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](https://pages.semanticscholar.org/coronavirus-research)

- **Elsevier Coronavirus Research Repository:**
  [https://coronavirus.1science.com/](https://coronavirus.1science.com/)

- **Coronavirus Tweets:**

- **CS472 Data science and AI for COVID-19**

- **Google’s COVID-19 Public Datasets**
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
Welcome

A brief history of NLU

A golden age for NLU

A peek behind the curtain

Assignments

Course mechanics

Wrap-up

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

Stanford NLP class enrollment

- CS224D (DL NLP)
- CS224S (speech)
- CS276 (IR)
- CS224U (NLU)
- CS124 (undergrad HLT)
- CS224N (NLP)

h/t @StanfordNLP
The history of CS224u enrollments

Stanford NLP class enrollment

CS224D (DL NLP)
CS224S (speech)
CS276 (IR)
CS224U (NLU)
CS124 (ugrad HLT)
CS224N (NLP)

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

“Hey Siri, turn on the lights in the living room”
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
Interrogé sur le sujet, un responsable de l'administration américaine a répondu: “Les États-Unis ne mènent pas de surveillance électronique à destination des bureaux de la Banque mondiale et du FMI à Washington”.

Interrogé à ce sujet, un responsable de l'administration américaine a répondu: "The United States is not conducting electronic surveillance aimed at offices of the World Bank and IMF in Washington."
Search, and way beyond search
Search, and way beyond search

Severe acute respiratory syndrome
Also called: SARS

OVERVIEW

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
Search, and way beyond search
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.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model Description</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Retro-Reader on ALBERT (ensemble)</td>
<td>Stanford University</td>
<td>90.115</td>
<td>92.580</td>
</tr>
<tr>
<td>2</td>
<td>ALBERT + DAAAF + Verifier (ensemble)</td>
<td>PINGAN Omni-Sinitic</td>
<td>90.002</td>
<td>92.425</td>
</tr>
<tr>
<td>3</td>
<td>ALBERT (ensemble model)</td>
<td>Google Research &amp; TTIC</td>
<td>89.731</td>
<td>92.215</td>
</tr>
<tr>
<td>3</td>
<td>Albert_Verifier_AA_Net (ensemble)</td>
<td>QIANXIN</td>
<td>89.743</td>
<td>92.180</td>
</tr>
<tr>
<td>4</td>
<td>albert+transform+verify (ensemble)</td>
<td>qianxin</td>
<td>89.528</td>
<td>92.059</td>
</tr>
<tr>
<td>13</td>
<td>RoBERTa+Verify (single model)</td>
<td>CW</td>
<td>86.448</td>
<td>89.586</td>
</tr>
<tr>
<td>13</td>
<td>BERT + ConvLSTM + MTL + Verifier (ensemble)</td>
<td>Layer 6 AI</td>
<td>86.730</td>
<td>89.286</td>
</tr>
</tbody>
</table>

Rajpurkar et al. 2016
Stanford Natural Language Inference (SNLI)

SNLI leaderboard: Systems over time

F1 score


Human

Bowman et al. 2015
MultiNLI

MultiNLI leaderboard: Systems over time

Human: 92.6

Williams et al. 2018
WinogradNLI

WNLI leaderboard: Systems by GLUE ranking

Human: 95.9

Wang et al. 2018
### Forecasting

<table>
<thead>
<tr>
<th>Question</th>
<th>Predictions</th>
<th>Open</th>
<th>Close</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>80% median</td>
<td>Aug 9</td>
<td>Dec 30, 2019</td>
</tr>
</tbody>
</table>

By May 2020, will a single language model obtain an average score equal to or greater than 90% on the **SuperGLUE benchmark**?

**Forecast:**
- **111 predictions**
- **80% median**

**Status:**
- **8 interested**
- **Open**
- Closes Dec 30, 2019

**Human is 89.8. Current top score: 89.3**
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?

The main character can be used as a result of one of the flags in the cycle when it was used to specify the current value of the line.
Does Anne Hathaway News Drive Berkshire Hathaway's Stock?

MAR 18 2011, 10:50 AM ET 28

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

Gender Bias in Contextualized Word Embeddings

Jieyu Zhao
Tianlu Wang
Mark Yae
Ryan Cotterell
Vicente Ordonez
Kai-Wei

University of California, Los Angeles
{ jyzhao, kwchan
University of Virginia
{ tw8bc, vicente } @ virg
Allen Institute for Artificial Intelligence
marky@

Semantics derived automatically
from language corpora contain
bias.

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
Bet
Johns Hopkins University
vandurme@cs.jhu.edu

gelis Atlidakis 2, Roxana Geambasu 2, Daniel Hsu 2,
Mathias Humbert 1, Ari Juels 3, and Huang Lin 1

1Ecole Polytechnique Fédérale de Lausanne — 2Columbia University — 3CornellTech

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?

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
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

<table>
<thead>
<tr>
<th>System</th>
<th>Original</th>
<th>Adversarial</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReasoNet-E</td>
<td>81.1</td>
<td>39.4</td>
</tr>
<tr>
<td>SEDT-E</td>
<td>80.1</td>
<td>35.0</td>
</tr>
<tr>
<td>BiDAF-E</td>
<td>80.0</td>
<td>34.2</td>
</tr>
<tr>
<td>Mnemonic-E</td>
<td>79.1</td>
<td>46.2</td>
</tr>
<tr>
<td>Ruminating</td>
<td>78.8</td>
<td>37.4</td>
</tr>
<tr>
<td>jNet</td>
<td>78.6</td>
<td>37.9</td>
</tr>
<tr>
<td>Mnemonic-S</td>
<td>78.5</td>
<td>46.6</td>
</tr>
<tr>
<td>ReasoNet-S</td>
<td>78.2</td>
<td>39.4</td>
</tr>
<tr>
<td>MPCM-S</td>
<td>77.0</td>
<td>40.3</td>
</tr>
<tr>
<td>SEDT-S</td>
<td>76.9</td>
<td>33.9</td>
</tr>
<tr>
<td>RaSOR</td>
<td>76.2</td>
<td>39.5</td>
</tr>
<tr>
<td>BiDAF-S</td>
<td>75.5</td>
<td>34.3</td>
</tr>
<tr>
<td>Match-E</td>
<td>75.4</td>
<td>29.4</td>
</tr>
<tr>
<td>Match-S</td>
<td>71.4</td>
<td>27.3</td>
</tr>
<tr>
<td>DCR</td>
<td>69.4</td>
<td>37.8</td>
</tr>
<tr>
<td>Logistic</td>
<td>50.4</td>
<td>23.2</td>
</tr>
</tbody>
</table>
## SQuAD adversarial testing

<table>
<thead>
<tr>
<th>System</th>
<th>Original Rank</th>
<th>Adversarial Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReasoNet-E</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>SEDT-E</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>BiDAF-E</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Mnemonic-E</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Ruminating</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>jNet</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Mnemonic-S</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>ReasoNet-S</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>MPCM-S</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>SEDT-S</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>RaSOR</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>BiDAF-S</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>Match-E</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>Match-S</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>DCR</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>Logistic</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>
# NLI adversarial testing

<table>
<thead>
<tr>
<th>Premise</th>
<th>Relation</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A turtle danced.</td>
<td>entails</td>
<td>A turtle moved.</td>
</tr>
<tr>
<td>Every reptile danced.</td>
<td>neutral</td>
<td>A turtle ate.</td>
</tr>
<tr>
<td>Some turtles walk.</td>
<td>contradicts</td>
<td>No turtles move.</td>
</tr>
</tbody>
</table>
## NLI adversarial testing

<table>
<thead>
<tr>
<th>Premise</th>
<th>Relation</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A little girl kneeling in the dirt crying.</td>
<td>entails</td>
<td>A little girl is very sad.</td>
</tr>
<tr>
<td>A little girl is very unhappy.</td>
<td>entails</td>
<td>A little girl is very unhappy.</td>
</tr>
</tbody>
</table>

Glockner et al. 2018
## NLI adversarial testing

<table>
<thead>
<tr>
<th>Premise</th>
<th>Relation</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A woman is pulling a child on a sled in the snow.</td>
<td>entails</td>
<td>A child is sitting on a sled in the snow.</td>
</tr>
<tr>
<td>A child is pulling a woman on a sled in the snow.</td>
<td>neutral</td>
<td></td>
</tr>
</tbody>
</table>

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

- **Nick Bostrom**
  - *Superintelligence: Paths, Dangers, Strategies*

- **Daniel H. Wilson**
  - *How to Survive a Robot Uprising: Tips on Defending Yourself Against the Coming Rebellion*
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

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
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

<table>
<thead>
<tr>
<th></th>
<th>against</th>
<th>age</th>
<th>agent</th>
<th>ages</th>
<th>ago</th>
<th>agree</th>
<th>ahead</th>
<th>ain’t</th>
<th>air</th>
<th>aka</th>
<th>al</th>
</tr>
</thead>
<tbody>
<tr>
<td>against</td>
<td>2003</td>
<td>90</td>
<td>39</td>
<td>20</td>
<td>88</td>
<td>57</td>
<td>33</td>
<td>15</td>
<td>58</td>
<td>22</td>
<td>24</td>
</tr>
<tr>
<td>age</td>
<td>90</td>
<td>1492</td>
<td>14</td>
<td>39</td>
<td>71</td>
<td>38</td>
<td>12</td>
<td>4</td>
<td>18</td>
<td>4</td>
<td>39</td>
</tr>
<tr>
<td>agent</td>
<td>39</td>
<td>14</td>
<td>507</td>
<td>2</td>
<td>21</td>
<td>5</td>
<td>10</td>
<td>3</td>
<td>9</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>ages</td>
<td>20</td>
<td>39</td>
<td>2</td>
<td>290</td>
<td>32</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>ago</td>
<td>88</td>
<td>71</td>
<td>21</td>
<td>32</td>
<td>1164</td>
<td>37</td>
<td>25</td>
<td>11</td>
<td>34</td>
<td>11</td>
<td>38</td>
</tr>
<tr>
<td>agree</td>
<td>57</td>
<td>38</td>
<td>5</td>
<td>5</td>
<td>37</td>
<td>627</td>
<td>12</td>
<td>2</td>
<td>16</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>ahead</td>
<td>33</td>
<td>12</td>
<td>10</td>
<td>4</td>
<td>25</td>
<td>12</td>
<td>429</td>
<td>4</td>
<td>12</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>ain’t</td>
<td>15</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>11</td>
<td>2</td>
<td>4</td>
<td>166</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>air</td>
<td>58</td>
<td>18</td>
<td>9</td>
<td>6</td>
<td>34</td>
<td>16</td>
<td>12</td>
<td>0</td>
<td>746</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>aka</td>
<td>22</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>11</td>
<td>19</td>
<td>10</td>
<td>3</td>
<td>5</td>
<td>261</td>
<td>9</td>
</tr>
<tr>
<td>al</td>
<td>24</td>
<td>39</td>
<td>25</td>
<td>6</td>
<td>38</td>
<td>14</td>
<td>7</td>
<td>3</td>
<td>11</td>
<td>9</td>
<td>861</td>
</tr>
</tbody>
</table>
Assign/Bake-off: Word similarity

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

; ;

...
Assign/Bake-off: Word similarity

<table>
<thead>
<tr>
<th>Reweighting</th>
<th>Dimensionality reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>probabilities</td>
<td>LSA</td>
</tr>
<tr>
<td>length norm.</td>
<td>GloVe</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>word2vec</td>
</tr>
<tr>
<td>O/E</td>
<td>autoencoders</td>
</tr>
<tr>
<td>PMI</td>
<td></td>
</tr>
<tr>
<td>Positive PMI</td>
<td></td>
</tr>
</tbody>
</table>
## Assign/Bake-off: Word similarity

<table>
<thead>
<tr>
<th>Reweighting</th>
<th>Dimensionality reduction</th>
<th>Vector comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probabilities</td>
<td>LSA</td>
<td>Euclidean</td>
</tr>
<tr>
<td>Length norm.</td>
<td>GloVe</td>
<td>Cosine</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>word2vec</td>
<td>Dice</td>
</tr>
<tr>
<td>O/E</td>
<td>autoencoders</td>
<td>KL</td>
</tr>
<tr>
<td>PMI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Assign/Bake-off: Word similarity

<table>
<thead>
<tr>
<th>word</th>
<th>similar word</th>
<th>similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>sunlight</td>
<td>50</td>
</tr>
<tr>
<td>automobile</td>
<td>car</td>
<td>50</td>
</tr>
<tr>
<td>river</td>
<td>water</td>
<td>49</td>
</tr>
<tr>
<td>food</td>
<td>gull</td>
<td>20</td>
</tr>
<tr>
<td>gate</td>
<td>hotel</td>
<td>20</td>
</tr>
<tr>
<td>dessert</td>
<td>head</td>
<td>7</td>
</tr>
<tr>
<td>born</td>
<td>hockey</td>
<td>7</td>
</tr>
</tbody>
</table>
### Assign/Bake-off: Word similarity

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pairs</th>
<th>Task-type</th>
<th>Best score</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordSim-353</td>
<td>353</td>
<td>Relatedness</td>
<td>82.8</td>
<td>Speer et al. 2017</td>
</tr>
<tr>
<td>MTurk-771</td>
<td>771</td>
<td>Relatedness</td>
<td>81.0</td>
<td>Speer et al. 2017</td>
</tr>
<tr>
<td>MEN</td>
<td>3,000</td>
<td>Relatedness</td>
<td>86.6</td>
<td>Speer et al. 2017</td>
</tr>
<tr>
<td>SimVerb-3500-dev</td>
<td>500</td>
<td>Similarity</td>
<td>61.1</td>
<td>Mrkšić et al. 2016</td>
</tr>
<tr>
<td>SimVerb-3500-test</td>
<td>3,000</td>
<td>Similarity</td>
<td>62.4</td>
<td>Mrkšić et al. 2016</td>
</tr>
</tbody>
</table>

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.
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

You design and tune the network

You decide how to put these vectors together by writing `vector_combo_func`. (Simplest is concatenation.)

You decide where these two vectors come from by writing `vector_func`. They can be random or they can come from a VSM you like.

0.9

(hidden vec)

hippo mammal
### Assign/Bake-off: Contextual color describers

<table>
<thead>
<tr>
<th>Context</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Context" /></td>
<td>blue</td>
</tr>
<tr>
<td><img src="image2" alt="Context" /></td>
<td>The darker blue one</td>
</tr>
<tr>
<td><img src="image3" alt="Context" /></td>
<td>dull pink not the super bright one</td>
</tr>
<tr>
<td><img src="image4" alt="Context" /></td>
<td>Purple</td>
</tr>
<tr>
<td><img src="image5" alt="Context" /></td>
<td>blue</td>
</tr>
</tbody>
</table>

*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

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation</td>
<td>5%</td>
</tr>
<tr>
<td>Homeworks and bake-offs</td>
<td>30%</td>
</tr>
<tr>
<td>Literature review</td>
<td>10%</td>
</tr>
<tr>
<td>Experimental protocol</td>
<td>15%</td>
</tr>
<tr>
<td>Video presentation of project</td>
<td>10%</td>
</tr>
<tr>
<td>Final project paper</td>
<td>30%</td>
</tr>
</tbody>
</table>
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


References


