Natural Language Processing

Tokenisation and embeddings

Marco Kuhlmann
Department of Computer and Information Science

This session

- Tokenisation and embeddings (Q&A)
- The neural bigram model revisited
- Bias in word representations (in-class assignment)
- Outlook on Unit 2

Tokenisation and embeddings (Q&A)

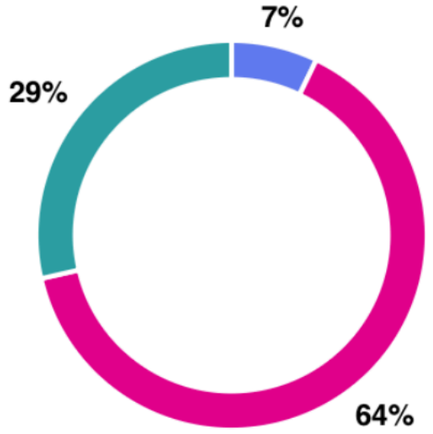
Lecture 1.4, question 5

5. Which of the following is an instance of “pre-training and fine-tuning”? (1 point)

More details

64% of respondents answered this question correctly.

initialise the embedding layer with random weights + train the network on a prediction task	2
initialise the embedding layer with trained weights + train the full network on a prediction task	18 ✓
initialise the embedding layer with trained weights + train the other parts on a prediction task	8



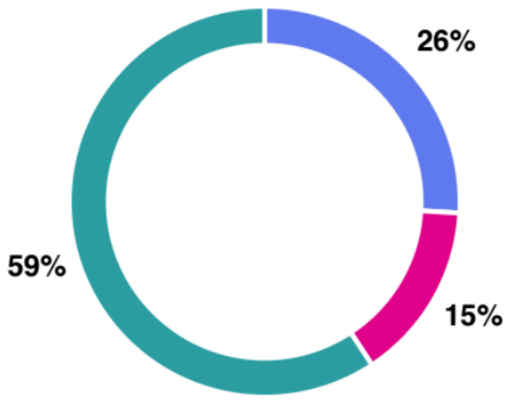
Lecture 1.6, question 1

1. The standard skip-gram model (without negative sampling) can be viewed as a classification problem with k classes. What would be a realistic value for k? (1 point)

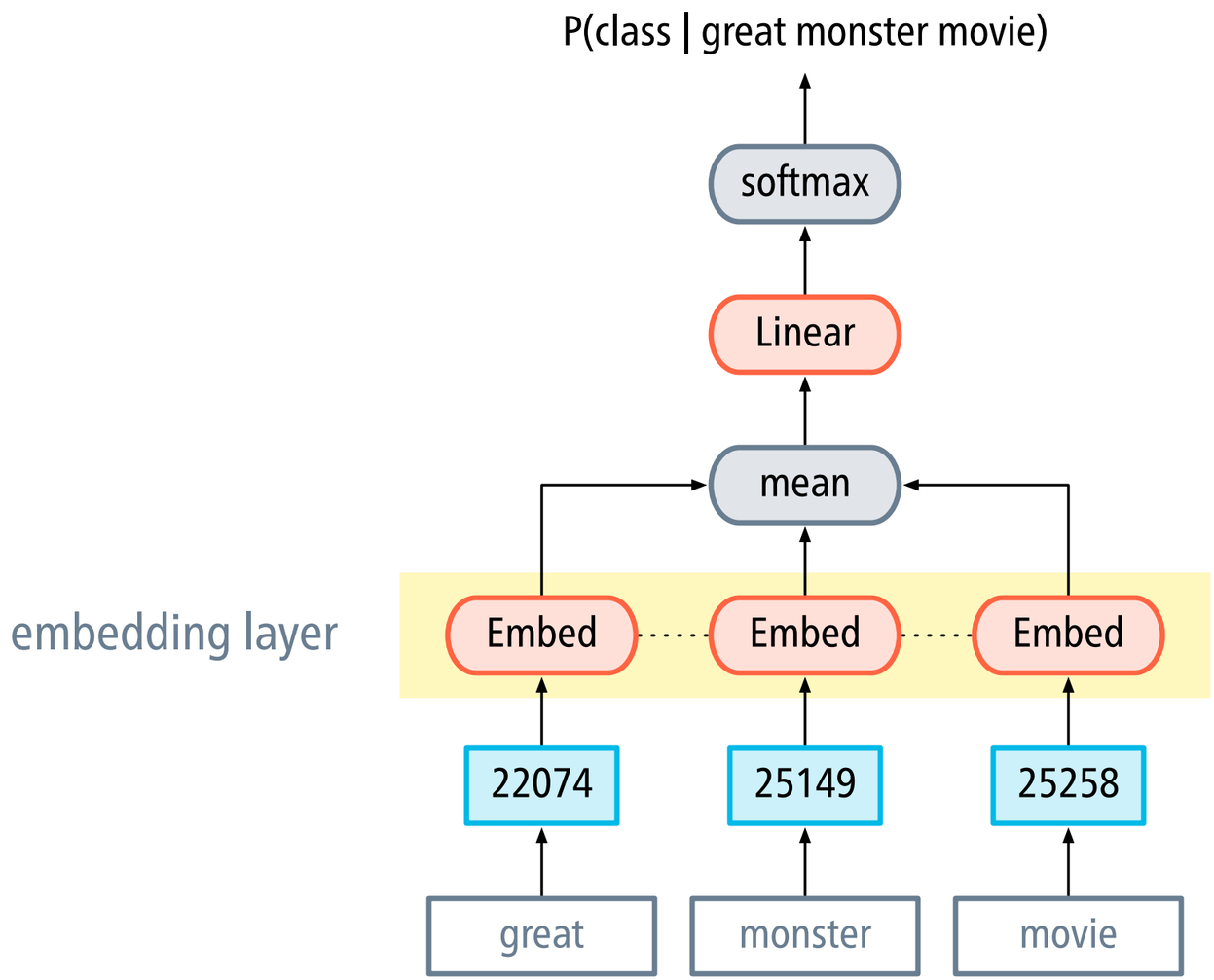
More details

59% of respondents answered this question correctly.

2	7
2,000	4
20,000	16 ✓

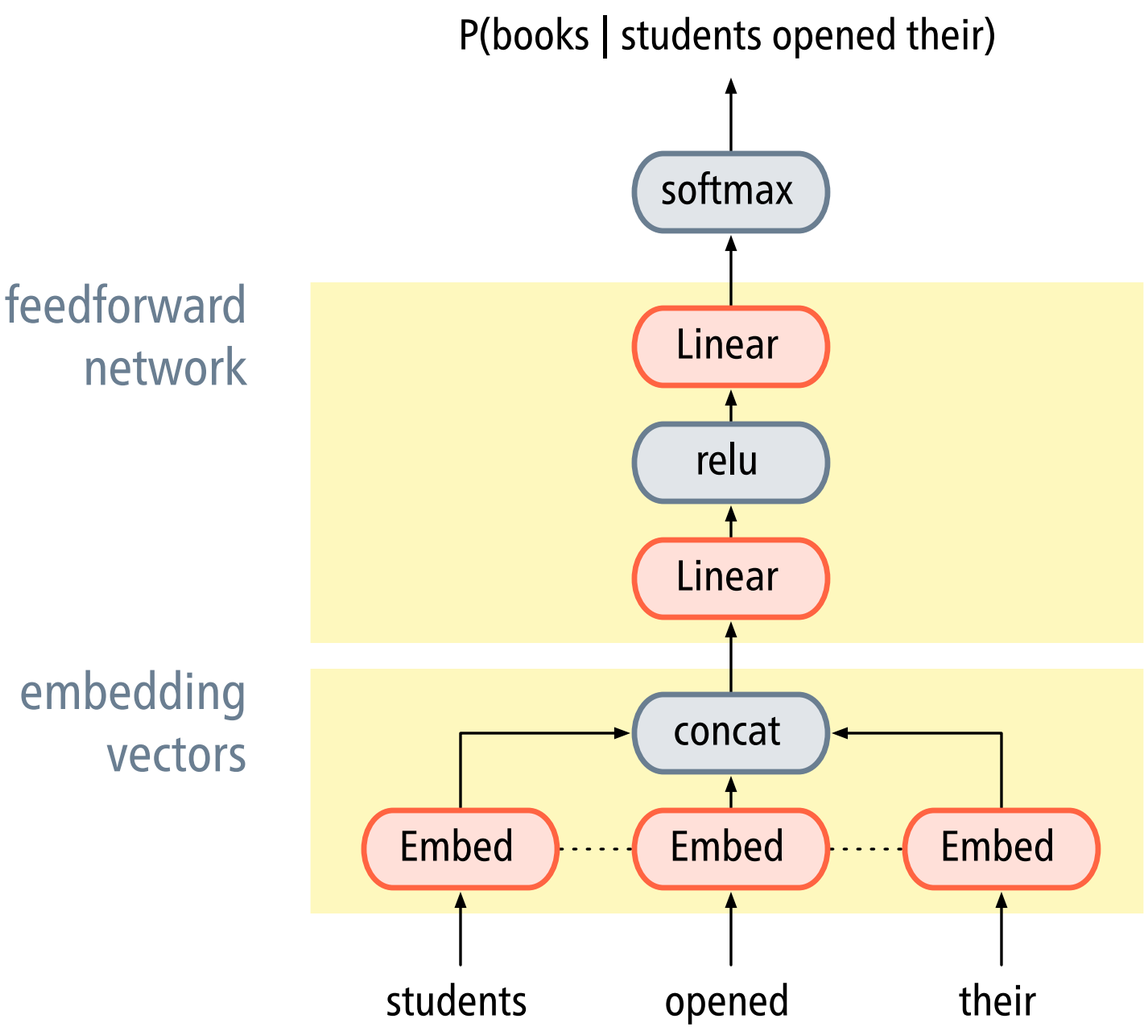


Bag-of-words classifier



The neural bigram model revisited

A neural four-gram model



Bias in word representations

Embedding bias and occupation participation

Figure 1 from Garg et al. (2018)

Partner discussion

- **Partner A:** "The results of Garg et al. clearly show that word embeddings contain harmful biases. There is a risk that we build these biases into our models. We should therefore develop methods for de-biasing embeddings."

- **Partner B:** "The results of Garg et al. simply show statistical correlations in the data; I would not call them harmful biases. The results suggest that word embeddings make an interesting tool for data-driven research in the social sciences."

Vem styr debatten om migrationen?

20 november 2018

Mikael Sönne

Hur har det offentliga samtalet om invandring förändrats i Sverige? Och vem ligger bakom den förändringen - politikerna, medierna eller allmänheten i sociala medier? Det ska ett nytt forskningsprojekt vid LiU försöka ta reda på.

"Att kunna analysera fritt skriven text frigör forskningen", säger Marc Keuschnigg. B ild: Mikael Sönne

In-class assignment

https://forms.office.com/e/R447gwBmcX

Outlook on Unit 2