAI2020

Select, Answer and Explain: Interpretable Multi-hop Reading Comprehension over Multiple Documents

Answer Prediction

$$\hat{\mathbf{Y}} = f_{span}(\mathbf{H}^i) \in \mathbb{R}^{L\times 2},$$

$$L^{span} = \frac{1}{2}(CE(\hat{\mathbf{Y}}[:,0], \mathbf{y}^{start}) + CE(\hat{\mathbf{Y}}[:,1], \mathbf{y}^{end}))$$

Supporting sentence prediction

$$\mathbf{S}^j = \mathbf{H}[j^s : j^e, :] \in \mathbb{R}^{L^j \times d}$$

$$\boldsymbol{\alpha}^{\boldsymbol{j}} = \sigma(f_{att}(\mathbf{S}^j) + \hat{\mathbf{Y}}[j^s : j^e, 0] + \hat{\mathbf{Y}}[j^s : j^e, 1])$$

$$\mathbf{s}^j = \sum_{k=0}^{L^j} \alpha_k^j \mathbf{S}^j[k, :] \in \mathbb{R}^{1\times d},$$

SAE

AAAI2020

Select, Answer and Explain: Interpretable Multi-hop Reading Comprehension over Multiple Documents

Table 1: Results comparison between our proposed SAE system with other methods. * indicates unpublished models.

	Model	Ans		Sup		Joint	
		EM	F_1	EM	F_1	EM	F_1
Dev	Baseline(Yang et al. 2018)	44.44	58.28	21.95	66.66	11.56	40.86
	QFE(Nishida et al. 2019)	53.70	68.70	58.80	84.70	35.40	60.60
	DFGN(Xiao et al. 2019)	55.66	69.34	53.10	82.24	33.68	59.86
	SAE(ours)	61.32	74.81	58.06	85.27	39.89	66.45
	SAE-oracle(ours)	63.48	77.16	62.80	89.29	42.77	70.13
	SAE-large(ours)	67.70	80.75	63.30	87.38	46.81	72.75
Test	Baseline(Yang et al. 2018)	45.46	58.99	22.24	66.62	12.04	41.37
	QFE(Nishida et al. 2019)	53.86	68.06	57.75	84.49	34.63	59.61
	DFGN(Xiao et al. 2019)	56.31	69.69	51.50	81.62	33.62	59.82
	SAE(ours)	60.36	73.58	56.93	84.63	38.81	64.96
	SAE-large(ours)	66.92	79.62	**61.53**	86.86	**45.36**	71.45
	C2F Reader*	67.98	81.24	60.81	87.63	44.67	72.73

SAE

AAAI2020

Select, Answer and Explain: Interpretable Multi-hop Reading Comprehension over Multiple Documents

Table 2: Ablation study results on HotpotQA dev set. PR(0,1) stands for giving 0 score to non-gold documents and 1 score to all gold documents when preparing pairwise labels, and PR(0,1,2) stands for giving 2 score to the gold document with answer span.

	EM_S	$Recall_S$	Acc_{span}	joint EM	joint F_1
BERT only	70.65	89.16	90.08	31.87	59.33
+MHSA	87.07	94.65	92.54	38.54	65.00
+PR(0,1)	89.76	94.75	94.53	39.53	65.44
+PR(0,1,2)	91.40	95.61	95.86	39.89	66.45

Table 3: Ablation study results on HotpotQA dev set.

	joint EM	joint F_1
full model	39.89	66.45
-mixed attn	39.59	66.28
-attn sum	38.04	65.33
-GNN	38.46	65.53
-type 1 edge	38.15	65.00
-type 2 edge	39.55	66.13
-type 3 edge	39.32	66.03
-type 2&3 edge	39.16	65.76

HGN

ACL2019

Hierarchical Graph Network for Multi-hop Question Answering

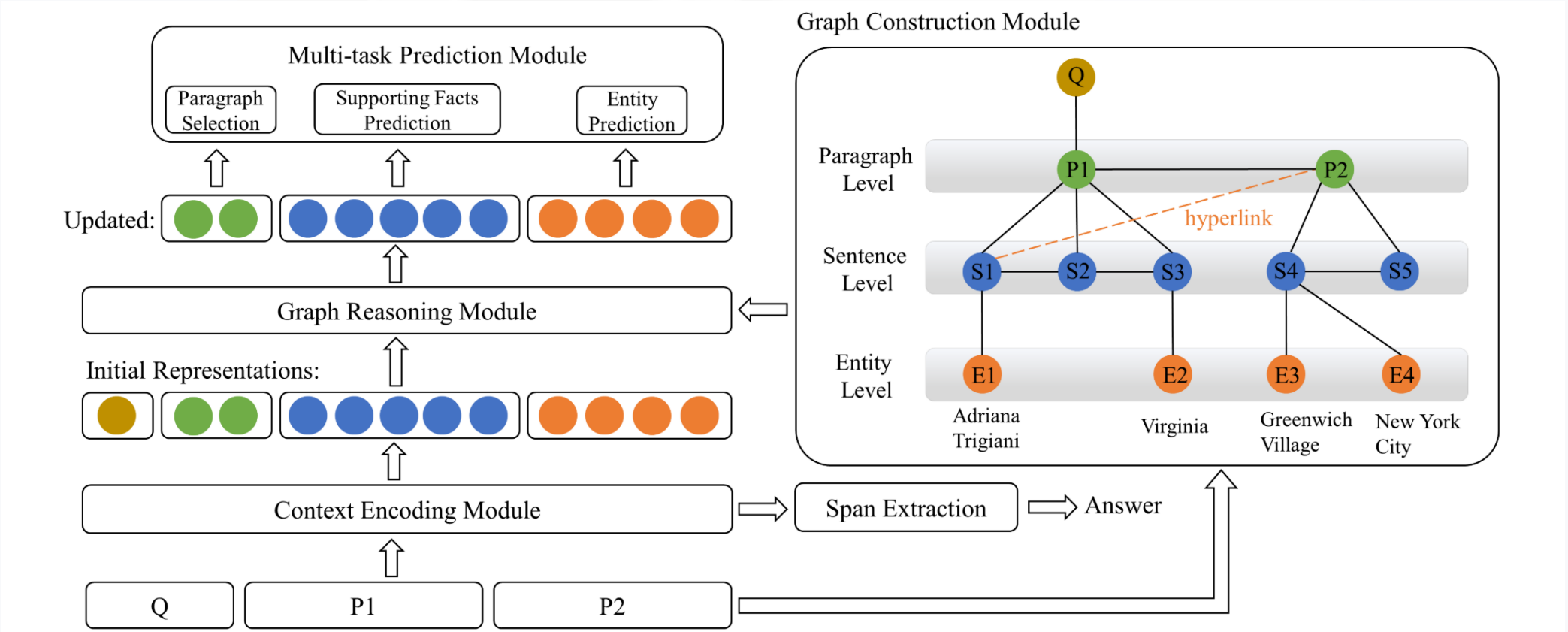


Figure 2: Model architecture of the proposed Hierarchical Graph Network. The constructed graph corresponds to the example in Figure 1. Green, blue, orange, and brown colors represent paragraph, sentence, entity, and question nodes, respectively. Some entities and hyperlinks are omitted for illustration simplicity.

HGN

ACL2019

Hierarchical Graph Network for Multi-hop Question Answering

Paragraph Selector:

use a paragraph ranking model to select paragraphs with top-N ranking scores in each step. This paragraph ranking model is based on a pretrained BERT encoder, followed by a binary classification layer, to predict whether an input paragraph contains the ground-truth supporting facts or not

Graph Constructor Rules:

(i) edges between question node and paragraph nodes;
(ii) edges between question node and its corresponding entity nodes (entities appearing in the question, not shown for simplicity);
(iii) edges between paragraph nodes and their corresponding sentence nodes (sentences within the paragraph);
(iv) edges between sentence nodes and their linked paragraph nodes (linked through hyperlinks);
(v) edges between sentence nodes and their corresponding entity nodes (entities appearing in the sentences);
(vi) edges between paragraph nodes;
(vii) edges between sentence nodes that appear in the same paragraph.

HGN

ACL2019

Hierarchical Graph Network for Multi-hop Question Answering

Context Encoding

$$\mathbf{p}_i = \mathrm{MLP}_1([\mathbf{M}[P^{(i)}_{start}][d:]; \mathbf{M}[P^{(i)}_{end}][:d]])$$
$$\mathbf{s}_i = \mathrm{MLP}_2([\mathbf{M}[S^{(i)}_{start}][d:]; \mathbf{M}[S^{(i)}_{end}][:d]])$$
$$\mathbf{e}_i = \mathrm{MLP}_3([\mathbf{M}[E^{(i)}_{start}][d:]; \mathbf{M}[E^{(i)}_{end}][:d]])$$
$$\mathbf{q} = \text{max-pooling}(\mathbf{Q}),$$

Graph Reasoning

$$\mathbf{h}'_i = \sigma\Big(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W} \mathbf{h}_j\Big)$$

$$\alpha_{ij} = \frac{\exp(f(\mathbf{W}_{e_{ij}}[\mathbf{h}_i; \mathbf{h}_j]))}{\sum_{k \in \mathcal{N}_i} \exp(f(\mathbf{W}_{e_{ik}}[\mathbf{h}_i; \mathbf{h}_k]))}$$

HGN

ACL2019

Hierarchical Graph Network for Multi-hop Question Answering

Multi-Task Prediction

(i) paragraph selection based on paragraph nodes;
(ii) supporting facts prediction based on sentence nodes;
(iii) answer prediction based on entity nodes.

$$\mathbf{o}_{sent} = \mathrm{MLP}_4(\mathbf{S}'),\ \mathbf{o}_{para} = \mathrm{MLP}_5(\mathbf{P}')$$

$$\mathbf{o}_{entity} = \mathrm{MLP}_6(\mathbf{E}')$$

$$\mathcal{L}_{joint} = \mathcal{L}_{span} + \lambda_1 \mathcal{L}_{entity} + \lambda_2 \mathcal{L}_{sent} + \lambda_3 \mathcal{L}_{para} + \lambda_4 \mathcal{L}_{type} ,$$

HGN

ACL2019

Hierarchical Graph Network for Multi-hop Question Answering

Model	Ans		Sup		Joint	
	EM	F1	EM	F1	EM	F1
DecompRC (Min et al., 2019b)	55.20	69.63	-	-	-	-
ChainEx (Chen et al., 2019)	61.20	74.11	-	-	-	-
Baseline Model (Yang et al., 2018)	45.60	59.02	20.32	64.49	10.83	40.16
QFE (Nishida et al., 2019)	53.86	68.06	57.75	84.49	34.63	59.61
DFGN (Xiao et al., 2019)	56.31	69.69	51.50	81.62	33.62	59.82
LQR-Net† (Anonymous, 2020a)	60.20	73.78	56.21	84.09	36.56	63.68
P-BERT†	61.18	74.16	51.38	82.76	35.42	63.79
SAE†	60.36	73.58	56.93	84.63	38.81	64.96
TAP2†	64.99	78.59	55.47	85.57	39.77	69.12
EPS+BERT†	65.79	79.05	58.50	86.26	42.47	70.48
HGN (ours)	**66.07**	**79.36**	**60.33**	**87.33**	**43.57**	**71.03**

Table 1: Results on the test set of HotpotQA in the Distractor setting. HGN achieves state-of-the-art results at the time of submission (Sep. 27, 2019). (†) indicates unpublished work. BERT-wwm is used for context encoding. Leaderboard: https://hotpotqa.github.io/.

HGN

ACL2019

Hierarchical Graph Network for Multi-hop Question Answering

Model	Ans F1	Sup F1	Joint F1
DFGN (paper)	69.38	82.23	59.89
DFGN			
+ threshold-based	71.90	83.57	63.04
+ 2 para. (ours)	72.53	83.57	63.87
+ 4 para. (ours)	**72.67**	83.34	63.63
BERT-base			
+ threshold-based	71.95	82.79	62.43
+ 2 para. (ours)	72.42	83.64	63.94
+ 4 para. (ours)	**72.67**	**84.86**	**64.24**

Table 4: Results with selected paragraphs on the dev set in the Distractor setting.

Model	Ans F1	Sup F1	Joint F1
w/o Graph	80.58	85.83	71.02
PS Graph	80.94	87.59	72.61
PSE Graph	80.70	**88.00**	72.79
Hier. Graph	**81.00**	87.93	**73.01**

Table 5: Ablation study on the effectiveness of the hierarchical graph on the dev set in the Distractor setting. RoBERTa-large is used for context encoding.

HGN

ACL2019

Hierarchical Graph Network for Multi-hop Question Answering

Objective	Ans F1	Sup F1	Joint F1
$\mathcal{L}_{joint}$	**81.00**	87.93	**73.01**
$-\mathcal{L}_{entity}$	80.86	**87.99**	72.87
$-\mathcal{L}_{para}$	80.89	87.71	72.73
$-\mathcal{L}_{entity}$ & $\mathcal{L}_{para}$	80.76	87.78	72.70

Table 6: Ablation study on the proposed multi-task loss. RoBERTa-large is used for context encoding.

Model	Ans F1	Sup F1	Joint F1
DFGN (BERT-base)	69.38	82.23	59.89
EPS (BERT-wwm)†	79.05	86.26	70.48
HGN (BERT-base)	74.07	85.62	66.01
HGN (BERT-wwm)	79.69	87.38	71.45
HGN (RoBERTa)	**81.00**	**87.93**	**73.01**

Table 7: Results with different pre-trained language models on the dev set in the Distractor setting. (†) is unpublished work with results on the test set, using BERT whole word masking (wwm).

THANK YOU!

2020 / 4 / 13