SOTA NLP Yomikai: Multi-hop Reading Comprehension through Question Decomposition and Rescoring

Min et al. (ACL 2019, long)

Reader: Saku Sugawara

2019-09-28

Abstract

- “Decomposition” system for multi-hop reading comprehension 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?

- Original question
- Span decomposed
- P1,P2: premises
- Answer
- Q1,Q2: sub questions

Why this paper?

Personal interest: evaluation of language understanding (and the quality of NLU datasets)

- 2016 Yomikai
  - Chen et al. (2016) A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task [paper]
  - Their system achieved the ceiling for performance on this task
- 2017 Yomikai
  - Jia & Liang (2017) Adversarial Examples for Evaluating Reading Comprehension Systems [paper]
  - SOTA systems are easily fooled by distracting sentences
- 2018 Yomikai (if I could attend)
  - Min et al. (2018) Efficient and Robust Question Answering from Minimal Context over Documents [paper]
  - Most questions in existing datasets can be answered with a small set of sentences (SQuAD v1: 92% of questions are answerable with a single sentence)

Background: Multi-hop RC

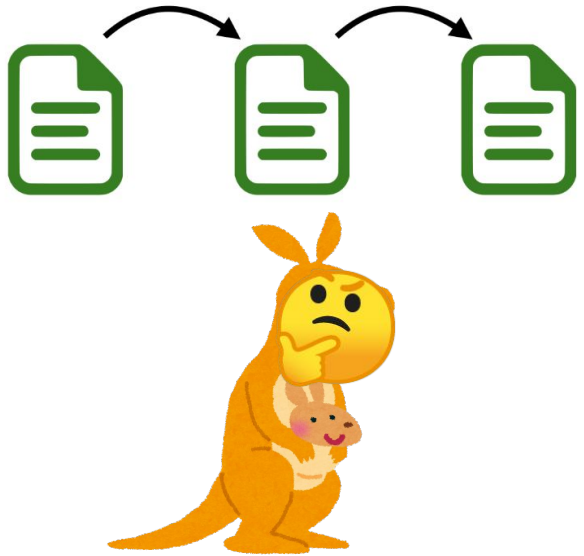
- Multi-hop Reading Comprehension
  - First(?): WikiHop dataset by Welbl et al. (2018) [web] [paper]
  - Requiring reasoning over multiple documents/evidence

The Hanging Gardens, in **[Mumbai]**, also known as Pherozeshah Mehta Gardens, are terraced gardens … They provide sunset views over the **[Arabian Sea]** …

Mumbai (also known as Bombay, the official name until 1995) is the capital city of the Indian state of Maharashtra. It is the most populous city in **India** …

The **Arabian Sea** is a region of the northern Indian Ocean bounded on the north by **Pakistan** and **Iran**, on the west by northeastern **Somalia** and the Arabian Peninsula, and on the east by **India** …

Q: (Hanging gardens of Mumbai, country, ?)
Options: {Iran, **India**, Pakistan, Somalia, …}



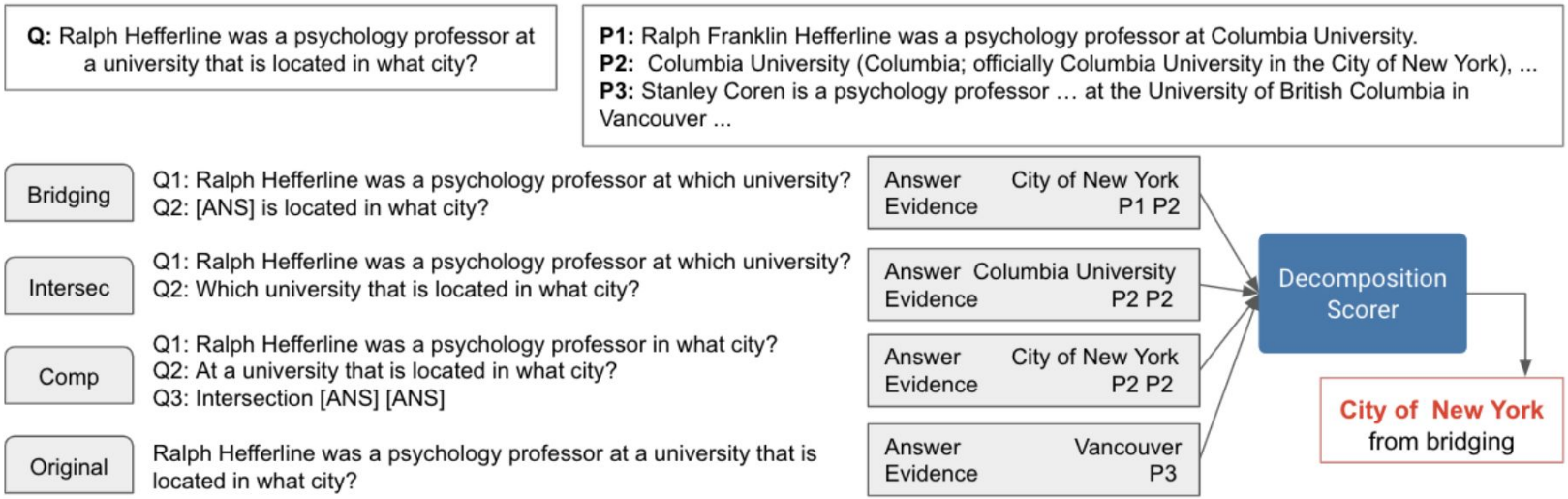
Background: HotpotQA

- HotpotQA dataset (Yang et al., EMNLP 2018) [web] [paper]
  - Crowdsourced using paragraphs in Wikipedia
  - Has “supporting facts” (annotated sentences)
  - Question types: bridging, comparison -> compositionality in questions

Reasoning Type	%	Example(s)
Inferring the *bridge entity* to complete the 2nd-hop question (Type I)	42	**Paragraph A:** The 2015 Diamond Head Classic was a college basketball tournament ... *Buddy Hield was named the tournament's MVP*. **Paragraph B:** *Chavano Rainier "Buddy" Hield* is a Bahamian professional basketball player for the **Sacramento Kings** of the NBA... **Q:** Which team does the player named 2015 Diamond Head Classic's MVP play for?
Comparing two entities (Comparison)	27	**Paragraph A:** LostAlone were a British rock band ... consisted of *Steven Battelle, Alan Williamson, and Mark Gibson*... **Paragraph B:** Guster is an American alternative rock band ... Founding members *Adam Gardner, Ryan Miller, and Brian Rosenworcel* began... **Q:** Did LostAlone and Guster have the same number of members? (**yes**)

Overview of how it works

1. Decompose the original Q into sub Qs according to a few reasoning types
2. Answer sub Qs
3. Score the answers and output the final answer (the scorer is trained)



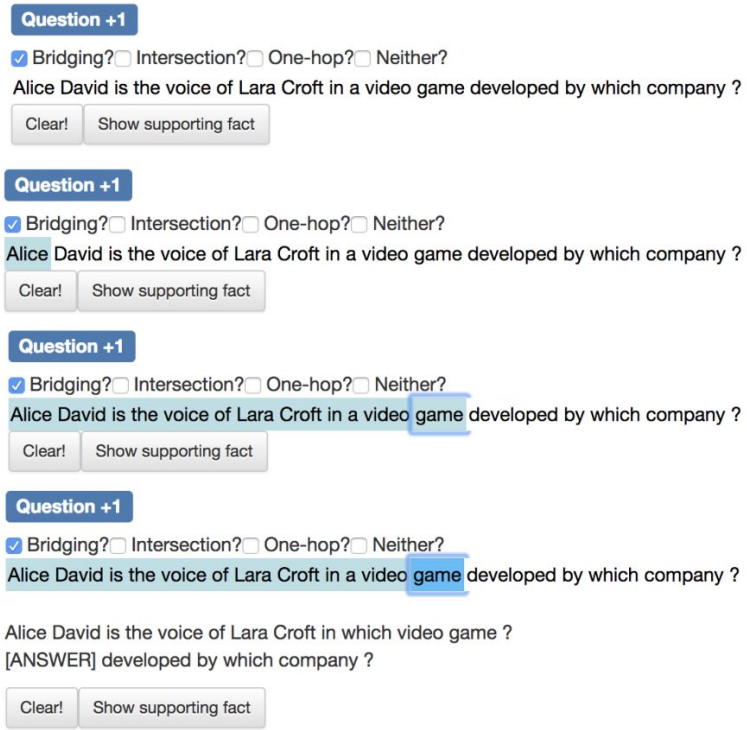
Pointer and Decomposition

- Train ***Pointer*** functions that point to a few indices
  - Bridging (3 indices) = (start, article, end) of bridging entity (clause)
  - Intersection (2 indices) = (start, end) of common information
  - Comparison (4 indices) = (start, end) of two entities (by Spacy NER)

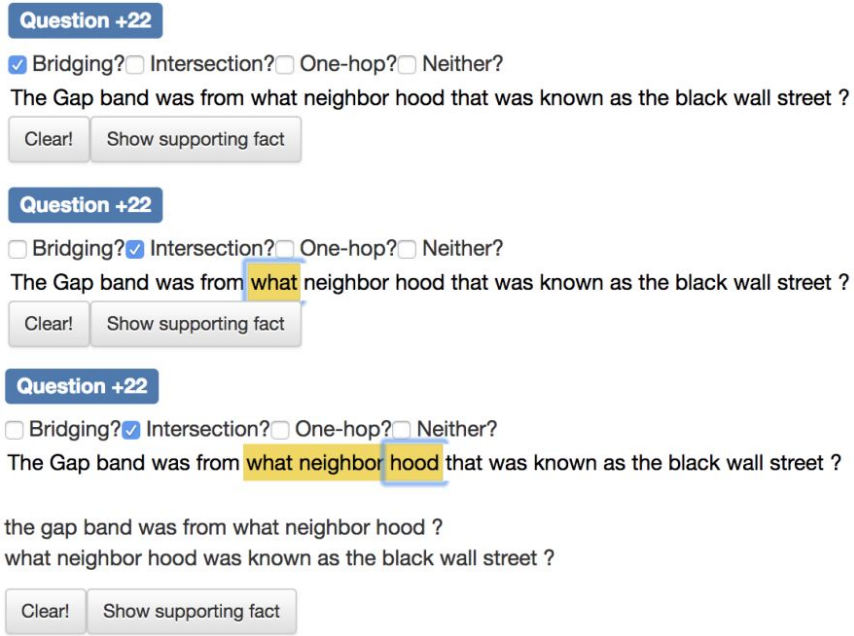
Type	**Bridging (47%)** requires finding the first-hop evidence in order to find another, second-hop evidence.
Q	Which team does the **player** named 2015 Diamond Head Classic's MVP play for?
Q1	Which player named 2015 Diamond Head Classic's MVP?
Q2	Which team does `ANS` play for?
Type	**Intersection (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`)

Span Annotation

Bridging (3 points)



Intersection (2 points)



Comparison Types

- Comparison types are hard-coded…
  - Numeric, logical, string match

  cf. DROP dataset (Dua et al., 2019)

procedure FIND_OPERATION(question, entity1, entity2)
coordination, preconjunct ← f(question, entity1, entity2)
Determine if the question is *either* question or *both* question from coordination and preconjunct
head entity ← f_{head}(question, entity1, entity2)
if *more, most, later, last, latest, longer, larger, younger, newer, taller, higher* in question **then**
  if head entity exists **then** discrete_operation ← Which is greater
  else discrete_operation ← Is greater
else if *less, earlier, earliest, first, shorter, smaller, older, closer* in question **then**
  if head entity exists **then** discrete_operation ← Which is smaller
  else discrete_operation ← Is smaller
else if head entity exists **then**
  discrete_operation ← Which is true
else if question is not yes/no question and asks for the property in common **then**
  discrete_operation ← Intersection
else if question is yes/no question **then**
  Determine if question asks for logical comparison or string comparison
  if question asks for logical comparison **then**
    if *either* question **then** discrete_operation ← Or
    else if *both* question **then** discrete_operation ← And
  else if question asks for string comparison **then**
    if asks for same? **then** discrete_operation ← Is equal
    else if asks for difference? **then** discrete_operation ← Not equal
return discrete_operation

Operation & Example
Type: Numeric
Is greater (ANS) (ANS) → yes or no
Is smaller (ANS) (ANS) → yes or no
Which is greater (ENT, ANS) (ENT, ANS) → ENT
Which is smaller (ENT, ANS) (ENT, ANS) → ENT
Did the Battle of Stones River occur before the Battle of Saipan?
Q1: The Battle of Stones River occur when? → 1862
Q2: The Battle of Saipan River occur when? → 1944
Q3: Is smaller (the Battle of Stones River, 1862) (the Battle of Saipan, 1944) → yes
Type: Logical
And (ANS) (ANS) → yes or no
Or (ANS) (ANS) → yes or no
Which is true (ENT, ANS) (ENT, ANS) → ENT
In between Atsushi Ogata and Ralpha Smart who graduated from Harvard College?
Q1: Atsushi Ogata graduated from Harvard College? → yes
Q2: Ralpha Smart graduated from Harvard College? → no
Q3: Which is true (Atsushi Ogata, yes) (Ralpha Smart, no) → Atsushi Ogata
Type: String
Is equal (ANS) (ANS) → yes or no
Not equal (ANS) (ANS) → yes or no
Intersection (ANS) (ANS) → string
Are Cardinal Health and Kansas City Southern located in the same state?
Q1: Cardinal Health located in which state? → Ohio
Q2: Cardinal Health located in which state? → Missouri
Q3: Is equal (Ohio) (Missouri) → no

Results

- 1hop train = trained only on single-hop Questions (provided in the dataset)
- single/multi = all three models of single-hop BERT obtain non-negative F1
- Settings
  - Distractor = two gold + eight distracting paragraphs
  - Full wiki = the first paragraphs of all Wikipedia articles

	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**
1hop train	61.73	61.57	62.36	79.38	46.53	39.17	35.30	54.57	50.03	29.83
BERT	67.08	69.41	57.81	82.98	53.38	38.40	34.77	52.85	46.14	31.74
1hop train	56.27	62.77	30.40	**87.21**	29.64	29.97	32.15	21.29	47.14	15.18
BiDAF	58.28	59.09	55.05	-	-	34.36	30.42	50.70	-	-

Development 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

Test set

Error Analysis

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?

1. Questions are not necessarily ***compositional***
2. Error question
3. Counting is still difficult
   cf. BERT-calculator https://arxiv.org/abs/1909.00109 , MTMSN https://arxiv.org/abs/1908.05514

SOTA NLP Yomikai: Compositional Questions Do Not Necessitate Multi-hop Reasoning

Min et al. (ACL 2019, short) [Paper]

Reader: Saku Sugawara
2019-09-28

Abstract

Question: What is the former name of the animal whose habitat the Réserve Naturelle Lomako Yokokala was established to protect?
Paragraph 5: The Lomako Forest Reserve is found in Democratic Republic of the Congo. It was established in 1991 especially to protect the habitat of the Bonobo apes.
Paragraph 1: The bonobo ("Pan paniscus"), formerly called the **pygmy chimpanzee** and less often, the dwarf or gracile chimpanzee, is an endangered great ape and one of the two species making up the genus "Pan".

[Paper]

ぼのぼ氏

- Answer type is animal. But only one of 10 paragraph is about an animal!
- The authors propose a single-paragraph BERT
  - Input = a single paragraph
  - Output = answer span (as usual) and 4 scalars for span/yes/no/empty
- It achieves a comparable performance with other SOTA methods

Really Multi-hop?

Reasoning Type	%	Example(s)
Inferring the *bridge entity* to complete the 2nd-hop question (Type I)	42	**Paragraph A:** The 2015 Diamond Head Classic was a college basketball tournament ... *Buddy Hield was named the tournament's MVP*. **Paragraph B:** *Chavano Rainier "Buddy" Hield* is a Bahamian professional basketball player for the **Sacramento Kings** of the NBA... **Q:** Which team does the player named 2015 Diamond Head Classic's MVP play for?
Comparing two entities (Comparison)	27	**Paragraph A:** LostAlone were a British rock band ... consisted of *Steven Battelle, Alan Williamson, and Mark Gibson*... **Paragraph B:** Guster is an American alternative rock band ... Founding members *Adam Gardner, Ryan Miller, and Brian Rosenworcel* began... **Q:** Did LostAlone and Guster have the same number of members? (**yes**)
Locating the **answer entity** by checking multiple properties (Type II)	15	**Paragraph A:** Several *current and former members of the Pittsburgh Pirates* ... John Milner, **Dave Parker**, and Rod Scurry... **Paragraph B:** **David Gene Parker**, *nicknamed "The Cobra"*, is an American former player in Major League Baseball... **Q:** Which former member of the Pittsburgh Pirates was nicknamed "The Cobra"?
Inferring about the property of an entity in question through a *bridge entity* (Type III)	6	**Paragraph A:** *Marine Tactical Air Command Squadron 28* is a United States Marine Corps aviation command and control unit based at *Marine Corps Air Station Cherry Point*... **Paragraph B:** *Marine Corps Air Station Cherry Point* ... is a United States Marine Corps airfield located in **Havelock, North Carolina**, USA ... **Q:** What city is the Marine Air Control Group 28 located in?
Other types of reasoning that require more than two supporting facts (Other)	2	**Paragraph A:** ... the towns of Yodobashi, **Okubo, Totsuka, and Ochiai town** *were merged into Yodobashi ward*. ... *Yodobashi Camera is a store with its name taken from the town and ward*. **Paragraph B**: *Yodobashi Camera* Co., Ltd. is *a major Japanese retail chain specializing in electronics, PCs, cameras and photographic equipment*. **Q:** Aside from Yodobashi, what other towns were merged into the ward which gave the major Japanese retail chain specializing in electronics, PCs, cameras, and photographic equipment it's name?

Examples from the original paper

- single "team name"?
  - no guarantee
- difficult
- XXX, nicknamed "Cobra"
- single "city name"?
  - no guarantee
- XXX were merged into...

Observations 1

Development set:

- Single-para BERT = **67.08** F1 / DecompRC = **70.57** F1

Human on 200 bridging questions

- (8 distractor para +) 2 gold para = **87.37** F1 / 1 gold para = **82.06** F1

Type	Question	%
Multi-hop	Ralph Hefferline was a psychology professor at a university that is located in what city?	27
Weak distractors	What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?	35
Redundant evidence	Kaiser Ventures corporation was founded by an American industrialist who became known as the father of modern American shipbuilding?	26
Non-compositional 1-hop	When was Poison's album 'Shut Up, Make Love' released?	8

Types of bridging questions: GOOD = 27%, BAD = 35+26+8 = 69%

Observations 2

Type	Question	%	F1
Multi-hop	Who was born first, Arthur Conan Doyle or Penelope Lively?	45	54.46
Context-dependent	Are Hot Rod and the Memory of Our People both magazines?	36	56.16
Single-hop	Which writer was from England, Henry Roth or Robert Erskine Childers?	17	70.54

Types of comparison questions and the performance of single-para BERT

- Multi-hop and Context-dependent = around chance accuracy
  → enough difficult for the single-paragraph model (& single hop is easy)

How can we ensure multi-hop reasoning?

- New adversarial distractor paras by model-based filtering (e.g., HellaSWAG)
  → the accuracy declines but it recovers when it is re-trained
  → Future work: develop improved method for distractor collection?

Summary and Kansou

- Question decomposition: interesting!
  - But compositional questions seem ***unnatural***
  - Can we create natural questions that require multi-hop reasoning?
    - Combination of knowledge base, math & logical operations, … ?

- Recent trends and future direction (?)
  - Large datasets (CNN/DailyMail, SQuAD etc.) (2015~2017)
  - Strong general models (GPT, BERT, XLNet, RoBERTa, ALBERT...)
  - Skill-oriented datasets (2018~)
    - SQuAD v2, CoQA, HotpotQA, DROP, CosmosQA, QuoRef, ...
  - Skill-specific architecture on top of strong general models (this etc.)

  → Comprehensive datasets & more general models?（できるのか？）