Numerical Reasoning in Question Answering

Department of Computer Science, Yonsei University

Seungone Kim

louisdebroglie@yonsei.ac.kr

Referenced Papers

- Prerequisites / Additional Papers
  - Gated self-matching networks for reading comprehension and question answering [ACL 2017]
  - Bidirectional attention flow for machine comprehension [ICLR 2017]
  - Simple and effective multi-paragraph reading comprehension [ACL 2018]
  - Fast and accurate reading comprehension by combining self-attention and convolution [ICLR 2018]
  - Reinforced mnemonic reader for machine reading comprehension [IJCAI 2018]
  - Drop : A reading comprehension benchmark requiring discrete reasoning over paragraphs [NAACL 2019]

  - HotpotQA : A dataset for diverse, explainable multi-hop question answering [EMNLP 2018]
  - HoVer : A dataset for many-hop fact extraction and claim verification [EMNLP 2020]
  - HybridQA : A dataset of multi-hop question answering over tabular and textual data [EMNLP 2020]
  - Task-oriented dialogue as dataflow synthesis [TACL 2020]
  - TaPas : Weakly supervised table parsing via pre-training [ACL 2020]
  - Span selection pretraining for question answering [ACL 2020]

- Key Papers
  - A Multi-Type Multi-Span Network for Reading Comprehension that Requires Discrete Reasoning [EMNLP-IJCNLP 2019]
  - Neural Symbolic Reader : Scalable Integration of Distributed and Symbolic Representations for Reading Comprehension [ICLR 2020]
  - ReasonBERT : Pre-trained to Reason with Distant Supervision [EMNLP 2021]

A Multi-Type Multi-Span Network for Reading Comprehension that Requires Discrete Reasoning

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li
National University of Defense Technology, Changsha, China
{huminghao09,pengyuxing,huangzhen,dsli}@nudt.edu.cn

EMNLP-IJCNLP 2019

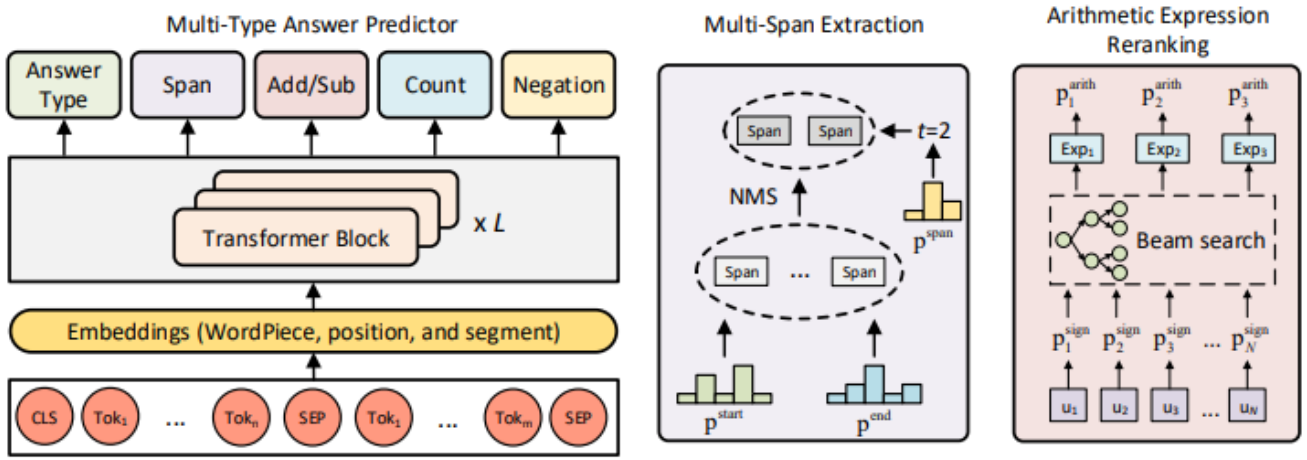
A Multi-Type Multi-Span Network for Reading Comprehension that Requires Discrete Reasoning [EMNLP-IJCNLP 2019]

- The performance of RC model degrades significantly when they are applied to **more realistic scenarios**
  - Unlike SQuAD and CNNDM, DROP is substantially more challenging in three ways.
  - Ex1) Answers are involved with **various types**
  - Ex2) **Multiple text strings** are correct answers
  - Ex3) **Discrete Reasoning** abilities are required.

- Existing approaches(NAQANet) has three problems
  - Although extending one-type answer prediction to **multi-type prediction** that supports span extraction, counting, and addition/subtraction, they have not fully considered all potential types.
  - E.g., What percent are not non-families? => Negation operation is needed!
  - Previous methods are designed to **produce one single span** as the answer
  - E.g., Which ancestral groups are smaller than 11%? => Italian, English, Polish
  - Prior work learns to predict signed numbers for obtaining an arithmetic expression that can be executed by a symbolic system.
  - However, the **prediction of each signed number is isolated**, and the expression's context information has not been considered.

A Multi-Type Multi-Span Network for Reading Comprehension that Requires Discrete Reasoning [EMNLP-IJCNLP 2019]

- Multi-Type Multi-Span Network (MTMSN)
  - The model combines a **multi-type answer predictor** designed to support various answer types with a **multi-span extraction method** for dynamically producing one or multiple text spans.
  - Ex) Span from the text, Arithmetic expression, Count number, Negation on Numbers, Answer Type
  - **Arithmetic expression reranking mechanism** is proposed to rank expression candidates for further confirming the prediction.
  - Based on beam search decoding algorithm.



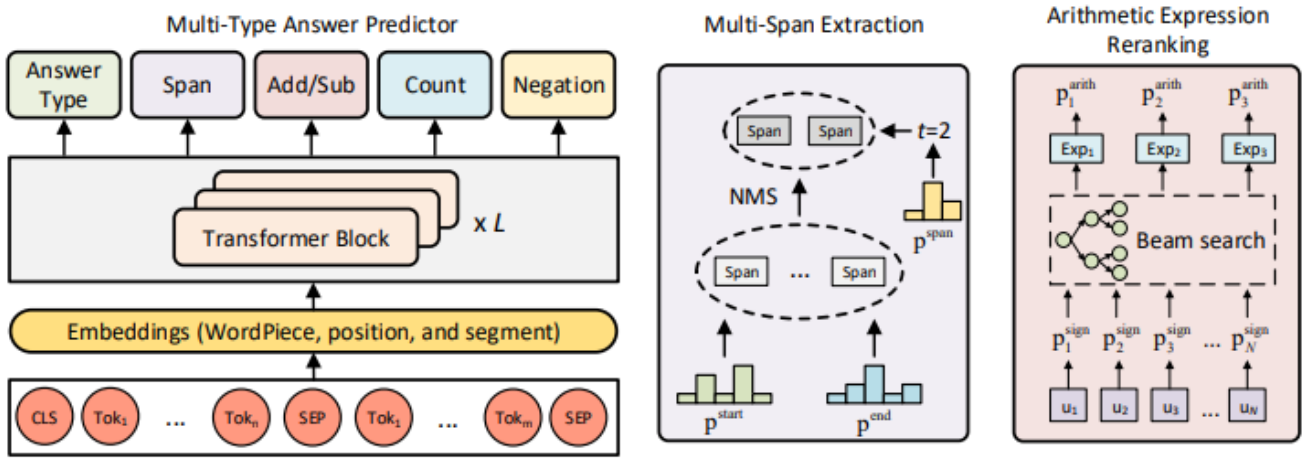
A Multi-Type Multi-Span Network for Reading Comprehension that Requires Discrete Reasoning [EMNLP-IJCNLP 2019]

- Multi-Type Answer Predictor
  - Like NAQANet, the authors design a **multi-type answer predictor** to selectively produce different kinds of answers.

$$\mathbf{p}^{\text{type}} = \text{softmax}(\text{FFN}([\mathbf{h}^{\mathbf{Q}_2}; \mathbf{h}^{\mathbf{P}_2}; \mathbf{h}^{\text{CLS}}]))$$

  - To **extract answer** either from the passage or from the question, the authors combine the gating mechanism of Wang et al., (2017) with standard decoding strategy of Seo et al., (2017) to predict the starting and ending positions across the entire sequence.

$$\begin{aligned}
\bar{\mathbf{M}}^{\text{start}} &= [\mathbf{M}_2; \mathbf{M}_0; \mathbf{g}^{\mathbf{Q}_2} \otimes \mathbf{M}_2; \mathbf{g}^{\mathbf{Q}_0} \otimes \mathbf{M}_0], \\
\bar{\mathbf{M}}^{\text{end}} &= [\mathbf{M}_2; \mathbf{M}_1; \mathbf{g}^{\mathbf{Q}_2} \otimes \mathbf{M}_2; \mathbf{g}^{\mathbf{Q}_1} \otimes \mathbf{M}_1], \\
\mathbf{p}^{\text{start}} &= \text{softmax}(\mathbf{W}^S \bar{\mathbf{M}}^{\text{start}}), \\
\mathbf{p}^{\text{end}} &= \text{softmax}(\mathbf{W}^E \bar{\mathbf{M}}^{\text{end}})
\end{aligned}$$



A Multi-Type Multi-Span Network for Reading Comprehension that Requires Discrete Reasoning [EMNLP-IJCNLP 2019]

- Multi-Type Answer Predictor
  - For each number mentioned in the passage, the authors gather its corresponding representation to **assign a plus, minus, or zero** for three-way classification.

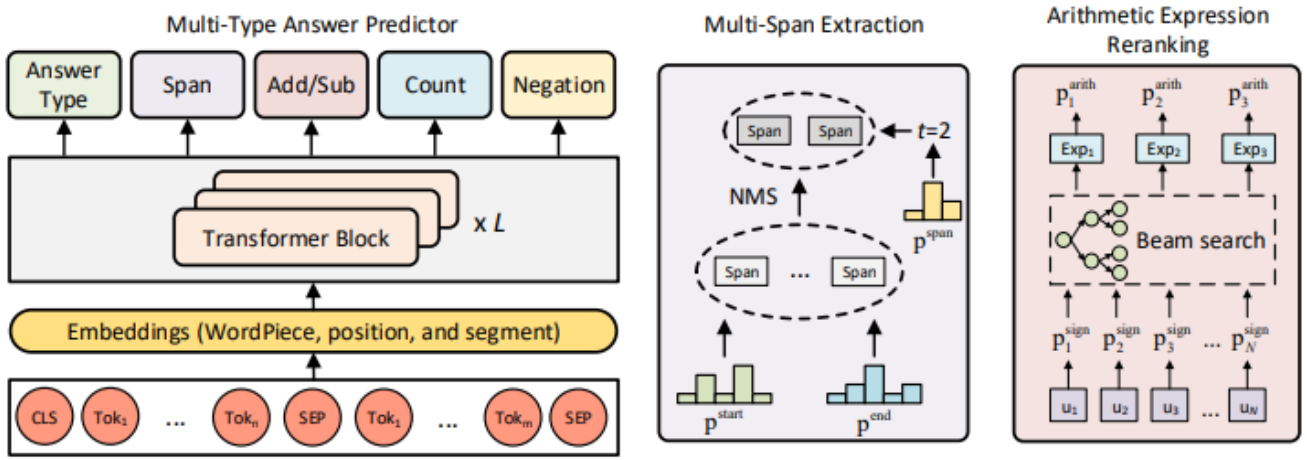
$$\mathbf{p}_i^{\text{sign}} = \text{softmax}(\text{FFN}([\mathbf{u}_i; \mathbf{h}^{\mathbf{Q}_2}; \mathbf{h}^{\mathbf{P}_2}; \mathbf{h}^{\text{CLS}}]))$$

  - Model produces a vector that summarizes the important information among all numbers and then perform multi-class classification to **count**.

$$\mathbf{p}^{\text{count}} = \text{softmax}(\text{FFN}([\mathbf{h}^{\mathbf{U}}; \mathbf{h}^{\mathbf{Q}_2}; \mathbf{h}^{\mathbf{P}_2}; \mathbf{h}^{\text{CLS}}]))$$

  - For each number, the model performs a two-way classification to indicate whether a **negation operation** should be performed.

$$\mathbf{p}^{\text{span}} = \text{softmax}(\text{FFN}([\mathbf{h}^{\mathbf{Q}_2}; \mathbf{h}^{\mathbf{P}_2}; \mathbf{h}^{\text{CLS}}]))$$



A Multi-Type Multi-Span Network for Reading Comprehension that Requires Discrete Reasoning [EMNLP-IJCNLP 2019]

- Multi-Span Extraction
  - By directly predicting the number of spans and model it as a classification problem, **multi-span problem** can be solved.

$$\mathbf{p}^{\text{span}} = \text{softmax}(\text{FFN}([\mathbf{h}^{\mathbf{Q}_2}; \mathbf{h}^{\mathbf{P}_2}; \mathbf{h}^{\text{CLS}}]))$$

  - To extract non-overlapped spans to the specific amount, the authors adopt **non-maximum suppression(NMS) algorithm**.
  - (1) The model first proposes a set of top-K spans S according to the descending order of the span score, and then predict the amount of extracted span t.
  - (2) Initialize a new set $\hat{S}$ of size t, and add elements from S one-by-one according to span score, if the degree of overlap is below a certain level.
  - (3) The degree of overlap is measured by text-level F1 function.

- Arithmetic Expression Reranking
  - The sign of each number is only determined by the number representation and global representations.
  - In this case, the **context information of expression itself has not been considered**, and the model might predict some obviously wrong expressions.
  - The authors construct an **expression representation** by taking both the numbers and signs into account.

$$\begin{aligned}
\boldsymbol{\alpha}_i^V &= \text{softmax}(\mathbf{W}^V(\mathbf{V}_i + \mathbf{C}_i)), \\
\mathbf{h}_i^{\mathbf{V}} &= \boldsymbol{\alpha}_i^V(\mathbf{V}_i + \mathbf{C}_i), \\
\mathbf{p}_i^{\text{arith}} &= \text{softmax}(\text{FFN}([\mathbf{h}_i^{\mathbf{V}}; \mathbf{h}^{\mathbf{Q}_2}; \mathbf{h}^{\mathbf{P}_2}; \mathbf{h}^{\text{CLS}}]))
\end{aligned}$$

  - By using **beam search** to produce top-ranked arithmetic expressions, the authors obtain several highly confident expression candidates.

A Multi-Type Multi-Span Network for Reading Comprehension that Requires Discrete Reasoning [EMNLP-IJCNLP 2019]

- Training / Inference / Results
  - The model is trained using **MML training objective**(finding all possible annotations).
  - Due to exponential search space, the authors only search addition/subtraction of three numbers at most.
  - Two objective functions for the multi-span component : **distantly-supervised loss** to maximize the probabilities of all matching spans, **classification loss** that maximizes the probability of span amount.
  - At test time, the model first chooses the answer type, and then performs specific prediction strategy.
  - Cf) (EM/F1) Andor et al., : 78.14/81.78 ; Numnet : 64.56/67.97 ; GenBERT : 68.6/72.4

Model	Dev		Test	
	EM	F1	EM	F1
Heuristic Baseline (Dua et al., 2019)	4.28	8.07	4.18	8.59
Semantic Role Labeling (Carreras and Màrquez, 2004)	11.03	13.67	10.87	13.35
BiDAF (Seo et al., 2017)	26.06	28.85	24.75	27.49
QANet+ELMo (Yu et al., 2018)	27.71	30.33	27.08	29.67
$\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019)	30.10	33.36	29.45	32.70
NAQANet (Dua et al., 2019)	46.20	49.24	44.07	47.01
$\text{NABERT}_{\text{BASE}}$	55.82	58.75	-	-
$\text{NABERT}_{\text{LARGE}}$	64.61	67.35	-	-
$\text{MTMSN}_{\text{BASE}}$	68.17	72.81	-	-
$\text{MTMSN}_{\text{LARGE}}$	**76.68**	**80.54**	**75.85**	**79.88**
Human Performance (Dua et al., 2019)	-	-	92.38	95.98

Table 1: The performance of MTMSN and other competing approaches on DROP dev and test set.

DROP Dataset Results (Current – 2022.02.06)

Rank	Model	F1 ↑	Paper	Code	Result	Year	Tags
1	**QDGAT** (ensemble)	88.38	Question Directed Graph Attention Network for Numerical Reasoning over Text			2020	
2	**BERT+Calculator** (ensemble)	81.78	Giving BERT a Calculator: Finding Operations and Arguments with Reading Comprehension			2019	
3	**NeRd**	81.71	Neural Symbolic Reader: Scalable Integration of Distributed and Symbolic Representations for Reading Comprehension			2020	
4	**TASE-BERT**	80.7	A Simple and Effective Model for Answering Multi-span Questions			2019	
5	**MTMSN Large**	79.88	A Multi-Type Multi-Span Network for Reading Comprehension that Requires Discrete Reasoning			2019	
6	**GenBERT** (+ND+TD)	72.4	Injecting Numerical Reasoning Skills into Language Models			2020	
7	**NumNet**	67.97	NumNet: Machine Reading Comprehension with Numerical Reasoning			2019	
8	**NAQA Net**	47.01	DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs			2019	
9	**GPT-3 175B** (Few-Shot)	36.5	Language Models are Few-Shot Learners			2020	
10	**BERT**	32.7	DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs			2019	

NEURAL SYMBOLIC READER: SCALABLE INTEGRATION OF DISTRIBUTED AND SYMBOLIC REPRESENTATIONS FOR READING COMPREHENSION

Xinyun Chen *
UC Berkeley
xinyun.chen@berkeley.edu

Chen Liang, Adams Wei Yu, Denny Zhou
Google Brain
{crazydonkey,adamsyuwei,dennyzhou}@google.com

Dawn Song
UC Berkeley
dawnsong@cs.berkeley.edu

Quoc V. Le
Google Brain
qvl@google.com

ICLR 2020

Neural Symbolic Reader : Scalable Integration of Distributed and Symbolic Representations for Reading Comprehension [ICLR 2020]

- Recent works on Numerical Reasoning cannot easily **scale to multiple domains** or **multi-step complex reasoning**
  - (1) They usually rely on handcrafted and specialized modules for each type of question.
  - (2) They don't support compositional applications of the operators, so it is hard to perform reasoning of more than one step.

- Integrating distributed representations with **symbolic operations** is essential for RC requiring complex reasoning
  - (1) By introducing a set of span selection operators, the **compositional programs**, usually executed against structured data such as DBs in semantic parsing, can now be executed over text
  - (2) The same architecture can be **applied to different domains** by simply extending the set of symbolic operators.

Neural Symbolic Reader : Scalable Integration of Distributed and Symbolic Representations for Reading Comprehension [ICLR 2020]

Passage	**Question & Answer**
Multiple spans	
...the population was spread out with **26.20% under the age of 18**, 9.30% from 18 to 24, **26.50% from 25 to 44**, **23.50% from 45 to 64**, and 14.60% who were 65 years of age or older...	**Question:** Which groups in percent are larger than 16%? **Program:** PASSAGE_SPAN(26,30), PASSAGE_SPAN(46,48), PASSAGE_SPAN(55,57) **Result:** 'under the age of 18', '25 to 44', '45 to 64'
Date	
When major general Nathanael Greene took command in the south, Marion and lieutenant colonel Henry Lee were ordered in January **1781**... On **August 31**, Marion rescued a small American force trapped by 500 British soldiers...	**Question:** When did Marion rescue the American force? **Program:** PASSAGE_SPAN(71,71), PASSAGE_SPAN(72,72), PASSAGE_SPAN(32,32) **Result:** 'August', '31', '1781'
Numerical operations	
...Lassen county had a population of **34,895**. The racial makeup of Lassen county was **25,532 (73.2%) white** (U.S. census), **2,834 (8.1%) African American** (U.S. census)...	**Question:** How many people were not either solely white or solely African American? **Program:** DIFF(9,SUM(10,12)) **Result:** 34895 - (25532 + 2834) = 6529

Neural Symbolic Reader : Scalable Integration of Distributed and Symbolic Representations for Reading Comprehension [ICLR 2020]

Counting	
...the Bolshevik party came to power in November 1917 through the **simultaneous election in the soviets** and an **organized uprising supported by military mutiny**...	**Question:** How many factors were involved in bringing the Bolsheviks to power? **Program:** COUNT(PASSAGE_SPAN(62, 66), PASSAGE_SPAN(69, 74)) **Result:** COUNT('simultaneous election in the soviets', 'organized uprising supported by military mutiny') = 2
Sorting	
...Jaguars kicker **Josh Scobee** managed to get a 48-yard field goal...with kicker Nate Kaeding getting a 23-yard field goal...	**Question:** Who kicked the longest field goal? **Program:** ARGMAX(KV(PASSAGE_SPAN(50,53),VALUE(9)), KV(PASSAGE_SPAN(92,94),VALUE(11))) **Result:** ARGMAX(KV('Josh Scobee', 48), KV('Nate Kaeding', 23)) = 'Josh Scobee'
...Leftwich flipped a **1-yard touchdown pass** to Wrighster...Leftwich threw a 16- yard touchdown pass to Williams for a 38-0 lead...	**Question:** How many yards was the shortest touchdown pass? **Program:** MIN(VALUE(17), VALUE(19)) **Result:** MIN(1, 16) = 1

Neural Symbolic Reader : Scalable Integration of Distributed and Symbolic Representations for Reading Comprehension [ICLR 2020]

- Neural Symbolic Reader (NeRd)
  - Reader : Encodes passages and questions into vector representations (BERT)
  - Programmer : Generates programs, which are executed to produce answers. (LSTM + Attention)

  - Compared to previous approaches that use a **unified programmer component** to generate programs for multi-step reasoning, NeRd can simply extend the operator set in the **domain specific language** to adapt to a different domain.

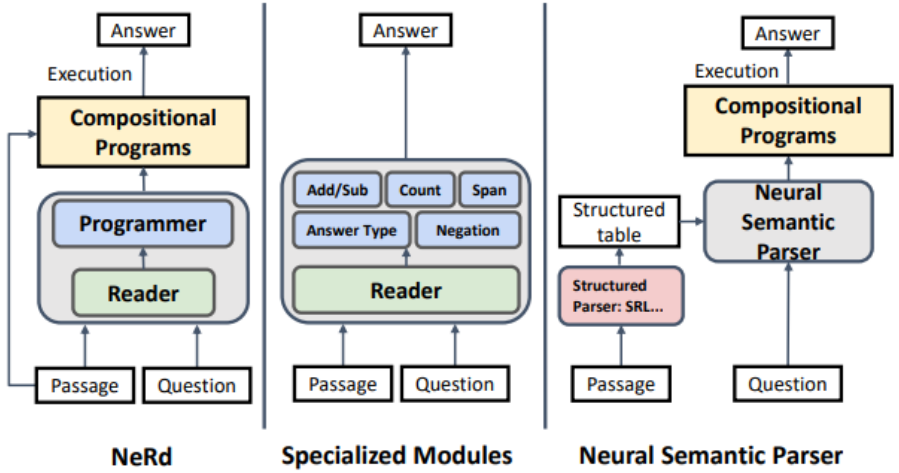


Figure 1: Comparison of NeRd with previous approaches for reading comprehension requiring complex reasoning. The components in grey boxes are the neural architectures. Previous works mainly take two approaches: (1) augmenting pre-trained language model such as BERT with specialized modules for each type of questions, which is hard to scale to multiple domains or multi-step complex reasoning; (2) applying neural semantic parser to the structured parses of the passage, which suffers severely from the cascade error. In contrast, the neural architecture of NeRd is domain-agnostic, which includes a *reader*, e.g., BERT, and a *programmer*, e.g., LSTM, to generate compositional programs that are directly executed over the passages.

Operator	Arguments	Outputs	Description
PASSAGE_SPAN QUESTION_SPAN	**v0**: the start index. **v1**: the end index.	a span.	Select a span from the passage or question.
VALUE	**v0**: an index.	a number.	Select a number from the passage.
KEY-VALUE (KV)	**v0**: a span. **v1**: a number.	a key-value pair.	Select a key (span) value (number) pair from the passage.
DIFF SUM	**v0**: a number or index. **v1**: a number or index.	a number.	Compute the difference or sum of two numbers.
COUNT	**v**: a set of spans.	a number.	Count the number of given spans.
MAX MIN	**v**: a set of numbers.	a number.	Select the maximum / minimum among the given numbers.
ARGMAX ARGMIN	**v**: a set of key-value pairs.	a span.	Select the key (span) with the highest / lowest value.

Table 1: Overview of our domain-specific language. See Table 2 for the sample usage.

Neural Symbolic Reader : Scalable Integration of Distributed and Symbolic Representations for Reading Comprehension [ICLR 2020]

- Domain Specific Language
  - The main challenge of applying such operations for reading comprehension is that the model needs to manipulate unstructured data, and parsing the text into structured representations may introduce a lot of **cascade errors**.
  - To address this issue the key insight is to introduce the **span selection operators**, so that all the arithmetics, counting and sorting operators can be applied to text.
  - E.g., PASSAGE_SPAN, QUESTION_SPAN, VALUE, KEY-VALUE + COUNT, ARGMAX
  - In summary, the introduction of span selection enables the application of discrete reasoning operators to text.

Operator	Arguments	Outputs	Description
PASSAGE_SPAN QUESTION_SPAN	**v0**: the start index. **v1**: the end index.	a span.	Select a span from the passage or question.
VALUE	**v0**: an index.	a number.	Select a number from the passage.
KEY-VALUE (KV)	**v0**: a span. **v1**: a number.	a key-value pair.	Select a key (span) value (number) pair from the passage.
DIFF SUM	**v0**: a number or index. **v1**: a number or index.	a number.	Compute the difference or sum of two numbers.
COUNT	**v**: a set of spans.	a number.	Count the number of given spans.
MAX MIN	**v**: a set of numbers.	a number.	Select the maximum / minimum among the given numbers.
ARGMAX ARGMIN	**v**: a set of key-value pairs.	a span.	Select the key (span) with the highest / lowest value.

Table 1: Overview of our domain-specific language. See Table 2 for the sample usage.

Neural Symbolic Reader : Scalable Integration of Distributed and Symbolic Representations for Reading Comprehension [ICLR 2020]

- Data Augmentation for Cold Start
  - It is often hard and expensive to obtain program annotations that represent the reasoning behind the answers.
  - **Cold Start Problem** means that the training cannot get started when there isn't any program available.

  - To obtain program annotations from question-answer pairs, we first follow previous work to find programs for questions answerable by span selection or arithmetic operations via an **exhaustive search**.
  - Because the space becomes too large for an exhaustive search, the authors apply data augmentation to address the **search space explosion problem**.

  - For **counting problems**, the authors augment the span selection questions by replacing the interrogatives
  - "What areas have a Muslim population of more than 50000 people?" => "How many areas …"

  - For **sorting problems**, the authors extract the key-value pairs by applying CoreNLP for entity recognition, and then heuristically find an associated number for each entity.
  - If including them as the argument of any sorting operator yields the correct answer, then such programs are added to the training set.

  - Although the programs found for counting and sorting through this data augmentation process is noisy, it helps **bootstrap** the training.
  - Throughout training, the authors also use the model to decode programs, and add those leading to correct answers into the training set.

Neural Symbolic Reader : Scalable Integration of Distributed and Symbolic Representations for Reading Comprehension [ICLR 2020]

- Hard EM with Thresholding against Spurious Programs
  - After collecting a set of programs for each question-answer pair, another obstacle is the **spurious program problem**, a phenomenon that a wrong program accidentally predicts a right answer.
  - E.g., DROP : average 9.8 programs that return correct answers
  - To filter out spurious programs, the authors adopt **hard EM** due to its simplicity and efficiency.
  - This approach uses the current model to select the program with the highest model probability among the ones that return the correct answer, and then maximizes the likelihood of the selected program.

Algorithm 1 Hard EM with Thresholding

Input: question-answer pairs $\{(x_i, y_i)\}_{i=1}^{N}$,
a model p_θ, initial threshold α_0, decay factor γ
for each (x_i, y_i) **do**
  $Z_i \leftarrow \text{DataAugmentation}(x_i, y_i)$
$T \leftarrow 0$
repeat
  $\alpha \leftarrow \alpha_0 * \gamma^T$
  $\mathcal{D} \leftarrow \emptyset$
  for each (x_i, y_i) **do**
    $z_i^* = \arg\max_k p_\theta(z_i^k|x_i), z_i^k \in Z_i$
    if $p_\theta(z_i^*) > \alpha$ or $T = 0$ and $|Z_i| = 1$ **then**
      $\mathcal{D} \leftarrow \mathcal{D} \cup (x_i, z_i^*)$
  Update θ by maximizing $\sum_{\mathcal{D}} \log p_\theta(z^*|x)$
  $T \leftarrow T + 1$
until converge or early stop

Neural Symbolic Reader : Scalable Integration of Distributed and Symbolic Representations for Reading Comprehension [ICLR 2020]

- Results

	Overall Dev		**Overall Test**		**Number** (62%)		**Span** (32%)		**Spans** (4.4%)		**Date** (1.6%)	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
NAQANet	46.75	50.39	44.24	47.77	44.9	45.0	58.2	64.8	0.0	27.3	32.0	39.6
$\text{NABERT}_{\text{LARGE}}$	64.61	67.35	–	–	63.8	64.0	75.9	80.6	0.0	22.7	55.7	60.8
$\text{MTMSN}_{\text{LARGE}}$	76.68	80.54	75.85	79.85	80.9	81.1	77.5	82.8	25.1	62.8	55.7	69.0
BERT-Calc	78.09	81.65	76.96	80.53	82.0	82.1	78.8	83.4	5.1	45.0	58.1	61.8
NeRd	**78.55** ± 0.27	**81.85** ± 0.20	**78.33**	**81.71**	**82.4** ± 0.3	**82.6** ± 0.2	76.2 ± 0.4	81.8 ± 0.2	**51.3** ± 0.8	**77.6** ± 1.2	**58.3** ± 1.8	67.2 ± 1.7

Table 4: Results on DROP dataset. On the development set, we present the mean and standard error of 10 NeRd models, and the test result of a single model. For all models, the performance breakdown of different question types is on the development set. Note that the training data of BERT-Calc model (Andor et al., 2019) for test set evaluation is augmented with CoQA (Reddy et al., 2019).

ReasonBERT: Pre-trained to Reason with Distant Supervision

Xiang Deng*[1], Yu Su[1], Alyssa Lees[2], You Wu[2], Cong Yu[2], and Huan Sun*[1]

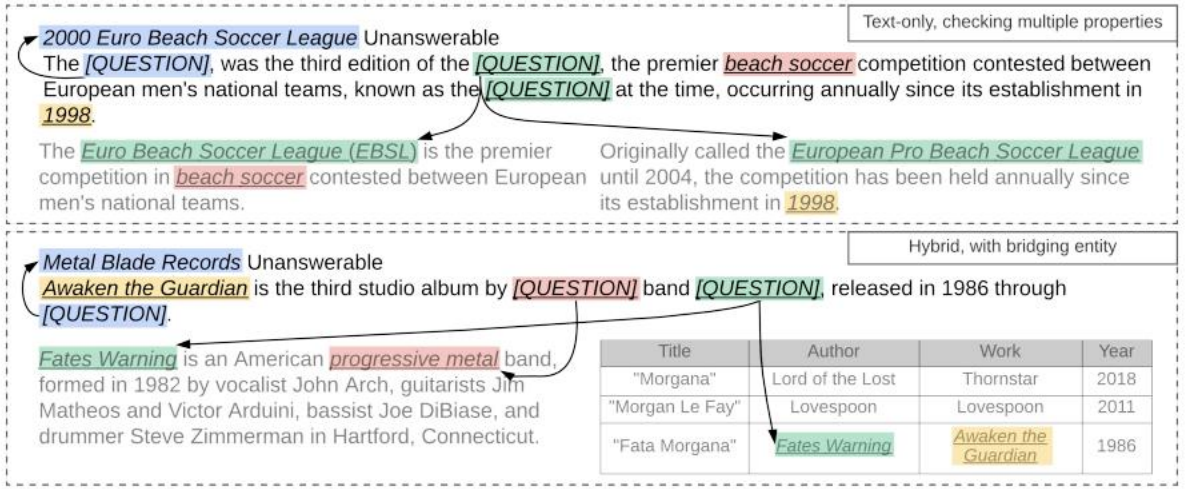
[1]The Ohio State University, Columbus, OH
{deng.595,su.809,sun.397}@osu.edu
[2]Google Research, New York, NY
{alyssalees,wuyou,congyu}@google.com

EMNLP 2021

ReasonBERT : Pretrained to Reason with Distant Supervision [EMNLP 2021]

- Existing pre-training methods only harvest **learning signals from local contexts** of naturally occurring texts
  - E.g., "Obama has two ___ , Malia and Sasha" => "daughters"
  - Many tasks require **reasoning beyond local texts** : Multi-hop QA, hybrid QA, Fact verification with multiple pieces of evidence, dialogue systems

- ReasonBERT : pre-training method to augment LMs for explicitly reasoning **long-range relations** and **multiple contexts**
  - ReasonBERT pairs a query sentence with **multiple relevant pieces of evidence** drawn from possibly different places.
  - Define a new LM pre-training objective, **span reasoning**, to recover entity spans that are masked out from the query sentence by jointly reasoning over the query sentence and the relevant evidence.
  - In addition to text, the authors also include **tables** as evidence to further empower LMs to reason over hybrid contexts.



ReasonBERT : Pretrained to Reason with Distant Supervision [EMNLP 2021]

- One major challenge lies in how to create a **large set of query-evidence pairs** for pretraining
  - Unlike existing self-supervised pre-training methods, examples with complex reasoning cannot be easily harvested from naturally occurring texts.
  - Instead, the authors draw inspiration from **distant supervision**, which assumes that "any sentence containing a pair of entities that are known to participate in a relation is likely to express that relation".
  - Specifically, given a query sentence containing an entity pair, if we mask one of the entities, another sentence or table that contains the same pair of entities can likely be **used as evidence to recover** the masked entity.
  - The authors collect multiple pieces of evidence that are jointly used to recover the masked entities in the query sentence, allowing to scatter the masked entities among different pieces of evidence to **mimic different types of reasoning**.

ReasonBERT : Pretrained to Reason with Distant Supervision [EMNLP 2021]

- Distant Supervision(DS) for Pre-training
  - The authors use **English Wikipedia** as the data source for pretraining.

  - **Text**
  - Extract sentences and tables from Wikipedia pages and then identify **salient spans**(e.g., named entities) from them using existing hyperlinks.
  - Since Wikipedia pages generally do not contain links to themselves,
    the authors additionally detect self-mentions by searching the names and aliases of the **topic entity** for each page.
  - **Temporal and numeric expressions** are identified using an existing NER tool(spark-nlp).

  - **Tables**
  - Extract tables that are labeled as <wikitable> from Wikipedia with no more than 500 cells.
  - **Real-world entities** are detected using existing **hyperlinks**.

  - Traditional NER tools are not tailored to work well on tables.
  - Instead, for a cell that does not contain hyperlinks, the authors match the complete cell value with sentences that are closely related to the table, sourced either from the **same page** or a page containing a **hyperlink** pointing to the current page.
  - If the matched span in the sentence contains a named entity, the authors consider the same entity as being linked to the cell as well.

ReasonBERT : Pretrained to Reason with Distant Supervision [EMNLP 2021]

- Distant Supervision(DS) for Pre-training
  - The key issue is **how to find evidence** that contains not only the answer entity, but also the relational information for inference.
  - Given a sentence as a query, the authors first **extract pairs of entities** in it.
  - For each entity pair, the authors then find **other sentences and tables that also contain the same pair as evidence**.

  - 1) First, the authors only consider **entity pairs** that contain at least one real-world entity.
  - 2) For textual evidence, the **entity pair** needs to contain the topic entity of the Wikipedia page, which is more likely to have relations to other entities.
  - 3) For tabular evidence, the authors consider only **entity pairs** that are in the same row of the table,
    but they do not need to contain the topic entity, as in many cases the topic entity is not present in tables.
  - 4) In both cases, the **query and evidence** should come from the same page, or the query contains a hyperlink pointing to the evidence page.
  - 5) For tabular evidence, the authors also allow for the case where the table contains a hyperlink pointing to the query page.

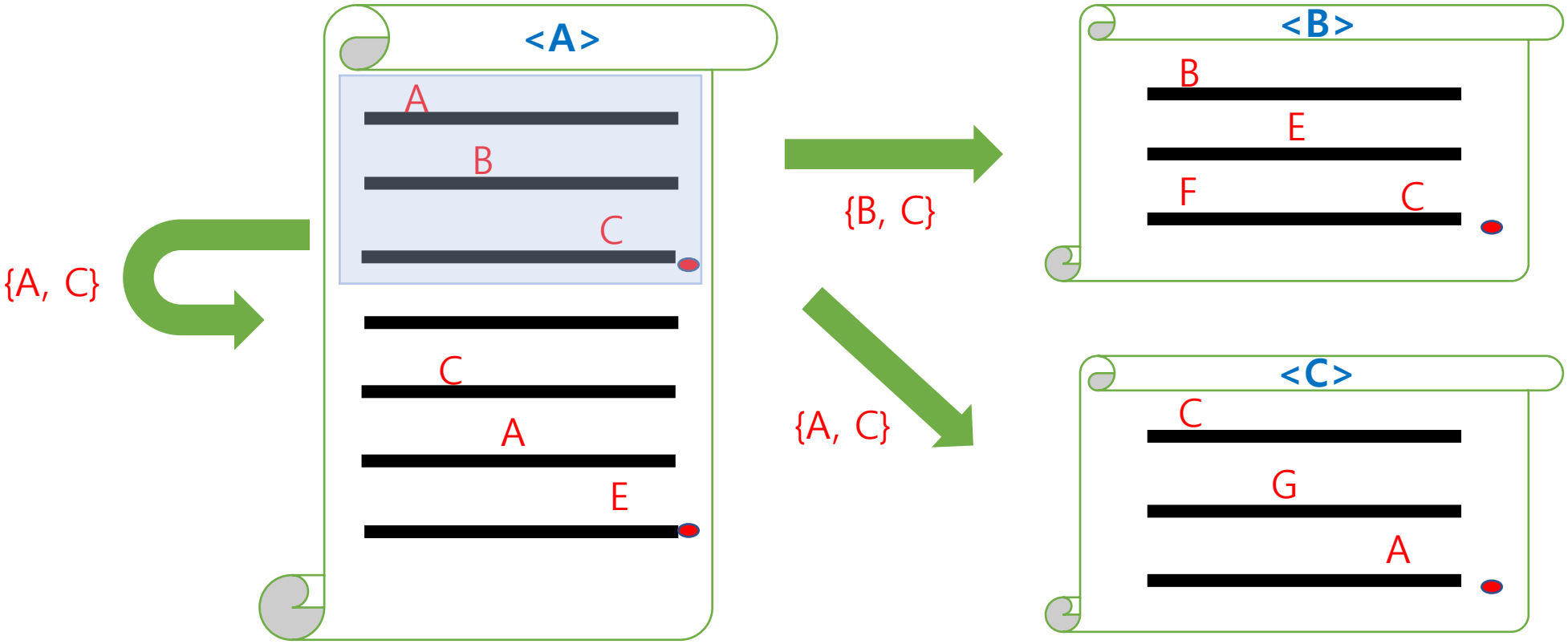
Query	Evidence
"*I Thought I Lost You*" was nominated for Broadcast Film Critics Association Award for Best Song and Golden Globe Award for Best Original Song, but lost both to *Bruce Springsteen*'s "*The Wrestler*" from *The Wrestler* (2008).	On January 11, 2009, *Springsteen* won the Golden Globe Award for Best Song for "*The Wrestler*", from the Darren Aronofsky *film by the same name*. "*I Thought I Lost You*" was nominated for the Broadcast Film Critics Association Award for Best Song at the 14th Broadcast Film Critics Association Award, but lost to *Bruce Springsteen*'s "*The Wrestler*" from *The Wrestler* (2008).

Query	Evidence
Rowland Barran (7 August 1858 – 6 August 1949) was an English *Liberal Party* politician and *Member of Parliament*.	*Rowland Barran* was the youngest son of Sir John Barran, a pioneer in clothing manufacture and *Member of Parliament* for Leeds and Otley.

Year	Member	Party
1885	William Jackson	Conservative
1902	*Rowland Barran*	*Liberal*
1918	Alexander Farquharson	Coalition Liberal

ReasonBERT : Pretrained to Reason with Distant Supervision [EMNLP 2021]

- Distant Supervision(DS) for Pre-training



ReasonBERT : Pretrained to Reason with Distant Supervision [EMNLP 2021]

- Distant Supervision(DS) for Pre-training

The *Citroen C6* is an *executive car* produced by the French car maker *Citroen* from 2005 to *2012*.	The *C6* was aimed as a stylish alternative to *executive cars*, like the BMW 5 Series and the Audi A6, and it has been described as "spaceship that rides on air", "charmingly idiosyncratic" and "refreshingly different". In *2012*, *Citroen* announced plans to enter the World Touring Car Championship. The team transformed a DS3 WRC into a laboratory vehicle to help with early development, while ...
Leaving the *Market Street subway* at Ferry Portal heading south, the T Third Street follows *The Embarcadero* south of Market Street, then veers onto King Street in front of *Oracle Park* until it reaches the *Caltrain* station terminal.	the 4th & King *Caltrain* station is 1.5 blocks from the stadium, and the *Oracle Park* Ferry Terminal is outside the east edge of the ballpark beyond the center field bleachers. the southwestern end of the *Market Street subway* connects to the much-older Twin Peaks Tunnel, and the northeastern end connects to surface tracks along the *The Embarcadero*.

Donovan also appeared in the *1980* film *Breaker Morant*, but in a subsidiary role, rather than as the title character.

Title	Year	Role
Cop Shop (TV series)	1978-1980	Detective Sgt. Vic
Breaker Morant	*1980*	Captain Simon Hunt
Smash Palace	1981	Traffic Officer

Try a Little Kindness is the sixteenth album by American singer/guitarist *Glen Campbell*, released in *1970*.

At the height of his popularity, a *1970* biography by Freda Kramer, The *Glen Campbell* Story, was published.

Day	Album	Artist	Notes
1	On the Boards	Taste	-
26	Chicago	Chicago	aka Chicago II
-	*Try a Little Kindness*	*Glen Campbell*	-

ReasonBERT : Pretrained to Reason with Distant Supervision [EMNLP 2021]

- Distant Supervision(DS) for Pre-training
  - A naïve way to construct pre-training examples is to sample a single piece of evidence for the query, and mask a shared entity as "answer".
  - However, this only simulates simple single-hop questions.

  - The authors construct complex pre-training examples that require the model to conduct multi-hop reasoning.
  - They combine **multiple pieces of evidence** in each pre-training example and **predict multiple masked entities simultaneously**.
  - They start by sampling up to two entity pairs from the query sentence and one piece of evidence(sentence or table) for each entity pair.
  - Then, they mask one entity in each pair as the "answer" to predict.

ReasonBERT : Pretrained to Reason with Distant Supervision [EMNLP 2021]

- Distant Supervision(DS) for Pre-training
  - The resulting pre-training examples fall into three categories
  - 1) Two disjoint entity pairs {(a,b), (c,d)} are sampled from the query, and one entity from each pair, and one entity from each pair, e.g., {a,c} is masked.

    => Combination of two single-hop questions
  - 2) Two sampled entity pairs {(a,b), (b,c)} share a common entity b, and b is masked.

    => Model needs to find two sets of entities that respectively satisfy the relationship with a and c, and take an intersection.
  - 3) Two sampled entity pairs {(a,b), (b,c)} share a common entity b, and {b,c} are masked.

    => Model needs to first identify b and then recover c based on its relationship with b.
  - The authors also mask an entity from the query that is not shown in the evidence to simulate unanswerable cases.

ReasonBERT : Pretrained to Reason with Distant Supervision [EMNLP 2021]

- Pretraining : Span Reasoning Objective
  - For text-only setting, BERT is used, whereas in the hybrid setting, TAPAS is used that uses extra row/column position token.
  - Given a sample for pretraining, the authors replace the masked entities in the query with special [QUESTION] token, and the task is to recover these masked entities from the given evidence.
  - $[\,[CLS], q, [SEP], E\,]$
  - With each [QUESTION] token for its entity, the authors predict the start and end position within the evidence.

$$P(s|q,E) = \frac{\exp\left(\mathbf{x}_s^\top \mathbf{S}\mathbf{x}_{a_i}\right)}{\sum_k \exp\left(\mathbf{x}_k^\top \mathbf{S}\mathbf{x}_{a_i}\right)}$$
$$P(e|q,E) = \frac{\exp\left(\mathbf{x}_e^\top \mathbf{E}\mathbf{x}_{a_i}\right)}{\sum_k \exp\left(\mathbf{x}_k^\top \mathbf{E}\mathbf{x}_{a_i}\right)} \quad (1)$$

$$L_{SR} = -\sum_{a_i \in \mathcal{A}} \left(\log P\left(s_{a_i}|q,E\right) + \log P\left(e_{a_i}|q,E\right)\right) \quad (2)$$

  - If no answer can be found in evidence, s, e is set to point to the [CLS] token.
  - The authors also include MLM objective in pre-training to leverage other tokens in the input that are not entities.
  - The final loss is the sum of span reasoning loss and masked language modeling loss.
  - ReasonBERT-B (based on BERT-Base), ReasonBERT-R (based on RoBERTa), ReasonBERT-T (based on TAPAS)

ReasonBERT : Pretrained to Reason with Distant Supervision [EMNLP 2021]

- Experiments
  - The authors conduct experiment with MRQA, HotpotQA, NQTables, HybridQA.
  - The authors compare with SpanBERT, SSPT, Splinter, TAPAS

Train. Size	Model	SQuAD	TriviaQA	NQ	NewsQA	SearchQA	HotpotQA	Average
16	BERT	9.9±0.6	15.4±1.3	20.5±1.5	6.5±1.2	16.8±1.2	9.6±1.6	13.1
	RoBERTa	10.3±1.1	21.0±3.1	22.5±2.1	6.7±2.0	23.4±3.5	11.2±1.0	15.9
	SpanBERT	15.7±3.6	27.4±4.1	24.3±2.1	8.1±1.4	24.1±3.2	16.3±2.0	19.3
	SSPT	10.8±1.2	21.2±3.8	23.7±4.1	6.5±1.9	25.8±2.6	9.1±1.5	16.2
	Splinter	16.7±5.9	23.9±3.8	25.1±2.8	11.6±1.0	23.6±4.5	15.1±3.5	19.3
	Splinter*	**54.6**	18.9	27.4	**20.8**	26.3	24.0	28.7
	ReasonBERT$_B$	33.2±4.0	37.2±2.6	33.1±2.7	11.8±2.3	**46.1**±5.2	22.4±2.8	30.6
	ReasonBERT$_R$	41.3±5.5	**45.5**±5.8	**33.6**±3.9	16.2±3.2	45.8±4.5	**34.1**±2.9	**36.1**
128	BERT	21.5±1.4	23.9±0.8	31.7±0.8	11.3±1.3	32.6±2.3	14.0±0.8	22.5
	RoBERTa	48.8±4.2	36.0±2.9	36.4±2.0	22.8±2.4	41.3±2.0	35.2±1.4	36.7
	SpanBERT	61.2±4.7	48.8±6.6	38.8±2.6	31.0±5.3	50.0±3.7	44.0±2.3	45.7
	SSPT	41.5±5.0	30.3±3.7	35.0±2.4	14.0±3.6	42.8±3.5	23.7±3.4	31.2
	Splinter	55.0±10.3	45.7±4.1	41.1±2.7	33.9±2.8	48.8±3.7	46.9±7.1	45.2
	Splinter*	**72.7**	44.7	46.3	**43.5**	47.2	54.7	51.5
	ReasonBERT$_B$	58.5±2.2	56.2±0.6	46.7±2.6	27.8±0.6	60.8±1.7	45.2±2.3	49.2
	ReasonBERT$_R$	66.7±2.9	**62.1**±0.9	**49.8**±1.6	35.7±1.5	**62.3**±1.7	**57.2**±0.6	**55.6**

All	BERT	88.8	73.6	78.7	67.5	82.0	76.2	77.8
	RoBERTa	92.0	78.1	80.6	**71.9**	85.2	79.1	81.2
	SpanBERT	**92.5**	**79.9**	80.7	71.1	84.8	**80.7**	**81.6**
	SSPT	91.1	77.0	80.0	69.7	83.3	79.7	80.1
	Splinter	92.4	79.7	80.3	70.8	84.0	80.6	81.3
	Splinter*	92.2	76.5	**81.0**	71.3	83.0	**80.7**	80.8
	ReasonBERT$_B$	90.3	77.5	79.9	68.7	83.7	80.5	80.1
	ReasonBERT$_R$	91.4	78.9	80.8	71.4	**85.3**	80.6	81.4

ReasonBERT : Pretrained to Reason with Distant Supervision [EMNLP 2021]

- Experiments

Model	Recall		1%		Full	
	Top 2	Top 3	F1	EM	F1	EM
$\text{HGN}_{\text{RoBERTa-Large}}$	-	-	-	-	82.2	-
HGN_{BERT}	-	-	-	-	74.8	-
BERT	92.4	96.9	39.8	28.6	71.9	57.9
RoBERTa	93.1	97.5	56.0	43.1	76.3	62.9
SpanBERT	93.6	97.7	56.5	44.1	76.3	62.9
SSPT	93.9	97.9	54.7	41.8	75.4	61.5
Splinter	94.1	97.9	57.0	44.2	76.5	62.5
$\text{ReasonBERT}_{\text{B}}$	93.8	97.8	57.6	45.3	77.2	63.4
$\text{ReasonBERT}_{\text{R}}$	**94.0**	**98.0**	**63.1**	**50.2**	**78.1**	**64.8**

Table 5: Results on HotpotQA.

Model	Dev		Test	
	F1	EM	F1	EM
RoBERTa	58.9	52.8	63.6	58.1
$\text{ReasonBERT}_{\text{R}}$	61.9	56.4	66.3	60.9
TAPAS	64.9	57.8	65.9	59.6
$\text{ReasonBERT}_{\text{T}}$	**69.2**	**63.5**	**72.5**	**67.3**

Table 6: Results on NQTables.

Model	Cell Selection		Dev		Test	
	Top 1	Top 2	F1	EM	F1	EM
$\text{HYBRIDER}_{\text{BERT-Base}}$	-	-	50.9	43.7	50.2	42.5
$\text{HYBRIDER}_{\text{BERT-Large}}$	68.5	-	50.7	44.0	50.6	43.8
TAPAS+RoBERTa	73.3	79.7	64.0	57.3	63.3	56.1
ReasonBERT	**76.1**	**81.3**	**67.2**	**60.3**	**65.3**	**58.0**

Table 7: Results on HybridQA.

ReasonBERT : Pretrained to Reason with Distant Supervision [EMNLP 2021]

- Ablation Studies
  - **Combining multiple pieces of evidence** and **predicting multiple masked spans** simultaneously brings the most gain, especially under the few-shot setting.
  - This is because this setting allows to simulate complex reasoning chains and encourage the model to do deep reasoning.

Model	1024		Full	
	F1	EM	F1	EM
ReasonBERT$_R$	65.2	52.8	79.2	65.8
– MLM	63.7	51.3	77.7	64.0
– Unanswerable Ent.	64.4	51.8	78.4	65.0
– Multiple Evidences	60.8	48.6	77.8	64.5

Table 8: Ablation study on HotpotQA.

ANY QUESTIONS?