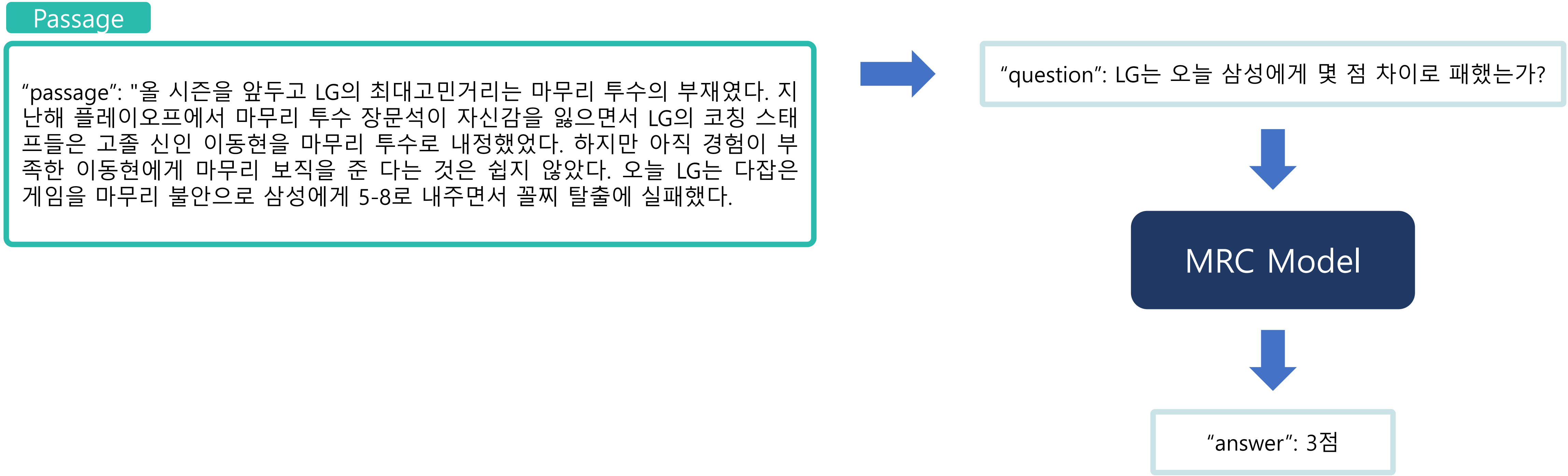
NumNet: Machine Reading Comprehension with Numerical Reasoning

(Qiu Ran et al, 2019)

김경민

들어가기 전에… MRC란?

➡ 주어진 Passage(context)를 보고 질문에 대한 정답을 하는 것



Passage

“passage”: "올 시즌을 앞두고 LG의 최대고민거리는 마무리 투수의 부재였다. 지난해 플레이오프에서 마무리 투수 장문석이 자신감을 잃으면서 LG의 코칭 스태프들은 고졸 신인 이동현을 마무리 투수로 내정했었다. 하지만 아직 경험이 부족한 이동현에게 마무리 보직을 준 다는 것은 쉽지 않았다. 오늘 LG는 다잡은 게임을 마무리 불안으로 삼성에게 5-8로 내주면서 꼴찌 탈출에 실패했다.

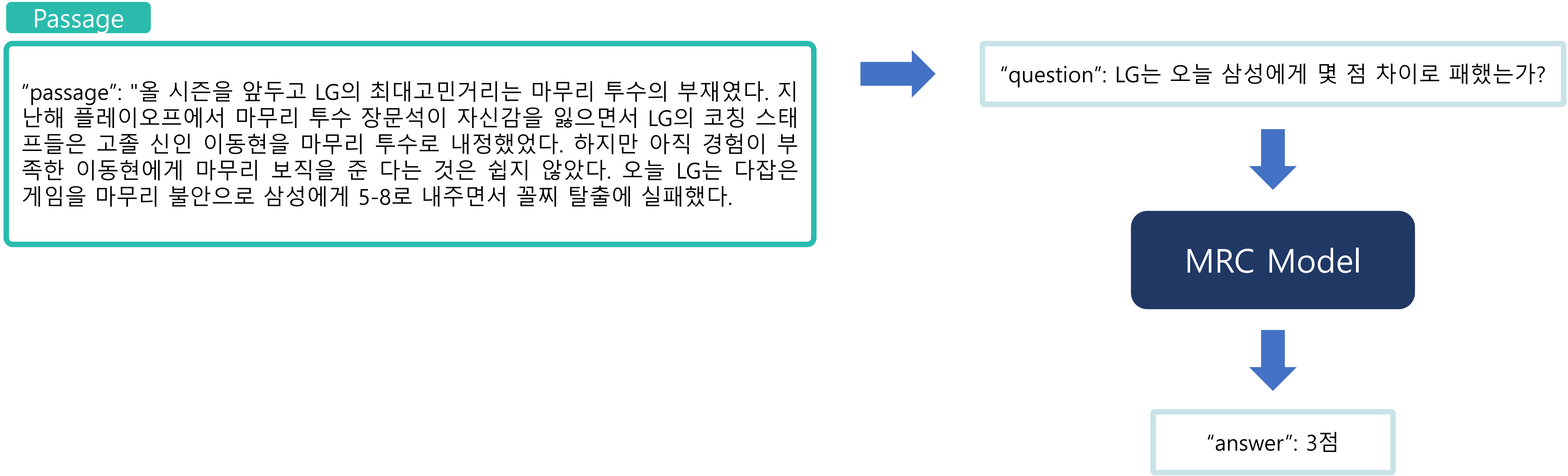
“question”: LG는 오늘 삼성에게 몇 점 차이로 패했는가?

MRC Model

“answer”: 3점

들어가기 전에... MRC란?

- 주어진 Passage(context)를 보고 질문에 대한 정답을 하는 것
- 한 단계 더 나아가, 주어진 Passage를 보고 질문에 대한 **(수치적 추론이 가능한)** 정답을 하는 것



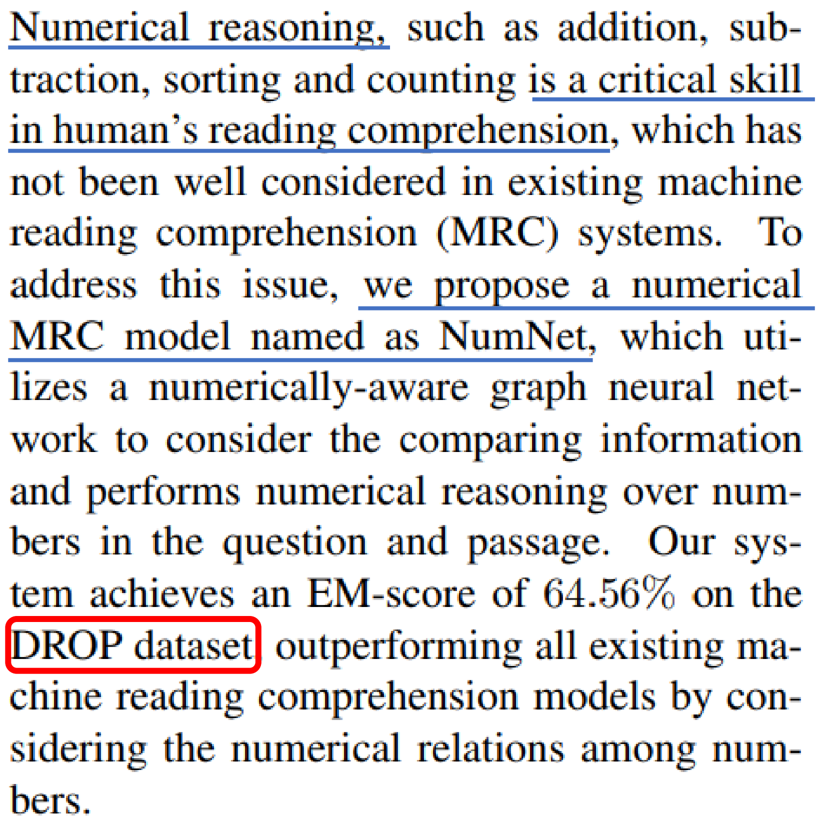
Passage

"passage": "올 시즌을 앞두고 LG의 최대고민거리는 마무리 투수의 부재였다. 지난해 플레이오프에서 마무리 투수 장문석이 자신감을 잃으면서 LG의 코칭 스태프들은 고졸 신인 이동현을 마무리 투수로 내정했었다. 하지만 아직 경험이 부족한 이동현에게 마무리 보직을 준 다는 것은 쉽지 않았다. 오늘 LG는 다잡은 게임을 마무리 불안으로 삼성에게 5-8로 내주면서 꼴찌 탈출에 실패했다.

"question": LG는 오늘 삼성에게 몇 점 차이로 패했는가?

"answer": 3점

Abstract



Numerical reasoning, such as addition, subtraction, sorting and counting is a critical skill in human's reading comprehension, which has not been well considered in existing machine reading comprehension (MRC) systems. To address this issue, we propose a numerical MRC model named as NumNet, which utilizes a numerically-aware graph neural network to consider the comparing information and performs numerical reasoning over numbers in the question and passage. Our system achieves an EM-score of 64.56% on the DROP dataset outperforming all existing machine reading comprehension models by considering the numerical relations among numbers.

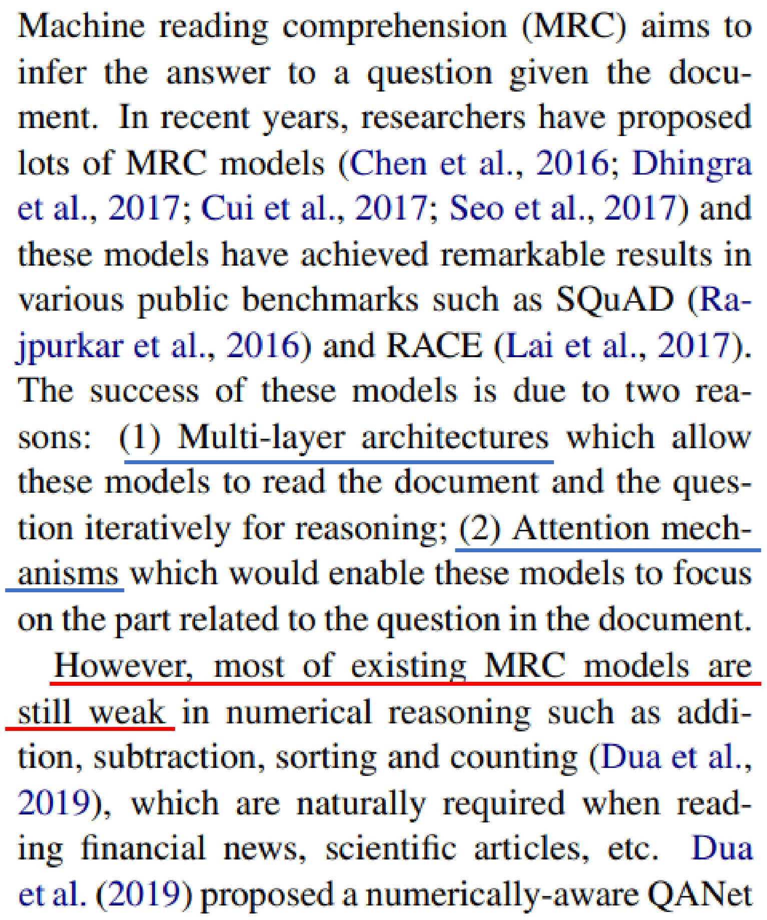
덧셈, 뺄샘, 정렬, 카운틱과 같은 수치적 추론은
인간의 독해 스킬에서 굉장히 중요하지만, **현재 존재하는**
MRC 시스템은 그 문제를 다루지 않는다.

↓

그래서 우리는 Numnet 을 제안하고,
이 모델은 **수치적 추론 문제**를 다룰 수 있다.

Introduction

Introduction



Machine reading comprehension (MRC) aims to infer the answer to a question given the document. In recent years, researchers have proposed lots of MRC models (Chen et al., 2016; Dhingra et al., 2017; Cui et al., 2017; Seo et al., 2017) and these models have achieved remarkable results in various public benchmarks such as SQuAD (Rajpurkar et al., 2016) and RACE (Lai et al., 2017). The success of these models is due to two reasons: (1) Multi-layer architectures which allow these models to read the document and the question iteratively for reasoning; (2) Attention mechanisms which would enable these models to focus on the part related to the question in the document.

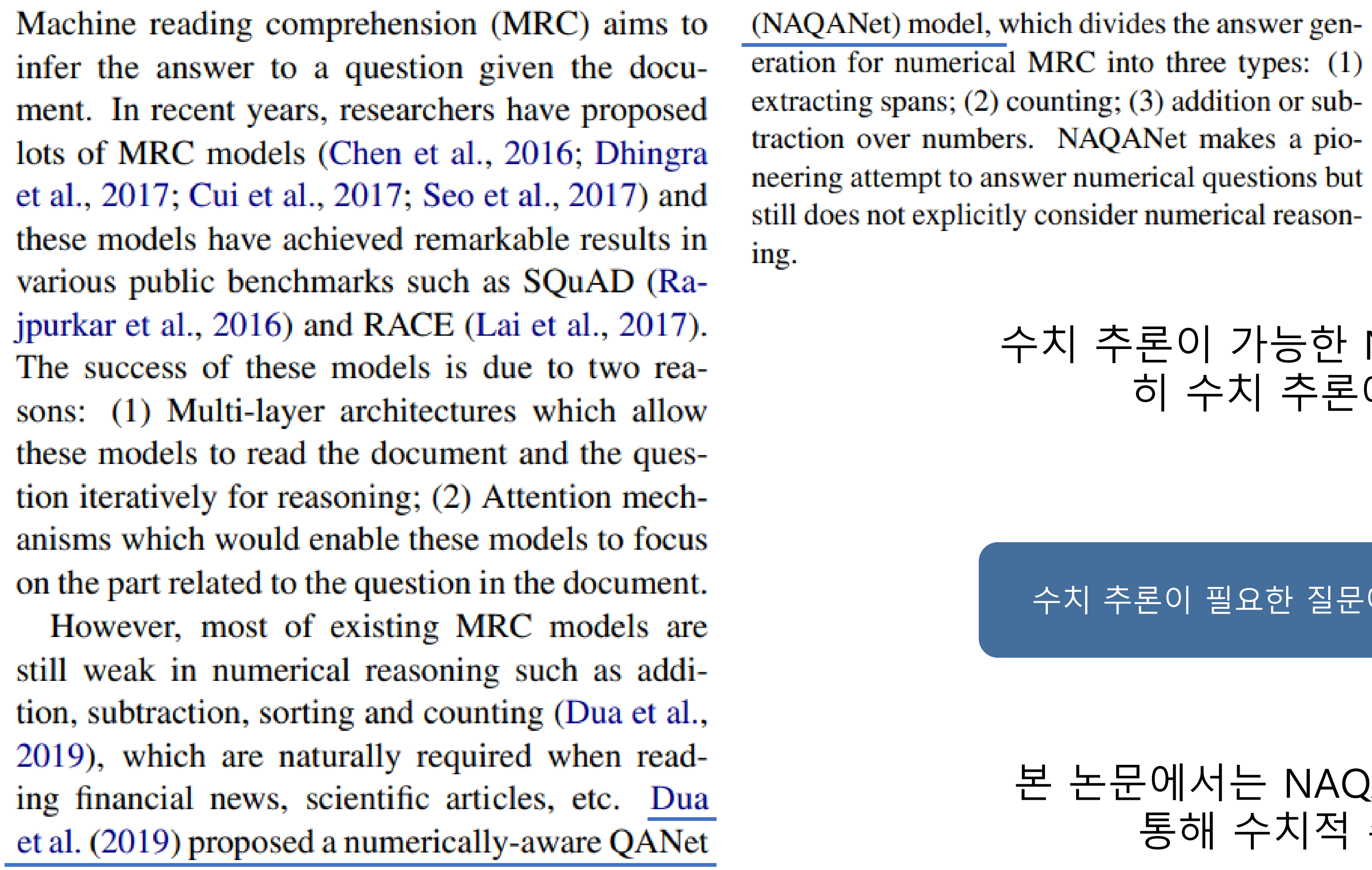
However, most of existing MRC models are still weak in numerical reasoning such as addition, subtraction, sorting and counting (Dua et al., 2019), which are naturally required when reading financial news, scientific articles, etc. Dua et al. (2019) proposed a numerically-aware QANet

최근 MRC 딥러닝 모델은 SQuAD, RACE와 같은 공개된 벤치마크 데이터셋에서 굉장히 좋은 성과를 보였는데, 그 성공의 이유는

1. Multi-layer 구조
2. Attention 메커니즘

그런데, 이러한 모델들도, **수치 추론**이 필요한 P&Q에서는 여전히 약하다.

Introduction



Machine reading comprehension (MRC) aims to infer the answer to a question given the document. In recent years, researchers have proposed lots of MRC models (Chen et al., 2016; Dhingra et al., 2017; Cui et al., 2017; Seo et al., 2017) and these models have achieved remarkable results in various public benchmarks such as SQuAD (Rajpurkar et al., 2016) and RACE (Lai et al., 2017). The success of these models is due to two reasons: (1) Multi-layer architectures which allow these models to read the document and the question iteratively for reasoning; (2) Attention mechanisms which would enable these models to focus on the part related to the question in the document.

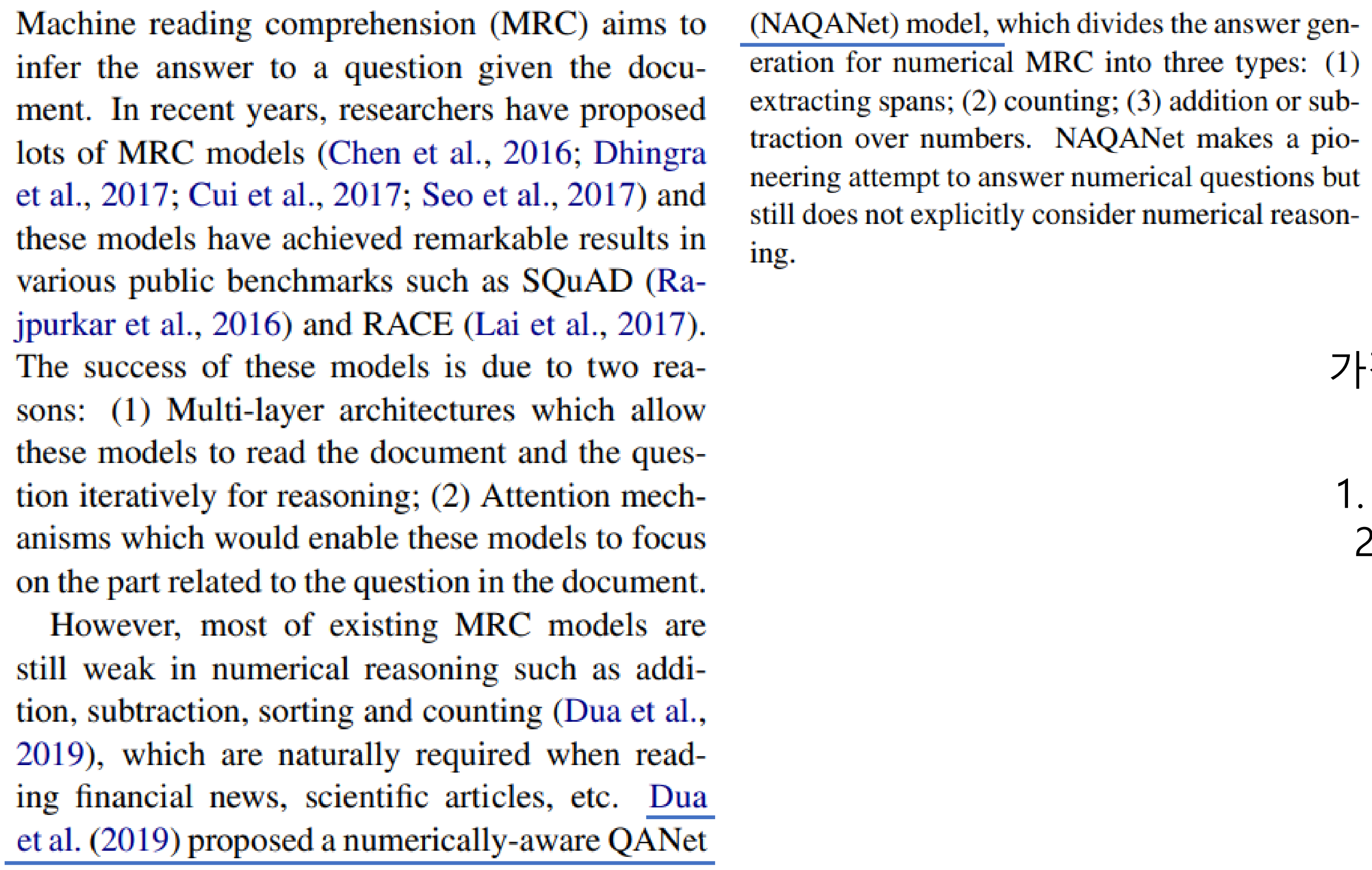
However, most of existing MRC models are still weak in numerical reasoning such as addition, subtraction, sorting and counting (Dua et al., 2019), which are naturally required when reading financial news, scientific articles, etc. Dua et al. (2019) proposed a numerically-aware QANet (NAQANet) model, which divides the answer generation for numerical MRC into three types: (1) extracting spans; (2) counting; (3) addition or subtraction over numbers. NAQANet makes a pioneering attempt to answer numerical questions but still does not explicitly consider numerical reasoning.

수치 추론이 가능한 NAQANet 모델은 있으나, 이 모델 또한 여전히 수치 추론에 대해 명시적으로 고려하지 않는다.

수치 추론이 필요한 질문에 답하기 위해서는, **수치 비교를 수행**할 수 있어야 한다.

본 논문에서는 NAQANet 모델을 사용해서, 숫자의 비교 과정을 통해 수치적 추론이 가능한 NumNet 모델을 제안

Introduction



Machine reading comprehension (MRC) aims to infer the answer to a question given the document. In recent years, researchers have proposed lots of MRC models (Chen et al., 2016; Dhingra et al., 2017; Cui et al., 2017; Seo et al., 2017) and these models have achieved remarkable results in various public benchmarks such as SQuAD (Rajpurkar et al., 2016) and RACE (Lai et al., 2017). The success of these models is due to two reasons: (1) Multi-layer architectures which allow these models to read the document and the question iteratively for reasoning; (2) Attention mechanisms which would enable these models to focus on the part related to the question in the document.

However, most of existing MRC models are still weak in numerical reasoning such as addition, subtraction, sorting and counting (Dua et al., 2019), which are naturally required when reading financial news, scientific articles, etc. Dua et al. (2019) proposed a numerically-aware QANet (NAQANet) model, which divides the answer generation for numerical MRC into three types: (1) extracting spans; (2) counting; (3) addition or subtraction over numbers. NAQANet makes a pioneering attempt to answer numerical questions but still does not explicitly consider numerical reasoning.

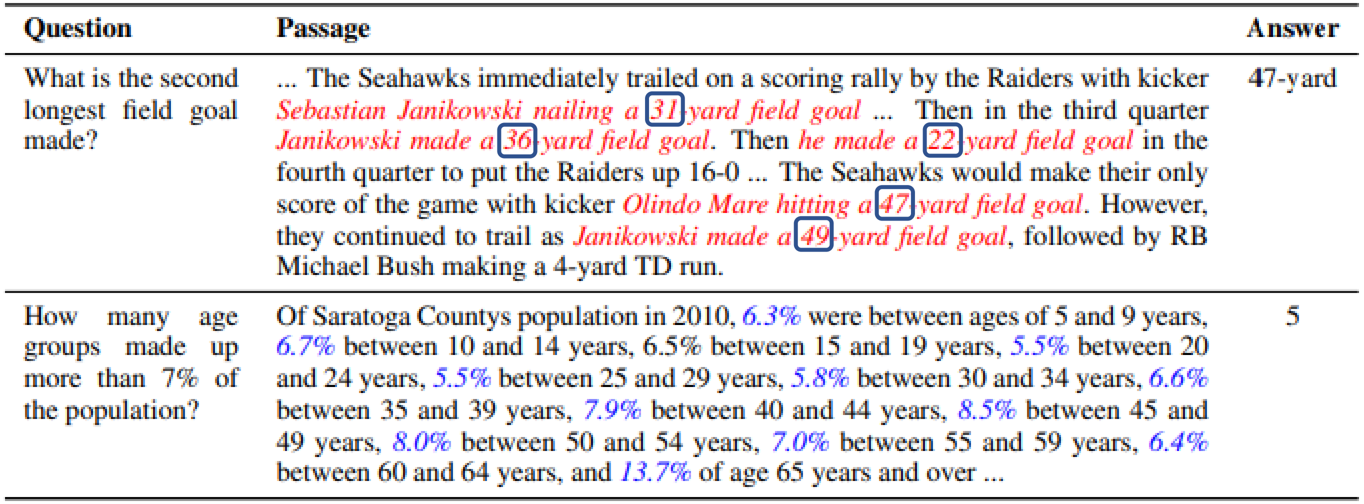
가장 크게 고려하는 두 가지

1. Numerical Comparison
2. Numerical Condition

Introduction

1. Numerical Comparison

만일 MRC가 "49 > 47 > 36 > 31 > 22" 을 알고있다면,
두 번째로 가장 긴 field goal이 47이라는 것을 answer로 답할 수 있을 것



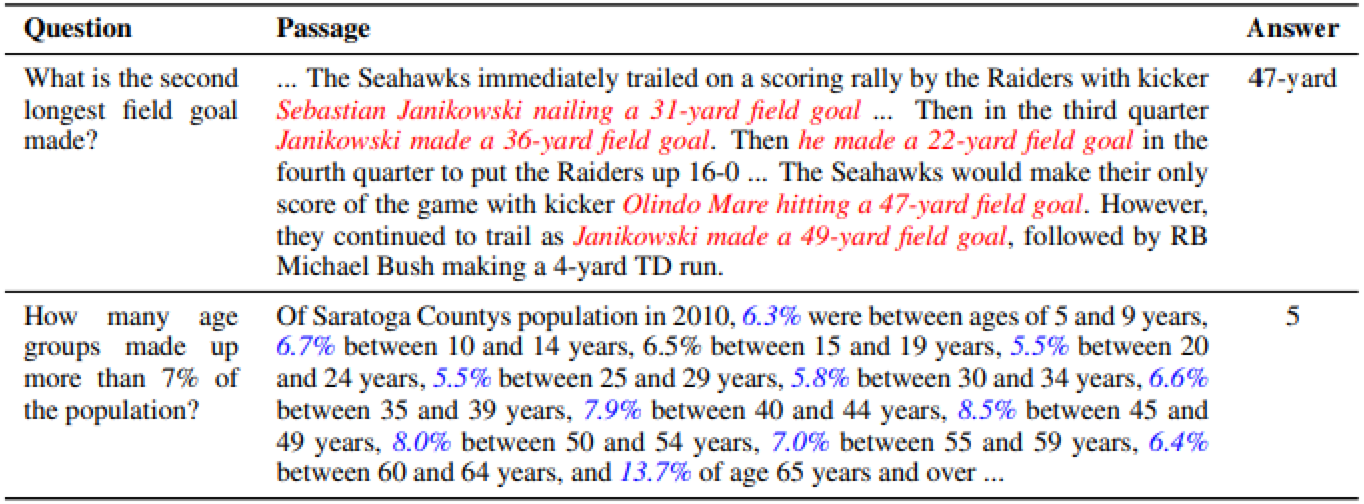
Question	Passage	Answer
What is the second longest field goal made?	... The Seahawks immediately trailed on a scoring rally by the Raiders with kicker *Sebastian Janikowski nailing a 31-yard field goal* ... Then in the third quarter *Janikowski made a 36-yard field goal*. Then *he made a 22-yard field 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
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

Table 1: Example questions from the DROP dataset which require numerical comparison. We highlight the relevant parts in the passage to infer the answer.

Introduction

2. Numerical Condition

질문에 해당하는 **그룹 수를 계산**하기 위해, 인구의 **7% 이상**을 구성하는 연령 그룹을 알아야 답할 수 있을 것



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
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

Table 1: Example questions from the DROP dataset which require numerical comparison. We highlight the relevant parts in the passage to infer the answer.

Related Work

MRC

- 데이터셋 : CNN/Daily Mail, SQuAD, RACE, Trivia-QA
- models : Attentive Reader, BiDAF, Interactive AoA Reader, Gated Attention Reader, R-Net, DCN, QANet

Arithmetic Word Problem Solving

- 기존 수학관련 연구들
- 결국 얘기하고자 하는 것 : 모든 존재하는 AWP 시스템은 적은 양의 벤치마크 데이터셋만 존재, AWP보다 real world 문제를 다루는 MRC가 더욱 챌린징하다.

Methodology

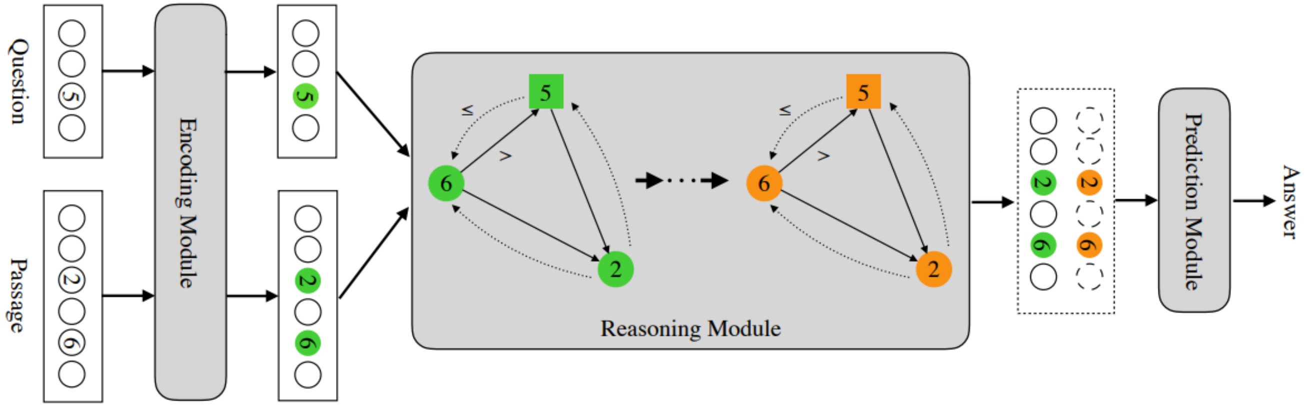
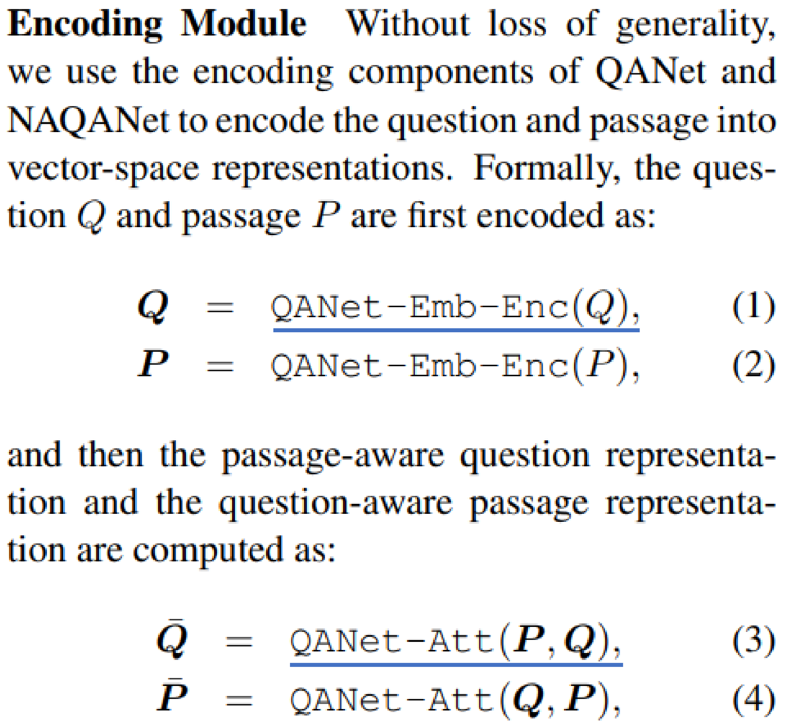


Figure 1: **The framework of our NumNet model.** Our model consists of an encoding module, a reasoning module and a prediction module. The numerical relations between numbers are encoded with the topology of the graph. For example, the edge pointing from "6" to "5" denotes "6" is greater than "5". And the reasoning module leverages a numerically-aware graph neural network to perform numerical reasoning on the graph. As numerical comparison is modeled explicitly in our model, it is more effective for answering questions requiring numerical reasoning such as addition, counting, or sorting over numbers.

Methodology



Encoding Module Without loss of generality, we use the encoding components of QANet and NAQANet to encode the question and passage into vector-space representations. Formally, the question Q and passage P are first encoded as:

$$\boldsymbol{Q} = \underline{\texttt{QANet-Emb-Enc}(Q),} \quad (1)$$

$$\boldsymbol{P} = \texttt{QANet-Emb-Enc}(P), \quad (2)$$

and then the passage-aware question representation and the question-aware passage representation are computed as:

$$\bar{\boldsymbol{Q}} = \underline{\texttt{QANet-Att}(\boldsymbol{P}, \boldsymbol{Q}),} \quad (3)$$

$$\bar{\boldsymbol{P}} = \texttt{QANet-Att}(\boldsymbol{Q}, \boldsymbol{P}), \quad (4)$$

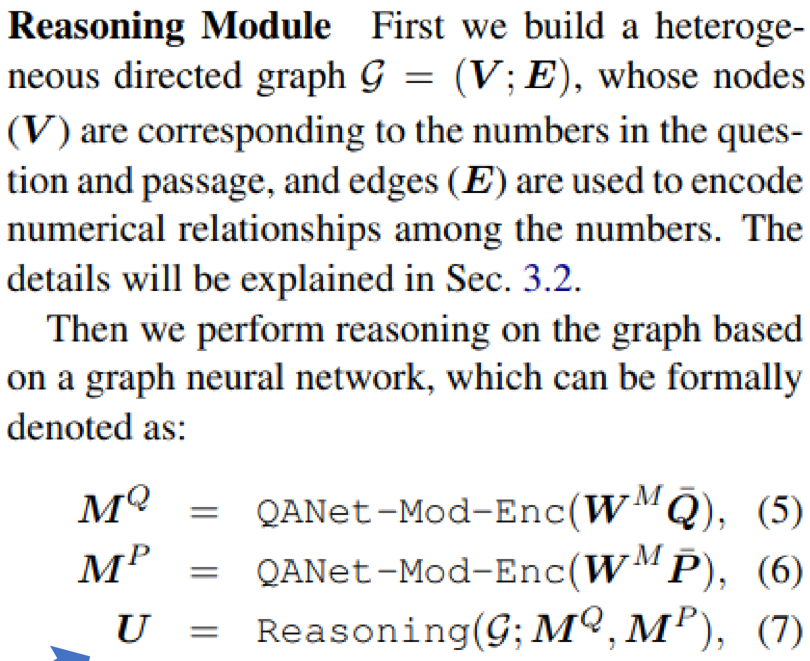
Three Layers

1. Convolution layer
2. Self-attention layer
3. Feed-forward layer

Attention Layer

Methodology

V = 노드(P&Q에 존재하는 숫자 정보)
E = 숫자 간 관계



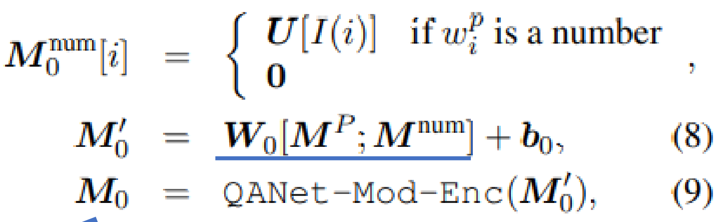
Reasoning Module First we build a heterogeneous directed graph $\mathcal{G} = (\boldsymbol{V}; \boldsymbol{E})$, whose nodes ($\boldsymbol{V}$) are corresponding to the numbers in the question and passage, and edges ($\boldsymbol{E}$) are used to encode numerical relationships among the numbers. The details will be explained in Sec. 3.2.

Then we perform reasoning on the graph based on a graph neural network, which can be formally denoted as:

$$\begin{aligned} \boldsymbol{M}^Q &= \texttt{QANet-Mod-Enc}(\boldsymbol{W}^M \bar{\boldsymbol{Q}}), && (5) \\ \boldsymbol{M}^P &= \texttt{QANet-Mod-Enc}(\boldsymbol{W}^M \bar{\boldsymbol{P}}), && (6) \\ \boldsymbol{U} &= \texttt{Reasoning}(\mathcal{G}; \boldsymbol{M}^Q, \boldsymbol{M}^P), && (7) \end{aligned}$$

U : 숫자 정보를
표현한 것

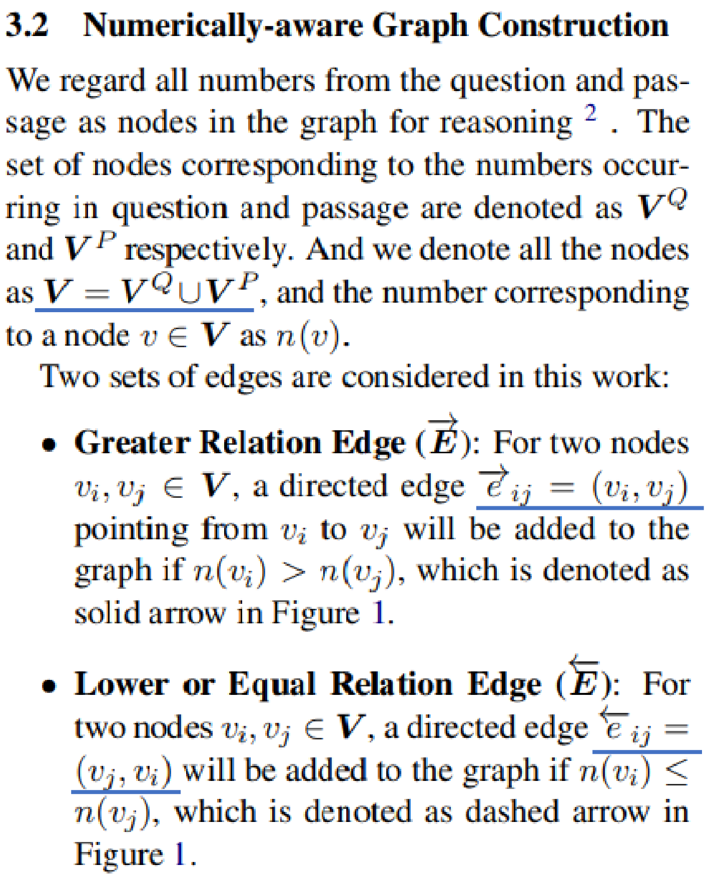
$I(i)$: passage 단어 중 숫자 word에
해당하는 노드 인덱스



$$\begin{aligned} \boldsymbol{M}_0^{\text{num}}[i] &= \begin{cases} \boldsymbol{U}[I(i)] & \text{if } w_i^p \text{ is a number} \\ \boldsymbol{0} \end{cases}, \\ \boldsymbol{M}_0' &= \underline{\boldsymbol{W}_0[\boldsymbol{M}^P; \boldsymbol{M}^{\text{num}}]} + \boldsymbol{b}_0, && (8) \\ \boldsymbol{M}_0 &= \texttt{QANet-Mod-Enc}(\boldsymbol{M}_0'), && (9) \end{aligned}$$

M_0 = numerically-aware passage representation

Methodology



3.2 Numerically-aware Graph Construction

We regard all numbers from the question and passage as nodes in the graph for reasoning [2] . The set of nodes corresponding to the numbers occurring in question and passage are denoted as $\boldsymbol{V}^Q$ and $\boldsymbol{V}^P$ respectively. And we denote all the nodes as $\boldsymbol{V} = \boldsymbol{V}^Q \cup \boldsymbol{V}^P$, and the number corresponding to a node $v \in \boldsymbol{V}$ as $n(v)$.

Two sets of edges are considered in this work:

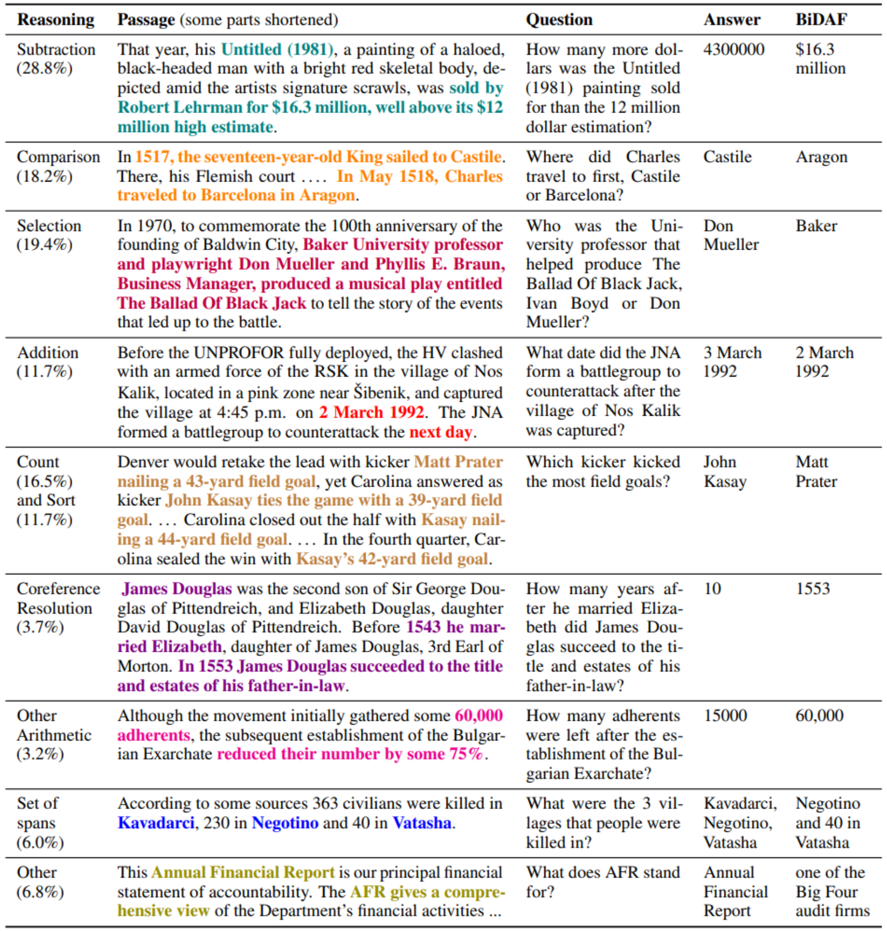
- **Greater Relation Edge ($\overrightarrow{\boldsymbol{E}}$):** For two nodes $v_i, v_j \in \boldsymbol{V}$, a directed edge $\overrightarrow{e}_{ij} = (v_i, v_j)$ pointing from v_i to v_j will be added to the graph if $n(v_i) > n(v_j)$, which is denoted as solid arrow in Figure 1.
- **Lower or Equal Relation Edge ($\overleftarrow{\boldsymbol{E}}$):** For two nodes $v_i, v_j \in \boldsymbol{V}$, a directed edge $\overleftarrow{e}_{ij} = (v_j, v_i)$ will be added to the graph if $n(v_i) \leq n(v_j)$, which is denoted as dashed arrow in Figure 1.

V^P, V^Q : P & Q의 숫자 집합

$V = V^Q \cup V^P$

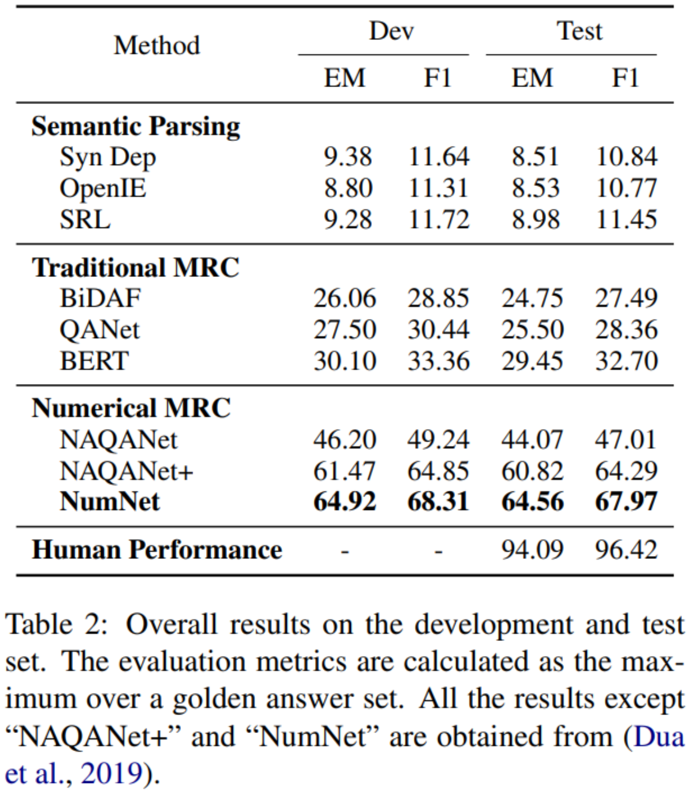
Experiments

Datasets：**DROP (Discrete Reasoning Over Paragraphs)**



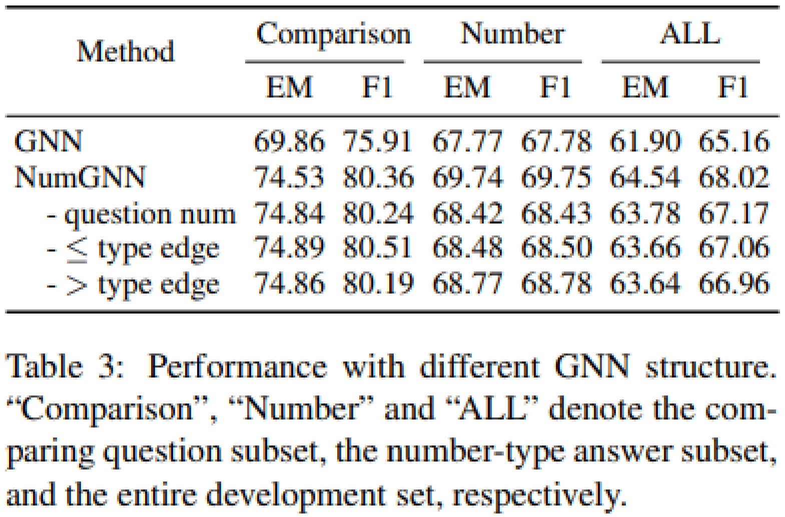
Reasoning	**Passage** (some parts shortened)	**Question**	**Answer**	**BiDAF**
Subtraction (28.8%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million
Comparison (18.2%)	In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile	Aragon
Selection (19.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, **Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack** to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller	Baker
Addition (11.7%)	Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on **2 March 1992**. The JNA formed a battlegroup to counterattack the **next day**.	What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?	3 March 1992	2 March 1992
Count (16.5%) and Sort (11.7%)	Denver would retake the lead with kicker **Matt Prater nailing a 43-yard field goal**, yet Carolina answered as kicker **John Kasay ties the game with a 39-yard field goal**. ... Carolina closed out the half with **Kasay nailing a 44-yard field goal**. ... In the fourth quarter, Carolina sealed the win with **Kasay's 42-yard field goal**.	Which kicker kicked the most field goals?	John Kasay	Matt Prater
Coreference Resolution (3.7%)	**James Douglas** was the second son of Sir George Douglas of Pittendreich, and Elizabeth Douglas, daughter David Douglas of Pittendreich. Before **1543 he married Elizabeth**, daughter of James Douglas, 3rd Earl of Morton. **In 1553 James Douglas succeeded to the title and estates of his father-in-law**.	How many years after he married Elizabeth did James Douglas succeed to the title and estates of his father-in-law?	10	1553
Other Arithmetic (3.2%)	Although the movement initially gathered some **60,000 adherents**, the subsequent establishment of the Bulgarian Exarchate **reduced their number by some 75%**.	How many adherents were left after the establishment of the Bulgarian Exarchate?	15000	60,000
Set of spans (6.0%)	According to some sources 363 civilians were killed in **Kavadarci**, 230 in **Negotino** and 40 in **Vatasha**.	What were the 3 villages that people were killed in?	Kavadarci, Negotino, Vatasha	Negotino and 40 in Vatasha
Other (6.8%)	This **Annual Financial Report** is our principal financial statement of accountability. The **AFR gives a comprehensive view** of the Department's financial activities ...	What does AFR stand for?	Annual Financial Report	one of the Big Four audit firms

Experiments



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

Table 2: Overall results on the development and test set. The evaluation metrics are calculated as the maximum over a golden answer set. All the results except "NAQANet+" and "NumNet" are obtained from (Dua et al., 2019).



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

Table 3: Performance with different GNN structure. "Comparison", "Number" and "ALL" denote the comparing question subset, the number-type answer subset, and the entire development set, respectively.

결과 : 그래프 정보를 고려한 NumGNN을 사용했을 때 성능 향상이 크다.

⇒ 수치 정보를 비교한다는 것은 결국 수치 추론에 효과적으로 도움이 될 수 있다.

+ 그래프에서 비교 관계를 같이 인코딩하는 것도 모델 성능 향상에 도움이 된다.

Experiments

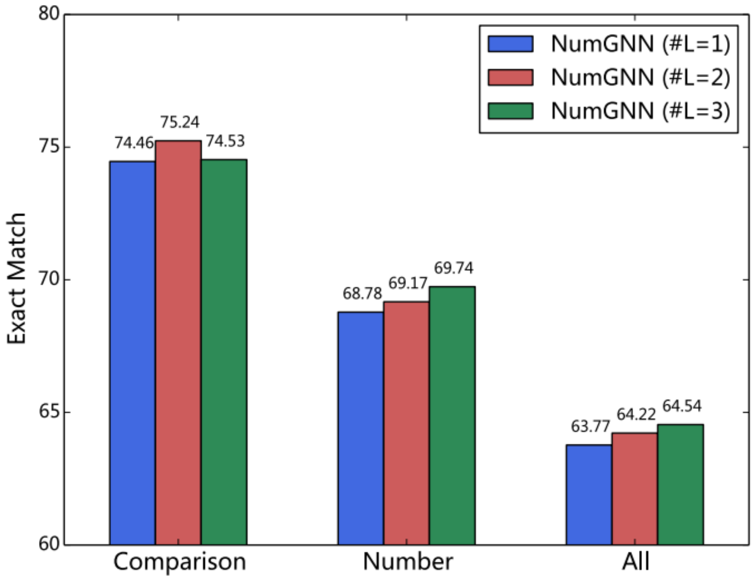


Figure 2: Effect of GNN layer numbers (# L).

비교하는 질문에 대해서는 NumGNN (L=2) 에서 최고 성능

- ex:) 두 번째로 나이가 많은 선수는 누구인가?

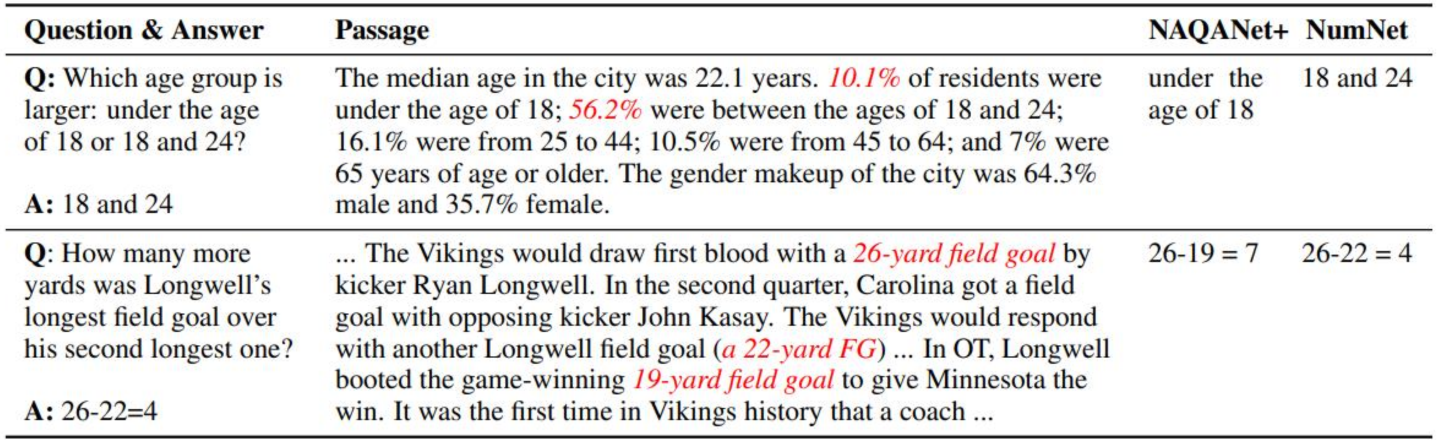
그 외 나머지 dev-set 에서는 레이어 수와 동일하게 성능이 증가

그러나,

graph reasoning step (K)이 4 이상에선 성능의 안정성 떨어짐 => GNN 그래프의 본질적인 over smoothing 문제로 판단

Experiments

Case study: Passage의 숫자 정보를 통합할 경우, 왜 더 좋은 answer를 생성할 수 있는지



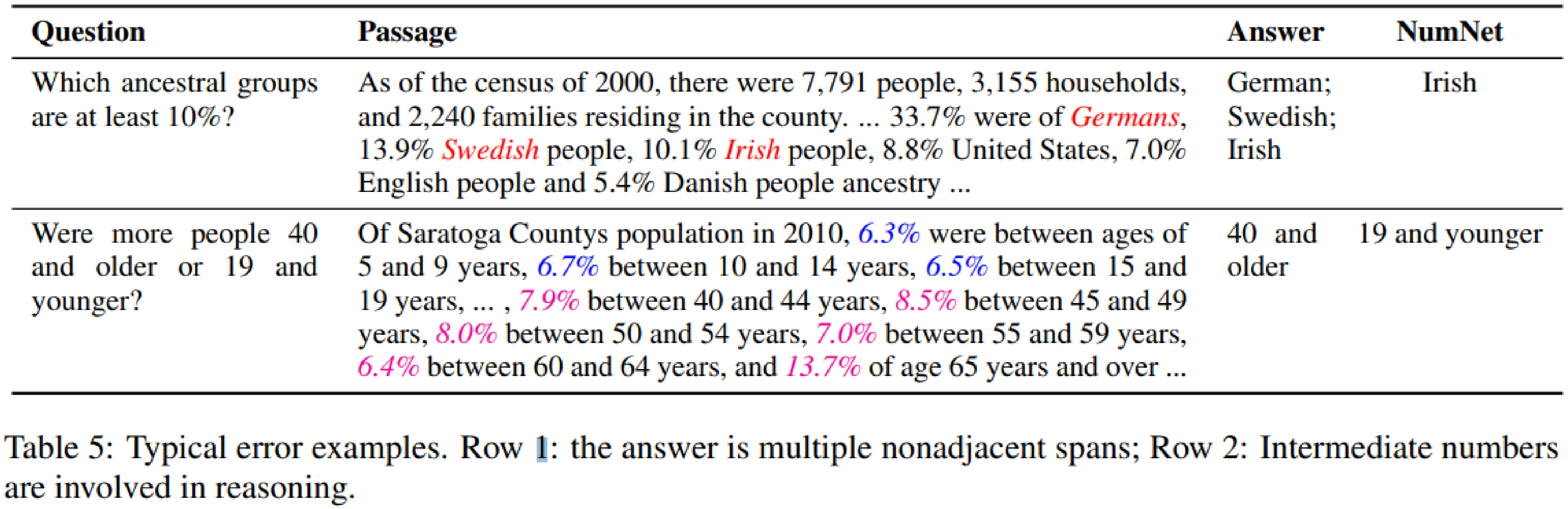
Question & Answer	Passage	NAQANet+	NumNet
Q: Which age group is larger: under the age of 18 or 18 and 24? **A:** 18 and 24	The median age in the city was 22.1 years. *10.1%* of residents were under the age of 18; *56.2%* were between the ages of 18 and 24; 16.1% were from 25 to 44; 10.5% were from 45 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

Table 4: Cases from the DROP dataset. We demonstrate the predictions of NAQANet+ and our NumNet model. Note that the two models only output the arithmetic expressions but we also provide their results for clarity.

결국 Passage의 수치 정보가 중요

Experiments

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

Table 5: Typical error examples. Row 1: the answer is multiple nonadjacent spans; Row 2: Intermediate numbers are involved in reasoning.

한계:

1. 다수의 인접하지 않은 span이 정답일 경우
2. P & Q에 존재하지 않지만 연산을 통해 정답을 생성해내야하는 경우

Discussion & Conclusion

- Numerically-aware 그래프와 NumGNN의 결합으로 NumNet은 수치 추론 능력을 할 수 있음
- Numerically-aware 그래프는 숫자를 노드로 인코딩, 숫자 비교에 필요한 엣지로 관계를 인코딩

But,

아직 미흡한 한계점이 존재

1. 미리 정의된 reasoning 그래프를 사용하므로 그래프에 표시되지 않은 중간 숫자가 포함된 추론이 어려움 (동적 그래프 필요)
2. AWP에 비해 한정적인 산술 표현식 처리만 가능
3. 더욱 정교한 symbolic reasoning 방법

Q & A