Know What You Don’t Know: Unanswerable Questions for SQuAD

ACL2018論文紹介

2018年9月10日

武田浩一
名古屋大学大学院情報学研究科
価値創造研究センター

© 2018 Nagoya University

Know What You Don't Know: Unanswerable Questions for SQuAD

Pranav Rajpurkar, Robin Jia, Percy Liang (Computer Science Department, Stanford University)

Article: Endangered Species Act
Paragraph: "*... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

質問1,2はともに正解が提示文脈に含まれない
(表層的に解答不能質問と判定するのは困難)

遠距離教師あり学習や正解を含まない文脈を使った自動質問生成は、解答不能性が検知されやすいため crowdsourcingによって人手で作成

- SQuADの各設問の元となるWikipedia記事の各パラグラフに対して最大5件の解答不能質問
- 質問作成数の少ない(25件以下)workerの入力は破棄
- 今回改めてSQuAD/SQuADRUnの人の正解率を精査

System	SQuAD test EM	SQuAD test F1	SQuADRUn dev EM	SQuADRUn dev F1	SQuADRUn test EM	SQuADRUn test F1
BNA BiDAF	68.0	77.3	59.8	62.6	59.2	62.1
DocQA	72.1	81.0	61.9	64.8	59.3	62.3
DocQA + ELMo	**78.6**	**85.8**	**65.1**	**67.6**	**63.4**	**66.3**
Human	82.3	91.2	86.3	89.0	86.9	89.5
Human–Machine Gap	3.7	5.4	**21.2**	**21.4**	**23.5**	**23.2**

SQuADRUnのSoTA手法と人間の性能比較(約23ポイントの差)
従来の正解以外に「解答不能」を推定可能

- 抽出型のQAデータセットでは正解が提示文脈に含まれる設問が大多数(あるいは自動生成された容易に検出可能な解答不能質問)
- 既存のSQuADデータセット(10万質問)に、解答可能な質問と似た表現の解答不能質問を約5万件追加したデータセットを作成(SQuADRUnあるいはSQuAD 2.0として公開)
  - SQuADと同じ文脈に対し解答不能な**関連**質問を追加
  - 正解と**同じ型**の、ありそうな解答(plausible answer)が文脈に含まれる(人やシステムの誤答の半数に合致するほど巧妙)
- 読解(reading comprehension)の新たなタスク評価データとしての活用を期待
  - **敵対的な設問に対する質問応答手法**

Reasoning	Description	Example	Percentage
Negation	Negation word inserted or removed.	Sentence: "*Several hospital pharmacies have decided to outsource high risk preparations ...*" Question: "*What types of pharmacy functions have **never** been outsourced?*"	9%
Antonym	Antonym used.	S: "*the extinction of the dinosaurs... allowed the tropical rainforest to spread out across the continent.*" Q: "*The extinction of what led to the **decline** of rainforests?*"	20%
Entity Swap	Entity, number, or date replaced with other entity, number, or date.	S: "*These values are much greater than the 9–88 cm as projected ... in its Third Assessment Report.*" Q: "*What was the projection of sea level increases in the **fourth assessment report**?*"	21%
Mutual Exclusion	Word or phrase is mutually exclusive with something for which an answer is present.	S: "*BSkyB... waiv[ed] the charge for subscribers whose package included two or more premium channels.*" Q: "*What service did BSkyB **give away for free unconditionally**?*"	15%
Impossible Condition	Asks for condition that is not satisfied by anything in the paragraph.	S: "*Union forces left Jacksonville and confronted a Confederate Army at the Battle of Olustee... Union forces then retreated to Jacksonville and held the city for the remainder of the war.*" Q: "*After what battle did Union forces leave Jacksonville **for good**?*"	4%
Other Neutral	Other cases where the paragraph does not imply any answer.	S: "*Schuenemann et al. concluded in 2011 that the Black Death... was caused by a variant of Y. pestis...*" Q: "*Who **discovered** Y. pestis?*"	24%
Answerable	Question is answerable (i.e. dataset noise).		7%

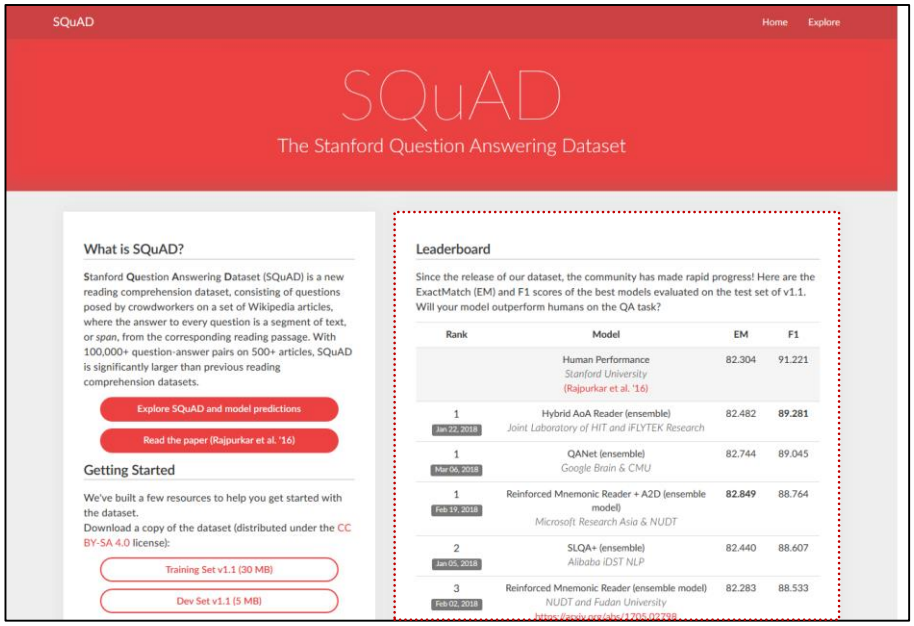
正解を含まない文脈と質問の分類
(SQuADRUn中のランダム100質問)

名古屋大学 NAGOYA UNIVERSITY

スタンフォード大学の質問応答データセットSQuAD

スタンフォード大学の質問応答データセットSQuAD

https://rajpurkar.github.io/SQuAD-explorer/



Wikipediaの**500**以上の項目について、**10万件**以上の質問とその解答をデータセットとしたもの

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, grau-pel and hail...

What causes precipitation to fall? gravity

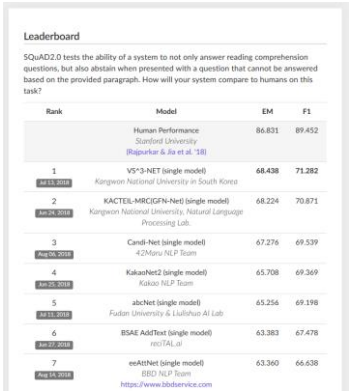
解答は与えられたパッセージ中のspan (i, j)になる

SQuAD1.1 Leaderboard

Since the release of SQuAD1.0, the community has made rapid progress, with the best models now rivaling human performance on the task. Here are the ExactMatch (EM) and F1 scores evaluated on the test set of v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	**84.454**	**90.490**
2 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
3 Jun 20, 2018	MARS (ensemble) *YUANFUDAO research NLP*	83.982	89.796
4 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737
5 Jun 21, 2018	MARS (single model) *YUANFUDAO research NLP*	83.122	89.224

EM(完全一致)とF1(適合率と再現率の調和平均)で評価。上位システムは頻繁に更新される。State-of-the-ArtはEMで人の成績を上回る。



SQuAD 2.0の評価も開始

名古屋大学 NAGOYA UNIVERSITY

質問応答評価タスクと主要なデータセット

1999年: TREC8でQAタスクを設定。その後2007年まで継続し、Yahoo! Answerの質問を使ったLiveQAタスクに移行して現在に至る

What date in 1989 did East Germany open the Berlin Wall? (質問)
LA012890-0072 (正解を支持する情報源)
LA122489-0101 (正解を支持する情報源)
Nov 9 (正解)

2001年: NTCIR-3でQAタスクを設定。その後現在までQA関連タスクを継続
2001-2008年:米国高等研究開発局(ARDA) のAQUAINTプログラムで質問応答技術の研究を推進
2003-2009年: CLEFでQAタスクを設定
2009-2014年: DARPAがMachine Reading Program(読解)を開始
2011年: IBM WatsonがJeopardy!に登場
2014-2017年: DARPAがBig Mechanism Program(読解)を開始。RAS関連の信号伝達パスウェイ解明
2017年: NIPS2017でHuman–Computer Question Answering (HCQA) Competition開催

- Microsoft MCTest (2013) https://github.com/mcobzarenco/mctest
  660話の子供向けの平易な文による架空のstoryに各4問の選択(複数可)質問 Self-contained
- Microsoft MARCO (2016) http://www.msmarco.org/dataset.aspx Bing質問と解答の対
- DeepMind (2015) https://github.com/deepmind/rc-data　固有表現を個別にID化したものを提供
  Daily Mail (ニュース記事と穴埋め(cloze)問題) 196,961記事 879,450質問
  CNN (ニュース記事と穴埋め問題) 90,266記事 380,298質問
- Facebook bAbI (2015) https://github.com/facebook/bAbI-tasks
  20種類のタスクを設定、限定語彙＋自動質問生成により各1000(学習用)+1000(評価用)質問を提供
  SimpleQuestions (Freebaseの構造化データベースを根拠とする質問) 108,442質問
  他にChildren's Book Test, WikiMoviesなどのデータセットが公開されている
- Stanford Univ. SQuAD (2016 前述)
- 他にNewsQA (CNNのspan推定版)、QALD(Question-Answering for Linked Data)やAllen AI Question Answeringなど

State-of-the-Art手法の精度改善が飽和ぎみ、敵対的な設問(Jia and Liang 2017)での精度劣化への取組み