Seq2seq Pre-training
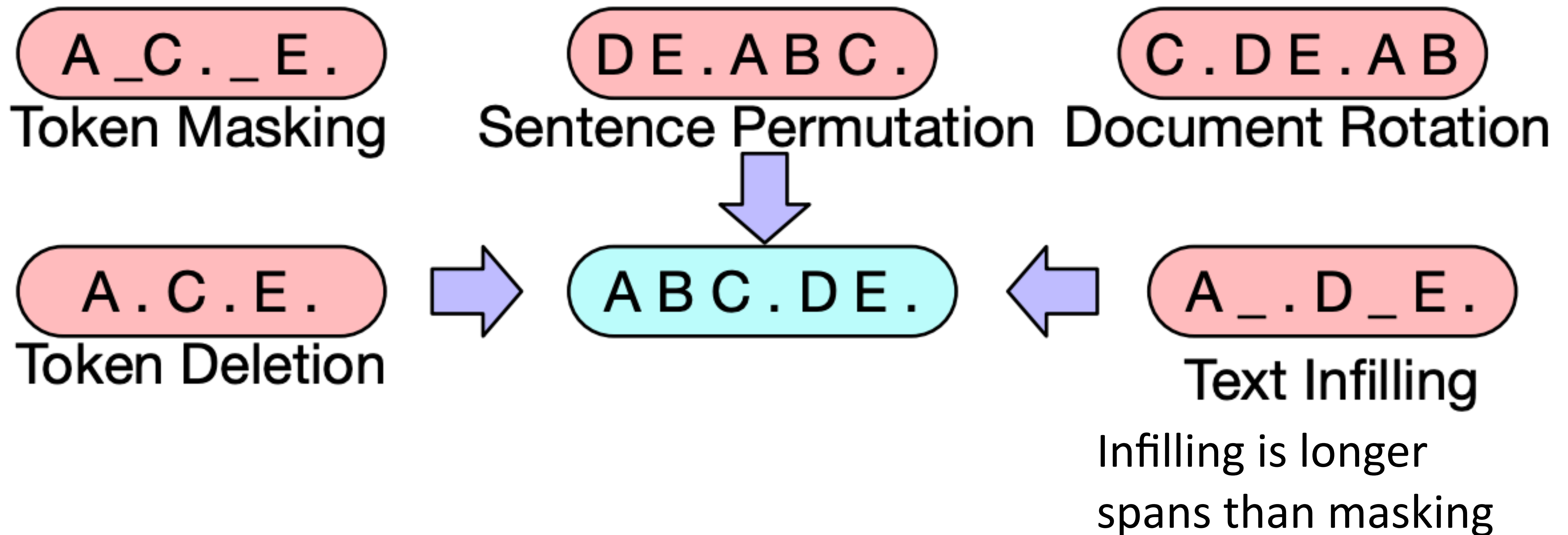
- LMs P($\mathbf{w}$): trained unidirectionally

- Masked LMs: trained bidirectionally but with masking

- How can we pre-train a model for P($\mathbf{y}|\mathbf{x}$)?

- Why was BERT effective? Predicting a mask requires some kind of text "understanding":

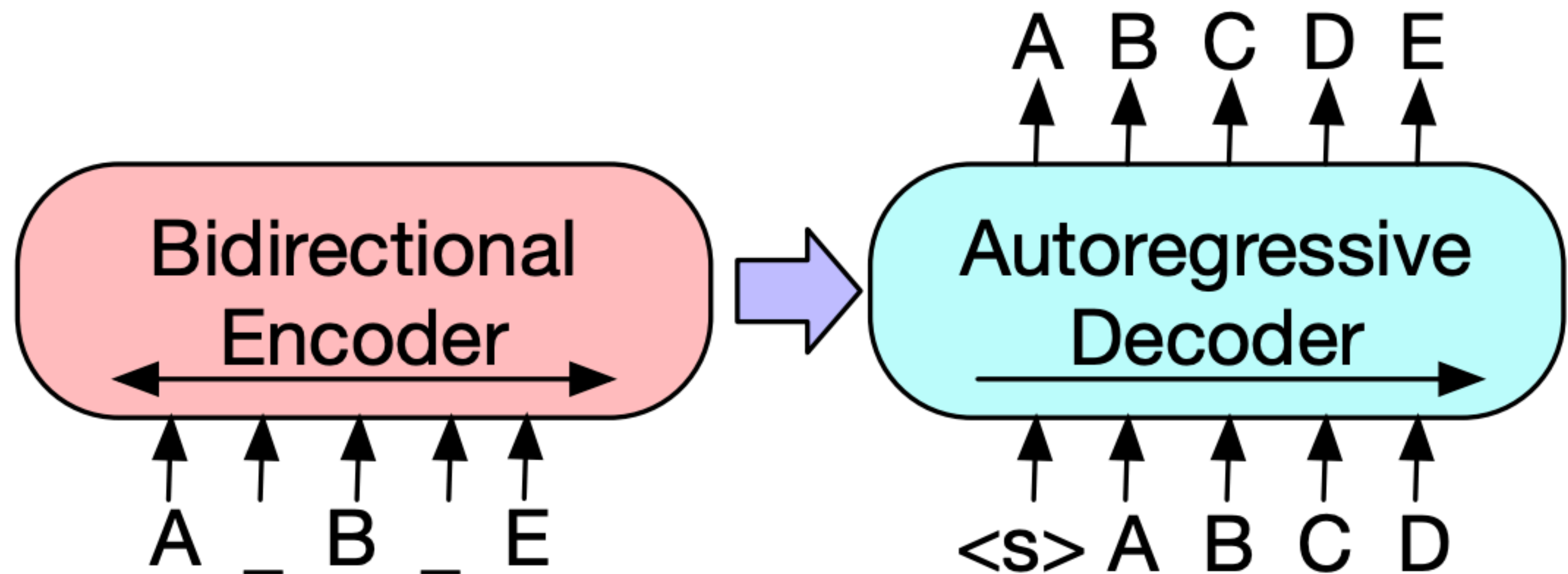- What would it take to impart the same "skills" for sequence prediction?

BART

Token Masking: A _ C . _ E .

Sentence Permutation: D E . A B C .

Document Rotation: C . D E . A B

Token Deletion: A . C . E .

ABC.DE.

Text Infilling: A _ . D _ E .

Infilling is longer
spans than masking

‣ Several possible strategies for corrupting a sequence are explored in
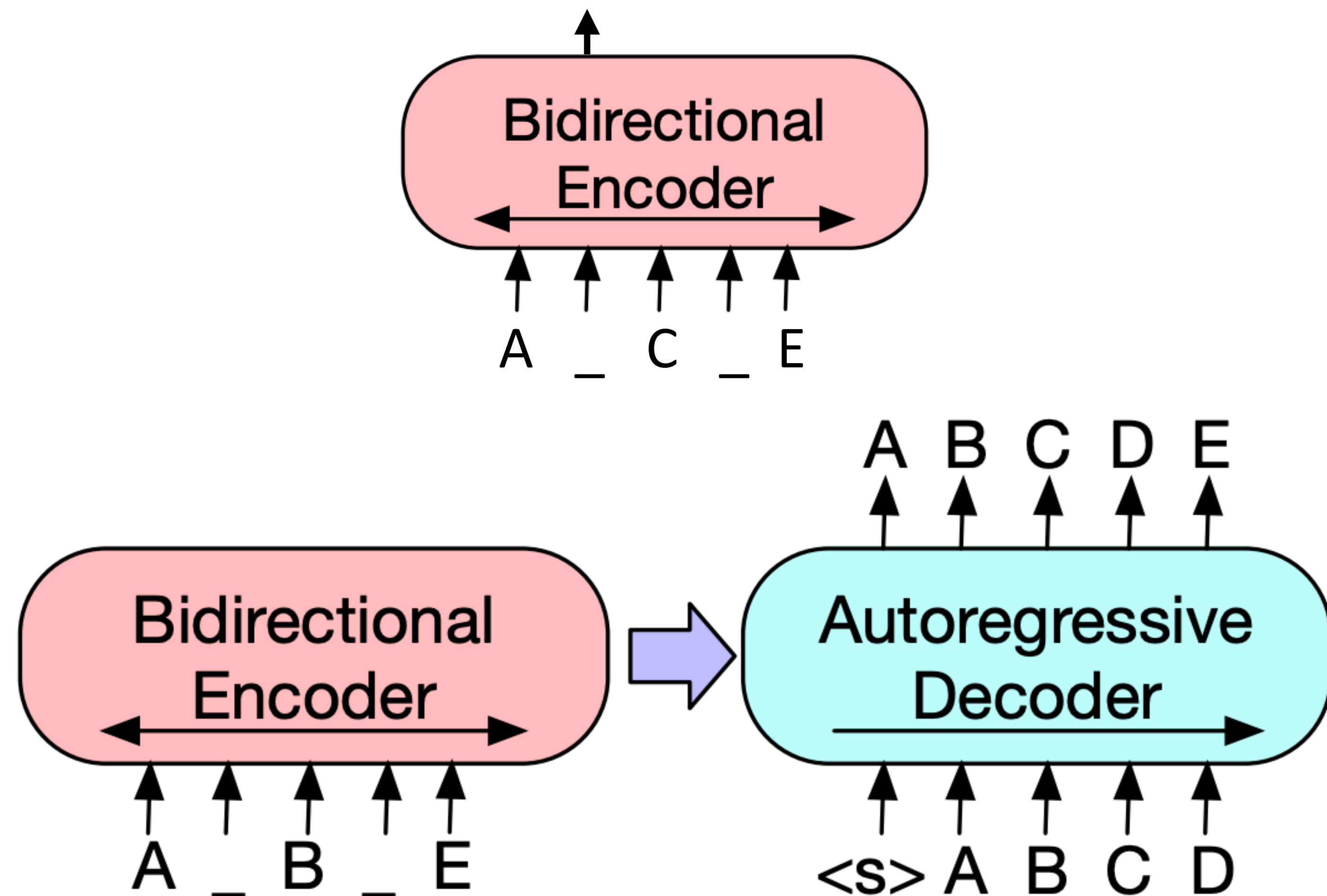 the BART paper

Lewis et al. (2019)

BART

‣ Sequence-to-sequence Transformer trained on this data: permute/ make/delete tokens, then predict full sequence autoregressively

Lewis et al. (2019)

BERT vs. BART

‣ BERT: only parameters are an encoder, trained with masked language modeling objective. Cannot generate text or do seq2seq tasks

‣ BART: both an encoder and a decoder. Can also use just the encoder wherever we would use BERT

Lewis et al. (2019)

BART for Summarization

‣ **Pre-train** on the BART task: take random chunks of text, noise them according to the schemes described, and try to "decode" the clean text

‣ **Fine-tune** on a summarization dataset: a news article is the input and a summary of that article is the output (usually 1-3 sentences depending on the dataset)

This is the first time anyone has been recorded to run a full marathon of 42.195 kilometers (approximately 26 miles) under this pursued landmark time. It was not, however, an officially sanctioned world record, as it was not an "open race" of the IAAF. His time was 1 hour 59 minutes 40.2 seconds. Kipchoge ran in Vienna, Austria. It was an event specifically designed to help Kipchoge break the two hour barrier.

↓

Kenyan runner Eliud Kipchoge has run a marathon in less than two hours.

Lewis et al. (2019)