Transfer Learning in Biomedical NLP: A Case Study with BERT

Yifan Peng

NCBI/NLM/NIH

U.S. National Library of Medicine

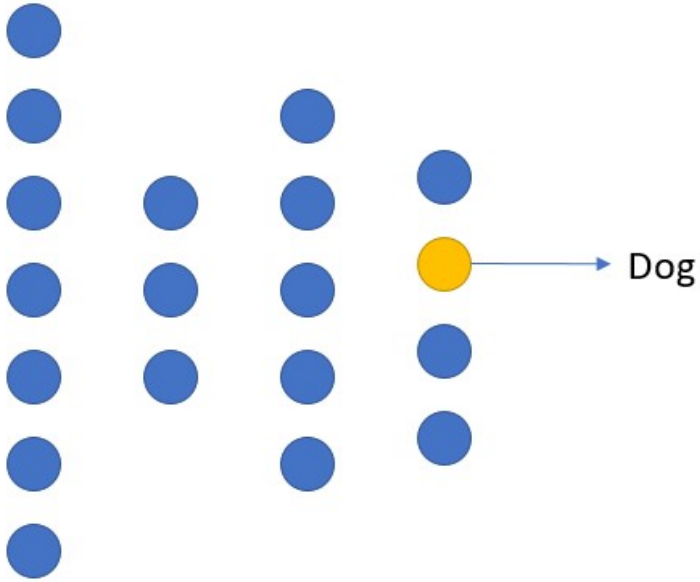
Transfer learning

- A technique that allows to reutilize an already trained model on a specific dataset and adapt it to a different dataset

- In the field of computer vision, researchers have repeatedly shown the value of transfer learning
  - pre-training
  - fine-tuning

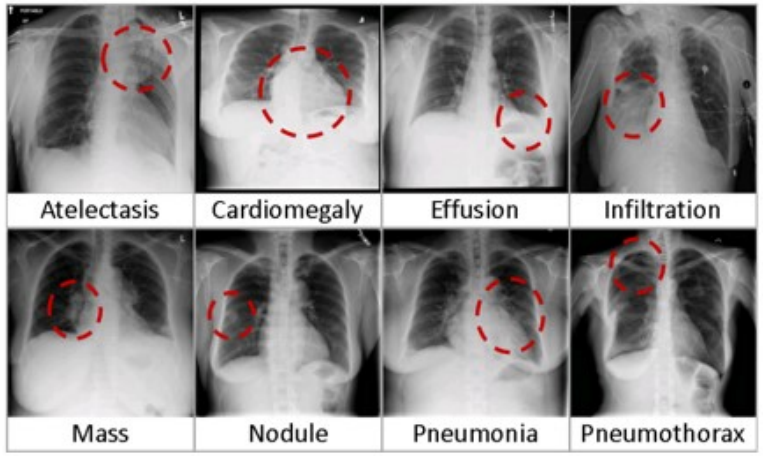
Transfer learning (1/2 steps)

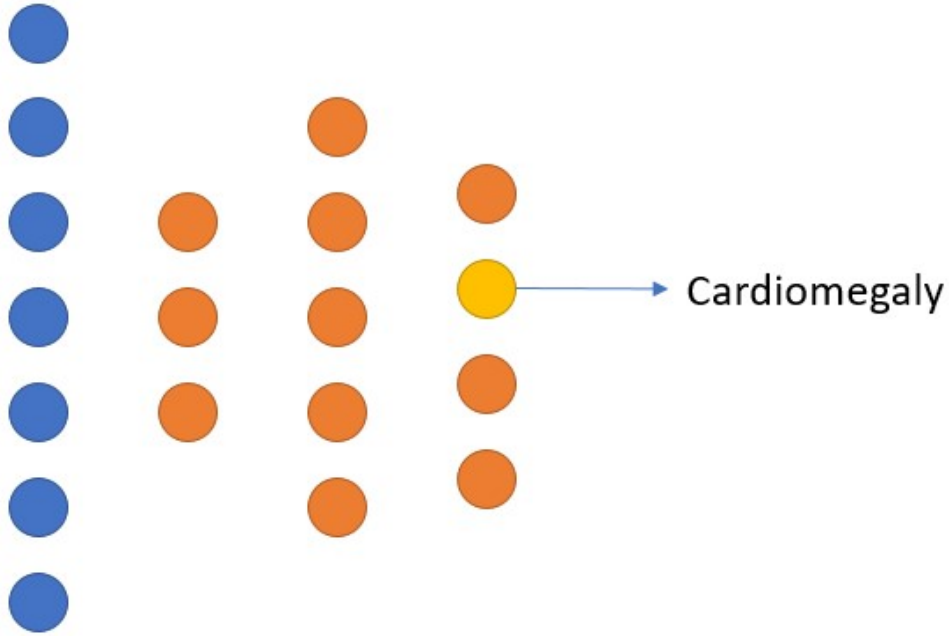
- Pre-training: Use a large training set to learns network parameters and saves them for later use (e.g., ImageNet).



Transfer learning (2/2 steps)

- Fine-tuning: use the pretrained network on the base dataset and train all (or part of) layers in the target dataset





Wang et al., ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. CVPR. 2017.

Transfer learning

- Less difficult to train a complex network
- Speed up the convergence speed of the training

- Pre-training process of natural language

Outlines

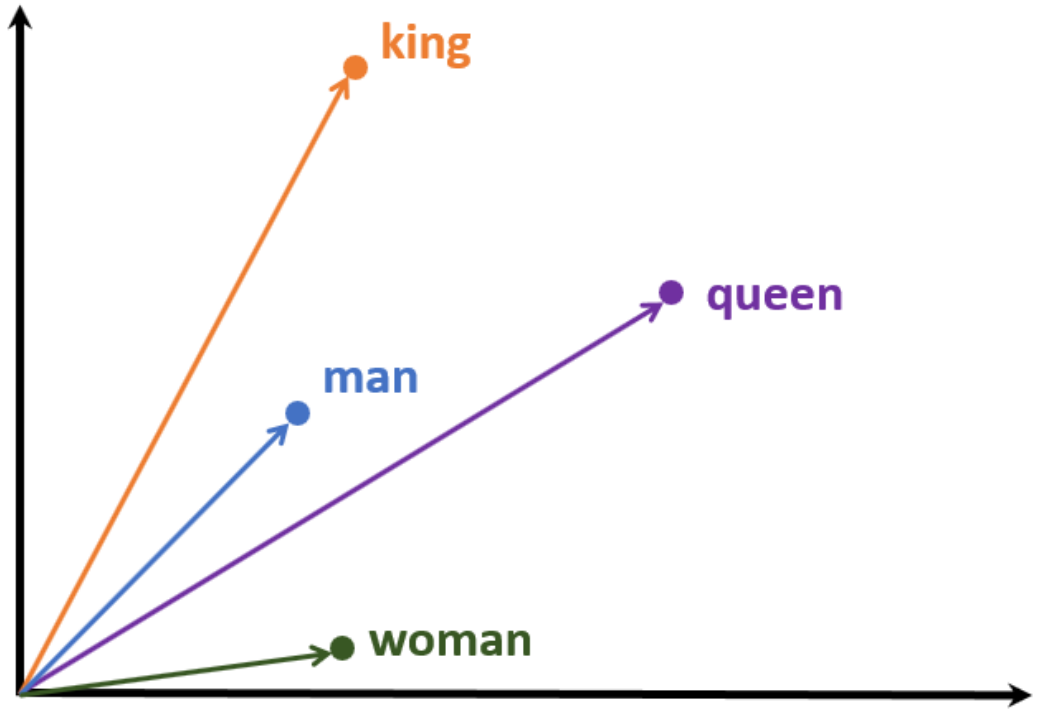
- Word embedding
- ELMo
- BERT

- How BERT's idea are gradually formed?
- What has been innovated?
- Why the effect is good?

- Pre-trained BERT models
- How to use pre-trained BERT
- Performance comparison
- BLUE Benchmark

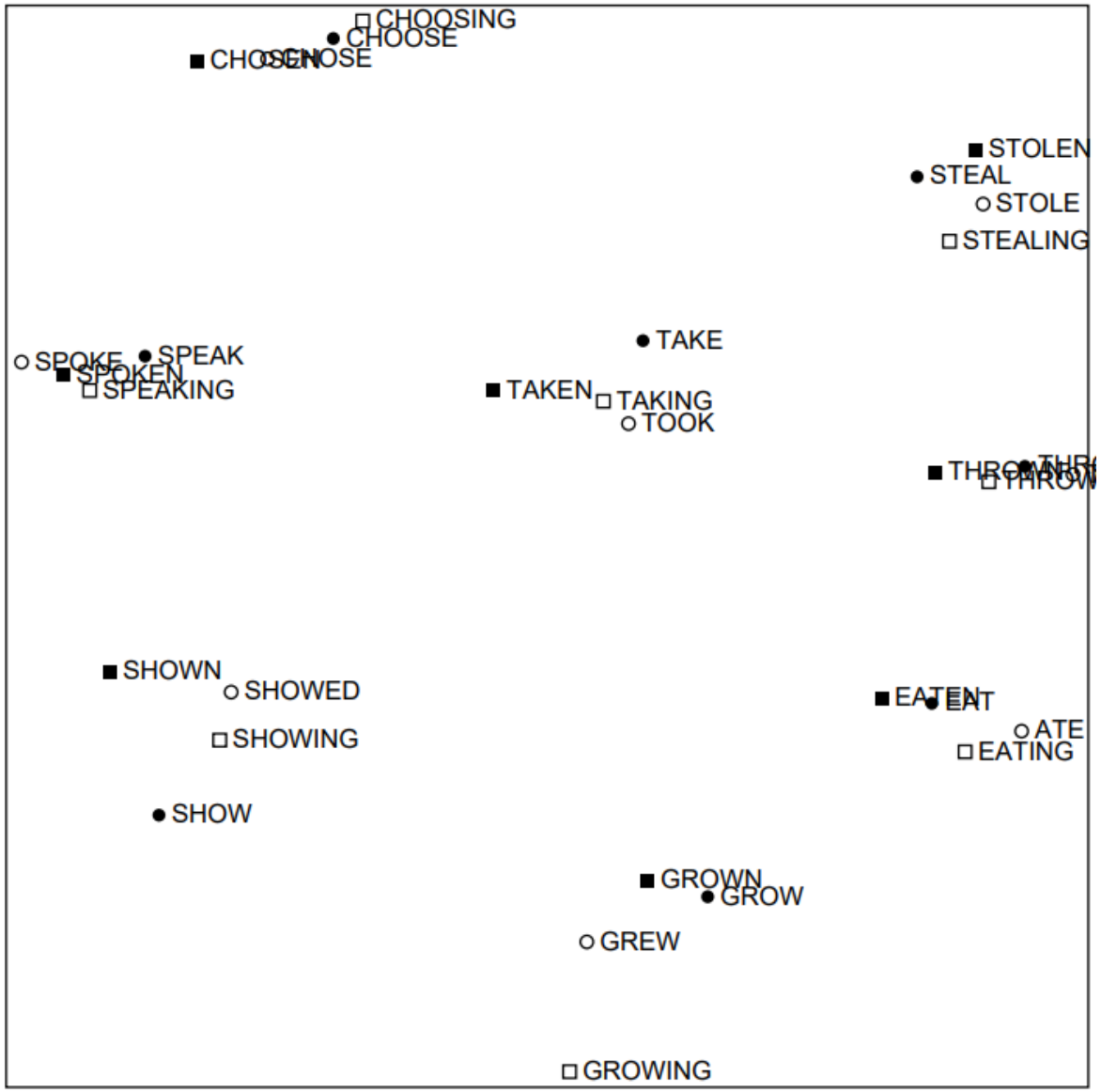
Word Embeddings

- Idea: store “most” of the important information in a fixed, small number of dimensions. For example, a 200-dimentional vector



Word Embeddings

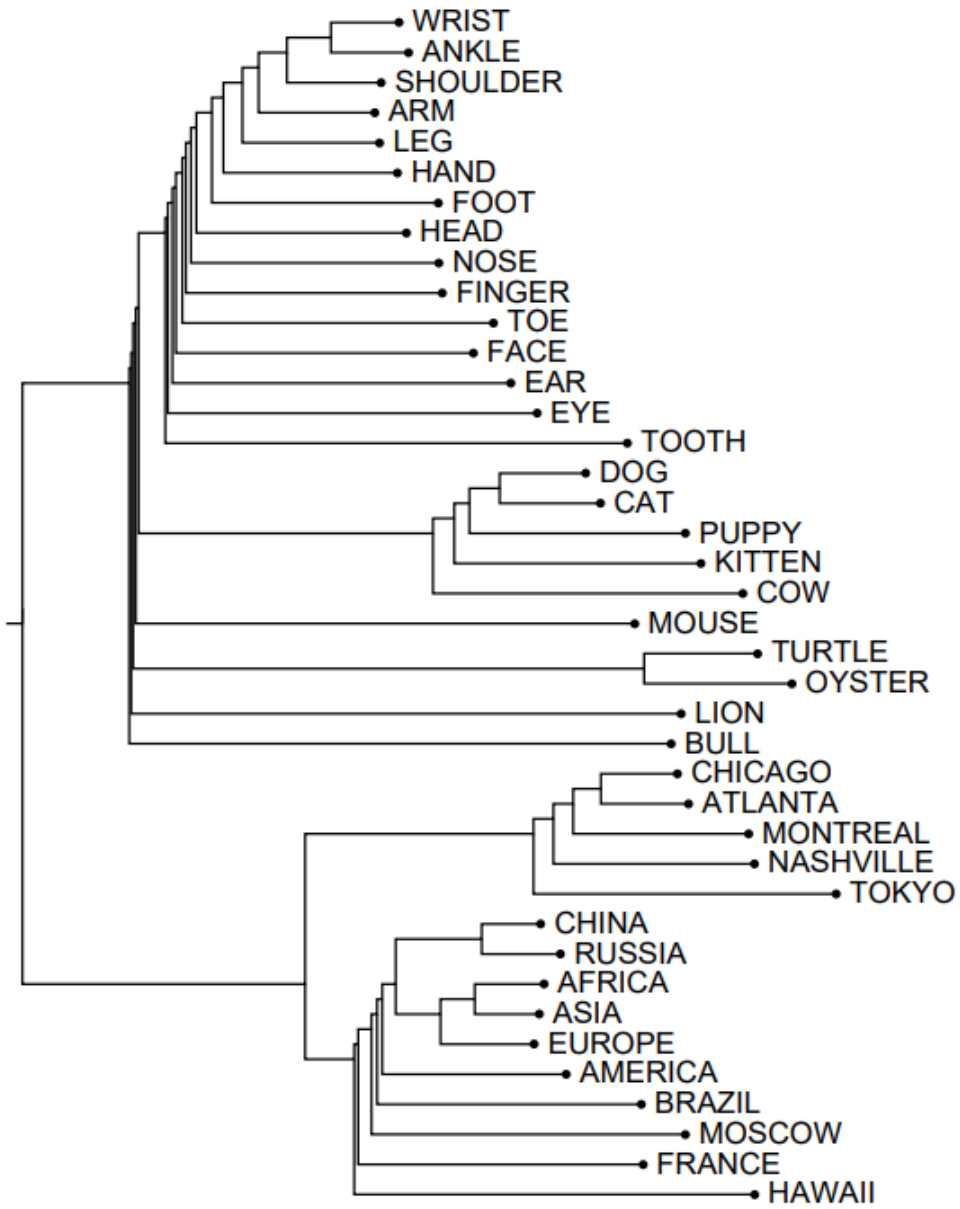
- Multidimensional scaling of present, past, progressive, and past participle forms for eight verb families.



Rohde et al. An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. 2005

Word Embedding

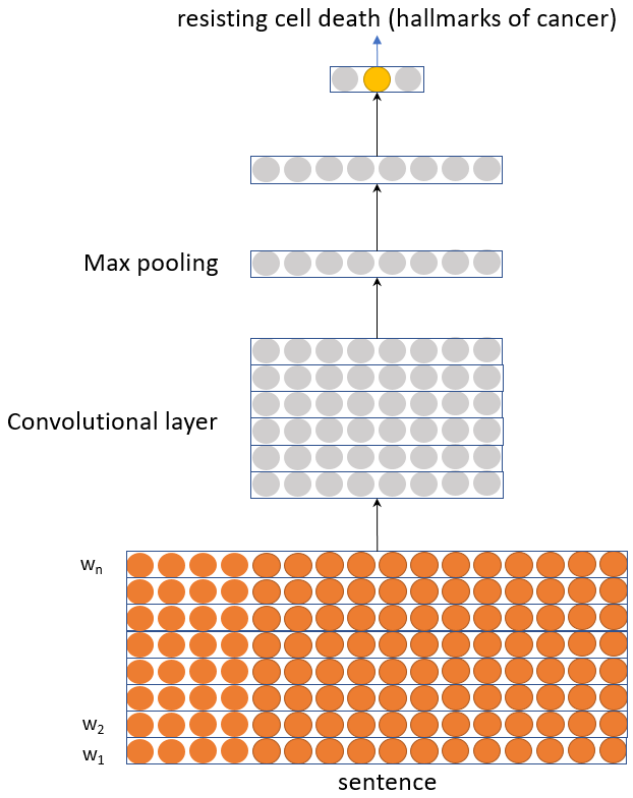
- Hierarchical clustering for three noun classes using distances based on vector correlations.



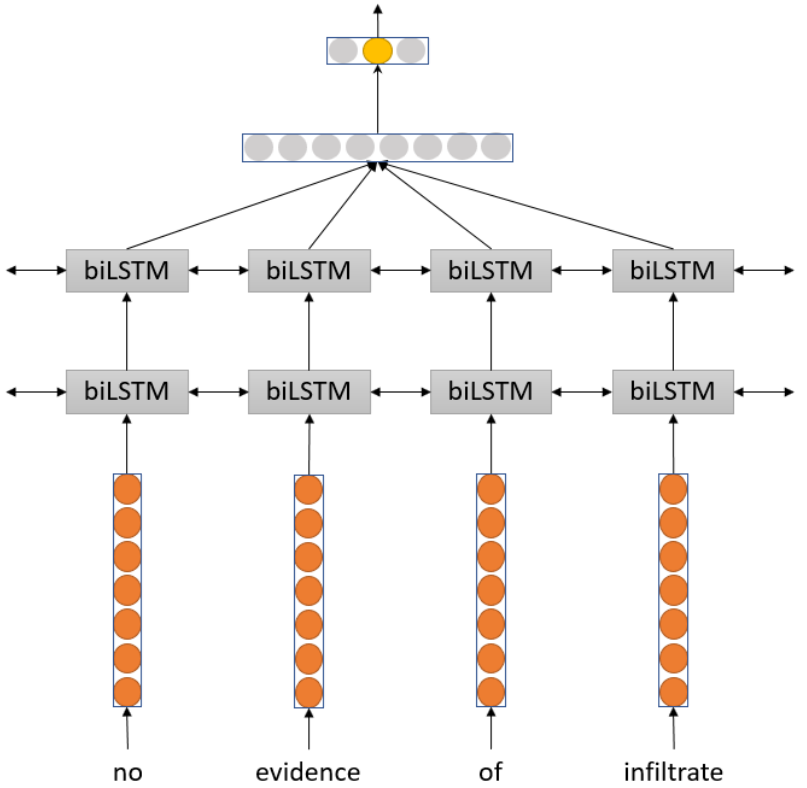
Rohde et al. An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. 2005

How to use Word Embeddings

Convolutional Neural Network



Recurrent Neural Network



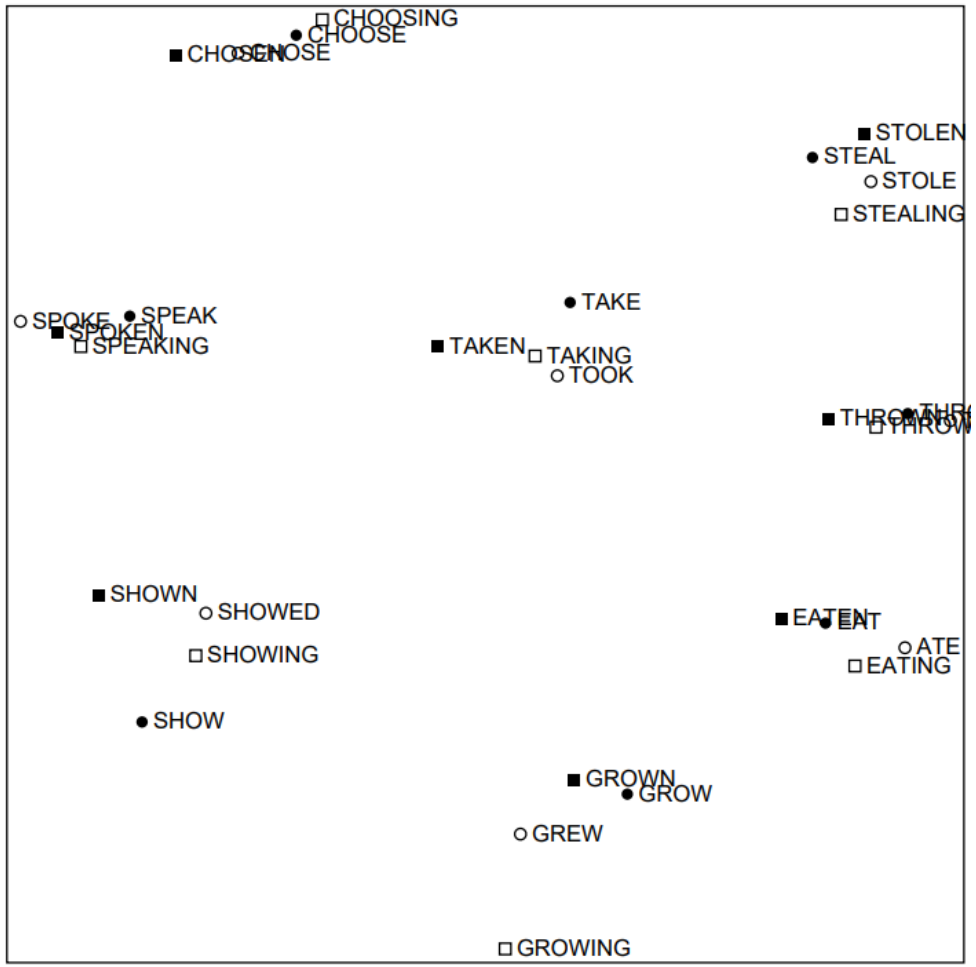
Peng et al. Extracting chemical-protein relations with ensembles of SVM and deep learning models. *Database*. 2018.

Problems of Word Embeddings

Polysemous problem

- About 100 above the **bank** of river…
- The **bank** has plan to branch through the country…

Static Word Embedding can't solve the problem of polysemous words.



From Word Embedding to ELMo

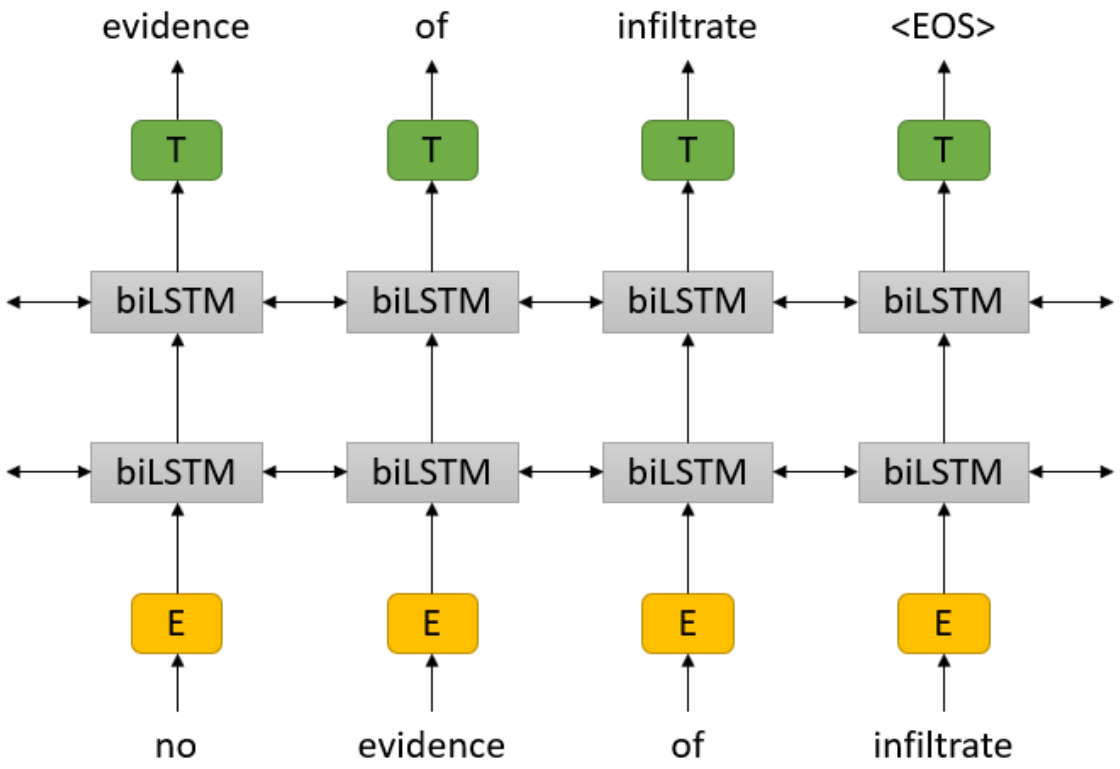
- “Embedding from Language Models”
- “Deep contextualized word representation”

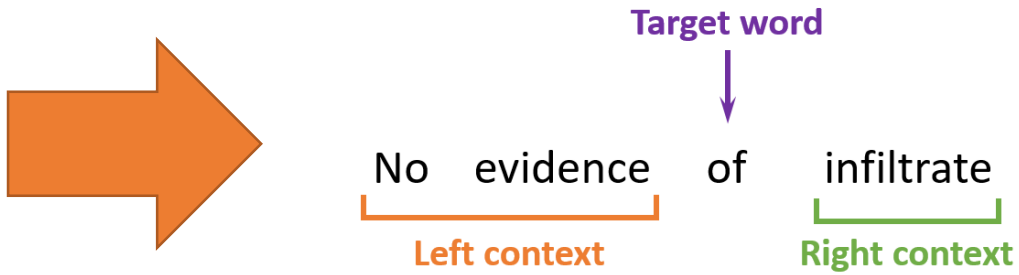
Adjust the Word Embedding representation of the word according to the semantics of the **context** word

Peters et al., Deep contextualized word representations. NAACL. 2018

ELMo

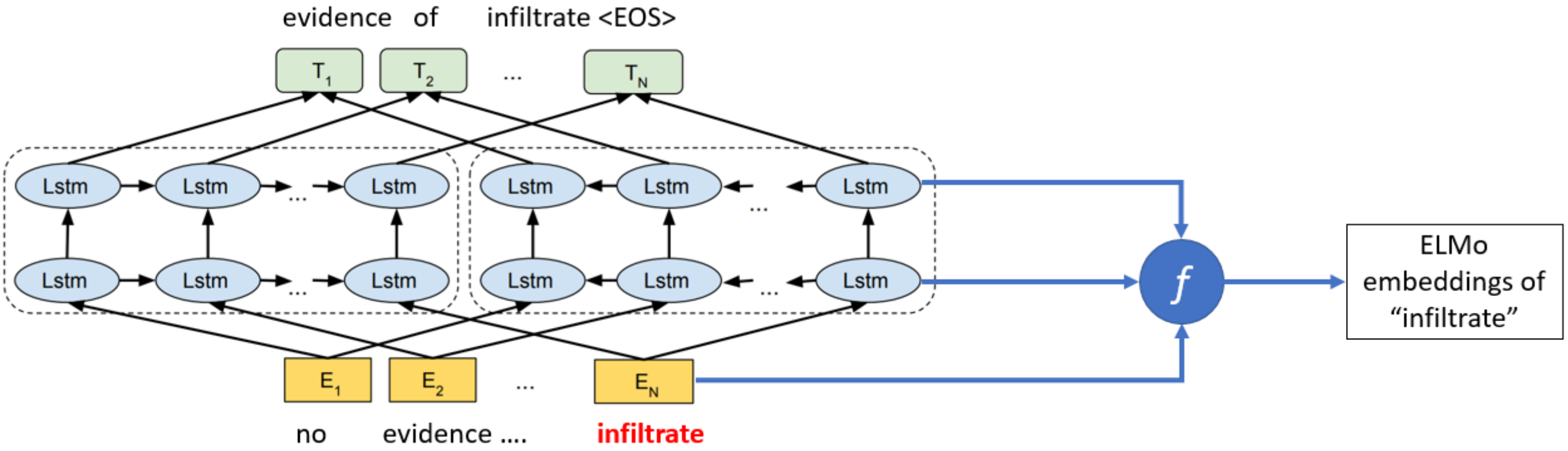
- A typical two-stage process
  - The first stage is to use the language model for pre-training.





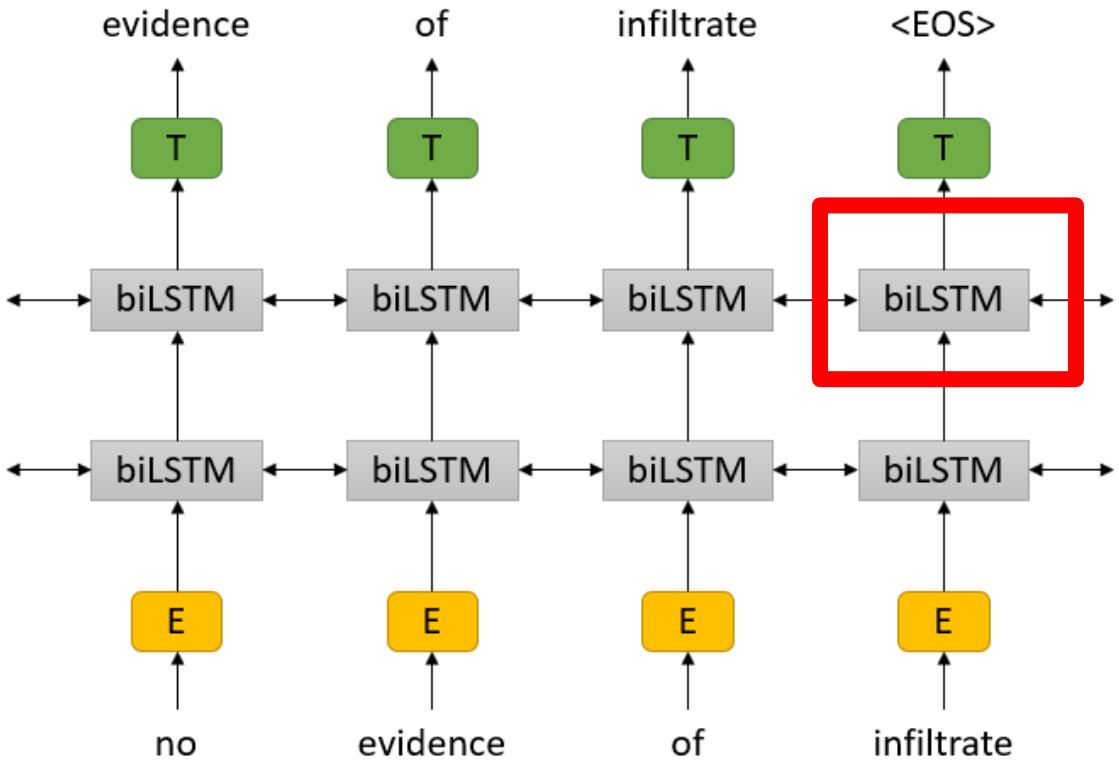
ELMo

- A typical two-stage process
  - The first stage is to use the language model for pre-training.
  - The second stage is to extract the Embeddings of each layer.

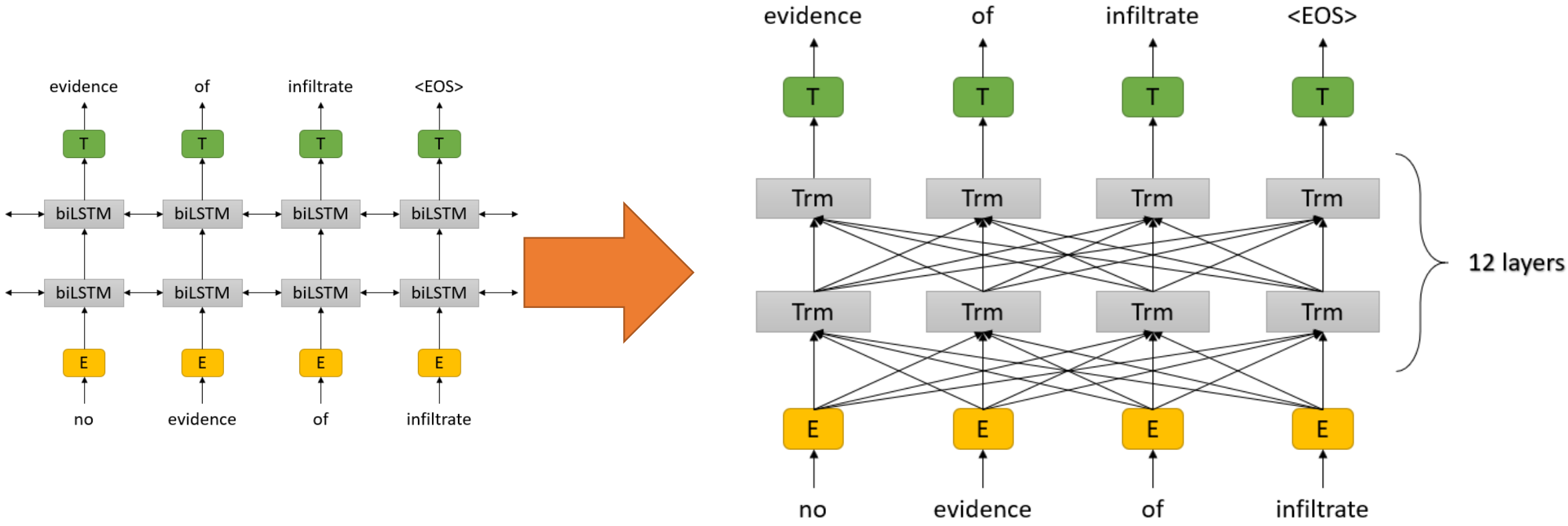


Problems of ELMo

- Hard to capture long distance information
- Computational expensive



From ELMo to BERT

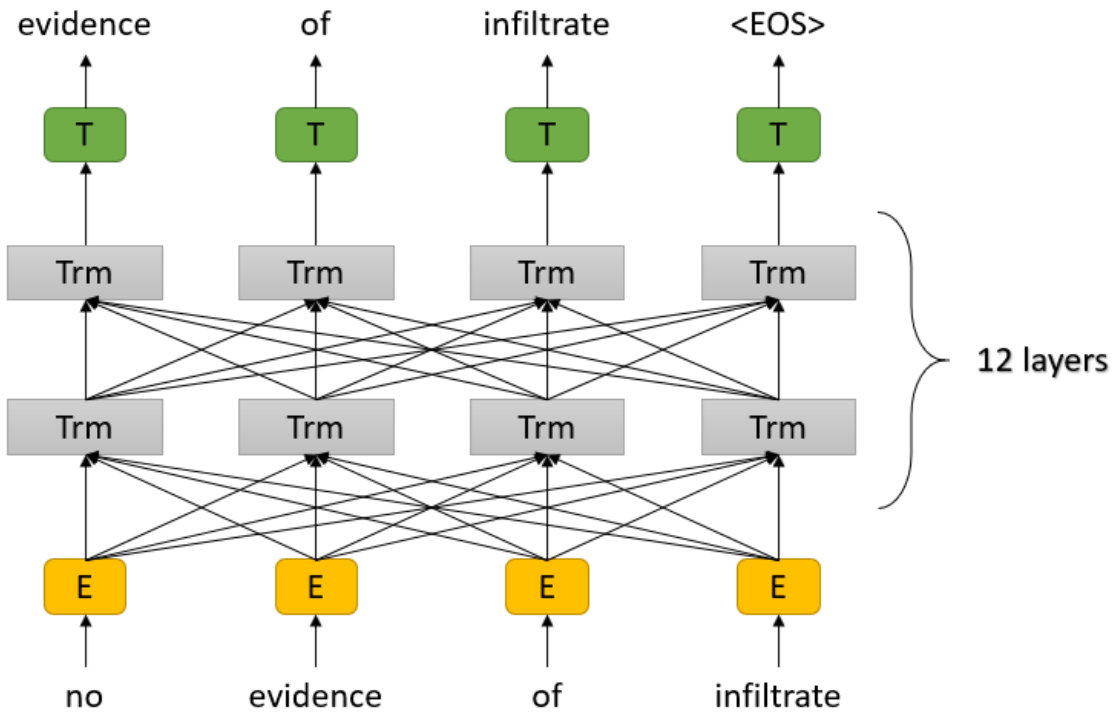


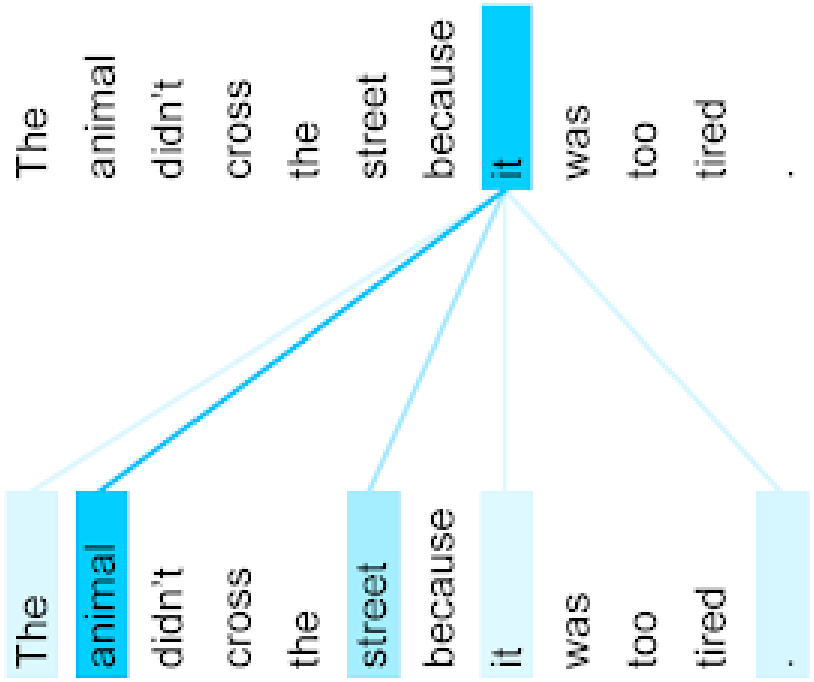
Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL. 2019

Transformer

Why transformer?

A self-attention mechanism which directly models relationships between all words in a sentence, regardless of their respective position.

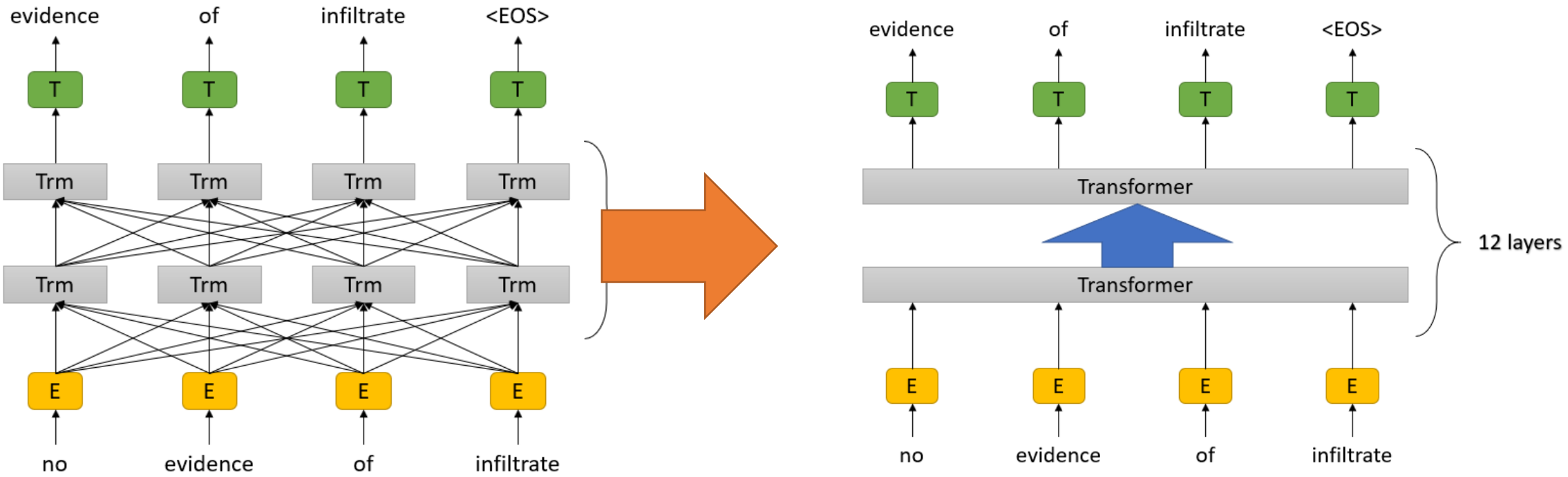




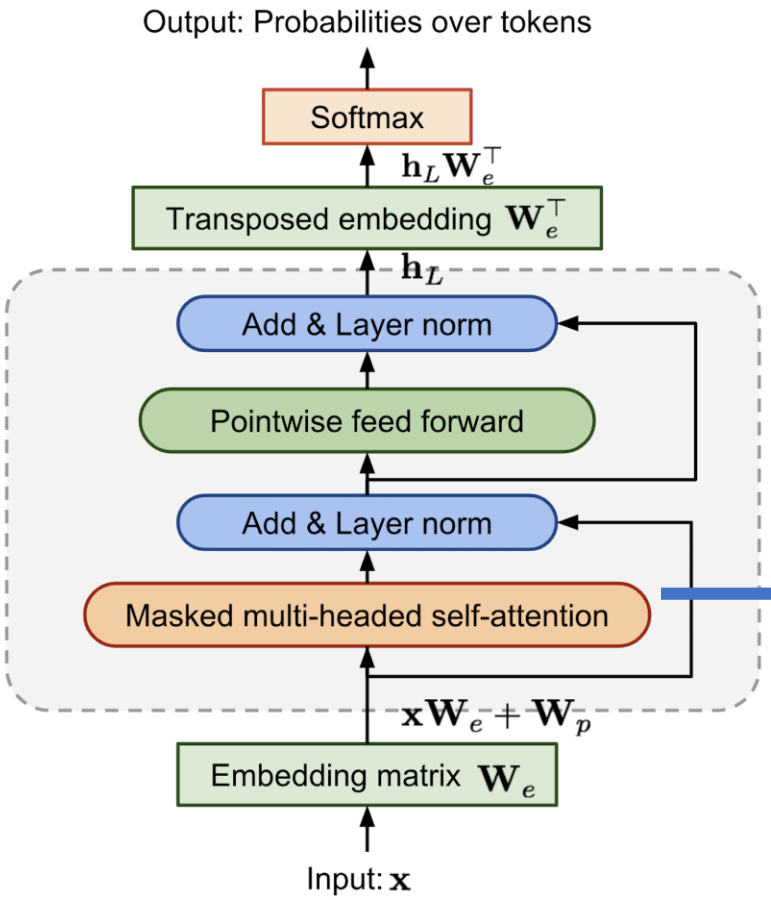
https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

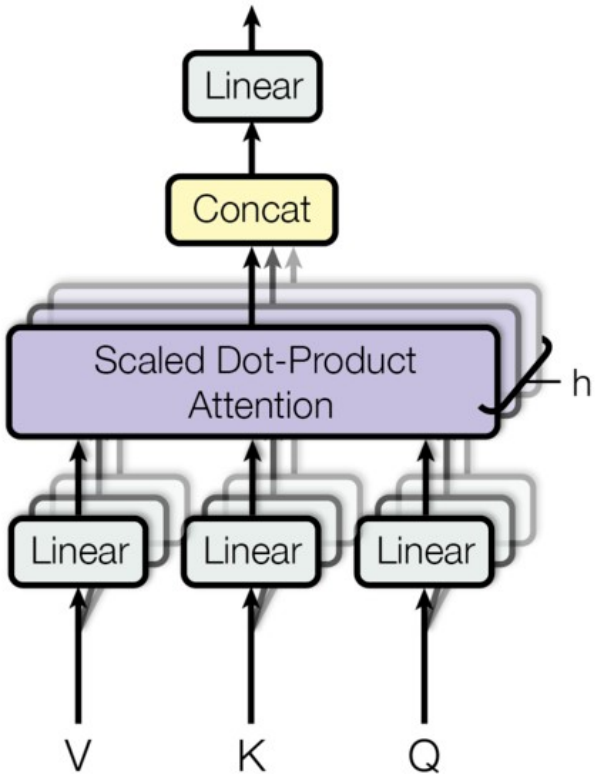
Why transformer?

In parallel. Much faster and more space-efficient

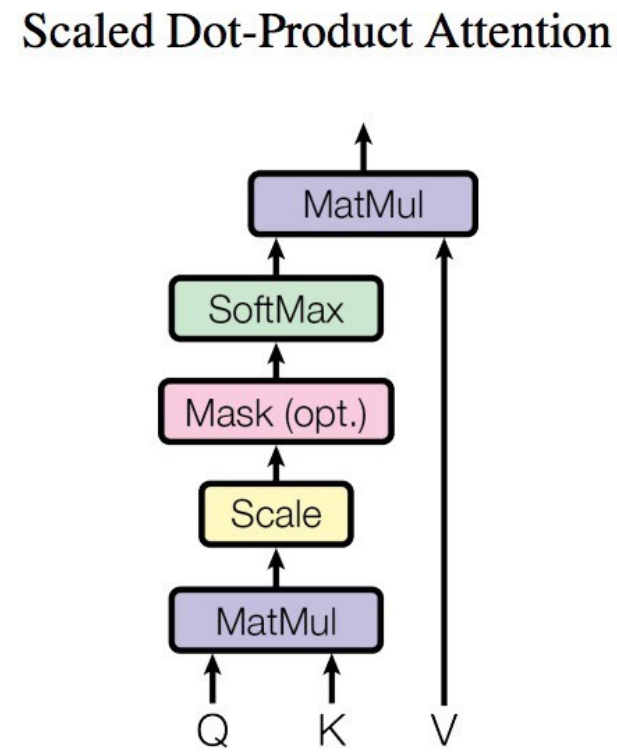


Transformer





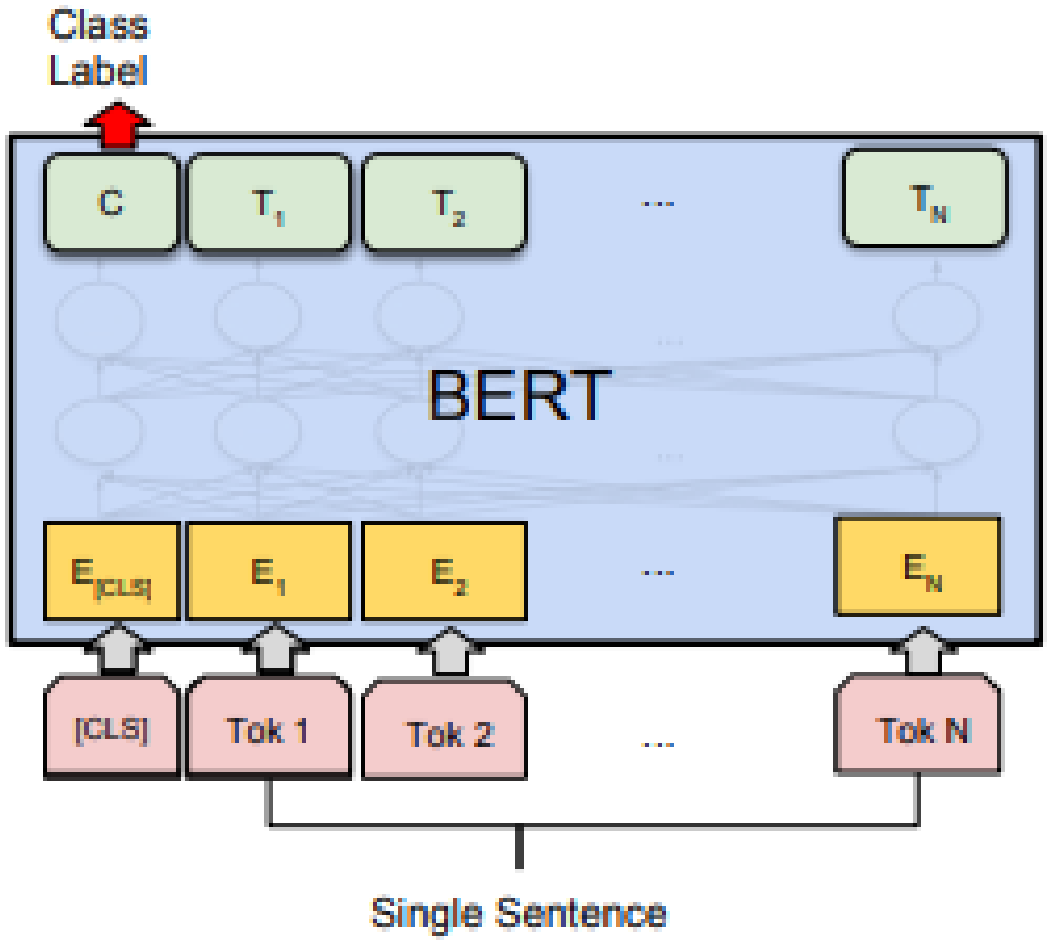
$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,...,\text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

How to use BERT?

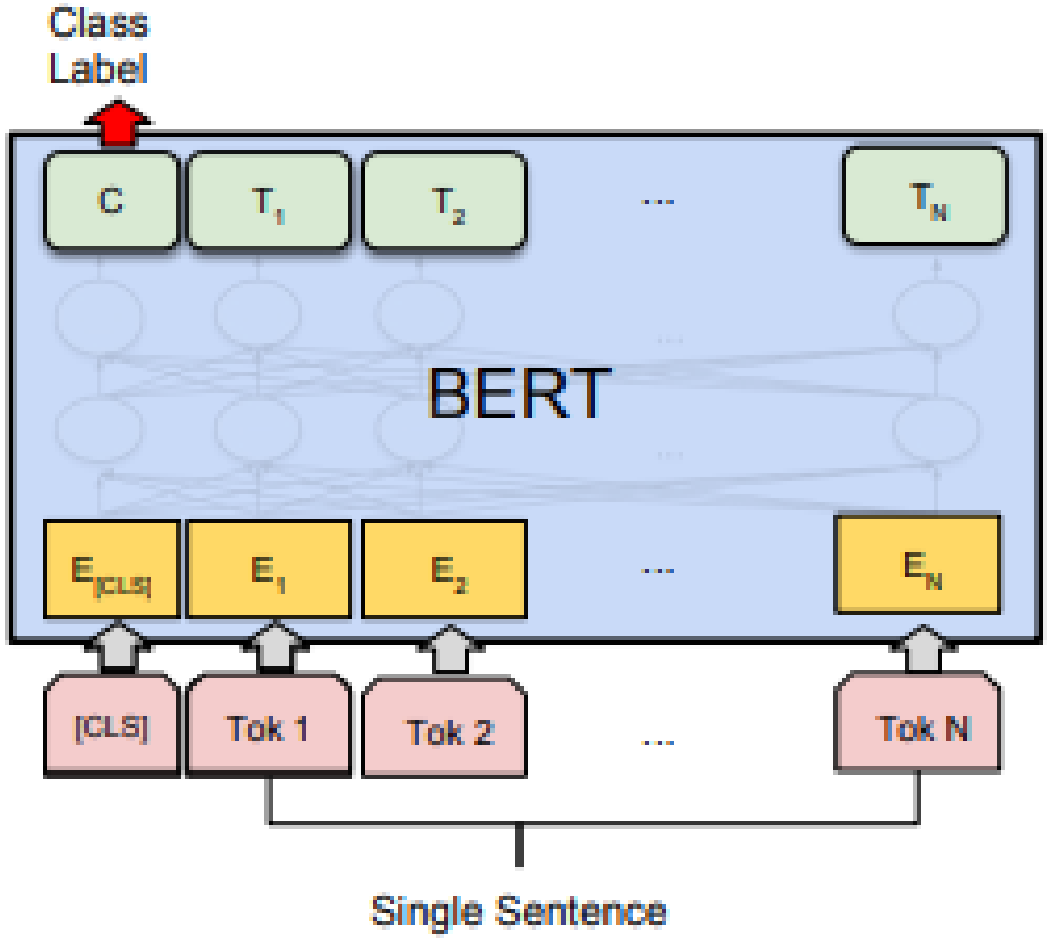
- Sentence classification
- Sentence similarity
- Named entity recognition
- Relation extraction (???)



How to use BERT - Sentence classification

Assign tags or categories to text according to its content.

- Organizing millions of cancer-related references from PubMed into the Hallmarks of Cancer



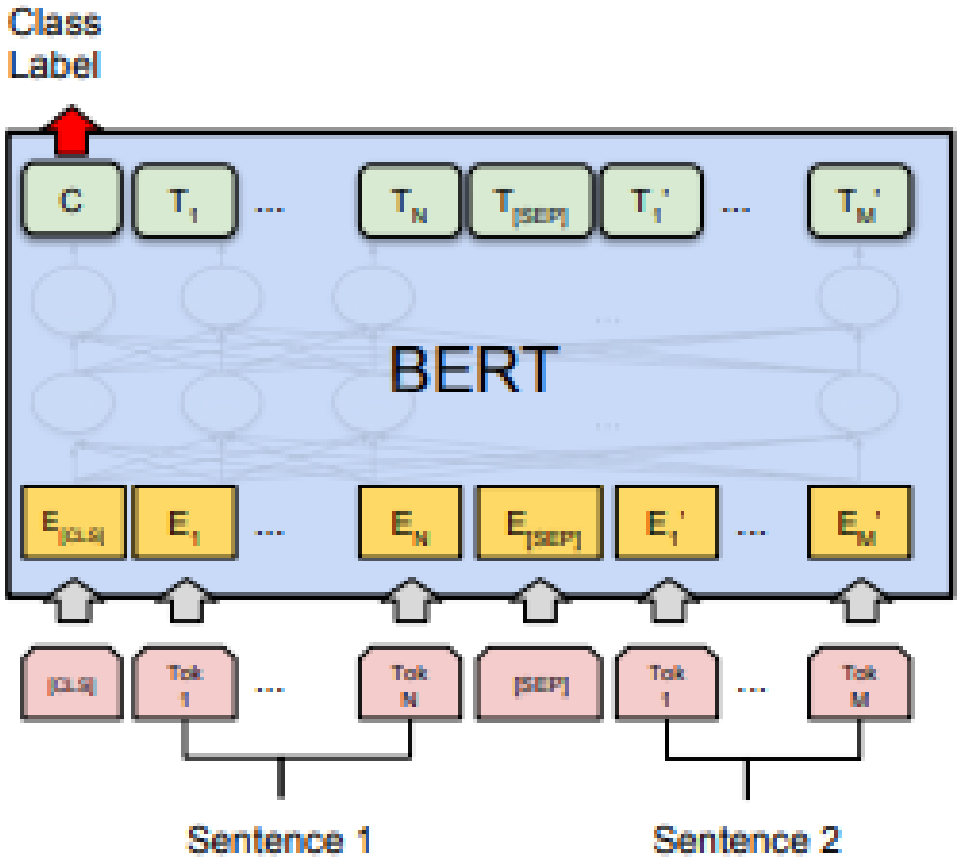
How to use BERT - Sentence similarity

Predict similarity scores based on the sentence pairs.

- The above was discussed with the patient, and she voiced understanding of the content and plan.

VS

- The patient verbalized understanding of the information and was satisfied with the plan of care.



Romanov, et al. Lessons from Natural Language Inference in the Clinical Domain. EMNLP. 2018.

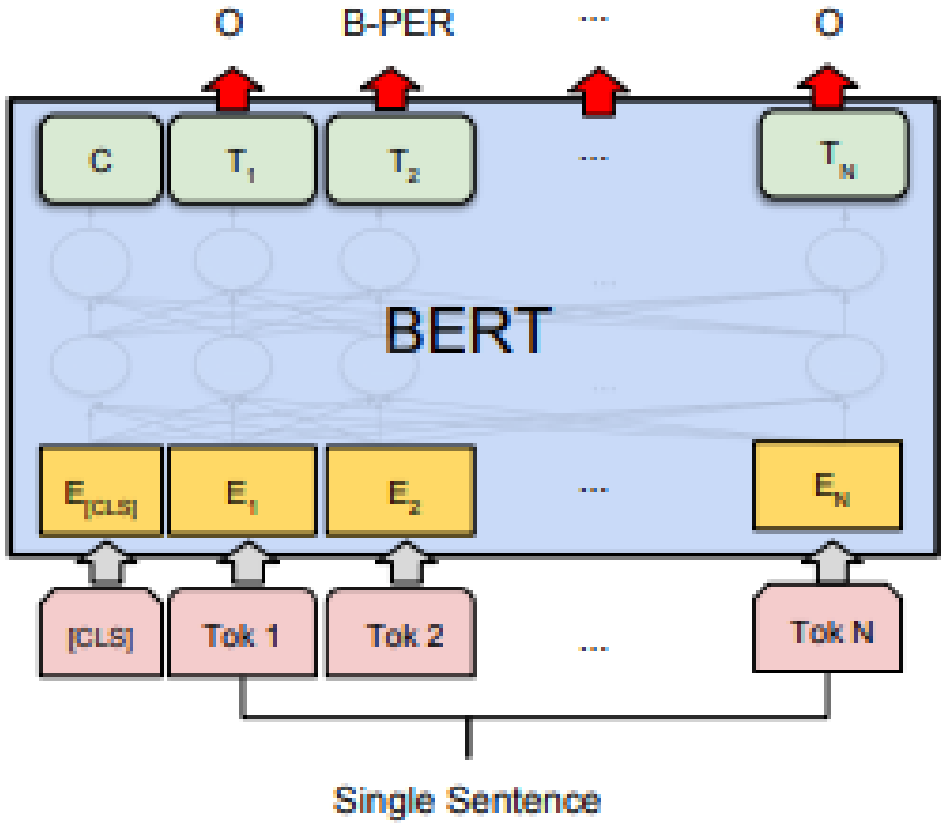
How to use BERT - Named entity recognition

Locate and classify named entity mentions in text into pre-defined categories

4 The Gram-negative pathogen Organism Salmonella enterica serovar Typhimurium experiences a number of acidic environments both inside and outside animal hosts.

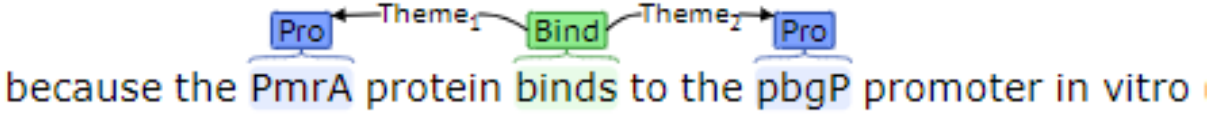
5 Growth in mild acid (pH 5.8) promotes transcription of genes activated by the response regulator Pro PmrA, but the signalling pathway(s) that mediates this response has thus far remained unexplored.

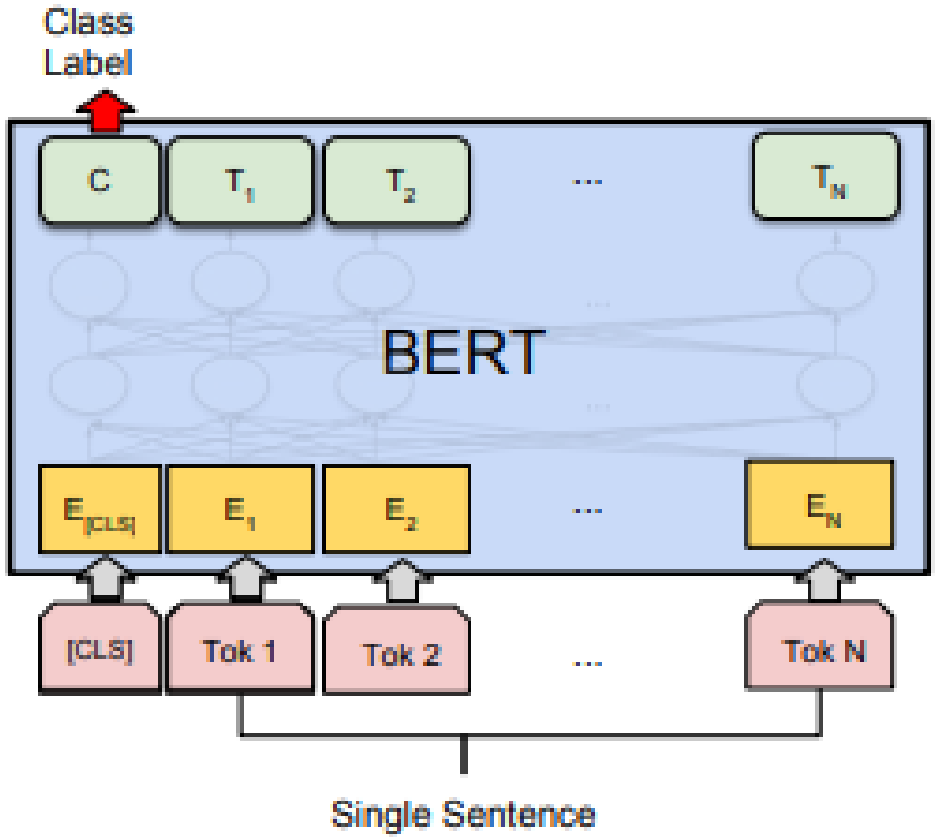
6 Here we report that this activation requires both Pro PmrA's cognate sensor kinase Pro PmrB, which had been previously shown to respond to Chem Fe3+ and Chem Al3+, and Pro PmrA's post-translational activator Pro PmrD.



How to use BERT - Relation extraction

Extract semantic relationships from a text





Performance comparison

Task	Corpus	Train	Dev	Test	Metrics	Domain
Sentence similarity	MedSTS	675	75	318	Pearson	clinical
	BIOSSES	64	16	20	Pearson	biomedical
Named entity recognition	BC5CDR-disease	4182	4244	4424	F1	biomedical
	BC5CDR-chemical	5203	5347	5385	F1	biomedical
	ShAReCLEFE	4628	1075	5195	F1	clinical
Relation Extraction	DDI	2937	1004	979	micro F1	biomedical
	ChemProt	4154	2416	3458	micro F1	biomedical
	i2b2 2010	3110	11	6293	F1	clinical
Document classification	HoC	1108	157	315	F1	biomedical
Inference	MedNLI	11232	1395	1422	accuracy	clinical

Pre-trained models

- General domain
- PubMed
- Clinical notes (MIMIC-III)

Corpus	Words	Domain
Books Corpus	800M	General
English Wikipedia	2,500M	General
PubMed abstract	**4,000M**	**Biomedical**
MIMIC-III	**500M**	**Clinical**

Performance comparison

Task	Corpus	State-of-the-art	ELMo	BERT (General)	BERT (PubMed)	BERT (PubMed +MIMIC)
Sentence similarity	MedSTS	83.6	68.6	84.1	84.5	**84.8**
	BIOSSES	84.8	60.2	84.7	89.3	**91.6**
Named entity recognition	BC5CDR-disease	82.6	83.9	83.2	**86.6**	85.4
	BC5CDR-chemical	91.4	91.5	91.2	**93.5**	92.4
	ShARe/CLEFE	70.0	75.6	75.7	75.4	**77.1**
Relation extraction	DDI	72.9	78.8	76.3	78.1	**79.4**
	Chem-Prot	64.1	66.6	67.5	**72.5**	69.2
	i2b2	73.7	71.2	71.9	74.4	**76.4**
Document classification	HoC	81.5	80.0	81.9	**85.3**	83.1
Inference	MedNLI	73.5	71.4	78.2	82.2	**84.0**

BLUE Benchmark

Biomedical **L**anguage **U**nderstanding **E**valuation (BLUE) benchmark

- Contains diverse range of text genres (biomedical literature and clinical notes)
- Highlight common biomedicine text-mining challenges
- Promote development of language representations in biomedicine domain

BERT limitation

- Length of sentences
- Data size
- Stability

Summary

- Word embeddings
- ELMo
  - Deep contextualized word representation
- BERT
  - Transformer
- Pre-trained BERT models
- How to use BERT
- Performance comparison and benchmark

Resources

- BioWordVec: https://github.com/ncbi-nlp/BioWordVec
- BioSentVec: https://github.com/ncbi-nlp/BioSentVec

(coming soon)

- NCBI BERT: https://github.com/ncbi-nlp/NCBI_BERT
  - Base (PubMed)
  - Base (PubMed + MIMIC-III)
  - Large (PubMed)
  - Large (PubMed + MIMIC-III)
- BLUE benchmark: https://github.com/ncbi-nlp/BLUE_Benchmark

Acknowledgment

- BERT and ELMo
- Shared tasks and datasets
  - BIOSSTS, MedSTS, BioCreative V chemical-disease relation task, ShARe/CLEF eHealth task, DDI extraction 2013 task, BioCreative VI CHEMPROT, i2b2 2010 shared task, Hallmarks of Cancers corpus
- NIH.AI workshop organizers

Thank you!

yifan.peng@nih.gov