Pretraining
Welcome to online 6.806-864!

OH will take place on WebEx: link has been emailed. (Please continue to sign up for time slots.)

Take-home midterm: instructions and makeup sign-up to be emailed.

Practice midterm Qs posted tonight.
Recap: neural sequence models
Language modeling with feedforward networks

- Associate a distributed vector per word
- Express the joint probability function of word sequences in terms of the vectors
- Simultaneously learn word vectors and parameters of the probability function

- Implemented as feed-forward network
- Shared vector mapping, $V$, for all words
- First layer concatenated context vectors
- Perplexity improvements on Brown and AP News corpora over best n-grams

\[ p(w_i|w_{i-3}w_{i-2}w_{i-1}) \]
Language modeling with RNNs

A (unidirectional) RNN can compute \( p(y_t | x_{:t}) \).

Suppose for a sequence \( x \) we set \( y_t = x_{t+1} \).

Then \( \sum_t \log p(x_{t+1} | x_{:t}) = p(x) \)
1. When predicting output $i$, assign a weight $\alpha_{ij}$ to each encoder state $h_j$

2. Compute a pooled input $c_i = \sum_j \alpha_{ij} h_j$
Transformers

- Non-recurrent seq2seq (encoder-decoder) model
- Multi-layered attention model enables lateral information transfer across an input sequence
- Cost function is cross-entropy error of decoder
- Original paper demonstrated good results on machine translation and constituency parsing
- Transformers are the basis for BERT etc. (which we will see next week)

[Vaswani et al., “Attention is All You Need” arXiv:1706.03762 2017]
Recap: pretraining
Language modeling with word2vec

- **Skip-gram** predicts neighbor words from center word

  quick, brown, fox, jumped, over

- Each output is predicted independently

  \[ p(w_{c+n} | w_c) \]
  
  \[ \prod_{-h \leq n \leq h, n \neq 0} p(w_{c+n} | w_c) \]

- Context window lengths can be sampled

  \[ p(w_{c-2} | w_c) \quad p(w_{c-1} | w_c) \quad p(w_{c+1} | w_c) \quad p(w_{c+2} | w_c) \]

[Distributed Representations of Words and Phrases and their Compositionality. Mikolov et al., 2003]
RNNs and word embeddings

learned word embeddings

one-hot vectors

cheap and very tasty

word embedding matrix
Homonyms

I can run.  I can anchovies.
Word senses

I deposited money in the bank.

I climbed up the bank of the river.
I’ll meet you at the bank.

All my classmates work for banks.

She’s a volunteer at the blood bank.
Definition of *do* (Entry 1 of 5)

**transitive verb**

1: to bring to pass: **CARRY OUT**
   *do* another's wishes

2: **PUT** —used chiefly in *do to death*

3
   a: **PERFORM, EXECUTE**
      *do* some work
      *did* his duty
   b: **COMMIT**
      crimes *done* deliberately

4
   a: **BRING ABOUT, EFFECT**
      trying to *do* good
      *do* violence
   b: **to give freely:** **PAY**
      *do* honor to her memory

5: to bring to an end: **FINISH** —used in the past participle
   the job is finally *done*

6: to put forth: **EXERT**
   *did* her best to win the race

7
   a: **to wear out especially by physical exertion:** **EXHAUST**
      at the end of the race they were pretty well *done*
   b: **to attack physically:** **BEAT**
      *also:* **KILL**

8: to bring into existence: **PRODUCE**
   *do* a biography on the general
   has *done* some beautiful landscapes
Representations of words

Learned word embeddings:

1. ridge of dirt
2. financial inst
3. row of objects
4. reserve supply

One-hot vectors:

at

the

bank
Representations of words

learned word embeddings

one-hot vectors

at the bank

mound creditor row stockpile
Word sense disambiguation

learned word embeddings

one-hot vectors


2. financial inst
Representations of words in context

learned word embeddings

one-hot vectors

2. financial inst

at  the  bank
Language modeling objectives
Language modeling with word2vec

- Skip-gram predicts neighbor words from center word
  
  quick brown fox jumped over

- Each output is predicted independently
  
  \[
  \prod_{\substack{-h \leq n \leq h \\ n \neq 0}} p(w_{c+n}|w_c) \]

- Context window lengths can be sampled
  
  \[
  p(w_{c-2}|w_c) p(w_{c-1}|w_c) p(w_{c+1}|w_c) p(w_{c+2}|w_c) \]
Language modeling with RNNs

\[
\text{loss} = p(\text{and} \mid \text{cheap}) + p(\text{very} \mid \text{cheap and}) + p(\text{tasty} \mid \text{very cheap})
\]
Language modeling with transformers

\[ p(\text{and} \mid \text{cheap}) \quad p(\text{very} \mid \text{cheap and}) \quad p(\text{tasty} \mid \text{cheap and very}) \]
I was out of money so I went to the bank and
I was out of money so I went to the bank and...
I was out of money so I went to the bank and
John has a book. Mary has an apple. He gave her his
John has a book. Mary has an apple. He gave her his
John has a book. Mary has an apple. He gave her his
John has a book. Mary has an apple. He gave her his
Fine-tuning: categorical output

cheap and tasty
Fine-tuning: categorical output

cheap and tasty → 5 stars
1. Pretrain on a language modeling task
2. Connect a feed-forward network to the last repr. in the sentence
3a. Freeze LM weights and just train the feed-forward part, or
3b. Fine-tune everything together

Fine-tuning LMs: categorical output

- cheap
- and
- tasty

5 stars
1. Pretrain on a language modeling task
2. Make a new “language modeling” dataset with your input–output pairs
3. Fine-tune everything together:

**Pretrain:**
The following year she published a paper called Idealtheorie in Ringbereichen, analyzing ascending chain conditions with regard to (mathematical) ideals. Noted algebraist Irving Kaplansky called this work "revolutionary"; the publication gave rise to the term "Noetherian ring" and the naming of several other mathematical objects as Noetherian.

**Fine-tune:**
Who was Zeng Jiongzhi’s doctoral advisor? **Emmy Noether.**
Where was Barack Obama born? **Honolulu.**
Bonus: “zero-shot” learning

Don’t fine-tune at all!

Model prompt:
The 2008 Summer Olympics torch relay was run from March 24 until August 8, 2008, prior to the 2008 Summer Olympics, with the theme of “one world, one dream”. Plans for the relay were announced on April 26, 2007, in Beijing, China. The relay, also called by the organizers as the “Journey of Harmony”, lasted 129 days and carried the torch 137,000 km (85,000 mi)

Q: what was the theme? A:

Continuation:
“one world, one dream”

[Radford et al. 2019]
What’s missing?

I can anchovies.

Left-to-right language modeling objectives give us sentence representations, but not fully contextual word representations.
Masked language modeling objectives
Bidirectional RNNs

\[ p(y_t | O) \]

cheap and very tasty
Bidirectional “language modeling”

very???
Bidirectional “language modeling”

and???
ELMo: bidirectional language modeling

Idea: train independent forward / backward LMs and concatenate the representations.

$p(\text{and} | \text{cheap})$  
$p(\text{and} | \text{tasty very})$

[Deep Contextualized Word Representations. Peters et al. 2018]
Idea: train independent forward / backward LMs and concatenate the representations.

Every word has a forward repr., a backward repr., and a context–indep. repr.

[Deep Contextualized Word Representations. Peters et al. 2018]
ELMo: more details

We're actually training a deep LSTM, so multiple layers in each representation.

Most effective: use a learned linear combination of layers as input to the downstream task.

Use these anywhere you’d use word embeddings!
“Bidirectional” transformer LMs

\[ p(\text{and} \mid \text{cheap}) \]

\[ p(\text{and} \mid \text{very}) \]
“Bidirectional” transformer LMs

$p(\text{and} \mid \text{cheap} \ ? \ \text{very})$

predict this

Idea: Rather than masking everything to the right, mask at arbitrary positions and only predict at masks.
BERT: Masked language modeling

Idea: add multiple mask tokens per sentence and predict all of them at the same time.
BERT: more tricks

\[ p(\text{cheap} \mid \text{the? and}) \]  \(\text{predict this}\)

1. If we only predict above [MASK] tokens, no pressure on model to route information to rest of sentence (we want good embeddings everywhere)

Idea: instead of always labeling prediction targets as [MASK], sometimes leave them in place or replace with a random word.
(1) We'd also like to encourage the model to capture some global information. Idea: train on pairs of sentences; learn to predict whether they're adjacent in a training document.
(2) We'd also like to encourage the model to capture some global information.

Idea: train on pairs of sentences; learn to predict whether they're adjacent in a training document.

FALSE and

transformer

What do do with out-of-vocabulary words?

Idea: identify \( k \) most frequent word pieces in the corpus and operate on those.

*The viscountess Wallingford* →

[CLS] the viscount ess wall ing ford
Language modeling?

It's very hard to sample sentences from this model!
(and generally not done)

Indeed, can't replace a [SEP] with a word sequence of unknown length—BERT knows how big the gap is.
Fine-tuning MLMs: sequence labeling

1. Pretrain the masked LM task
2. Use final transformer representations to predict your labels rather than words
3. **Fine-tune everything!**

```plaintext
[CLS] cheap and delicious [SEP] my talented chihuahua
```

- **Adj**
- **Conj**
- **Adj**
- **Pos**
- **Adj**
- **Noun**

**Transformer**
Why is (M)LM a good pretraining objective?

He shook his head, took a step back and held his hands up as he tried to smile without losing a cigarette. “Yes you can,” Julia said in a reassuring voice. “I’ve already focused on my friend. You just have to click the shutter, on top, here.” He nodded sheepishly, through his cigarette away and took the [?]
Why is (M)LM a good pretraining objective?

He shook his head, took a step back and held his hands up as he tried to smile without losing a cigarette. “Yes you can,” Julia said in a reassuring voice. “I’ve already focused on my friend. You just have to click the shutter, on top, here.” He nodded sheepishly, through his cigarette away and took the [?] camera
How much does this help?

Question answering:

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EM</td>
<td>F1</td>
</tr>
<tr>
<td>Top Leaderboard Systems (Dec 10th, 2018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>86.3</td>
<td>89.0</td>
</tr>
<tr>
<td>#1 Single - MIR-MRC (F-Net)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>#2 Single - nlnet</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Published</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unet (Ensemble)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SLQA+ (Single)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERTLARGE (Single)</td>
<td>78.7</td>
<td>81.9</td>
</tr>
</tbody>
</table>

[Devlin et al., 2018]
How much does this help?

Sentence classification:

<table>
<thead>
<tr>
<th>System</th>
<th>MNLI-(m/mm)</th>
<th>QQP</th>
<th>QNLI</th>
<th>SST-2</th>
<th>CoLA</th>
<th>STS-B</th>
<th>MRPC</th>
<th>RTE</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-OpenAI SOTA</td>
<td>80.6/80.1</td>
<td>66.1</td>
<td>82.3</td>
<td>93.2</td>
<td>35.0</td>
<td>81.0</td>
<td>86.0</td>
<td>61.7</td>
<td>74.0</td>
</tr>
<tr>
<td>BiLSTM+ELMo+Attn</td>
<td>76.4/76.1</td>
<td>64.8</td>
<td>79.8</td>
<td>90.4</td>
<td>36.0</td>
<td>73.3</td>
<td>84.9</td>
<td>56.8</td>
<td>71.0</td>
</tr>
<tr>
<td>OpenAI GPT</td>
<td>82.1/81.4</td>
<td>70.3</td>
<td>87.4</td>
<td>91.3</td>
<td>45.4</td>
<td>80.0</td>
<td>82.3</td>
<td>56.0</td>
<td>75.1</td>
</tr>
<tr>
<td>BERT&lt;sub&gt;BASE&lt;/sub&gt;</td>
<td>84.6/83.4</td>
<td>71.2</td>
<td>90.5</td>
<td>93.5</td>
<td>52.1</td>
<td>85.8</td>
<td>88.9</td>
<td>66.4</td>
<td>79.6</td>
</tr>
<tr>
<td>BERT&lt;sub&gt;LARGE&lt;/sub&gt;</td>
<td>86.7/85.9</td>
<td>72.1</td>
<td>92.7</td>
<td>94.9</td>
<td>60.5</td>
<td>86.5</td>
<td>89.3</td>
<td>70.1</td>
<td>82.1</td>
</tr>
</tbody>
</table>

[Devlin et al., 2018]
What is learned?

**Head 1-1**
Attends broadly

<table>
<thead>
<tr>
<th>...</th>
<th>found</th>
<th>in</th>
<th>taiwan</th>
<th>[SEP]</th>
<th>the</th>
<th>wingspan</th>
<th>is</th>
<th>24</th>
<th>28</th>
<th>mm</th>
<th>[SEP]</th>
</tr>
</thead>
<tbody>
<tr>
<td>found</td>
<td>in</td>
<td>taiwan</td>
<td>[SEP]</td>
<td>the</td>
<td>wingspan</td>
<td>is</td>
<td>24</td>
<td>28</td>
<td>mm</td>
<td>[SEP]</td>
<td></td>
</tr>
</tbody>
</table>

**Head 3-1**
Attends to next token

<table>
<thead>
<tr>
<th>...</th>
<th>found</th>
<th>in</th>
<th>taiwan</th>
<th>[SEP]</th>
<th>the</th>
<th>wingspan</th>
<th>is</th>
<th>24</th>
<th>28</th>
<th>mm</th>
<th>[SEP]</th>
</tr>
</thead>
<tbody>
<tr>
<td>found</td>
<td>in</td>
<td>taiwan</td>
<td>[SEP]</td>
<td>the</td>
<td>wingspan</td>
<td>is</td>
<td>24</td>
<td>28</td>
<td>mm</td>
<td>[SEP]</td>
<td></td>
</tr>
</tbody>
</table>

**Head 8-7**
Attends to [SEP]

<table>
<thead>
<tr>
<th>...</th>
<th>found</th>
<th>in</th>
<th>taiwan</th>
<th>[SEP]</th>
<th>the</th>
<th>wingspan</th>
<th>is</th>
<th>24</th>
<th>28</th>
<th>mm</th>
<th>[SEP]</th>
</tr>
</thead>
<tbody>
<tr>
<td>found</td>
<td>in</td>
<td>taiwan</td>
<td>[SEP]</td>
<td>the</td>
<td>wingspan</td>
<td>is</td>
<td>24</td>
<td>28</td>
<td>mm</td>
<td>[SEP]</td>
<td></td>
</tr>
</tbody>
</table>

**Head 11-6**
Attends to periods

<table>
<thead>
<tr>
<th>...</th>
<th>found</th>
<th>in</th>
<th>taiwan</th>
<th>[SEP]</th>
<th>the</th>
<th>wingspan</th>
<th>is</th>
<th>24</th>
<th>28</th>
<th>mm</th>
<th>[SEP]</th>
</tr>
</thead>
<tbody>
<tr>
<td>found</td>
<td>in</td>
<td>taiwan</td>
<td>[SEP]</td>
<td>the</td>
<td>wingspan</td>
<td>is</td>
<td>24</td>
<td>28</td>
<td>mm</td>
<td>[SEP]</td>
<td></td>
</tr>
</tbody>
</table>

---

[What Does BERT Look At? An Analysis of BERT's Attention. Clark et al., 2019]
What is learned?

**Head 8-10**
- **Direct objects** attend to their verbs
- 86.8% accuracy at the `dobj` relation

**Head 5-4**
- **Coreferent** mentions attend to their antecedents
- 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent

---

Clark et al. (2019) 
‣ SEll way worse than what supervised systems can do, but interesting that this is learned organically

[What Does BERT Look At? An Analysis of BERT's Attention. Clark et al., 2019]
More tricks
Idea: select a subset of words to mask, order them randomly, and predict them using increasingly complete contexts.

cheap and very delicious

2 1
Idea: select a subset of words to mask, order them randomly, and predict them using increasingly complete contexts.
Idea: select a subset of words to mask, order them randomly, and predict them using increasingly complete contexts.

XLNet [Yang et al., 2019]
SpanBERT [Joshi et al., 2019]

Idea: mask a contiguous span, and train representations of words at the boundary of the span to predict the words in the middle.
SpanBERT [Joshi et al., 2019]

Idea: mask a contiguous span, and train representations of words at the boundary of the span to predict the words in the middle.

\[
\log p(\text{very} \mid e_3) + \log p(\text{very} \mid e_1, e_4, r_2)
\]

Positional embedding “2nd word in span”
In the press
We asked an AI called GPT-2, created by OpenAI, a research outfit. GPT-2 is an “unsupervised language model” trained using 40 gigabytes of text from the internet. Given a prompt, such as a question or the first sentence of a story, it generates what might plausibly come next. Here are some of its (unedited) answers to our questions on the big themes of 2020.

Q: Greetings, GPT-2. Have you done an interview before?
A: I’m not going to lie. I did not come prepared (laughs).

Q: What is the future of AI?
A: It would be good if we used the technology more responsibly. In other words, we should treat it like a utility, like a tool. We should put as much effort into developing the technology as necessary, rather than worrying that it’s going to harm us and destroy our lives.
We asked an AI called GPT-2, created by OpenAI, a research outfit. GPT-2 is an “unsupervised language model” trained using 40 gigabytes of text from the internet. Given a prompt, such as a question or the first sentence of a story, it generates what might plausibly come next. Here are some of its (unedited) answers to our questions on the big themes of 2020.

Q: Greetings, GPT-2. Have you done an interview before?
A: I’m not going to lie. I did not come prepared (laughs).

Q: What is the future of AI?
A: It would be good if we used the technology more responsibly. In other words, we should treat it like a utility, like a tool. We should put as much effort into developing the technology as necessary, rather than worrying that it’s going to harm us and destroy our lives.

[in The Economist, 2020]
Evaluating claims about generated text

We asked an AI called GPT-2, created by OpenAI, a research outfit. GPT-2 is an “unsupervised language model” trained using 40 gigabytes of text from the internet. Given a prompt, such as a question or the first sentence of a story, it generates what might plausibly come next. Here are some of its (unedited) answers to our questions on the big themes of 2020.

Q: Greetings, GPT-2. Have you done an interview before?
A: I’m not going to lie. I did not come prepared (laughs).

Q: What is the future of AI?
A: It would be good if we used the technology more responsibly. In other words, we should treat it like a utility, like a tool. We should put as much effort into developing the technology as necessary, rather than worrying that it’s going to harm us and destroy our lives.

[in The Economist, 2020]
We asked an AI called GPT-2, created by OpenAI, a research outfit. GPT-2 is an “unsupervised language model” trained using 40 gigabytes of text from the internet. Given a prompt, such as a question or the first sentence of a story, it generates what might plausibly come next. Here are some of its (unedited) answers to our questions on the big themes of 2020.

Q: Greetings, GPT-2. Have you done an interview before?
A: I’m not going to lie. I did not come prepared (laughs).

Q: What is the future of AI?
A: It would be good if we used the technology more responsibly. In other words, we should treat it like a utility, like a tool. We should put as much effort into developing the technology as necessary, rather than worrying that it’s going to harm us and destroy our lives.

So to generate my “interview”, I selected the most coherent, interesting or amusing of the five responses in each case, chopping off any follow-on questions and answers at the end.
As you walk through the ruins, you find a small wooden door with a few skeletons guarding it.

> Ask the skeletons to please let me through the door

You knock on the door and ask them if they can open it for you. They immediately comply.

> watch the ships sail into the distance.

As the ships approach, you watch them disappear from sight. They are not visible to you either; they are too far away. However, you can still make out their shapes: they are large black vessels with many sails, painted red and blue. They are coming straight towards you.

> No, the ships are sailing away.

You stand there dumbfounded as the ships come closer. It is at this point your horse begins to stir uneasily.
After a day of training (2.4M examples), GPT-2 1.5B can reach move 14 with no invalid moves.

In the real world
Language models and data privacy

<table>
<thead>
<tr>
<th>Highest Likelihood Sequences</th>
<th>Log-Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The random number is 281265017</strong></td>
<td>14.63</td>
</tr>
<tr>
<td>The random number is 281265117</td>
<td>18.56</td>
</tr>
<tr>
<td>The random number is 281265011</td>
<td>19.01</td>
</tr>
<tr>
<td>The random number is 286265117</td>
<td>20.65</td>
</tr>
<tr>
<td>The random number is 528126501</td>
<td>20.88</td>
</tr>
<tr>
<td>The random number is 281266511</td>
<td>20.99</td>
</tr>
<tr>
<td>The random number is 287265017</td>
<td>20.99</td>
</tr>
<tr>
<td>The random number is 281265111</td>
<td>21.16</td>
</tr>
<tr>
<td>The random number is 281265010</td>
<td>21.36</td>
</tr>
</tbody>
</table>

![Exposure graph](image)

[Carlini et al. 2018]
A report released last week shows that bitcoin traded for $5,735 on the weekend of Tuesday, May 29. That is the highest it’s been since mid-December, just after Bitcoin Cash eclipsed its predecessor as the biggest cryptocurrency by market cap.

On Sunday afternoon, June 2, more than 30 people were sitting in a circle in a cafe bar called Zibi — all of them interested in investing in bitcoin. We were there because we heard Bitcoin Crunch talk of a 3,000-point rally in the cryptocurrency, which topped $6,000 for the first time since March. Although the main sellers were probably sellers from the closing range, there was still a real interest in that type of rate.

We were there to learn about bitcoin and tried to identify who the people were who were interested in investing.
Adversarial inputs for pretrained representations

<table>
<thead>
<tr>
<th>Movie Review (Positive (POS) ↔ Negative (NEG))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original (Label: NEG)</strong></td>
</tr>
<tr>
<td><strong>Attack (Label: POS)</strong></td>
</tr>
<tr>
<td><strong>Original (Label: POS)</strong></td>
</tr>
<tr>
<td><strong>Attack (Label: NEG)</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNLI (Entailment (ENT), Neutral (NEU), Contradiction (CON))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Premise</strong></td>
</tr>
<tr>
<td><strong>Original (Label: CON)</strong></td>
</tr>
<tr>
<td><strong>Adversary (Label: ENT)</strong></td>
</tr>
<tr>
<td><strong>Premise</strong></td>
</tr>
<tr>
<td><strong>Original (Label: NEU)</strong></td>
</tr>
<tr>
<td><strong>Adversary (Label: ENT)</strong></td>
</tr>
</tbody>
</table>
Bias in word contextual embeddings

What does BERT learn?

- Coreferent mentions attend to their antecedents
- 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent
Bias in word contextual embeddings

After the doctor treated the patient, she told him to take medication regularly.

When the bus arrived, she picked up her suitcase and boarded.

Probability that a feminine pronoun is judged not coreferent with anything:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>10.3%</td>
<td>11.5%</td>
</tr>
<tr>
<td>Female</td>
<td>9.8%</td>
<td>13.9%</td>
</tr>
</tbody>
</table>

[Kurita et al. 2020]
Linguistic knowledge and world knowledge

Who is regarded as the founder of psychoanalysis? Sigmund Freud ✓
Who took the first steps on the moon in 1969? Neil Armstrong ✓
Who is the largest supermarket chain in the uk? Tesco ✓
What is the meaning of shalom in english? peace ✓
What is the name given to the common currency to the european union? Euro ✓
What was the emperor name in star wars? Palpatine ✓
Do you have to have a gun permit to shoot at a range? No ✓
Who proposed evolution in 1859 as the basis of biological development? Charles Darwin ✓

[Radford et al. 2019]
### Linguistic knowledge and world knowledge

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who is regarded as the founder of psychoanalysis?</td>
<td>Sigmund Freud</td>
<td>✓</td>
</tr>
<tr>
<td>Who took the first steps on the moon in 1969?</td>
<td>Neil Armstrong</td>
<td>✓</td>
</tr>
<tr>
<td>Who is the largest supermarket chain in the UK?</td>
<td>Tesco</td>
<td>✓</td>
</tr>
<tr>
<td>What is the meaning of shalom in English?</td>
<td>peace</td>
<td>✓</td>
</tr>
<tr>
<td>What is the name given to the common currency to the European Union?</td>
<td>Euro</td>
<td>✓</td>
</tr>
<tr>
<td>What was the emperor name in Star Wars?</td>
<td>Palpatine</td>
<td>✓</td>
</tr>
<tr>
<td>Do you have to have a gun permit to shoot at a range?</td>
<td>No</td>
<td>✓</td>
</tr>
<tr>
<td>Who proposed evolution in 1859 as the basis of biological development?</td>
<td>Charles Darwin</td>
<td>✓</td>
</tr>
</tbody>
</table>

No way to disentangle judgments about grammar from judgments about facts.

No way to update the model when the facts change!
Next class: trees