MANCHESTER
1824

The University of Manchester

Contextualised Word Embeddings

Viktor Schlegel

Recap

- Word embeddings: static map from string to vector, trained on co-occurrence in a large training corpus

Static map

f('cat') = ↗

f trained on large corpus
Based on co-occurrence of words

"Static map"

cat	↗
dog	↗
bank	↖
president	↘
…	…

However: same mappings for different words

“Yesterday I had some duck!”

Different POS!

“You need to duck under his punch!”

“I sat by the river bank.”

Different senses!

“I work for the bank.”

Context

So far, we learned how to obtain context when training on a task-specific dataset.

Why not move this contextualisation in the word embeddings?

Language modelling: "Autocomplete"

Task: probability of 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 ✓

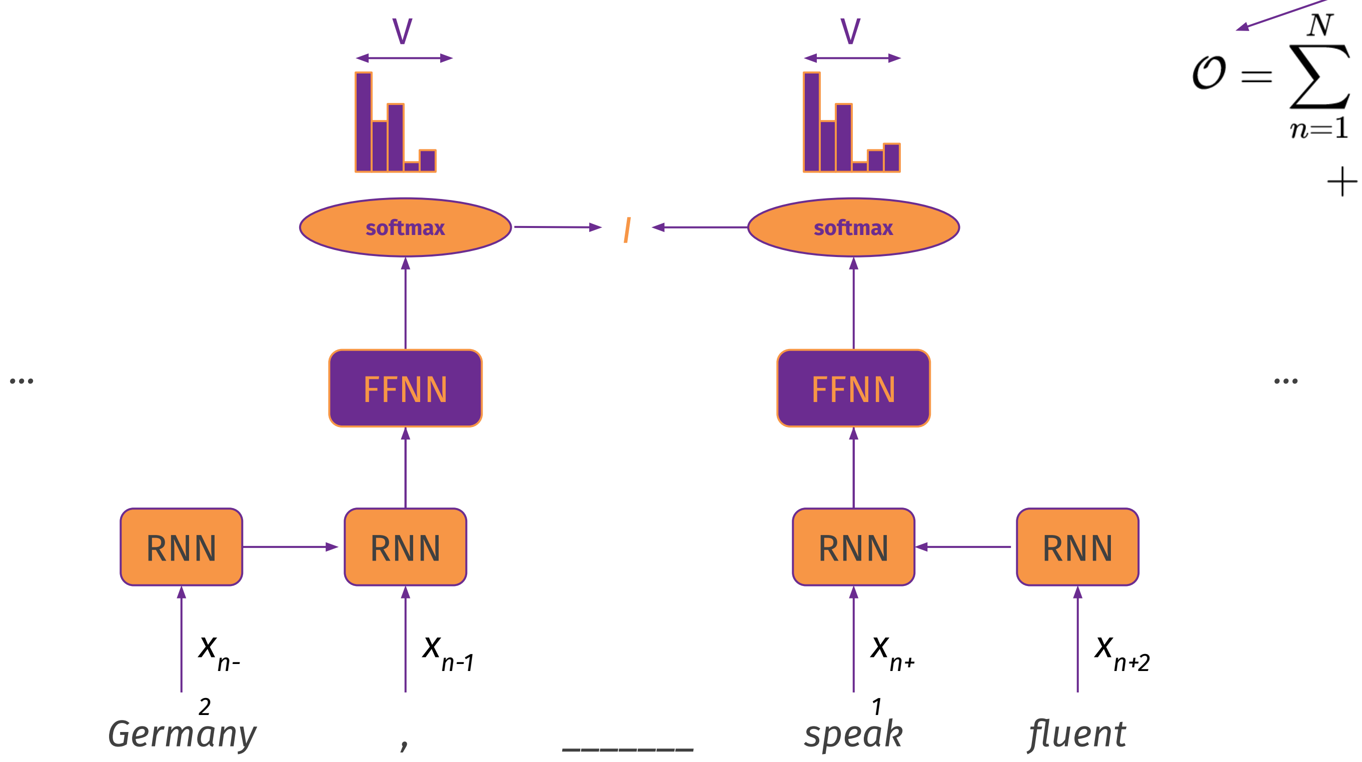
semantic information ✓

self supervision ✓

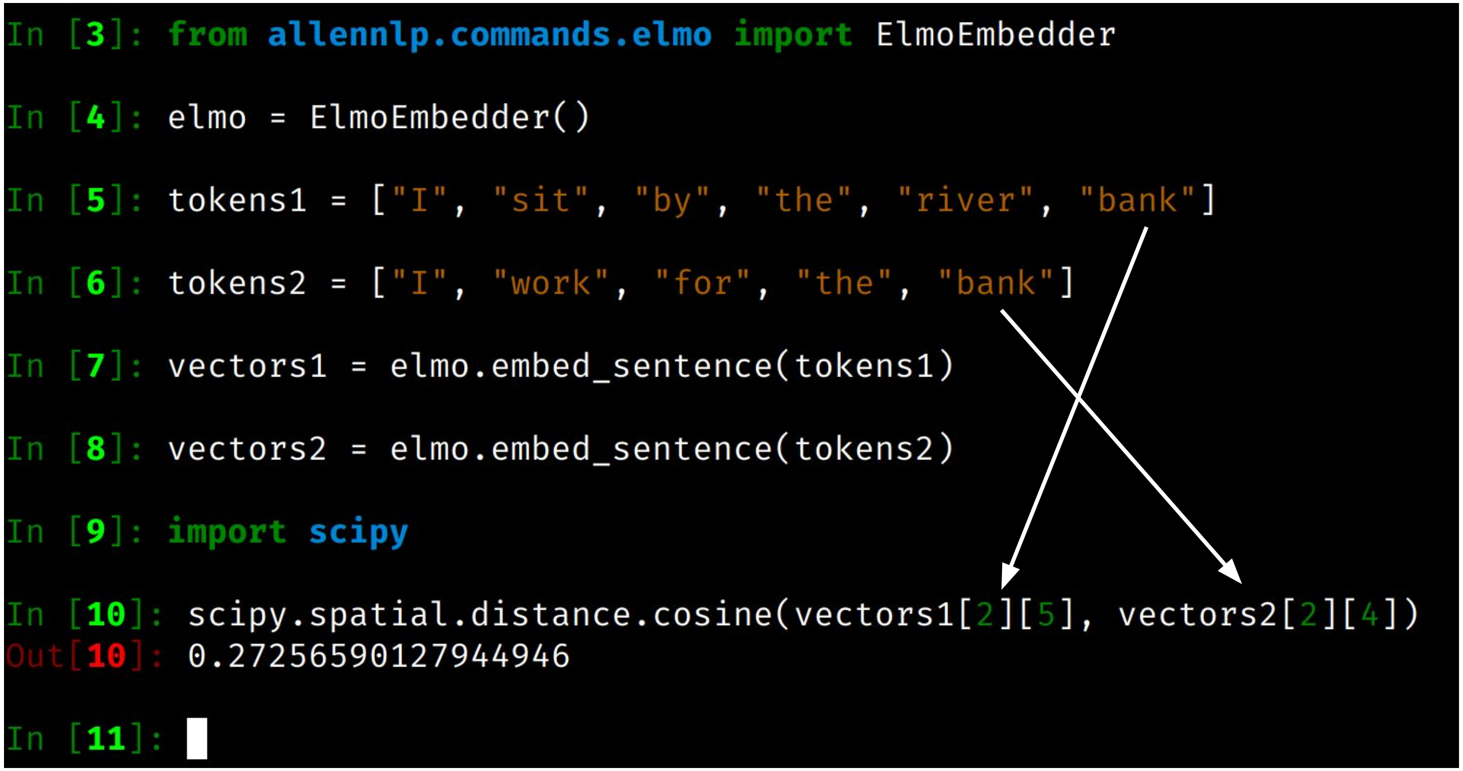
Language modelling with BiLSTM: ELMo

Maximise log likelihood of expected token jointly for both directions

$$\mathcal{O} = \sum_{n=1}^{N} log(P(x_n|x_1,\ldots,x_{n-1}) + log(P(x_n|x_{n+1},\ldots,x_N)$$



Elmo: Contextualised word embeddings



```
In [3]: from allennlp.commands.elmo import ElmoEmbedder

In [4]: elmo = ElmoEmbedder()

In [5]: tokens1 = ["I", "sit", "by", "the", "river", "bank"]

In [6]: tokens2 = ["I", "work", "for", "the", "bank"]

In [7]: vectors1 = elmo.embed_sentence(tokens1)

In [8]: vectors2 = elmo.embed_sentence(tokens2)

In [9]: import scipy

In [10]: scipy.spatial.distance.cosine(vectors1[2][5], vectors2[2][4])
Out[10]: 0.2725659012794946

In [11]: 
```

```
d[word] = vector
f(word, context) = contextualised_vector
```

As a result

- More expressive representations, because contextualised on current context as opposed to “static” word vectors that reflect co-occurrences in training data only
- Better results if used as representation for many natural language processing tasks

Transformer language models

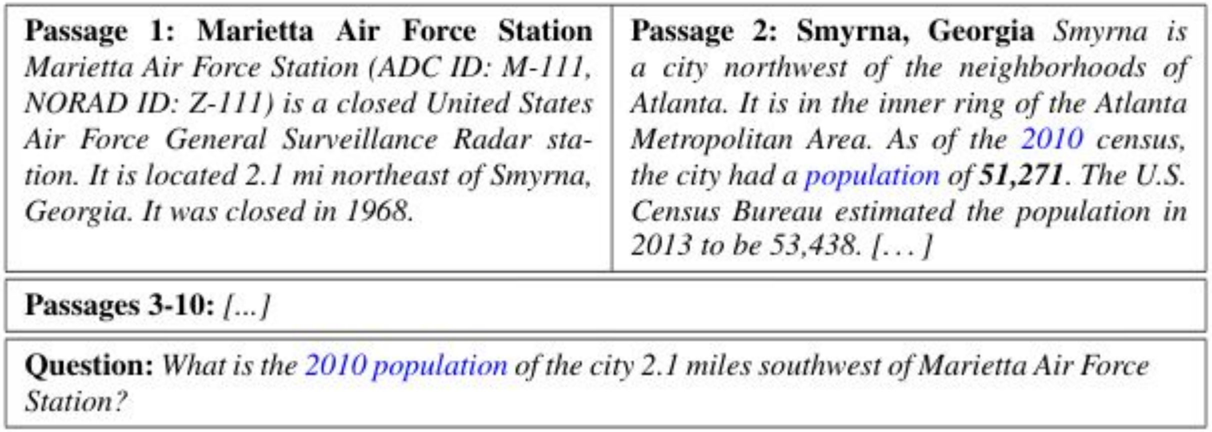
Self-attention to replace RNN:

- self-attention does not require BPTT
- optimisation is faster, so can optimise bigger language models on more data
- Bigger language models lead to improvements on almost all NLP tasks
- Train even bigger language models

Problems: Spurious correlations

Neural networks excel at exploiting **statistical patterns** in data

- No device to distinguish between spurious and robust correlations



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. It is in the inner ring of the Atlanta Metropolitan Area. 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. [...]*

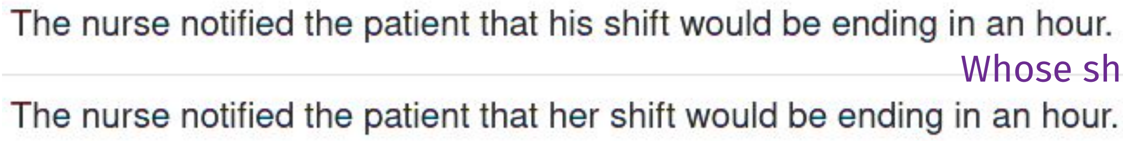
Passages 3-10: *[...]*

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

Problems: Spurious correlations

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 prevail
- However: Majority is not always right!



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

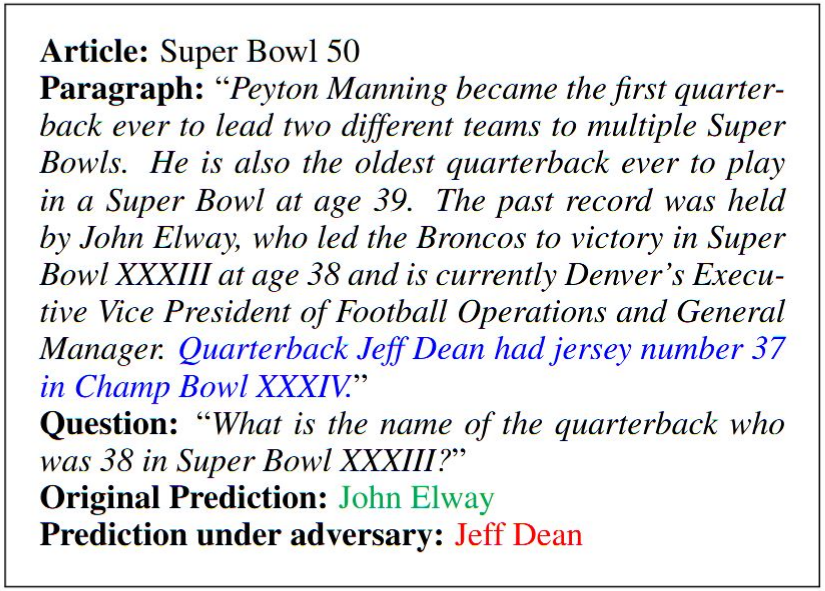
The nurse notified the patient that her shift would be ending in an hour.

Whose shift will end in an hour?

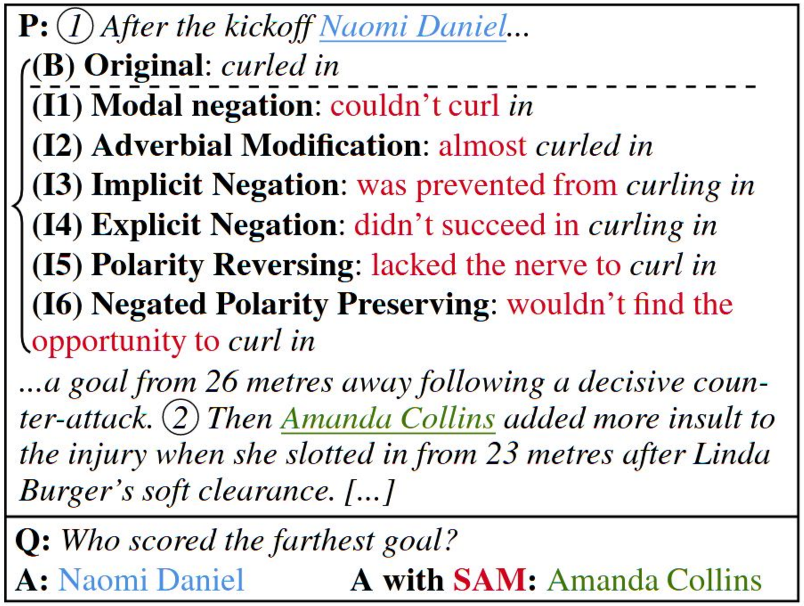
Problems: "Out of distribution" generalisation

Neural networks excel at exploiting **statistical patterns** in data

- If the training data is not representative of the application scenario, the patterns are different



Article: Super Bowl 50
Paragraph: "*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?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

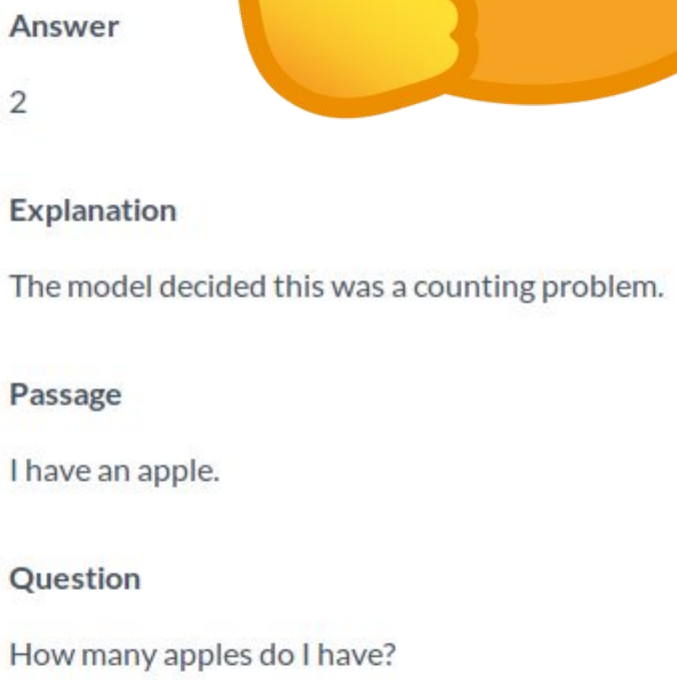


P: ① *After the kickoff Naomi Daniel...*
(B) Original: *curled in*
(I1) Modal negation: couldn't curl *in*
(I2) Adverbial Modification: almost *curled in*
(I3) Implicit Negation: was prevented from *curling in*
(I4) Explicit Negation: didn't succeed in *curling in*
(I5) Polarity Reversing: lacked the nerve to *curl in*
(I6) Negated Polarity Preserving: wouldn't find the opportunity to *curl in*
...a goal from 26 metres away following a decisive counter-attack. ② *Then Amanda Collins added more insult to the injury when she slotted in from 23 metres after Linda Burger's soft clearance. [...]*

Q: *Who scored the farthest goal?*
A: Naomi Daniel **A with SAM:** Amanda Collins

Problems: interpretability

It is not obvious, which statistical patterns are exploited and for which examples this is going to fail.



Answer

2

Explanation

The model decided this was a counting problem.

Passage

I have an apple.

Question

How many apples do I have?

What I learned in this lecture is…

- High-level overview of deep learning
- Expressing NLP tasks as sequence processing
  - Sequence Classification
  - Sequence Labelling
  - Span extraction
  - Sequence to sequence translation

What I learned in this lecture is...

- RNN: process token *sequentially* by incorporating previous information
- Vanishing gradients problem
- LSTM & GRU: manage memory state as solution
- BiRNN: process text sequentially left-to-right and right-to-left
- Attention: for each word in a sentence, learn the relevance of all other tokens

What I learned in this lecture is...

- Language models produce embeddings as functions of the word and the context
- Deep learning based approaches are good in applications where the training data represents the evaluation data well, things get more tricky if it’s not the case

MANCHESTER
1824
The University of Manchester

Links & References

- Jurafsky & Martin: Speech and Language Processing, Chapter 9: Sequence Processing with Recurrent Networks
- Goodfellow, Bengio and Courville: Chapter 10 Sequence Modeling: Recurrent and Recursive Nets
- Denny Britz: Vanishing Gradients problem
- Chris Olah: Understanding LSTMs
- Hochreiter & Schmidhuber: Long Short-term Memory
- Cho et al: On the properties of neural machine translation: Encoder-Decoder approaches (GRU)
- Peters et al: Deep contextualised word representations (ELMo)