Contextual Token Representations

ULMfit, OpenAI GPT, ELMo, BERT, XLM

Noe Casas

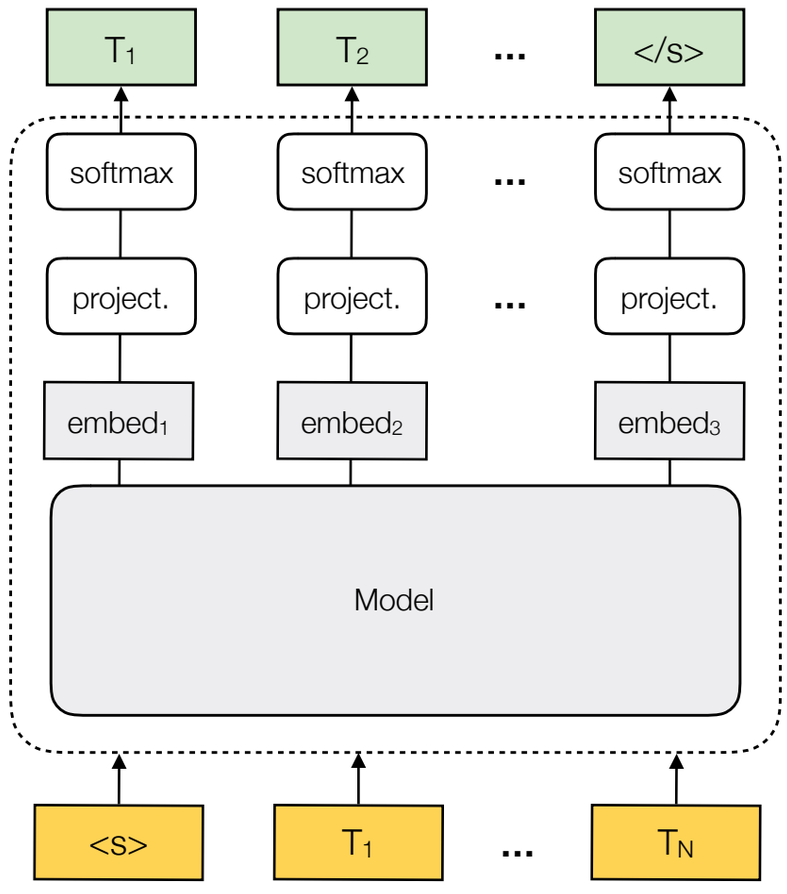
Background: Language Modeling

- Data: Monolingual Corpus

- Task: predict next token given previous tokens (causal):

$$P(T_i \mid T_1 \ldots T_{i-1})$$

- Usual models: LSTM, Transformer.



Contextual embeddings: intuition

- Same word can have different meaning depending on the context. Example:

  - Please, **type** everything in lowercase.
  - What **type** of flowers do you like most?

- Classic word embeddings offer the same vector representation regardless of the context.

- Solution: create word representations that depend on the context.

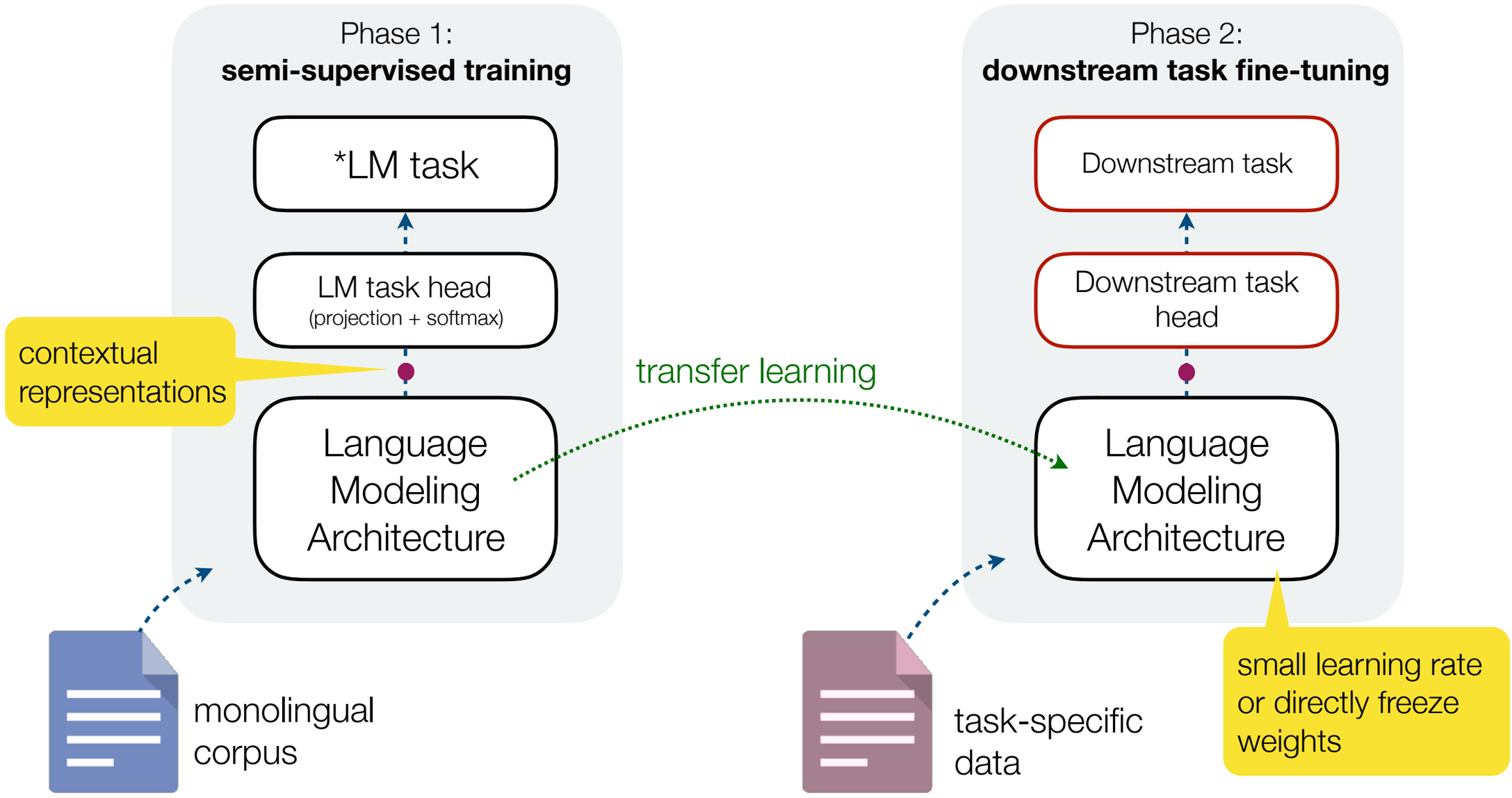
Articles

Model Alias	Org.	Article Reference
ULMfit	fast.ai	*Universal Language Model Fine-tuning for Text Classification* Howard and Ruder
ELMo	AllenNLP	*Deep contextualized word representations* Peters et al.
OpenAI GPT	OpenAI	*Improving Language Understanding by Generative Pre-Training* Radford et al.
BERT	Google	*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* Devlin et al.
XLM	Facebook	*Cross-lingual Language Model Pretraining* Lample and Conneau

Overview

- Train model in one of multiple tasks that lead to word representations.
- Release pre-trained models.
- Use pre-trained models, options:
    A. Fine-tune model on final task.
    B. Directly encode token representations with model.

Overview (graphical)

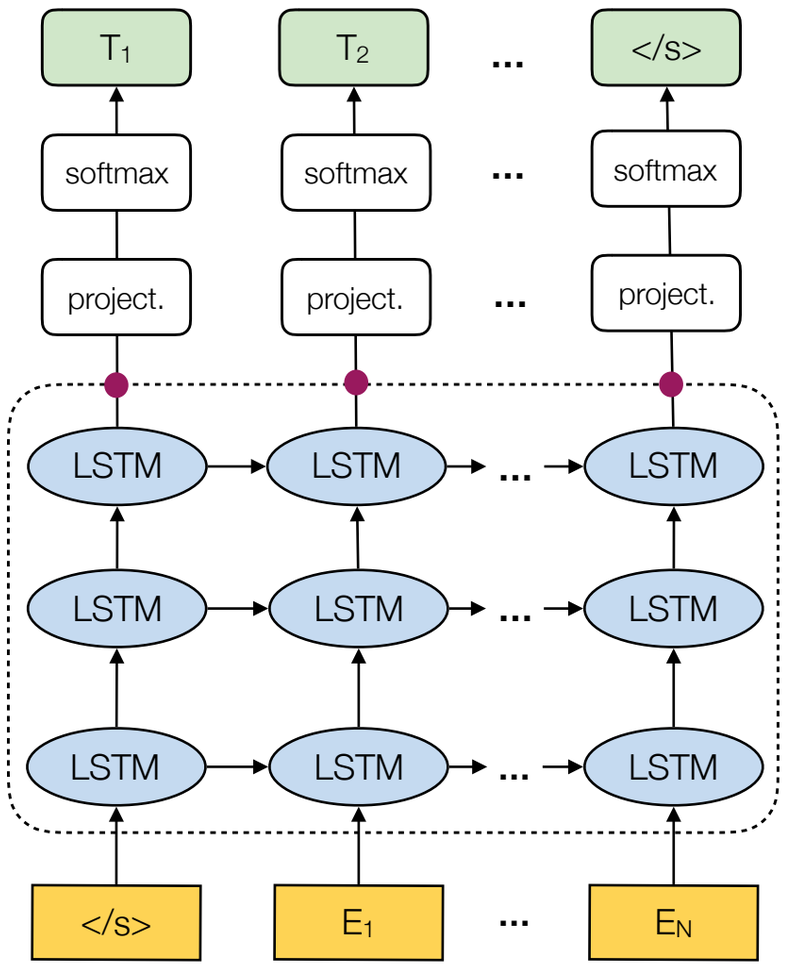


Differences

Alias	Model	Token	Tasks	Language
ULMfit	LSTM	word	Causal LM	English
ELMo	LSTM	word	Bidirectional LM	English
OpenAI GPT	Transformer	subword	Causal LM + Classification	English
BERT	Transformer	subword	Masked LM + Next sentence prediction	Multilingual
XLM	Transformer	subword	Causal LM +Masked LM + Translation LM	Multilingual

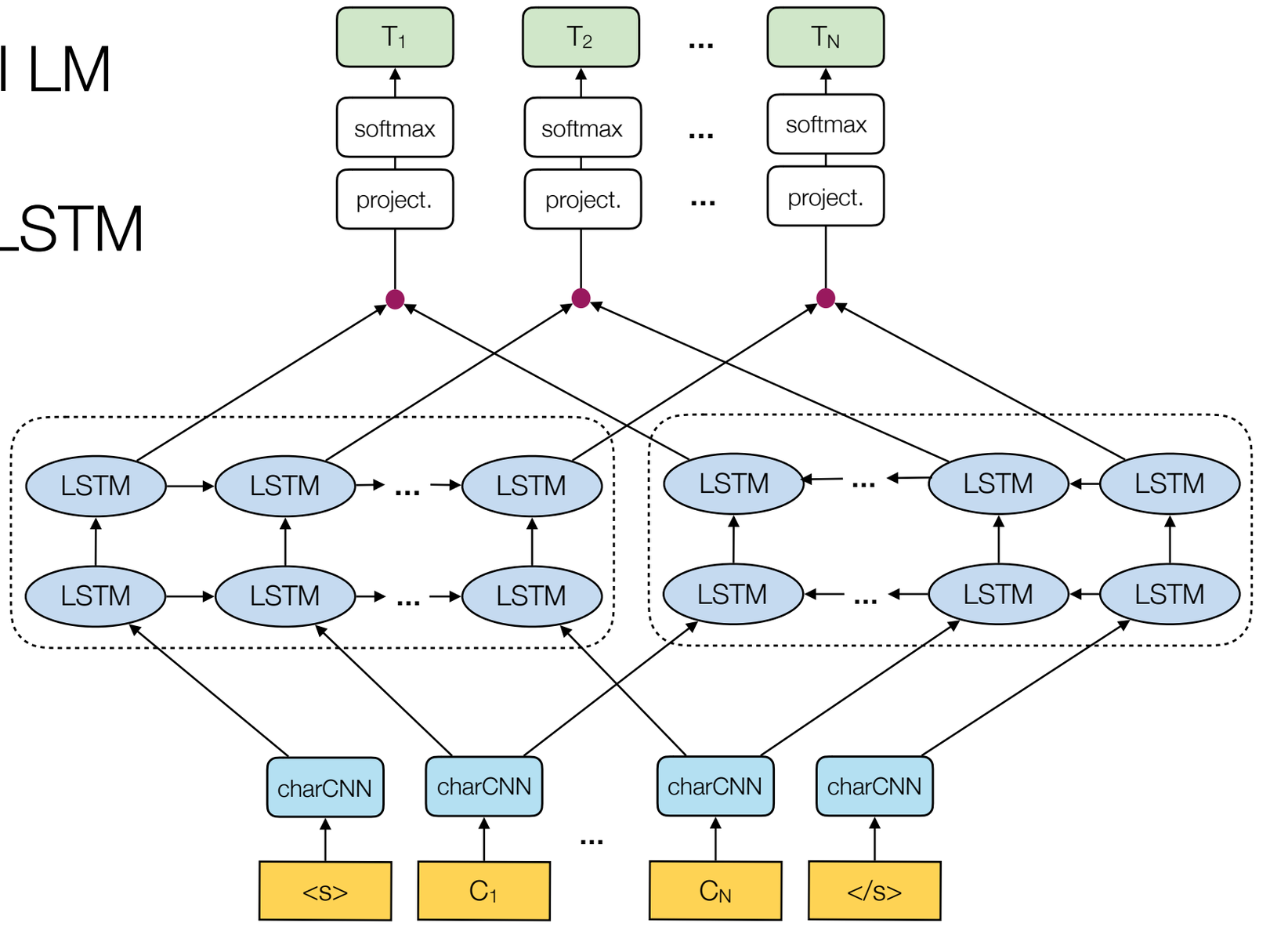
ULMFiT

- **Task**: causal LM
- **Model**: 3-layer LSTM
- **Tokens**: words



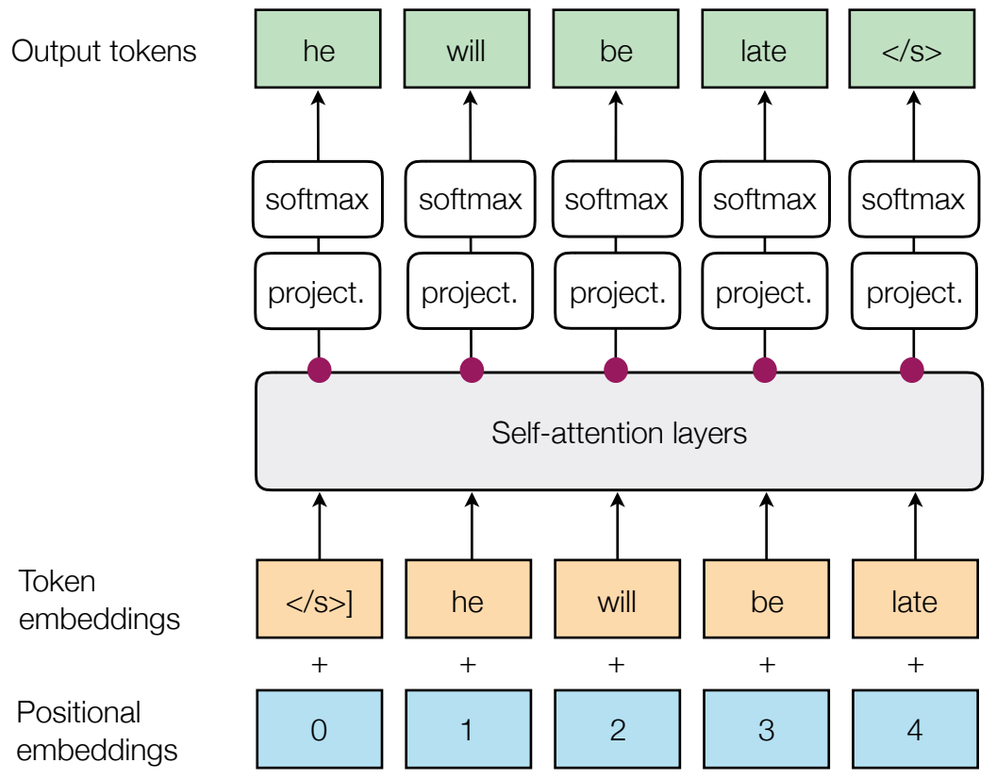
ELMO

- **Task**: bidirectional LM
- **Model**: 2-layer biLSTM
- **Tokens**: words

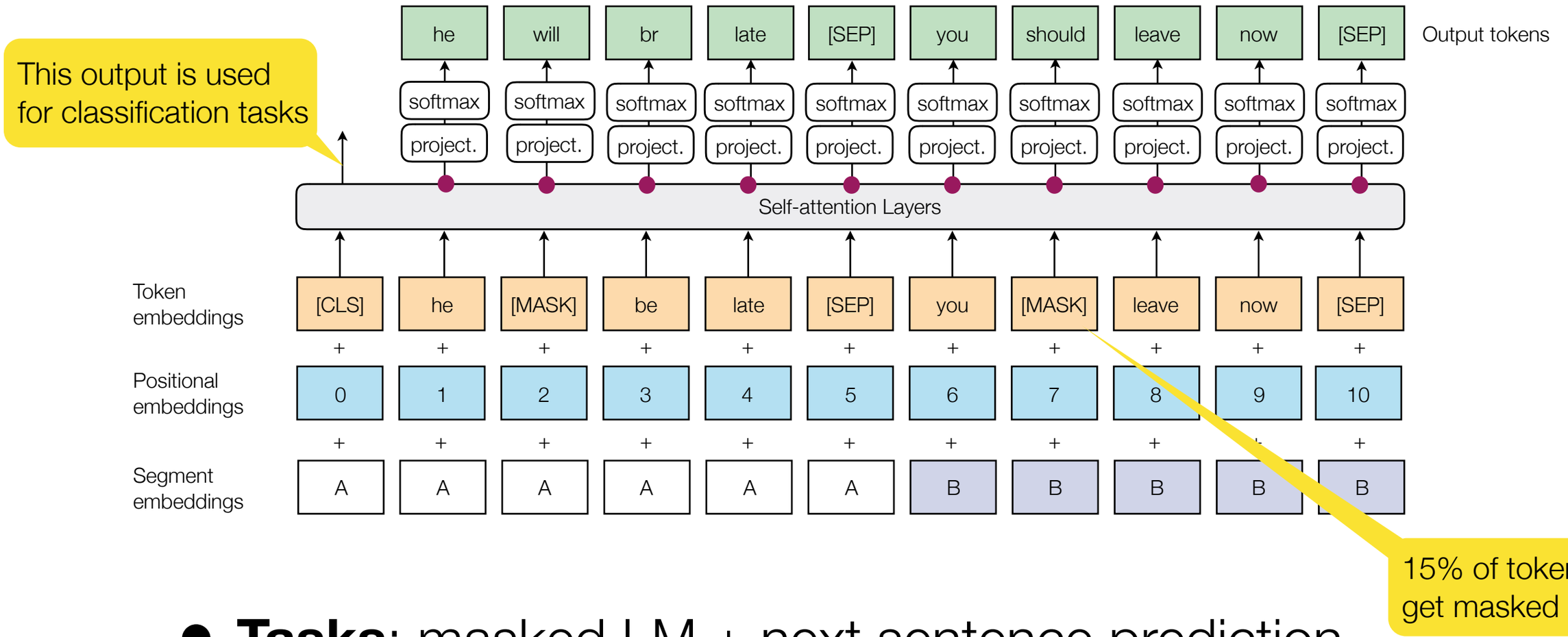


OpenAI GPT

- **Task**: causal LM
- **Model**: self-attention layers
- **Tokens**: subwords

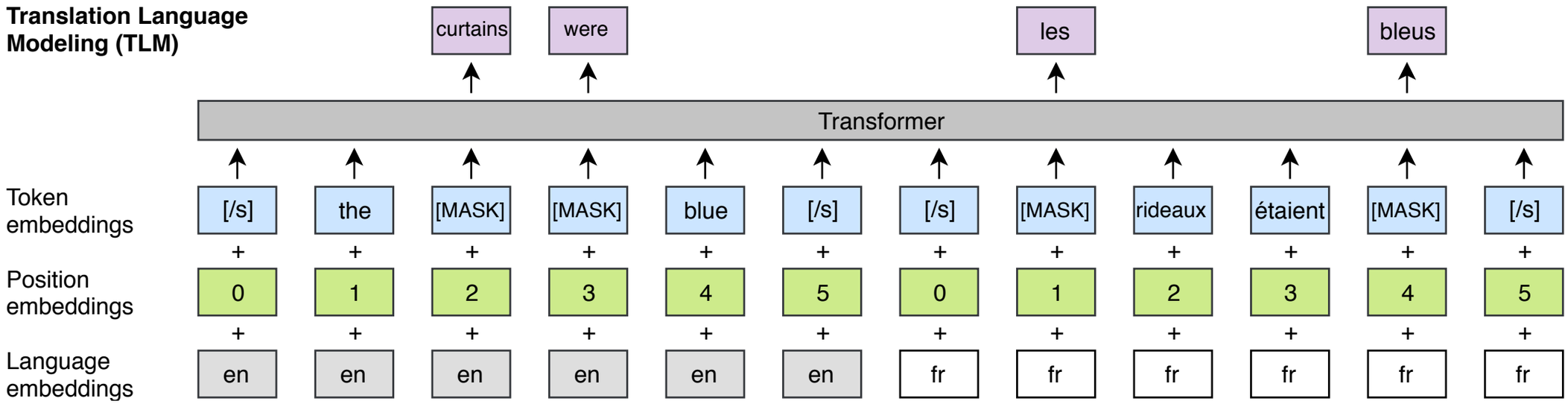


BERT



- **Tasks**: masked LM + next sentence prediction
- **Model**: self-attention layers
- **Tokens**: subwords

XLM



- **Tasks**: LM + masked LM + Translation LM
- **Model**: self-attention layers
- **Tokens**: subwords

Masked LM with parallel sentences

Projection and softmax are omitted

*figure from “Cross-lingual Language Model Pretraining”

Downstream Tasks

- Natural Language Inference (NLI) or Cross-lingual NLI.
- Text classification (e.g. sentiment analysis).
- Next sentence prediction.
- Supervised and Unsupervised Neural Machine Translation (NMT).
- Question Answering (QA).
- Named Entity Recognition (NER).

Further reading

- “Looking for ELMo's friends: Sentence-Level Pretraining Beyond Language Modeling”, Bowman et al., 2018
- “What do you learn from context? Probing for sentence structure in contextualized word representations”, Tenney et al., 2018.
- “Assessing BERT’s Syntactic Abilities”, Goldberg, 2018
- “Learning and Evaluating General Linguistic Intelligence”, Yogatama et al., 2019.

Differences with other representations

Note the differences of contextual token representations with:

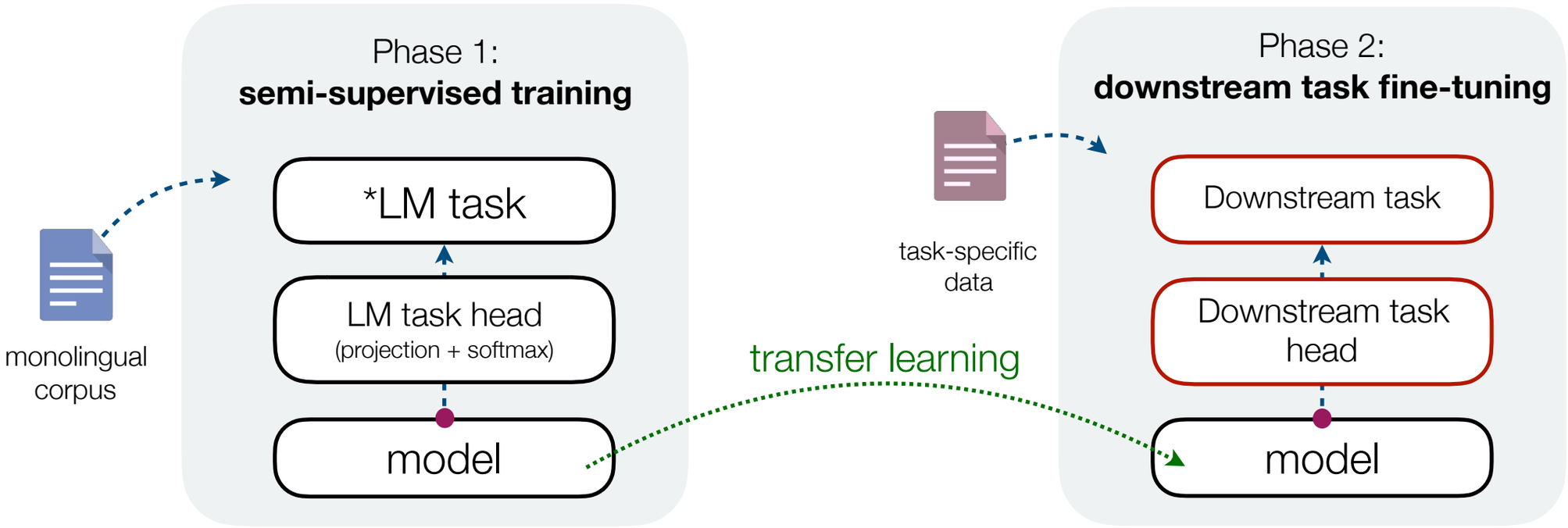
- Non-word representations like in (CoVe): *Learned in Translation: Contextualized Word Vectors* by McCann et al. 2017 [salesforce].

- Fixed-size sentence representations like in *Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond* by Artetxe and Schewnk, 2018 [facebook].

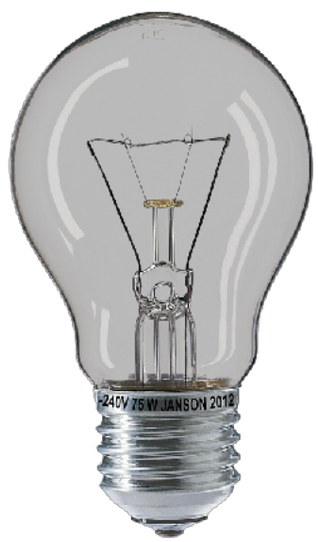
Other resources

- https://nlp.stanford.edu/seminar/details/jdevlin.pdf
- http://jalammar.github.io/illustrated-bert/
- https://medium.com/dissecting-bert/dissecting-bert-part2-335ff2ed9c73
- https://github.com/huggingface/pytorch-pretrained-BERT

Summary



Alias	Model	Token	Tasks	Language
ULMfit	LSTM	word	Causal LM	English
ELMo	LSTM	word	Bidirectional LM	English
OpenAI GPT	Transformer	subword	Causal LM + Classification	English
BERT	Transformer	subword	Masked LM + Next sentence prediction	Multilingual
XLM	Transformer	subword	Causal LM +Masked LM + Translation LM	Multilingual



Bonus slides

Are these really token representations?

- They are a linear projection away from token space.
- Word-level nearest neighbours in corpus finds same word with same usage.