Recent Advances in Transferable Representation Learning

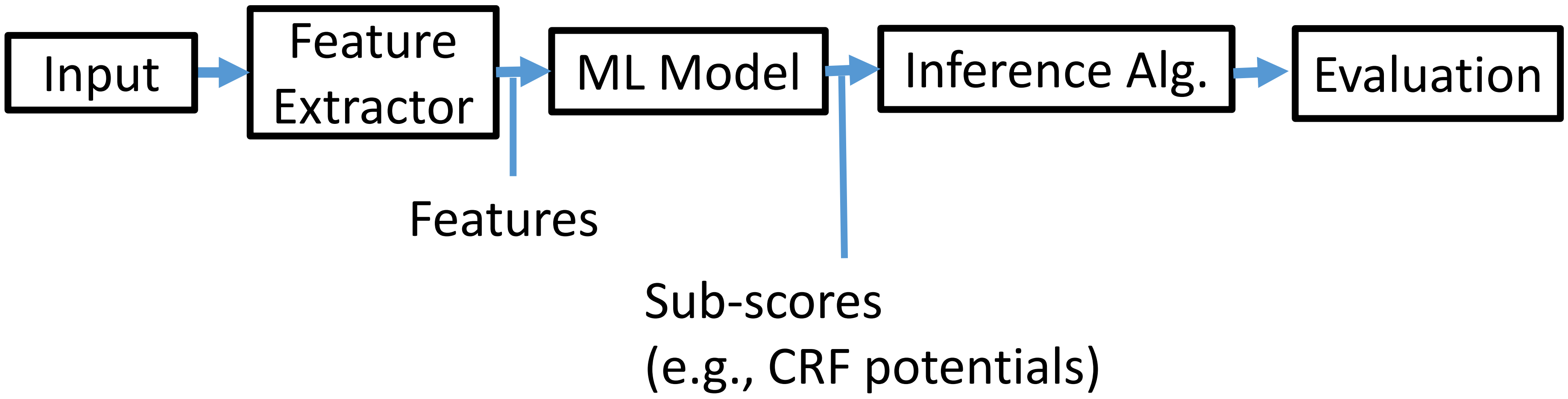
The Basics of Embeddings

Muhao Chen, **Kai-Wei Chang**, Dan Roth

AAAI 2020 Tutorial

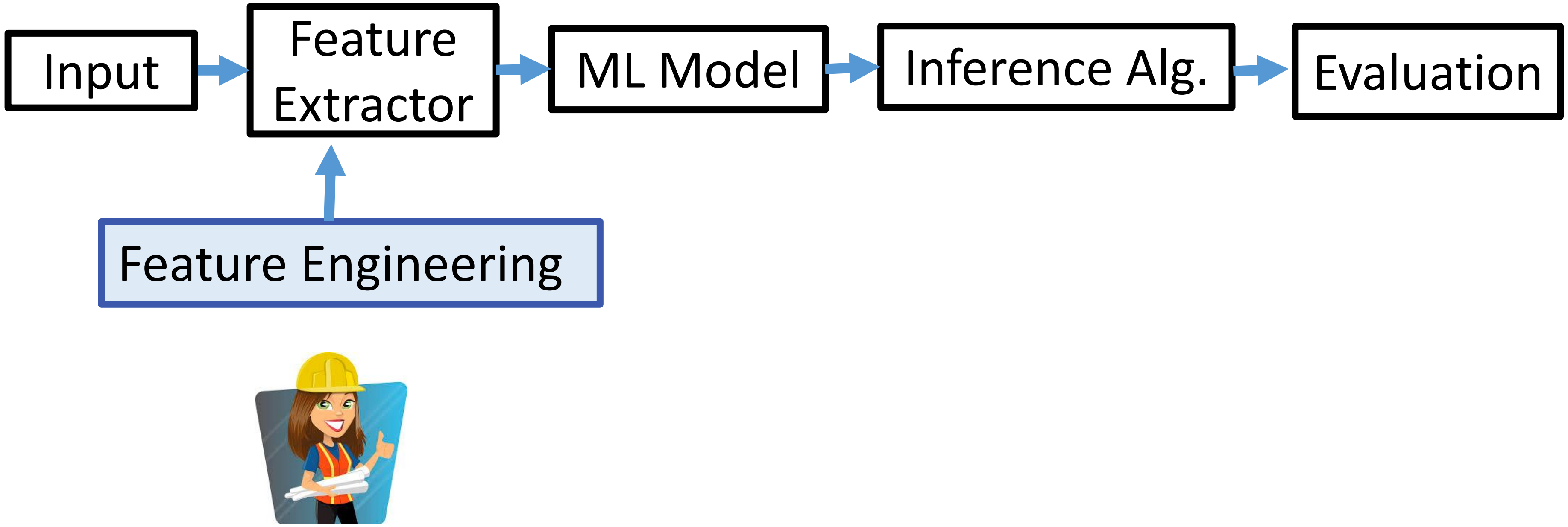
Machine Learning Pipelines

Traditional ML framework



Machine Learning Pipelines

Traditional ML framework



Machine Learning Pipelines

Traditional ML Framework

Input → Feature Extractor → ML Model → Inference Alg. → Evaluation

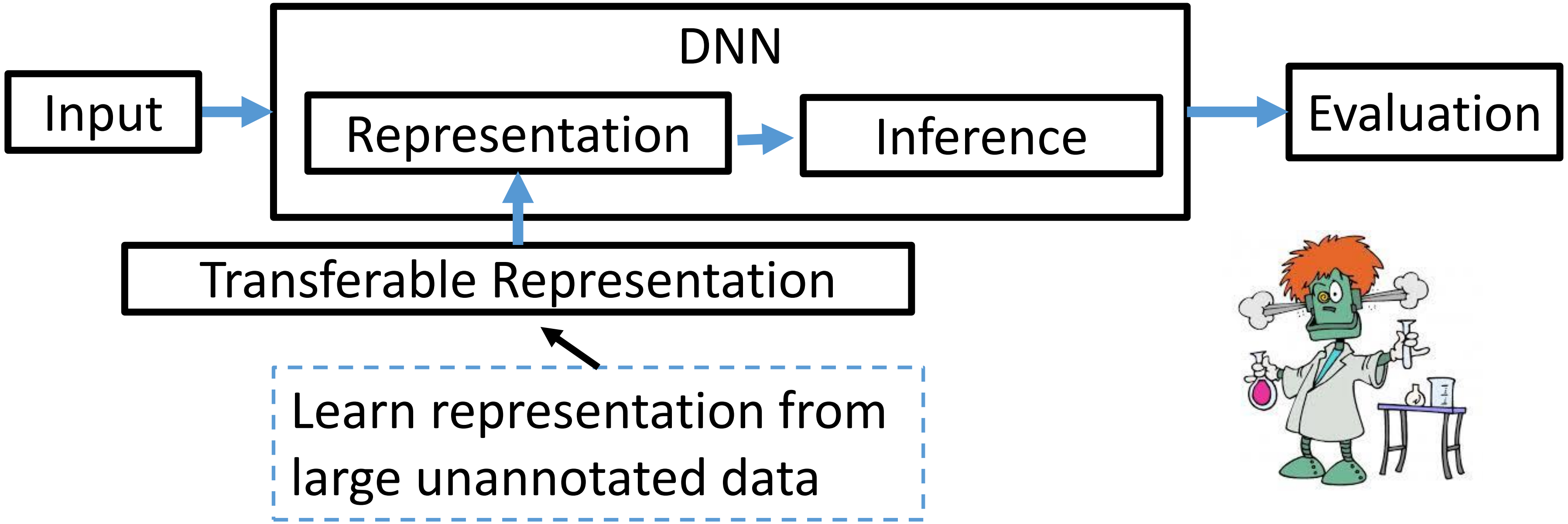
Representation Learning

Deep Learning Framework

Input → DNN [Representation → Inference] → Evaluation

Transferable Representation

Deep Learning Framework



SQuAD2.0 (Rajpurkar & Jia et al. '18)

Packet switching contrasts with another principal networking paradigm, circuit switching, a method which pre-allocates dedicated network bandwidth specifically for each communication session, each having a constant bit rate and latency between nodes. In cases of billable services, such as cellular communication services, circuit switching is characterized by a fee per unit of connection time, even when no data is transferred, while packet switching may be characterized by a fee per unit of information transmitted, such as characters, packets, or messages.

Q: Packet Switching contrast with what other principal

A: circuit switching

A History of Word Representation

How to Represent Words?

Credit: https://www.flickr.com/photos/182229932@N07/48688109908

Idea 1: Representing Words as Discrete Symbols

- Represent word as a "one-hot" vector, e.g.,

$$\overrightarrow{happy} = \begin{array}{ccccccc} [\,0 & 0 & 0 & 1 & 0 & .... & 0\,] \\ \text{egg} & \text{student} & \text{glad} & \text{happy} & \text{buy} & ... & \text{milk} \end{array}$$

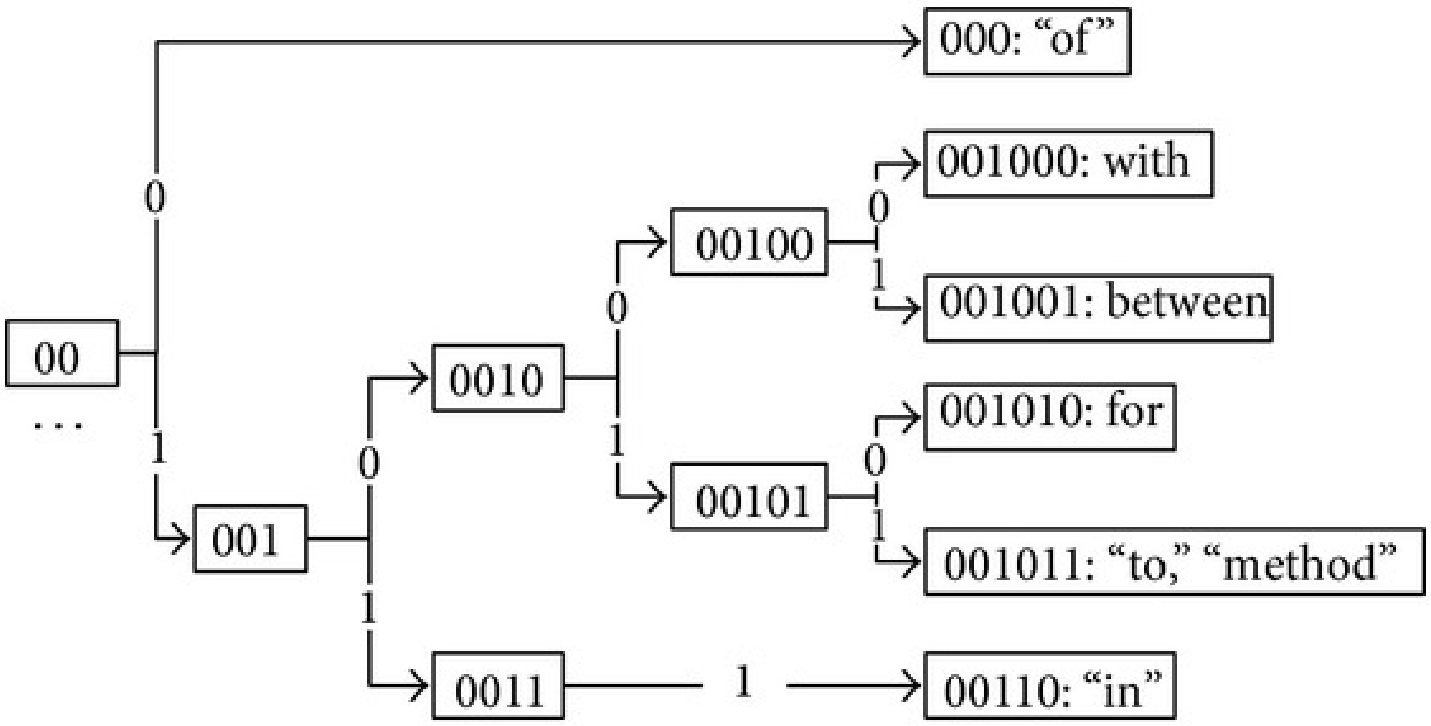
$$\overrightarrow{glad} = \begin{array}{ccccccc} [\,0 & 0 & 1 & 0 & 0 & ... & 0\,] \\ \text{egg} & \text{student} & \text{glad} & \text{happy} & \text{buy} & ... & \text{milk} \end{array}$$

- How large is this vector?
  - Vector dimension = number of words in vocabulary PTB data: ~50k, Google 1T data: 13M
- Issue: no notion of similarity.

$\overrightarrow{happy}$ and $\overrightarrow{glad}$ are orthogonal.

Idea 2: Similarity = Clustering

- Brown Cluster https://en.wikipedia.org/wiki/Brown_clustering
- Dictionary: e.g., WordNet (Miller 1995)

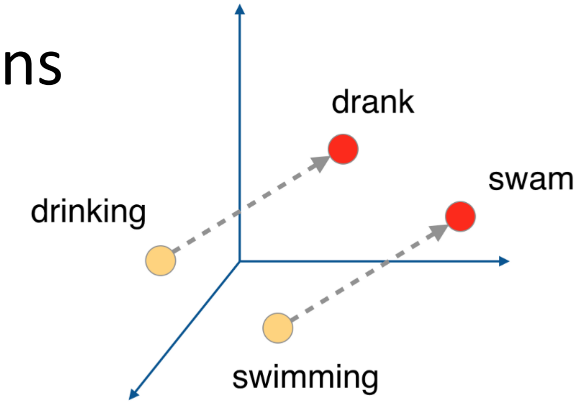


Idea 3: Vector Representations

- Discrete ⇒ continuous: A dense vector for each word
- Words with similar meaning are closer in the embedding space
- Word meanings are vector of "basic concepts"
  - ☐ The "basic concepts" might not be explicit

$$
\begin{array}{llllll}
v_{king} = & [\ 0.8 & 0.9 & 0.1 & 0 & \ldots\ ] \\
v_{queen} = & [\ 0.8 & 0.1 & 0.8 & 0 & \ldots\ ] \\
v_{apply} = & [\ 0.1 & 0.2 & 0.1 & 0.8 & \ldots\ ]
\end{array}
$$

  - ☐ Difference between word vectors captures their relations
    [Mikolov+ 13, Pennington+ 14]



How to Learn word Vectors

- Distributional hypothesis:
  “You shall know a word by the company it keeps” (J. R. Firth 1957: 11)
  linguistic items with similar distributions have similar meanings.

context words

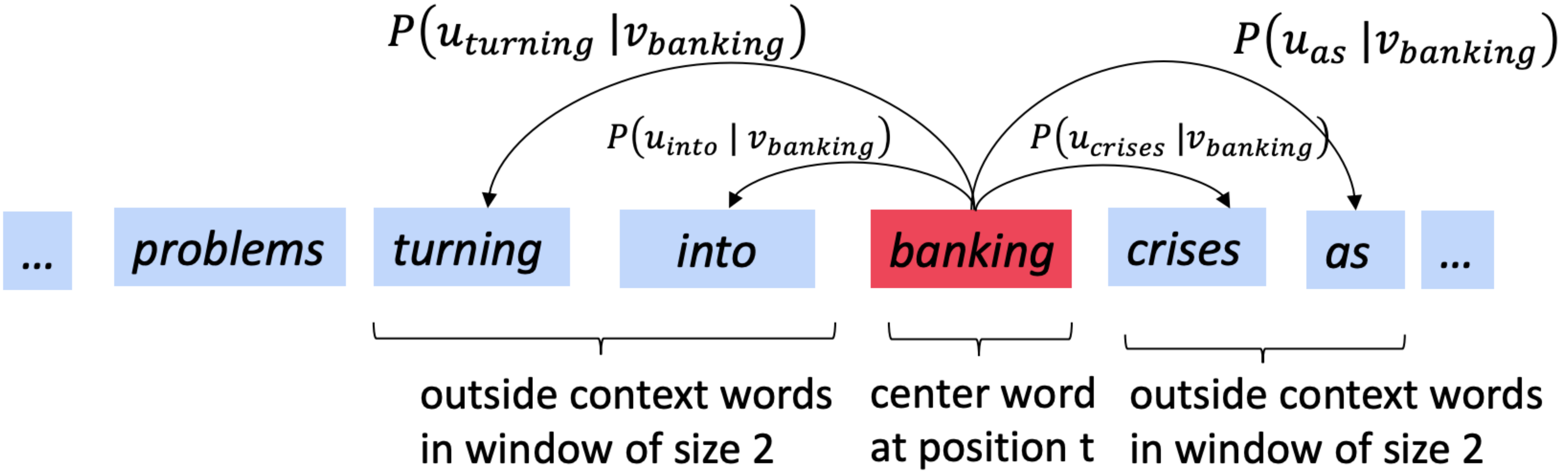
he curtains open and the	stars	shining in on the barely
ars and the cold , close	stars	" . And neither of the w
rough the night with the	stars	shining so brightly , it
made in the light of the	stars	. It all boils down , wr
surely under the bright	stars	, thrilled by ice-white
sun , the seasons of the	stars	? Home , alone , Jay pla
m is dazzling snow , the	stars	have risen full and cold
un and the temple of the	stars	, driving out of the hug
in the dark and now the	stars	rise , full and amber a
bird on the shape of the	stars	over the trees in front
But I could n’t see the	stars	or the moon , only the
they love the sun , the	stars	and the stars . None of
r the light of the shiny	stars	. The plash of flowing w
man ’s first look at the	stars	; various exhibits , aer
rief information on both	stars	and constellations, inc

context words

How to Learn word Vectors

- Learn word representations based on co-occurrences
- E.g., Word2vec [Mikolov+ 2013]

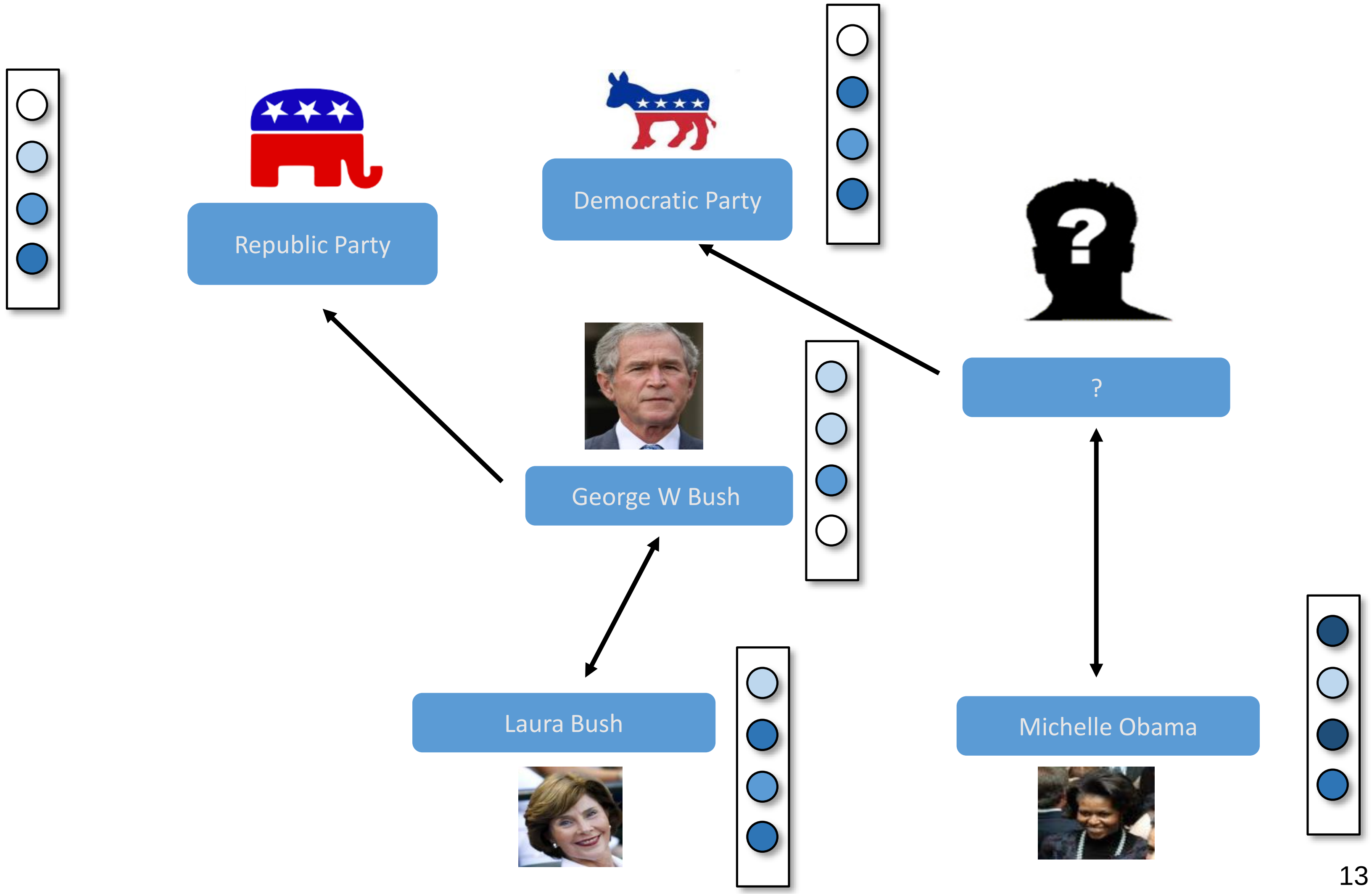
$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$



Slide credit: Stanford cs224n

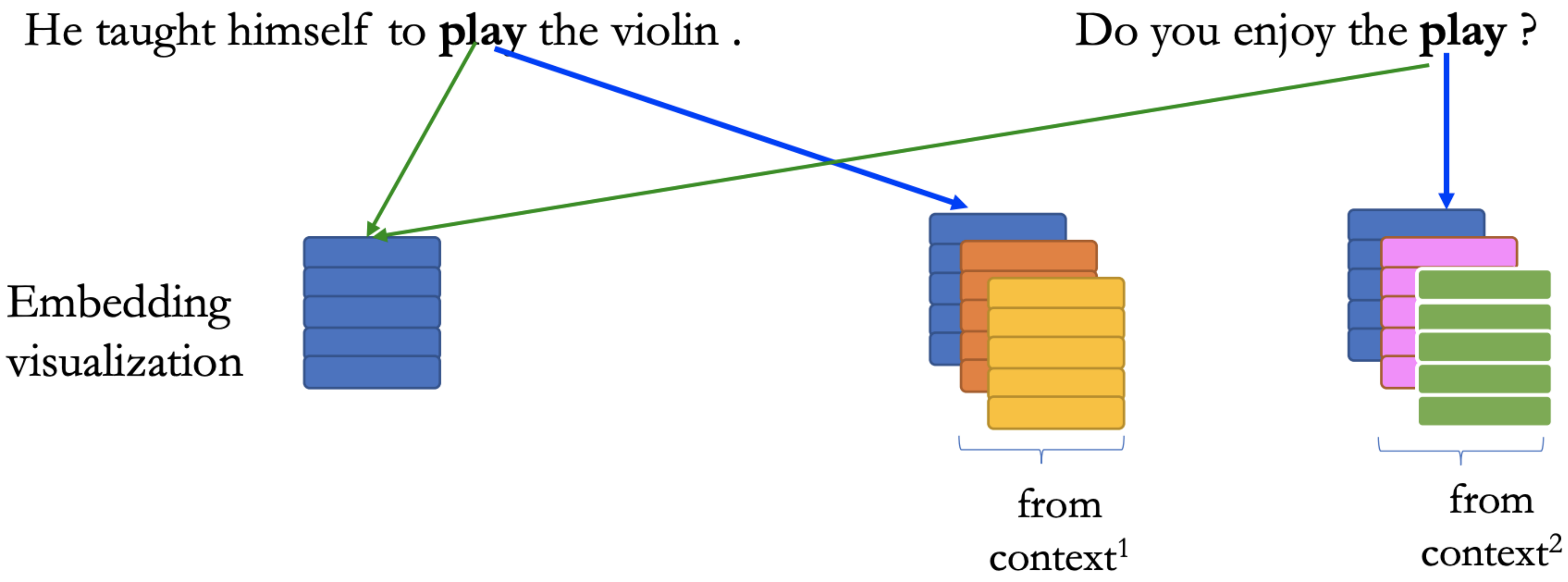
Continuous representations for entities

Embeddings can be learned from Freebase, Dbpedia, YAGO, NELL, etc.



Contextualized Word Representations

- Most words have multiple meanings
- Can we encode word also based on the surrounding contexts?

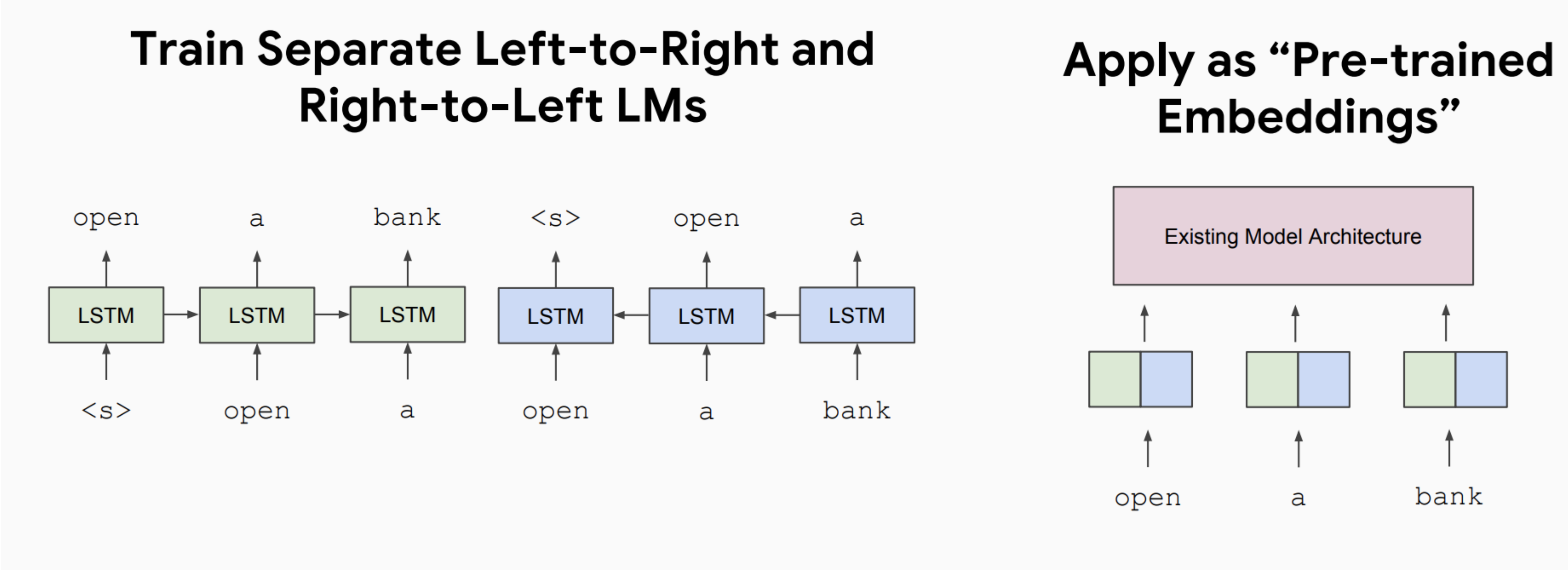


Word Embedding

Contextulaized Word Embedding

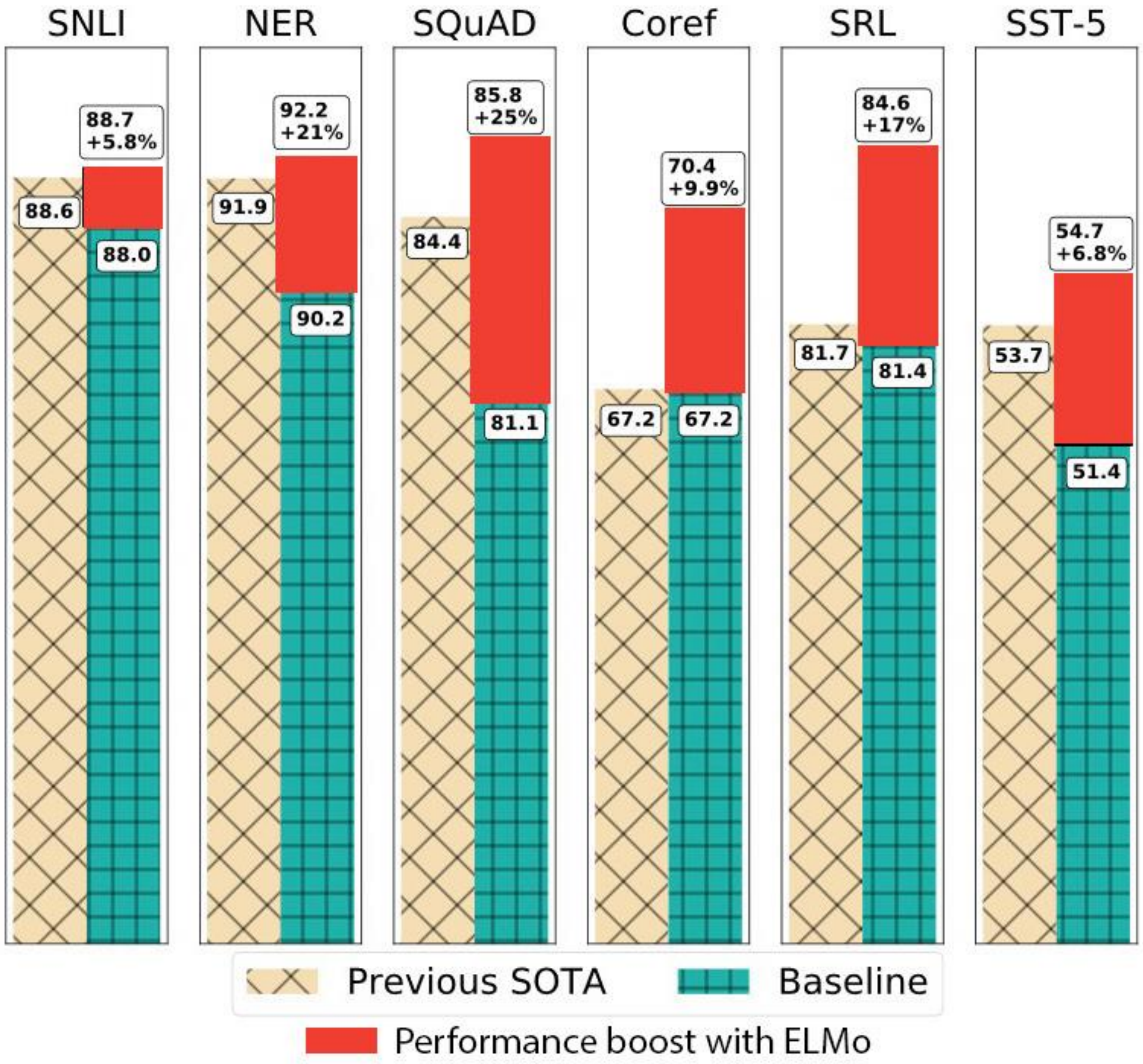
Embeddings from Language Models (ELMo)

Deep contextualized word representations (Peters+2018)

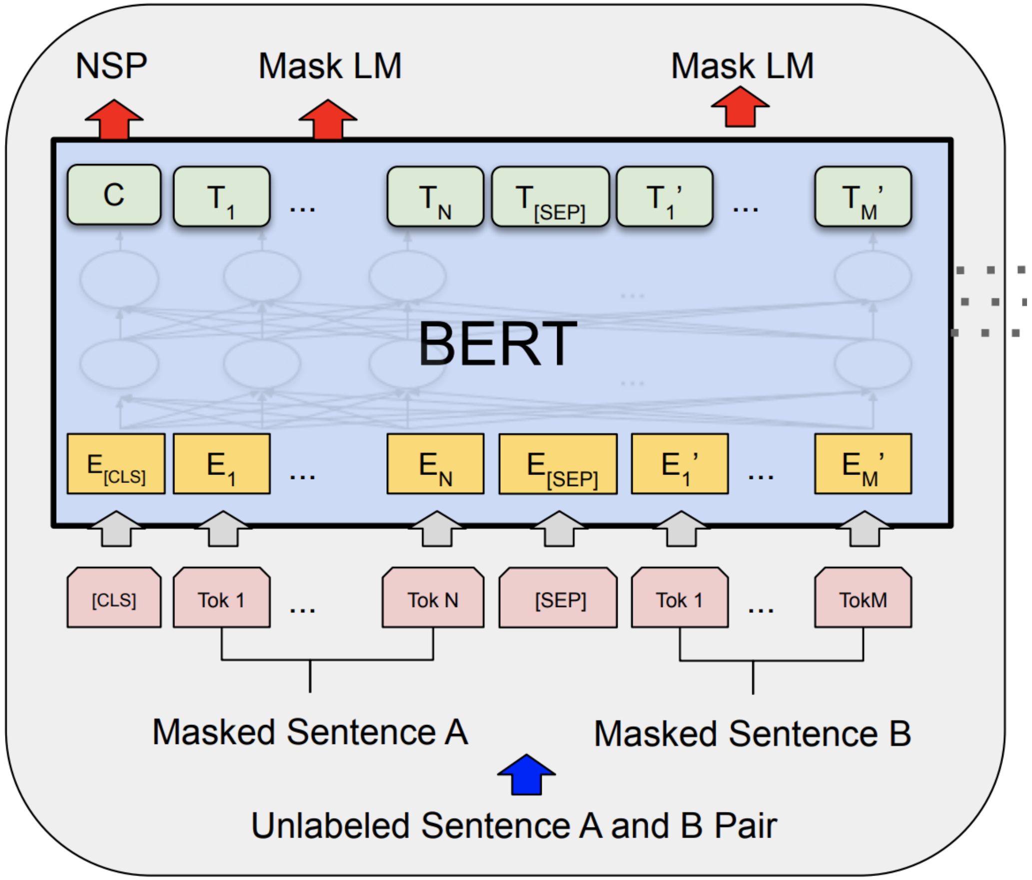


Learning transferable representations using language model objective.

Performance Boost with ELMo

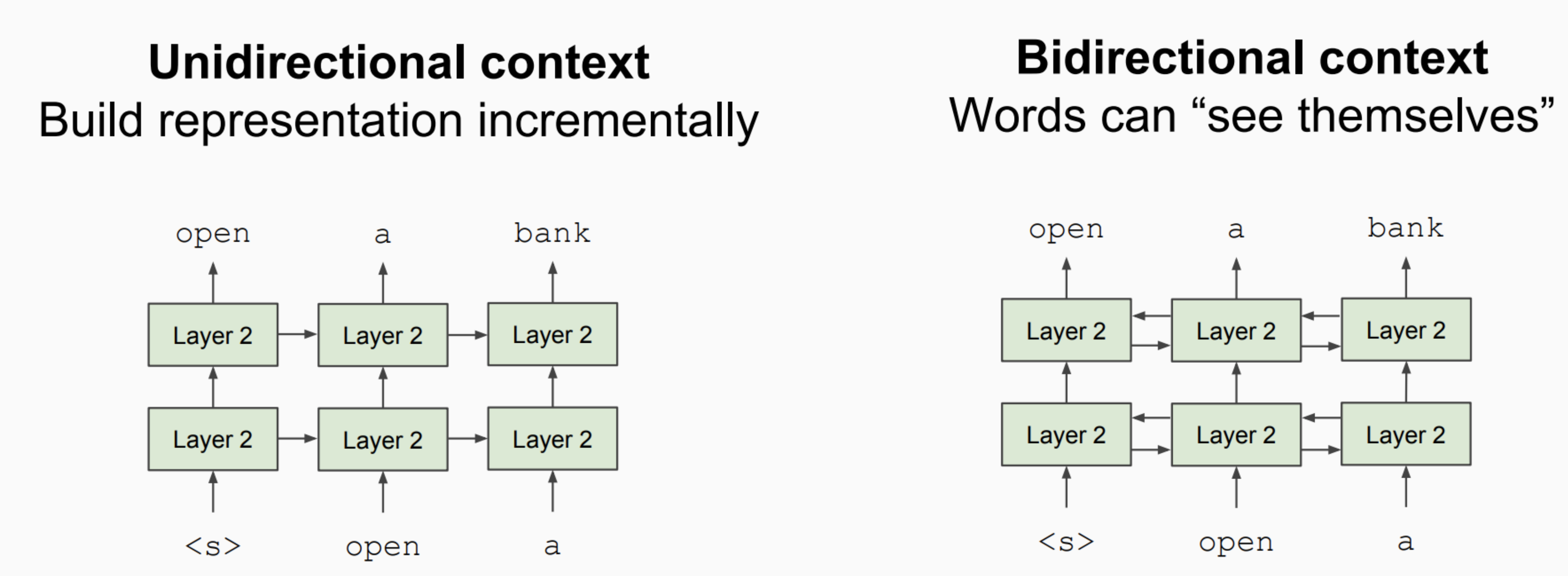


Bidirectional Encoder Representations from Transformers (BERT)



Masked Language Model

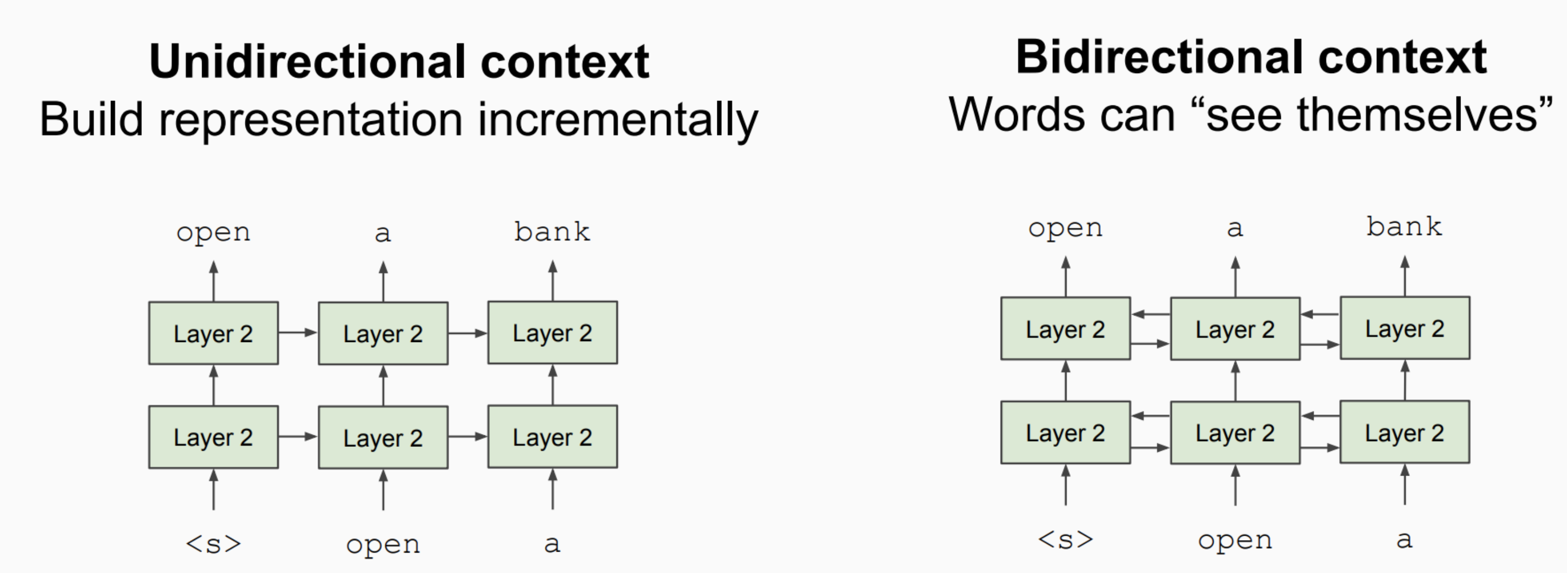
- How to jointly capture the context information from both directions?



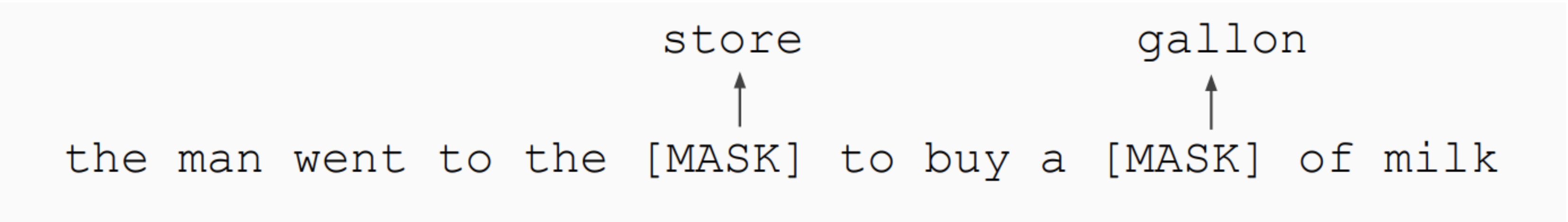
Slides from Jacob Devlin

Masked Language Model

- How to jointly capture the context information from both directions?

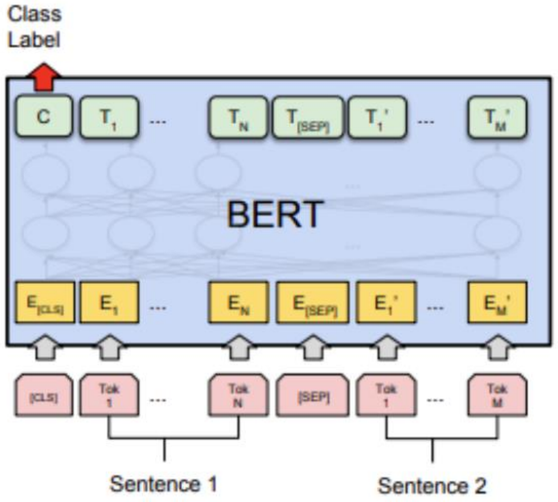


- Masked Language Model: Mask out k% of the input words

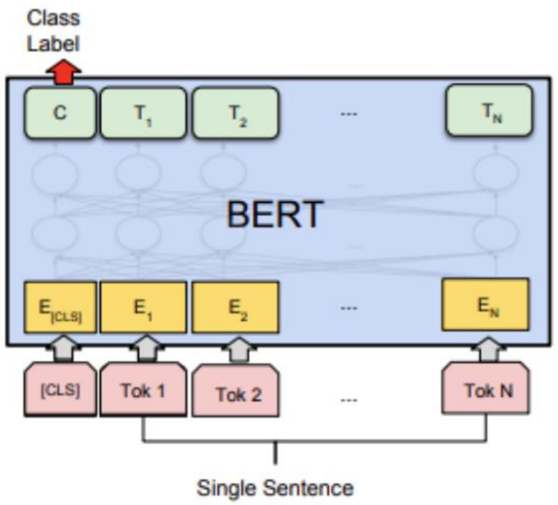


Slides from Jacob Devlin

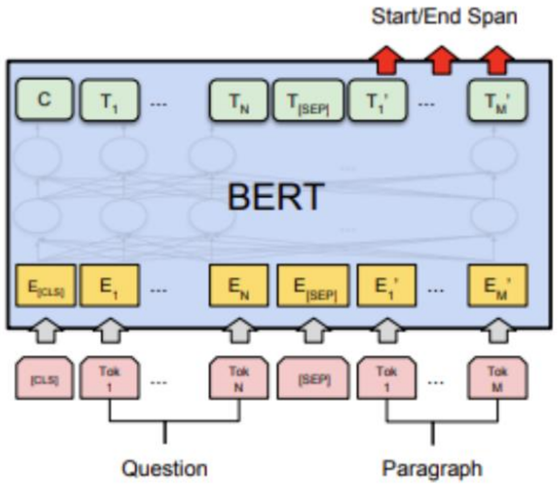
Fine-Tuning Procedure



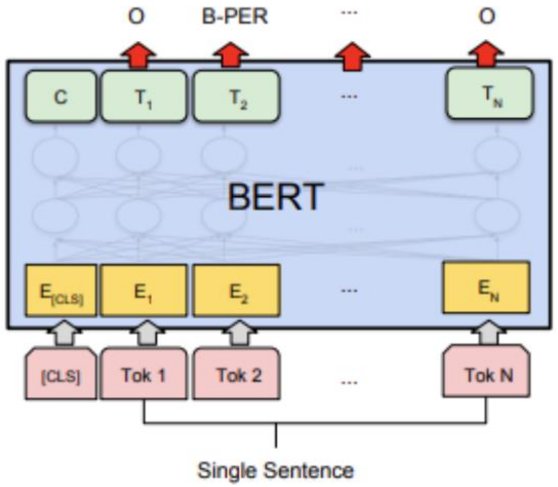
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER