CS769 Advanced NLP

Pre-trained Sentence and Contextualized Word Representations

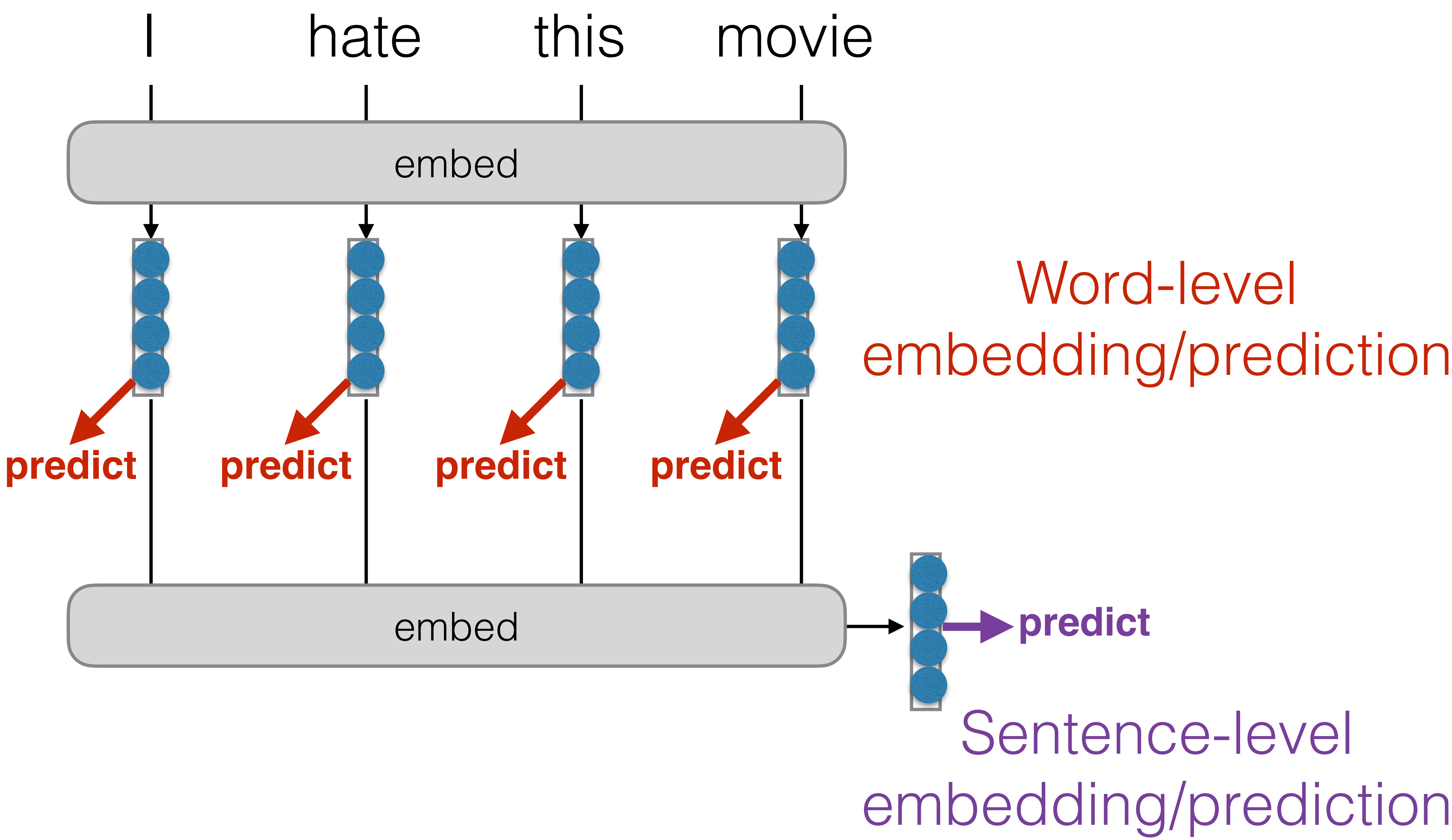
Junjie Hu

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Slides adapted from Graham, Antonis

https://junjiehu.github.io/cs769-spring22/

Remember: Neural Models



Goal for Today

- Discuss **contextualized word** and **sentence** representations
- Briefly Introduce **tasks**, **datasets** and **methods**
- Introduce different **training objectives**
- Talk about **multitask/transfer learning**

Multi-task Learning Overview

Types of Learning

- **Multi-task learning** is a general term for training on multiple tasks
- **Transfer learning** is a type of multi-task learning where we only really care about one of the tasks
- **Pre-training** is a type of transfer learning where one pre-training objective is used first

Plethora of Tasks in NLP

- In NLP, there are a plethora of tasks, each requiring different varieties of data
  - **Only text:** e.g. language modeling
  - **Naturally occurring data:** e.g. machine translation
  - **Hand-labeled data:** e.g. most analysis tasks
- And each in many languages, many domains!

Rule of Thumb 1: Multitask to Increase Data

- Perform multi-tasking when one of your two tasks has many fewer data

- **General domain → specific domain**
  (e.g. web text → medical text)

- **High-resourced language → low-resourced language**
  (e.g. English → Telugu)

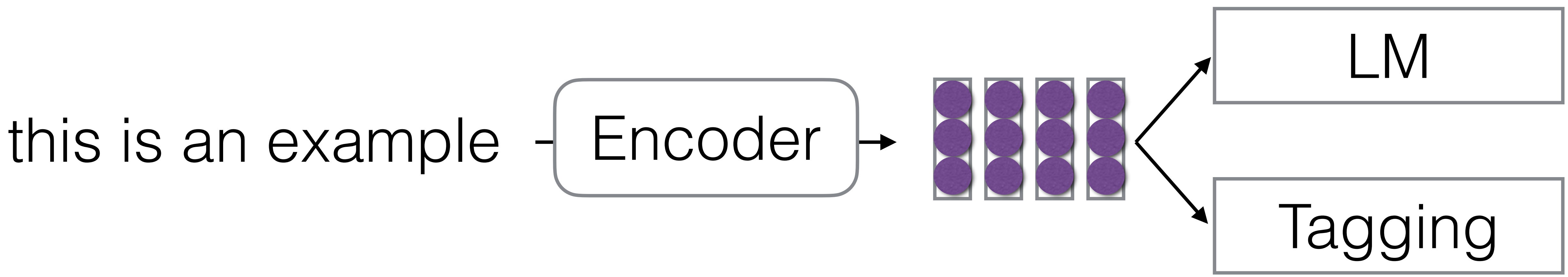
- **Plain text → labeled text**
  (e.g. LM -> parser)

Rule of Thumb 2:

- Perform multi-tasking when your **tasks are related**
- e.g. predicting eye gaze and summarization (Klerke et al. 2016)

Standard Multi-task Learning

- Train representations to do well on multiple tasks at once

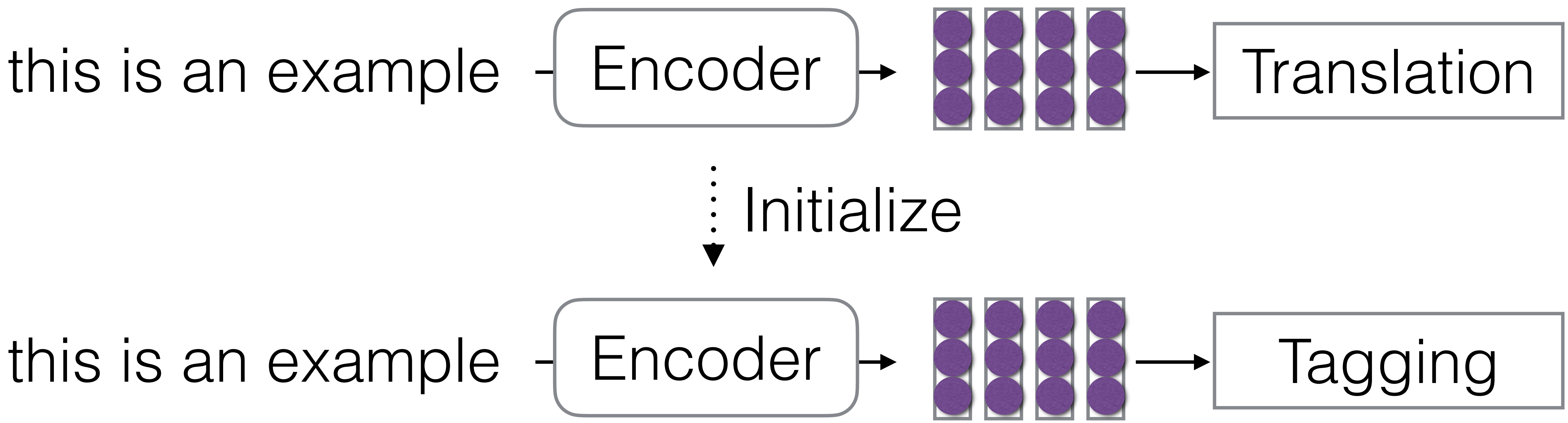


- In general, as simple as randomly choosing minibatch from one of multiple tasks

- Many many examples, starting with Collobert and Weston (2011)

Pre-training

- First train on one task, then train on another



- Widely used in word embeddings (Turian et al. 2010), sentence encoders (Dai et al. 2015) or contextualized word representations (Melamud et al. 2016)

Thinking about Multi-tasking, and Pre-trained Representations

- Many methods have names like ELMo, BERT, RoBERTa, XLNet along with pre-trained models
- These often refer to a combination of
  - **Model:** The underlying neural network architecture
  - **Training Objective:** What objective is used to pre-train
  - **Data:** What data the authors chose to use to train the model
- Remember that these are often conflated (and don't need to be)!

Tasks Using Sentence Representations

Where would we need/use Sentence Representations?

- Sentence Classification
- Paraphrase Identification
- Semantic Similarity
- Entailment
- Retrieval

Sentence Classification

- Classify sentences according to various traits
- Topic, sentiment, subjectivity/objectivity, etc.

I hate this movie

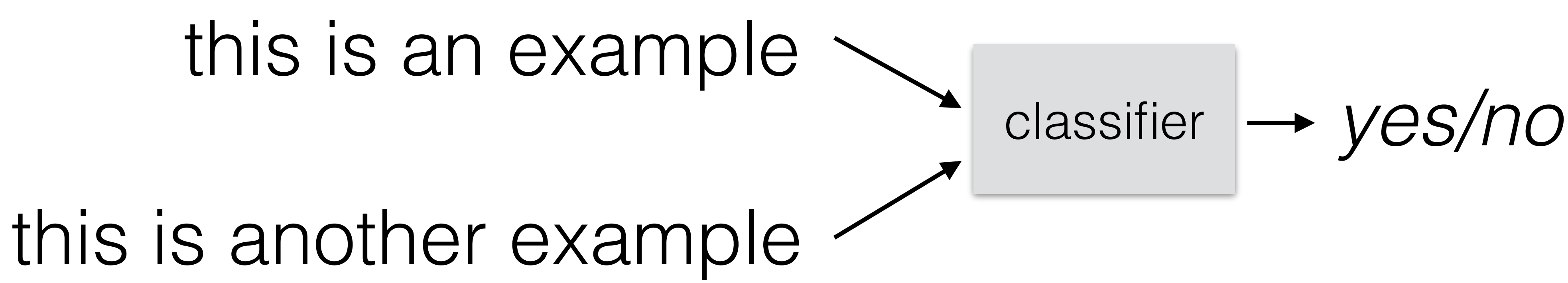
very good
good
neutral
bad
very bad

I love this movie

very good
good
neutral
bad
very bad

Sentence Pair Classification

- Classify over multiple sentences



Paraphrase Identification
(Dolan and Brockett 2005)

- Identify whether A and B mean the same thing

Charles O. Prince, 53, was named as Mr. Weill's successor.

↕

Mr. Weill's longtime confidant, Charles O. Prince, 53, was named as his successor.

- **Note:** *exactly* the same thing is too restrictive, so use a loose sense of similarity

Semantic Similarity/Relatedness
(Marelli et al. 2014)

- Do two sentences mean something similar?

Relatedness score	Example
1.6	A: "*A man is jumping into an empty pool*" B: "*There is no biker jumping in the air*"
2.9	A: "*Two children are lying in the snow and are making snow angels*" B: "*Two angels are making snow on the lying children*"
3.6	A: "*The young boys are playing outdoors and the man is smiling nearby*" B: "*There is no boy playing outdoors and there is no man smiling*"
4.9	A: "*A person in a black jacket is doing tricks on a motorbike*" B: "*A man in a black jacket is doing tricks on a motorbike*"

- Like paraphrase identification, but with shades of gray.

Textual Entailment
(Dagan et al. 2006, Marelli et al. 2014)

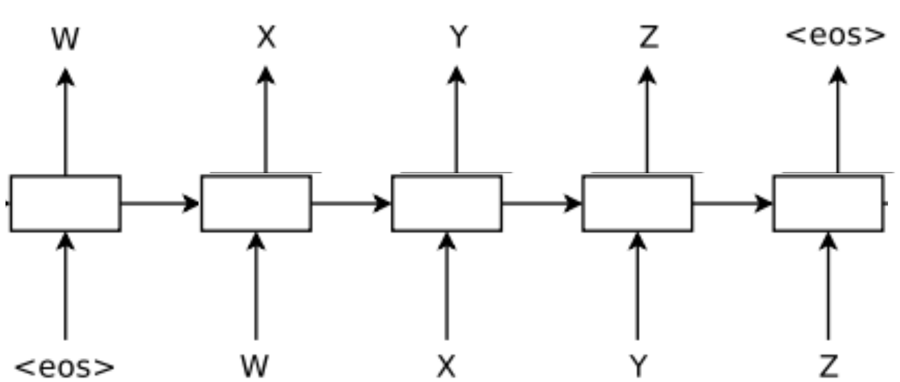
- **Entailment:** if A is true, then B is true (c.f. paraphrase, where opposite is also true)
  - The woman bought a sandwich for lunch
    → The woman bought lunch
- **Contradiction:** if A is true, then B is not true
  - The woman bought a sandwich for lunch
    → The woman did not buy a sandwich
- **Neutral:** cannot say either of the above
  - The woman bought a sandwich for lunch
    → The woman bought a sandwich for dinner

Training Sentence Representations

Language Model+Transfer

(Dai and Le 2015)

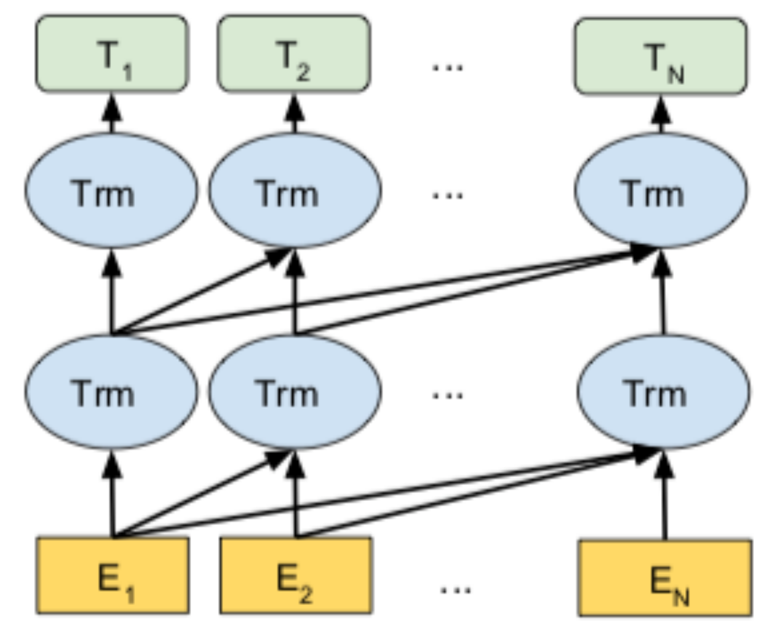
- **Model:** LSTM
- **Objective:** LM objective
- **Data:** Classification data itself, or Amazon reviews



- **Downstream:** On text classification, initialize weights and continue training

"GPT" (Radford et al. 2018)

- **Model:** Masked self-attention
- **Objective:** LM objective
- **Data:** BooksCorpus

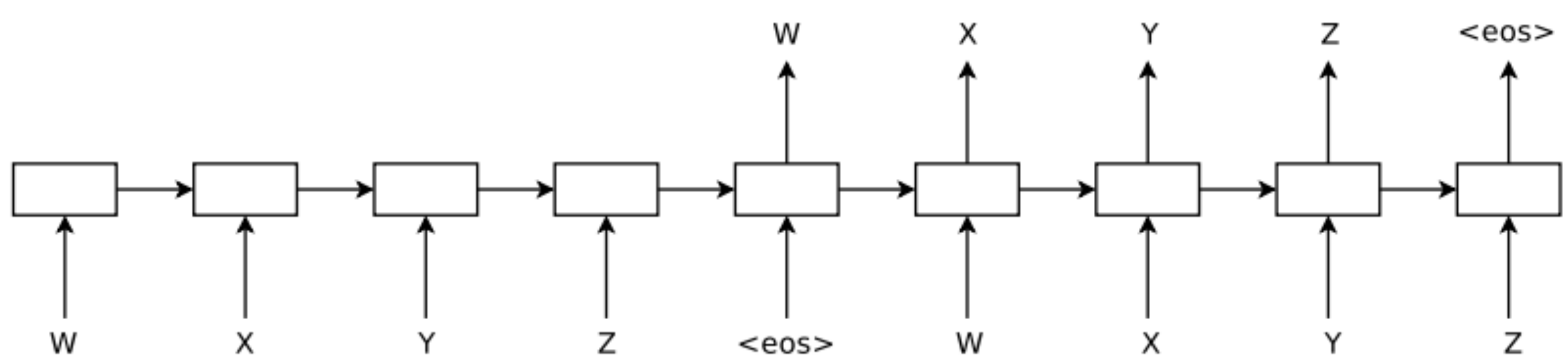


Downstream: Some task fine-tuning, other tasks additional multi-sentence training

Auto-encoder+Transfer

(Dai and Le 2015)

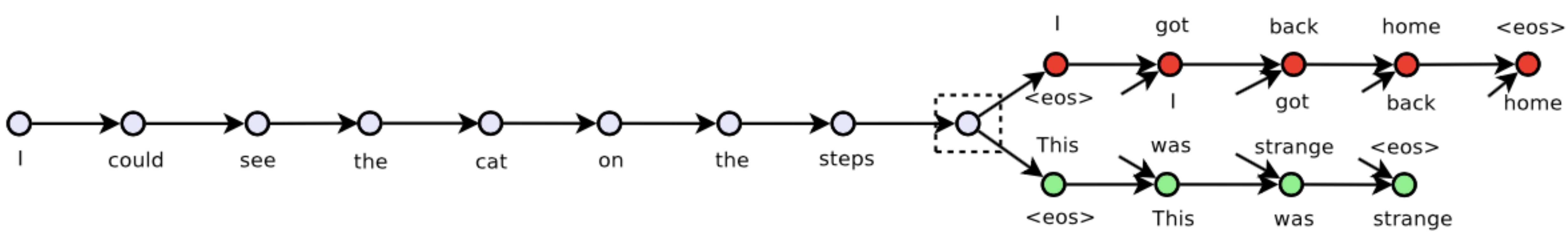
- **Model:** LSTM
- **Objective:** From single sentence vector, re-construct the sentence
- **Data:** Classification data itself, or Amazon reviews



- **Downstream:** On text classification, initialize weights and continue training

Sentence-level Context Prediction+Transfer: "Skip-thought Vectors" (Kiros et al. 2015)

- **Model:** LSTM
- **Objective:** Predict the surrounding sentences
- **Data:** Books, important because of context



- **Downstream Usage:** Train logistic regression on [|u-v|; u*v] (component-wise) for sentence pair classification (u, v are two sentence embeddings)

Paraphrase-based Contrastive Learniing

(Wieting et al. 2015)

- **Model:** Try many different ones
- **Objective:** Predict whether two phrases are paraphrases or not from
- **Data:** Paraphrase database (http://paraphrase.org), created from bilingual data
- **Downstream Usage:** Sentence similarity, classification, etc.
- **Result:** Interestingly, LSTMs work well on in-domain data, but word averaging generalizes better

Large Scale Paraphrase Data (ParaNMT-50MT)

(Wieting and Gimpel 2018)

- **Automatic construction of large paraphrase DB**
  - Get large parallel corpus (English-Czech)
  - Translate the Czech side using a SOTA NMT system
  - Get automated score and annotate a sample
- Corpus is **huge but includes noise**, 50M sentences (about 30M are high quality)
- Trained representations work quite well and generalize

Entailment+Transfer "InferSent"

(Conneau et al. 2017)

- Previous objectives use no human labels, but what if:
- **Objective:** supervised training for a task such as entailment learn generalizable embeddings?
  - Task is more difficult and requires capturing nuance → yes?, or data is much smaller → no?
- **Model:** Bi-LSTM + max pooling
- **Data:** Stanford NLI, MultiNLI
- **Results:** Tends to be better than unsupervised objectives such as SkipThought

Sentence Transformers
(Reimers and Gurevych 2019)

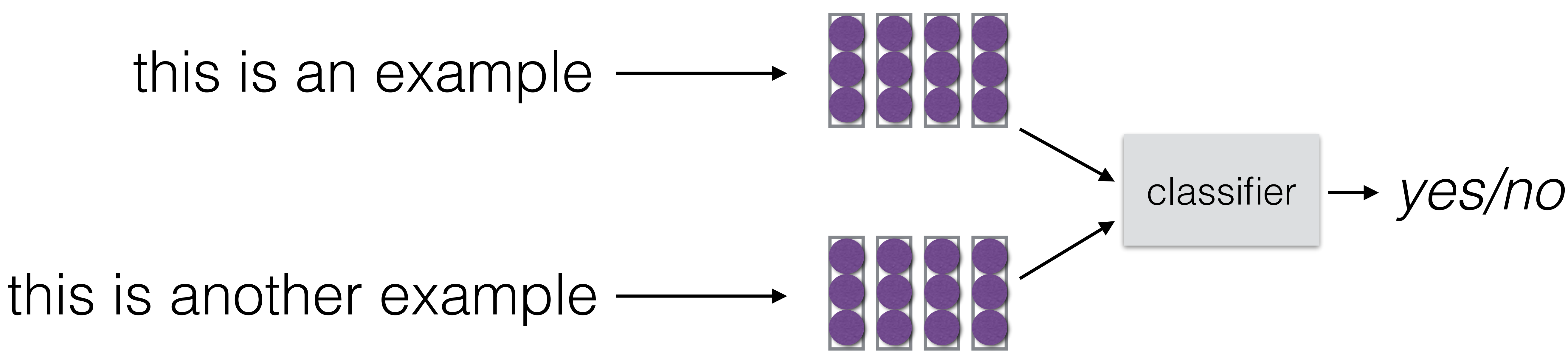
- A toolkit that implements a large number of sentence representations (e.g. BERT, paraphrase)

https://www.sbert.net/

Contextualized Word Representations

Contextualized Word Representations

- Instead of one vector per sentence, one vector per word!



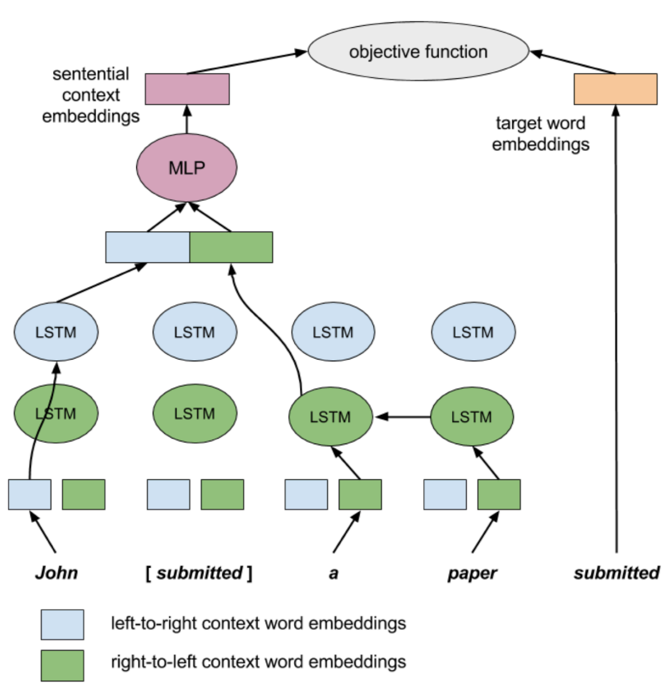
How to train this representation?

Central Word Prediction

context2vec

(Melamud et al. 2016)

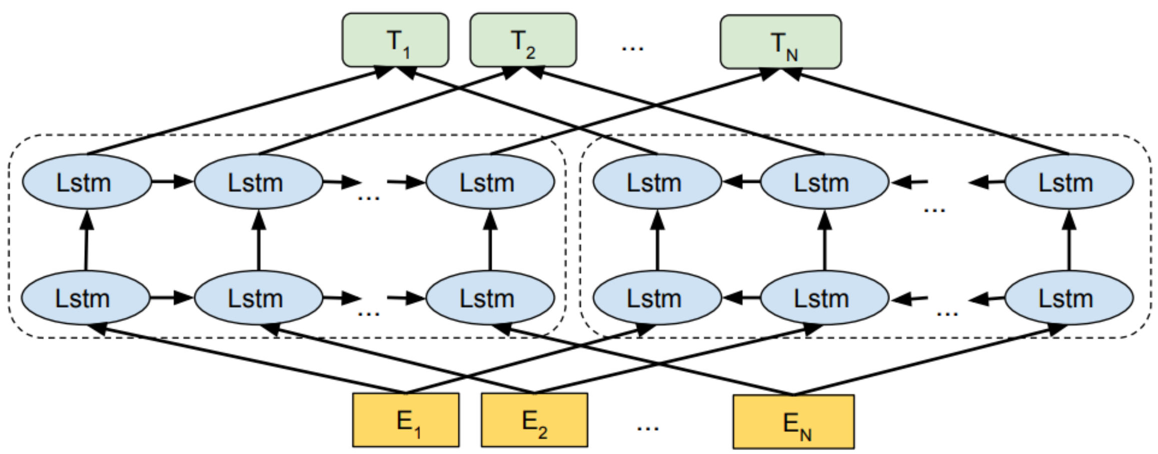
- **Model:** Bi-directional LSTM
- **Objective:** Predict the word given context
- **Data:** 2B word ukWaC corpus
- **Downstream:** use vectors for sentence completion, word sense disambiguation, etc.



ELMo

(Peters et al. 2018)

- **Model:** Multi-layer bi-directional LSTM
- **Objective:** Predict the next word left->right, next word right->left independently

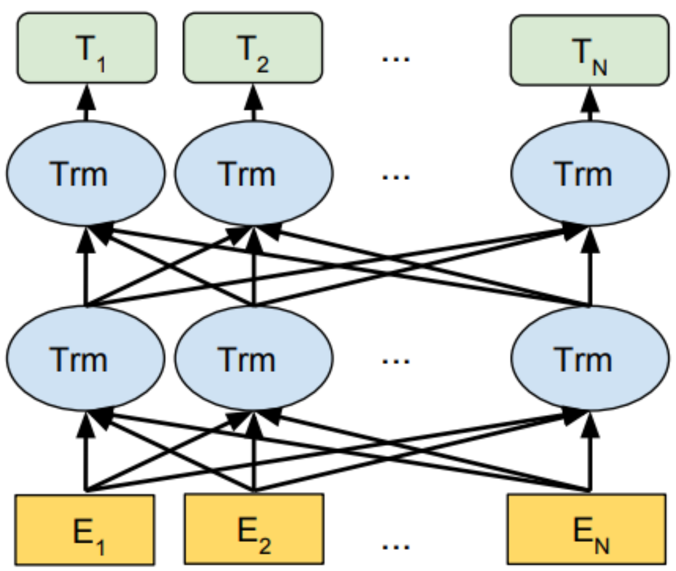


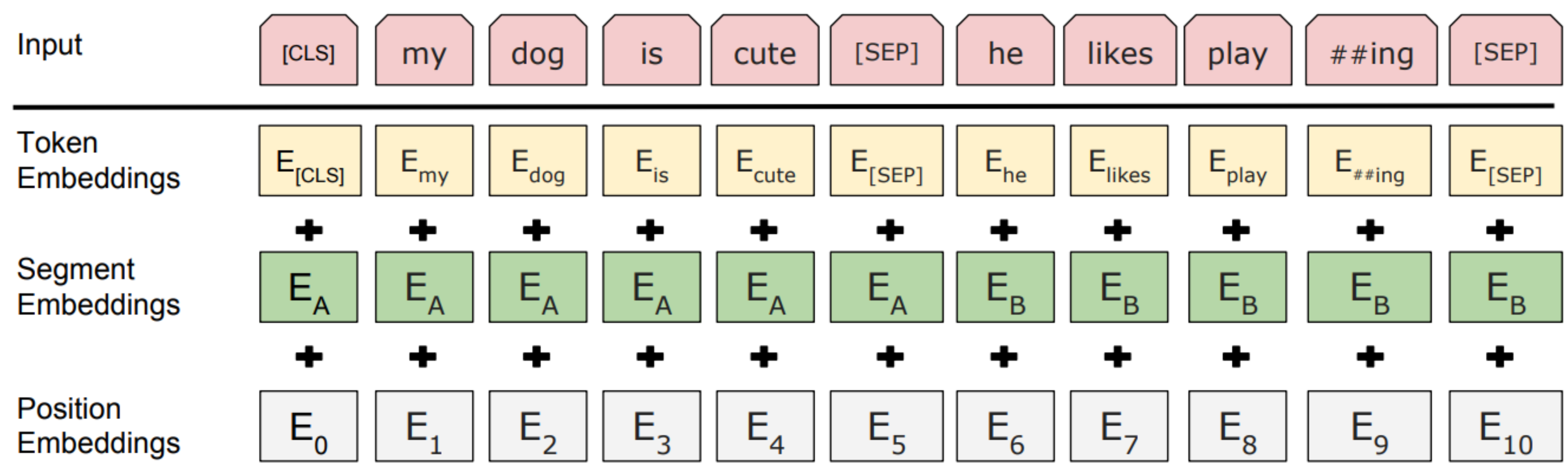
- **Data:** 1B word benchmark LM dataset
- **Downstream:** Finetune the weights of the linear combination of layers on the downstream task

Masked Word Prediction (BERT)

(Devlin et al. 2018)

- **Model:** Multi-layer self-attention. Input sentence or pair, w/ [CLS] token, subword representation





- **Objective:** Masked word prediction + next-sentence prediction
- **Data:** BooksCorpus + English Wikipedia (16GB)

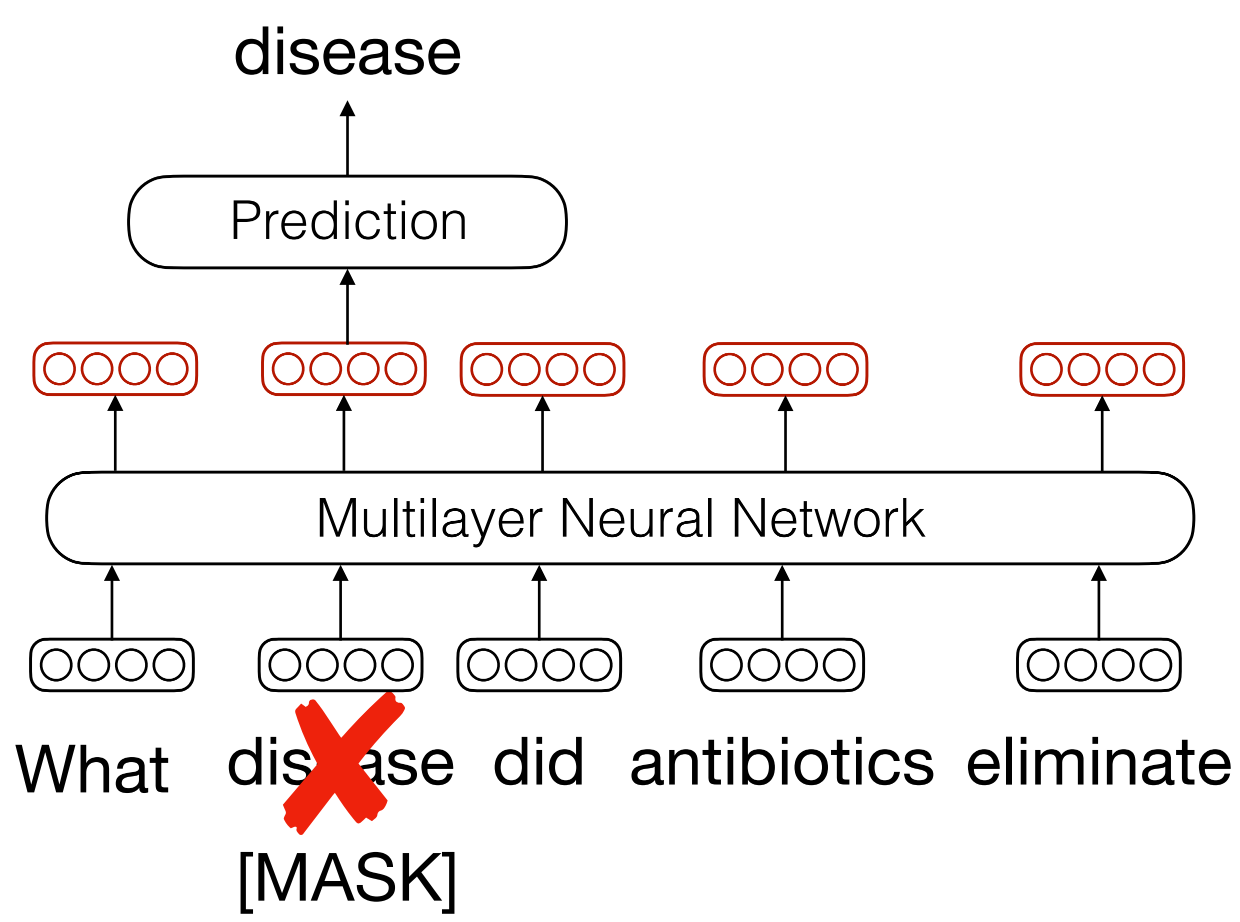
Masked Word Prediction
(Devlin et al. 2018)

1. Select 15% of words at random in a sequence
2. For these selected words:
   - 80% of the time: substitute selected word with [MASK]
   - 10% of the time: substitute selected word with random word
   - 10% of the time: no change
3. Predict all the masked words

- Like context2vec, but **better suited for multi-layer self attention**

Masked Word Prediction
(Devlin et al. 2018)

$$P(x_{\text{mask}} | x_{\text{unmasked}})$$



Consecutive Sentence Prediction

(Devlin et al. 2018)

1. classify two sentences as consecutive or not:
   - 50% of training data (from OpenBooks) is "consecutive"

Input = `[CLS] the man [MASK] to the store [SEP]`
`penguin [MASK] are flight ##less birds [SEP]`
Label = `NotNext`

Input = `[CLS] the man went to [MASK] store [SEP]`
`he bought a gallon [MASK] milk [SEP]`
Label = `IsNext`

Hyperparameter Optimization/Data (RoBERTa)

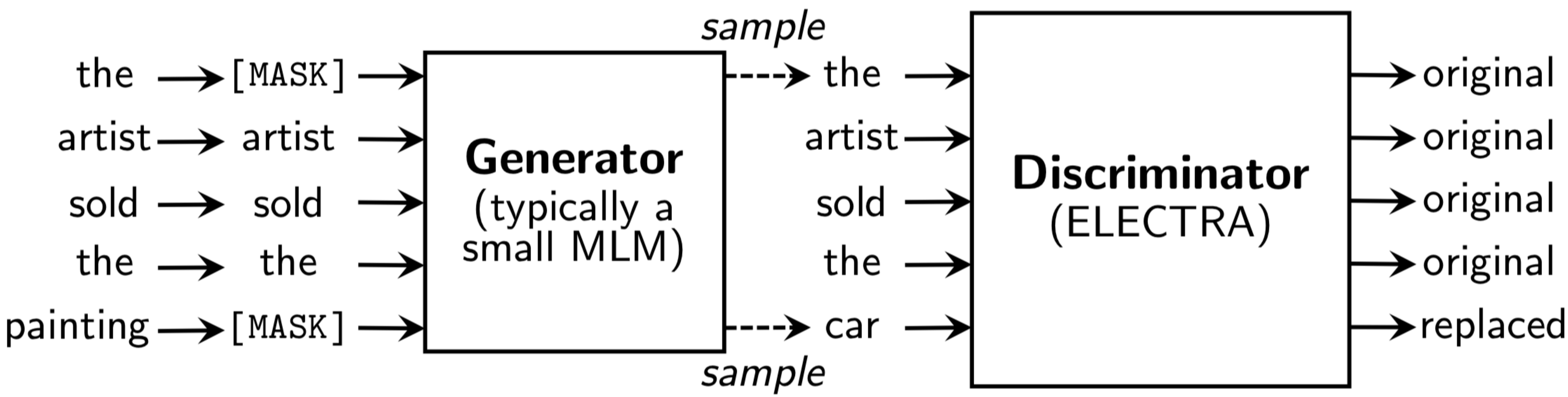
(Liu et al. 2019)

- **Model:** Same as BERT
- **Objective:** Same as BERT, but *train longer* and *drop sentence prediction* objective
- **Data:** BooksCorpus & English Wikipedia (16GB) + CC-News (76GB) + OpenWebText (38GB) + Stories (31GB)
- **Results:** are empirically much better

Distribution Discrimination (ELECTRA)

(Clark et al. 2020)

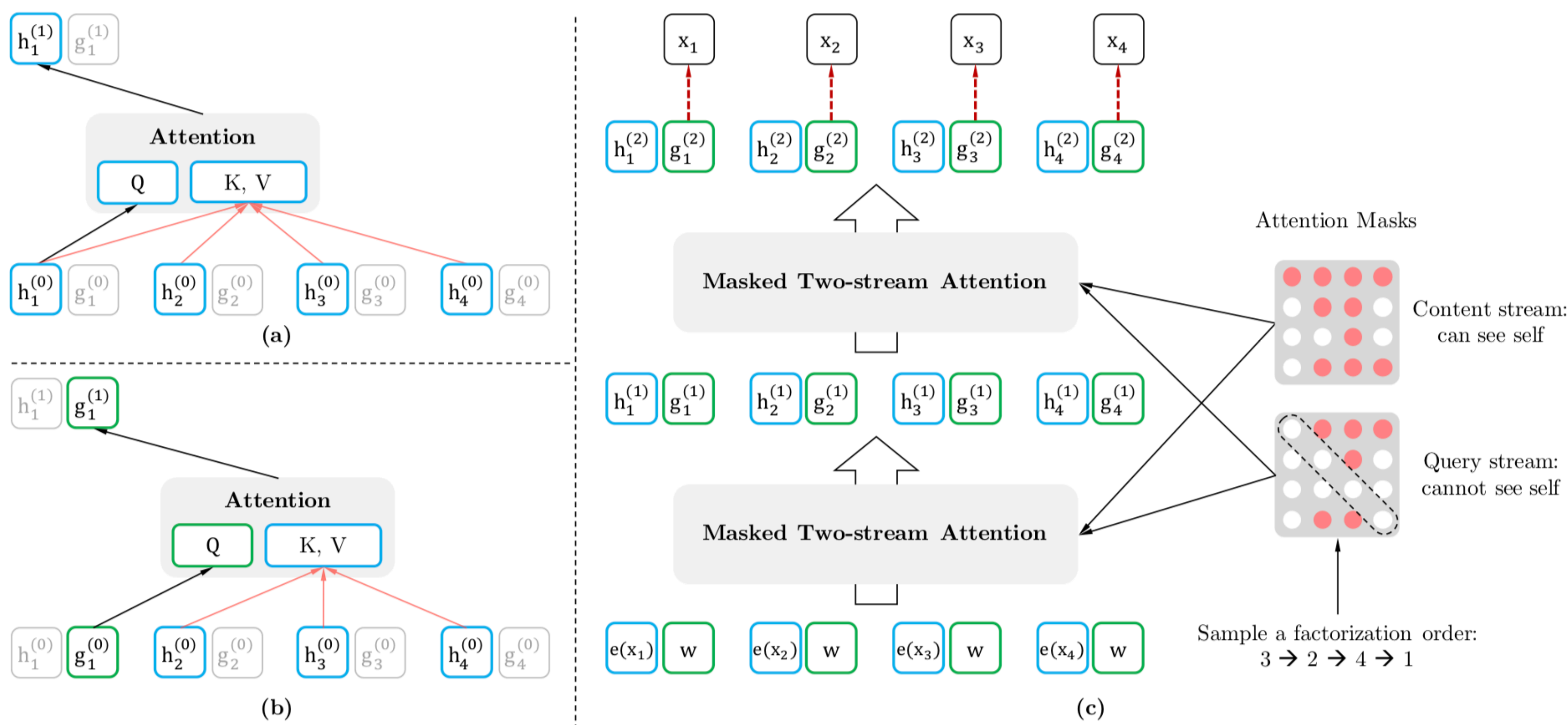
- **Model:** Same as BERT
- **Objective:** Sample words from language model, try to discriminate which words are sampled



- **Data:** Same as BERT, or XL-Net (next) for large models
- **Result:** Training much more efficient!

Permutation-based Auto-regressive Model + Long Context (XL-Net) (Yang et al. 2019)

- **Model:** Same as BERT, but include longer context
- **Objective:** Predict words in order, but different order every time



- **Data:** 39B tokens from Books, Wikipedia and Web

Compact Pre-trained Models

- Large models are expensive, can we make them smaller?
- **ALBERT (Lan et al. 2019):** Smaller embeddings, and parameter sharing across all layers
- **DistilBERT (Sanh et al. 2019):** Train a model to match the distribution of regular BERT

Which Method is Better?

Which Model?

- Wieting et al. (2015) find that simple word averaging is more robust out-of-domain
- Devlin et al. (2018) compare unidirectional and bi-directional transformer, but no comparison to LSTM like ELMo (for performance reasons?)
- Yang et al. (2019) have ablation where similar data to BERT is used and improvements are shown

Which Training Objective?

- Zhang and Bowman (2018) control for training data, and find that bi-directional LM seems better than MT encoder
- Devlin et al. (2018) find next-sentence prediction objective good compliment to LM objective, but Liu et al. (2019) find not

Which Data?

- Zhang and Bowman (2018) find that more data is probably better, but results preliminary.
- Yang et al. (2019) show some improvements by adding much more data from web, but not 100% consistent.
- Data with context is probably essential.

Questions?