How Well Do Multi-hop Reading Comprehension Models Understand Date Information?

Xanh Ho, Saku Sugawara, and Akiko Aizawa (contact: xanh@nii.ac.jp)

AACL-IJCNLP 2022, Taiwan (Online)

1. Background & Motivation

- Multi-hop *Machine Reading Comprehension (MRC)* requires a model to read multiple paragraphs to answer a given question
- Existing multi-hop MRC datasets: QAngaroo, HotpotQA, 2Wiki, MuSiQue, …
- Two types of questions:
  - *Bridge*: have a bridge entity that connects two paragraphs
  - *Comparison*: compare two entities on a specific aspect (e.g., compare two people about their date of birth)
- Tang et al. (2021) explored sub-questions of the bridge questions for model evaluation

Limitations:

- It is unclear about the ability of multi-hop models to perform step-by-step reasoning when finding an answer to a comparison question
- The current form of the reasoning process information of comparison questions does not describe the full path from question to answer

2. What We Did?

- Introduce a **HieraDate** dataset with three probing tasks:
  - Extraction task
  - Reasoning task
  - Robustness task
- Use our dataset to evaluate two leading models:
  - HGN (Fang et al., 2020)
  - NumNet+ (Ran et al., 2019)

  on two settings: with and without fine-tuning
- Other experiments:
  - Whether the number of required reasoning skills in each question type affects **question difficulty**?
  - Whether our probing questions are useful for improving ***QA performance***?
  - Whether our dataset can be used for ***data augmentation***?

Question: Who lived longer, Maceo Anderson or Jacek Karpiński?

Paragraph A: ***Maceo Anderson***

[1] Maceo Anderson (September 3, 1910 – July 4, 2001 in Los Angeles, California) expressed an interest in dancing at … . [2] ....

Paragraph B: ***Jacek Karpiński***

[3] Jacek Karpiński (9 April 1927 – 21 February 2010) was a Polish pioneer in computer engineering and … . [4] ....

Answer: Maceo Anderson

What is the date of birth of Maceo Anderson?
What is the date of death of Maceo Anderson?
What is the date of birth of Jacek Karpiński?
What is the date of death of Jacek Karpiński?
} Extraction Task

Reasoning Task:

How old was Maceo Anderson when they died?
How old was Jacek Karpiński when they died?
Full-date version: Is a 90-year-10-month-1-day-old person older than a 82-year-10-month-12-day-old person?
Year-only version: Is a 90-year-old person older than a 82-year-old person?

Robustness Task:

Who lived shorter, Maceo Anderson or Jacek Karpiński?

Example of a question in our dataset.

3. Dataset Construction & Information

Dataset Construction

- Use two datasets: HotpotQA & 2Wiki
- Include three main steps:

1. Obtain Date Questions
   Use predefined keywords: born first, lived longer, …
2. Generate Sub-questions & Sub-answers
   - 2Wiki: templates
   - HotpotQA: templates & munual annotation
3. Construct HieraDate
   - Dev. & test: dev. sets of HotpotQA and 2Wiki
   - Train: training set of 2Wiki

HieraDate Information

- Each main question has the extraction, reasoning, and robustness tasks

Split	Main	Extraction	Reasoning	Robustness
Train	8745	21340	19415	8745
Dev.	549	1346	1222	549
Test	549	1346	1222	549

- Two main types of questions:
  - *Combined reasoning*: requires both comparison and numerical reasoning
  - *Comparison reasoning*: requires only comparison reasoning

Dataset Quality Check

- Randomly select 100 samples from the test set
- Instruct graduate students to conduct the annotation
  (one sample is annotated by two annotators)
- Input & Output:
  - Input: the context and a list of questions
  - Output: all answers
- Result:
  - Human upper bound is 100% for all tasks
  - Human average is slightly low
  - Manually investigation: annotators made careless mistakes in several examples
  - Confirm that these examples are answerable and reasonable

	Main		Extraction		Reasoning		Robustness	
	EM	F1	EM	F1	EM (num)	F1 (comp)	EM	F1
Human (avg.)	94.00	94.90	99.16	99.53	100	98.06	95.5	95.9
Human UB	100	100	100	100	100	100	100	100

4. Models

- Existing models cannot perform all three tasks
  → Evaluate them under two groups:
  - Focus on comparison reasoning: HGN (Fang et al., 2020)
  - Focus on numerical reasoning: NumNet+ (Ran et al., 2019
- HGN:
  - Is designed for HotpotQA
  - Can answer yes/no questions
  - Cannot answer numerical questions
- NumNet+:
  - Is designed for DROP
  - Cannot answer yes/no questions
  - Can answer numerical questions
- Both models can perform on the extraction and robustness tasks

5. Results

Date Understanding Evaluation

- The models are not fine-tuned on our dataset:
  - Extraction: HGN performs quite well; NumNet+ performs worse
  - Reasoning: both HGN and NumNet+ fail; reasons:
    - The forms of reasoning questions are new to these models
    - These models do not possess the reasoning abilities as humans do
  - Robustness: comparable with the results of the main multi-hop questions
- The models are fine-tuned on our dataset:
  - All scores improve
  - Reasoning:
    - HGN reaches the highest score in the comparison reasoning task
    - **Full-date version**: NumNet+ does not have abilities to subtract two dates
    - **Year-only version**: can perform subtraction in the form of numbers

Fine-tuning	Model	Main		Extraction		Reasoning		Robustness	
		EM	F1	EM	F1	EM (num)	F1 (comp)	EM	F1
X	HGN	66.85	76.15	94.58	96.14	*N/A*	53.08	71.95	81.64
	NumNet+	67.94	71.57	1.26	47.93	22.79 (F1)	*N/A*	69.58	71.91
✓	HGN	78.87	82.69	96.06	97.14	*N/A*	100	76.68	78.58
	NumNet+	95.08	95.20	96.36	97.73	35.96 (F1)	*N/A*	94.90	94.93

Whether the number of required reasoning skills in each question type affects question difficulty?

Scores of comparison reasoning questions > those of combined reasoning questions

- HGN: 85.7 vs. 72.3 F1
- NumNet+: 98.8 vs. 81.6 F1

→ Combined reasoning questions are more difficult than comparison reasoning questions

Whether our probing questions are useful for improving QA performance?

Train HGN and NumNet+ on six combinations of the main and probing tasks

- Each task in our dataset helps to improve QA performance
- When training the models on all tasks: improve significantly in both HGN and NumNet+ compared with the models trained on the main questions only
  - HGN: 82.7 vs. 77.1 F1
  - NumNet+: 94.9 vs. 84.6 F1

→ our probing questions help to improve the QA performance

Whether our dataset can be used for data augmentation?

- Train HGN and NumNet+ on two settings:
  1. The original dataset
  2. The original dataset and our dataset
- **Results**:
  - No significant change on the original datasets (e.g., 81.1 → 81.4 F1 for HotpotQA)
  - The improvement in our dataset is significant (e.g., 76.3 → 84.9 F1)
  - All models that are trained on setting #2 are better in our robustness task

→ our dataset can be used as augmentation data for improving the robustness of the models trained on HotpotQA, 2Wiki, and DROP

6. Conclusion

- Propose a new multi-hop RC dataset for comprehensively evaluating the ability of existing models
- Evaluate top-performing models
- The models may not possess the ability to subtract two dates even when fine-tuned on our dataset
- Our probing questions could help to improve QA performance and can be used for data augmentation

Check out more of our work at:
https://arxiv.org/abs/2210.05208
https://github.com/Alab-NII/HieraDate

[1] Zhilin Yang, Peng Qi, Saizheng Zhang, et al., HotpotQA: A dataset for diverse, explainable multi-hop question answering, EMNLP 2018.
[2] Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, and Zhiyuan Liu, NumNet: Machine reading comprehension with numerical reasoning, EMNLP-IJCNLP 2019.
[3] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa, Constructing a multihop QA dataset for comprehensive evaluation of reasoning steps, COLING 2020.
[4] Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu, Hierarchical graph network for multi-hop question answering, EMNLP 2020.
[5] Yixuan Tang, Hwee Tou Ng, and Anthony Tung, Do multi-hop question answering systems know how to answer the single-hop sub-questions? EACL 2021.