Introduction and course overview

Christopher Potts

Stanford Linguistics

CS 224U: Natural language understanding
April 6

Welcome

Bill MacCartney

Chris Potts

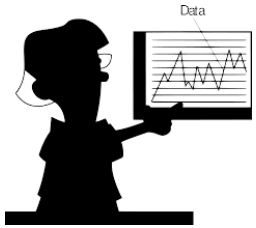
Adam Keppler

Nishit Asnani

Rohan Badlani

Michael Hahn

John Kamalu

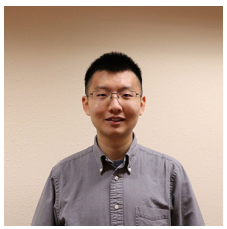


Mandy Lu

Jonathan Mak

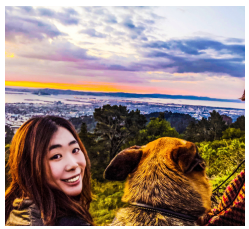
Chetanya Rastogi

Kaushik Ram Sadagopan



Zijian Wang

Sahil Yakhmi



Kaylie Zhu

COVID-19 accommodations 😷

CS224u will be a fully online course for the entire quarter:

- The class meetings will be video seminars (discussion encouraged!), which will be recorded and put on Canvas.
- Office hours will also be by video using a queue system.
- We will rely even more than usual on our discussion forum to exchange ideas, address challenges, and collaborate with each other.

COVID-19 and NLU

- CORD-19:
  https://pages.semanticscholar.org/coronavirus-research
- Elsevier Coronavirus Research Repository:
  https://coronavirus.1science.com/
- Coronavirus Tweets:
  https://www.kaggle.com/smid80/coronavirus-covid19-tweets
- CS472 Data science and AI for COVID-19
  https://sites.google.com/view/data-science-covid-19
- Google's COVID-19 Public Datasets
  https://console.cloud.google.com/marketplace/details/bigquery-public-datasets/covid19-public-data-program

Plan for today

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

Advances in NLU

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

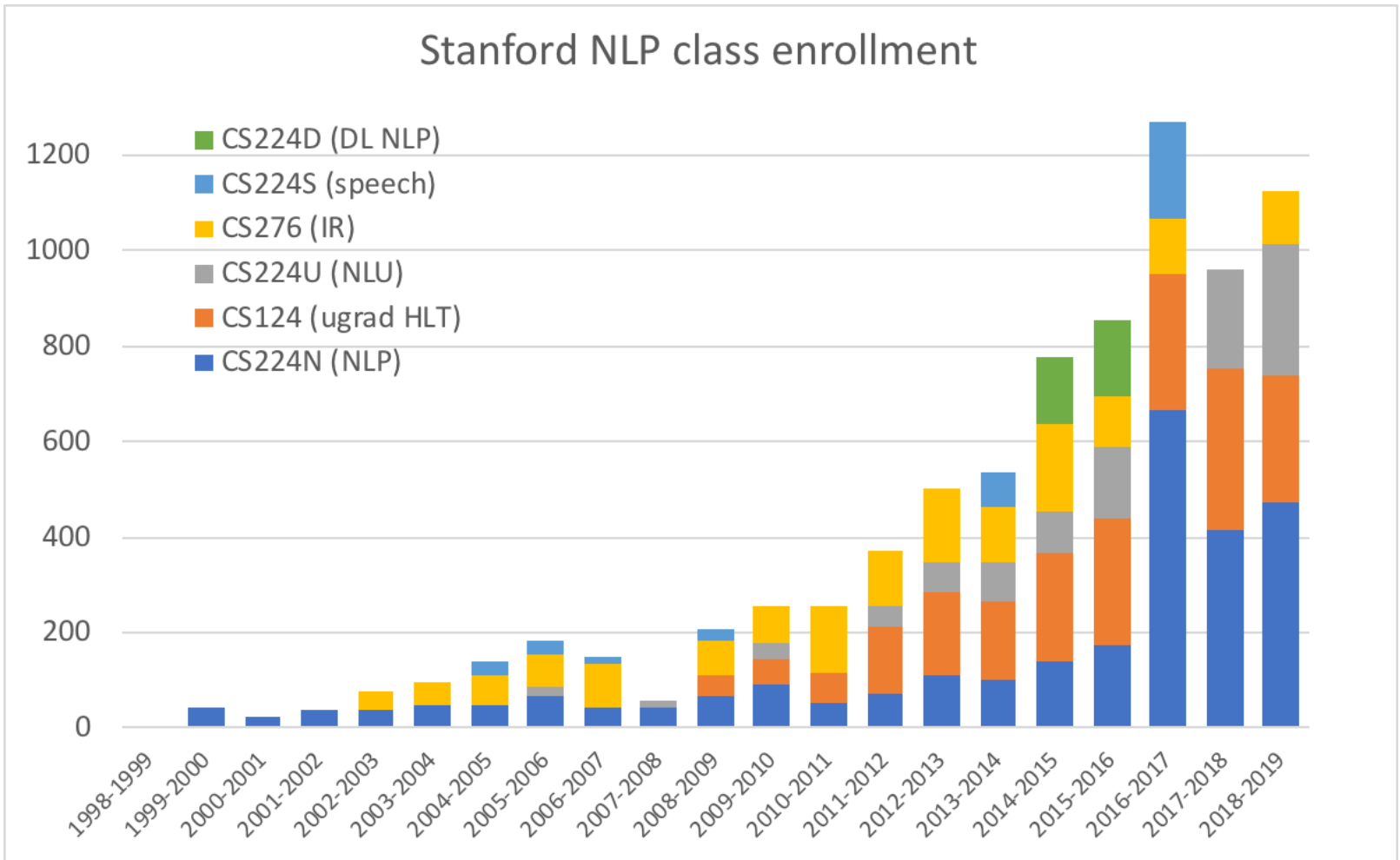
A brief history of NLU approaches

- 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 (**LRLDG**) systems; restricted applications.
- Mid-1990s: Machine learning revolution in NLP leads to a decrease in NLU work.
- Late 2000s: **LRLDG** systems re-emerge, now with *learning*.
- Mid-2010s: NLU returns to center stage, with deep learning the most prevalent set of techniques. **LRLDG** systems go into decline.
- 2020–: [predictions?]

A brief history of NLU technologies

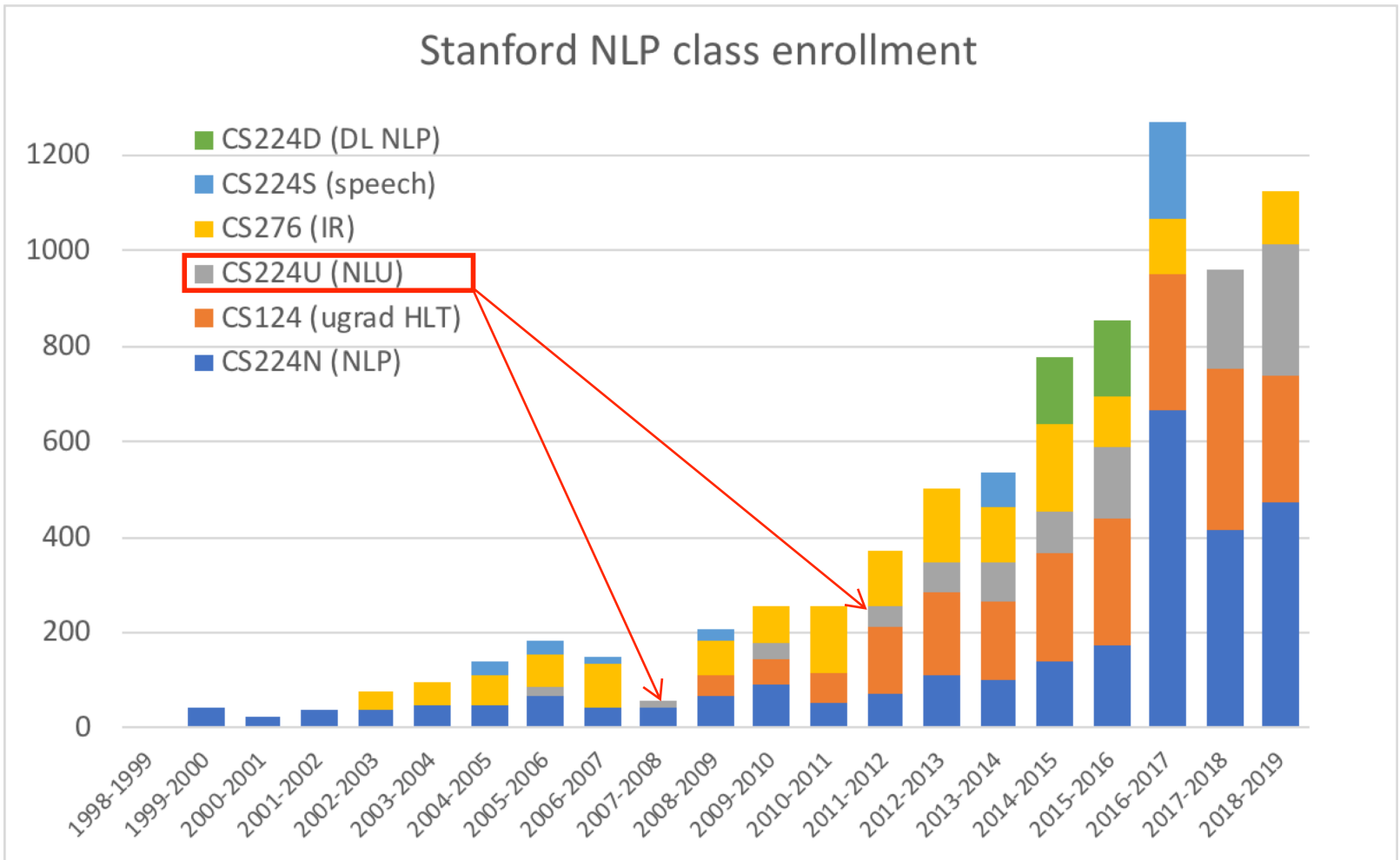
- 1966: Eliza
- 1988: Latent Semantic Analysis patent
- January 2011: IBM Watson beats Jeopardy! champions
- October 2011: Apple Siri launches in beta
- April 2014: Microsoft Cortana demoed
- November 2014: Amazon Alexa
- May 2016: Google Assistant

The history of CS224u enrollments



h/t @StanfordNLP

The history of CS224u enrollments



h/t @StanfordNLP

The history of CS224u topics

2012

1. WordNet
2. Word sense disambiguation
3. Vector-space models
4. Dependency parsing for NLU
5. Relation extraction
6. Semantic role labeling
7. Semantic parsing
8. Textual inference
9. Sentiment analysis
10. Semantic composition with vectors
11. Text segmentation
12. Dialogue

2020

1. Vector-space models
2. Sentiment analysis
3. Relation extraction
4. Natural Language Inference
5. Grounding
6. Contextual word representations
7. Adversarial testing
8. Methods and metrics

A golden age for NLU

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

Artificial assistants

The promise of these artificial assistants

You: Any good burger joints around here?

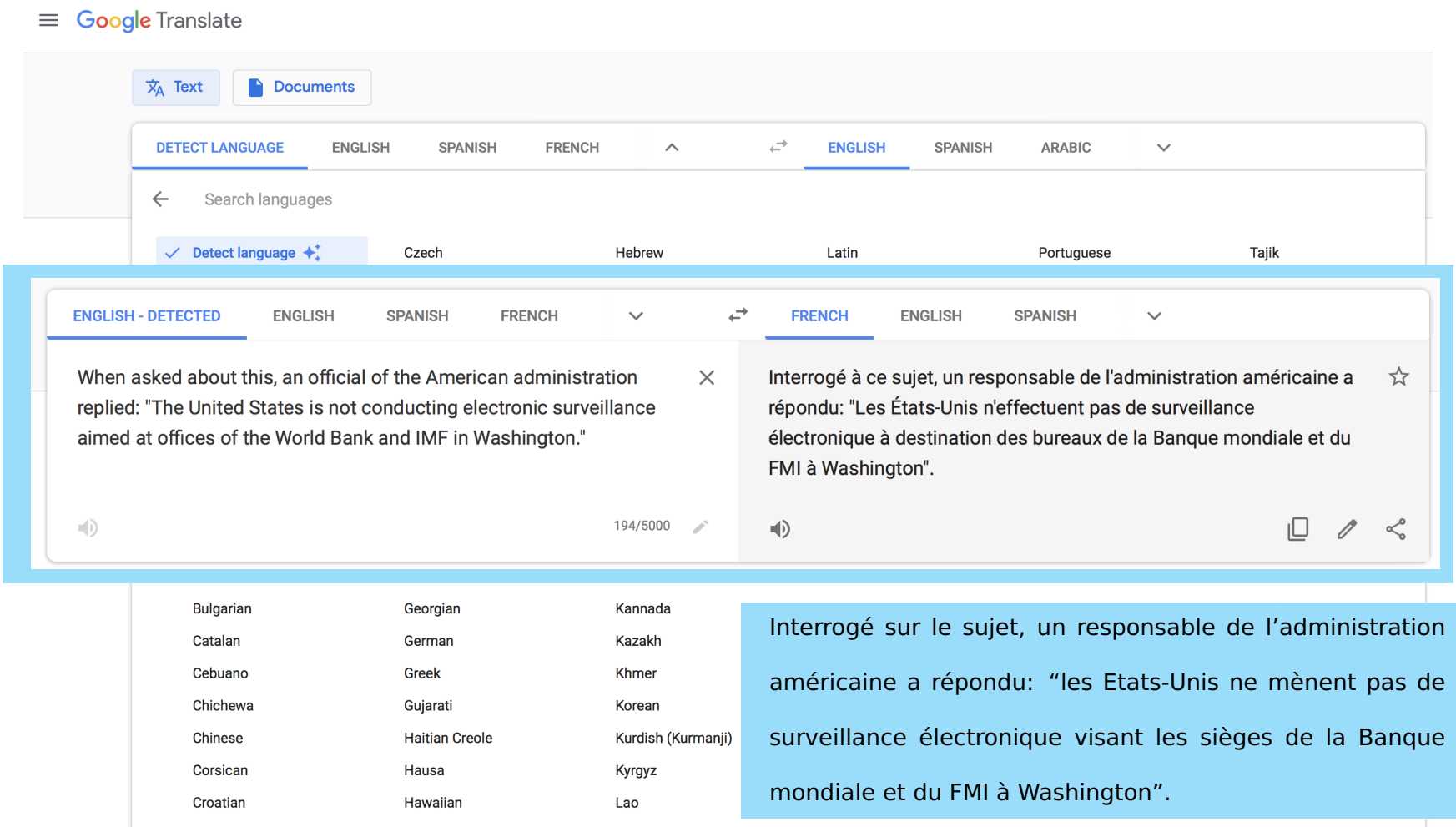
Siri: I found a number of burger restaurants near you.

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

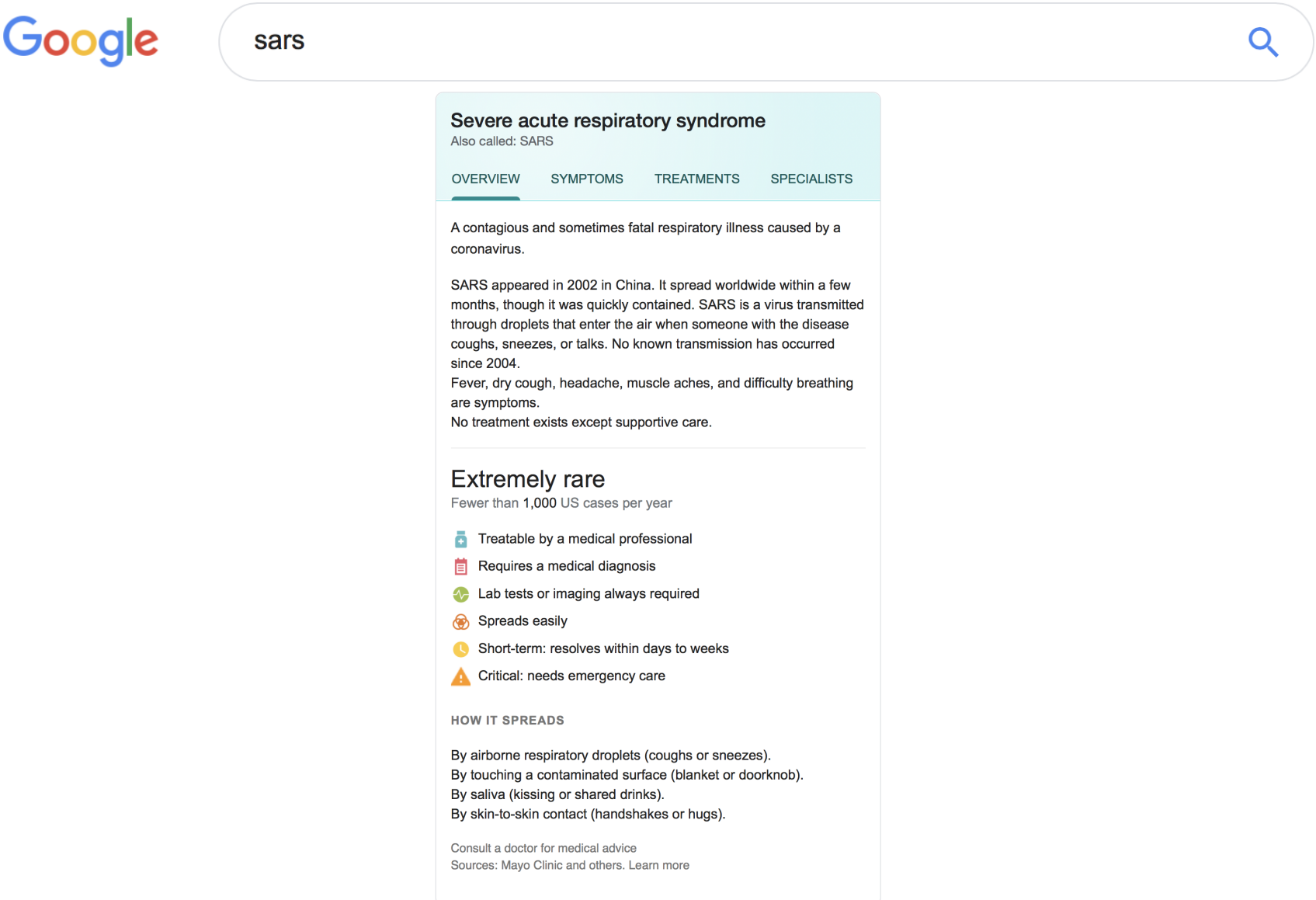
Translation



Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".

Search, and way beyond search

Search, and way beyond search



Search, and way beyond search

Search, and way beyond search

Google

parasite

Parasite

R 2019 · Drama/Mystery · 2h 12m

Play trailer on YouTube

8.6/10	99%	4/4
IMDb	Rotten Tomatoes	Roger Ebert

90% liked this movie

Google users

Greed and class discrimination threaten the newly formed symbiotic relationship between the wealthy Park family and the destitute Kim clan.

Release date: October 5, 2019 (USA)

Director: Bong Joon-ho

Hangul: 기생충

Awards: Academy Award for Best Picture, Palme d'Or, MORE

Nominations: Cannes Best Actress Award, MORE

Search, and way beyond search

how to bike to my office

```
(TravelQuery
  (Destination /m/0d6lp)
  (Mode BIKE))
```

angelina jolie net worth

```
(FactoidQuery
  (Entity /m/0f4vbz)
  (Attribute /person/net_worth))
```

weather friday austin tx

```
(WeatherQuery
  (Location /m/0vzm)
  (Date 2013-12-13))
```

text my wife on my way

```
(SendMessage
  (Recipient 0x31cbf492)
  (MessageType SMS)
  (Subject "on my way"))
```

play sunny by boney m

```
(PlayMedia
  (MediaType MUSIC)
  (SongTitle "sunny")
  (MusicArtist /m/017mh))
```

is REI open on sunday

```
(LocalQuery
  (QueryType OPENING_HOURS)
  (Location /m/02nx4d)
  (Date 2013-12-15))
```

Stanford Question Answering Dataset (SQuAD)

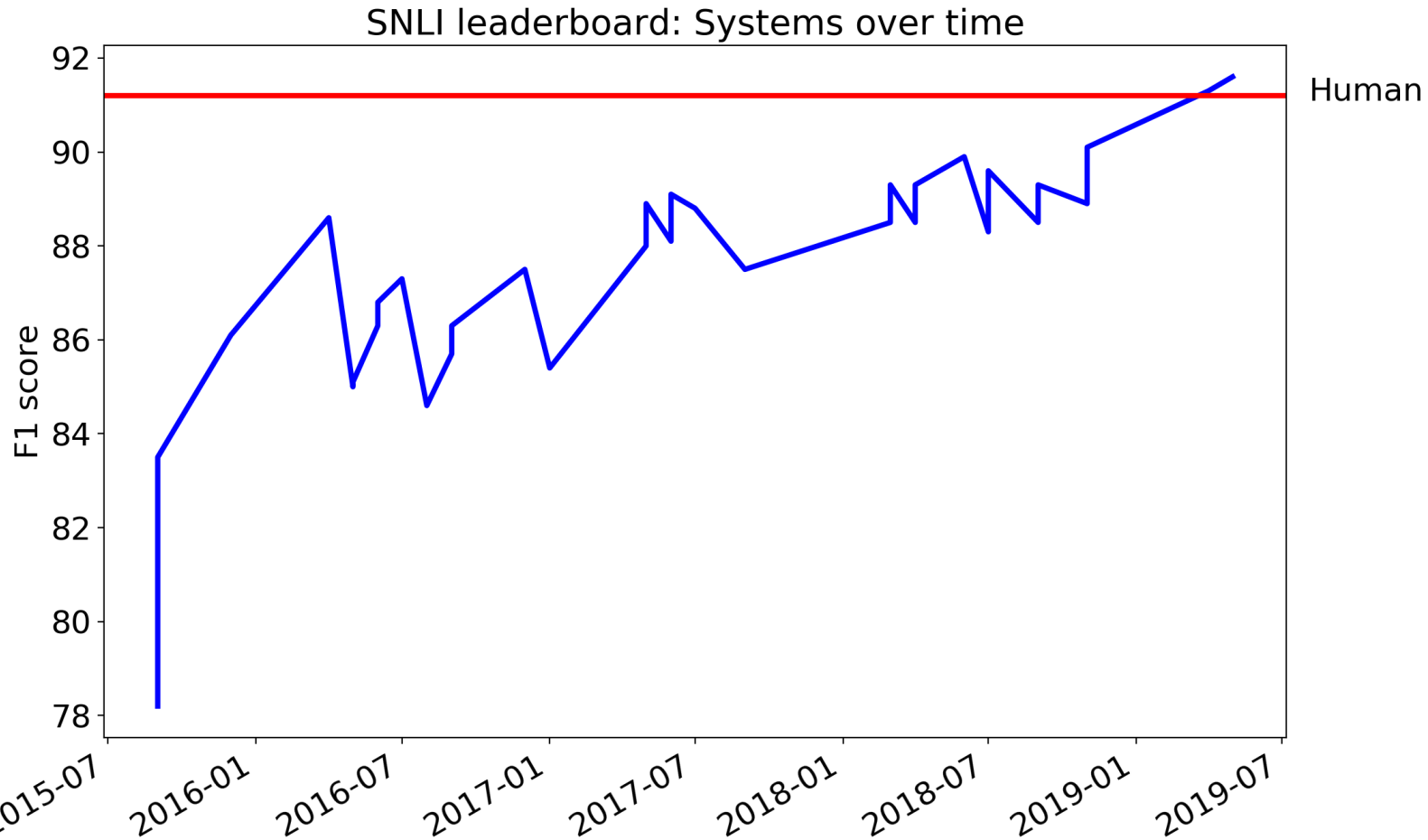
Leaderboard

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 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	**90.115**	**92.580**
2 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
3 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
3 Feb 25, 2020	Albert_Verifier_AA_Net (ensemble) *QIANXIN*	89.743	92.180
4 Jan 23, 2020	albert+transform+verify (ensemble) *qianxin*	89.528	92.059
⋮	⋮	⋮	⋮
13 Nov 12, 2019	RoBERTa+Verify (single model) *CW*	86.448	89.586
13 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286

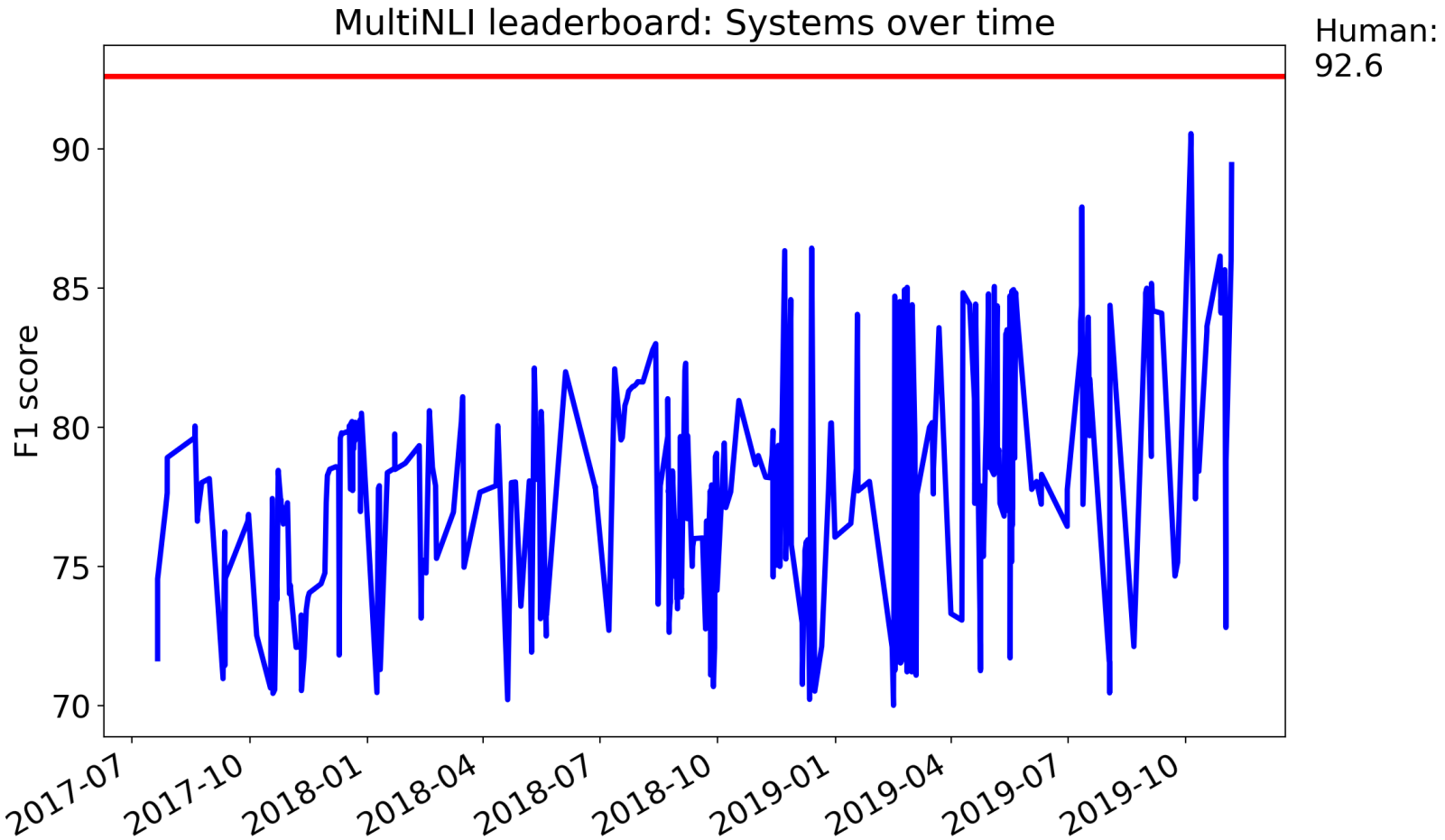
Rajpurkar et al. 2016

Stanford Natural Language Inference (SNLI)



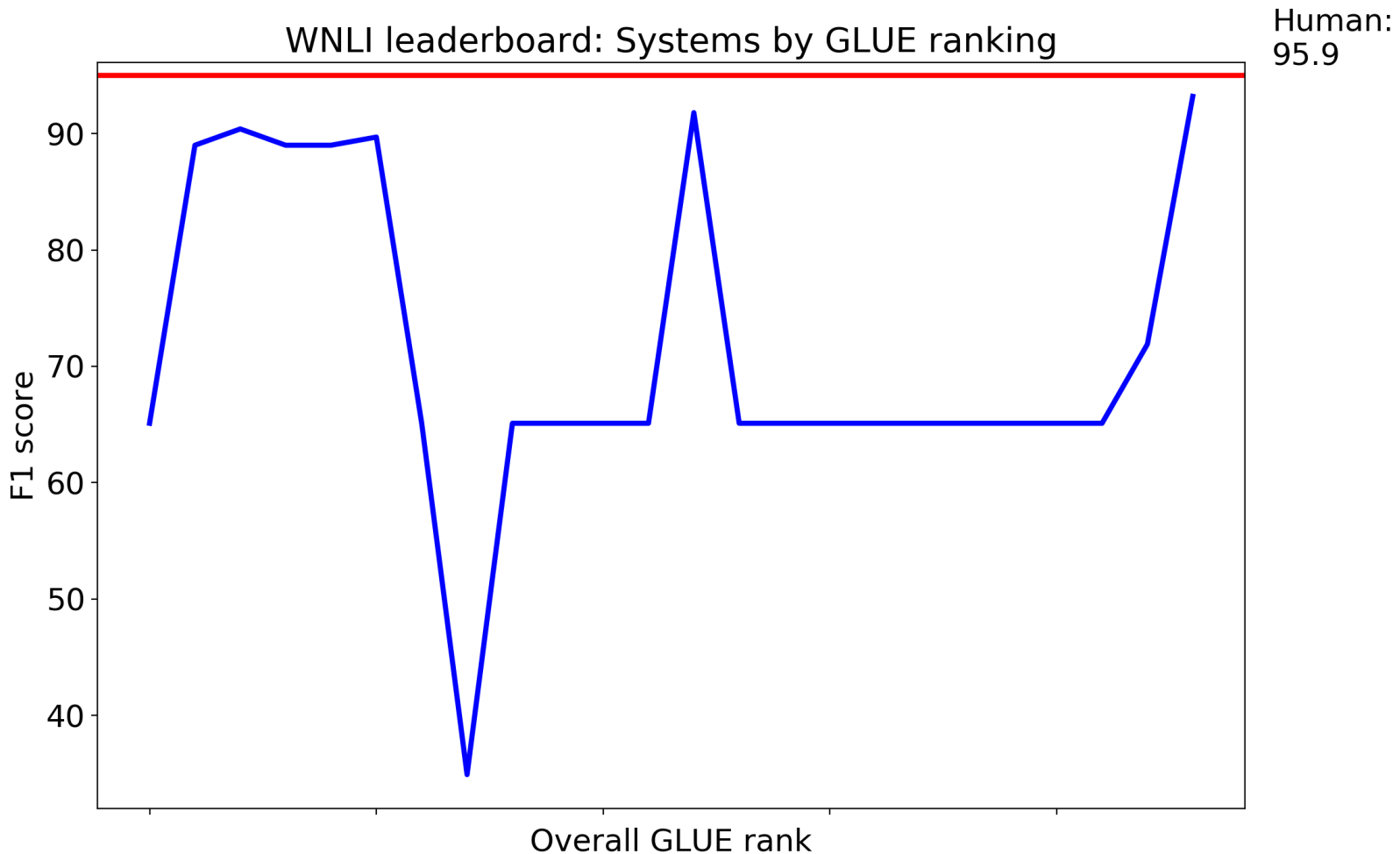
Bowman et al. 2015

MultiNLI



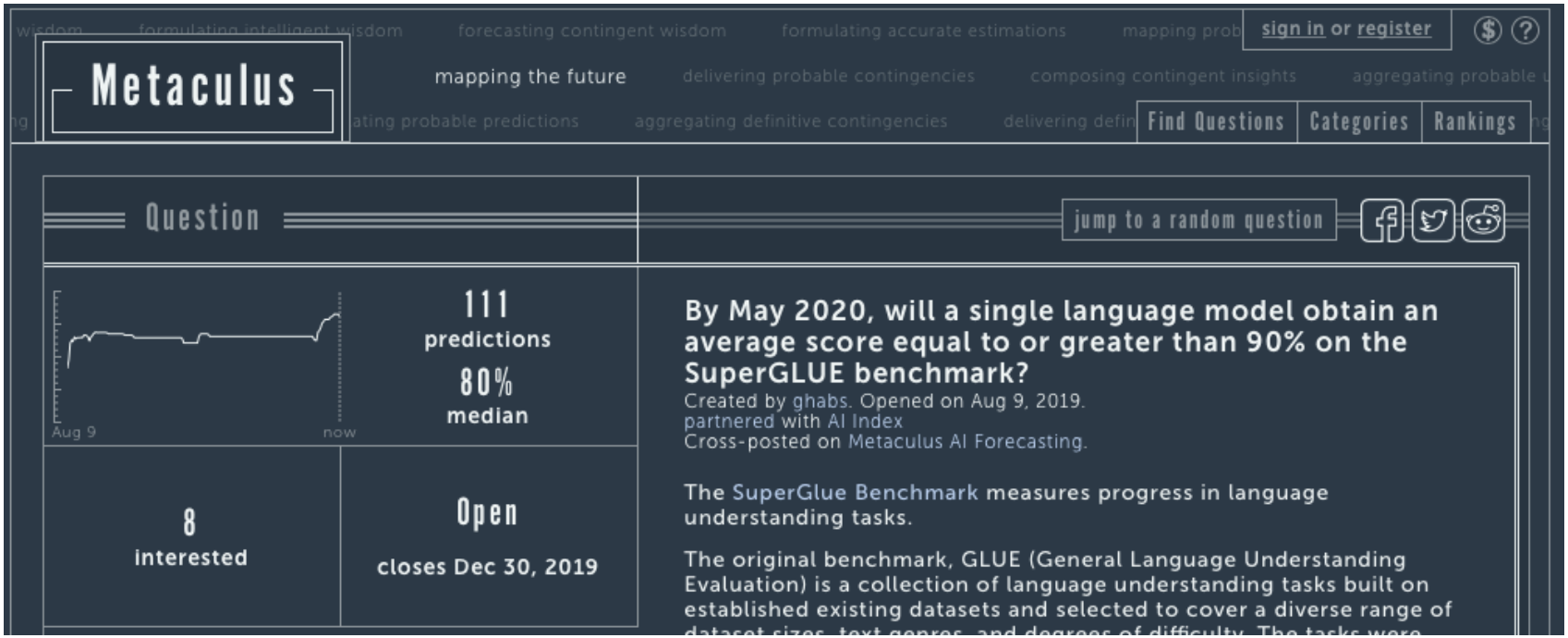
Williams et al. 2018

WinogradNLI

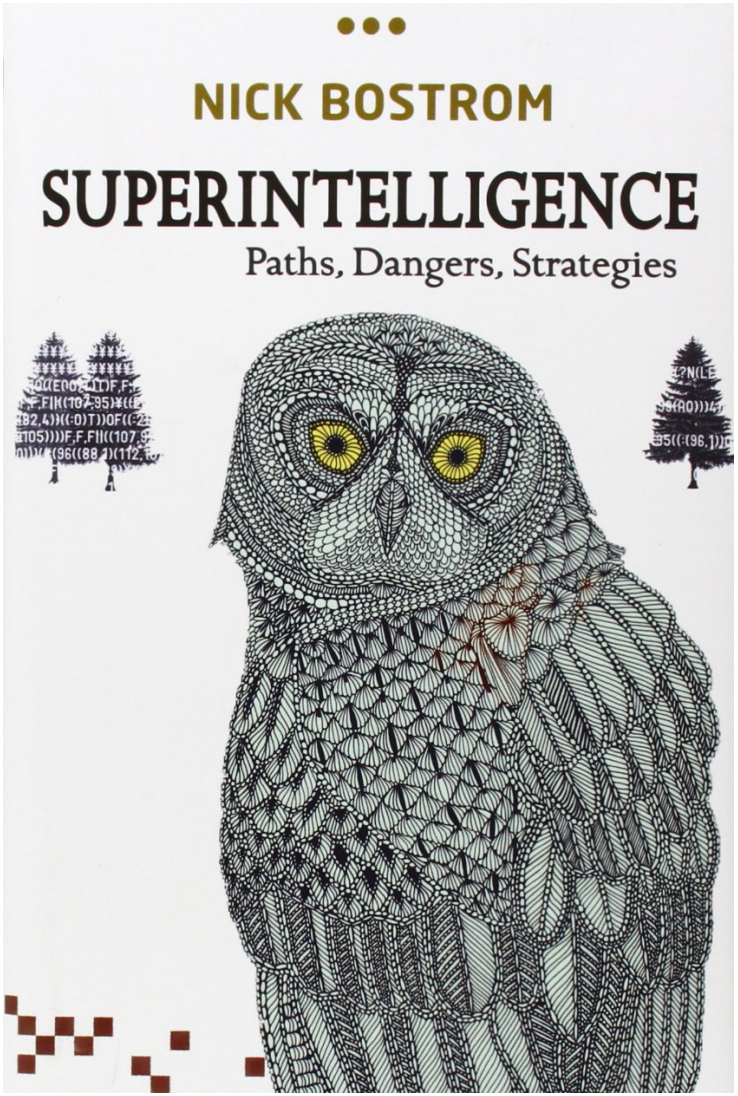


Wang et al. 2018

Forecasting



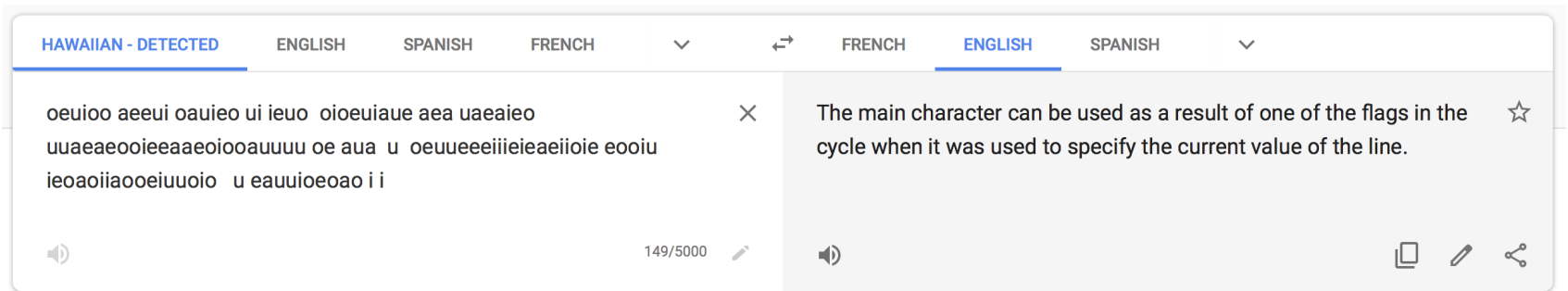
Human is 89.8. Current top score: 89.3


NICK BOSTROM
SUPERINTELLIGENCE
Paths, Dangers, Strategies

A peek behind the curtain

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

Translation: Garbage in, fluent text out?



Does Anne Hathaway News Drive Berkshire Hathaway's Stock?

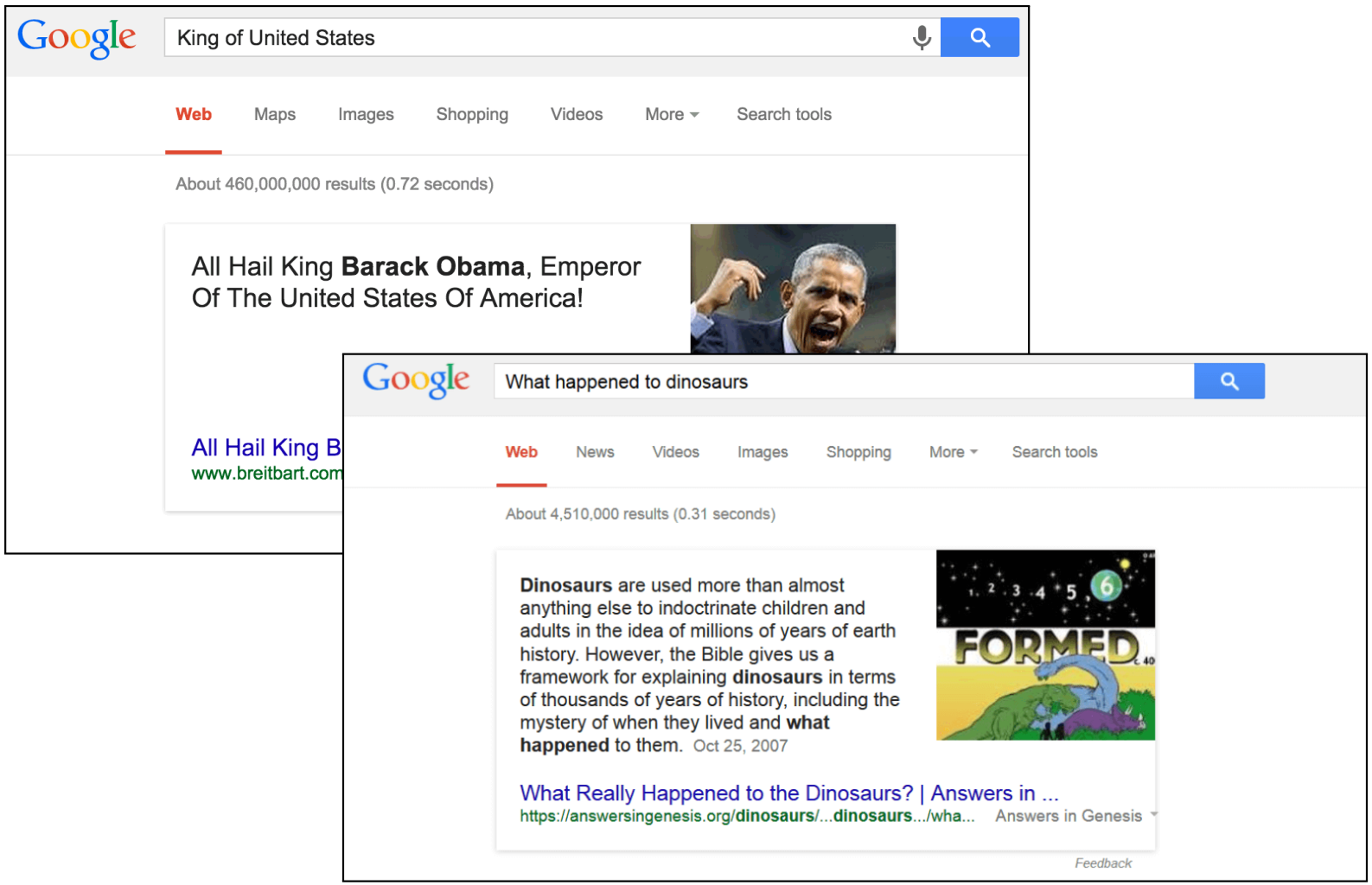
MAR 18 2011, 10:50 AM ET 28 Share 257 Tweet 471 +1 7 Recommend 616

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”

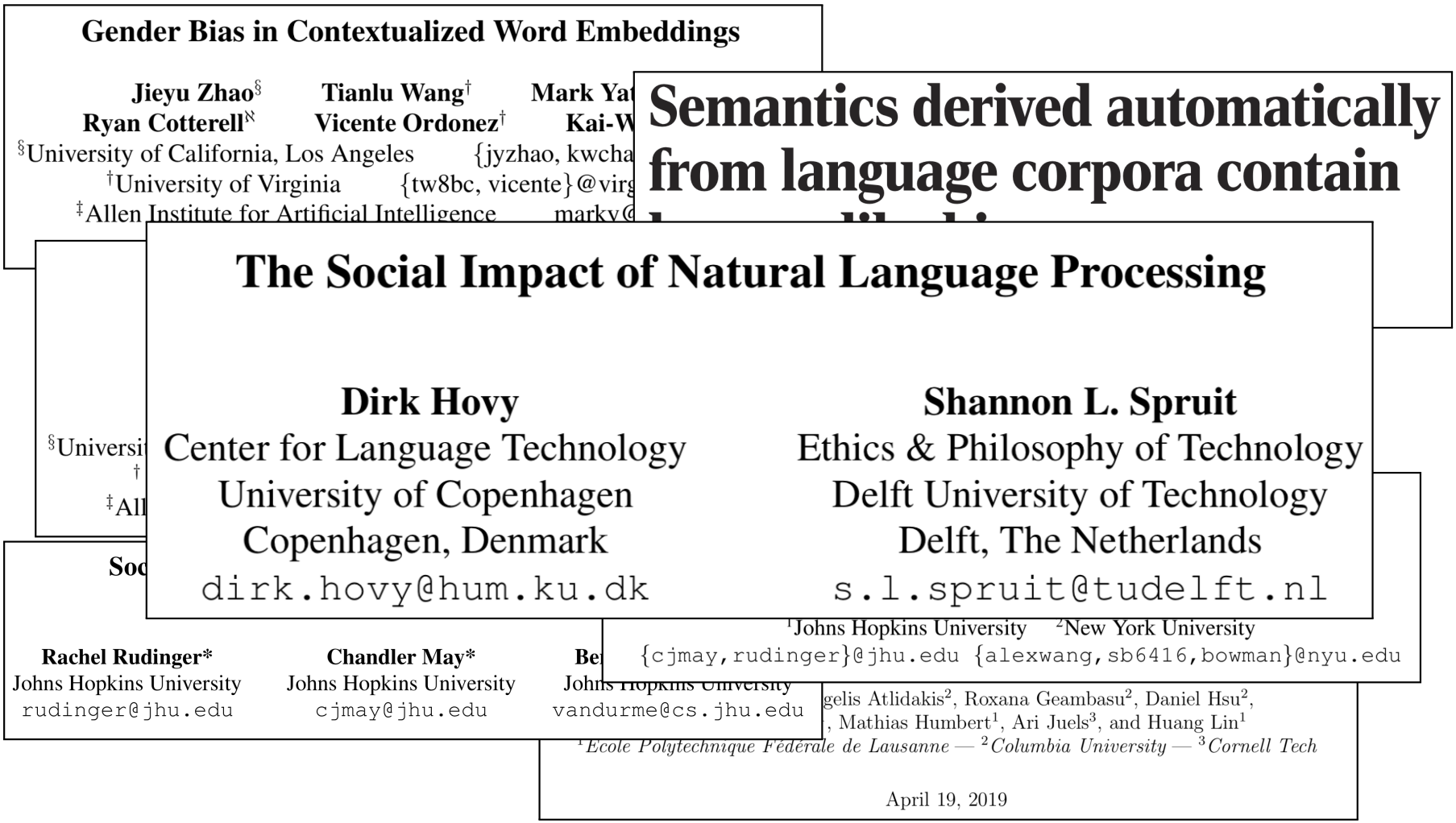
In 2008, when a newspaper accidentally republished a 2002 bankruptcy story, automated trading systems reacted in seconds, and $1B in market value evaporated within 12 minutes.

Misleading automatic curation



https://searchengineland.com

Bias perpetuation



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?

Jia and Liang 2017

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

Jia and Liang 2017

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 Jeff Dean 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

Jia and Liang 2017

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 Jeff Dean 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: Jeff Dean

Jia and Liang 2017

SQuAD adversarial testing

Passage

Quarterback Jeff Dean 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

Jia and Liang 2017

SQuAD adversarial testing

Passage

Quarterback Jeff Dean 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: Jeff Dean

Jia and Liang 2017

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.

Glockner et al. 2018

NLI adversarial testing

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

Nie et al. 2019

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?

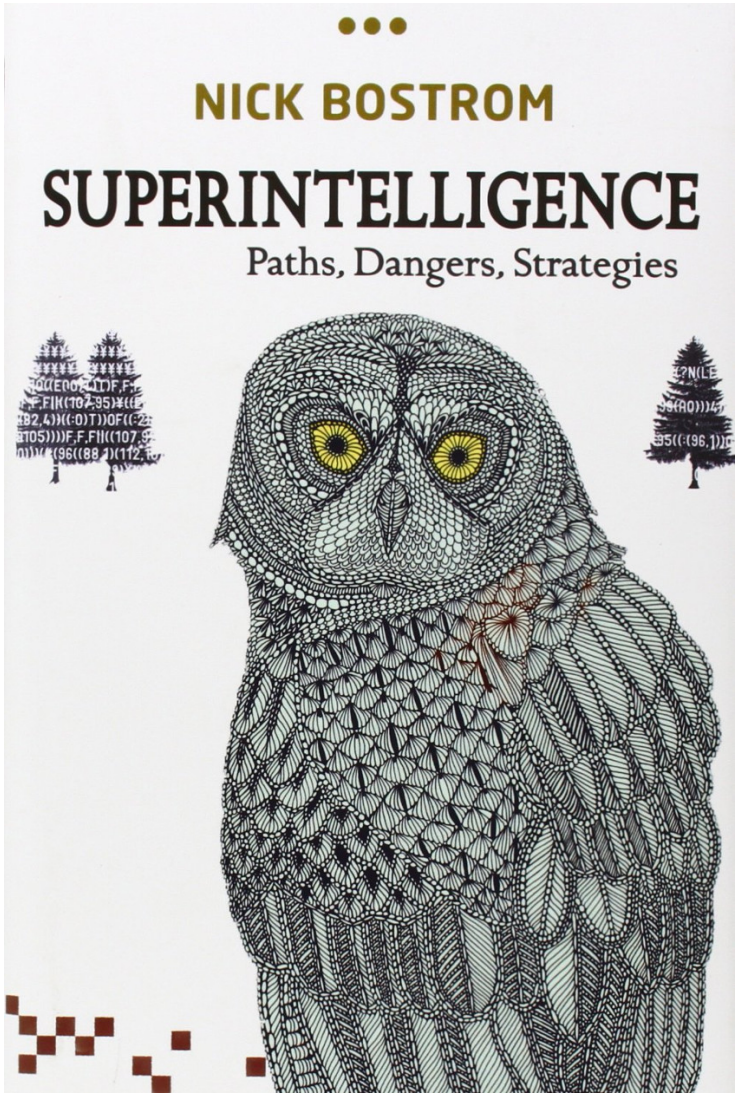
[. . .]

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

Two perspectives



Levesque (2013)

- “This paper is about the science of AI. Unfortunately, the technology of AI that gets all the attention.”
- “AI is the study of intelligent behaviour in computational terms.”
- “Should baseball players be allowed to glue small wings onto their caps?”
- “We need to return to our roots in Knowledge Representation and Reasoning for language and from language.”

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.

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

Assignments, bake-offs, and projects

High-level summary

Topics

1. Vector-space models
2. Sentiment analysis
3. Relation extraction
4. NLI
5. Grounding
6. Contextual word representations
7. Adversarial testing
8. Methods and metrics

Assignments/bake-offs

1. Word similarity
2. Relation extraction with distant supervision
3. Word-level entailment
4. Generating color descriptions in context

Final projects

1. Literature review
2. Experiment protocol
3. Short video presentation
4. Final paper

Assignments and bake-offs

1. There are four regular assignments. The first is due April 20, and they are weekly after that.
2. Each assignment culminates in a bake-off: 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 bake-off entries earn the additional point.
5. Winning bake-off 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/Bake-off: Word similarity

	against	age	agent	ages	ago	agree	ahead	ain't	air	aka	al
against	2003	90	39	20	88	57	33	15	58	22	24
age	90	1492	14	39	71	38	12	4	18	4	39
agent	39	14	507	2	21	5	10	3	9	8	25
ages	20	39	2	290	32	5	4	3	6	1	6
ago	88	71	21	32	1164	37	25	11	34	11	38
agree	57	38	5	5	37	627	12	2	16	19	14
ahead	33	12	10	4	25	12	429	4	12	10	7
ain't	15	4	3	3	11	2	4	166	0	3	3
air	58	18	9	6	34	16	12	0	746	5	11
aka	22	4	8	1	11	19	10	3	5	261	9
al	24	39	25	6	38	14	7	3	11	9	861

Assign/Bake-off: Word similarity

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

Assign/Bake-off: Word similarity

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

Dimensionality reduction
LSA
GloVe
word2vec
autoencoders
⋮

Assign/Bake-off: Word similarity

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

Dimensionality reduction
LSA
GloVe
word2vec
autoencoders
⋮

Vector comparison
Euclidean
Cosine
Dice
KL
⋮

Assign/Bake-off: Word similarity

sun	sunlight	50
automobile	car	50
river	water	49
food	gull	20
gate	hotel	20
dessert	head	7
born	hockey	7

Assign/Bake-off: Word similarity

Dataset	Pairs	Task-type	Best score	Paper
WordSim-353	353	Relatedness	82.8	Speer et al. 2017
MTurk-771	771	Relatedness	81.0	Speer et al. 2017
MEN	3,000	Relatedness	86.6	Speer et al. 2017
SimVerb-3500-dev	500	Similarity	61.1	Mrkšić et al. 2016
SimVerb-3500-test	3,000	Similarity	62.4	Mrkšić et al. 2016
And two held-out datasets for bake-off assessment				

Assign/Bake-off: Relation extraction

Obama was born in Honolulu, Hawaii

From 1964 to 1967, former President Barack Obama resided in Honolulu's Manoa neighborhood.

Barack Obama, the 44th president of the United States, was born on August 4, 1961 in Honolulu, Hawaii to Barack Obama, Sr., and Stanley Ann Dunham.

President Barack Obama holds hands with daughters Malia and Sasha during a family vacation in Honolulu.

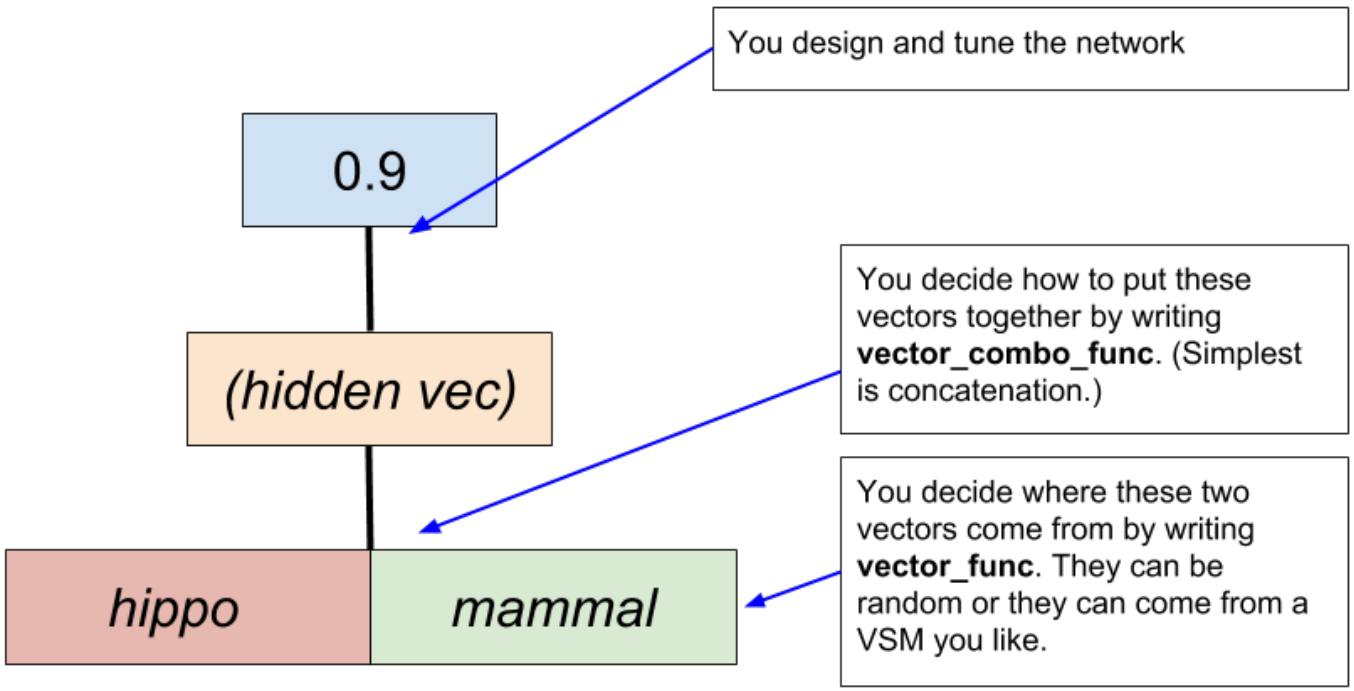
$$\begin{pmatrix} \textbf{Barack Obama} \\ R \\ \textbf{Honolulu} \end{pmatrix}$$

Assign/Bake-off: Word-level entailment

Train		
turtle	animal	1
turtle	desk	0
ingredient	element	1
pain	joint	0
	⋮	
Test		
dog	mammal	1
grenade	cycling	0
	⋮	

Train and test have disjoint *vocabs*.

Assign/Bake-off: Word-level entailment

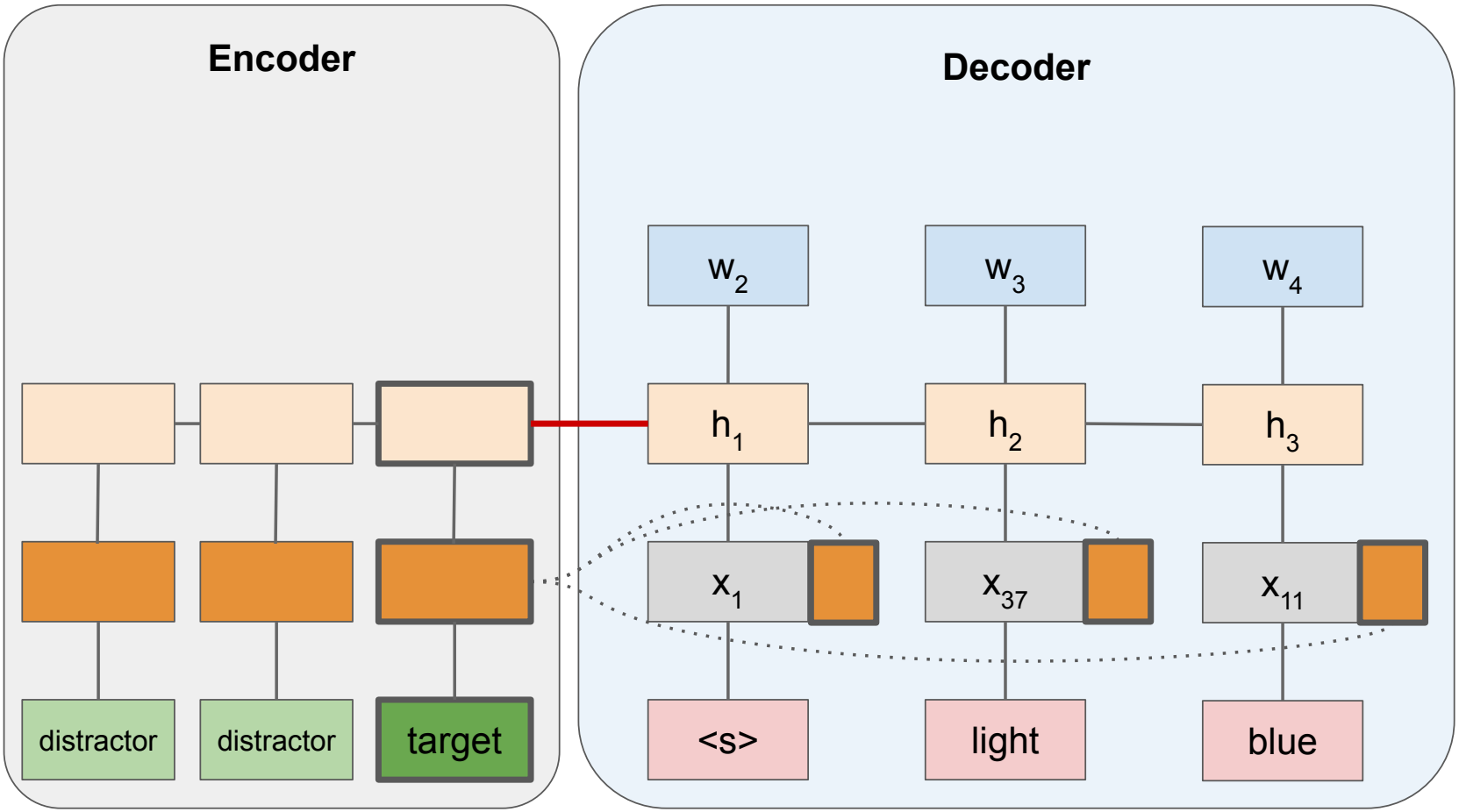


Assign/Bake-off: Contextual color describers

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

Monroe et al. 2017, 2018

Assign/Bake-off: Contextual color describers



Monroe et al. 2017, 2018

A note on grading original systems

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

1. Any system that performs extremely well on the bake-off 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 bake-off 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. Short video presentation
   d. Final paper
4. Exceptional final projects (and some videos) 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 brief history of NLU
2. A golden age for NLU
3. A peek behind the curtain
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://us.edstem.org/courses/326/discussion/

Gradescope
For submitting work; details sent out soon.

Teaching team
cs224u-spr1920-staff@lists.stanford.edu

Components

Participation	5%
Homeworks and bake-offs	30%
Literature review	10%
Experimental protocol	15%
Video presentation of project	10%
Final project paper	30%

An all-video course for 2020

Lectures

- Delivered by Zoom at the scheduled time.
- Discussion encouraged.
- Recorded and placed on Canvas shortly after.

Office hours

- All by Zoom.
- See the course Canvas for team members' scheduled times and Zoom links.

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`

The one and only quiz!

1. We will have exactly one required “quiz”.
2. The quiz is entirely devoted to course requirements and related details.
3. The sole purpose of the quiz is to create a clear incentive for you to study the website and understand your rights and obligations.
4. The quiz is administered on Canvas. You can take it as many times as you like – our goal is not to evaluate you but rather to ensure that you acquire this information.
5. It is due April 29 and cannot be turned in late. The quiz will be incorporated into your participation grade.

Take-home exam

The take-home exam is cancelled!

AWS credits

1. Thanks to AWS Educate, we can provide every enrolled student with a $100 AWS credit.
2. All members of winning bake-off teams will receive additional $100 credits as prizes.
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. Consider doing the quiz as a way of getting to know your rights and obligations for this course.

3. Start working with `vsm_01_distributional.ipynb`. If this material is new to you, consider watching the associated screencasts (linked from the course site).

4. For corresponding with the teaching team: cs224u-spr1920-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.

5. Next time: vector space models of meaning!

References I

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.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Max Glockner, Vered Shwartz, 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.

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.

Hector J. Levesque. 2013. On our best behaviour. In *Proceedings of the Twenty-third International Conference on Artificial Intelligence*, Beijing.

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.

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.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148. 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.

References II

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.

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.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 4444–4451. AAAI Press.

Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. FairTest: Discovering wnwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416. IEEE.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. 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.

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 North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.