What’s the talk about

1. The past and the Present of NLP
2. Applications of state of the art NLP
3. (some) opportunities and open research questions
Rule- and feature based approaches

Input
The Pats win the AFC East for the 9th straight year. The Patriots trailed 24-16 at the end of the third quarter. They scored on a 46-yard field goal with 4:00 left in the game to pull within 24-19. Then, with 56 seconds remaining, Dion Lewis scored on an 8-yard run and the Patriots added a two-point conversion to go ahead 27-24. The game ended on a Roethlisberger interception. Steelers wide receiver Antonio Brown left in the first half with a bruised calf.

Feature vector
0 1 1 1 0 1 1

Rules
if... then...
if... then...
if... then...
if... then...

Dataset
The Pats win the AFC East for the 9th straight year. The Patriots trailed 24-16 at the end of the third quarter. They scored on a 46-yard field goal with 4:00 left in the game to pull within 24-19. Then, with 56 seconds remaining, Dion Lewis scored on an 8-yard run and the Patriots added a two-point conversion to go ahead 27-24. The game ended on a Roethlisberger interception. Steelers wide receiver Antonio Brown left in the first half with a bruised calf.

Symbol level (strings)

Symbol-level operations
RegEx
Gazetteers
Lexica

Pipeline
→ POS tagging
→ Dependency Parsing
→ NER
→ etc

Prediction
SPAM

Decision

Gold Standard
The Pats win the AFC East for the 9th straight year. The Patriots trailed 24-16 at the end of the third quarter. They scored on a 46-yard field goal with 4:00 left in the game to pull within 24-19. Then, with 56 seconds remaining, Dion Lewis scored on an 8-yard run and the Patriots added a two-point conversion to go ahead 27-24. The game ended on a Roethlisberger interception. Steelers wide receiver Antonio Brown left in the first half with a bruised calf.

archetypal example:
Spam detection

Feature
\( (x, r, y) \) covers all words in \( s \)
The last preposition in \( r \) is for
The last preposition in \( r \) is on
The last preposition in \( r \) is of
\( len(s) \leq 10 \) words
There is a WH-word to the left of \( r \)
\( r \) matches VW*P from Figure 1
Embedding-based approaches

The Patriots win the AFC East for the 9th straight year. The Patriots trailed 24-16 at the end of the third quarter. They scored on a 46-yard field goal with 4:00 left in the game to pull within 24-19. Then, with 56 seconds remaining, Dion Lewis scored on an 8-yard run and the Patriots added a two-point conversion to go ahead 27-24. The game ended on a Roethlisberger interception. Steelers wide receiver Antonio Brown left in the first half with a bruised calf.
Distributional Representation

“Tell me with whom thou art found, and I will tell thee who thou art.” - JW Goethe

“Words that occur in similar contexts are similar.”
Place words in a high-dimensional vector space
→ Move words closer that appear in similar contexts
→ Move words apart that do not appear in similar contexts

As observed in some big corpus
Embeddings

Static map from words to their (high dimensional) distributional representation.

\[ \text{jet} \approx \frac{1}{4} \times (\text{bee} + \text{rocket}) \]
Tasks

Exploit similarity (low distance) in high-dimensional vector space

- topic modelling
- visualisations
- clustering
- information retrieval (K-nearest neighbours)
What is a neural network

\[ h_t = g(W x_t) \]

Matrix multiplication

(Non linear) activation function

“hidden vector”
Training
Embedding-based approaches

1. embed inputs
2. obtain prediction
3. calculate error between prediction & actual value
4. attribute error to parameters in neural networks (backpropagation)
5. Change parameters to reduce the error (gradient descent)

The Pats win the AFC East for the 9th straight year. The Patriots trailed 24-16 at the end of the third quarter. They scored on a 46-yard field goal with 4:00 left in the game to pull within 24-19. Then, with 56 seconds remaining, Dion Lewis scored on an 8-yard run and the Patriots added a two-point conversion to go ahead 27-24. The game ended on a Roethlisberger interception. Steelers wide receiver Antonio Brown left in the first half with a bruised calf.
What that means

Collect examples as input/output pairs

Mature software exists to perform the computations

1. embed inputs
2. obtain prediction
3. calculate error between prediction & actual value
4. attribute error to parameters in neural networks (backpropagation)
5. Change parameters to reduce the error (gradient descent)
What that means

Minimal supervision

• No task-specific knowledge needs to be encoded in the neural network
• neural networks learns to perform task from input-output examples

... as long as you have the data
Tasks: Document-Level Classification

Assign (one or multiple) fixed category to a piece of text
- Paraphrasing
- Abuse (e.g. offensive language) detection
- Fact Verification
- Stance detection
- Sentiment Analysis
- Emotion Classification
Fact Verification

Claim: The Rodney King riots took place in the most populous county in the USA.

[wiki/Los_Angeles_Riots]
The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[wiki/Los_Angeles_County]
Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported
Stance detection

SDQC support classification. Example 1:

u1: We understand there are two gunmen and up to a dozen hostages inside the cafe under siege at Sydney. ISIS flags remain on display #7News [support]
  u2: @u1 not ISIS flags [deny]
  u3: @u1 sorry - how do you know it’s an ISIS flag? Can you actually confirm that? [query]
  u4: @u3 no she can’t cos it’s actually not [deny]
  u5: @u1 More on situation at Martin Place in Sydney, AU –LINK– [comment]
  u6: @u1 Have you actually confirmed its an ISIS flag or are you talking shit [query]

SDQC support classification. Example 2:

u1: These are not timid colours; soldiers back guarding Tomb of Unknown Soldier after today’s shooting #StandforCanada –PICTURE– [support]
  u2: @u1 Apparently a hoax. Best to take Tweet down. [deny]
  u3: @u1 This photo was taken this morning, before the shooting. [deny]
  u4: @u1 I don’t believe there are soldiers guarding this area right now. [deny]
  u5: @u4 wondered as well. I’ve reached out to someone who would know just to confirm that. Hopefully get response soon. [comment]
  u4: @u5 ok, thanks. [comment]
## Sentiment Analysis

<table>
<thead>
<tr>
<th>Tweet Content</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>@ThisIsDeep_ you are about as deep as a turd in a toilet bowl. Internet culture is garbage and you are bladder cancer.</td>
<td>-4</td>
</tr>
<tr>
<td>A paperless office has about as much chance as a paperless bathroom</td>
<td>-3</td>
</tr>
<tr>
<td>Today will be about as close as you'll ever get to a &quot;PERFECT 10&quot; in the weather world! Happy Mother's Day! Sunny and pleasant! High 80.</td>
<td>3</td>
</tr>
<tr>
<td>I missed voting due to work. But I was behind the Austrian entry all the way, so to speak. I might enter next year. Who knows?</td>
<td>1</td>
</tr>
</tbody>
</table>

Emotion Detection


Utterance

**Pheobe**: Oh my god, he’s lost it. He’s totally lost it.

**Monica**: What?

**Ross**: Or! Or, we could go to the bank, close our accounts and cut them off at the source.

**Chandler**: You’re a genius!

**Joey**: Aww, man, now we won’t be bank buddies!

Emotion

- t-4: [oth]
- t-3: [sur]
- t-2: [neu]
- t-1: [joy]
- t: [sad]
Tasks: Token-Level Classification

Assign (one or multiple) fixed category to each word:
- part of speech tagging
- named entity recognition

Model Output

Entities
Viktor Schlegel is a Lecturer at the University of Manchester

The quick brown fox jumped over the lazy dog.
### Tasks: Token-Level Classification

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Extractions for <code>walk</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>As I walk through the valley of the shadow of death, I take a look at my life and realise I've nothing left.</td>
<td><img src="image" alt="Extraction" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Extractions for <code>take</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>As I walk through the valley of the shadow of death, I take a look at my life and realise I've nothing left.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Extractions for <code>realise</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>As I walk through the valley of the shadow of death, I take a look at my life and realise I've nothing left.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Extractions for <code>I’ve</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>As I walk through the valley of the shadow of death, I take a look at my life and realise I've nothing left.</td>
</tr>
</tbody>
</table>
Token-Level Classification: Language modelling

Task: Predict next word given $n$ previous words.

$$P(x_{n+1} | x_1, \ldots, x_n)$$

“I grew up in Germany, i speak fluent _____”

syntax information
semantic information
Language models for embeddings

self supervision

Masked sentence

I grew up in Germany, I speak fluent [MASK].

minimise

Masked Word Prediction

Actual masked Word

potato

German
The Pats win the AFC East for the 9th straight year. The Patriots trailed 24-16 at the end of the third quarter. They scored on a 46-yard field goal with 4:00 left in the game to pull within 24-19. Then, with 56 seconds remaining, Dion Lewis scored on an 8-yard run and the Patriots added a two-point conversion to go ahead 27-24. The game ended on a Roethlisberger interception. Steelers wide receiver Antonio Brown left in the first half with a bruised calf.

State-of-the-art NLP

0: Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." 2018
**Tasks: Generation**

Given a piece of text, generate the next token

- Translation
- Abstractive summarisation
- Question Answering

\[ P(x_{n+1} | x_1, \ldots, x_n) \]
Abstractive summarisation

How to Help Save Rivers

Method 1: Reducing Your Water Usage

1. Take quicker showers to conserve water. One easy way to conserve water is to cut down on your shower time. Practice cutting your showers down to 10 minutes, then 7, then 5. Challenge yourself to take a shorter shower every day.

2. Wait for a full load of clothing before running a washing machine. Washing machines take up a lot of water and electricity, so running a cycle for a couple of articles of clothing is inefficient. Hold off on laundry until you can fill the machine.

3. Turn off the water when you're not using it. Avoid letting the water run while you’re brushing your teeth or shaving. Keep your hoses and faucets turned off as much as possible. When you need them, use them sparingly.

Article 1:
One easy way to conserve water is to cut down on your shower time. Practice cutting your showers down to 10 minutes, then 7, then 5. Challenge yourself to take a shorter shower every day. Washing machines take up a lot of water and electricity, so running a cycle for a couple of articles of clothing is inefficient. Hold off on laundry until you can fill the machine. Avoid letting the water run while you’re brushing your teeth or shaving. Keep your hoses and faucets turned off as much as possible. When you need them, use them sparingly.

...  

Summary 1:
Take quicker showers to conserve water. Wait for a full load of clothing before running a washing machine. Turn off the water when you’re not using it.
Open research questions

The Pats win the AFC East for the 9th straight year. The Patriots trailed 24-16 at the end of the third quarter. They scored on a 46-yard field goal with 4:00 left in the game to pull within 24-19. Then, with 56 seconds remaining, Dion Lewis scored on an 8-yard run and the Patriots added a two-point conversion to go ahead 27-24. The game ended on a Roethlisberger interception. Steelers wide receiver Antonio Brown left in the first half with a bruised calf.
The Pats win the AFC East for the 9th straight year. The Patriots trailed 24-16 at the end of the third quarter. They scored on a 46-yard field goal with 4:00 left in the game to pull within 24-19. Then, with 56 seconds remaining, Dion Lewis scored on an 8-yard run and the Patriots added a two-point conversion to go ahead 27-24. The game ended on a Roethlisberger interception. Steelers wide receiver Antonio Brown left in the first half with a bruised calf.

Who was injured during the match?
(a) Rob Gronkowski (b) Ben Roethlisberger (c) Dion Lewis (d) Antonio Brown

The Patriots champion the cup for ___ consecutive seasons.

What was the final score of the game?

How many points ahead were the Patriots by the end of the game? 3
Why Question Answering/MRC

MRC/QA as interface

- **HCI interface to exploration**

  ```sql
  SELECT name FROM mountains WHERE country = UK
  ORDER BY height
  LIMIT 1
  ```
  “What is the highest mountain in the UK?”

- **Interface to other tasks**

  What is being done? → walk  
  Who walks through the valley? → I  
  Where do I walk? → through the valley of the shadow of death
What was the final score of the game?
Neural networks are black boxes
What was the final score of the game?
We don’t really know what these models learn.

1. What do gold standards (not) evaluate?
2. Which capabilities do the models (not) acquire?
A (systematic) look into the data...

(a) Lexical (grey background) and syntactic (white background) linguistic features

(b) Answer Types
(c) Answer Quality
(d) Required Knowledge
What’s missing?

E.g.: 
Restrictivity:
“Brady almost scored a TD” vs “Brady scored a TD”

Factivity:
“He was probably involved in the plot” vs “He was definitely involved in the plot”

Discourse relations:
Conditionals:
“If they have milk, I will buy 3 bottles of milk” vs “I will buy 3 bottles of milk.”

Conjunctions:
“Mary ate an apple. John ate an apple. John ate a pear.”
“Who ate an apple?” vs “Who ate an apple and a pear?”

Common Theme: 
Phenomena that preserve the lexical surface form while altering the meaning are missing!

⇒ cannot say, whether systems learn to process them...

...look similar but mean different things.
2. Which capabilities do the models (not) acquire?

“Skill-based” Evaluation

- Skill specific Instructions
- Crowd workers
- ~ 100k Questions
- Passages
- Answers
- Training Set
- Evaluation Set

Dua et al: DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. NAACL 2019

Yang et al: HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. EMNLP 2018


MRC: Expectation...

<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>ALBERT + DAADF + Verifier (ensemble)</td>
<td>90.002</td>
<td>92.425</td>
</tr>
<tr>
<td></td>
<td>PINGAN Omni-Sinitic</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MRC: vs reality

Answer
2

Explanation
The model decided this was a counting problem.

Passage
I have an apple.

Question
How many apples do I have?
Weaknesses...

Passage 1: Marietta Air Force Station
Marietta Air Force Station (ADC ID: M-111, NORAD ID: Z-111) is a closed United States Air Force General Surveillance Radar station. It is located 2.1 mi northeast of Smyrna, Georgia. It was closed in 1968.

Passage 2: Smyrna, Georgia
Smyrna is a city northwest of the neighborhoods of Atlanta. [...] As of the 2010 census, the city had a population of 51,271. The U.S. Census Bureau estimated the population in 2013 to be 53,438. [...] 

Question: What is the 2010 population of the city 2.1 miles southwest of Marietta Air Force Station?

- Survey 121 papers for
  - Reported Data and model “weaknesses”
  - Reported Methods to reveal them
  - Reported Methods to overcome them
- Classify, categorise, detect common themes

words in question colocated with answer
=> unwanted cues

low-bias models rely on strongest signal in data: learn to “pay attention” to the surface form
=> sophisticated word matching

Undetected with usual evaluation methodology
We still don’t really know what these systems learn.

All we can say “It learned to succeed at this dataset”.
Dataset represents task?

E.g. in Chess, task:
outperform humans consistently
dataset: millions of games
“Skill based” evaluation: Challenge sets

Construct challenge set:
- Hand-craft/generate examples that require skill of interest to solve
- Evaluate (optimised) MRC system on those examples
  - “Good” performance: system “learned” it?
  - “Bad” performance: need to discount
    - Unknown words?
    - Different topic?
    - Different passage lengths?
    - ???
    - Doesn’t learn the skill?
For example: Semantic altering Modifications (SAM)

Understanding MRC models can process SAM is interesting, because those phenomena require some sort of deeper comprehension beyond the exploitation of lexical cues.

Brady scored a 45-yard TD. Brady scored another 25-yard touchdown after 5 minutes. Brady scored 2 20-yard touchdowns in the second half.

Brady *almost* scored a 45-yard TD. Brady scored another 25-yard TD touchdown after 5 minutes. Brady scored 2 20-yard touchdowns in the second half.

What is the longest TD run?

*Compare numbers next to ‘TD/touchdown’, pick highest!*

Doesn’t work anymore!
Can MRC systems process SAM?

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Average DICE</th>
<th>SQUAD</th>
<th>HOTPOTQA</th>
<th>NEWSQA’</th>
<th>DROP’</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EM/F1</td>
<td>EM/F1</td>
<td>EM/F1</td>
<td>EM/F1</td>
</tr>
<tr>
<td>bidaf</td>
<td>11 ± 3</td>
<td>67.2/76.9</td>
<td>44.6/57.9</td>
<td>40.0/54.3</td>
<td>50.8/56.8</td>
</tr>
<tr>
<td>bert-base</td>
<td>13 ± 2</td>
<td>76.3/84.9</td>
<td>50.7/64.9</td>
<td>46.6/62.5</td>
<td>50.5/58.2</td>
</tr>
<tr>
<td>bert-large</td>
<td>15 ± 2</td>
<td>81.9/89.4</td>
<td>54.4/68.7</td>
<td>49.1/65.7</td>
<td>62.2/68.7</td>
</tr>
<tr>
<td>roberta-base</td>
<td>15 ± 2</td>
<td>82.4/89.9</td>
<td>51.9/66.4</td>
<td>50.8/66.9</td>
<td>63.5/69.3</td>
</tr>
<tr>
<td>roberta-large</td>
<td>18 ± 1</td>
<td>86.4/93.3</td>
<td>58.6/72.9</td>
<td>54.4/71.1</td>
<td>77.3/82.8</td>
</tr>
<tr>
<td>albert-base</td>
<td>14 ± 2</td>
<td>82.8/90.3</td>
<td>55.4/69.7</td>
<td>49.7/65.7</td>
<td>60.7/67.0</td>
</tr>
<tr>
<td>albert-large</td>
<td>16 ± 1</td>
<td>85.4/92.1</td>
<td>59.4/73.7</td>
<td>52.5/68.9</td>
<td>69.3/75.1</td>
</tr>
<tr>
<td>albert-xl</td>
<td>27 ± 2</td>
<td>87.1/93.5</td>
<td>62.4/76.2</td>
<td>54.2/70.4</td>
<td>76.4/81.8</td>
</tr>
<tr>
<td>albert-xxl</td>
<td>27 ± 1</td>
<td>88.2/94.4</td>
<td>65.9/79.5</td>
<td>54.3/71.0</td>
<td>78.4/84.5</td>
</tr>
<tr>
<td>t5-small</td>
<td>10 ± 1</td>
<td>76.8/85.8</td>
<td>51.8/65.6</td>
<td>47.3/63.3</td>
<td>60.4/66.1</td>
</tr>
<tr>
<td>t5-base</td>
<td>16 ± 1</td>
<td>82.4/90.6</td>
<td>61.0/74.4</td>
<td>52.4/68.8</td>
<td>69.0/74.9</td>
</tr>
<tr>
<td>t5-large</td>
<td>20 ± 1</td>
<td>86.3/93.1</td>
<td>65.0/78.5</td>
<td>53.4/70.0</td>
<td>70.1/75.3</td>
</tr>
<tr>
<td>average</td>
<td><strong>19 ± 0</strong></td>
<td>76.4/83.2</td>
<td>53.1/65.9</td>
<td>47.1/62.1</td>
<td>61.5/67.0</td>
</tr>
<tr>
<td>albert-xl-comb</td>
<td><strong>20 ± 2</strong></td>
<td>85.3/92.2</td>
<td>60.6/74.3</td>
<td>53.6/70.4</td>
<td>76.9/82.4</td>
</tr>
<tr>
<td>random</td>
<td>5 ± 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>learned</td>
<td>98 ± 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Baseline Passage: Bob baked some cookies for Alice. […]
Intervention Passage: Bob baked Alice some cookies. […]
Question: What did Bob bake? Answer: some cookies (not Alice)

Baseline Passage: Bob drew a picture of mom for Alice. […]
Intervention Passage: Bob drew Alice a picture of mom. […]
Question: Who did Bob draw? Answer: mom (not Alice)

Also useful for other phenomena! (e.g. dative alteration)
We still still don’t really know what these models learn.

All we can say “models didn’t learn skill x”.
Diminishing returns

- NLP progress is driven by size (model/data)
- bigger models require more compute
- Participation barrier


Raffel et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR 2020
Being represented

- if a group is not represented in training data, data driven solution “won’t work” for them
  - e.g. face recognition system trained on a non-diverse dataset
  - or NLP system trained only on standard english

<table>
<thead>
<tr>
<th>Original</th>
<th>When is the suspended team scheduled to return?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adversary</td>
<td>When are the suspended team schedule to returned?</td>
</tr>
<tr>
<td>Prediction</td>
<td><strong>Before:</strong> 2018        <strong>After:</strong> No answer</td>
</tr>
</tbody>
</table>

Majority isn’t always right!

Neural networks excel at exploiting statistical patterns in data

- No device to distinguish between spurious and robust correlations
- With more training data, so the hope, robust correlations will dominate over spurious
- However: Majority is not always right!

The nurse notified the patient that his shift would be ending in an hour.

Whose shift will end in an hour?

The nurse notified the patient that her shift would be ending in an hour.
Majority isn’t always right!

⇒ Debiasing representations

What if there’s no data?

- Language models work so well because there’s a lot of data to learn a rich representation
- what if a language has little (no) data available?
  ➞ few-shot/transfer learning
  ➞ cross-lingual representations (can help)
  ➞ cross-lingual few-shot learning
In conclusion

- Neural network based approaches model tasks end-to-end
  - And can learn to perform a task based on input-output examples
  - If there’s enough input-output examples
- Many tasks can be modelled as input-output examples
- More training data and bigger models usually means better performance
In conclusion

- Neural networks excel at inferring statistical patterns from training data
  - Superb performance if task is well represented
  - Unpredictable performance if application differs from training data
  - To interpret their behaviour on a fine-grained level, we can evaluate their inferred capabilities
  - We use task-specific data that requires a capability to be solved successfully to establish the capability
In conclusion

- Deep learning research has impact beyond the scientific community
  - Higher participation barrier can lead to underrepresentation
  - Bias in data can be exacerbated