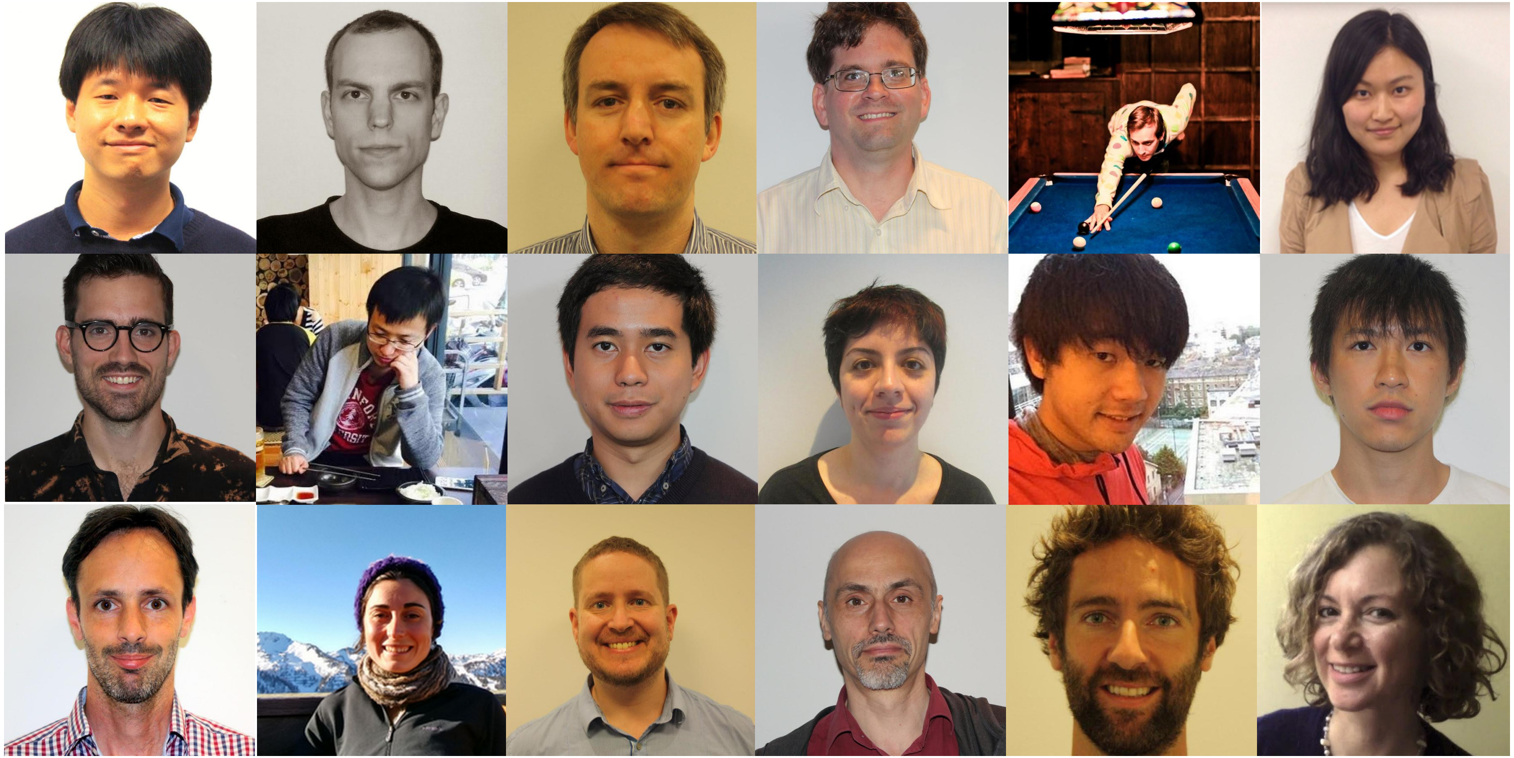
Data Driven Reading Comprehension

Phil Blunsom

In collaboration with Karl Moritz Hermann, Tomáš Kočiský, Ed Grefenstette and the DeepMind Natural Language Group

DeepMind

The DeepMind Language Group

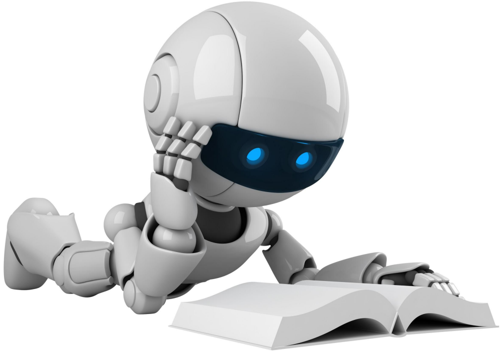


Reading Comprehension

We aim to build models that can read a text, represent the information contained within it, and answer questions based on this representation

There are two broad motivations for doing this,

1. To build QA applications or products,
2. To evaluate language understanding algorithms.



MC Test

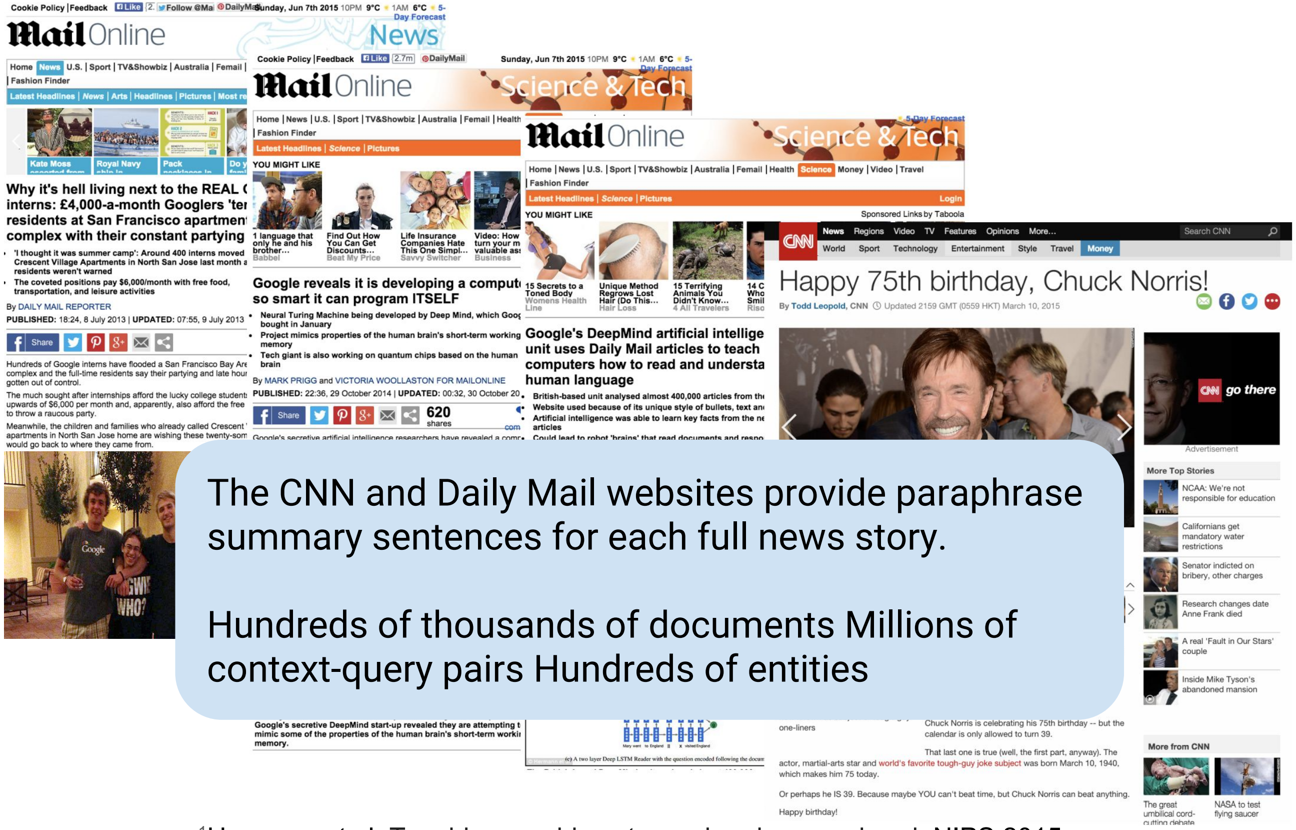
James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back. One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home. …

Where did James go after he went to the grocery store?

1. his deck
2. his freezer
3. a fast food restaurant
4. his room

[1] Richardson. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. EMNLP 2013

The CNN and Daily Mail datasets: aims



Cookie Policy | Feedback Like 2. Follow @Mai DailyMa Sunday, Jun 7th 2015 10PM 9°C 1AM 6°C 5-Day Forecast

Mail Online News

Home | News | U.S. | Sport | TV&Showbiz | Australia | Femail | Fashion Finder

Latest Headlines | News | Arts | Headlines | Pictures | Most re

Kate Moss | Royal Navy | Pack | Do y

Why it's hell living next to the REAL (interns: £4,000-a-month Googlers 'ter residents at San Francisco apartmen complex with their constant partying

- 'I thought it was summer camp': Around 400 interns moved Crescent Village Apartments in North San Jose last month a residents weren't warned
- The coveted positions pay $6,000/month with free food, transportation, and leisure activities

By DAILY MAIL REPORTER

PUBLISHED: 18:24, 8 July 2013 | UPDATED: 07:55, 9 July 2013

Share

Hundreds of Google interns have flooded a San Francisco Bay Are complex and the full-time residents say their partying and late hour gotten out of control.

The much sought after internships afford the lucky college student upwards of $6,000 per month and, apparently, also afford the free to throw a raucous party.

Meanwhile, the children and families who already called Crescent ' apartments in North San Jose home are wishing these twenty-som would go back to where they came from.

Cookie Policy | Feedback Like 2.7m DailyMail Sunday, Jun 7th 2015 10PM 9°C 1AM 6°C 5-Day Forecast

Mail Online Science & Tech

Home | News | U.S. | Sport | TV&Showbiz | Australia | Femail | Health | Fashion Finder

Latest Headlines | Science | Pictures

YOU MIGHT LIKE

1 language that only he and his brother... Babbel | Find Out How You Can Get Discounts... Beat My Price | Life Insurance Companies Hate This One Simpl... Savvy Switcher | Video: How turn your m valuable ass Business

Google reveals it is developing a comput so smart it can program ITSELF

- Neural Turing Machine being developed by Deep Mind, which Goog bought in January
- Project mimics properties of the human brain's short-term working memory
- Tech giant is also working on quantum chips based on the human brain

By MARK PRIGG and VICTORIA WOOLLASTON FOR MAILONLINE

PUBLISHED: 22:36, 29 October 2014 | UPDATED: 00:32, 30 October 20

Share 620 shares

Google's secretive artificial intelligence researchers have revealed a comp

Google's secretive DeepMind start-up revealed they are attempting t mimic some of the properties of the human brain's short-term worki memory.

Mail Online Science & Tech 5-Day Forecast

Home | News | U.S. | Sport | TV&Showbiz | Australia | Femail | Health | Science | Money | Video | Travel | Fashion Finder

Latest Headlines | Science | Pictures Login

YOU MIGHT LIKE Sponsored Links by Taboola

15 Secrets to a Toned Body Womens Health Line | Unique Method Regrows Lost Hair (Do This... Hair Loss | 15 Terrifying Animals You Didn't Know... 4 All Travelers | 14 C Who Smil Riso

Google's DeepMind artificial intellige unit uses Daily Mail articles to teach computers how to read and understa human language

- British-based unit analysed almost 400,000 articles from the
- Website used because of its unique style of bullets, text an
- Artificial intelligence was able to learn key facts from the ne articles
- Could lead to robot 'brains' that read documents and respo

(c) A two layer Deep LSTM Reader with the question encoded following the docum

CNN News Regions Video TV Features Opinions More... Search CNN

World Sport Technology Entertainment Style Travel Money

Happy 75th birthday, Chuck Norris!

By Todd Leopold, CNN Updated 2159 GMT (0559 HKT) March 10, 2015

CNN go there

Advertisement

More Top Stories

- NCAA: We're not responsible for education
- Californians get mandatory water restrictions
- Senator indicted on bribery, other charges
- Research changes date Anne Frank died
- A real 'Fault in Our Stars' couple
- Inside Mike Tyson's abandoned mansion

one-liners

Chuck Norris is celebrating his 75th birthday -- but the calendar is only allowed to turn 39.

That last one is true (well, the first part, anyway). The actor, martial-arts star and world's favorite tough-guy joke subject was born March 10, 1940, which makes him 75 today.

Or perhaps he IS 39. Because maybe YOU can't beat time, but Chuck Norris can beat anything.

Happy birthday!

More from CNN

The great umbilical cord-cutting debate | NASA to test flying saucer

The CNN and Daily Mail websites provide paraphrase summary sentences for each full news story.

Hundreds of thousands of documents Millions of context-query pairs Hundreds of entities

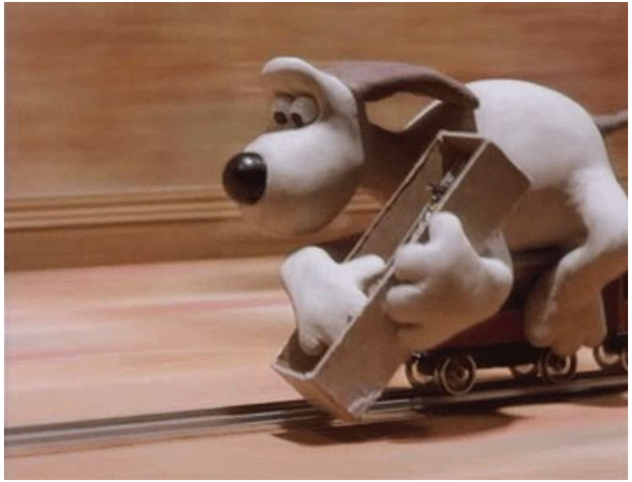
[1] Hermann et al. Teaching machines to read and comprehend. NIPS 2015

The CNN and Daily Mail datasets: large scale RC

MC Test

500 stories, 2k questions.

CNN and Daily Mail Corpora



~300k stories and >1M questions.

The CNN and Daily Mail datasets: the data

The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." …

Cloze-style question:

Query: Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.

Answer: Oisin Tymon

The CNN and Daily Mail datasets: the data

From the Daily Mail:

- The hi-tech bra that helps you beat breast **X**
- Could Saccharin help beat **X** ?
- Can fish oils help fight prostate **X** ?

Any n-gram language model train on the Daily Mail would correctly predict (**X** = cancer)

The CNN and Daily Mail datasets: anonymisation

We aimed to design the task to avoid shortcuts such as QA by language modelling or correlation:

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

... delexicalised

(*ent23*) *ent7* are on course for a first ever *ent15* title after a thrilling semifinal victory over *ent34*, 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 *ent12*, *ent17* hit a six right at the death to confirm victory and send the *ent83* crowd into raptures. It is the first time they have ever reached a *ent15* final.

Question:
_____ reach *ent3 ent15* final?

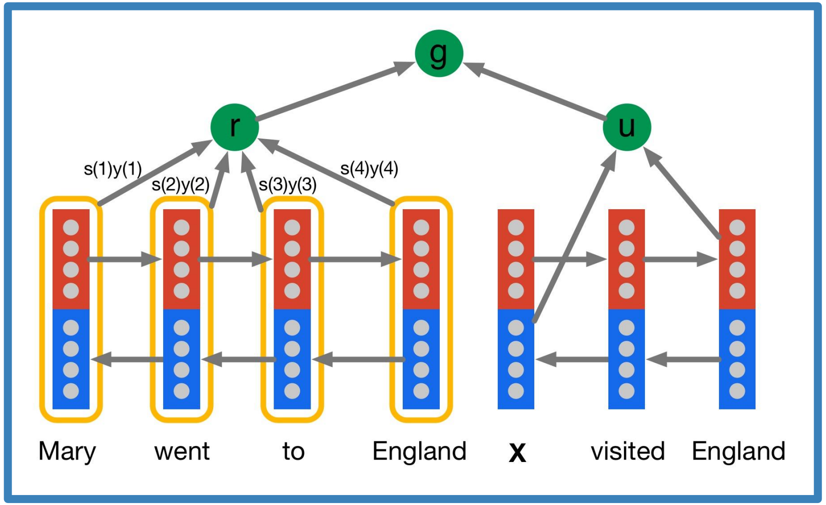
Answer:
ent7

The CNN and Daily Mail datasets: models

We proposed a simple attention based approach:

- Separate encodings for query and context tokens
- Attend over context token encodings
- Predict based on joint weighted attention and query representation

The Attentive Reader



The CNN and Daily Mail datasets: the good and bad

Good: we recognised that there must be a level of indirection between annotators producing questions and the text from which the questions are answered.

Bad: many of the automatically generated questions are of poor quality or ambiguous.[1]

[1] Chen et al. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. ACL 2016

The CNN and Daily Mail datasets: the good and bad

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.

The CNN and Daily Mail datasets: the good and bad

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.

Desiderata for Reading Comprehension Data sets

Applications

If our aim is to build a product or application, we must acquire data as close to the real use case as possible, i.e. representative questions and document contexts.

If we artificially generate data we risk introducing spurious correlations, which overparameterised neural networks are excellent at exploiting.

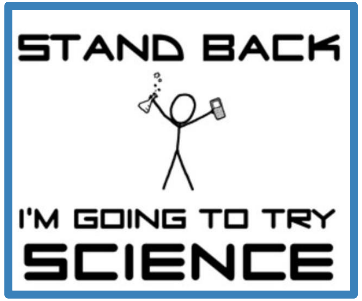
Desiderata for Reading Comprehension Data sets

Language Understanding

Any data annotation process will introduce spurious artifacts into the data.

Performance on a language understanding evaluation can thus be factored into two components, 1) that which measures true understanding, 2) and that which captures overfitting to the artifacts.

If our aim is to evaluate language understanding systems we must not train on data collected with the same annotation process as our evaluation set.



Stanford Question Answering Dataset (SQuAD)

Question answer pairs crowdsourced on ~500 Wikipedia articles.

Answers are spans in the context passage.

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...

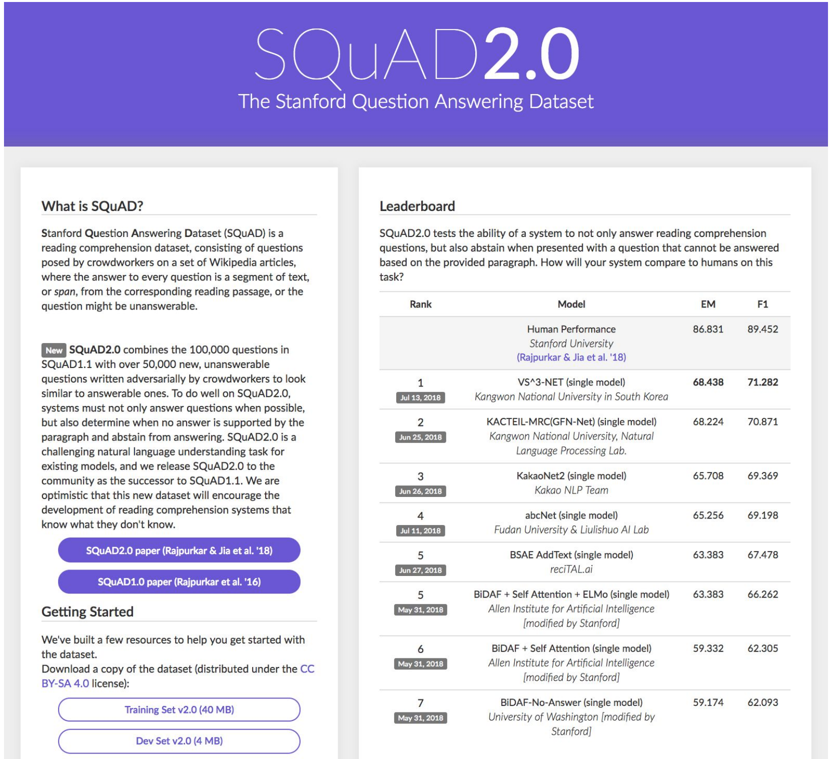
Question:
Which parts of the Earth are included in the lithosphere?

[1] Rajpurkar et al. SQuAD: 100,000+ Questions for Machine Comprehension of Text. EMNLP 2016

Stanford Question Answering Dataset (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.



SQuAD2.0

The Stanford Question Answering Dataset

What is SQuAD?

Stanford **Q**uestion **A**nswering **D**ataset (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):

Training Set v2.0 (40 MB)

Dev Set v2.0 (4 MB)

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?

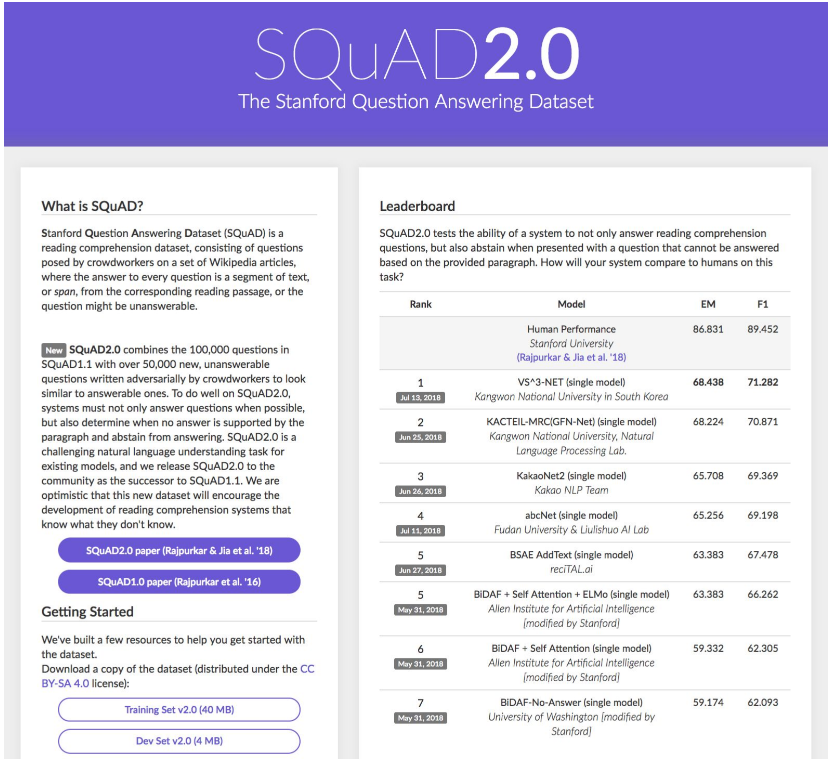
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jul 13, 2018	VS^3-NET (single model) *Kangwon National University in South Korea*	**68.438**	**71.282**
2 Jun 25, 2018	KACTEIL-MRC(GFN-Net) (single model) *Kangwon National University, Natural Language Processing Lab.*	68.224	70.871
3 Jun 26, 2018	KakaoNet2 (single model) *Kakao NLP Team*	65.708	69.369
4 Jul 11, 2018	abcNet (single model) *Fudan University & Liulishuo AI Lab*	65.256	69.198
5 Jun 27, 2018	BSAE AddText (single model) *reciTAL.ai*	63.383	67.478
5 May 31, 2018	BiDAF + Self Attention + ELMo (single model) *Allen Institute for Artificial Intelligence [modified by Stanford]*	63.383	66.262
6 May 31, 2018	BiDAF + Self Attention (single model) *Allen Institute for Artificial Intelligence [modified by Stanford]*	59.332	62.305
7 May 31, 2018	BiDAF-No-Answer (single model) *University of Washington [modified by Stanford]*	59.174	62.093

[1] https://rajpurkar.github.io/SQuAD-explorer/

Stanford Question Answering Dataset (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.



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?

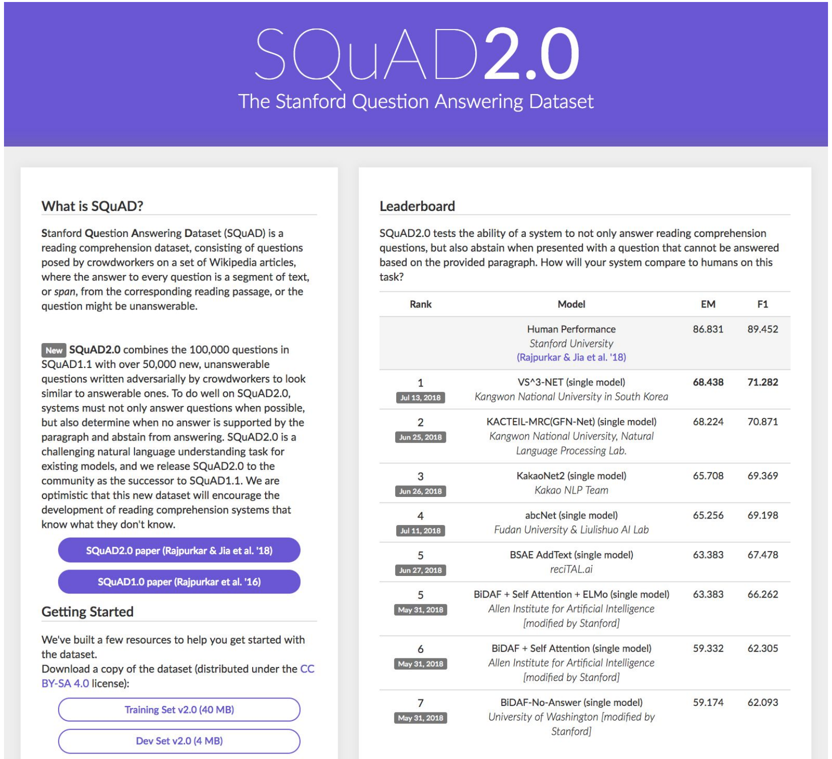
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jul 13, 2018	VS^3-NET (single model) *Kangwon National University in South Korea*	**68.438**	**71.282**
2 Jun 25, 2018	KACTEIL-MRC(GFN-Net) (single model) *Kangwon National University, Natural Language Processing Lab.*	68.224	70.871
3 Jun 26, 2018	KakaoNet2 (single model) *Kakao NLP Team*	65.708	69.369
4 Jul 11, 2018	abcNet (single model) *Fudan University & Liulishuo AI Lab*	65.256	69.198
5 Jun 27, 2018	BSAE AddText (single model) *reciTAL.ai*	63.383	67.478
5 May 31, 2018	BiDAF + Self Attention + ELMo (single model) *Allen Institute for Artificial Intelligence [modified by Stanford]*	63.383	66.262
6 May 31, 2018	BiDAF + Self Attention (single model) *Allen Institute for Artificial Intelligence [modified by Stanford]*	59.332	62.305
7 May 31, 2018	BiDAF-No-Answer (single model) *University of Washington [modified by Stanford]*	59.174	62.093

[1]https://rajpurkar.github.io/SQuAD-explorer/

Stanford Question Answering Dataset (SQuAD)

SQUAD provides a great resource for experimenting with machine learning models.

However, just like the CNN/DailyMail corpus, it does not satisfy the requirements for building applications, nor for evaluating language understanding systems.



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?

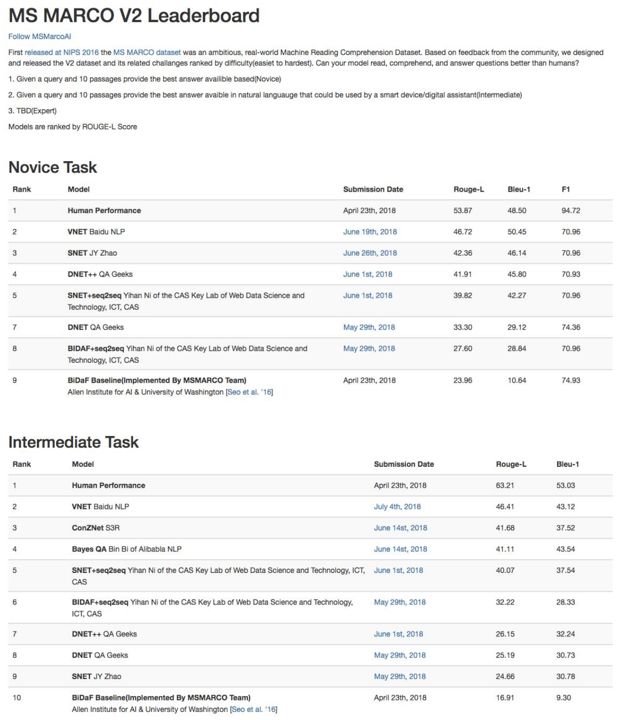
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jul 13, 2018	VS^3-NET (single model) *Kangwon National University in South Korea*	**68.438**	**71.282**
2 Jun 25, 2018	KACTEIL-MRC(GFN-Net) (single model) *Kangwon National University, Natural Language Processing Lab.*	68.224	70.871
3 Jun 26, 2018	KakaoNet2 (single model) *Kakao NLP Team*	65.708	69.369
4 Jul 11, 2018	abcNet (single model) *Fudan University & Liulishuo AI Lab*	65.256	69.198
5 Jun 27, 2018	BSAE AddText (single model) *reciTAL.ai*	63.383	67.478
5 May 31, 2018	BiDAF + Self Attention + ELMo (single model) *Allen Institute for Artificial Intelligence [modified by Stanford]*	63.383	66.262
6 May 31, 2018	BiDAF + Self Attention (single model) *Allen Institute for Artificial Intelligence [modified by Stanford]*	59.332	62.305
7 May 31, 2018	BiDAF-No-Answer (single model) *University of Washington [modified by Stanford]*	59.174	62.093

[1] https://rajpurkar.github.io/SQuAD-explorer/

MS Marco

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

Answers are not restricted to be subspans of the documents, and some questions are not answerable from the context.



MS MARCO V2 Leaderboard

Follow MSMarcoAI

First released at NIPS 2016 the MS MARCO 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 challanges ranked by difficulty(easiet to hardest). Can your model read, comprehend, and answer questions better than humans?

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

2. Given a query and 10 passages provide the best answer avaible in natural languauge 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 23th, 2018	53.87	48.50	94.72
2	**VNET** Baidu NLP	June 19th, 2018	46.72	50.45	70.96
3	**SNET** JY Zhao	June 26th, 2018	42.36	46.14	70.96
4	**DNET++** QA Geeks	June 1st, 2018	41.91	45.80	70.93
5	**SNET+seq2seq** Yihan Ni of the CAS Key Lab of Web Data Science and Technology, ICT, CAS	June 1st, 2018	39.82	42.27	70.96
7	**DNET** QA Geeks	May 29th, 2018	33.30	29.12	74.36
8	**BIDAF+seq2seq** Yihan Ni of the CAS Key Lab of Web Data Science and Technology, ICT, CAS	May 29th, 2018	27.60	28.84	70.96
9	**BiDaF Baseline(Implemented By MSMARCO Team)** Allen Institute for AI & University of Washington [Seo et al. '16]	April 23th, 2018	23.96	10.64	74.93

Intermediate Task

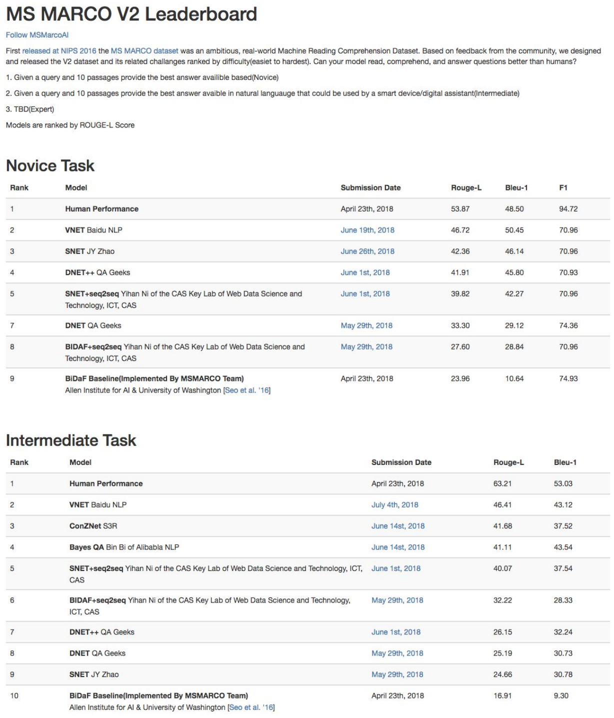
Rank	Model	Submission Date	Rouge-L	Bleu-1
1	**Human Performance**	April 23th, 2018	63.21	53.03
2	**VNET** Baidu NLP	July 4th, 2018	46.41	43.12
3	**ConZNet** S3R	June 14st, 2018	41.68	37.52
4	**Bayes QA** Bin Bi of Alibabla NLP	June 14st, 2018	41.11	43.54
5	**SNET+seq2seq** Yihan Ni of the CAS Key Lab of Web Data Science and Technology, ICT, CAS	June 1st, 2018	40.07	37.54
6	**BIDAF+seq2seq** Yihan Ni of the CAS Key Lab of Web Data Science and Technology, ICT, CAS	May 29th, 2018	32.22	28.33
7	**DNET++** QA Geeks	June 1st, 2018	26.15	32.24
8	**DNET** QA Geeks	May 29th, 2018	25.19	30.73
9	**SNET** JY Zhao	May 29th, 2018	24.66	30.78
10	**BiDaF Baseline(Implemented By MSMARCO Team)** Allen Institute for AI & University of Washington [Seo et al. '16]	April 23th, 2018	16.91	9.30

[1] Nguyen et al. MS MARCO: A Human Generated Mahine Reading Comprehension Dataset. NIPS 2016

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 V2 Leaderboard

Follow MSMarcoAI

First released at NIPS 2016 the MS MARCO 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 challanges ranked by difficulty(easiet to hardest). Can your model read, comprehend, and answer questions better than humans?

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

2. Given a query and 10 passages provide the best answer avaible in natural languauge 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 23th, 2018	53.87	48.50	94.72
2	**VNET** Baidu NLP	June 19th, 2018	46.72	50.45	70.96
3	**SNET** JY Zhao	June 26th, 2018	42.36	46.14	70.96
4	**DNET++** QA Geeks	June 1st, 2018	41.91	45.80	70.93
5	**SNET+seq2seq** Yihan Ni of the CAS Key Lab of Web Data Science and Technology, ICT, CAS	June 1st, 2018	39.82	42.27	70.96
7	**DNET** QA Geeks	May 29th, 2018	33.30	29.12	74.36
8	**BIDAF+seq2seq** Yihan Ni of the CAS Key Lab of Web Data Science and Technology, ICT, CAS	May 29th, 2018	27.60	28.84	70.96
9	**BiDaF Baseline(Implemented By MSMARCO Team)** Allen Institute for AI & University of Washington [Seo et al. '16]	April 23th, 2018	23.96	10.64	74.93

Intermediate Task

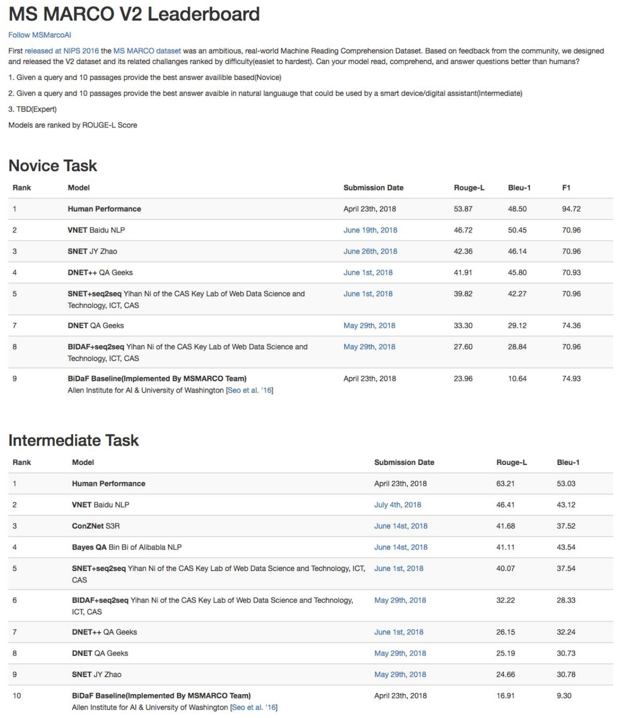
Rank	Model	Submission Date	Rouge-L	Bleu-1
1	**Human Performance**	April 23th, 2018	63.21	53.03
2	**VNET** Baidu NLP	July 4th, 2018	46.41	43.12
3	**ConZNet** S3R	June 14st, 2018	41.68	37.52
4	**Bayes QA** Bin Bi of Alibabla NLP	June 14st, 2018	41.11	43.54
5	**SNET+seq2seq** Yihan Ni of the CAS Key Lab of Web Data Science and Technology, ICT, CAS	June 1st, 2018	40.07	37.54
6	**BIDAF+seq2seq** Yihan Ni of the CAS Key Lab of Web Data Science and Technology, ICT, CAS	May 29th, 2018	32.22	28.33
7	**DNET++** QA Geeks	June 1st, 2018	26.15	32.24
8	**DNET** QA Geeks	May 29th, 2018	25.19	30.73
9	**SNET** JY Zhao	May 29th, 2018	24.66	30.78
10	**BiDaF Baseline(Implemented By MSMARCO Team)** Allen Institute for AI & University of Washington [Seo et al. '16]	April 23th, 2018	16.91	9.30

[1] Nguyen et al. MS MARCO: A Human Generated Mahine Reading Comprehension Dataset. NIPS 2016

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!



MS MARCO V2 Leaderboard

Follow MSMarcoAI

First released at NIPS 2016 the MS MARCO 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 challanges ranked by difficulty(easiet to hardest). Can your model read, comprehend, and answer questions better than humans?

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

2. Given a query and 10 passages provide the best answer avaible in natural languauge 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 23th, 2018	53.87	48.50	94.72
2	**VNET** Baidu NLP	June 19th, 2018	46.72	50.45	70.96
3	**SNET** JY Zhao	June 26th, 2018	42.36	46.14	70.96
4	**DNET++** QA Geeks	June 1st, 2018	41.91	45.80	70.93
5	**SNET+seq2seq** Yihan Ni of the CAS Key Lab of Web Data Science and Technology, ICT, CAS	June 1st, 2018	39.82	42.27	70.96
7	**DNET** QA Geeks	May 29th, 2018	33.30	29.12	74.36
8	**BIDAF+seq2seq** Yihan Ni of the CAS Key Lab of Web Data Science and Technology, ICT, CAS	May 29th, 2018	27.60	28.84	70.96
9	**BiDaF Baseline(Implemented By MSMARCO Team)** Allen Institute for AI & University of Washington [Seo et al. '16]	April 23th, 2018	23.96	10.64	74.93

Intermediate Task

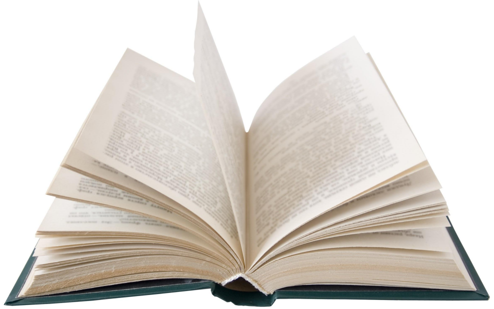
Rank	Model	Submission Date	Rouge-L	Bleu-1
1	**Human Performance**	April 23th, 2018	63.21	53.03
2	**VNET** Baidu NLP	July 4th, 2018	46.41	43.12
3	**ConZNet** S3R	June 14st, 2018	41.68	37.52
4	**Bayes QA** Bin Bi of Alibabla NLP	June 14st, 2018	41.11	43.54
5	**SNET+seq2seq** Yihan Ni of the CAS Key Lab of Web Data Science and Technology, ICT, CAS	June 1st, 2018	40.07	37.54
6	**BIDAF+seq2seq** Yihan Ni of the CAS Key Lab of Web Data Science and Technology, ICT, CAS	May 29th, 2018	32.22	28.33
7	**DNET++** QA Geeks	June 1st, 2018	26.15	32.24
8	**DNET** QA Geeks	May 29th, 2018	25.19	30.73
9	**SNET** JY Zhao	May 29th, 2018	24.66	30.78
10	**BiDaF Baseline(Implemented By MSMARCO Team)** Allen Institute for AI & University of Washington [Seo et al. '16]	April 23th, 2018	16.91	9.30

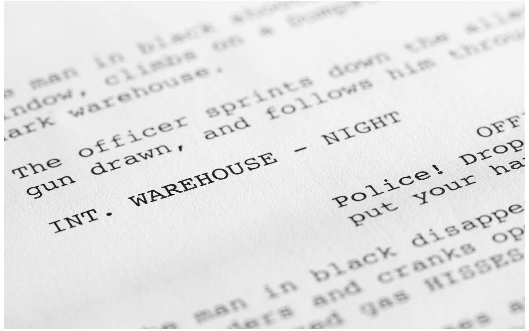
[1] Nguyen et al. MS MARCO: A Human Generated Mahine Reading Comprehension Dataset. NIPS 2016

Narrative QA: aims

Understanding language goes beyond reading and answering literal questions on factual content.

Narratives present many interesting challenges, requiring models to represent and reason over characters and temporal relationships.





e man in black sho
ndow, climbs on a Dumps
ark warehouse.

The officer sprints down the alle
gun drawn, and follows him throu

INT. WAREHOUSE - NIGHT

OFF

Police! Drop
put your ha

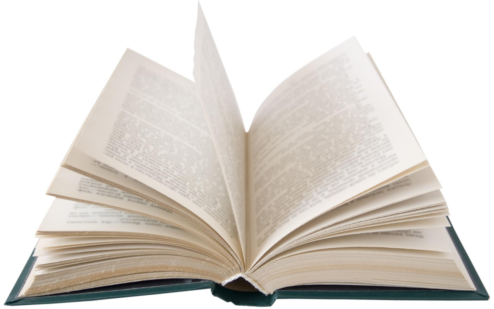
man in black disappe
ders and cranks op
gas HISSES

[1]Kocisky et al. The NarrativeQA Reading Comprehension Challenge. TACL 2018

Narrative QA: construction

Documents are **books** and **movie scripts**

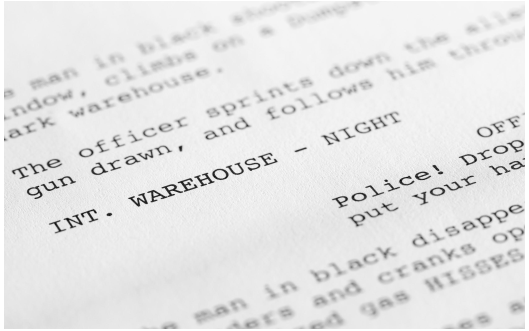
- Complex, long, self-contained narratives
- Contain dialogue



Questions from **abstractive summaries**

- Summary → 30 questions with 2 answers each

Answers are human generated



e man in black slo
ndow, climbs on a Dumpst
ark warehouse.

The officer sprints down the alle
gun drawn, and follows him throu

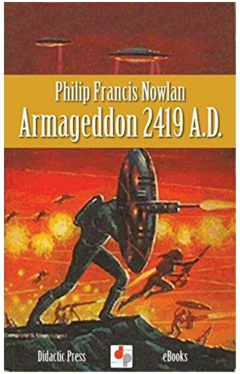
INT. WAREHOUSE - NIGHT

OFF
Police! Drop
put your ha

man in black disappe
ders and cranks op
ed gas HISSES

[1] Kocisky et al. The NarrativeQA Reading Comprehension Challenge. TACL 2018

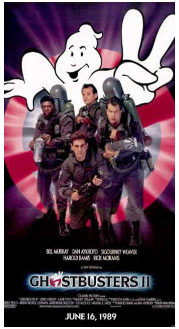
Narrative QA: examples



Question: In what year did Rogers awaken from his deep slumber?
Answer: 2419

Summary snippet: ...Rogers remained in sleep for 492 years. He awakes in 2419 and,...
Story snippet: I should state therefore, that I, Anthony Rogers, am, so far as I know, the only man alive whose normal span of eighty-one years of life has been spread over a period of 573 years. To be precise, I lived the first twenty-nine years of my life between 1898 and 1927; the other fifty-two since 2419. The gap between these two, a period of nearly five hundred years, I spent in a state of suspended animation, free from the ravages of katabolic processes, and without any apparent effect on my physical or mental faculties. When I began my long sleep, man had just begun his real conquest of the air...

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: composition

	Train	Validation	Test
# documents	1,102	115	355
# questions	32,747	3,461	10,557
Avg. # tokens in summaries	659	638	654
Avg. # tokens in scripts	29,934	29,515	29,900
Avg. # tokens in books	95,632	94,506	86,329
Max # tokens	430,061	418,265	404,641

Narrative QA: question categories

Why doesn’t Pozdnyshev run after the violinist? [**Why/reason 9.4%**]

How did Jake survive being shot? [**How/method 8.1%**]

When does Reiko realize the curse is still unbroken? [**Event 4.4%**]

What is Nora’s relationship to Michael? [**Relation 1.3%**]

Person 30.5%, Description 24.5%, Location 9.7%, Entity 4.0%, Object 3.7% Numeric 3.0%, Duration 1.7%

Narrative QA: benchmarks

We evaluated models with Bleu, Rouge, and Mean Reciprocal Rank. None of these are ideal.

Models that have performed well on tasks such as MS Marco also give similar performance when answering questions directly from the summaries.

Such models do not scale to the task of answering questions from full narratives, so we experimented with an initial IR step to retrieve candidate passages.

All the models we tried were unable to answer a significant number of questions posed on the full narratives.

Narrative QA: the good and bad

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: the good and bad

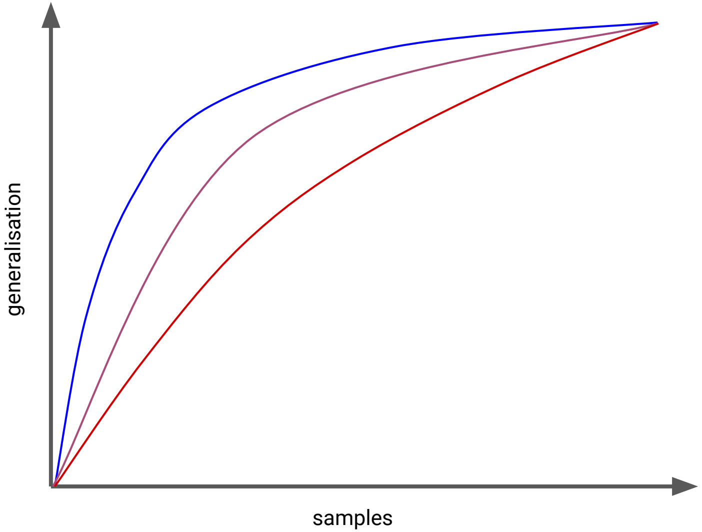
Good: The relatively small number of narratives for training models forces researchers to approach this task from a transfer learning perspective.

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.

Transfer learning and Sample Complexity

We need to move away from training narrow supervised RC language understanding models towards a transfer learning paradigm.

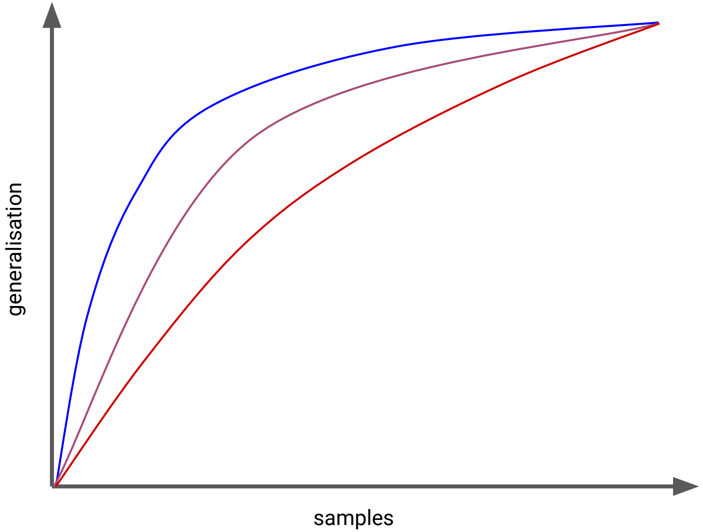
This will require to us to let go of our CL obsession with single point accuracy comparisons and embrace sample complexity as the true goal.



Transfer learning and Sample Complexity

Multitask learning is not the answer.

It is easy to grow neural networks with the number of tasks, thus allowing models to overtrain to the idiosyncrasies of each task.



Summary

The marriage of machine learning models, large data sets, and reading comprehension has produced a vibrant research environment.

We must move beyond this paradigm, embracing a variety of language and focusing on representation and inference over local context and correlation.

Thank You.

Data Driven Reading Comprehension

Phil Blunsom

In collaboration with Karl Moritz Hermann, Tomáš Kočiský, Ed Grefenstette and the DeepMind Natural Language Group

DeepMind