# Using the Transformer

# Ernie (Sun et al., 2019a, 2019b, 2021)


**Learning goals**

- Understand the differences to BERT
- Dynamic Masking

# ERNIE VS. BERT

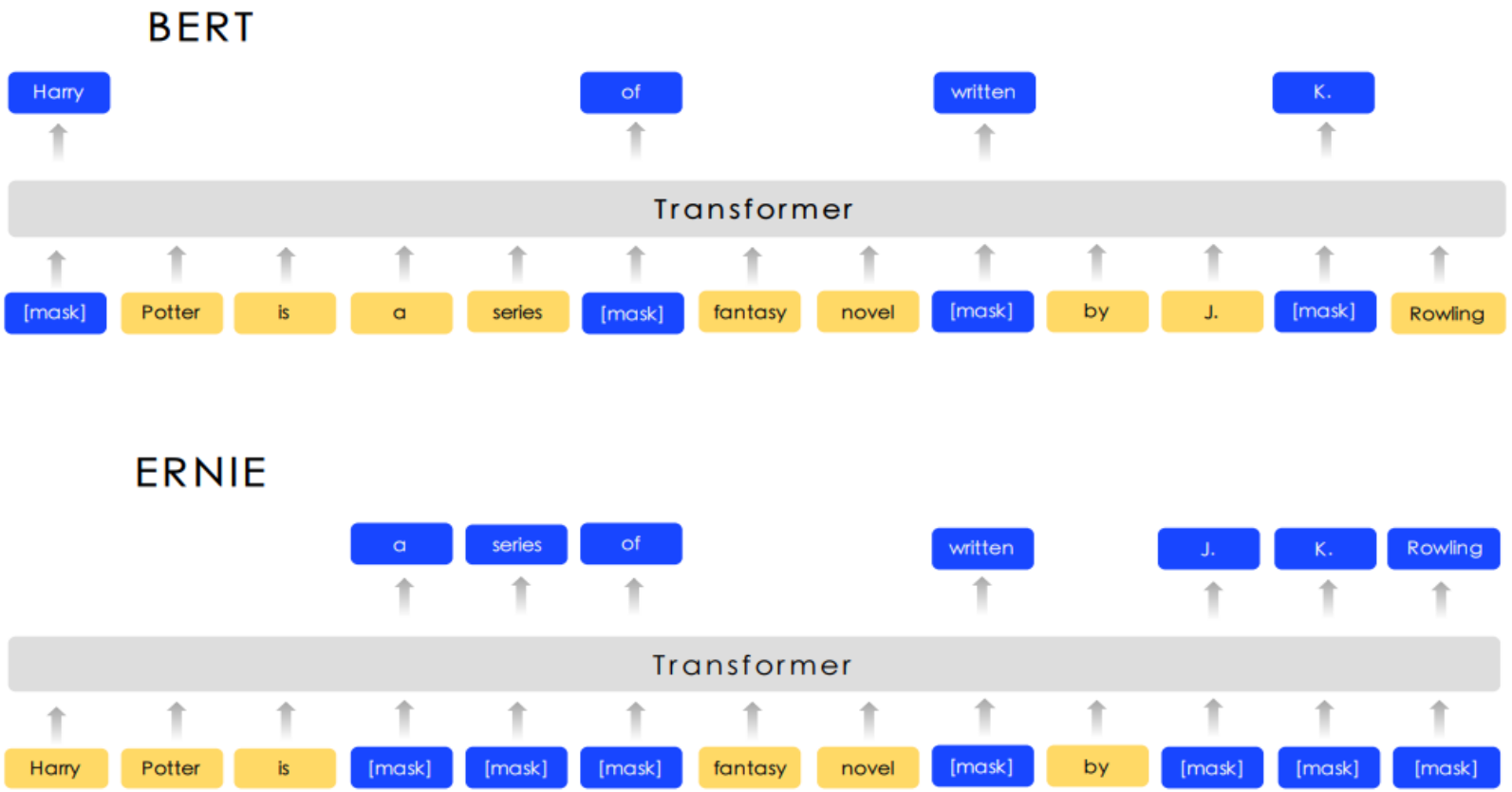


Figure 1: The different masking strategy between BERT and ERNIE

Source: Sun et al. (2019)

# ROBERTA

**Improvements in Pre-Training:**

- Authors claim that BERT is seriously undertrained
- Change of the `MASK`ing strategy
  $\rightarrow$ BERT masks the sequences once before pre-training
  $\rightarrow$ RoBERTa uses dynamic `MASK`ing
  $\Rightarrow$ RoBERTa sees the same sequence `MASK`ed differently
- RoBERTa does not use the additional NSP objective during pre-training
- 160 GB of pre-training resources instead of 13 GB
- Pre-training is performed with larger batch sizes (8k)

# DYNAMIC VS. STATIC MASKING ▸ LIU ET AL., 2019

**Static Masking (BERT):**

- Apply `MASK`ing procedure to pre-training corpus once
- (additional for BERT: Modify the corpus for NSP)
- Train for approximately 40 epochs

**Dynamic Masking (RoBERTa):**

- Duplicate the training corpus *ten* times
- Apply `MASK`ing procedure to each duplicate of the pre-training corpus
- Train for 40 epochs
- Model sees each training instance in ten different "versions" (each version four times) during pre-training