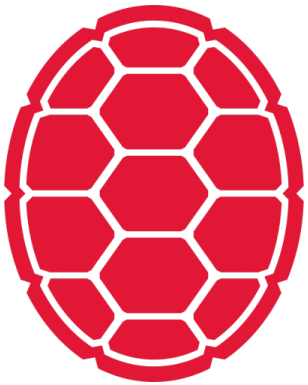


Question Answering

Natural Language Processing: Jordan Boyd-Graber

University of Maryland

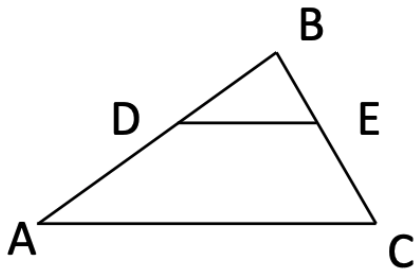
MACHINE READING TASKS

Adapted from slides by Phil Blunsom

What we're not talking about

Text Input

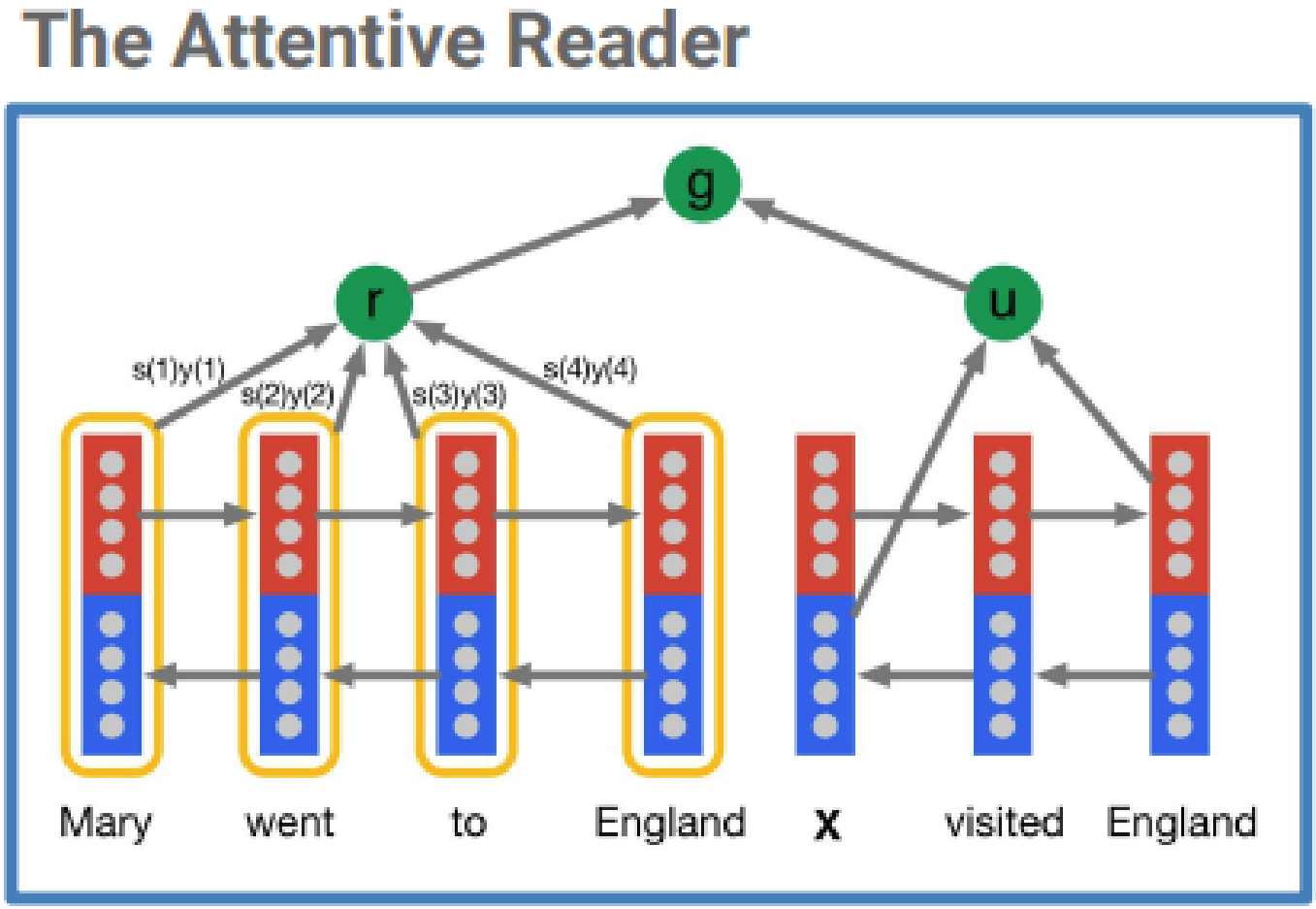
In triangle ABC, line DE is parallel with line AC, DB equals 4, AD is 8, and DE is 5. Find AC.
(a) 9 (b) 10 (c) 12.5 (d) 15 (e) 17



Logical form

$IsTriangle(ABC) \wedge Parallel(AC, DE) \wedge Equals(LengthOf(DB), 4) \wedge Equals(LengthOf(AD), 8) \wedge Equals(LengthOf(DE), 5) \wedge Find(LengthOf(AC))$

Machine Reading Framework



CNN/Daily Mail Datasets

[1]Hermann et al. Teaching machines to read and comprehend. NIPS 2015

CNN/Daily Mail Datasets

lexicalised ...

(CNN) New Zealand are on course for a first ever World Cup title after a thrilling semifinal victory over South Africa, secured off the penultimate ball of the match.

Chasing an adjusted target of 298 in just 43 overs after a rain interrupted the match at Eden Park, Grant Elliott hit a six right at the death to confirm victory and send the Auckland crowd into raptures. It is the first time they have ever reached a world cup final.

Question:
_____ reach cricket Word Cup final?

Answer:
New Zealand

CNN/Daily Mail Datasets

Good

we aimed to factor out world knowledge through entity anonymisation so models could not rely on correlations rather than understanding.

Bad

The generation process and entity anonymisation reduced the task to multiple choice and introduced additional noise.

CNN/Daily Mail Datasets

Good

posing reading comprehension as a large scale conditional modelling task made it accessible to machine learning researchers, generating a great deal of subsequent research.

Bad

while this approach is reasonable for building applications, it is entirely the wrong way to develop and evaluate natural language understanding.

Narrative QA

Narrative QA: examples

Question: How is Oscar related to Dana?
Answer: He is 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.

Narrative QA

Good

A challenging evaluation that tests a range of language understanding, particularly temporal aspects of narrative, and also scalability as current models cannot represent and reason over full narratives

Bad

Performing well on this task is clearly well beyond current models, both representationally and computationally. As such it will be hard for researchers to hill climb on this evaluation.

Narrative QA

Good

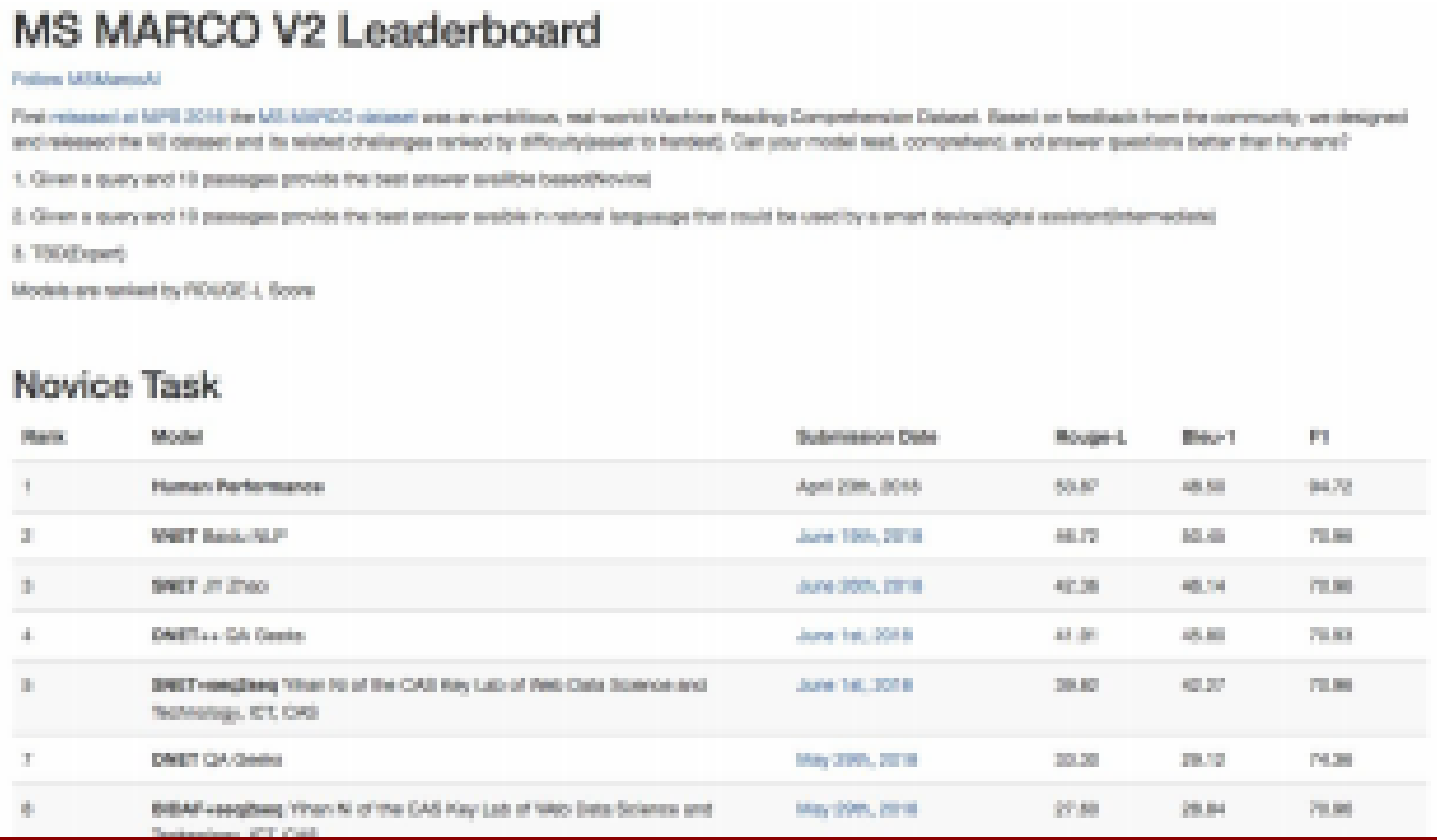
The relatively small number of narratives for training models forces researchers to approach this task from a transfer learning perspective.

Bad

The relatively small number of narratives means that this dataset is not of immediate use for those wanting to build supervised models for applications.

MS Marco

Questions are mined from a search engine and matched with candidate answer passages using IR techniques.



MS MARCO V2 Leaderboard

Follow MSMarcoAI

First released at NIPS 2016 the MS MSPCO dataset was an ambitious, real-world Machine Reading Comprehension Dataset. Based on feedback from the community, we designed and released the V2 dataset and its related challenges ranked by difficulty(easiest to hardest). Can your model read, comprehend, and answer questions better than humans?

1. Given a query and 10 passages provide the best answer possible based(Novice)

2. Given a query and 10 passages provide the best answer possible in natural language that could be used by a smart device/digital assistant(Intermediate)

3. TBD(Expert)

Models are ranked by ROUGE-L Score

Novice Task

Rank	Model	Submission Date	Rouge-L	Bleu-1	F1
1	Human Performance	April 28th, 2016	53.87	46.98	84.72
2	S-NET Baidu NLP	June 19th, 2018	68.79	62.49	75.86
3	S-NET JY Zhao	June 26th, 2018	42.28	46.14	75.86
4	DNET++ QA Geeks	June 1st, 2018	41.91	45.89	75.83
5	S-NET+seq2seq Yihan Ni of the ICAS Key Lab of Web Data Science and Technology, ICT, CAS	June 1st, 2018	39.82	42.37	75.86
7	DNET QA Geeks	May 29th, 2018	33.28	28.12	74.39
8	BiDAF+seq2seq Yihan Ni of the CAS Key Lab of Web Data Science and Technology, ICT, CAS	May 29th, 2018	27.69	28.84	75.86

MS Marco

Good

The reliance on real queries creates a much more useful resource for those interested in applications.

Bad

People rarely ask interesting questions of search engines, and the use of IR techniques to collect candidate passages limits the usefulness of this dataset for evaluating language understanding.

MS Marco

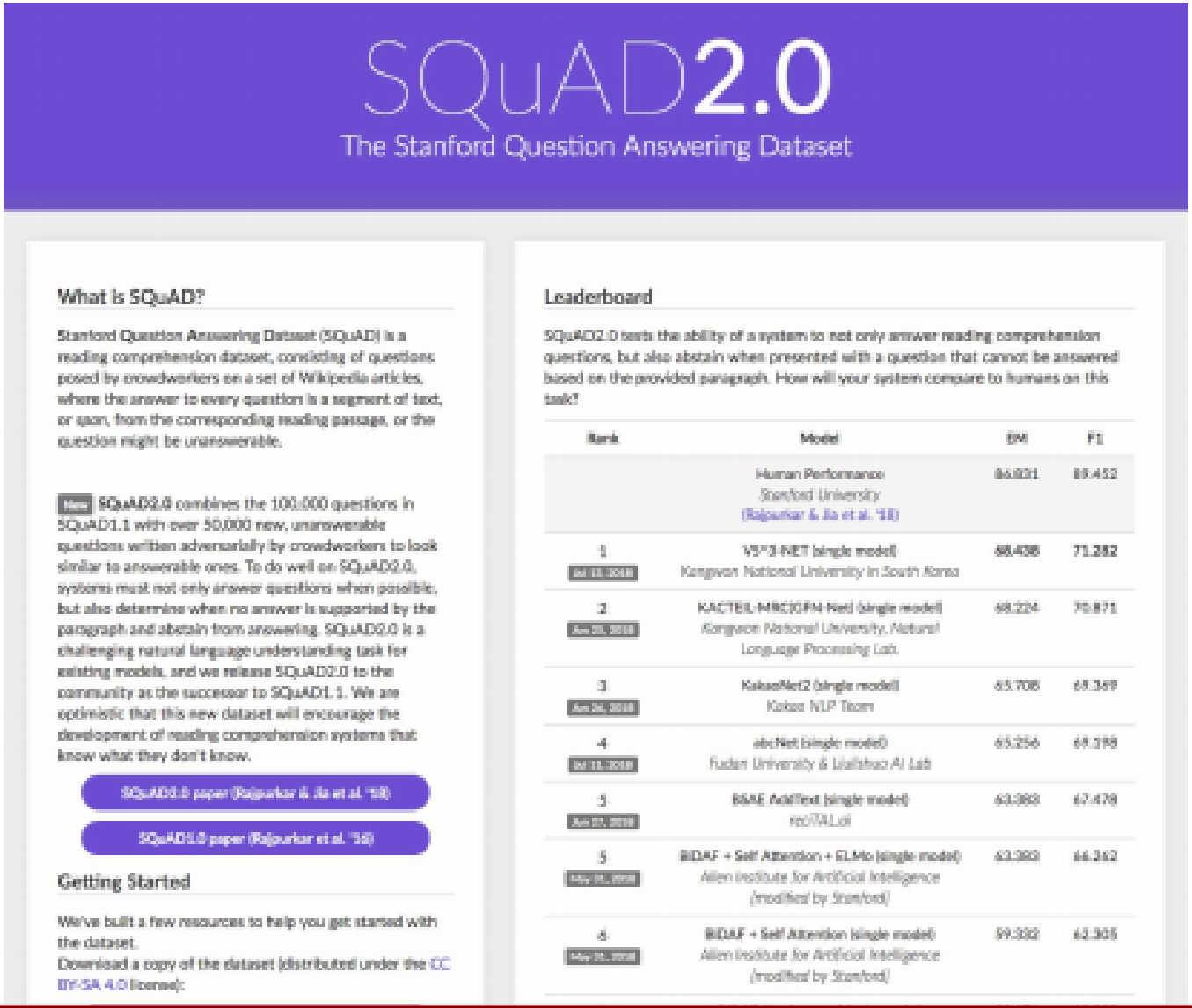
Good

Unrestricted answers allow a greater range of questions.

Bad

How to evaluate freeform answers is an unsolved problem. Bleu is not the answer!

SQuAD



SQuAD2.0

The Stanford Question Answering Dataset

What is SQuAD?

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.

New SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 new, unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering. SQuAD2.0 is a challenging natural language understanding task for existing models, and we release SQuAD2.0 to the community as the successor to SQuAD1.1. We are optimistic that this new dataset will encourage the development of reading comprehension systems that know what they don't know.

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

SQuAD1.0 paper (Rajpurkar et al. '16)

Getting Started

We've built a few resources to help you get started with the dataset.

Download a copy of the dataset (distributed under the CC BY-SA 4.0 license):

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jul 11, 2018	V5*3-NET (single model) Kangwon National University in South Korea	68.408	71.282
2 Jun 03, 2018	KACTEL-MRC(SFN-Net) (single model) *Kangwon National University, Natural Language Processing Lab.*	68.224	70.871
3 Jun 26, 2018	KakaoNet2 (single model) Kakao NLP Team	65.708	68.369
4 Jul 11, 2018	ahcNet (single model) *Fudan University & Liulishuo AI Lab*	65.256	68.198
5 Jun 27, 2018	BiAE AddText (single model) rpc/TAL.ai	63.989	67.478
5 May 21, 2018	BiDAF + Self Attention + GLMo (single model) *Allen Institute for Artificial Intelligence (modified by Stanford)*	63.960	66.262
6 May 21, 2018	BiDAF + Self Attention (single model) *Allen Institute for Artificial Intelligence (modified by Stanford)*	59.930	62.905

SQuAD

In the 1960s, a series of discoveries, the most important of which was seafloor spreading, showed that the Earth's lithosphere, which includes the crust and rigid uppermost portion of the upper mantle, is separated into a number of tectonic plates that move across the plastically deforming, solid, upper mantle, which is called the asthenosphere. There is an intimate coupling between the movement of the plates on the surface and the convection of...

Question:
Which parts of the Earth are included in the lithosphere?

SQuAD

Good

Very scalable annotation process that can cheaply generate large numbers of questions per article.

Bad

Annotating questions directly from the context passages strongly skews the data distribution. The task then becomes reverse engineering the annotators, rather than language understanding.

SQuAD

Good

The online leaderboard allows easy benchmarking of systems and motivates competition.

Bad

Answers as spans reduces the task to multiple choice, and doesn't allow questions with answers latent in the text.

SQuAD

Good

Human upperbound sets reasonable goal.

Bad

Allows mischaracterization of what it means to “read”.

Quiz Bowl

Quiz Bowl

Good

Free data from experts

Bad

Sometimes can be trivially solved with pattern matching

Quiz Bowl

Good	Bad
Based on already known knowledge	Only tied to readable data *post hoc*

Quiz Bowl

Good

Human comparison makes sense

Bad

More cumbersome computer evaluation

Where next?

- Robustness
- Logical reasoning
- Domain adaptation