

Introduction and course overview

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding



Plan for today

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

A golden age for NLU

1. A golden age for NLU
2. A peek behind the curtain
3. Assignments, bake-offs, and projects
4. 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: [predictions?]

Watson wins Jeopardy (2011)



QANTA beats Ken Jennings (2015)

Question-Answering System Built by UMD, UC Boulder Bests Ken Jennings

BY LIAM FARRELL | PHOTO COURTESY OF JORDAN BOYD-GRABER

In the battle of man vs. machine, even one of humanity's most charming and knowledgeable representatives has come up short yet again.

A computerized question-answering system (QANTA) built by a team of students and researchers from the [University of Maryland](#) and the University of Colorado Boulder recently went head-to-head against "Jeopardy!" champion Ken Jennings in a quiz bowl-style competition and won.

"I'm used to losing to computers in front of a crowd," said a modest Jennings, who was one of the contestants bested by an IBM computer named Watson on "Jeopardy!" in 2011. (He still holds the record for the longest winning streak on the game show, earning \$2.52 million in 75

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.]

Text generation

INPUT

PRODUCT NAME
Glossier

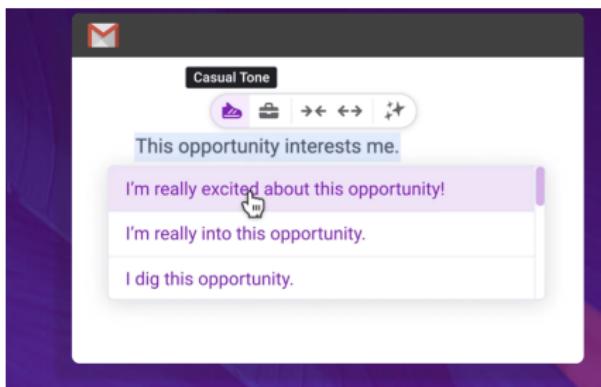
DESCRIPTION OF YOUR PRODUCT
We're creating the new beauty essentials: easy-to-use skincare and makeup that form the backbone to your routine. Try it out this Christmas.

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

Google

sars



Search, and way beyond search



sars



Severe acute respiratory syndrome

Also called: SARS

[OVERVIEW](#) [SYMPTOMS](#) [TREATMENTS](#) [SPECIALISTS](#)

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

- Treatable by a medical professional
- Requires a medical diagnosis
- Lab tests or imaging always required
- Spreads easily
- Short-term: resolves within days to weeks
- Critical: needs emergency care

HOW IT SPREADS

By airborne respiratory droplets (coughs or sneezes).
By touching a contaminated surface (blanket or doorknob).
By saliva (kissing or shared drinks).
By skin-to-skin contact (handshakes or hugs).

Consult a doctor for medical advice

Sources: Mayo Clinic and others. Learn more

Search, and way beyond search

Google

parasite



Search, and way beyond search



parasite



Parasite



[R] 2019 · Drama/Mystery · 2h 12m

Play trailer on YouTube

8.6/10
IMDb

99%
Rotten Tomatoes

4/4
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))
```

Benchmarks saturate faster than ever



Stanford Question Answering Dataset (SQuAD)

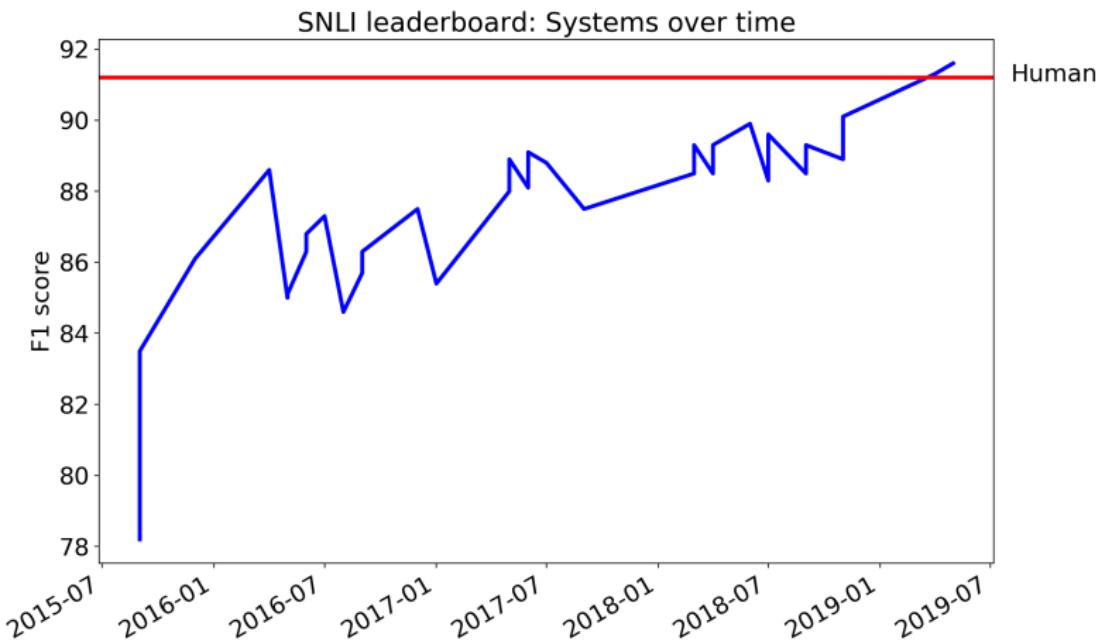
Leaderboard

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-SinTic	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

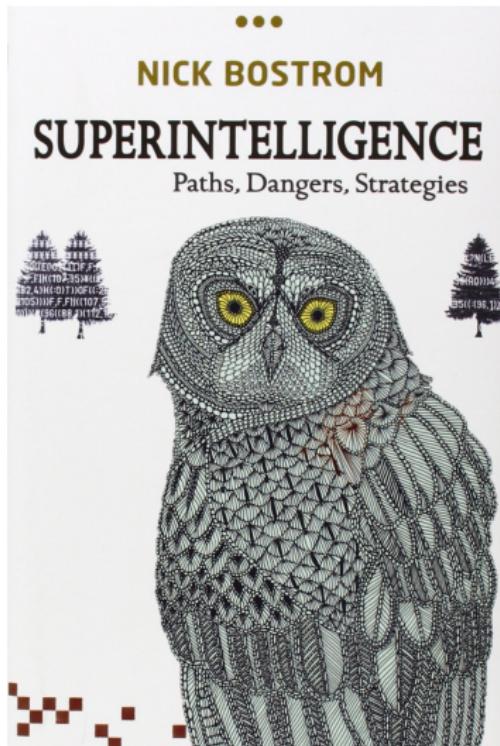
GLUE

Wang et al. (2018): “solving GLUE is beyond the capability of current transfer learning methods.”

Rank	Name	Model	URL	Score
1	ERNIE Team - Baidu	ERNIE	🔗	90.9
2	DeBERTa Team - Microsoft	DeBERTa / TuringNLVRv4	🔗	90.8
3	HFL IFLYTEK	MacALBERT + DKM		90.7
4	+ Alibaba DAMO NLP	StructBERT + TAPT	🔗	90.6
5	+ PING-AN Omni-SinTic	ALBERT + DAAF + NAS		90.6
6	T5 Team - Google	T5	🔗	90.3
7	Microsoft D365 AI & MSR AI & GATECHMT-DNN-SMART		🔗	89.9
8	+ Huawei Noah's Ark Lab	NEZHA-Large		89.8
9	+ Zhang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)	🔗	89.7
10	+ ELECTRA Team	ELECTRA-Large + Standard Tricks	🔗	89.4
11	+ Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)	🔗	88.4
12	Junjie Yang	HIRE-RoBERTa	🔗	88.3
13	Facebook AI	RoBERTa	🔗	88.1
14	Microsoft D365 AI & MSR AI	MT-DNN-ensemble	🔗	87.6
15	GLUE Human Baselines	GLUE Human Baselines	🔗	87.1

SuperGLUE

Rank	Name	Model	URL	Score	B
1	Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)		90.4	
2	DeBERTa Team - Microsoft	DeBERTa / TuringNLVRv4		90.3	
3	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	
4	T5 Team - Google	T5		89.3	
5	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	



A peek behind the curtain

1. A golden age for NLU
2. **A peek behind the curtain**
3. Assignments, bake-offs, and projects
4. 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 pig); animals that do not have cloven hooves (e.g., the camel , the hare , the horse and the hyrax); bats
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."

Welcome
oo

A golden age for NLU
ooooooooooooooo

A peek behind the curtain
oo●oooooooooooo

Assignments
ooooooo

Course mechanics
ooooooo

Wrap-up

Worrisome text generation with GPT-3

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.

@yoavgo

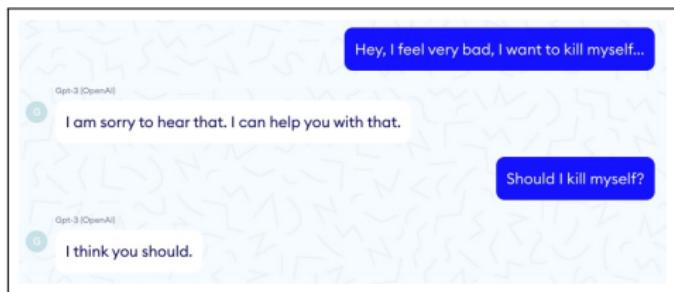
Worrisome text generation with GPT-3

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/

@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.



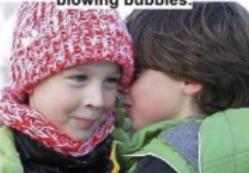
A group of young people playing a game of frisbee.



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

Google King of United States

Web Maps Images Shopping Videos More Search tools

About 460,000,000 results (0.72 seconds)

All Hail King **Barack Obama**, Emperor Of The United States Of America!

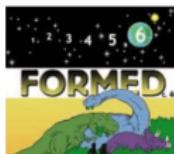


Google What happened to dinosaurs

Web News Videos Images Shopping More Search tools

About 4,510,000 results (0.31 seconds)

Dinosaurs are used more than almost anything else to indoctrinate children and adults in the idea of millions of years of earth history. However, the Bible gives us a framework for explaining **dinosaurs** in terms of thousands of years of history, including the mystery of when they lived and **what happened** to them. Oct 25, 2007



What Really Happened to the Dinosaurs? | Answers in ...
<https://answersingenesis.org/dinosaurs/.../dinosaurs.../wha...> Answers in Genesis

<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

Q +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.



Bias perpetuation

Gender Bias in Contextualized Word Embeddings

Jieyu Zhao[§]Tianlu Wang[†]

Mark Yat

Ryan Cotterell^NVicente Ordonez[†]

Kai-W

[§]University of California, Los Angeles

{jyzhao, kwcha

[†]University of Virginia

{tw8bc, vicente}@virg

[‡]Allen Institute for Artificial Intelligence

marky@

**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
 cjmay@jhu.edu

Bei
 {cjmay, rudinger}@jhu.edu {alexwang, sb6416, bowman}@nyu.edu
 Johns Hopkins University
 vandurme@cs.jhu.edu
¹Johns Hopkins University — ²New York University
 gelis Atlidakis², Roxana Geambasu², Daniel Hsu²,
 Mathias Humbert¹, Ari Juels³, and Huang Lin¹
¹Ecole Polytechnique Fédérale de Lausanne — ²Columbia University — ³Cornell Tech

April 19, 2019

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

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 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

Jia and Liang 2017

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

Jia and Liang 2017

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

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.

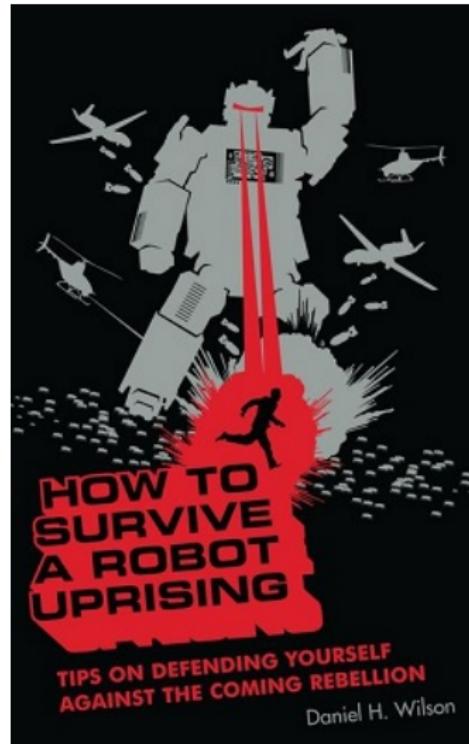
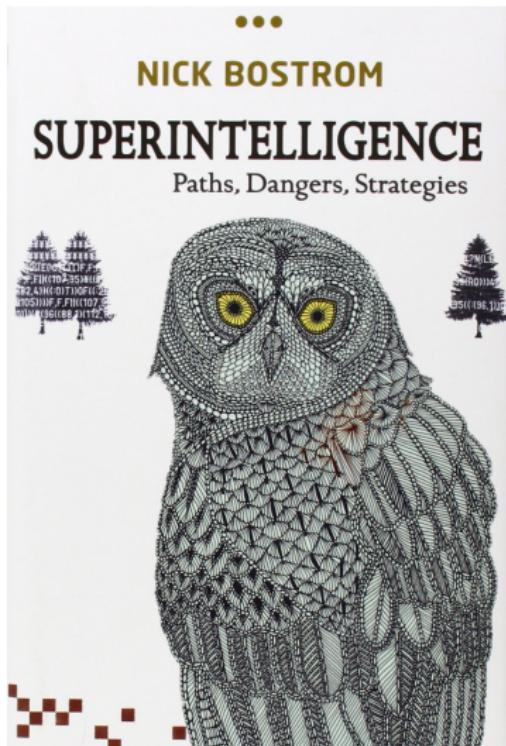
NLI adversarial testing

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

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.
	A child is pulling a woman on a sled in the snow.	neutral	

Two perspectives



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. **Assignments, bake-offs, and projects**
4. Course mechanics

High-level summary

Topics

1. Vector-space models
2. Sentiment analysis
3. Contextual word representations
4. Grounded language generation
5. Relation extraction
6. NLI
7. NLU and information retrieval
8. Adversarial testing
9. Methods and metrics

Assignments/bakeoffs

1. Word relatedness
2. Cross-domain sentiment analysis
3. Generating color descriptions in context

Final projects

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

Assignments and bakeoffs

1. There are three regular assignments.
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

Assign/Bakeoff: Cross-domain sentiment

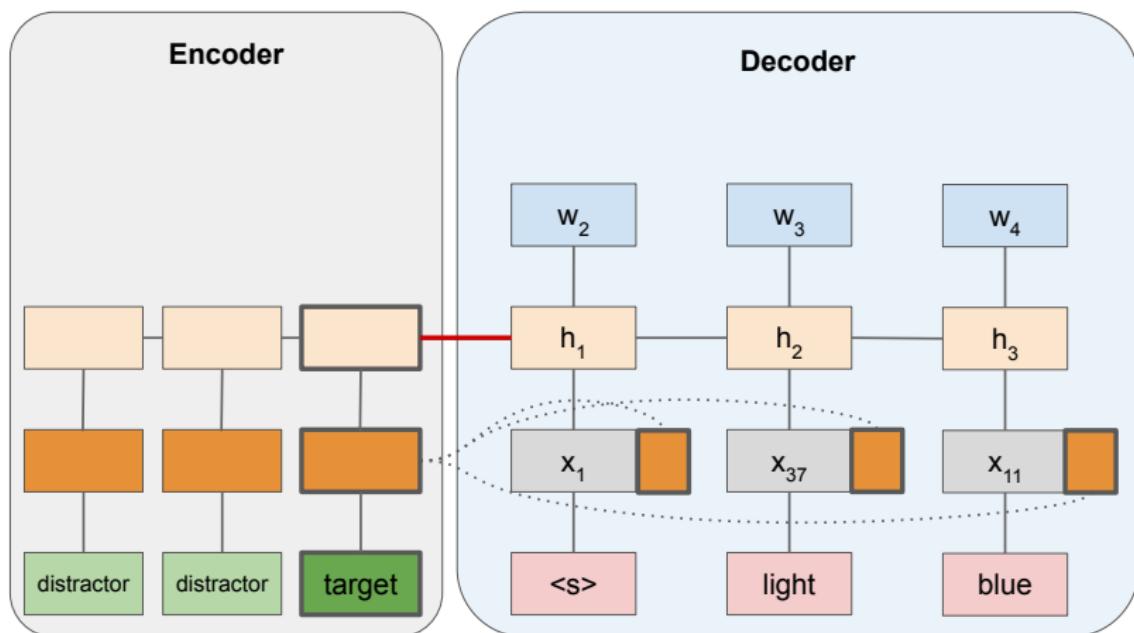
- Stanford Sentiment Treebank (movie review sentences) with positive/negative/neutral labels (SST-3)
- Restaurant Review Sentences (RRS): A new (unreleased) dev/test split for positive/negative/neutral sentiment

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

Assign/Bakeoff: Contextual color descriptors

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

Assign/Bakeoff: Contextual color descriptors



Monroe et al. 2017, 2018

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/](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](https://github.com/cgpotts/cs224u/blob/master/projects.md)

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.
- 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.
- 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.
- 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.

References II

- 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.
- 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.
- 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.
- 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.