Multi-hop Reading Comprehension through Question Decomposition and Rescoring

Sewon Min[1], Victor Zhong[1], Luke Zettlemoyer[1], Hannaneh Hajishirzi[1,2]

University of Washington[1], Allen Institute of Artificial Intelligence[2]

Code & demo available at https://github.com/shmsw25/DecompRC

Abstract

- Multi-hop Reading Comprehension (RC) requires answering questions by reasoning over multiple pieces of evidence.
- Our Idea: Decompose a multi-hop question into a series of single-hop sub-questions.

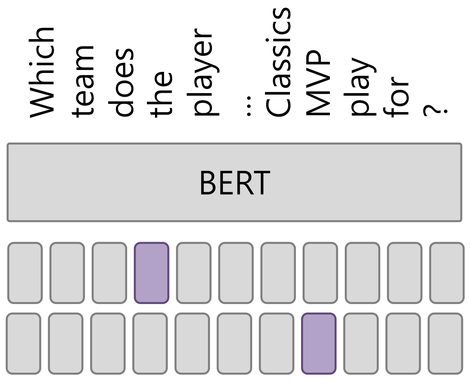
1) Intuitive Solution | 2) Leverage Single-hop RC | 3) Explainability of model decisions

Challenges & Our solutions

1) No annotation for single-hop sub-questions
   - ▻ **Recast the generation into span prediction**: 400 examples are enough to generate high quality sub-questions.
2) How to avoid cascading errors?
   - ▻ **Rescorer with global reasoning**: all decompositions & potential answers are helpful in deciding the most suitable option.

Q Which team does the player named 2015 Diamond Head Classics MVP play for?

P1 The 2015 Diamond Head Classic … Buddy Hield was named the tournament's MVP.
P2 Chavano Rainier Buddy Hield is a Bahamian professional basketball player for the Sacramento Kings …



1) Decompose question using spans from Pointer (trained on 400 examples)

Q1 Which player named 2015 Diamond Head Classics MVP?
Q2 Which team does [ANSWER] play for?

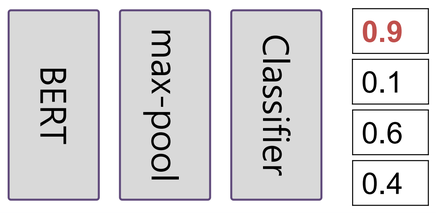
Single-hop Reading Comprehension Model

2) Answer each sub-question using well-studied single-hop RC Model (trained on SQuAD)

Answer to Q1 = Buddy Hield
Answer to Q2.replace([ANSWER], Buddy Hield) = **Sacramento Kings**

3) Obtain answers for each decomposition type and feed them into Rescorer

Bridging -> Sacramento Kings
Intersection -> Buddy Hield
Comparison -> Sacramento Kings
Original -> Atlanta Hawks



Type: Bridging (47%)
Q Which team does the player named 2015 Diamond Head Classics MVP play for?
Q1 Which player named 2015 Diamond Head Classics MVP?
Q2 Which team does [ANSWER] play for?

Type Intersection (23%)
Q Stories USA starred / which actor and comedian / from 'The Office'?
Q1 Stories USA starred which actor and comedian?
Q2 Which actor and comedian from 'The Office'?

Type Comparison (22%)
Q Who was born earlier, Emma Bull or Virginia Woolf ?
Q1 Emma Bull was born when? / Q2 Virginia Woolf was born when?
Q3 Which is smaller (Emma Bull, [ANSWER]) (Virginia Woolf, [ANSWER])

Related Work

- End-to-end models for multi-hop QA (Dhingra et al 18, Zhong et al 19).
- Talmor and Berant (18) proposed a model which uses semantic parsing annotations and decide on decomposition in pipeline approach.

Experiments

Dataset: HotpotQA (Yang et al 18) / Metric: F1 score

Model	Distractor setting	Fullwiki setting
DecompRC	70.57	43.26
DecompRC-1hop train	61.73	39.17
BERT	67.08	38.40
BERT-1hop train	56.27	29.97
BiDAF	58.28	34.36

'-1hop train': only trained on 1hop questions (semi-supervised setting)

Span-based vs. Free-form subQs / Rescorer vs. Pipeline

Robert Smith founded the multinational company headquartered in what city?

Span-based
Q1 Robert Smith founded which multinational company?
Q2 [ANSWER] headquartered in what city?

Free-form
Q1 Which multinational company was founded by Robert Smith?
Q2 Which city contains a headquarter of [ANSWER]?

Model	F1
Span-based via Pointer 200	65.44
Span-based via Ponter 400	69.44
Span-based via Human	70.41
Free-form via Human	70.76

Model	F1
No Scorer	61.73
Pipeline	63.59
DecompRC	70.57
Oracle	76.75

Modifying input paragraphs

Hotpot contains 2 gold paragraphs + 8 distractors (top 8 via TF-IDF).
We alternate distractors by choosing the next top 8 via TF-IDF.

Model	Original	Modified
DecompRC	70.57	59.07
DecompRC-1hop train	61.73	58.30
BERT (end-to-end)	67.08	44.68

Inverted comparison questions

For binary comparison questions, we automatically invert 667 question.
e.g. "Who was born earlier, A or B?" -> "Who was born later, A or B?"

Model	Original	Inverted	Joint
DecompRC	67.80	65.78	55.80
BERT (end-to-end)	54.65	32.49	19.27

Limitations

Q What country is the Selun located in?
P1 Selun lies between … in the canton of St. Gallen.
P2 The canton of St. Gallen is a canton of **Switzerland**.

Q Which pizza chain has locations in more cities, Round Table Pizza or Marion's Piazza?
P1 **Round Table Pizza** is a large chain of pizza parlors in the western US.
P2 Marion's Piazza … operates 9 restaurants throughout the greater Dayton area.
Q1 Round Table Pizza has locations in how many cities?
Q2 Marion 's Piazza has locations in how many cities?

Conclusion

- We propose DecompRC to decompose a multi-hop question into single-hop sub-questions.
- Using span prediction, our model produces human-level sub-questions with only 400 annotations.
- In addition, decomposition scorer outperforms previous pipeline approach.
- We show DecompRC achieves SOTA on HotpotQA and is more robust than end-to-end model.

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". NAACL 2019.
Bhuwan Dhingra, Qiao Jin, Zhilin Yang, William W Cohen and Ruslan Salakhutdinov. "Neural models for reasoning over multiple mentions using coreference". NAACL 2018.
Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev and Percy Liang, " SQuAD: 100,000+ questions for machine comprehension of text", EMNLP 2016
Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi and Hannaneh Hajishirzi, "Bidirectional Attention Flow for Machine Comprehension", ICLR 2017
Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. NAACL 2018.
Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov and Manning, "HotpotQA: a dataset for diverse, explainable multi-hop question answering", EMNLP 2018
Victor Zhong, Caiming Xiong, Nitish Keskar and Richard Socher. "Coarse-grain fine-grain coattention network for multi-evidence question answering". ICLR 2019.

Contact

@sewon__min