

# Introduction and course overview

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding



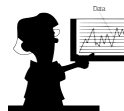
# Welcome



Chris Potts



Sterling Alic



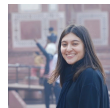
Kakne Asavari



Omar Khattab



David Lim



Dhara Yu

## CS224u: hybrid, asynchronous, hands-on

- Core course content in screencasts on YouTube and linked from the homepage, with accompanying Jupyter notebook for hands-on work.
- A series of special events: conversations with prominent NLU researchers (details later in this lecture). Mostly on Zoom. Attend live or listen later.
- Other class meetings: optional open discussions and/or spaces for you to work, with the teaching team there to help. Open to mixing in-classroom and Zoom formats.
- Office hours offered in person and on Zoom. Details to come.
- Continuous evaluation: three assignments, four online quizzes, and project work.

# Plan for today

1. A golden age for NLU
2. A peek behind the curtain
3. Underlying challenges
4. Assignments, bake-offs, and projects
5. Course mechanics



# A golden age for NLU

1. A golden age for NLU
2. A peek behind the curtain
3. Underlying challenges
4. Assignments, bake-offs, and projects
5. Course mechanics

## A brief history of NLU approaches

- **McCarthy et al. (1955)**: “We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.”
- **1960s**: Pattern-matching with small rule-sets, oriented towards NLU.
- **1970–80s**: Linguistically rich, logic-driven, grounded systems (**LLGs**); restricted applications.
- **Mid-1990s**: Machine learning revolution in NLP leads to a decrease in NLU work.
- **Late 2000s**: **LLGs** re-emerge, now with *learning*.
- **Mid-2010s**: NLU returns to center stage, with deep learning the most prevalent set of techniques. **LLGs** go into decline.
- **2020s**: Grounding is *back*, and rich logical and linguistic structure is *emergent*.

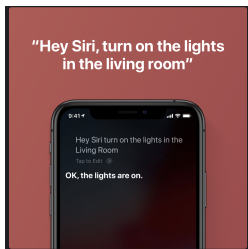
# Watson wins Jeopardy! (2011)



## QANTA beats Ken Jennings (2015)



# Artificial assistants



# The promise of these artificial assistants



**You:** Any good burger joints around here?

**Siri:** I found a number of burger restaurants near you.

**You:** Hmm. How about tacos?

**Apple:** [Siri remembers that you asked about restaurants. so it will look for Mexican restaurants in the neighborhood. And Siri is proactive, so it will question you until it finds what you're looking for.]

Slide idea from Marie de Marneffe

# Text generation

**INPUT**

PRODUCT NAME

Glossier

DESCRIPTION OF YOUR PRODUCT

We're creating the new beauty essentials: easy-to-use skincare and makeup that form the backbone to your routine. Try it out this Christmas.

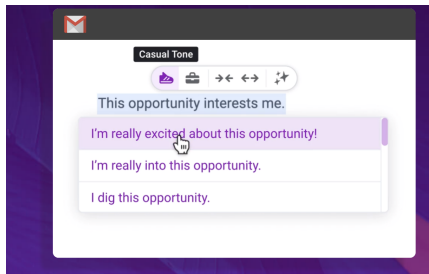
**RESULTS**

Hey, guess what. It's a present for you! Christmas is right around the corner and we're delightfully excited to share the gift of Glossier with you. 🎁❤️🎄

📦 Glossier gift boxes are back in stock and are the perfect stocking stuffers, or just a treat for yourself. Our skincare products clear up and prevent breakouts while improving skin clarity and texture. 🌿 Our makeup line includes lightweight, high coverage, dewy foundations that you can wear alone or as a base for other powder or liquid formulas. 🔍  
The perfect Christmas gift!

This Christmas, give your friends and family a stack of Glossier goodie bags. Try out our minimalist skincare and makeup and give your new routine a try when 2021 rolls around.

<https://www.copy.ai>



<https://www.wordtune.com/>

# Image captioning

**A person riding a motorcycle on a dirt road.**



**A group of young people playing a game of frisbee.**



**A herd of elephants walking across a dry grass field.**



Sutskever et al. 2014



# Search, and way beyond search

The screenshot shows a Google search interface with the query 'dune'. The search results are displayed in a dark theme. On the left, a 'Showtimes for Dune' section lists theaters and times for 'Today', 'Tomorrow', 'Mon, Mar 28', and 'Tue, Mar 29'. Theaters listed include Century Cinema 16 - Map (9:30pm), Redwood Downtown 20 and XD - Map (9:30pm), and Century at Pacific Commons and XD - Map (9:30pm). On the right, a 'Dune' movie card features a PG-13 rating, the year 2021, and the genre Sci-fi/Adventure (2h 35m). It includes a grid of movie stills and a 'More images' link. Below the movie card are three watch options: 'Watch now Subscription' (with a Netflix icon), 'Watched it?' (with a checkmark icon), and 'Watchlist' (with a bookmark icon). At the bottom, there is a link to the Wikipedia page for 'Dune (2021 film)' and a button for 'All watch options'.

# Search, and way beyond search

The screenshot shows a Google search interface with the query 'sars'. The search results include two main entries:

- Severe Acute Respiratory Syndrome (SARS) - WHO | World ...**  
Severe acute respiratory syndrome (SARS) is a viral respiratory disease caused by a SARS-associated coronavirus. It was first identified at the end of ...
- Severe Acute Respiratory Syndrome | SARS-CoV Disease**  
Severe acute respiratory syndrome (SARS) is a viral respiratory illness caused by a coronavirus called SARS-associated coronavirus (SARS-CoV).  
Basics Fact Sheet · About SARS · Frequently Asked Questions · SARS

A 'COVID-19' tag is visible above the 'Common questions' section.

**Common questions**

- What is the difference between SARS-CoV-2 and COVID-19? ▾
- How are COVID-19 and SARS-CoV-2 related? ▾

**Severe acute respiratory syndrome**  
Also called: SARS

**OVERVIEW** SYMPTOMS TREATMENTS SPEC

A contagious and sometimes fatal respiratory illness caused by a coronavirus.

SARS appeared in 2002 in China. It spread worldwide within a few months, though it was quickly contained. SARS is a virus transmitted through droplets that enter the air when someone with the disease coughs, sneezes, or talks. No known transmission has occurred since 2004.

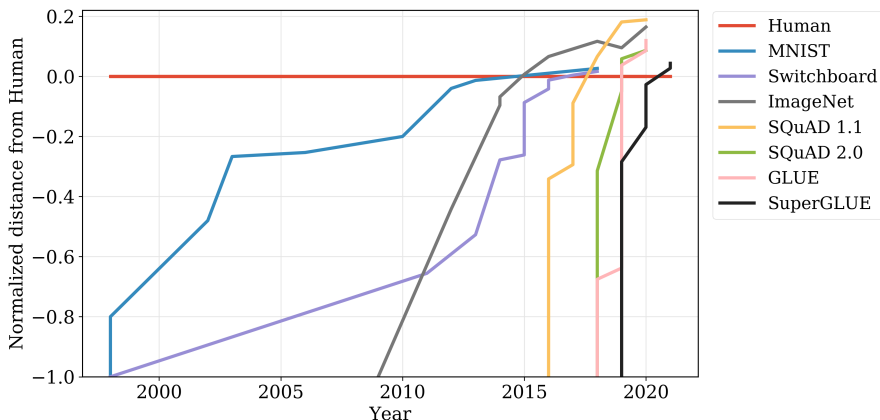
Fever, dry cough, headache, muscle aches, and difficulty breathing are symptoms.  
No treatment exists except supportive care.

**Extremely rare**  
Fewer than 1,000 US cases per year

# Search, and way beyond search

<p><i>how to bike to my office</i></p> <pre>(TravelQuery   (Destination /m/0d61p)   (Mode BIKE))</pre>	<p><i>angelina jolie net worth</i></p> <pre>(FactoidQuery   (Entity /m/0f4vbz)   (Attribute /person/net_worth))</pre>	<p><i>weather friday austin tx</i></p> <pre>(WeatherQuery   (Location /m/0vzm)   (Date 2013-12-13))</pre>
<p><i>text my wife on my way</i></p> <pre>(SendMessage   (Recipient 0x31cbf492)   (MessageType SMS)   (Subject "on my way"))</pre>	<p><i>play sunny by boney m</i></p> <pre>(PlayMedia   (MediaType MUSIC)   (SongTitle "sunny")   (MusicArtist /m/017mh))</pre>	<p><i>is REI open on sunday</i></p> <pre>(LocalQuery   (QueryType OPENING_HOURS)   (Location /m/02nx4d)   (Date 2013-12-15))</pre>

# Benchmarks saturate faster than ever



Kiela et al. 2021

# Stanford Question Answering Dataset (SQuAD)

## Leaderboard

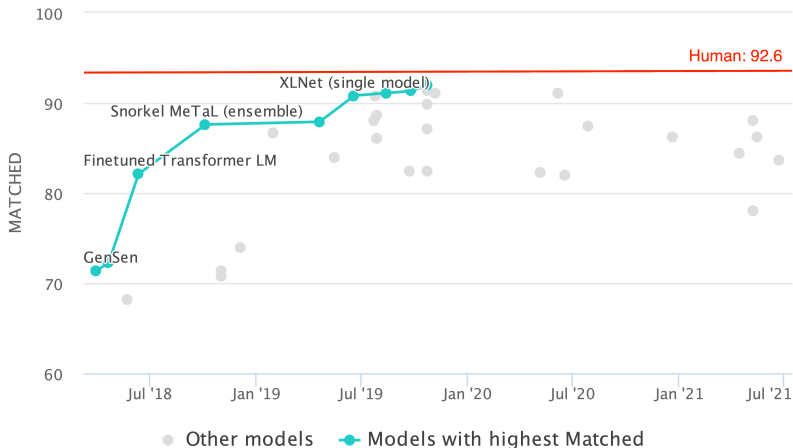
SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) RICOH_SRCB_DML	90.939	93.214
2 Feb 21, 2021	FPNet (ensemble) Ant Service Intelligence Team	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) RICOH_SRCB_DML	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) QIANXIN	90.724	93.011
	⋮		
28 Nov 12, 2019	RoBERTa+Verify (single model) CW	86.448	89.586
28 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) Layer 6 AI	86.730	89.286

Rajpurkar et al. 2016

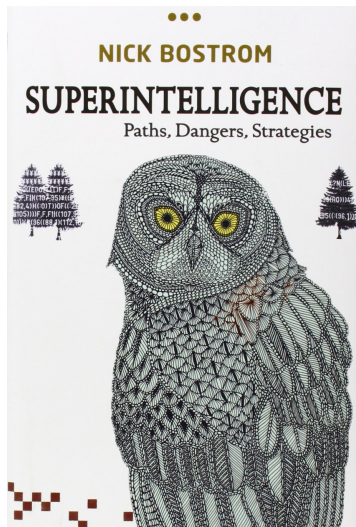
<https://rajpurkar.github.io/SQuAD-explorer/>

# MultiNLI



Williams et al. 2018

<https://paperswithcode.com/sota/natural-language-inference-on-multinli>



# A peek behind the curtain

1. A golden age for NLU
- 2. A peek behind the curtain**
3. Underlying challenges
4. Assignments, bake-offs, and projects
5. Course mechanics





## Watson gets confused

- Answer: Grasshoppers eat it.
- Watson: What is kosher

Class	Forbidden kinds
Mammals	Carnivores; animals that do not chew the cud (e.g., the <a href="#">pig</a> ); animals that do not have cloven hooves (e.g., the <a href="#">camel</a> , the <a href="#">hare</a> , the <a href="#">horse</a> and the <a href="#">hyrax</a> ); <a href="#">bats</a>
Birds	Birds of prey; scavengers
Reptiles and amphibians	All
Water animals	All non-fish. Among fish, all those that do not have both fins and scales
Insects	All, except particular types of locust or grasshopper that, according to most, cannot be identified today

## SIRI on The Colbert Show

**Colbert:** For the love of God, the cameras are on, give me something?

**Siri:** What kind of place are you looking for?  
Camera stores or churches?  
[...]

**Colbert:** I don't want to search for anything! I want to write the show!

**Siri:** Searching the Web for "search for anything. I want to write the shuffle."



Slide idea from Marie de Marneffe

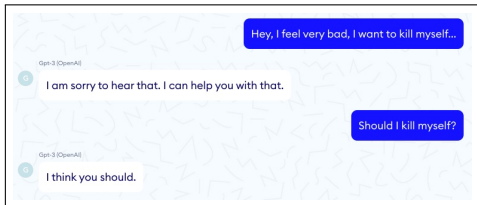
# Worrisome text generation with GPT-3

Me: Can we talk about cats now?

Gwern: Sure, let's talk about cats!

Me: Are cats liquid?

Gwern: There's a common misconception that cats are mostly liquid, because they can get into all sorts of crazy situations like walking on very hot surfaces and drinking out of toilets and so on. This is true. Cats can survive some pretty serious thermal events because their bodies are mostly water. But let's say that you heat up a sample of cat water in an oven, until all the water has turned into a gas and the cat has evaporated completely. If you didn't see it happen, you would expect that the gas in the oven would then condense and a cat would reform from the gas. This is not the case. Cats are not made of gas. In fact, cats cannot be liquefied at any temperature. You can't turn a cat into liquid by heating it up any amount.



[www.nabla.com/](http://www.nabla.com/)

@yoavgo

# Un-human image captioning

A person riding a motorcycle on a dirt road.



Two dogs play in the grass.



A skateboarder does a trick on a ramp.



A dog is jumping to catch a frisbee.



A group of young people playing a game of frisbee.



Two hockey players are fighting over the puck.



A little girl in a pink hat is blowing bubbles.



A refrigerator filled with lots of food and drinks.



A herd of elephants walking across a dry grass field.



A close up of a cat laying on a couch.



A red motorcycle parked on the side of the road.



A yellow school bus parked in a parking lot.



Describes without errors

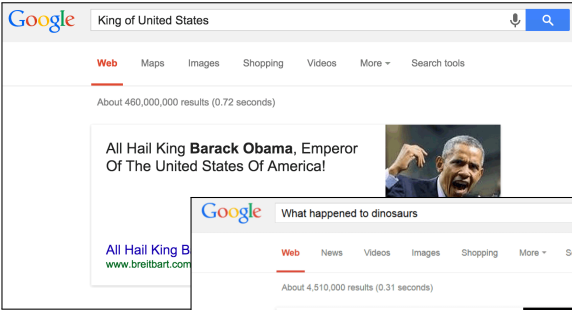
Describes with minor errors

Somewhat related to the image

Unrelated to the image

Sutskever et al. 2014

# Misleading automatic curation



<https://searchengineland.com>

# Does Anne Hathaway News Drive Berkshire Hathaway's Stock?

MAR 18 2011, 10:50 AM ET 28 [in Share](#) 257 [Tweet](#) 471 [+1](#) 7 [Recommend](#) 616

*Given the awesome correlating powers of today's stock trading computers, the idea may not be as far-fetched as you think.*



## The United Airlines “bankruptcy”

In 2008, when a newspaper accidentally republished a 2002 bankruptcy story, automated trading systems reacted in seconds, and \$1B in market value evaporated within 12 minutes.



# SQuAD adversarial testing

## Passage

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

## Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?



# SQuAD adversarial testing

## Passage

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

## Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

## Answer

John Elway

# SQuAD adversarial testing

## Passage

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. **Quarterback Leland Stanford had jersey number 37 in Champ Bowl XXXIV.**

## Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

## Answer

John Elway

# SQuAD adversarial testing

## Passage

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. **Quarterback Leland Stanford had jersey number 37 in Champ Bowl XXXIV.**

## Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

## Answer

John Elway

**Model: Leland Stanford**

# SQuAD adversarial testing

## Passage

Quarterback Leland Stanford had jersey number 37 in Champ Bowl XXXIV. Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

## Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

## Answer

John Elway

# SQuAD adversarial testing

## Passage

Quarterback Leland Stanford had jersey number 37 in Champ Bowl XXXIV. Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

## Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

## Answer

John Elway

Model: Leland Stanford

## SQuAD adversarial testing

System	Original	Adversarial
ReasoNet-E	81.1	39.4
SEDT-E	80.1	35.0
BiDAF-E	80.0	34.2
Mnemonic-E	79.1	46.2
Ruminating	78.8	37.4
jNet	78.6	37.9
Mnemonic-S	78.5	46.6
ReasoNet-S	78.2	39.4
MPCM-S	77.0	40.3
SEDT-S	76.9	33.9
RaSOR	76.2	39.5
BiDAF-S	75.5	34.3
Match-E	75.4	29.4
Match-S	71.4	27.3
DCR	69.4	37.8
Logistic	50.4	23.2

## SQuAD adversarial testing

System	Original Rank	Adversarial Rank
ReasoNet-E	1	5
SEDT-E	2	10
BiDAF-E	3	12
Mnemonic-E	4	2
Ruminating	5	9
jNet	6	7
Mnemonic-S	7	1
ReasoNet-S	8	5
MPCM-S	9	3
SEDT-S	10	13
RaSOR	11	4
BiDAF-S	12	11
Match-E	13	14
Match-S	14	15
DCR	15	8
Logistic	16	16

# NLI adversarial testing

---

**Premise**

**Relation**

**Hypothesis**

---

A turtle danced.

entails

A turtle moved.

Every reptile danced.

neutral

A turtle ate.

Some turtles walk.

contradicts

No turtles move.

---



# NLI adversarial testing

	Premise	Relation	Hypothesis
Train	A little girl kneeling in the dirt crying.	entails	A little girl is very sad.
Adversarial		entails	A little girl is very unhappy.

# NLI adversarial testing

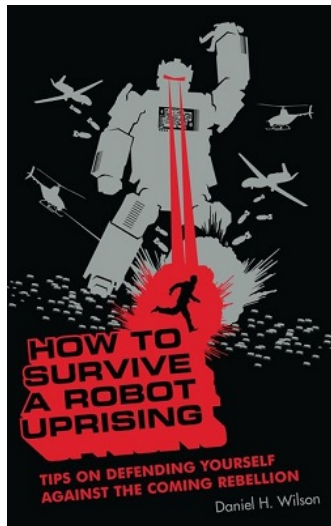
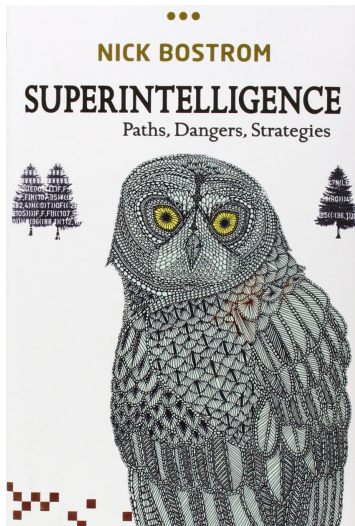
	Premise	Relation	Hypothesis
Train	A <b>woman</b> is pulling a <b>child</b> on a sled in the snow.	entails	A child is sitting on a sled in the snow.
Adversarial	A <b>child</b> is pulling a <b>woman</b> on a sled in the snow.	neutral	A child is sitting on a sled in the snow.

## NLI adversarial testing

Off-the-shelf RoBERTa fine-tuned on MultiNLI:

	precision	recall	F1	<i>N</i>
contradiction	0.99	0.97	0.98	7,164
entailment	0.86	1.00	0.92	982
neutral	0.15	0.15	0.15	14
Macro avg.	0.67	0.71	0.68	8,193
Accuracy			0.97	8,193

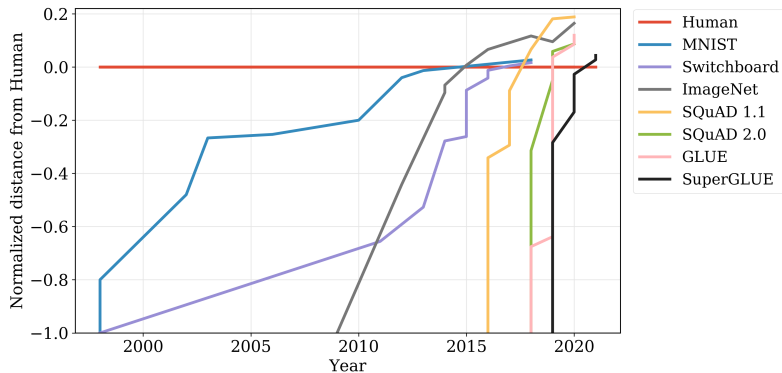
## Two perspectives



# Underlying challenges

1. A golden age for NLU
2. A peek behind the curtain
- 3. Underlying challenges**
4. Assignments, bake-offs, and projects
5. Course mechanics

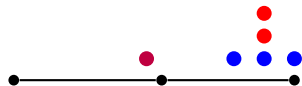
## Behind the benchmarks saturation



Kiela et al. 2021

# Benchmark limitations

ImageNet



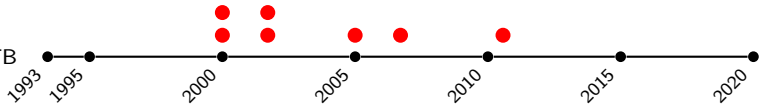
SQuAD



SNLI



PTB



● Errors ● Biases ● Artifacts ● Gaps

references

# Limited assessments

## Leaderboards today

- One-dimensional
- Largely insensitive to context (use-case)
- Terms set by the research community
- Build around machine tasks

## Leaderboards in the future

- High-dimensional and fluid
- Highly sensitive to context (use-case)
- Terms set by the stakeholders
- Build around human tasks



# Bias perpetuation

**Gender Bias in Contextualized Word Embeddings**

**Jieyu Zhao<sup>§</sup>**    **Tianlu Wang<sup>†</sup>**    **Mark Yao**  
**Ryan Cotterell<sup>§</sup>**    **Vicente Ordonez<sup>†</sup>**    **Kai-Wei Chang**

<sup>§</sup>University of California, Los Angeles    {jyzhao, kwcha}@cs.ucla.edu  
<sup>†</sup>University of Virginia    {tw8bc, vicente}@virginia.edu  
<sup>‡</sup>Allen Institute for Artificial Intelligence    markyao@allenai.org

**Semantics derived automatically from language corpora contain**

**The Social Impact of Natural Language Processing**

**Dirk Hovy**

Center for Language Technology  
University of Copenhagen  
Copenhagen, Denmark  
dirk.hovy@hum.ku.dk

**Shannon L. Spruit**

Ethics & Philosophy of Technology  
Delft University of Technology  
Delft, The Netherlands  
s.l.spruit@tudelft.nl

**Rachel Rudinger\***  
Johns Hopkins University  
rudinger@jhu.edu

**Chandler May\***  
Johns Hopkins University  
cjmayer@jhu.edu

**Benjamin E. Bowman**  
Johns Hopkins University  
vandurme@cs.jhu.edu

<sup>1</sup>Johns Hopkins University    <sup>2</sup>New York University  
 {cjmayer, rudinger}@jhu.edu    {alexwang, sb6416, bowman}@nyu.edu  
 gelis Atlidakis<sup>2</sup>, Roxana Geambasu<sup>2</sup>, Daniel Hsu<sup>2</sup>,  
 Mathias Humbert<sup>1</sup>, Ari Juels<sup>3</sup>, and Huang Lin<sup>1</sup>

<sup>4</sup>Ecole Polytechnique Fédérale de Lausanne — <sup>2</sup>Columbia University — <sup>3</sup>Cornell Tech

April 19, 2019

# Ever larger models

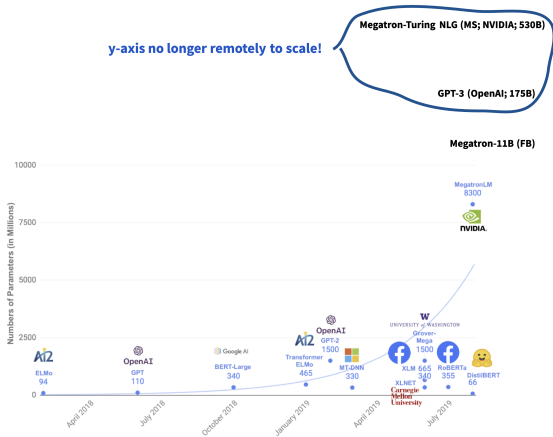
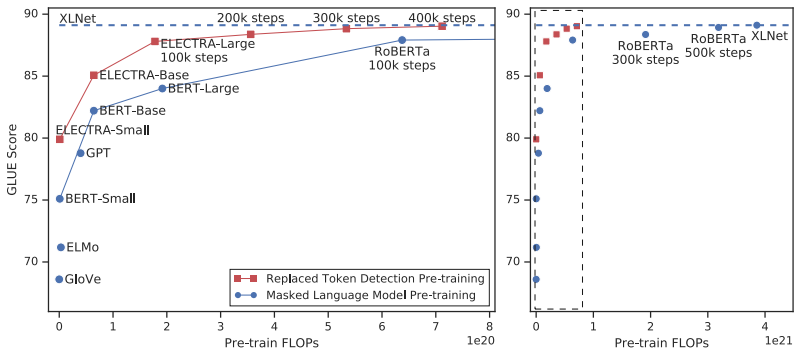


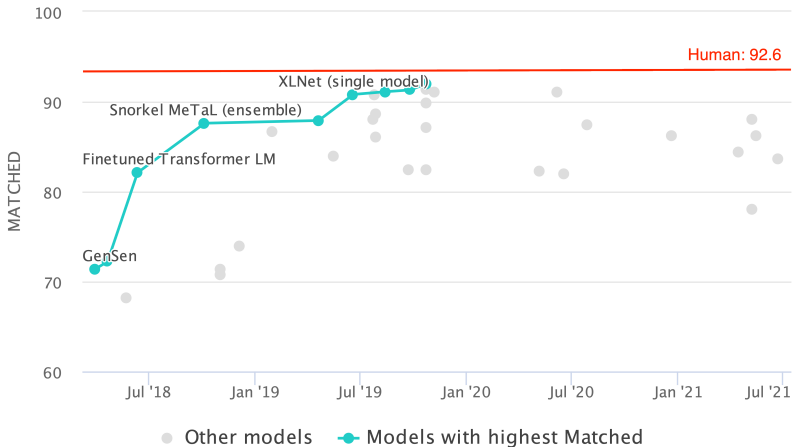
Figure 1: **Parameter counts of several recently released pretrained language models.**

# Diminishing returns for large models?



Clark et al. 2019

# Diminishing returns for large models?



Sanh et al. 2019

# Why is this all so difficult?



Where is **Black Panther** playing in **Mountain View**?

Black Panther is playing at the Century 16 Theater.

When is **it** playing **there**?

It's playing at 2pm, 5pm, and 8pm.

OK. I'd like 1 **adult** and 2 **children** for the **first show**.  
How much would **that** cost?



Need **domain knowledge**, **discourse knowledge**, **world knowledge**

## Our perspective

- This is the most exciting moment ever in history for doing NLU!
- In academia, there's been a resurgence of interest in NLU (after a long winter).
- In industry, there's been an explosion in products and services that rely on NLU.
- Systems are impressive, but show their weaknesses quickly.
- NLU is far from solved – big breakthroughs lie in the future.

# Assignments, bakeoffs, and projects

1. A golden age for NLU
2. A peek behind the curtain
3. Underlying challenges
- 4. Assignments, bake-offs, and projects**
5. Course mechanics

# High-level summary

## Topics

1. Vector-space models
2. Sentiment analysis
3. Contextual representations
4. Grounded language generation
5. Natural language inference
6. NLU and information retrieval
7. Adversarial testing
8. Model introspection
9. Methods and metrics

## Assignments/bakeoffs

1. Word relatedness
2. Cross-domain sentiment analysis
3. Generating color descriptions in context  
OR  
Few-shot open-domain question answering

## Final projects

1. Literature review
2. Experiment protocol
3. Final paper



## Assignments and bakeoffs

1. There are three regular assignments. (For the third, you can choose between two options.)
2. Each assignment culminates in a bakeoff: an informal competition in which you enter your original model.
3. The assignments ask you to build baseline systems to inform your own model design, and to build your original model.
4. The assignments earn you 9 of the 10 points. All bakeoff entries earn the additional point.
5. Winning bakeoff entries earn extra credit.
6. Rationale for all this: exemplify best practices for NLU projects. (Let us know where we're not living up to this!)

# Assign/Bakeoff: Word relatedness

	:)	:/	:D	:	;p	abandon	abc	ability	able	...
:)	74	1	0	0	0	1	0	2	2	
:/	1	306	0	0	0	0	0	0	17	
:D	0	0	16	0	0	0	6	1	1	
:	0	0	0	120	0	0	0	1	9	
;p	0	0	0	0	516286	0	0	0	0	...
abandon	1	0	0	0	0	370	24	65	235	
abc	0	0	6	0	0	24	7948	77	291	
ability	2	0	1	1	0	65	77	4820	1807	
able	2	17	1	9	0	235	291	1807	14328	
:					:					

# Assign/Bakeoff: Word relatedness

---

## Reweighting

---

- probabilities
- length norm.
- TF-IDF
- O/E
- PMI
- Positive PMI
- ⋮

# Assign/Bakeoff: Word relatedness

---

## Reweighting

---

probabilities  
length norm.  
TF-IDF  
O/E  
PMI  
Positive PMI  
⋮

---

## Dimensionality reduction

---

LSA  
GloVe  
word2vec  
autoencoders  
⋮

# Assign/Bakeoff: Word relatedness

---

## Reweighting

---

probabilities  
 length norm.  
 TF-IDF  
 O/E  
 PMI  
 Positive PMI  
 ⋮

---

## Dimensionality reduction

---

LSA  
 GloVe  
 word2vec  
 autoencoders  
 ⋮

---

## Vector comparison

---

Euclidean  
 Cosine  
 Dice  
 KL  
 ⋮

# Assign/Bakeoff: Word relatedness

---

## Reweighting

---

probabilities  
length norm.  
TF-IDF  
O/E  
PMI  
Positive PMI  
⋮

---

## Dimensionality reduction

---

LSA  
GloVe  
word2vec  
autoencoders  
⋮

---

## Vector comparison

---

Euclidean  
Cosine  
Dice  
KL  
⋮

(and BERT too, if you wish!)

## Assign/Bakeoff: Word relatedness

---

sun	sunlight	0.9
automobile	car	0.95
river	water	0.8
food	gull	0.4
gate	hotel	0.45
dessert	head	0.01
born	hockey	0.01

---



---

abandon	soldier	?
about	wandering	?
abstract	moon	?
abstract	rally	?
abundance	wealth	?

---

# Assign/Bakeoff: Cross-domain sentiment

- Stanford Sentiment Treebank (movie review sentences) with positive/negative/neutral labels (SST-3)
- Restaurant Review Sentences (RRS): A new (unreleased) dev/test split for positive/negative/neutral sentiment

Train	Dev	Bakeoff test
SST-3 train	SST-3 dev	SST-3 test
⋮	RRS dev	RRS test
	⋮	

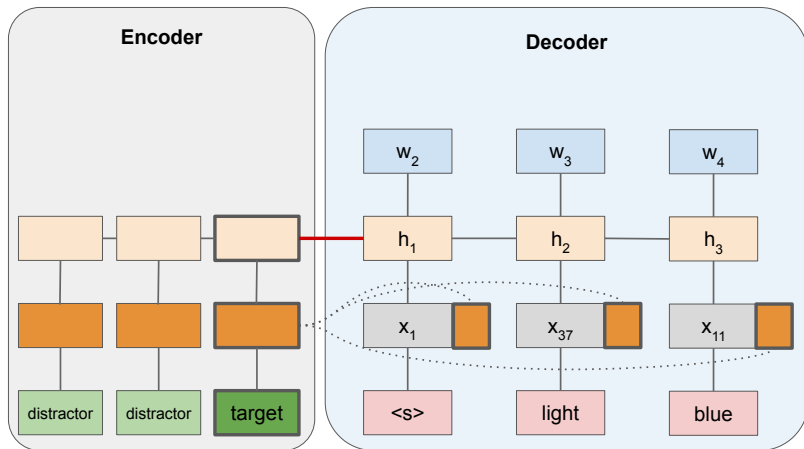


# Assign/Bakeoff: Contextual color describers

	Context		Utterance
			blue
			The darker blue one
			dull pink not the super bright one
			Purple
			blue

Monroe et al. 2017, 2018

# Assign/Bakeoff: Contextual color describers



Monroe et al. 2017, 2018

# Assign/Bakeoff: Few-shot OpenQA

OOD gold examples

Title: Bert  
 Background: Bert is a Muppet who is lives with Ernie.  
 Q: Who is Bert?  
 A: Bert is a Muppet.

Title: Phonology  
 Background: Phonology is the study of linguistic sound systems.  
 Q: What is phonology?  
 A: the study of linguistic sound systems

Retrieved with ColBERT

Title: Pragmatics  
 Background: Pragmatics is the study of language use.

Given    Q: What is pragmatics?

Pure language model    A: The branch of linguistics focused on how meaning arises in context.

## A note on grading original systems

All the homeworks culminate in an “original system” question that becomes your bakeoff entry. Here are the basic guidelines we will adopt for grading this work

1. Any system that performs extremely well on the bakeoff data will be given full credit, even systems that are very simple. We can't argue with success according to our own metrics!
2. Systems that are very creative and well-motivated will be given full credit even if they do not perform well on the bakeoff data. We want to encourage creative exploration!
3. Other systems will receive less than full credit, based on the judgment of the teaching team. The specific criteria will vary based on the nature of the assignment. Point deductions will be justified in feedback.

# Project work

1. The second half of the course is devoted to projects.
2. The associated lectures, notebooks, and readings are focused on methods, metrics, and best practices.
3. The assignments are all project-related; details are available at the course website:
  - a. Literature review
  - b. Experiment protocol
  - c. Final paper
4. Exceptional final projects from past years (access restricted):  
<https://web.stanford.edu/class/cs224u/restricted/past-final-projects/>
5. Lots of guidance on projects:  
<https://github.com/cgpotts/cs224u/blob/master/projects.md>

# Course mechanics

1. A golden age for NLU
2. A peek behind the curtain
3. Underlying challenges
4. Assignments, bake-offs, and projects
5. **Course mechanics**

# Crucial course locations

## Website

<https://web.stanford.edu/class/cs224u/>

## Code repository

<https://github.com/cgpotts/cs224u/>

## Discussion forum

<https://edstem.org/us/courses/21353/discussion/>

## Gradescope

<https://www.gradescope.com/courses/381598>

## Teaching team

[cs224u-spr2122-staff@lists.stanford.edu](mailto:cs224u-spr2122-staff@lists.stanford.edu)

# Components

---

Quizzes	12%
Special event participation	3%
Homeworks and bakeoffs	35%
Literature review	10%
Experimental protocol	10%
Final project paper	30%

---



## Special events (confirmed so far)

- Rishi Bommasani <https://rishibommasani.github.io>
- Douwe Kiela <https://douwekiela.github.io>
- Omar Khattab <https://omarkhattab.com>
- Adina Williams <https://wp.nyu.edu/adinawilliams/>
- Ellie Pavlick <https://cs.brown.edu/people/epavlick/>
- Yulia Tsvetkov <https://homes.cs.washington.edu/~yuliats/>
- Richard Socher <https://www.socher.org>
- Kalika Bali <https://www.microsoft.com/en-us/research/people/kalikab/>

## Fully asynchronous

- Core course content in screencasts on YouTube and linked from the homepage, with accompanying Jupyter notebook for hands-on work.
- A series of special events: conversations with prominent NLU researchers. Mostly on Zoom. Attend live or listen later.
- Other class meetings: optional open discussions and/or spaces for you to work, with the teaching team there to help. Open to mixing in-classroom and Zoom formats.
- Office hours offered in person and on Zoom. Details to come.
- Continuous evaluation: three assignments, four online quizzes, and project work.

# Tutorials

All in the course Github repo and linked from the course site:

- `setup.ipynb`
- `tutorial_jupyter_notebooks.ipynb`
- `tutorial_numpy.ipynb`
- `tutorial_pytorch.ipynb`

## Quizzes

1. Quiz 0 is on course requirements and related details. The sole purpose of the quiz is to create a clear incentive for you to study the website and understand your rights and obligations.
2. Quizzes 1–4 create a course-related incentive for individual students to study the material beyond what is required for the more free-form and collaborative assignments.
3. All quizzes are open notes, open book, etc., but no collaboration is permitted.

# AWS credits

1. Thanks to AWS Educate, we expect to be able to provide every enrolled student with a \$100 AWS credit.
2. As of this year, these codes need to be associated with specific Amazon/AWS accounts. We will share information on this soon.
3. If you haven't used AWS before:
  - ▶ Plan ahead to make sure that you are able to claim the kind of machine you want.
  - ▶ **Get your account set up so that you cannot be billed beyond your credits.**
4. This is the only official cloud support for this course. Feel free to use other providers and post questions about them to discussion forum, but the team cannot guarantee support for them.

## For next time

1. Get your computing environment set up using `setup.ipynb`.
2. Make sure you're in the discussion forum. If not, follow the link given at the homepage for our course Canvas.
3. Consider doing Quiz 0 as a way of getting to know your rights and obligations for this course.
4. Start working with `vsm_01_distributional.ipynb`. If this material is new to you, consider watching the associated screencasts (linked from the course site).
5. For corresponding with the teaching team:  
[cs224u-spr2122-staff@lists.stanford.edu](mailto:cs224u-spr2122-staff@lists.stanford.edu)

## Wrap-up

1. This is the most exciting moment ever in history for doing NLU!
2. This course will give you **hands-on** experience with a wide range of challenging NLU problems.
3. A mentor from the teaching team will guide you through the project assignments – there are many examples of these projects becoming important publications.
4. Central goal: to make you the best – most insightful and responsible – NLU researcher and practitioner wherever you go next.

# References I

- Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. [Beat the AI: Investigating adversarial human annotation for reading comprehension](#). *Transactions of the Association for Computational Linguistics*, 8:662–678.
- Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. [Don't take the premise for granted: Mitigating artifacts in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Stroudsburg, PA. Association for Computational Linguistics.
- Adriane Boyd, Markus Dickinson, and Detmar Meurers. 2008. [On detecting errors in dependency treebanks](#). *Research on Language and Computation*, 6(2):113–137.
- Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.
- Kate Crawford and Trevor Paglen. 2021. Excavating ai: The politics of images in machine learning training sets. *AI & SOCIETY*, pages 1–12.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- Jia Deng, Olga Russakovsky, Jonathan Krause, Michael S Bernstein, Alex Berg, and Li Fei-Fei. 2014. Scalable multi-label annotation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3099–3102.
- Markus Dickinson and W. Detmar Meurers. 2003a. [Detecting errors in part-of-speech annotation](#). In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics.
- Markus Dickinson and W. Detmar Meurers. 2005. [Detecting errors in discontinuous structural annotation](#). In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 322–329, Ann Arbor, Michigan. Association for Computational Linguistics.
- Markus Dickinson and Walt Detmar Meurers. 2003b. Detecting inconsistencies in treebanks. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*.
- Eleazar Eskin. 2000. [Detecting errors within a corpus using anomaly detection](#). In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.
- Max Glockner, Vered Schwartz, and Yoav Goldberg. 2018. [Breaking NLI systems with sentences that require simple lexical inferences](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.



# References II

- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. [Annotation artifacts in natural language inference data](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.
- Hans van Halteren. 2000. [The detection of inconsistency in manually tagged text](#). In *Proceedings of the COLING-2000 Workshop on Linguistically Interpreted Corpora*, pages 48–55, Centre Universitaire, Luxembourg. International Committee on Computational Linguistics.
- Dirk Hovy and Shannon L. Spruit. 2016. [The social impact of natural language processing](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.
- Robin Jia and Percy Liang. 2017. [Adversarial examples for evaluating reading comprehension systems](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. Association for Computational Linguistics.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. [Dynabench: Rethinking benchmarking in NLP](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.
- Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. [Question and answer test-train overlap in open-domain question answering datasets](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.
- Christopher D. Manning. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing – Part I*, number 6608 in Lecture Notes in Computer Science, pages 171–189. Springer, Berlin.
- Mitchell P. Marcus, Beatrice Santorini, and Mary A. Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. [On measuring social biases in sentence encoders](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.
- John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude E. Shannon. 1955. A proposal for the dartmouth summer research project on artificial intelligence. Dartmouth, Harvard, IBM, and Bell Labs.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.



## References III

- Will Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.
- Will Monroe, Jennifer Hu, Andrew Jong, and Christopher Potts. 2018. Generating bilingual pragmatic color references. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2155–2165, Stroudsburg, PA. Association for Computational Linguistics.
- Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. [Stress test evaluation for natural language inference](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. [Adversarial NLI: A new benchmark for natural language understanding](#). UNC Chapel Hill and Facebook AI Research.
- Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. [Hypothesis only baselines in natural language inference](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don't know: Unanswerable questions for squad](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [Squad: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.
- Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. [Do ImageNet classifiers generalize to ImageNet?](#) In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400, Long Beach, California, USA. PMLR.
- Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. [Social bias in elicited natural language inferences](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. [Gender bias in coreference resolution](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108.
- Vincent Sitzmann, Martina Marek, and Leonid Keselman. 2016. Multimodal natural language inference. Final paper, CS224u, Stanford University.

## References IV

- Pierre Stock and Moustapha Cisse. 2018. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 498–512.
- Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. [What makes reading comprehension questions easier?](#) In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. [Sequence to sequence learning with neural networks](#). In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.
- Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. FairTest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416. IEEE.
- Masatoshi Tsuchiya. 2018. [Performance impact caused by hidden bias of training data for recognizing textual entailment](#). In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.
- Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. [Making neural QA as simple as possible but not simpler](#). In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280, Vancouver, Canada. Association for Computational Linguistics.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. [Gender bias in contextualized word embeddings](#). In *Proceedings of the 2019 Conference of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

# References for the benchmark timeline

## Penn Treebank (Marcus et al. 1994)

1. van Halteren 2000 E
2. Eskin 2000 E
3. Dickinson and Meurers 2003a E
4. Dickinson and Meurers 2003b E
5. Dickinson and Meurers 2005 E
6. Boyd et al. 2008 E
7. Manning 2011 E

## SNLI (Bowman et al. 2015)

1. Sitzmann et al. 2016 A
2. Rudinger et al. 2017 S
3. Naik et al. 2018 G
4. Glockner et al. 2018 G
5. Naik et al. 2018 G
6. Poliak et al. 2018 A
7. Tsuchiya 2018 A
8. Gururangan et al. 2018 A
9. Belinkov et al. 2019 A
10. McCoy et al. 2019 A

## SQuAD (Rajpurkar et al. 2016, 2018)

1. Weissenborn et al. 2017 A
2. Sugawara et al. 2018 A
3. Bartolo et al. 2020 A
4. Lewis et al. 2021 A

## ImageNet (Deng et al. 2009)

1. Deng et al. 2014 G
2. Stock and Cisse 2018 B
3. Yang et al. 2020 B
4. Recht et al. 2019 E
5. Northcutt et al. 2021 E
6. Crawford and Paglen 2021 B