

Nov 2020



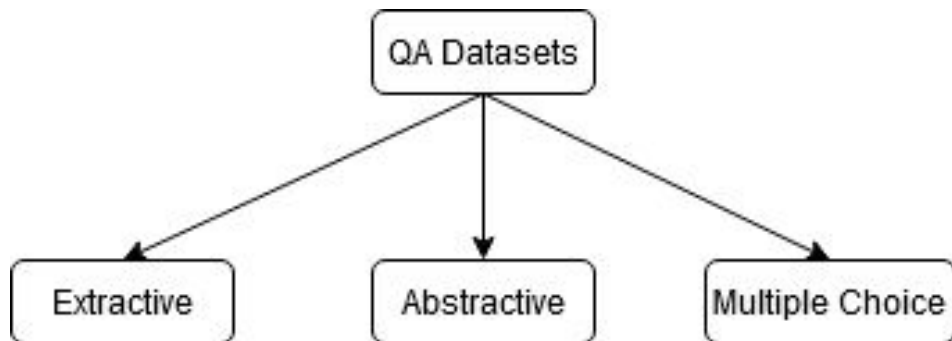
QA Datasets

CS 395T : Topics in Natural Language Processing

Tharun Mohandoss, Maohua Wang

tm35848, mw37285 The University of Texas at Austin

QA Dataset Classification

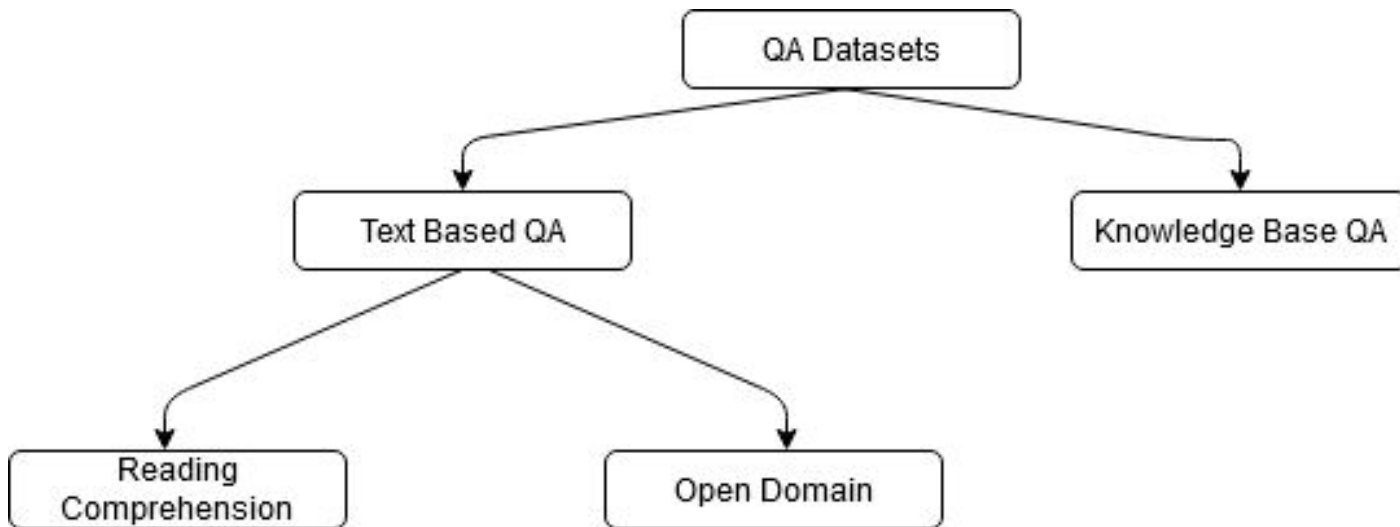


Extractive : Span in the evidence document

Abstractive : Free String/Answer needs to be generated

Multiple Choice : Pick from different options

QA Dataset Hierarchy



QA Dataset Knowledge Based

Single Relation Dataset: One fact in the KB used to answer the question. Eg : Simple Questions (Bordes Et al.[2015]). SQ contains 76K, 11K and 21K for training, development and test

Multi Relation Dataset: Answering a question may require combining information from multiple facts in the KB.

Eg : WebQSP(Yih et al., 2016). It contains 2848 training questions, 250 development questions and 1639 test questions, Complex Web Questions (CWQ)(Talmor and Berant 2018) (27k, 3k , 3k)



Reading Comprehension



RC Based

- **SQuAD(Rajpurkar et al. 2016) , Stanford Question Answering Dataset** : 100,000+ questions posed by crowdworkers on a set of Wikipedia articles where the answer to each question is a segment of text from the corresponding reading passage
- **SQuAD 2.0 (Rajpurkar et al. 2018)** : SQuAD 2.0 combines existing SQuAD data with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones
- **TREC** : Question Answering Track since 1999. Challenge to retrieve text snippets as answers for open domain closed-class questions.

RC Based

- Popular existing RC datasets had been solved by 2018/2019, Example : [Devlin et al.](#) SQuAD v2.0 F1 of 83.1%.
- More challenging dataset required
- Many datasets with some added additional complexities :
- Eg :
 - **DROP**
 - CoQA(Reddy et al. 2019), QuAC(Choi et al 2018)
 - TriviaQA(Joshi et al. 2017), HotPotQA(Yang et al.),
COMPLEXWEBQUESTIONS(Talmor and Berant 2018)
 - Duorc(Saha et al.) 2018, NarrativeQA (Kociksky et al. 2017)
 - OpenBookQA(Mihaylov et al.), ReCoRD(Zhang et al 2018)
 - ProPara(Mishra et al 2018), Mcscript(Ostermann et al. 2018)
 - ([Kashabi et al. 2018](#)),

RC Based

- Requiring Tracking of conversational State :
 - CoQA(Reddy et al. 2019) :
127k question and answers from 8k conversations about text passages from seven diverse domains. Questions are conversational
 - QuAC(Choi et al 2018) :(Question answering in context)
Its questions are often more open-ended, unanswerable, or only meaningful within the dialog context.

RC Based

CoQA :

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q₁: Who had a birthday?

A₁: Jessica

R₁: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q₂: How old would she be?

A₂: 80

R₂: she was turning 80

Q₃: Did she plan to have any visitors?

A₃: Yes

R₃: Her granddaughter Annie was coming over

Q₄: How many?

A₄: Three

R₄: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q₅: Who?

A₅: Annie, Melanie and Josh

R₅: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Figure 1: A conversation from the CoQA dataset. Each turn contains a question (Q_i), an answer (A_i) and a rationale (R_i) that supports the answer.

RC Based

QuAC:

Section:  **Daffy Duck, Origin & History**

STUDENT: **What is the origin of Daffy Duck?**

TEACHER: ↔ first appeared in Porky's Duck Hunt

STUDENT: **What was he like in that episode?**

TEACHER: ↔ assertive, unrestrained, combative

STUDENT: **Was he the star?**

TEACHER: ↔ No, barely more than an unnamed bit player in this short

STUDENT: **Who was the star?**

TEACHER: ↗ No answer

STUDENT: **Did he change a lot from that first episode in future episodes?**

TEACHER: ↔ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc

STUDENT: **How has he changed?**

TEACHER: ↔ Daffy was less anthropomorphic

STUDENT: **In what other ways did he change?**

TEACHER: ↔ Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.

STUDENT: **Why did they add the lisp?**

TEACHER: ↔ One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.

STUDENT: **Is there an "unofficial" story?**

TEACHER: ↔ Yes, Mel Blanc (...) contradicts that conventional belief

...

RC Based

TriviaQA (Joshi et al. 2017):

- has relatively complex, compositional questions
- has considerable syntactic and lexical variability between questions and corresponding answer-evidence sentences
- requires more cross sentence reasoning to find answers.

Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?

Answer: The Guns of Navarone

Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

RC Based

HotPotQA (Yang et al. 2018):

- Requires finding and reasoning over multiple supporting documents to answer
- the questions are diverse and not constrained to any pre-existing knowledge bases or knowledge schemas
- Provides sentence-level supporting facts required for reasoning, allowing QA systems to reason with strong supervision and explain the predictions
- Offers a new type of factoid comparison questions to test QA systems' ability to extract relevant facts and perform necessary comparison

RC Based

HotPotQA

Paragraph A, Return to Olympus:

[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:

[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?

A: Malfunkshun

Supporting facts: 1, 2, 4, 6, 7

RC Based

DuoRC (Saha et al. 2018) contains 186,089 unique question-answer pairs created from a collection of 7680 pairs of movie plots. Two version of the same movie were used i.e one from Wiki and other from IMDB. Questions from one and Answers from the other ensuring to make it more challenging and avoiding lexical overlap.

Movie: Twelve Monkeys

Shorter Plot Synopsis (Wikipedia)

A deadly virus released in 1996....[James Cole is a prisoner living in a subterranean compound beneath the ruins of Philadelphia.]^{Q1} [Cole is selected for a mission]^{Q2}, ...

[Cole arrives in Baltimore]^{Q3} in 1990, not 1996 as planned...[Goines denies any involvement with the group and says that in 1990 Cole originated the idea of wiping out humanity with a virus stolen from Goines' virologist father.]^{Q4}

Cole convinces himself... [Railly confronts him with evidence of his time travel.]^{Q5} [They decide to spend their remaining time together in the Florida Keys before the onset of the plague]^{Q6}. ...

[At the airport, Cole leaves a last message]^{Q7} [He is soon confronted by Jose, an acquaintance from his own time, who gives Cole a handgun]^{Q8} and ambiguously instructs him to follow orders. At the same time, Railly spots Dr. Peters....

Cole forces his way through a security checkpoint.... [Peters, aboard the plane with the virus]^{Q9}, ...

Longer Plot Synopsis (IMDB)

The time is the indeterminate future. A virus, deliberately released in 1996 ... One such prisoner is [James Cole, who after retrieving samples is given the chance to go back in time to 1996]^{Q2} and find information about the group believed responsible, known as "The Army of 12 Monkeys."

Throughout the ensuing episodes, Cole ... There he meets Jeffrey Goines, ... Cole is now racing against time... he wants to stay in 1996 with Dr. Railly, ...They [travel to Philadelphia]^{Q1}, eventually finding ... [Dr. Railly ... She becomes convinced that "The Army of 12 Monkeys" indeed poses a threat, and she persuades Cole to take up his cause again]^{Q5}.They travel to Jeffrey's...

[Jeffrey rambles about how Cole had given him the idea to release a virus that would destroy most of humanity.]^{Q4} Cole leaves, ...and then posts flyers declaring "We did it!" [Cole realizes that the "Army" is not the threat, and he leaves a phone message to that effect]^{Q7}.

Shortly after, [Jose, a fellow "volunteer" from the present, approaches Cole with orders for him to complete his mission and hands him a revolver]^{Q8}... In an airport, while attempting with Cole to elude capture, Dr. Railly recognizes [Dr Peters, a man who worked with Jeffrey Goines's father The man goes through airport screening and manages to persuade security that his biological samples]^{Q9}...

Q1: James Cole is a prisoner living in a subterranean shelter beneath what city? Philadelphia, Philadelphia

Q2: What is the name of the person selected for the mission? James Cole, James Cole

Q3: Where did Cole arrive in 1990? Baltimore, -

Q4: Who does Goines claim came up with the idea to exterminate humanity? Cole, Cole

Q5: What does Railly confront Cole with? Evidence of his time travel, The "Army of 12 Monkeys" poses a threat

Q6: Where do Cole and Railly decide to go before the plague? Florida Keys, -

Q7: Where does Cole leave his message? At the airport, on the phone

Q8: Who gives Cole a handgun? Jose, Jose

Q9: Peters is aboard the plane with what? Virus, biological samples

RC Based

OpenBookQA (Mihaylov et al. 2018): It has around 6000 questions that can be answered using 1326 elementary level science facts. The questions require combining an open book fact and some other common sense facts/s from other sources.

NarrativeQA(Kociksky et al. 2017) : It has questions that require a reader to read an entire books or movie script while understanding the narrative to answer. The motivation is that the model must require understanding the underlying narrative rather than rely on shallow pattern matching alone.

RC Based

OpenBookQA

Question:

Which of these would let the most heat travel through?

- A) a new pair of jeans.
- B) a steel spoon in a cafeteria.
- C) a cotton candy at a store.
- D) a calvin klein cotton hat.

Science Fact:

Metal is a thermal conductor.

Common Knowledge:

Steel is made of metal.

Heat travels through a thermal conductor.

Narrative QA

Title: Ghostbusters II

Question: How is Oscar related to Dana?

Answer: her son

Summary snippet: ...Peter's former girlfriend Dana Barrett has had a son, Oscar. . .

Story snippet:

DANA (setting the wheel brakes on the buggy)
Thank you, Frank. I'll get the hang of this eventually.

She continues digging in her purse while Frank leans over the buggy and makes funny faces at the baby, OSCAR, a very cute nine-month old boy.

FRANK (to the baby)

Hiya, Oscar. What do you say, slugger?

FRANK (to Dana)

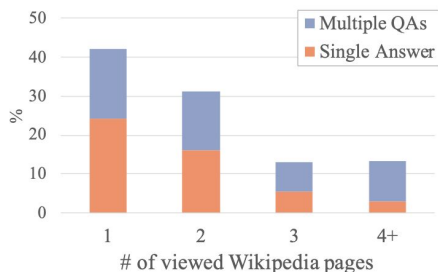
That's a good-looking kid you got there, Ms. Barrett.

Open Question Answering

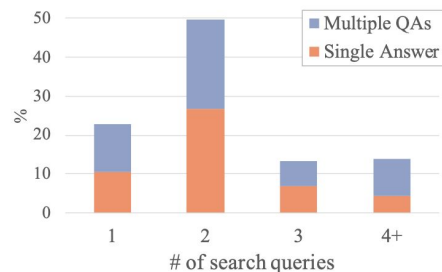
- OTT-QA (Open Table-and-Text QA) by Chen, [Wenhu, et al.](#)
 - expand on HybridQA
- XORQA (Cross-lingual Open-Retrieval QA) by [Asai, Akari, et al.](#)
 - 40k information seeking questions from 7 non-English languages
 - Professional translation
- AmbigQA by [Min, Sewon, et al.](#)
 - expands on NQ-Open
 - find all plausible answers to a question (50% of NQ-Open)
 - rewrite questions to resolve ambiguity

Open Question Answering

- OTT-QA (Open Table-and-Text QA) by Chen, [Wenhu, et al.](#)
 - expand on HybridQA
- XORQA (Cross-lingual Open-Retrieval QA) by [Asai, Akari, et al.](#)
 - 40k information seeking questions from 7 non-English languages
 - Professional translation
- AmbigQA by [Min, Sewon, et al.](#)
 - expands on NQ-Open
 - find all plausible answers to
 - rewrite questions to resolve



(a) Number of unique Wikipedia pages visited by crowdworkers.[†]



(b) Number of search queries written by crowdworkers.

Open Question Answering

- NarrativeQA
- ELI

Nov 2020



DROP : A Reading Comprehension Benchmark Requiring Discrete Reasoning over Paragraphs

CS 395T : Topics in Natural Language Processing

Tharun Mohandoss

tm35848, The University of Texas at Austin

DROP

- Crowdsourced, Adversarially created 96k Question benchmark
- To perform well, the model would need to resolve references to many input positions and perform discrete operations over them(+, - , sorting etc.)
- Requires a more thorough understanding of the content of paragraphs.

Motivation

- Popular RC datasets have been solved, Example : [Devlin et al.](#) SQuAD v2.0 F1 of 83.1%
- More challenging dataset required
- Existing systems are brittle (Eg. : [Robin Jia et al.](#))
- Need to push the field towards more comprehensive analysis of text.

Contributions

- Dataset collection : Created the 96k question benchmark through crowdsourcing while using [BiDAF](#) (Seo et al.) as an adversary to ensure that the questions are challenging
- Show that existing SOTA in RC and Semantic parsing literatures achieve only 32.7% F1 on this dataset.
- Present a new model that achieves 47% F1 by combining RC methods with numerical reasoning.

Related Work : QA Datasets

Related Work	Explanation/Additional Complexity
Reddy et al. 2019, Choi. et al 2018	Tracking Conversational State
Joshi et al. 2017, Yang et al. 2018, Talmor and Berant, 2018	Passage Retrieval
Sasha et al. 2018, Kociksky et al. 2018, Rajpurkar et al. 2018	Mismatched Passages and Questions
Mihaylov et al., 2018; Zhang et al. 2019	Integrating knowledge from external sources
Mishra et al., 2018; Ostermann et al., 2018	Tracking Entity State Changes
Welbl et al., 2018; Khashabi et al., 2018	Multistep reasoning of multiple documents
(Pampari et al., 2018; Suster and Daelemans, 2018	Medical domain datasets

Related Work : QA Datasets

- DROP has none of these additional complexities
- DROP focuses on passage understanding but adds an additional complexity of requiring discrete/numerical reasoning
- Algebraic word problem datasets Koncel-Kedziorski et al., 2015; Kushman et al., 2014; Hosseini et al., 2014; Clark et al., 2016; Ling et al., 2017 contain similar reasoning requirement but DROP is more open domain and requires deeper paragraph understanding

Related Work : Semantic Parsing

- Zelle and Mooney, 1996; Zettlemoyer and Collins, 2005; Berant et al., 2013a try to understand complex compositional question semantics in terms of grounded knowledge base or other environments.
- Questions in DROP are modeled based on WikiTableQuestions dataset (Pasupat and Liang, 2015) but DROP is for paragraph understanding.

Related Work : WikiTableQuestions

- Wikitable Questions : Each question in this is associated with a table from Wikipedia.
- 2108 HTML tables from wikipedia with 22k question-answer pairs
- Questions in DROP are modeled based on WikiTable Questions dataset (Pasupat and Liang, 2015) but DROP is for paragraph understanding.

Related Work : WikiTableQuestions

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

x_1 : "Greece held its last Summer Olympics in which year?"

y_1 : {2004}

x_2 : "In which city's the first time with at least 20 nations?"

y_2 : {Paris}

x_3 : "Which years have the most participating countries?"

y_3 : {2008, 2012}

x_4 : "How many events were in Athens, Greece?"

y_4 : {2}

x_5 : "How many more participants were there in 1900 than in the first year?"

y_5 : {10}

Related Work : Adversarial Dataset Construction

- Some recent works Paperno et al., 2016; Minervini and Riedel, 2018; Zellers et al., 2018; Zhang et al., 2019; Zellers et al., 2019 use adversarial baselines but these use the baseline to filter automatically generated samples
- DROP incorporates the adversarial framework in a crowd sourcing context.

Related Work : Neural Symbolic Reasoning

- DROP encourages solutions that combine neural and Symbolic/Discrete reasoning methods.
- Other works are : Reed and de Freitas (2016), Neelakantan et al. (2016),and Liang et al. (2017)

DROP : Dataset Collection

1. Extract passages from Wikipedia for which generating complex question is easy.
2. Create question-answer pairs using crowdsourcing while ensuring that questions require discrete reasoning.
3. Validate the development and test portions of DROP to ensure quality and report inter-annotator agreement.

Dataset Collection : Extracting Passages

1. Wikipedia passages with a narrative sequence with high proportion of numbers were chosen.
2. NFL game summaries + History articles + Any passage with 20+ numbers.
3. This process yields 7000 passages

Dataset Collection : Question Collection

1. Workers presented with 5 passages and require to produce 12 QA pairs.
2. Examples questions from semantic parsing literature were shown to elicit questions requiring more comprehension.
3. Only allowed to submit questions that BiDAF could not solve.
4. Only three types of answers allowed : spans of text from either question or passage, date and numbers.
5. Collected 96567 QA pairs.

Dataset Collection : Validation

- Two additional answers collected using crowdsourcing for each QA pair.
- Good resulting inter-annotator agreement.

DROP : Data Analysis

Answer Type	Percent	Example
NUMBER	66.1	12
PERSON	12.2	Jerry Porter
OTHER	9.4	males
OTHER ENTITIES	7.3	Seahawks
VERB PHRASE	3.5	Tom arrived at Acre
DATE	1.5	3 March 1992

Table 3: Distribution of answer types in training set, according to an automatic named entity recognition.

Question Analysis : Most frequent trigram pattern “Which team scored” appears only in 4% of the span type questions indicating the huge variety in linguistic constructs.

Answer Analysis : On average 2.14 spans are needed to be considered to answer a question and a majority of the answers are numerical values and proper nouns.

Evaluation : Metrics

Exact Match

- Removes articles and other simple normalization

F1 Score(Numeracy focused)

- SQuAD based F1 score modified to become 0 when there is a number mismatch between gold and predicted answers.

Baselines

Semantic Parsing : Grammar-constrained semantic parsing model built by Krishnamurthy et al. (2017)(KDG) for the WIKITABLEQUESTIONS tabular database.

In order to represent paragraphs as structured contexts in order to run KDG, the authors choose three paradigms/sentence representation schemes :

- Stanford dependencies (de Marneffe and Manning, 2008, SynDep)
- Open Information Extraction (Banko et al.,2007, Open IE)
- Semantic Role Labeling (Carreras and M`arquez,2005, SRL)

Logical Form Language

- Predicate argument structures, strings, dates, numbers + functions
- Used these to induce a grammar
- Context specific rules to produce strings occurring in both Question and Passage

Training : The KDG parser maximizes the marginal likelihood of a set of (possibly spurious) question logical forms that evaluate to the correct answer.

Baselines : RC

- BiDAF, QANet, QANet + ELMo, BERT
- These model require a few minor adaptations when training on DROP.

Baselines : Heuristics

- Question Only/ Paragraph Only
- Most frequent answers for each question word

NAQANET

- Numerically Aware QANET = NAQANET
- Combines Neural RC and Symbolic Reasoning
- First predicts answer type as a span from question/count/arithmetic expression
- Neural Architecture produces a partially executed logical form which symbolic reasoning system solves.

NAQANET : Model

- A few layers are added on top of the original QANet's architecture without the output layer(embedding + encoding + Passage question attention). They get Question representation \mathbf{Q} and question aware passage representation \overline{P} .
- 4 different output layers for each kind of output/answer that the model can produce i.e
 - Passage Span
 - Question Span
 - Count
 - Arithmetic Expression

Output Layer : Passage Span

- Apply three repetition of QANet encoder on \bar{P} to get \mathbf{M}_0 , \mathbf{M}_1 , and \mathbf{M}_2
- Starting and Ending position of passage computed as

$$\mathbf{p}^{\text{p-start}} = \text{softmax}(\text{FFN}([\mathbf{M}_0; \mathbf{M}_1])),$$

$$\mathbf{p}^{\text{p-end}} = \text{softmax}(\text{FFN}([\mathbf{M}_0; \mathbf{M}_2]))$$

Where FFN is a two layer feed forward network with RELU

Output Layer : Question Span

- First they Compute h^P as follows :

$$\alpha^P = \text{softmax}(\mathbf{W}^P \bar{\mathbf{P}}),$$

$$\mathbf{h}^P = \alpha^P \bar{\mathbf{P}}$$

- Then they compute starting and ending position in the questions

$$\mathbf{p}^{\text{q-start}} = \text{softmax}(\text{FFN}([\mathbf{Q}; \mathbf{e}^{|\mathbf{Q}|} \otimes \mathbf{h}^P])),$$

$$\mathbf{p}^{\text{q-end}} = \text{softmax}(\text{FFN}([\mathbf{Q}; \mathbf{e}^{|\mathbf{Q}|} \otimes \mathbf{h}^P]))$$

Output Layer : Count/Arithmetic Expression

- Count treated as a multiclass classification problem with the 10 digits from 0-9 as the possibilities.
- For arithmetic expression, they extract add numbers and assign a +/- or 0 to each number.
- They apply one more round of QANet encoder to M_2 to get M_3
- Then they select an index over concatenation of M_0 and M_3 to get a representation of each number in the passage. The i th number is represented as h_i^N and the probabilities of +/- and 0 are then,

$$\mathbf{p}_i^{\text{sign}} = \text{softmax}(\text{FFN}(\mathbf{h}_i^N))$$

Output Layer : Answer type prediction

- They use a categorical variable to decide between above four answer types with probabilities computed as

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

Weakly Supervised Training

- They find all executions that lead to the right answer and maximize the marginal likelihood of these executions

Results

Method	Dev		Test	
	EM	F ₁	EM	F ₁
Heuristic Baselines				
Majority	0.09	1.38	0.07	1.44
Q-only	4.28	8.07	4.18	8.59
P-only	0.13	2.27	0.14	2.26
Semantic Parsing				
Syn Dep	9.38	11.64	8.51	10.84
OpenIE	8.80	11.31	8.53	10.77
SRL	9.28	11.72	8.98	11.45

Method	Dev		Test	
	EM	F ₁	EM	F ₁
SQuAD-style RC				
BiDAF	26.06	28.85	24.75	27.49
QANet	27.50	30.44	25.50	28.36
QANet+ELMo	27.71	30.33	27.08	29.67
BERT	30.10	33.36	29.45	32.70
NAQANet				
+ Q Span	25.94	29.17	24.98	28.18
+ Count	30.09	33.92	30.04	32.75
+ Add/Sub	43.07	45.71	40.40	42.96
Complete Model	46.20	49.24	44.07	47.01
Human	-	-	94.09	96.42

Error Analysis

- Conducted error analysis on a random sample of 100 wrong answers
 - Arithmetic operations - 51%
 - Counting - 30%
 - Domain Knowledge and common sense - 23%
 - Co-reference - 6%

Phenomenon	Passage Highlights	Question	Answer	Our model
Subtraction + Coreference	... Twenty-five of his 150 men were sick, and his advance stalled ...	How many of Bartolom de Amsqueta’s 150 men were not sick?	125	145
Count + Filter	... Macedonians were the largest ethnic group in Skopje, with 338,358 inhabitants ... Then came ... Serbs (14,298 inhabitants), Turks (8,595), Bosniaks (7,585) and Vlachs (2,557) ...	How many ethnicities had less than 10000 people?	3	2
Domain knowledge	... Smith was sidelined by a torn pectoral muscle suffered during practice ...	How many quarters did Smith play?	0	2
Addition	... culminating in the Battle of Vienna of 1683, which marked the start of the 15-year-long Great Turkish War ...	What year did the Great Turkish War end?	1698	1668

Table 5: Representative examples from our model’s error analysis. We list the identified semantic phenomenon, the relevant passage highlights, a gold question-answer pair, and the erroneous prediction by our model.

Current SOTA

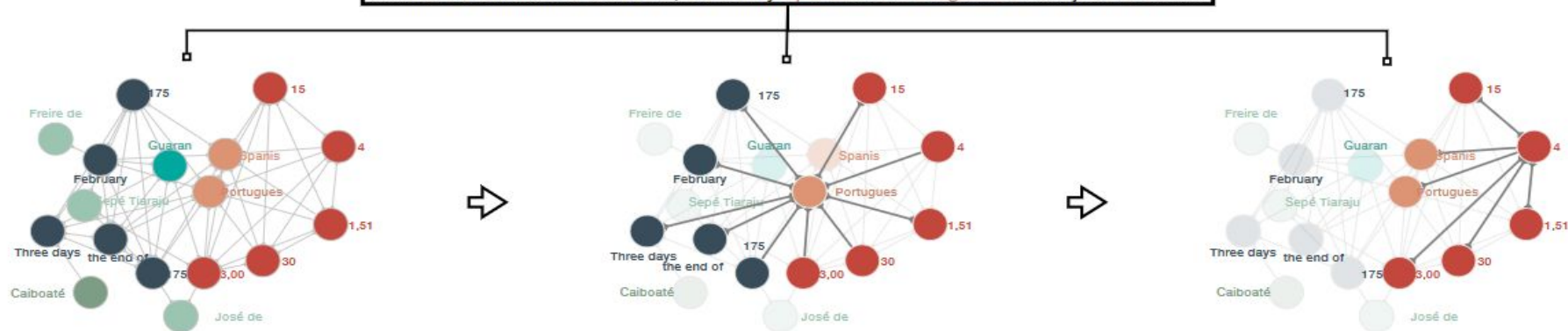
Question directed graph attention network (QDGAT)(Chen et al. 2020) :

- F1 : 0.9010, Exact Match : 0.8704
- They construct a heterogeneous directed graph where nodes can be different types of entities and different types numbers and edges encode different types of relations between nodes.
- Using contextual encoders such as BERT and ROBERTa to extract representations of numbers and entities in both question and passage to serve as the initial embeddings of each node in the graph.
- After several message passing iterations, QDGAT aggregates node information to answer the questions.

Table 1: Two MRC cases requiring numerical reasoning are illustrated. There are entities and numbers of different types. Both are emphasized by different colors: **entity**, **number**, **percentage**, **date**, **ordinal**. We explicitly encode the type information into our model and leverage the question representation to conduct the reasoning process.

Question	Passage	Answer
At the battle of Caiboatá how many Spanish and Portuguese were injured or killed?	... In 1754 Spanish and Portuguese military forces were dispatched to force the Guarani to leave the area ... Hostilities resumed in 1756 when an army of 3,000 Spanish , Portuguese , and native auxiliary soldiers under José de Andonaegui and Freire de Andrade was sent to subdue the Guarani rebels. On February 7, 1756 the leader of the Guarani rebels, Sepé Tiaraju , was killed in a skirmish with Spanish and Portuguese troops. ... 1,511 Guarani were killed and 152 taken prisoner, while 4 Spanish and Portuguese were killed and about 30 were wounded...	34
In which quarter did Stephen Gostkowski kick his shortest field goal of the game?	The Cardinals' east coast struggles continued in the second quarter as quarterback Matt Cassel completed a 15- yard touchdown pass to running back Kevin Faulk and an 11- yard touchdown pass to wide receiver Wes Welker , followed by kicker Stephen Gostkowski's 38- yard field goal. In the third quarter , Arizona's deficit continued to climb as Cassel completed a 76- yard touchdown pass to wide receiver Randy Moss , followed by Gostkowski's 35- and 24- yard field goal. In the fourth quarter , New England concluded its domination with Gostkowski's 30- yard	third

Question: At the battle of Caiboaté, how many Spanish and Portuguese were injured or killed?



Discussion

- Why report statistics for the BiDAF model when we used it as an adversary? Is it meaningful?
- Current SOTA was reached within 2 years of dataset release, does that mean that technology improved so much in the last two years or just that existing methods were modified slightly to overfit this dataset.

Nov 2020



Commonsense Datasets: ATOMIC and SOCIAL IQA

CS 395T : Topics in Natural Language Processing

Maohua Wang

The University of Texas at Austin

Related Works

- Commensense Benchmarks - WSC and COPA
 - expert curated, high quality but size too small
- Commonsense Knowledge Bases (knowledge graph)
 - good for reasoning and downstream applications
- Constrained or Adversarial Data Collection
 - SQuAD, 50k unanswerable questions
 - adversarial filtering of generated incorrect answers to minimize surface patterns(Zellers et al)

Contributions

- Social and emotional intelligence important in daily lives - helps to achieve human-like AI
- Models trained on text corpora limited by reporting bias of knowledge
- First large (38K questions) common sense QA dataset by crowdsourcing
- adversarial question-switched answers to minimize annotation artifacts
- State-of-the-art BERT model only achieving 64.5% - great room for improvement!
- Social IQA can be used for transfer learning to solve other common sense tasks

Examples

Inference to the past

REASONING ABOUT MOTIVATION

Tracy had accidentally pressed upon Austin in the small elevator and it was awkward.

Q Why did Tracy do this?

- A**
- (a) get very close to Austin
 - (b) squeeze into the elevator ✓
 - (c) get flirty with Austin

Inference to the future

REASONING ABOUT WHAT HAPPENS NEXT

Alex spilled the food she just prepared all over the floor and it made a huge mess.

Q What will Alex want to do next?

- A**
- (a) taste the food
 - (b) mop up ✓
 - (c) run around in the mess

REASONING ABOUT EMOTIONAL REACTIONS

In the school play, Robin played a hero in the struggle to the death with the angry villain.

Q How would others feel afterwards?

- A**
- (a) sorry for the villain
 - (b) hopeful that Robin will succeed ✓
 - (c) like Robin should lose

ATOMIC Dataset

- commonsense reasoning knowledge graph
- Relation types
 - if-event-then-mental-state
 - if-event-then-event
 - if-event-then-persona

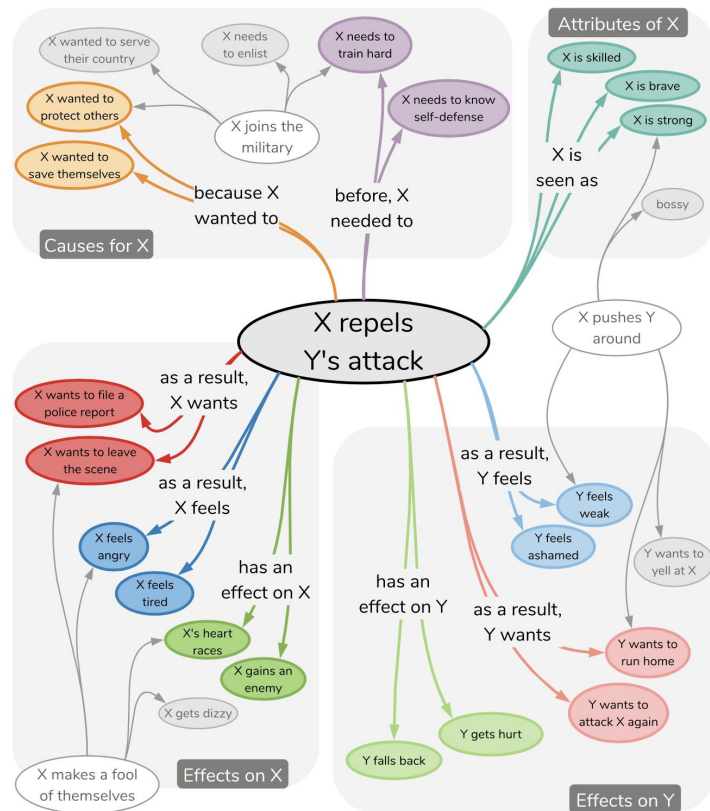


Figure 1: A tiny subset of ATOMIC, an atlas of machine commonsense for everyday events, causes, and effects.

Dataset Creation

- Pre-processing
 - PersonX spills ___ all over the floor
 - Alex spilled food all over the floor
- expands to <Context, Question, Answer>
- Post-processing - make the machine learn the right thing
 - Handwritten Incorrect Answers
 - similar in terms of words used, length, style
 - Question-Switching Answers
 - include answers to other questions but in a different context

Question-Switching Answers

- Select incorrect answers of different questions from within the same context.
- Make it harder for the model to hack through, either through incorrect answer bias or others

Alex spilt food all over the floor and it made a huge mess.

WHAT HAPPENS NEXT	WHAT HAPPENED BEFORE
What will Alex want to do next? <input checked="" type="checkbox"/> mop up <input checked="" type="checkbox"/> give up and order take out <input checked="" type="checkbox"/> have slippery hands <input checked="" type="checkbox"/> get ready to eat	What did Alex need to do before this? <input checked="" type="checkbox"/> have slippery hands <input checked="" type="checkbox"/> get ready to eat




Figure 2: Question-Switching Answers (QSA) are collected as the correct answers to the wrong question that targets a different type of inference (here, reasoning about what happens before instead of after an event).

Dataset Validation

- 3 workers to answer, 5 workers to validate
 - majority voting, otherwise discard the tuple
- human performance: 84-87%

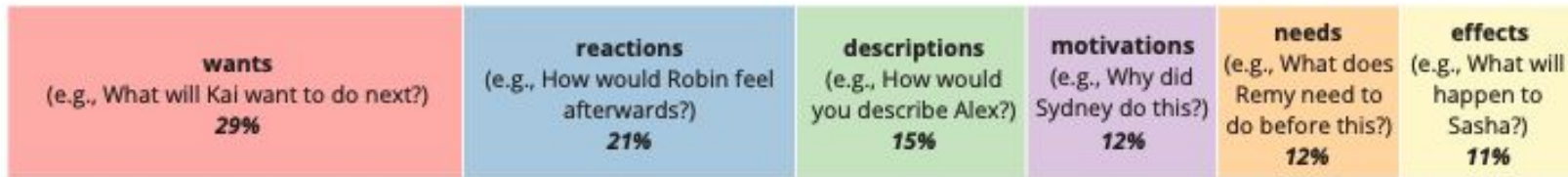


Figure 3: SOCIAL IQA contains several question types which cover different types of inferential reasoning. Question types are derived from ATOMIC inference dimensions.

Baseline Performance

- OpenAI-GPT and BERT
- BERT-large projected to match human performance with 1 million examples
 - but what does it mean if it achieves better performance than human?
- model succeeds at predicting motivation and actions but does not so well at involuntary effects
 - maybe better physics/ causation/ world model help?

Model	Accuracy (%)	
	Dev	Test
Random baseline	33.3	33.3
GPT	63.3	63.0
BERT-base	63.3	63.1
BERT-large	66.0	64.5
w/o context	52.7	–
w/o question	52.1	–
w/o context, question	45.5	–
Human	86.9*	84.4*

Table 2: Experimental results. We additionally perform an ablation by removing contexts and questions, verifying that both are necessary for BERT-large’s performance. Human evaluation results are obtained using 900 randomly sampled examples.

Transfer Learning

- Better performance
 - 3-5% higher accuracy
 - tighter spreads
- Can be used to improve previous model
 - different datasets compensate each other

Task	Model	Acc. (%)		
		best	mean	std
COPA	Sasaki et al. (2017)	71.2	–	–
	BERT-large	80.8	75.0	3.0
	BERT-SOCIAL IQA	83.4	80.1	2.0
WSC	Kocijan et al. (2019)	72.5	–	–
	BERT-large	67.0	65.5	1.0
	BERT-SOCIAL IQA	72.5	69.6	1.7
DPR	Peng et al. (2015)	76.4	–	–
	BERT-large	79.4	71.2	3.8
	BERT-SOCIAL IQA	84.0	81.7	1.2

Table 4: Sequential finetuning of BERT-large on SOCIAL IQA before the task yields state of the art results (bolded) on COPA (Roemmele et al., 2011), Winograd Schema Challenge (Levesque, 2011) and DPR (Rahman and Ng, 2012). For comparison, we include previous published state of the art performance.

Takeaway from SOCIAL IQA

- Emotion and common sense dataset
 - human performance at around 85%
- crowdsourcing
 - how to generate negative answers and adversarial data
 - question-switching answers
- 20% accuracy room of improvement!

Comparisons

- Worker disagreement
 - SocialQA discards jobs that cannot achieve majority voting
 - OTT-QA reassigns the job until accepted (71% overall rate)
 - AmbigQA tries to solve disagreement by breaking it down to smaller segments
- Worker types
 - generators and validators, maybe also coauthors
 - workers agree on most (90% in AmbigQA), not all

Dynabench

- QA Datasets saturate quickly
- Benchmarks usually have artifacts
- Benchmarks can be deceiving
- Research work usually overfits to benchmarks

Dynabench is a dynamic benchmarking platform developed with the goal to address these issues. Dynabench collects datasets adversarially using current SOTA models and releases them in multiple rounds.

QA Datasets Takeaway

- Earlier datasets such as SQuAD have been solved
- To encourage more thorough understanding and with the aim of creating harder tasks, several new datasets have been created.
- Each dataset adds unique new complexities to the QA task.