Natural Language Processing

Lecture 5.5

# BERT

Marco Kuhlmann
Department of Computer and Information Science

# Bidirectional Encoder Representations from Transformers

- The acronym **BERT** stands for 'Bidirectional Encoder Representations from Transformers'.

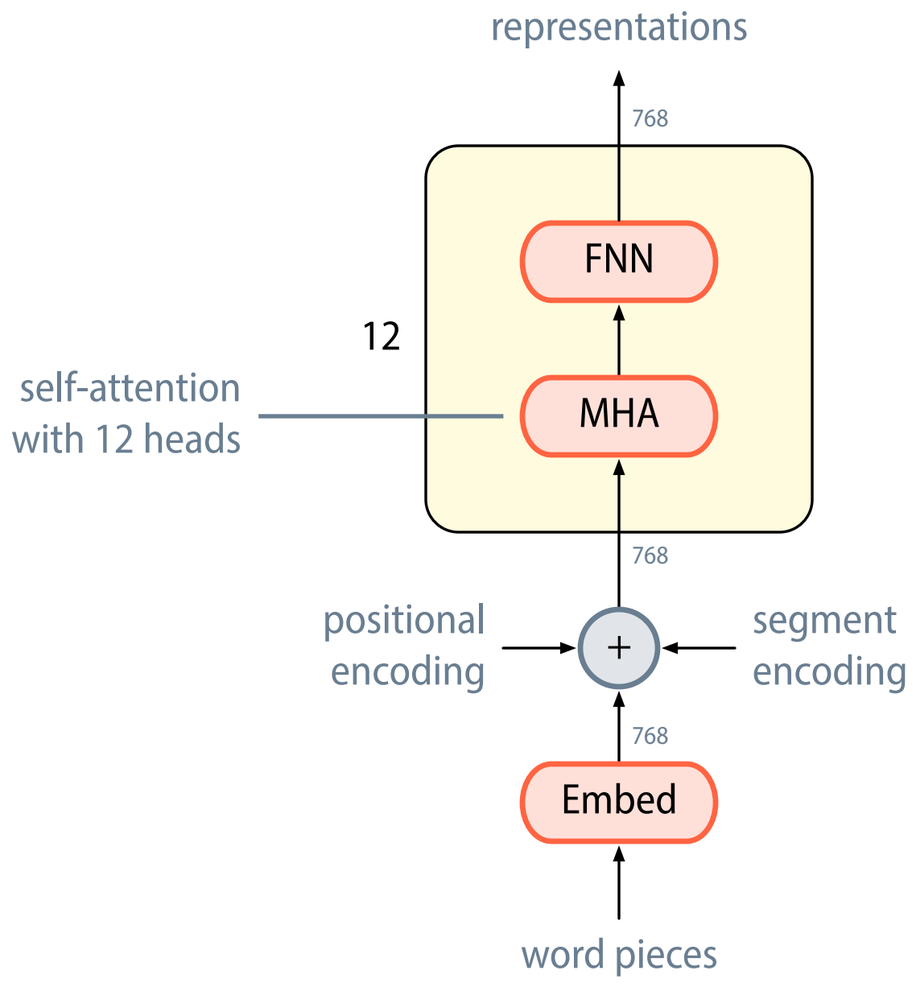
- BERT can be used to pre-train contextualised word representations on unlabelled text.

  original pre-training corpus had 3.3 billion words

- These pre-trained representations can then be fine-tuned on a wide range of different tasks.

Devlin et al. (2019)

# BERT (Base model)




Devlin et al. (2019)

# Pre-training tasks

- **Masked Language Model**

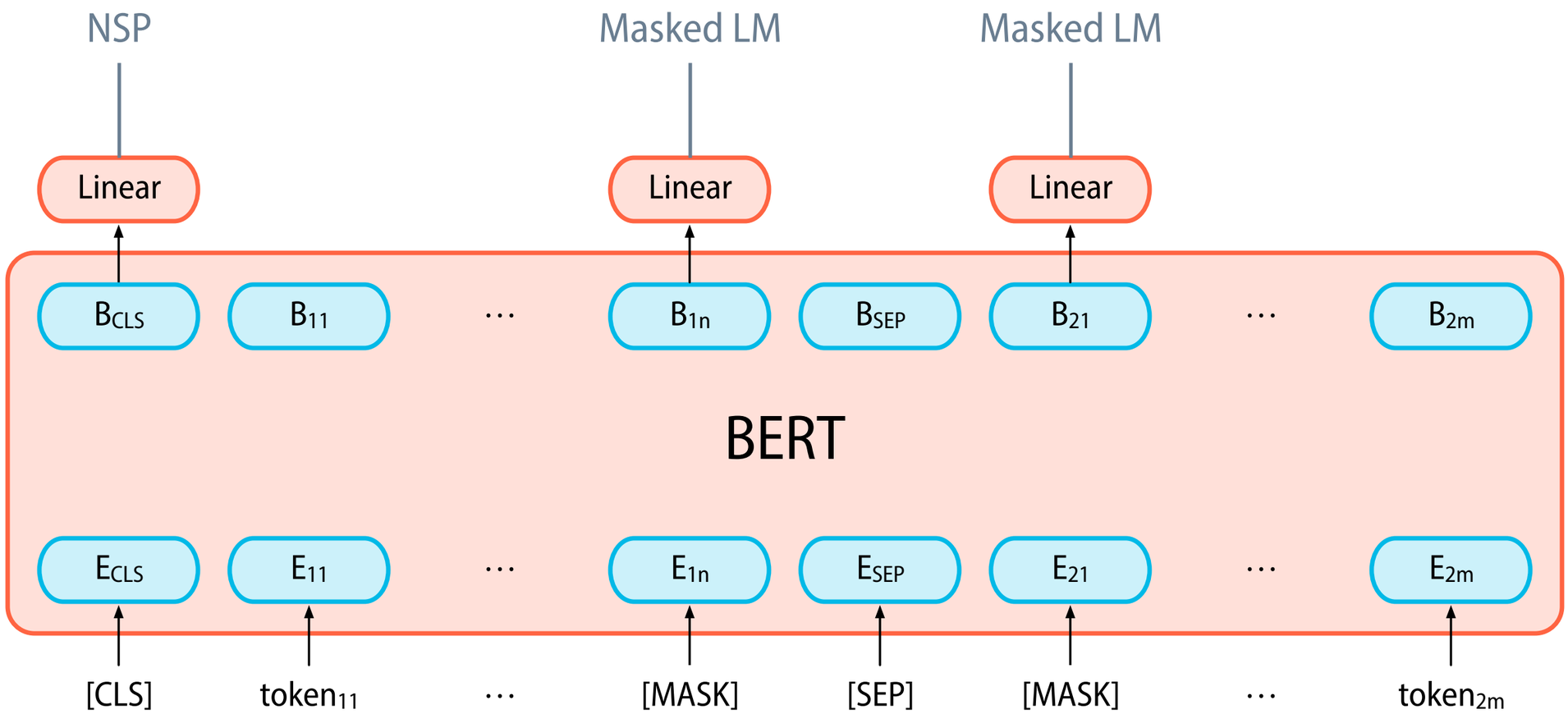
  During pre-training, 15% of all tokens are masked at random.
  The model is trained to predict the masked tokens.

  similar to standard LM pre-training tasks, but non-directional

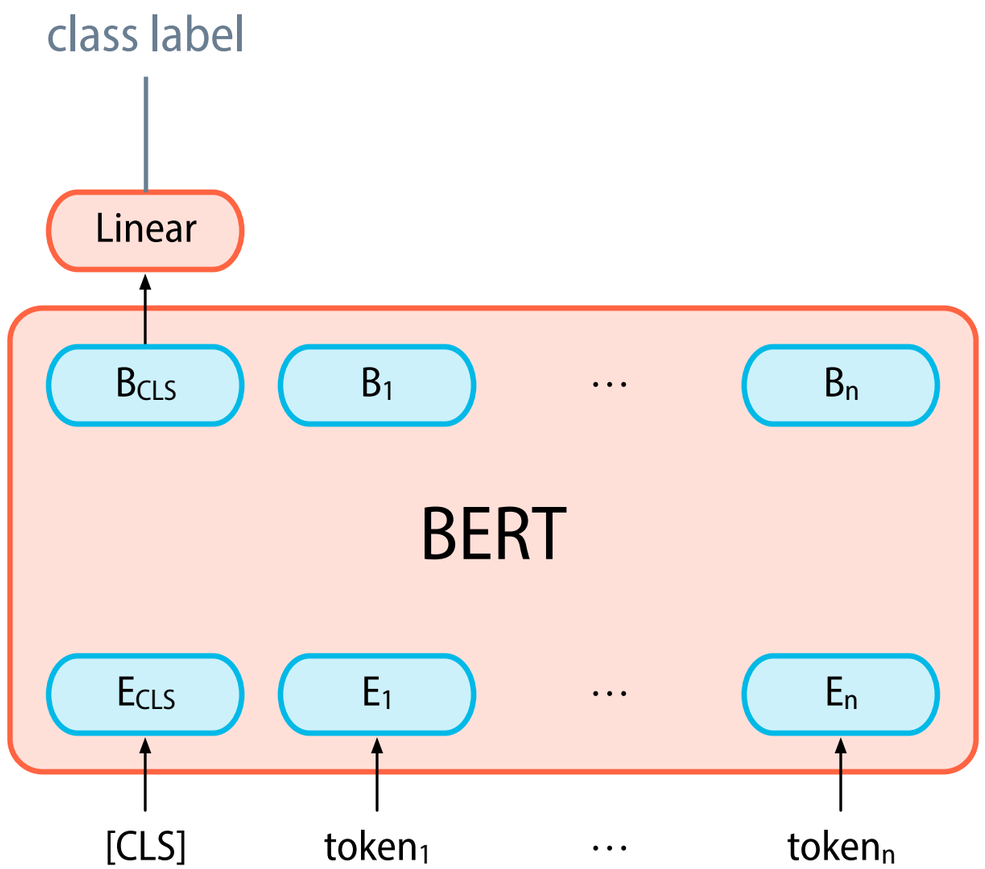
- **Next Sentence Prediction (NSP)**

  Pre-training uses sentence pairs. The model is trained to predict whether the two sentences are adjacent in the training data.

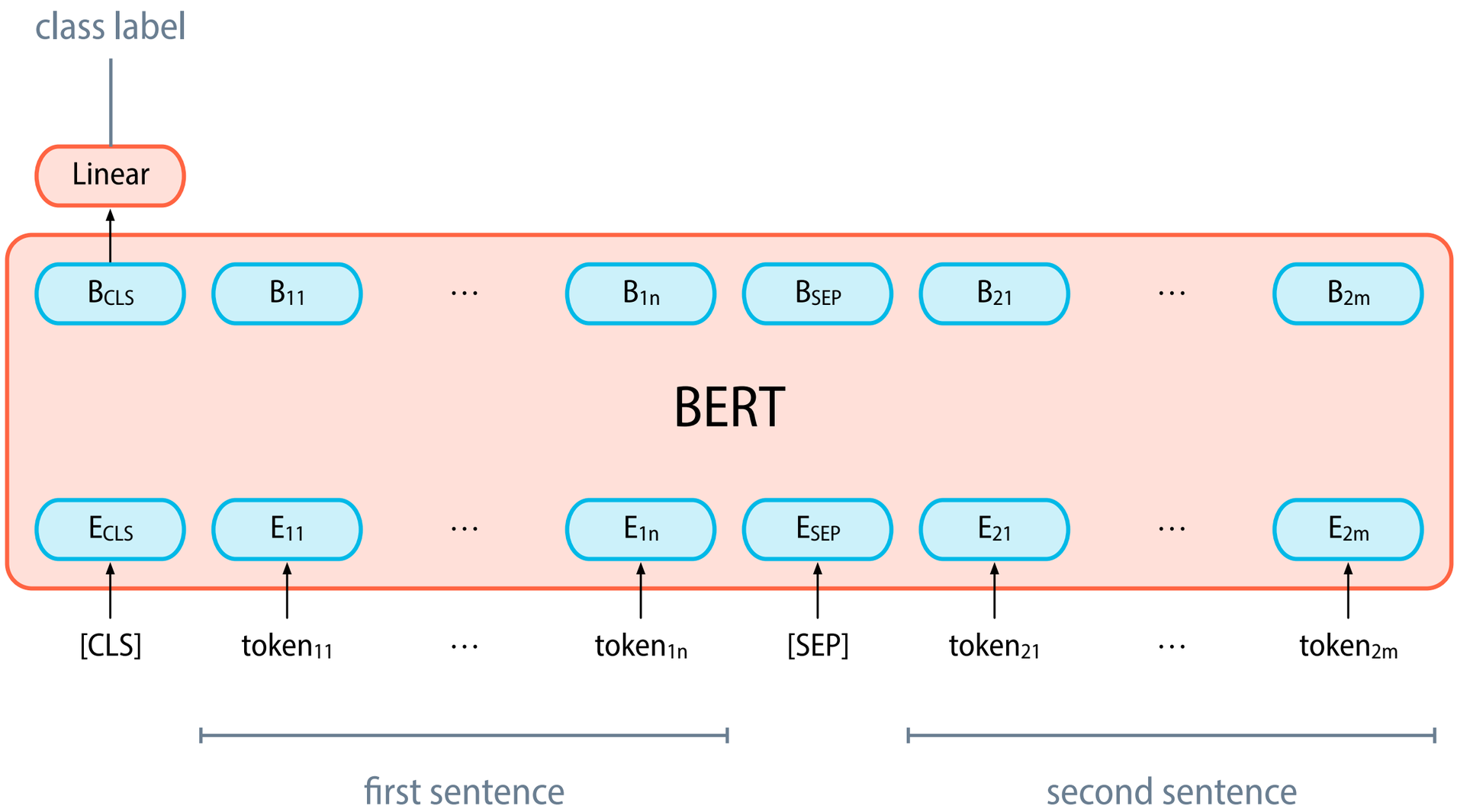
# Pre-training



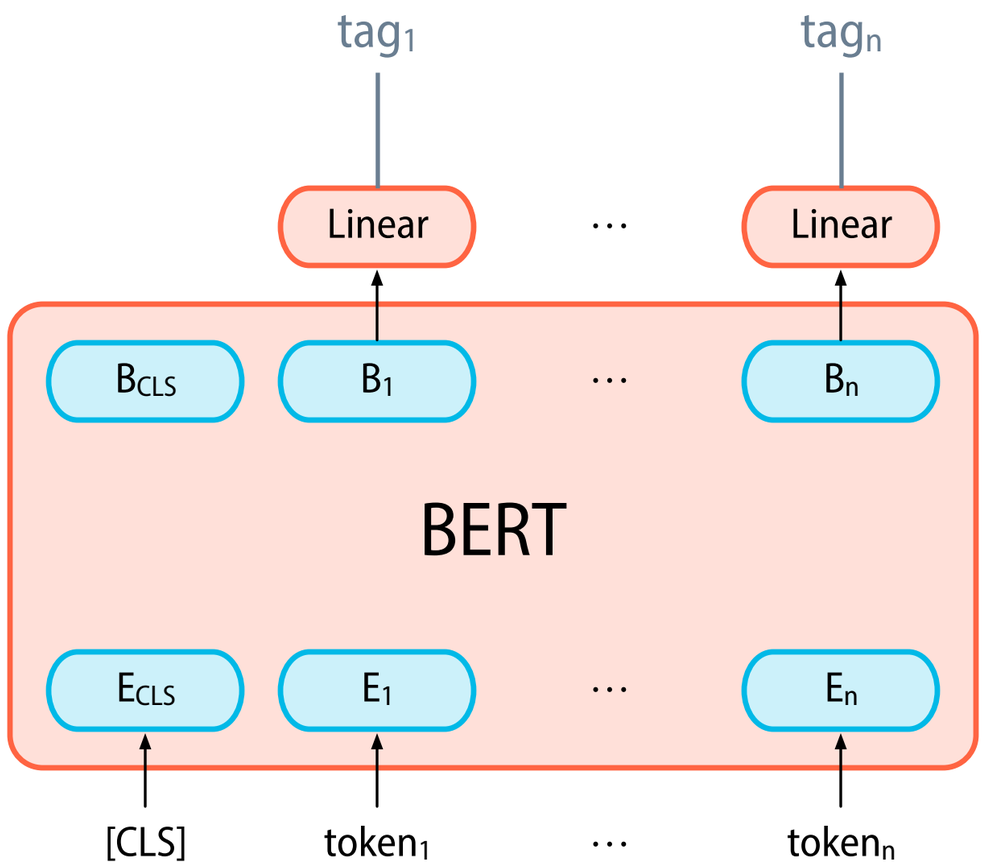
# Fine-tuning on a single-sentence classification task



# Fine-tuning on a sentence-pair classification task



# Fine-tuning on a sequence labelling task



# Performance

| | GLUE |
|---|---|
| Bi-LSTM + ELMo + Attention | 71.0 |
| Previous state-of-the-art | 74.0 |
| BERT (base) | 79.6 |
| BERT (large) | **82.1** |

GLUE test results, scored by the evaluation server – Devlin et al. (2019)

# Recent BERT-like models

- RoBERTa uses an improved recipe for pre-training BERT models and trains on a larger data set.

  Liu et al. (2019)

- DistilBERT distills BERT into a slightly less enormous model using a student–teacher approach.

  Sanh et al. (2019)

- The Swedish Royal Library has released a Swedish BERT.

  Available via HuggingFace Transformers.