Natural Language Processing & Artificial Intelligence

Discrete Reasoning Machine Reading Comprehension

NLP&AI LAB

김경민

Machine Reading Comprehension

- Discrete Reasoning

Numerical Reasoning - Numnet

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 않음

↓

수치 추론이 필요한 질문에 답하기 위해
숫자 간 **수치 비교를 수행**할 수 있어야 한다

↓

본 논문에서는 NAQANet 모델 기반의 숫자의 비교 과정을 통한
수치 추론이 가능한 NumNet 모델을 제안

가장 크게 고려하는 두 가지

1. Numerical Comparison
2. Numerical Condition

Numerical Reasoning - Numnet

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.

Numerical Reasoning - Numnet

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.

Numerical Reasoning - Numnet

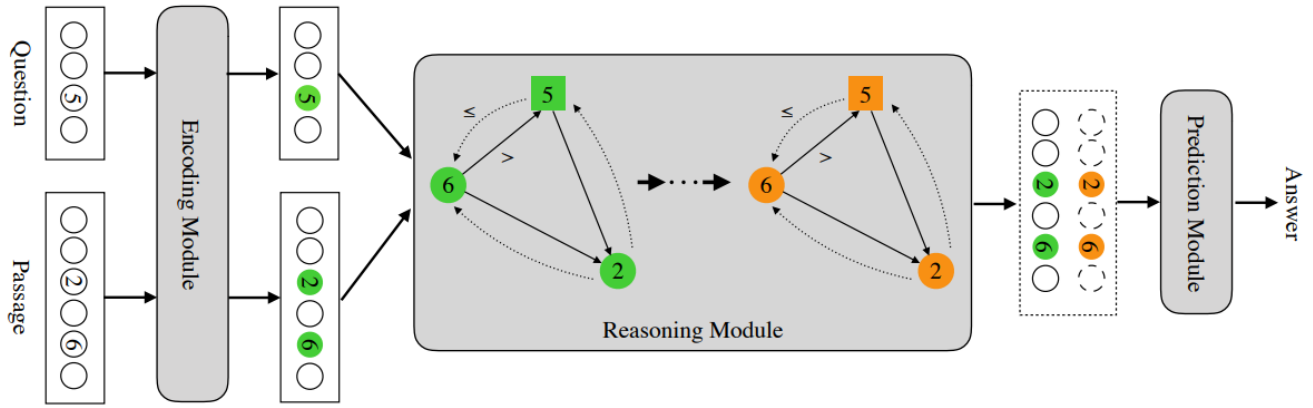


Figure 1: **The framework of our NumNet model.** Our model consists of an encoding module, a reasoning module and a prediction module. The numerical relations between numbers are encoded with the topology of the graph. For example, the edge pointing from "6" to "5" denotes "6" is greater than "5". And 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.

Numerical Reasoning - Numnet

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:

$$\boldsymbol{Q} = \texttt{QANet-Emb-Enc}(Q), \quad (1)$$
$$\boldsymbol{P} = \texttt{QANet-Emb-Enc}(P), \quad (2)$$

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

$$\bar{\boldsymbol{Q}} = \texttt{QANet-Att}(\boldsymbol{P}, \boldsymbol{Q}), \quad (3)$$
$$\bar{\boldsymbol{P}} = \texttt{QANet-Att}(\boldsymbol{Q}, \boldsymbol{P}), \quad (4)$$

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

4. Attention Layer

Numerical Reasoning - Numnet

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:

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

U : 숫자 정보

$I(i)$: number node index among passages

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

M_0 = numerically-aware passage representation

Numerical Reasoning - Numnet

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.

⇒ NumGNN을 활용하여 수치 정보를 비교한다는 것은 결국 수치 추론에 효과적

+ 그래프에서 비교 관계를 같이 인코딩하는 것도 모델 성능 향상에 도움

Numerical Reasoning - QDGAT

Machine reading comprehension (MRC) aims to develop AI models that can answer questions for text documents. Recently, the performance of MRC in public datasets has been improved dramatically due to the advanced pre-trained models, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2019).

However, pre-trained models are not explicitly aware of the concepts of numerical reasoning since numeracy supervision signals are rarely available during pre-training. The representations from these pre-trained models fall short in their ability to support downstream numerical reasoning. Yet such ability is critical for the comprehension of financial news and scientific articles, since basic numerical operations, such as addition, subtraction, sorting and counting, need to be conducted to extract the essential information (Dua et al., 2019).

PLM 모델은 Pre-trained 단계에서
수치 추론의 개념이 명시적으로 표현되지 않음

↓

PLM 모델의 표현만으로는 다운스트림 수치 추론을
지원하는 능력이 부족하지만, 수치 추론을 요구하는
경제 뉴스, 과학 기사에서는 굉장히 중요한 능력

Numerical Reasoning - QDGAT

dor et al., 2019; Geva et al., 2020; Chen et al., 2020), we argue that NumNet is insufficient for sophisticated numerical reasoning, since it lacks two critical ingredients for numerical reasoning:

1. **Number Type and Entity Mention.** The number comparison graph in NumNet is not able to identify different number types, and lacks the information of entities mentioned in the document that connect the number nodes.
2. **Direct Interaction with Question.** The graph reasoning module in NumNet leaves out the direct question representation, which may encounter difficulties in locating important numbers directed by the question as the pivot for numerical reasoning.

Numnet에 생략된 가장 중요한 두 가지

- Number Type and Entity Mention
- Direct Interaction with Question

1. Numnet의 숫자 비교 그래프는 다른 유형의 숫자 정보를 구분하지 않고, 숫자 노드 간 엔티티 부족
2. Numnet의 그래프 추론 모듈은 직접적인 질문 표현 (question representation) 을 생략, 이는 질문이 가리키는 중요한 숫자를 찾는 데 어려움

정확한 수치 정보 고려를 위해서는 서로 다른 유형의 숫자 정보 구분 능력 & 추가적인 엔티티 노드
Question representation 까지 고려한 추론 모듈 생성

Numerical Reasoning – QDGAT

Table 1: Two MRC cases requiring numerical reasoning are illustrated. There are entities and numbers of different types. Both are emphasized by different colors: entity, number, percentage, date, ordinal. We explicitly encode the type information into our model and leverage the question representation to conduct the reasoning process.

Question	Passage	Answer
At the battle of Caiboaté how many Spanish and Portuguese were injured or killed?	... In 1754 Spanish and Portuguese military forces were dispatched to force the Guarani to leave the area ... Hostilities resumed in 1756 when an army of 3,000 Spanish, Portuguese, and native auxiliary soldiers under José de Andonaegui and Freire de Andrade was sent to subdue the Guarani rebels. On February 7, 1756 the leader of the Guarani rebels, Sepé Tiaraju, was killed in a skirmish with Spanish and Portuguese troops. ... 1,511 Guarani were killed and 152 taken prisoner, while 4 Spanish and Portuguese were killed and about 30 were wounded...	34
In which quarter did Stephen Gostkowski kick his shortest field goal of the game?	The Cardinals' east coast struggles continued in the second quarter as quarterback Matt Cassel completed a 15-yard touchdown pass to running back Kevin Faulk and an 11-yard touchdown pass to wide receiver Wes Welker, followed by kicker Stephen Gostkowski's 38-yard field goal. In the third quarter, Arizona's deficit continued to climb as Cassel completed a 76-yard touchdown pass to wide receiver Randy Moss, followed by Gostkowski's 35- and 24-yard field goal. In the fourth quarter, New England concluded its domination with Gostkowski's 30-yard	third

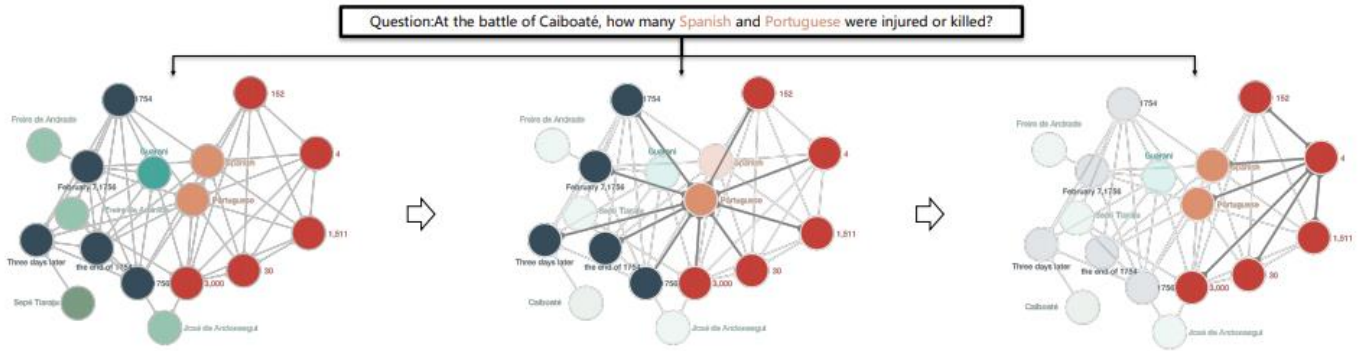
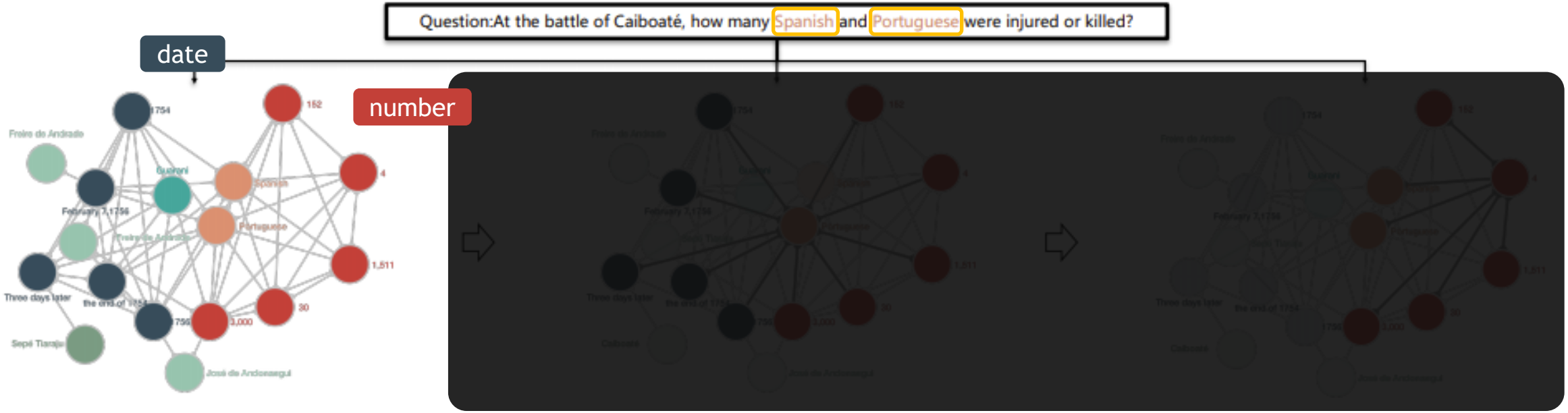
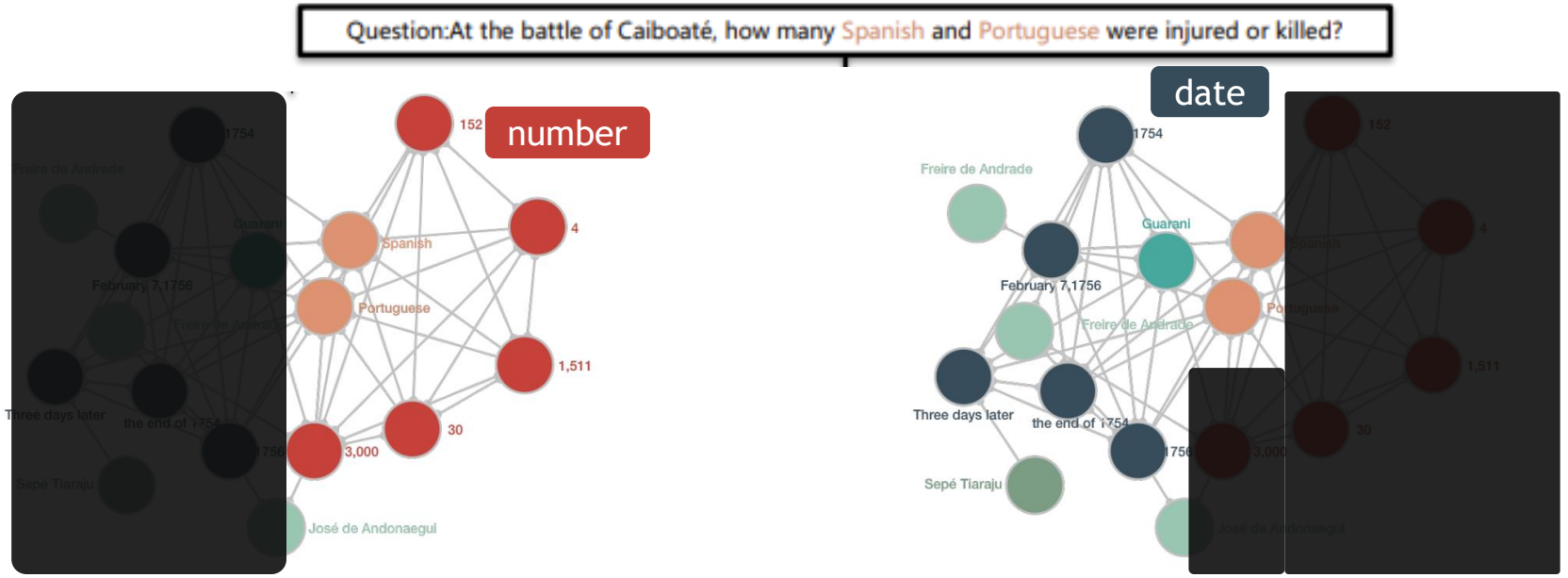


Figure 1: The constructed heterogeneous typed graph of the example in Table 1 is illustrated on the left. The red (dark blue) nodes are the numbers (dates) and the others are entities. The edges encode the relations among the numbers and entities: (1) The numbers with the same number type, e.g., date, are wired together. (2) The graph connects the numbers and the entities that are in the same sentence to indicate their co-occurrence. In the first round, the model pays attention to a sub-graph that contains the *Spanish* and *Portuguese* entities since they are mentioned in the question. In the update, the model learns to distinguish between the numbers and the dates and extracts the numbers related to the question. In the second round, the representations of the numbers are updated by the messages from the entities as well as the question to conduct the reasoning.

Numerical Reasoning - QDGAT

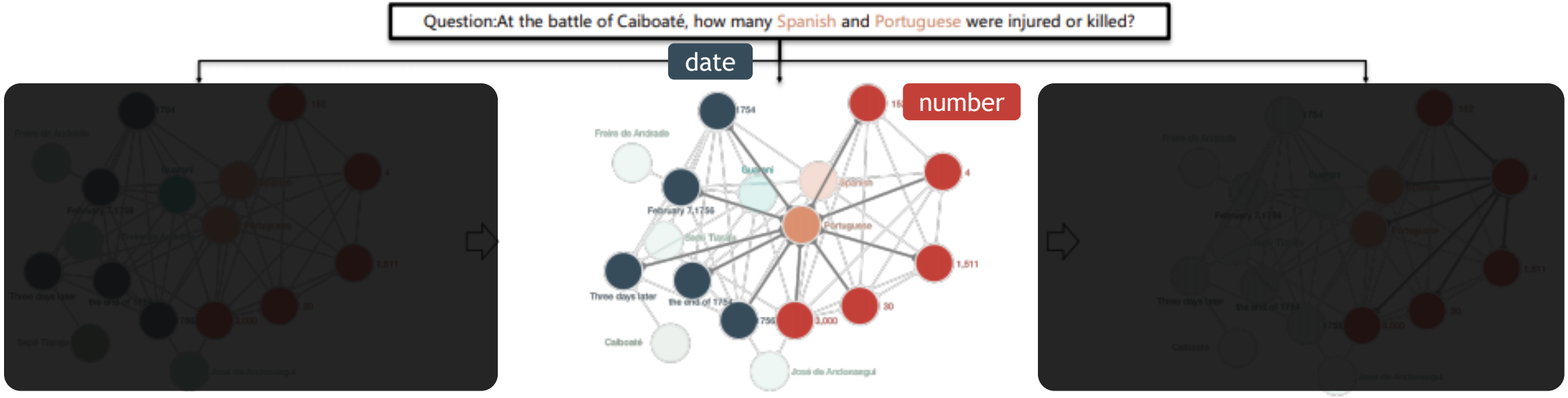


Numerical Reasoning - QDGAT



Question에 언급된 Entity를 포함하고있는 그래프에
집중해서 모델이 number와 date를 식별하도록 학습

Numerical Reasoning - QDGAT



Numerical Reasoning - QDGAT

Question:At the battle of Caiboaté, how many Spanish and Portuguese were injured or killed?

Representations of the numbers are updated by the messages from the entities as well as the question to conduct the reasoning

Numerical Reasoning – QDGAT

Method	Dev		Test	
	EM	F1	EM	F1
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
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
NAQANet	46.20	49.24	44.07	47.01
ALBERT-Calculator	80.22	83.98	79.85	83.56
NumNet	64.92	68.31	64.56	67.97
NumNet+ (RoBERTa)	81.07†	84.42†	81.52†	84.84†
NumNet+ (ensemble)	82.63†	85.59†	83.14†	86.16†
QDGAT (RoBERTa)	**82.74**	**85.85**	**83.23**	**86.38**
QDGAT$_p$ (RoBERTa)	**84.07**	**87.05**	**84.53**	**87.57**
QDGAT$_p$ (ensemble)	**85.31**	**88.10**	**85.46**	**88.38**
Human			94.09	96.42

Semantic Parsing

Traditional MRC

Numerical MRC

Numerical Reasoning – QDGAT

- Numnet – NAQANET + GNN
- QDGAT – RoBERTa + GAT

Possibility for performance improvement

1. Use pre-defined reasoning graph
   => dynamic graph
2. Span extraction for numerical reasoning resources can still be improved
3. Intrinsic over smoothing problem of GNN

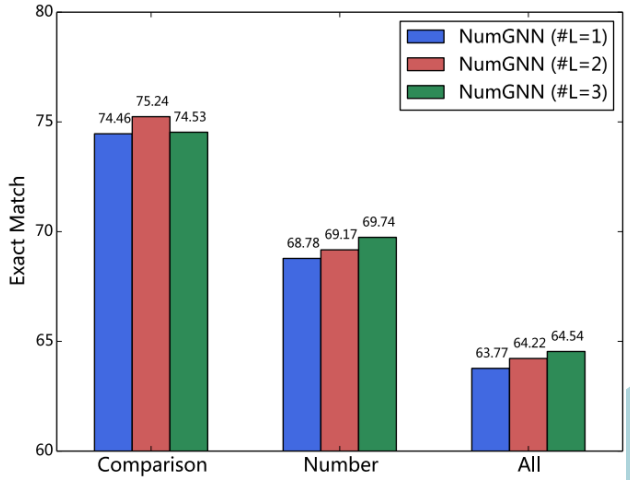


Figure 2: Effect of GNN layer numbers (# L).

Numerical Reasoning

As per the study in the cognitive system - “this abstract, notation-independent appreciation of numbers develops gradually over the first several years of life ... human infants appreciate numerical quantities at a non-symbolic level: They know approximately how many objects they see before them even though they do not understand number words or Arabic numerals.”, the concept of discrete number is gradually developed through the real-life experience (Cantlon et al., 2009).

“숫자 인식은 생후 첫 몇 년간 점진적으로 발전합니다. 유아기의 인간은 기호가 없는 수준에서 숫자의 수량을 인식합니다. 숫자를 이해하지 못하더라도 본인이 얼마나 많은 사물을 봤는지, 물체의 수를 대략적으로 알고 있습니다.”

Q & A