Multi-hop Reading Comprehension through Question Decomposition and Rescoring

Sewon Min, Victor Zhong, Luke Zettlemoyer, Hannaneh Hajishirzi
University of Washington/Allen Institute for AI, 2019

George Tolkachev (georgeto@seas.upenn.edu)
April 15, 2020

Penn Engineering

Contents

- Background
- Problem setup
- Previous work
- Overview of new model
- Experiments and results
- Shortcomings/limitations
- Potential for future work

Reading Comprehension

- Given a document, the model must answer questions regarding contents
- New datasets allow for evaluation of RC models
  - CNN/Daily Mail (Hermann et al., 2015)
  - TriviaQA (Joshi et al., 2017)
  - SQuAD (Rajpurkar et al., 2016)
- A single sentence suffices to answer most questions

Multi-hop Reading Comprehension

- Requires evidence from multiple paragraphs to answer questions
- Previous datasets contained relational queries as questions
- HotpotQA, meanwhile, contains diverse natural language questions

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

Q1 Which player named 2015 Diamond Head Classic's MVP?
Q2 Which team does ANS play for?

Previous Work: Talmor and Berant (2018)

- Proposed decomposing questions into smaller parts, and computed final answer from sequence of answers
- Key differences from present approach

Talmor and Berant (2018)	Min et al. (2019)
Decomposed questions that corresponded to relational queries	Answer natural language questions
Built distant supervision data for training their model	Training requires only 400 decomposition examples
Select decomposition method based solely on question	Perform several decompositions and select answer by rescoring

Examples of Previous Work

Question	**Split-1**	**Split-2**
"Find the actress who played Hailey Rogers, what label is she signed to"	*"the actress who played Hailey Rogers"*	*"Find* `VAR` *, what label is she signed to"*
"What are the colors of the sports team whose arena stadium is the AT&T Stadium"	*"the sports team whose arena stadium is the AT&T Stadium"*	*"What are the colors of* `VAR`*"*
"What amusement park is located in Madrid Spain and includes the stunt fall ride"	*"What amusement park is located in Madrid Spain and"*	*"park includes the stunt fall ride"*
"Which university whose mascot is The Trojan did Derek Fisher attend"	*"Which university whose mascot is The Trojan did"*	*"university Derek Fisher attend"*

Composit.	**Complex SPARQL query r'**	**Example (natural language)**
CONJ.	r. ?x pred$_1$ obj. **or** r. ?x pred$_1$?c. ?c pred$_2$ obj.	*"What films star Taylor Lautner and have costume designs by Nina Proctor?"*
SUPER.	r. ?x pred$_1$?n.ORDER BY DESC(?n) LIMIT 1	*"Which school that Sir Ernest Rutherford attended has the latest founding date?"*
COMPAR.	r. ?x pred$_1$?n. FILTER ?$n < V$	*"Which of the countries bordering Mexico have an army size of less than 1050?"*
COMP.	$r[e/y]$. ?y pred$_1$ obj.	*"Where is the end of the river that originates in Shannon Pot?"*

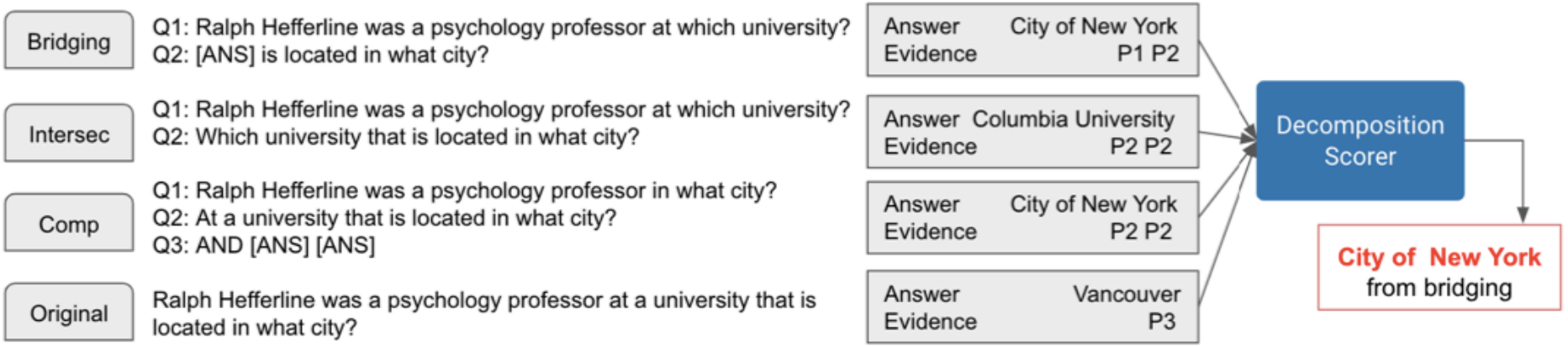
Proposed Model: DecompRC

- Decomposes question into smaller sub-questions based on span prediction and several reasoning types
- Employs a single-hop RC model to provide answer for each sub-question, and combines answers according to specific reasoning type
- Decomposition scorer reranks all decompositions and returns answer from top decomposition

Example

Q: Ralph Hefferline was a psychology professor at a university that is located in what city?

P1: Ralph Franklin Hefferline was a psychology professor at Columbia University.
P2: Columbia University (Columbia; officially Columbia University in the City of New York), ...
P3: Stanley Coren is a psychology professor … at the University of British Columbia in Vancouver ...



Question Decomposition

- 3 reasoning types: bridging, intersection, and comparison
- Decompose question according to reasoning types

Type	**Bridging (47%)** requires finding the first-hop evidence in order to find another, second-hop evidence.
Q	Which team does the **player** named 2015 Diamond Head Classic's MVP play for?
Q1	Which player named 2015 Diamond Head Classic's MVP?
Q2	Which team does ANS play for?
Type	**Intersection (23%)** requires finding an entity that satisfies two independent conditions.
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%)** requires comparing the property of two different entities.
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, ANS) (Virgina Woolf, ANS)

Decomposition Algorithm

- Train model Pointer_c to map question to *c* points using only 400 annotations
- Let $S = [s_1, \ldots, s_n]$ be an *n*-word sequence in input
- Encode S using BERT
- Select *c* indices that yield highest joint probability

$$\text{ind}_1, \ldots, \text{ind}_c = \underset{i_1 \leq \ldots \leq i_c}{\text{argmax}} \prod_{j=1}^{c} \mathbb{P}(i_j = \text{ind}_j)$$

procedure GENERATESUBQ(Q : question, Pointer_c)

/ Find q_1^b and q_2^b for Bridging */*

$\text{ind}_1, \text{ind}_2, \text{ind}_3 \leftarrow \text{Pointer}_3(Q)$

$q_1^b \leftarrow Q_{\text{ind}_1:\text{ind}_3}$

$q_2^b \leftarrow Q_{:\text{ind}_1}$: ANS : $Q_{\text{ind}_3:}$

article in $Q_{\text{ind}_2-5:\text{ind}_2} \leftarrow$ 'which'

/ Find q_1^i and q_2^i for Intersecion */*

$\text{ind}_1, \text{ind}_2 \leftarrow \text{Pointer}_2(Q)$

$s_1, s_2, s_3 \leftarrow Q_{:\text{ind}_1}, Q_{\text{ind}_1:\text{ind}_2}, Q_{\text{ind}_2:}$

if s_2 starts with wh-word **then**

$q_1^i \leftarrow s_1 : s_2,\ q_2^i \leftarrow s_2 : s_3$

else

$q_1^i \leftarrow s_1 : s_2,\ q_2^i \leftarrow s_1 : s_3$

/ Find q_1^c, q_2^c and q_3^c for Comparison */*

$\text{ind}_1, \text{ind}_2, \text{ind}_3, \text{ind}_4 \leftarrow \text{Pointer}_4(Q)$

$\text{ent}_1, \text{ent}_2 \leftarrow Q_{\text{ind}_1:\text{ind}_2}, Q_{\text{ind3:ind}_4}$

$op \leftarrow \text{find\_op}(Q, \text{ent}_1, \text{ent}_2)$

$q_1^c, q_2^c \leftarrow \text{form\_subq}(Q, \text{ent}_1, \text{ent}_2, op)$

$q_3^c \leftarrow op\ (\text{ent}_1, \text{ANS})\ (\text{ent}_2, \text{ANS})$

Single-hop Reading Comprehension

- BERT reading comprehension model (trained using SQuAD)
- Input: sub-question and N paragraphs $S_1, \ldots, S_N$
- Output: answer and evidence in the form of paragraph
- Possible values of answer are span, yes, or no
- For each paragraph S_i, compute four scores for each answer $[y_i^{\text{span}}; y_i^{\text{yes}}; y_i^{\text{no}}; y_i^{\text{none}}] = \max(U_i)W_1$, where

  U_i is the BERT encoding of the sub-question appended with S_i,

  W_1 is a parameter matrix, and $\max$ is max-pooling across input

Decomposition Scorer

- Scores all decompositions and outputs answer and evidence according to highest ranking decomposition
- Let x denote a concatenation of the original question, the reasoning type t, and answer_t and evidence_t
- Encode x using BERT to obtain matrix U_t
- Calculate score as $p_t = \text{sigmoid}(W_2^T \max(U_t))$, where W_2 is a trainable parameter matrix

Experimental Setup

- HotpotQA dataset, comprised of Wikipedia articles
- Evaluate DecompRC using two different settings: distractor (contains question and 10 paragraphs) and full-wiki (contains only question)
- Training set consists of easy (single-hop) and medium and hard (multi-hop) questions
- Dev and test set consist of only hard questions

Results

Dev set

Model	Distractor setting					Full wiki setting				
	All	Bridge	Comp	Single	Multi	All	Bridge	Comp	Single	Multi
DecompRC	**70.57**	**72.53**	**62.78**	**84.31**	**58.74**	**43.26**	**40.30**	**55.04**	**52.11**	**35.64**
BERT	67.08	69.41	57.81	82.98	53.38	38.40	34.77	52.85	46.14	31.74
BiDAF	58.28	59.09	55.05	-	-	34.36	30.42	50.70	-	-

Test set

Model	Dist F1	Open F1
DecompRC	69.63	40.65
Cognitive Graph	-	48.87
BERT Plus	69.76	-
MultiQA	-	40.23
DFGN+BERT	68.49	-
QFE	68.06	38.06
GRN	66.71	36.48
BiDAF	59.02	32.89

Evaluating Robustness

- Modified distractor paragraphs to observe how much performance would worsen
- Rewrote original questions so that correct answer was inverted
  - E.g. if original question is "Who was born earlier, Emma Bull or Virginia Woolf?", new question is "Who was born later?"

Model	F1
DecompRC	70.57 —> 59.07
BERT	67.08 —> 44.68

Model	Orig F1	Inv F1	Joint F1
DecompRC	67.80	65.78	55.80
BERT	54.65	32.49	19.27

Ablation Studies

- Compared span-based and human-written sub-questions

Question	Robert Smith founded the multinational company headquartered in what city?
Span-based	Q1: Robert Smith founded which multinational company? Q2: `ANS` headquartered in what city?
Free-form	Q1: Which multinational company was founded by Robert Smith? Q2: Which city contains a headquarter of `ANS`?

Sub-questions	F1
Span (Pointer$_c$ trained on 200)	65.44
Span (Pointer$_c$ trained on 400)	69.44
Span (human)	70.41
Free-form (human)	70.76

- Compared decomposition decision method based on scorer with oracle that provides an upper bound

Decomposition decision method	F1
Decomposition scorer (DecompRC)	70.57
Oracle	76.75

Shortcomings/Limitations

- Cannot decompose questions that require implicit multi-hop reasoning
- Answer may not be explicitly found within the text
- Incapable of other reasoning types such as counting

Q What country is the Selun located in? **P1** Selun lies between the valley of Toggenburg and Lake Walenstadt 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 United States. **P2** Marion's Piazza ... the company currently 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?
Q Which magazine had more previous names, Watercolor Artist or The General? **P1** Watercolor Artist, formerly Watercolor Magic, is an American bi-monthly magazine that focuses on ... **P2** The General (magazine): Over the years the magazine was variously called 'The Avalon Hill General', 'Avalon Hill's General', 'The General Magazine', or simply 'General'. **Q1** Watercolor Artist had how many previous names? **Q2** The General had how many previous names?

Conclusions and Further Work

- Approaching sub-question generation as a span prediction problem reduces the number of annotations necessary for training
- Decomposition scorer allows for comparison of the effectiveness of different decompositions
- Provides "reasons" in the form of sub-questions
- Future work can address other reasoning types, as well as experiment with unanswerable questions