NumNet: Machine Reading Comprehension with Numerical Reasoning

EMNLP-IJCNLP 2019

Qiu Ran[1*] , Yankai Lin[1*] , Peng Li[1] , Jie Zhou[1] , Zhiyuan Liu[2]

[1]**Pattern Recognition Center, WeChat AI, Tencent Inc, China**

[2]**Department of Computer Science and Technology, Tsinghua University, Beijing, China**

Part 1.

Introduction

- **Machine reading comprehension (MRC)**
  - Infer the answer to a question given the document
  - Achieve remarkable results in various public benchmarks such as SQuAD and RACE

  - The success of MRC models
    - Multi-layer architectures which allow these models to read the document and the question iteratively for reasoning
    - Attention mechanisms which would enable these models to focus on the part related to the question in the document

  - The limitations of MRC models
    - These models are still weak in numerical reasoning such as addition, subtraction, sorting and counting
    - Naturally require when reading financial news, scientific articles, etc.

Part 1.

Introduction

- **Previous Work: A numerically-aware QANet (NAQANet)**
  - Divides the answer generation for numerical MRC into three types:
    - (1) Extracting spans; (2) Counting; (3) Addition or Subtraction over numbers
  - Answer numerical questions but still does not explicitly consider numerical reasoning

Part 1.

Introduction

- **NumNet**
  - Integrate numerical reasoning into existing MRC models
  - How to perform numerical comparison in MRC systems, which is crucial for two common types of questions:

  - (1) Numerical Comparison:
  - (2) Numerical Condition:

Part 1.

Introduction

- **NumNet**
  - Integrate numerical reasoning into existing MRC models
  - How to perform numerical comparison in MRC systems, which is crucial for two common types of questions:

  - (1) Numerical Comparison:
    - The answers of the questions can be directly obtained via performing numerical comparison, such as sorting and comparison, in the documents
    - If the MRC system knows the fact that "$49 > 47 > 36 > 31 > 22$", it could easily extract that the second longest field goal is 47-yard

Question	Passage	Answer
What is the second longest field goal made?	... The Seahawks immediately trailed on a scoring rally by the Raiders with kicker *Sebastian Janikowski nailing a 31-yard field goal* ... Then in the third quarter *Janikowski made a 36-yard field goal*. Then *he made a 22-yard field goal* in the fourth quarter to put the Raiders up 16-0 ... The Seahawks would make their only score of the game with kicker *Olindo Mare hitting a 47-yard field goal*. However, they continued to trail as *Janikowski made a 49-yard field goal*, followed by RB Michael Bush making a 4-yard TD run.	47-yard

Part 1.

Introduction

- **NumNet**
  - Integrate numerical reasoning into existing MRC models
  - How to perform numerical comparison in MRC systems, which is crucial for two common types of questions:

  - (2) Numerical Condition:
    - The answers of the questions cannot be directly obtained through simple numerical comparison in the documents
    - Needs to know which age group made up more than 7% of the population to count the group number

Question	Passage	Answer
How many age groups made up more than 7% of the population?	Of Saratoga Countys population in 2010, *6.3%* were between ages of 5 and 9 years, *6.7%* between 10 and 14 years, 6.5% between 15 and 19 years, *5.5%* between 20 and 24 years, *5.5%* between 25 and 29 years, *5.8%* between 30 and 34 years, *6.6%* between 35 and 39 years, *7.9%* between 40 and 44 years, *8.5%* between 45 and 49 years, *8.0%* between 50 and 54 years, *7.0%* between 55 and 59 years, *6.4%* between 60 and 64 years, and *13.7%* of age 65 years and over ...	5

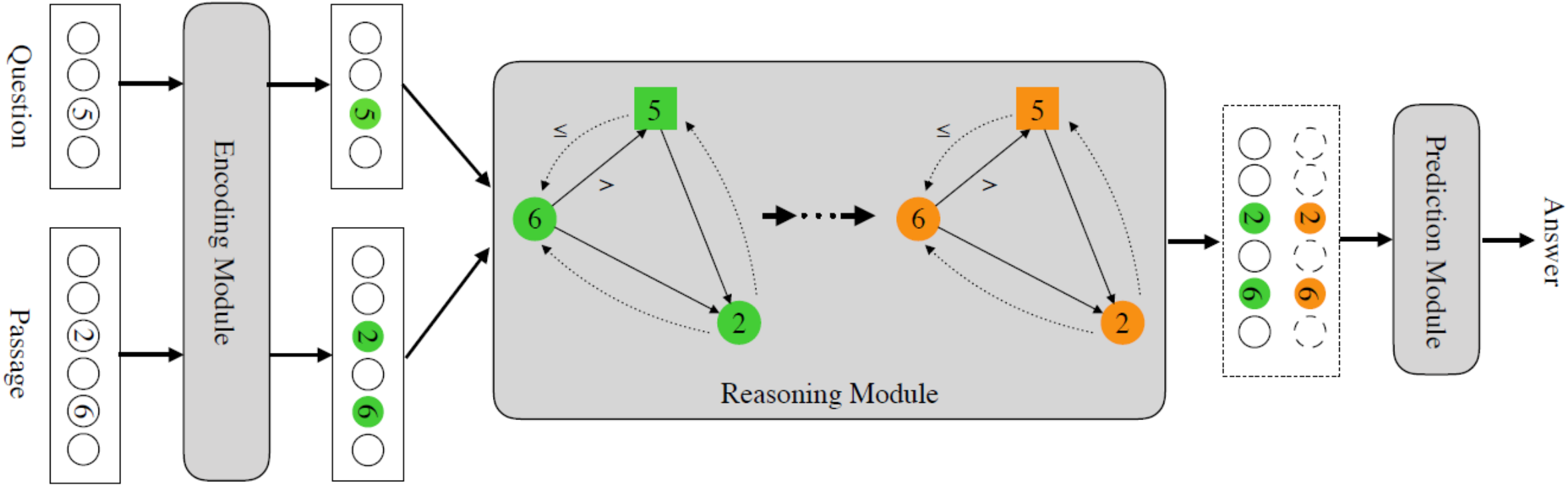
Introduction

- **NumNet**
  - Encode both the question and passages through an encoding module consisting of convolution layers, self-attention layers and feed-forward layers as well as a passage-question attention layer
  - Feed the question and passage representations into a numerically-aware graph neural network (NumGNN)
    - Further to integrate the comparison information among numbers into their representations
    - Utilize the numerically-aware representation of passages to infer the answer to the question
  - The experimental results
    - Dataset: A public numerical MRC dataset DROP
    - Achieve significant and consistent improvement
    - Effectively deal with questions requiring sorting with multi-layer NumGNN

Part 2.

Method

- **Framework**
  - The numerical relations between numbers are encoded with the topology of the graph
    - The edge pointing from “6” to “5” denotes “6” is greater than “5”
  - The reasoning module leverages a numerically-aware graph neural network



Method

- **Encoding Module**
  - Without loss of generality
  - Use the encoding components of QANet and NAQANet to encode the question and passage into vector-space representations

  - **First encoding**
  - Convolution, selfattention and feed-forward layers
  - With stacked embedding encoder layer

$$\texttt{QANet-Emb-Enc}(\cdot)$$

$$\boldsymbol{Q} = \texttt{QANet-Emb-Enc}(Q)$$
$$\boldsymbol{P} = \texttt{QANet-Emb-Enc}(P)$$

  - **The passage-aware question representation**
  - **The question-aware passage representation**
  - With context-query attention layer

$$\texttt{QANet-Att}(\cdot)$$

$$\bar{\boldsymbol{Q}} = \texttt{QANet-Att}(\boldsymbol{P}, \boldsymbol{Q})$$
$$\bar{\boldsymbol{P}} = \texttt{QANet-Att}(\boldsymbol{Q}, \boldsymbol{P})$$

Method

- **Reasoning Module**
  - A heterogeneous directed graph $\mathcal{G} = (\boldsymbol{V}; \boldsymbol{E})$
  - Nodes $\boldsymbol{V}$
    - They are corresponding to the numbers in the question and passage
  - Edges **E**
    - They are used to encode numerical relationships among the numbers

  - Reasoning on the graph based on a graph neural network

$$\boldsymbol{M}^Q = \texttt{QANet-Mod-Enc}(\boldsymbol{W}^M \bar{\boldsymbol{Q}})$$

$$\boldsymbol{M}^P = \texttt{QANet-Mod-Enc}(\boldsymbol{W}^M \bar{\boldsymbol{P}})$$

$$\boldsymbol{U} = \texttt{Reasoning}(\mathcal{G}; \boldsymbol{M}^Q, \boldsymbol{M}^P)$$

  - A shared weight matrix W^M
  - The representations of the nodes corresponding to the numbers U
  - Model encoder layer $\texttt{QANet-Mod-Enc}(\cdot)$

Method

- **Reasoning Module**
  - $\boldsymbol{U}$ only contains the representations of numbers, to tackle span-style answers containing non-numerical words
  - Concatenate $\boldsymbol{U}$ with M^P to produce numerically-aware passage representation M_0

$$\boldsymbol{M}^{\text{num}}[i] = \begin{cases} \boldsymbol{U}[I(i)] & \text{if } w_i^p \text{ is a number} \\ \mathbf{0} \end{cases}$$

$$\boldsymbol{M}_0' = \boldsymbol{W}_0[\boldsymbol{M}^P; \boldsymbol{M}^{\text{num}}] + \boldsymbol{b}_0$$

$$\boldsymbol{M}_0 = \texttt{QANet-Mod-Enc}(\boldsymbol{M}_0')$$

  - Matrix concatenation $[\cdot;\cdot]$
  - $W[k]$ denotes the k-th column of a matrix W
  - $I(i)$ denotes the node index corresponding to the passage word (number) w_i^p
  - W_0 A weight matrix
  - b_0 A bias vector

Method

- **Prediction Module**
  - Following NAQANet (Dua et al., 2019), divide the answers into four types
    Use a unique output layer for the conditional answer probability $\Pr(\text{answer}|\text{type})$
  - Passage Span/Question Span
    - The answer is a span of the passage
    - The answer probability is defined as the product of the probabilities of the start and end positions.
  - Count
    - The answer is obtained by counting, and it is treated as a multi-class classification problem over ten numbers (0-9), which covers most of the Count type answers in the DROP dataset
  - Arithmetic expression
    - The answer is the result of an arithmetic expression
    - The expression is obtained in three steps: (1) extract all numbers from the passage; (2) assign a sign (plus, minus or zero) for each number; (3) sum the signed numbers

Method

- **Prediction Module**
  - An extra output layer is used to predict the probability $\Pr(\text{type})$
  - Training time
    - The final answer probability is defined as the joint proba Pbility over all feasible answer type

$$\sum_{\text{type}} \Pr(\text{type}) \Pr(\text{answer}|\text{type})$$

    - The answer type annotation is not required and the probability $\Pr(\text{type})$ is learnt by the model
  - Test time
    - The model first selects the most probable answer type greedily and then predicts the best answer accordingly
  - Without loss of generality
    - Leverage the definition of the five output layers in (Dua et al., 2019), with M_0 and $\mathbf{Q}$ as inputs

Method

- **Prediction Module**
  - The major difference between our model and NAQANet
    - NAQANet does not have the reasoning module
    - i.e., M_0 is simply set as M^P
  - Numbers are treated as common words in NAQANet except in the prediction module
  - NAQANet may struggle to learn the numerical relationships between numbers, and potentially cannot well generalize to unseen numbers

Method

- **Numerically-aware Graph Construction**
  - Regard all numbers from the question and passage as nodes in the graph for reasoning
    - V^Q and V^P
    - All the nodes $\boldsymbol{V} = \boldsymbol{V}^Q \cup \boldsymbol{V}^P$
    - The number corresponding to a node $v \in \boldsymbol{V}$ as $n(v)$
  - Greater Relation Edge $\overrightarrow{\boldsymbol{E}}$
    - For two nodes $v_i, v_j \in \boldsymbol{V}$, A directed edge $\overrightarrow{e}_{ij} = (v_i, v_j)$
    - Pointing from v_i to vj will be added to the graph if $n(v_i) > n(v_j)$
    - Solid arrow
  - Lower or Equal Relation Edge $\overleftarrow{\boldsymbol{E}}$
    - For two nodes $v_i, v_j \in \boldsymbol{V}$, A directed edge $\overleftarrow{e}_{ij} = (v_j, v_i)$
    - The edge will be added to the graph if $n(v_i) \leq n(v_j)$
    - Dashed arrow

Method

- **Numerically-aware Graph Construction**
  - Theoretically, $\overrightarrow{E}$ and $\overleftarrow{E}$ are complement to each other
  - As a number may occur several times and represent different facts in a document
  - Add a distinct node for each occurrence in the graph to prevent potential ambiguity
  - More reasonable to use both $\overrightarrow{E}$ and $\overleftarrow{E}$ in order to encode the equal information among nodes

Method: Numerical Reasoning

- **Numerical Reasoning** $\texttt{Reasoning}(\cdot)$
  - Built the graph $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$

- **Initialization**
  - For each node $v_i^P \in \boldsymbol{V}^P$
  - Its representation is initialized as the corresponding column vector of M^P
  - Formally, the initial representation is $\boldsymbol{v}_i^P = \boldsymbol{M}^P[I^P(v_i^P)]$
  - $I^P(v_i^P)$ denotes the word index corresponding to v_i^P
  - The initial representation $\boldsymbol{v}_j^Q$ for a node $v_j^Q \in \boldsymbol{V}^Q$
    - Set as the corresponding column vector of M^Q
    - Denote all the initial node representations as $\boldsymbol{v}^0 = \{\boldsymbol{v}_i^P\} \cup \{\boldsymbol{v}_j^Q\}$

Method: Numerical Reasoning

- **One-step Reasoning**
  - Given the graph $\mathcal{G}$ and the node representations v, we use a GNN to perform reasoning in three steps

- **(1) Node Relatedness Measure**
  - As only a few numbers are relevant for answering a question generally
    Compute a weight for each node to by-pass irrelevant numbers in reasoning

  - Formally, the weight for node v_i

$$\alpha_i = \text{sigmoid}(\boldsymbol{W}_v \boldsymbol{v}[i] + b_v)$$

Method: Numerical Reasoning

- **(2) Message Propagation**
  - The role a number plays in reasoning is not only decided by itself,
  - but also related to the context,
  - Propagate messages from each node to its neighbors to help to perform reasoning
  - Edges corresponding to different numerical relations should be distinguished
  - Use relation-specific transform matrices in the message propagation

Propagation Function

$$\widetilde{\boldsymbol{v}}_i' = \frac{1}{|\mathcal{N}_i|}\left(\sum_{j\in\mathcal{N}_i} \alpha_j \boldsymbol{W}^{\mathrm{r}_{ji}} \boldsymbol{v}[j]\right)$$

The message representation of node $\widetilde{\boldsymbol{v}}_i'$

r_{ji} **is the relation assigned to edge** e_{ji}

Relation-specific transform matrices $\boldsymbol{W}^{\mathrm{r}_{ji}}$

$\mathcal{N}_i = \{j|(v_j, v_i) \in \boldsymbol{E}\}$ **is the neighbors of node** v_i

For each edge e_{ji}, r_{ji} is determined by the following two attributes:

- Number relation: $>$ *or* $\leq$
- Node types: q-q, p-p, q-p, p-q

$$\mathrm{r}_{ij} \in \{>, \leq\} \times \{\text{q-q}, \text{p-p}, \text{q-p}, \text{p-q}\}$$

Method: Numerical Reasoning

- **(3) Node Representation Update**
  - As the message representation obtained in the previous step only contains information from the neighbors
  - It needs to be fused with the node representation to combine with the information carried by the node itself

$$\boldsymbol{v}_i' = \mathrm{ReLU}(\boldsymbol{W}_f \boldsymbol{v}_i + \widetilde{\boldsymbol{v}}_i' + \boldsymbol{b}_f)$$

  - The entire one-step reasoning process as a single function

$$\boldsymbol{v}' = \texttt{Reasoning-Step}(\mathcal{G}, \boldsymbol{v})$$

  - The graph constructed in Sec. 3.2 has encoded the numerical relations via its topology, the reasoning process is numerically-aware

Method: Numerical Reasoning

- **Multi-step Reasoning**
  - By single-step reasoning, we can only infer relations between adjacent nodes
  - Relations between multiple nodes may be required for certain task (e.g., sorting)

$$\boldsymbol{v}^t = \texttt{Reasoning-Step}(\boldsymbol{v}^{t-1}) \quad t \geq 1$$

  - Suppose we perform K steps of reasoning, $\boldsymbol{v}^K$ is used as U

Experiment

- **Dataset and Evaluation Metrics**
  - DROP dataset (Dua et al., 2019)
    - Require numerical reasoning such as addition, counting, or sorting over numbers in the passages
  - Exact Match (EM)
  - Numerically-focused F1 scores
    - Set to be 0 when the predicted answer is mismatched for those questions with the numeric golden answer

- **Baselines**
  - Semantic parsing models
    - Syn Dep (Dua et al., 2019), OpenIE (Dua et al., 2019), SRL (Dua et al., 2019)
  - Traditional MRC models
    - BiDAF (Seo et al., 2017), QANet (Yu et al., 2018), BERT (Devlin et al., 2019)
  - Numerical MRC models
    - NAQANet (Dua et al., 2019), NAQANet+

Experiment

- **Overall results**

Method	Dev		Test	
	EM	F1	EM	F1
Semantic Parsing				
Syn Dep	9.38	11.64	8.51	10.84
OpenIE	8.80	11.31	8.53	10.77
SRL	9.28	11.72	8.98	11.45
Traditional MRC				
BiDAF	26.06	28.85	24.75	27.49
QANet	27.50	30.44	25.50	28.36
BERT	30.10	33.36	29.45	32.70
Numerical MRC				
NAQANet	46.20	49.24	44.07	47.01
NAQANet+	61.47	64.85	60.82	64.29
NumNet	**64.92**	**68.31**	**64.56**	**67.97**
Human Performance	-	-	94.09	96.42

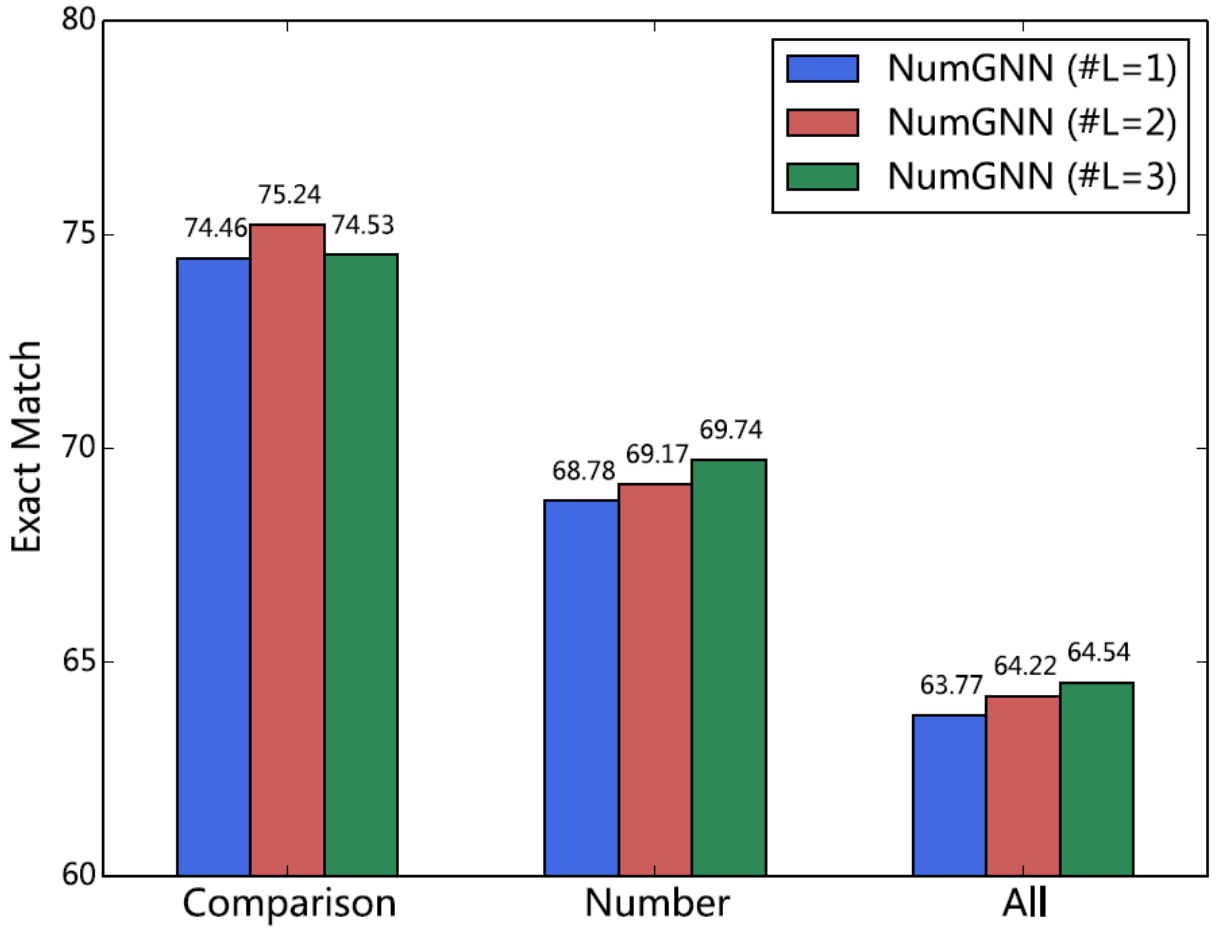
Experiment

- **Overall results**

Method	Comparison		Number		ALL	
	EM	F1	EM	F1	EM	F1
GNN	69.86	75.91	67.77	67.78	61.90	65.16
NumGNN	74.53	80.36	69.74	69.75	64.54	68.02
- question num	74.84	80.24	68.42	68.43	63.78	67.17
- $\leq$ type edge	74.89	80.51	68.48	68.50	63.66	67.06
- $>$ type edge	74.86	80.19	68.77	68.78	63.64	66.96

Method

- **Effect of GNN Structure**



Method

- **Case Study**

Question & Answer	Passage	NAQANet+	NumNet
Q: Which age group is larger: under the age of 18 or 18 and 24? **A:** 18 and 24	The median age in the city was 22.1 years. *10.1%* of residents were under the age of 18; *56.2%* were between the ages of 18 and 24; 16.1% were from 25 to 44; 10.5% were from 45 to 64; and 7% were 65 years of age or older. The gender makeup of the city was 64.3% male and 35.7% female.	under the age of 18	18 and 24
Q: How many more yards was Longwell's longest field goal over his second longest one? **A:** 26-22=4	... The Vikings would draw first blood with a *26-yard field goal* by kicker Ryan Longwell. In the second quarter, Carolina got a field goal with opposing kicker John Kasay. The Vikings would respond with another Longwell field goal (*a 22-yard FG*) ... In OT, Longwell booted the game-winning *19-yard field goal* to give Minnesota the win. It was the first time in Vikings history that a coach ...	26-19 = 7	26-22 = 4

Method

- **Error Analysis**

Question	Passage	Answer	NumNet
Which ancestral groups are at least 10%?	As of the census of 2000, there were 7,791 people, 3,155 households, and 2,240 families residing in the county. ... 33.7% were of *Germans*, 13.9% *Swedish* people, 10.1% *Irish* people, 8.8% United States, 7.0% English people and 5.4% Danish people ancestry ...	German; Swedish; Irish	Irish
Were more people 40 and older or 19 and younger?	Of Saratoga Countys population in 2010, *6.3%* were between ages of 5 and 9 years, *6.7%* between 10 and 14 years, *6.5%* between 15 and 19 years, ... , *7.9%* between 40 and 44 years, *8.5%* between 45 and 49 years, *8.0%* between 50 and 54 years, *7.0%* between 55 and 59 years, *6.4%* between 60 and 64 years, and *13.7%* of age 65 years and over ...	40 and older	19 and younger

Method

- **Ablation Studies**

Method	Comparison		Number		ALL	
	EM	F1	EM	F1	EM	F1
NAQANet+	69.11	75.62	66.92	66.94	61.11	64.54
- real number	66.87	73.25	45.82	45.85	47.82	51.22
- richer arithmetic expression	68.62	74.55	52.48	52.51	52.02	55.32
- passage-preferred	64.06	72.34	66.46	66.47	59.64	63.34
- data augmentation	65.28	71.81	67.05	67.07	61.21	64.60