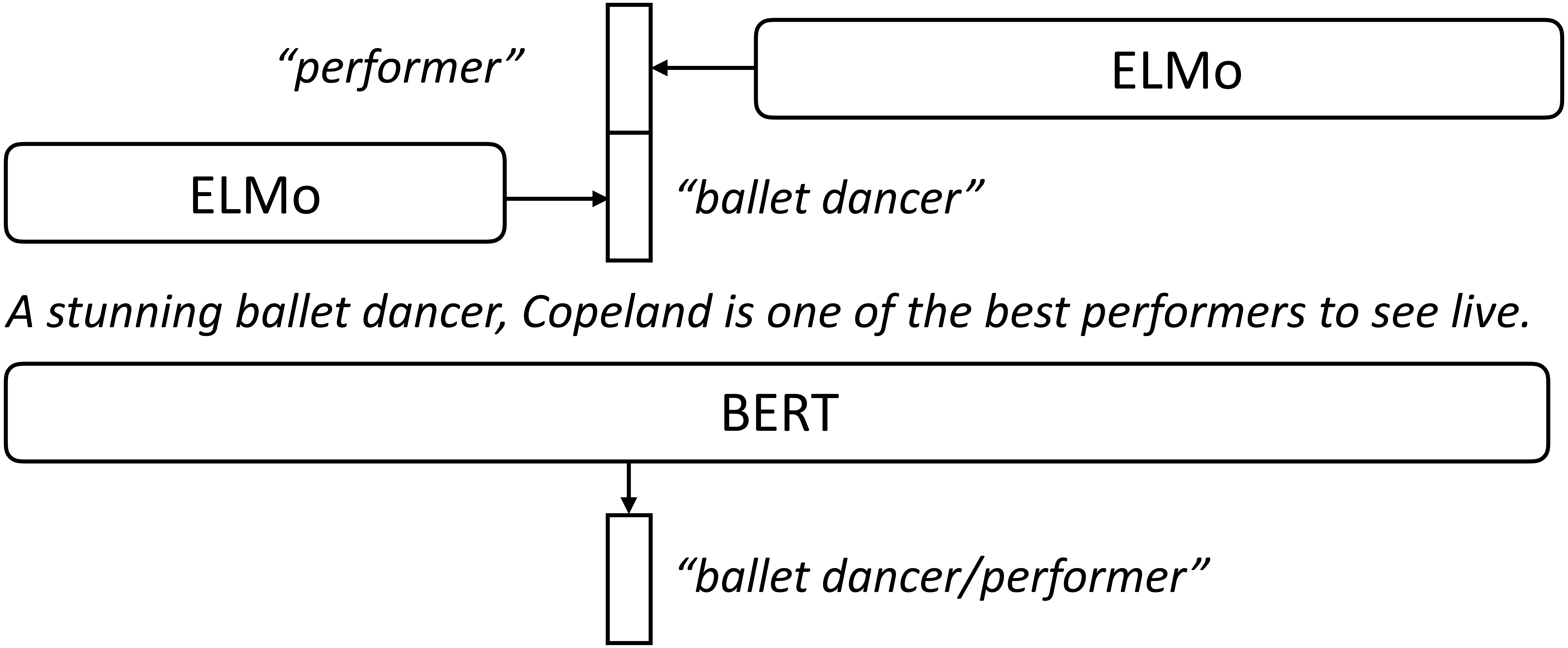
BERT

- AI2 made ELMo in spring 2018, GPT (transformer-based ELMo) was released in summer 2018, BERT came out October 2018
- Four major changes compared to ELMo:
  - Transformers instead of LSTMs
  - Bidirectional model with "Masked LM" objective instead of normal LM
  - Fine-tune instead of freeze at test time
  - Operates over word pieces (byte pair encoding)

BERT

- ELMo is a unidirectional model (as is GPT): we can concatenate two unidirectional models, but is this the right thing to do?
- ELMo reprs look at each direction in isolation; BERT looks at them jointly



"performer"

ELMo

ELMo

"ballet dancer"

A stunning ballet dancer, Copeland is one of the best performers to see live.

BERT

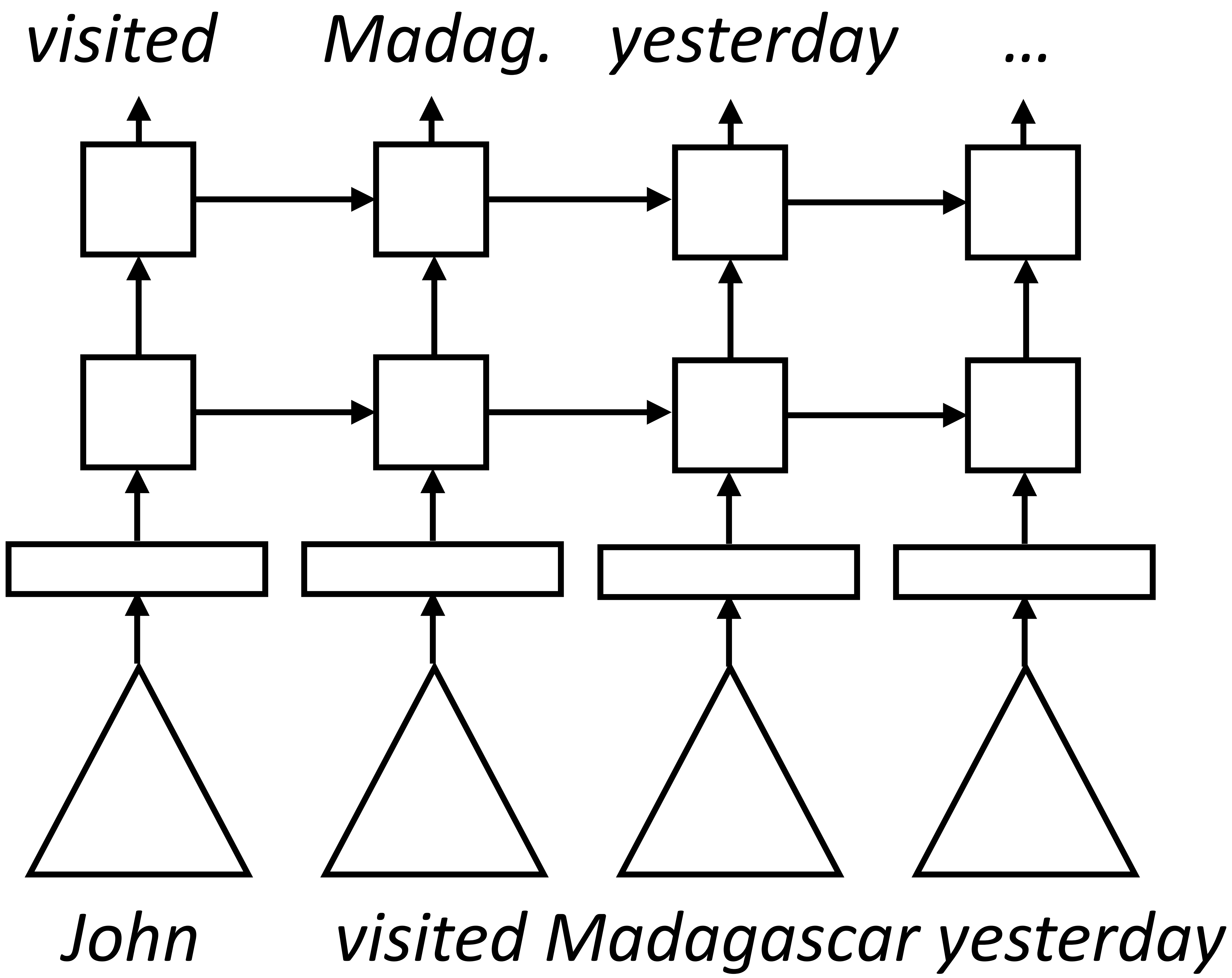
"ballet dancer/performer"

Devlin et al. (2019)

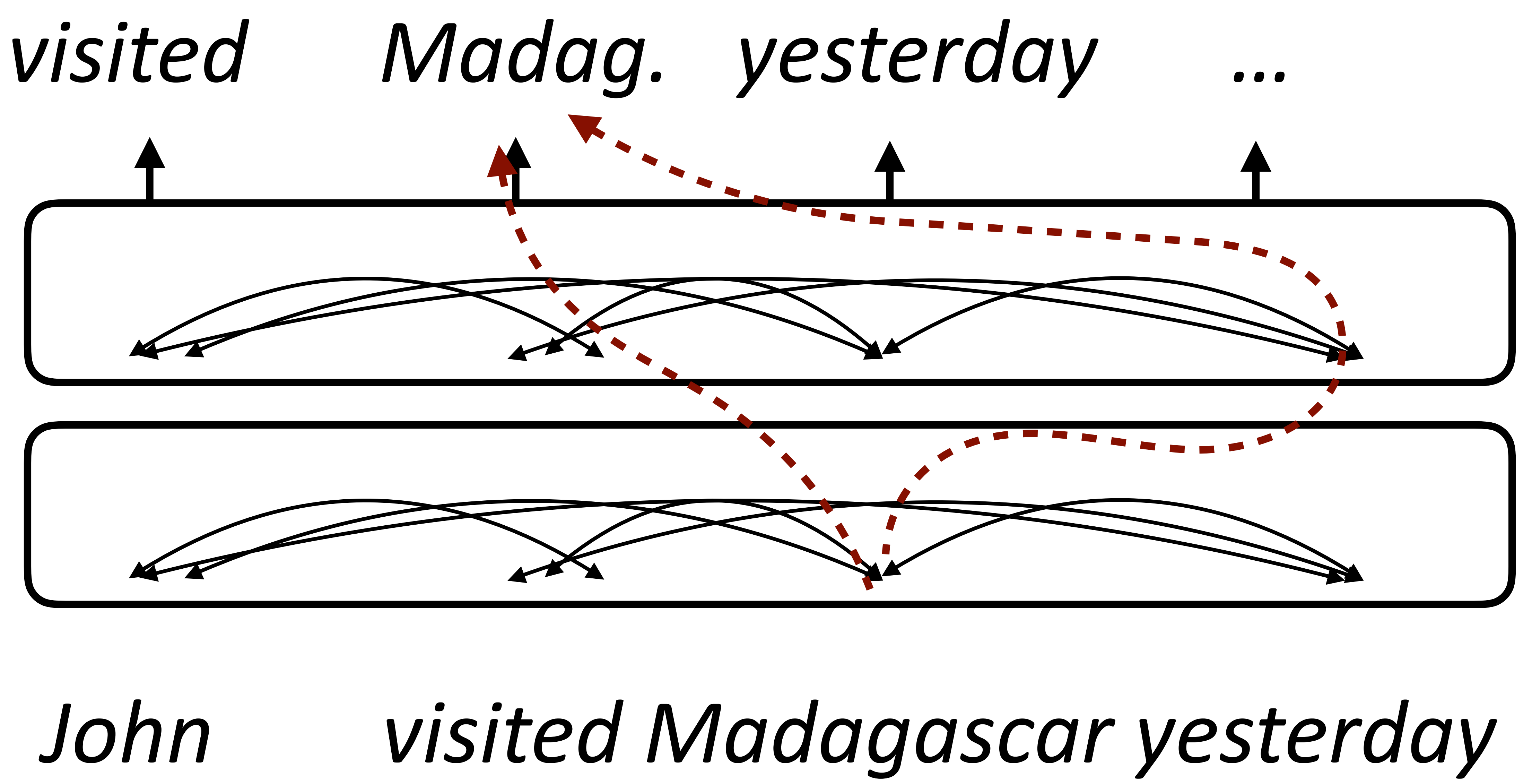
Bidirectional Modeling

- How to learn a "deeply bidirectional" model? What happens if we just replace an LSTM with a transformer?

ELMo (Language Modeling)



BERT



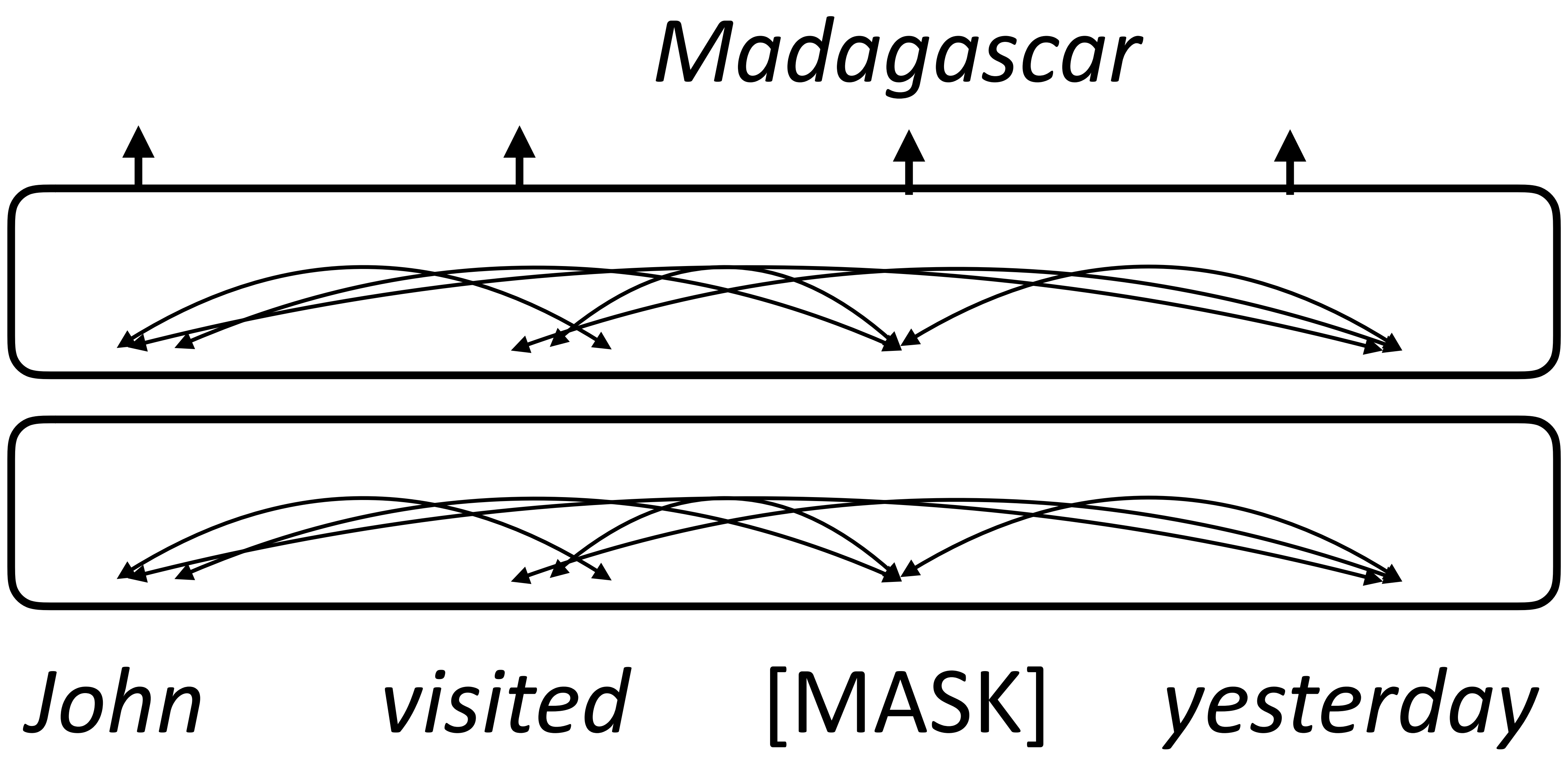
- You could do this with a "one-sided" transformer, but this "two-sided" model can cheat

Devlin et al. (2019)

Masked Language Modeling

- How to prevent cheating? Next word prediction fundamentally doesn't work for bidirectional models, instead do *masked language modeling*

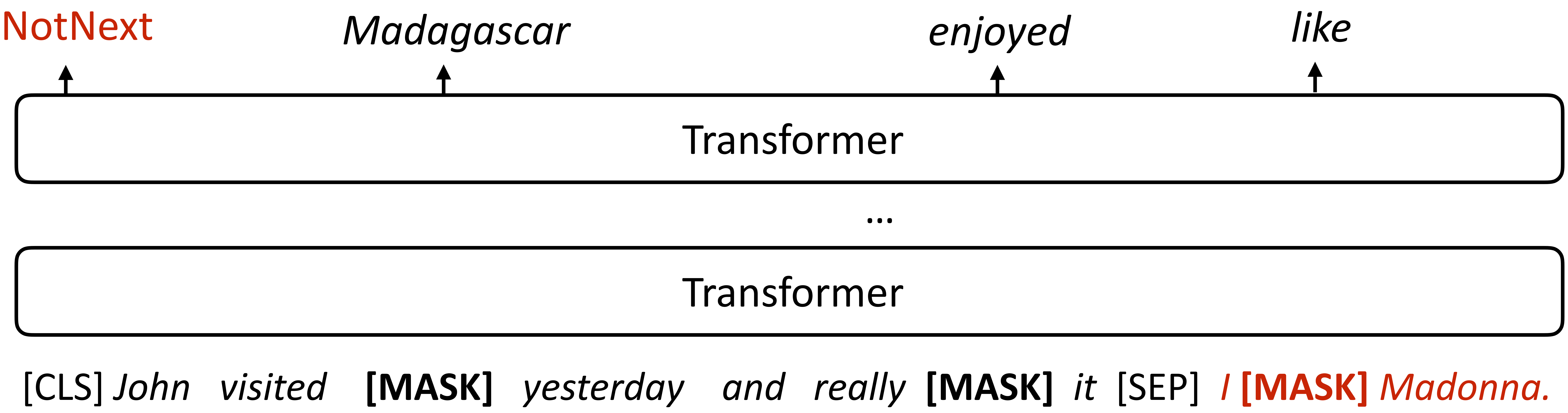
- BERT formula: take a chunk of text, mask out 15% of the tokens, and try to predict them



Devlin et al. (2019)

BERT Objective

- Input: [CLS] Text chunk 1 [SEP] Text chunk 2
- 50% of the time, take the true next chunk of text, 50% of the time take a random other chunk. Predict whether the next chunk is the "true" next
- BERT objective: masked LM + next sentence prediction


NotNext
Madagascar
enjoyed
like
Transformer
…
Transformer
[CLS] *John* *visited* **[MASK]** *yesterday* *and* *really* **[MASK]** *it* [SEP] *I* **[MASK]** *Madonna.*