Introduction and course overview

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding
Welcome
CS224u: hybrid, asynchronous, hands-on

- Core course content in screencasts on YouTube and linked from the homepage, with accompanying Jupyter notebook for hands-on work.

- A series of special events: conversations with prominent NLU researchers (details later in this lecture). Mostly on Zoom. Attend live or listen later.

- Other class meetings: optional open discussions and/or spaces for you to work, with the teaching team there to help. Open to mixing in-classroom and Zoom formats.

- Office hours offered in person and on Zoom. Details to come.

- Continuous evaluation: three assignments, four online quizzes, and project work.
Plan for today

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

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

- **McCarthy et al. (1955):** “We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.”
- **1960s:** Pattern-matching with small rule-sets, oriented towards NLU.
- **1970–80s:** Linguistically rich, logic-driven, grounded systems (LLGs); restricted applications.
- **Mid-1990s:** Machine learning revolution in NLP leads to a decrease in NLU work.
- **Late 2000s:** **LLGs** re-emerge, now with *learning*.
- **Mid-2010s:** NLU returns to center stage, with deep learning the most prevalent set of techniques. **LLGs** go into decline.
- **2020s:** Grounding is *back*, and rich logical and linguistic structure is *emergent*. 
Watson wins Jeopardy (2011)
QANTA beats Ken Jennings (2015)
Artificial assistants

"Hey Siri, turn on the lights in the living room"

Hey Siri turn on the lights in the Living Room
Tap to Edit ·

OK, the lights are on.
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. 🎄❤️🎉

耆 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
Search, and way beyond search

Severe Acute Respiratory Syndrome (SARS) - WHO | World ...
Severe acute respiratory syndrome (SARS) is a viral respiratory disease caused by a SARS-associated coronavirus. It was first identified at the end of ...

Severe Acute Respiratory Syndrome | SARS-CoV Disease
Severe acute respiratory syndrome (SARS) is a viral respiratory illness caused by a coronavirus called SARS-associated coronavirus (SARS-CoV).
Basics Fact Sheet · About SARS · Frequently Asked Questions · SARS

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

Kiela et al. 2021
Stanford Question Answering Dataset (SQuAD)

**Leaderboard**

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human Performance</td>
<td>86.831</td>
<td>89.452</td>
</tr>
<tr>
<td></td>
<td>Stanford University</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Rajpurkar &amp; Jia et al. '18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>IE-Net (ensemble)</td>
<td>90.939</td>
<td>93.214</td>
</tr>
<tr>
<td></td>
<td>RICOH_SRCB_DML</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>FPNet (ensemble)</td>
<td>90.871</td>
<td>93.183</td>
</tr>
<tr>
<td></td>
<td>Ant Service Intelligence Team</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>IE-NetV2 (ensemble)</td>
<td>90.860</td>
<td>93.100</td>
</tr>
<tr>
<td></td>
<td>RICOH_SRCB_DML</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SA-Net on Albert (ensemble)</td>
<td>90.724</td>
<td>93.011</td>
</tr>
<tr>
<td></td>
<td>QIANXIN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>RoBERTa+Verify (single model)</td>
<td>86.448</td>
<td>89.586</td>
</tr>
<tr>
<td></td>
<td>CW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>BERT + ConvLSTM + MTL + Verifier (ensemble)</td>
<td>86.730</td>
<td>89.286</td>
</tr>
<tr>
<td></td>
<td>Layer 6 AI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Rajpurkar et al. 2016

[https://rajpurkar.github.io/SQuAD-explorer/](https://rajpurkar.github.io/SQuAD-explorer/)
MultiNLI

Williams et al. 2018

https://paperswithcode.com/sota/natural-language-inference-on-multinli
A peek behind the curtain

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

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Forbidden kinds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mammals</td>
<td>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</td>
</tr>
<tr>
<td>Birds</td>
<td>Birds of prey; scavengers</td>
</tr>
<tr>
<td>Reptiles and amphibians</td>
<td>All</td>
</tr>
<tr>
<td>Water animals</td>
<td>All non-fish. Among fish, all those that do not have both fins and scales</td>
</tr>
<tr>
<td>Insects</td>
<td>All, except particular types of locust or grasshopper that, according to most, cannot be identified today</td>
</tr>
</tbody>
</table>
SIRI on The Colbert Show

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

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

[...]

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

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

---

*Slide idea from Marie de Marneffe*
Worrisome text generation with GPT-3

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

@yoavgo

www.nabla.com/
Un-human image captioning

Sutskever et al. 2014
Misleading automatic curation

https://searchengineland.com
Does Anne Hathaway News Drive Berkshire Hathaway's Stock?

MAR 18 2011, 10:50 AM ET 28  

Given the awesome correlating powers of today's stock trading computers, the idea may not be as far-fetched as you think.
The United Airlines “bankruptcy”

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

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

Question
What is the name of the quarterback who was 38 in Super Bowl XXXIII?
SQuAD adversarial testing

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

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

Answer
John Elway

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

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

<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>
## NL1 adversarial testing

<table>
<thead>
<tr>
<th>Premise</th>
<th>Relation</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>A little girl kneeling in the dirt crying.</td>
<td>entails</td>
</tr>
<tr>
<td>Adversarial</td>
<td></td>
<td>entails</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>
</table>
| **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
## NLI adversarial testing

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

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>F1</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>contradiction</td>
<td>0.99</td>
<td>0.97</td>
<td>0.98</td>
<td>7,164</td>
</tr>
<tr>
<td>entailment</td>
<td>0.86</td>
<td>1.00</td>
<td>0.92</td>
<td>982</td>
</tr>
<tr>
<td>neutral</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>14</td>
</tr>
<tr>
<td>Macro avg.</td>
<td>0.67</td>
<td>0.71</td>
<td>0.68</td>
<td>8,193</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td>0.97</td>
<td>8,193</td>
</tr>
</tbody>
</table>
Two perspectives

Nick Bostrom

Superintelligence
Paths, Dangers, Strategies

How to Survive a Robot Uprising
Tips on Defending Yourself Against the Coming Rebellion
Daniel H. Wilson
Underlying challenges

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

Kiela et al. 2021
Benchmark limitations

- **ImageNet**
- **SQuAD**
- **SNLI**
- **PTB**

Errors: Red circles
Biases: Blue circles
Artifacts: Orange circles
Gaps: Pink circles
Limited assessments

Leaderboards today

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

Leaderboards in the future

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

Gender Bias in Contextualized Word Embeddings

Jieyu Zhao§  Tianlu Wang†  Mark Yan‡
Ryan Cotterell§  Vicente Ordonez‡  Kai-Wei Chang‡
§University of California, Los Angeles  {jyzhao, kwchang}@cs.ucla.edu
†University of Virginia  {tw8bc, vicente}@virginia.edu
‡Allen Institute for Artificial Intelligence  marky@allenai.org

Semantics derived automatically from language corpora contain gendered biases

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

Bevil Vandurme
Johns Hopkins University
vandurme@cs.jhu.edu

Angelis Atliakos2, Roxana Geambasu2, Daniel Hsu2,
Mathias Humbert3, Ari Juels3, and Huang Lin4

1Ecole Polytechnique Fédérale de Lausanne — 2Columbia University — 3Cornell Tech

April 19, 2019
Ever larger models

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

Sanh et al. 2019
Diminishing returns for large models?

Figure 1: Replaced token detection pre-training consistently outperforms masked language model pre-training given the same compute budget. The left figure is a zoomed-in view of the dashed box.

approach is reminiscent of training the discriminator of a GAN, our method is not adversarial in that the generator producing corrupted tokens is trained with maximum likelihood due to the difficulty of applying GANs to text (Caccia et al., 2018).

We call our approach ELECTRA for "Efficiently Learning an Encoder that Classifies Token Replacements Accurately." As in prior work, we apply it to pre-train Transformer text encoders (Vaswani et al., 2017) that can be fine-tuned on downstream tasks. Through a series of ablations, we show that learning from all input positions causes ELECTRA to train much faster than BERT. We also show ELECTRA achieves higher accuracy on downstream tasks when fully trained.

Most current pre-training methods require large amounts of compute to be effective, raising concerns about their cost and accessibility. Since pre-training with more compute almost always results in better downstream accuracies, we argue an important consideration for pre-training methods should be compute efficiency as well as absolute downstream performance. From this viewpoint, we train ELECTRA models of various sizes and evaluate their downstream performance vs. their compute requirement. In particular, we run experiments on the GLUE natural language understanding benchmark (Wang et al., 2019) and SQuAD question answering benchmark (Rajpurkar et al., 2016). ELECTRA substantially outperforms MLM-based methods such as BERT and XLNet given the same model size, data, and compute (see Figure 1). For example, we build an ELECTRA-Small model that can be trained on 1 GPU in 4 days.

ELECTRA-Small outperforms a comparably small BERT model by 5 points on GLUE, and even outperforms the much larger GPT model (Radford et al., 2018). Our approach also works well at large scale, where we train an ELECTRA-Large model that performs comparably to RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019), despite having fewer parameters and using 1/4 of the compute for training. Training ELECTRA-Large further results in an even stronger model that outperforms ALBERT (Lan et al., 2019) on GLUE and sets a new state-of-the-art for SQuAD 2.0. Taken together, our results indicate that the discriminative task of distinguishing real data from challenging negative samples is more compute-efficient and parameter-efficient than existing generative approaches for language representation learning.

METHOD

We first describe the replaced token detection pre-training task; see Figure 2 for an overview. We suggest and evaluate several modeling improvements for this method in Section 3.2.
Diminishing returns for large models?

Sanh et al. 2019
Why is this all so difficult?

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

Black Panther is playing at the Century 16 Theater.

When is it playing there?

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

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

Need **domain knowledge**, **discourse knowledge**, **world knowledge**
Our perspective

• This is the most exciting moment ever in history for doing NLU!

• In academia, there’s been a resurgence of interest in NLU (after a long winter).

• In industry, there’s been an explosion in products and services that rely on NLU.

• Systems are impressive, but show their weaknesses quickly.

• NLU is far from solved – big breakthroughs lie in the future.
Assignments, bakeoffs, and projects

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

**Topics**

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

**Assignments/bakeoffs**

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

**Final projects**

1. Literature review
2. Experiment protocol
3. Final paper
Assignments and bakeoffs

1. There are three regular assignments. (For the third, you can choose between two options.)

2. Each assignment culminates in a bakeoff: an informal competition in which you enter your original model.

3. The assignments ask you to build baseline systems to inform your own model design, and to build your original model.

4. The assignments earn you 9 of the 10 points. All bakeoff entries earn the additional point.

5. Winning bakeoff entries earn extra credit.

6. Rationale for all this: exemplify best practices for NLU projects. (Let us know where we’re not living up to this!)
Assign/Bakeoff: Word relatedness

|   | :) | :/ | :D | :| | ;p | abandon | abc | ability | able | ... |
|---|----|----|----|---|---|-------|-----|--------|------|-----|
| :) | 74 | 1  | 0  | 0 | 0 | 1    | 0   | 2      | 2    |     |
| :/ | 1  | 306| 0  | 0 | 0 | 0    | 0   | 0      | 17   |     |
| :D | 0  | 0  | 16 | 0 | 0 | 0    | 6   | 1      | 1    |     |
| :| | 0  | 0  | 0  | 120| 0   | 0   | 1      | 9    |     |
| ;p | 0  | 0  | 0  | 0 | 516286| 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

<table>
<thead>
<tr>
<th>Reweighting</th>
<th>Dimensionality reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>probabilities length norm. TF-IDF O/E PMI Positive PMI</td>
<td>LSA GloVe word2vec autoencoders</td>
</tr>
</tbody>
</table>

Vector comparison
- Euclidean
- Cosine
- Dice
- KL
### Assign/Bakeoff: Word relatedness

<table>
<thead>
<tr>
<th><strong>Reweighting</strong></th>
<th><strong>Dimensionality reduction</strong></th>
<th><strong>Vector comparison</strong></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>
<tr>
<td>Positive PMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Assign/Bakeoff: Word relatedness

<table>
<thead>
<tr>
<th>Rewriteing</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>
<tr>
<td>Positive PMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(and BERT too, if you wish!)</td>
</tr>
</tbody>
</table>
### Assign/Bakeoff: Word relatedness

<table>
<thead>
<tr>
<th>Word</th>
<th>Related Word</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>sunlight</td>
<td>0.9</td>
</tr>
<tr>
<td>automobile</td>
<td>car</td>
<td>0.95</td>
</tr>
<tr>
<td>river</td>
<td>water</td>
<td>0.8</td>
</tr>
<tr>
<td>food</td>
<td>gull</td>
<td>0.4</td>
</tr>
<tr>
<td>gate</td>
<td>hotel</td>
<td>0.45</td>
</tr>
<tr>
<td>dessert</td>
<td>head</td>
<td>0.01</td>
</tr>
<tr>
<td>born</td>
<td>hockey</td>
<td>0.01</td>
</tr>
<tr>
<td>abandon</td>
<td>soldier</td>
<td>??</td>
</tr>
<tr>
<td>about</td>
<td>wandering</td>
<td>??</td>
</tr>
<tr>
<td>abstract</td>
<td>moon</td>
<td>??</td>
</tr>
<tr>
<td>abstract</td>
<td>rally</td>
<td>??</td>
</tr>
<tr>
<td>abundance</td>
<td>wealth</td>
<td>??</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Train</th>
<th>Dev</th>
<th>Bakeoff test</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST-3 train</td>
<td>SST-3 dev</td>
<td>SST-3 test</td>
</tr>
<tr>
<td>RRS dev</td>
<td>RRS test</td>
<td></td>
</tr>
</tbody>
</table>
Assign/Bakeoff: Contextual color describers

<table>
<thead>
<tr>
<th>Context</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>blue</td>
</tr>
<tr>
<td></td>
<td>The darker blue one</td>
</tr>
<tr>
<td></td>
<td>dull pink not the super bright one</td>
</tr>
<tr>
<td></td>
<td>Purple</td>
</tr>
<tr>
<td></td>
<td>blue</td>
</tr>
</tbody>
</table>

Monroe et al. 2017, 2018
Assign/Bakeoff: Contextual color describers

Monroe et al. 2017, 2018
Assign/Bakeoff: Few-shot OpenQA

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

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

**Title: Pragmatics**
Background: Pragmatics is the study of language use.
Q: What is pragmatics?
A: The branch of linguistics focused on how meaning arises in context.
A note on grading original systems

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

1. Any system that performs extremely well on the bakeoff data will be given full credit, even systems that are very simple. We can’t argue with success according to our own metrics!

2. Systems that are very creative and well-motivated will be given full credit even if they do not perform well on the bakeoff data. We want to encourage creative exploration!

3. Other systems will receive less than full credit, based on the judgment of the teaching team. The specific criteria will vary based on the nature of the assignment. Point deductions will be justified in feedback.
Project work

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

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

Website
https://web.stanford.edu/class/cs224u/

Code repository
https://github.com/cgpotts/cs224u/

Discussion forum
https://edstem.org/us/courses/21353/discussion/

Gradescope
https://www.gradescope.com/courses/381598

Teaching team
cs224u-spr2122-staff@lists.stanford.edu
Components

<table>
<thead>
<tr>
<th>Component</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quizzes</td>
<td>12%</td>
</tr>
<tr>
<td>Special event participation</td>
<td>3%</td>
</tr>
<tr>
<td>Homworks and bakeoffs</td>
<td>35%</td>
</tr>
<tr>
<td>Literature review</td>
<td>10%</td>
</tr>
<tr>
<td>Experimental protocol</td>
<td>10%</td>
</tr>
<tr>
<td>Final project paper</td>
<td>30%</td>
</tr>
</tbody>
</table>
Special events (confirmed so far)

- Rishi Bommasani  
  https://rishibommasani.github.io

- Douwe Kiela  
  https://douwekiela.github.io

- Omar Khattab  
  https://omarkhattab.com

- Adina Williams  
  https://wp.nyu.edu/adinawilliams/

- Ellie Pavlick  
  https://cs.brown.edu/people/epavlick/

- Yulia Tsvetkov  
  https://homes.cs.washington.edu/~yuliats/

- Richard Socher  
  https://www.socher.org

- Kalika Bali  
Fully asynchronous

• Core course content in screencasts on YouTube and linked from the homepage, with accompanying Jupyter notebook for hands-on work.

• A series of special events: conversations with prominent NLU researchers. Mostly on Zoom. Attend live or listen later.

• Other class meetings: optional open discussions and/or spaces for you to work, with the teaching team there to help. Open to mixing in-classroom and Zoom formats.

• Office hours offered in person and on Zoom. Details to come.

• Continuous evaluation: three assignments, four online quizzes, and project work.
Tutorials

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

- setup.ipynb
- tutorial_jupyter_notebooks.ipynb
- tutorial_numpy.ipynb
- tutorial_pytorch.ipynb
Quizzes

1. Quiz 0 is on course requirements and related details. The sole purpose of the quiz is to create a clear incentive for you to study the website and understand your rights and obligations.

2. Quizzes 1–4 create a course-related incentive for individual students to study the material beyond what is required for the more free-form and collaborative assignments.

3. All quizzes are open notes, open book, etc., but no collaboration is permitted.
AWS credits

1. Thanks to AWS Educate, we expect to be able to provide every enrolled student with a $100 AWS credit.

2. As of this year, these codes need to be associated with specific Amazon/AWS accounts. We will share information on this soon.

3. If you haven’t used AWS before:
   - Plan ahead to make sure that you are able to claim the kind of machine you want.
   - **Get your account set up so that you cannot be billed beyond your credits.**

4. This is the only official cloud support for this course. Feel free to use other providers and post questions about them to discussion forum, but the team cannot guarantee support for them.
For next time

1. Get your computing environment set up using setup.ipynb.

2. Make sure you’re in the discussion forum. If not, follow the link given at the homepage for our course Canvas.

3. Consider doing Quiz 0 as a way of getting to know your rights and obligations for this course.

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

5. For corresponding with the teaching team: cs224u-spr2122-staff@lists.stanford.edu
Wrap-up

1. This is the most exciting moment ever in history for doing NLU!

2. This course will give you **hands-on** experience with a wide range of challenging NLU problems.

3. A mentor from the teaching team will guide you through the project assignments – there are many examples of these projects becoming important publications.

4. Central goal: to make you the best – most insightful and responsible – NLU researcher and practitioner wherever you go next.


References

References III

References IV


## References for the benchmark timeline

### Penn Treebank (Marcus et al. 1994)
1. van Halteren 2000
2. Eskin 2000
3. Dickinson and Meurers 2003a
4. Dickinson and Meurers 2003b
5. Dickinson and Meurers 2005
7. Manning 2011

### SNLI (Bowman et al. 2015)
2. Rudinger et al. 2017
3. Naik et al. 2018
4. Glockner et al. 2018
5. Naik et al. 2018
6. Poliak et al. 2018
7. Tsuchiya 2018
8. Gururangan et al. 2018
9. Belinkov et al. 2019
10. McCoy et al. 2019

### SQuAD (Rajpurkar et al. 2016, 2018)
1. Weissenborn et al. 2017
2. Sugawara et al. 2018
3. Bartolo et al. 2020
4. Lewis et al. 2021

### ImageNet (Deng et al. 2009)
1. Deng et al. 2014
2. Stock and Cisse 2018
3. Yang et al. 2020
4. Recht et al. 2019
5. Northcutt et al. 2021
6. Crawford and Paglen 2021