University of BRISTOL

9.3 Fine-tuning

Edwin Simpson

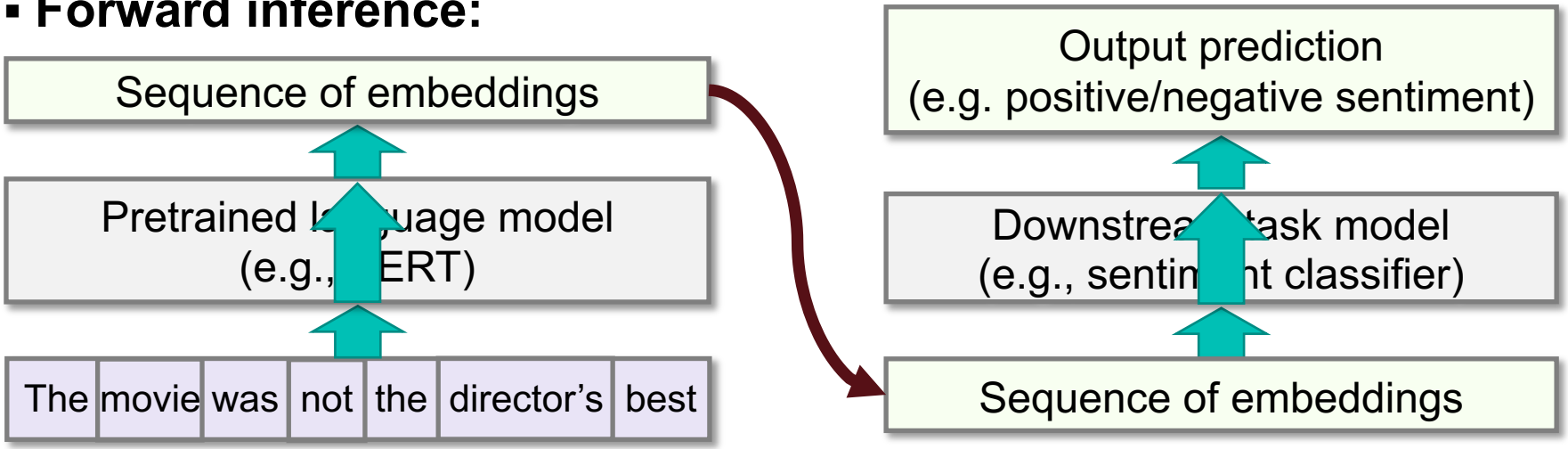
Department of Computer Science,

University of Bristol, UK.

bristol.ac.uk

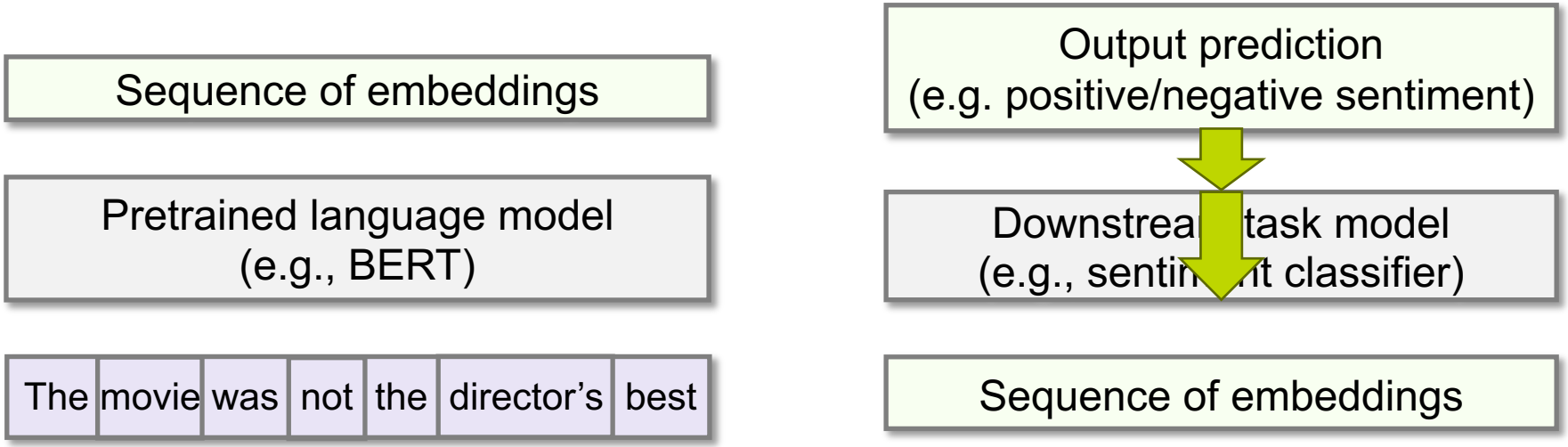
Applying BERT to Downstream Tasks

- Pretrained models like BERT can compute contextualised word embeddings for the text in our downstream task's dataset.
- As with skip-gram, we can **freeze** the embedding model.
- **Forward inference:**



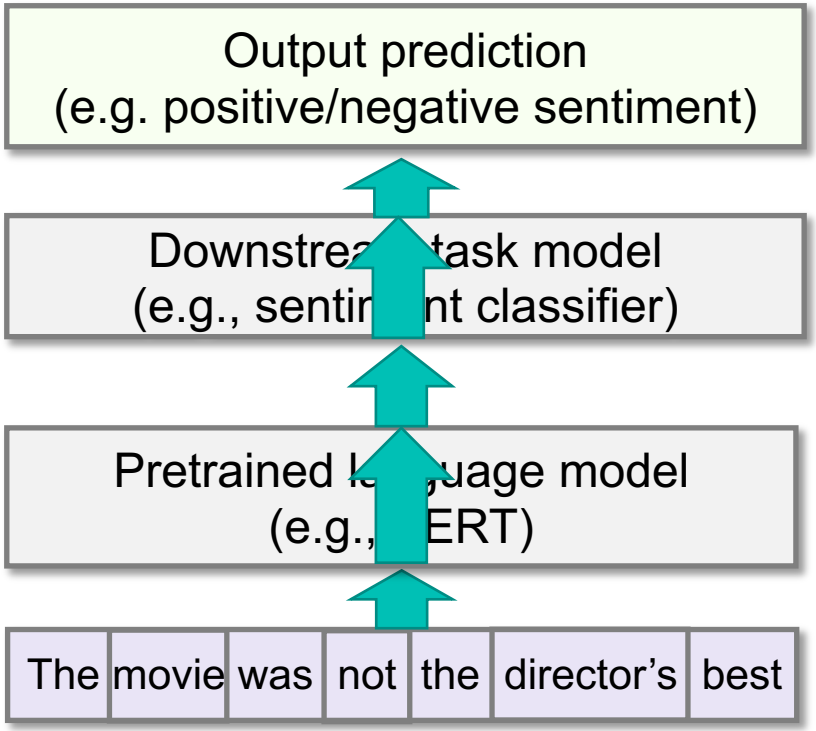
Applying BERT to Downstream Tasks

- If the embedding model is frozen, there is not further training of BERT on the downstream task.
- **Backpropagation:**



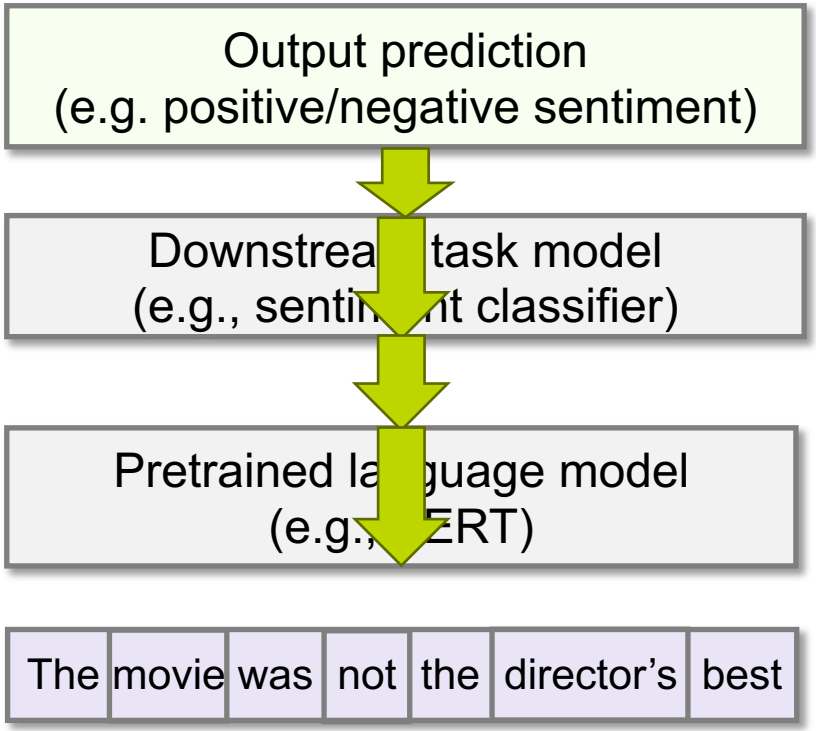
Fine-tuning BERT to Downstream Tasks

- As with skip-gram embeddings, we can also **fine-tune** the embeddings on the downstream task.
- The downstream model consists of additional layers on top of the pretrained BERT model.
- **Forward inference:**



Fine-tuning BERT to Downstream Tasks

- During training, the losses are propagated back through the BERT layers as well as the downstream layers.
- **Backpropagation:**



Frozen versus Fine-tuned BERT

To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks, Peters et al., 2019.

Frozen

- Training cost is much lower (~10x) as we only train weights on the final layers.
- Performance can be competitive.

Fine-tuned

- Can perform better as it adapts embeddings to downstream task.
- Danger of overfitting to small downstream datasets.

	NER (span F1)	Sentiment (accuracy)	Paraphrase detection (acc.)	Sentence similarity (Pearson corr.)
Frozen	.922	93.0	78.1	82.9
Fine-tuned	.924	93.5	84.8	87.1

Frozen versus Fine-tuned BERT

To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks, Peters et al., 2019.

When is fine-tuning effective?

- When pretraining and downstream tasks are similar.
- Fine-tuning uses fewer task-specific parameters so may overwrite information in the BERT layers to help it fit the training data.
- But it can also emphasise useful information from pretraining if tasks as similar.

	NER (span F1)	Sentiment (accuracy)	Paraphrase detection (acc.)	Sentence similarity (Pearson corr.)
Frozen	.922	93.0	78.1	82.9
Fine-tuned	.924	93.5	84.8	87.1

Summary

- Pretrained language models can provide embeddings as features for downstream tasks.
- We can also fine-tune BERT to adapt it to the downstream task.
- In fine-tuning, gradients are propagated back through the pretrained model and the weights are updated.
- This can increase performance at additional computational cost.