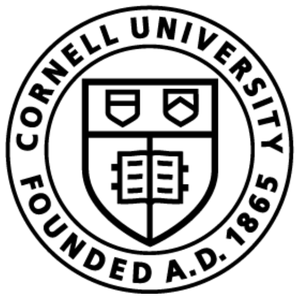
Lecture 9: Encoder-only Transformers + Evaluation



Cornell Bowers CIS
Computer Science

Claire Cardie, Tanya Goyal

CS 4740 (and crosslists): Introduction to Natural Language Processing

Announcements

- HW2 grades released.
  - The deadline for requesting regrades is Apr 10, 2026, 11.59 p.m.

- HW3 Milestone submission due April 13, 11.59 p.m.
- HW3 Final submission due April 20, 11.59 p.m.

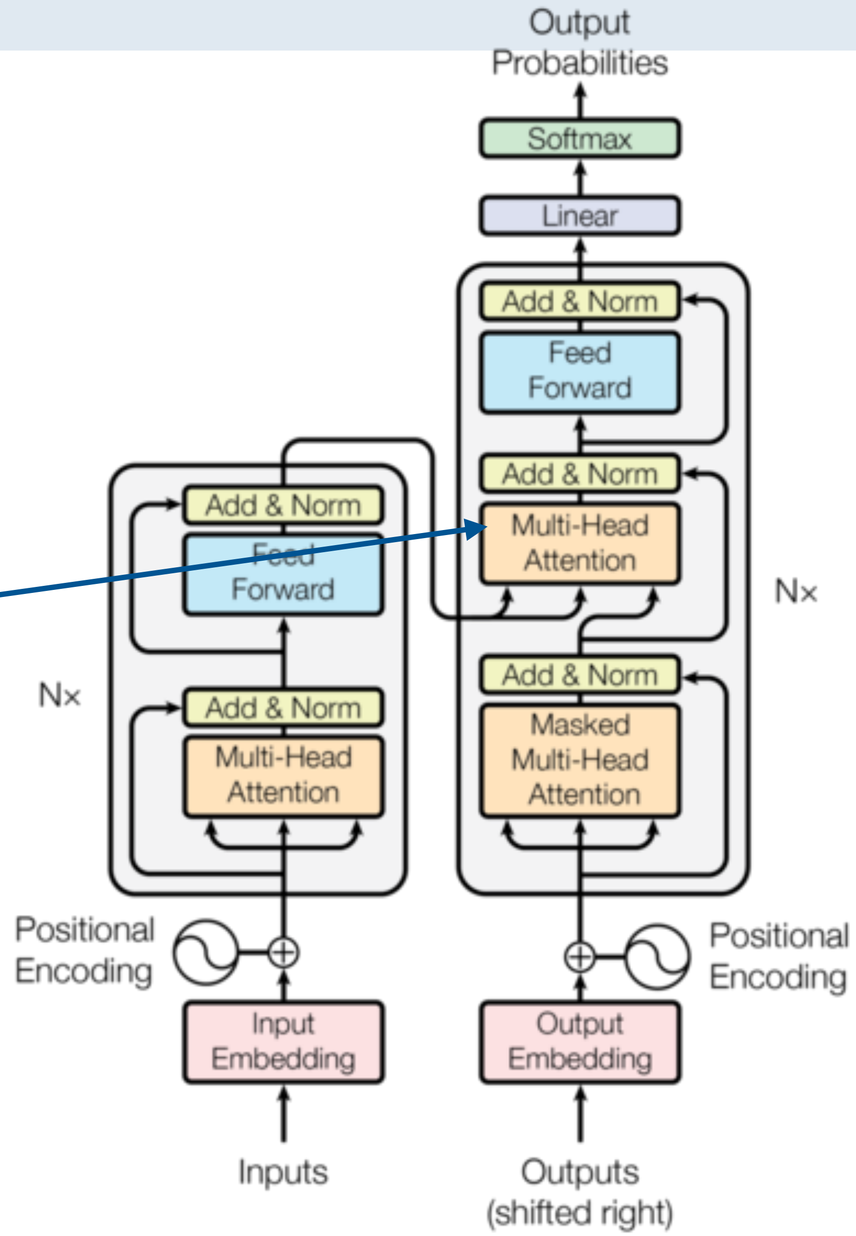
Recap: Encoder-Decoder Architecture

- What are the key differences between the transformer-based encoder and decoder architectures?

Encoder-Decoder Architecture

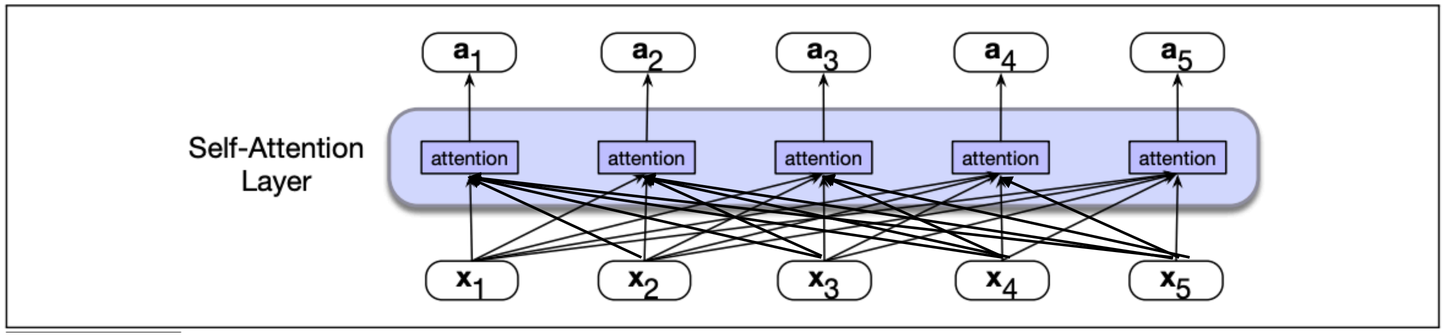
- Decoder has a cross attention layer to the encoder output layer

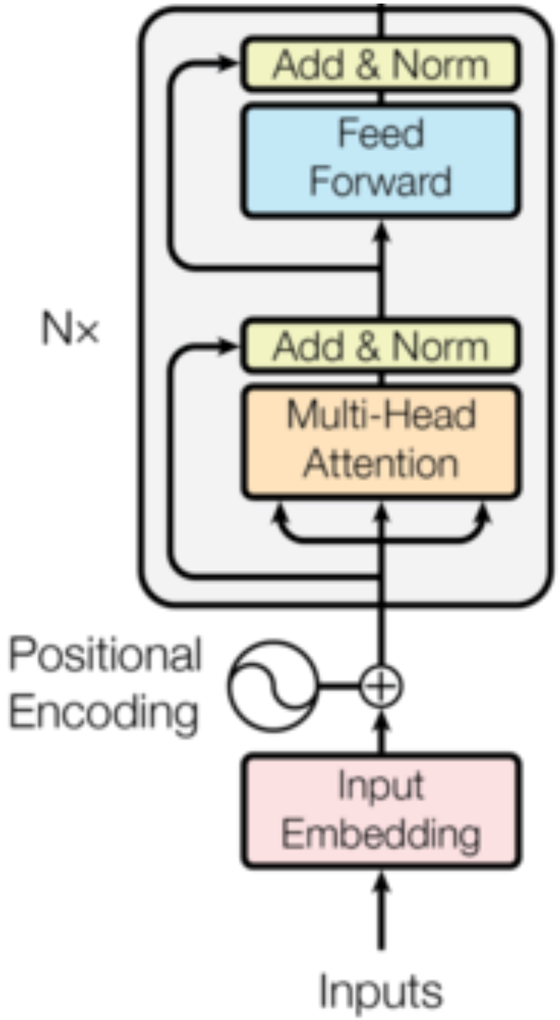
- Multi-head attention from decoder states at each layer to **output of the last layer** of the encoder.



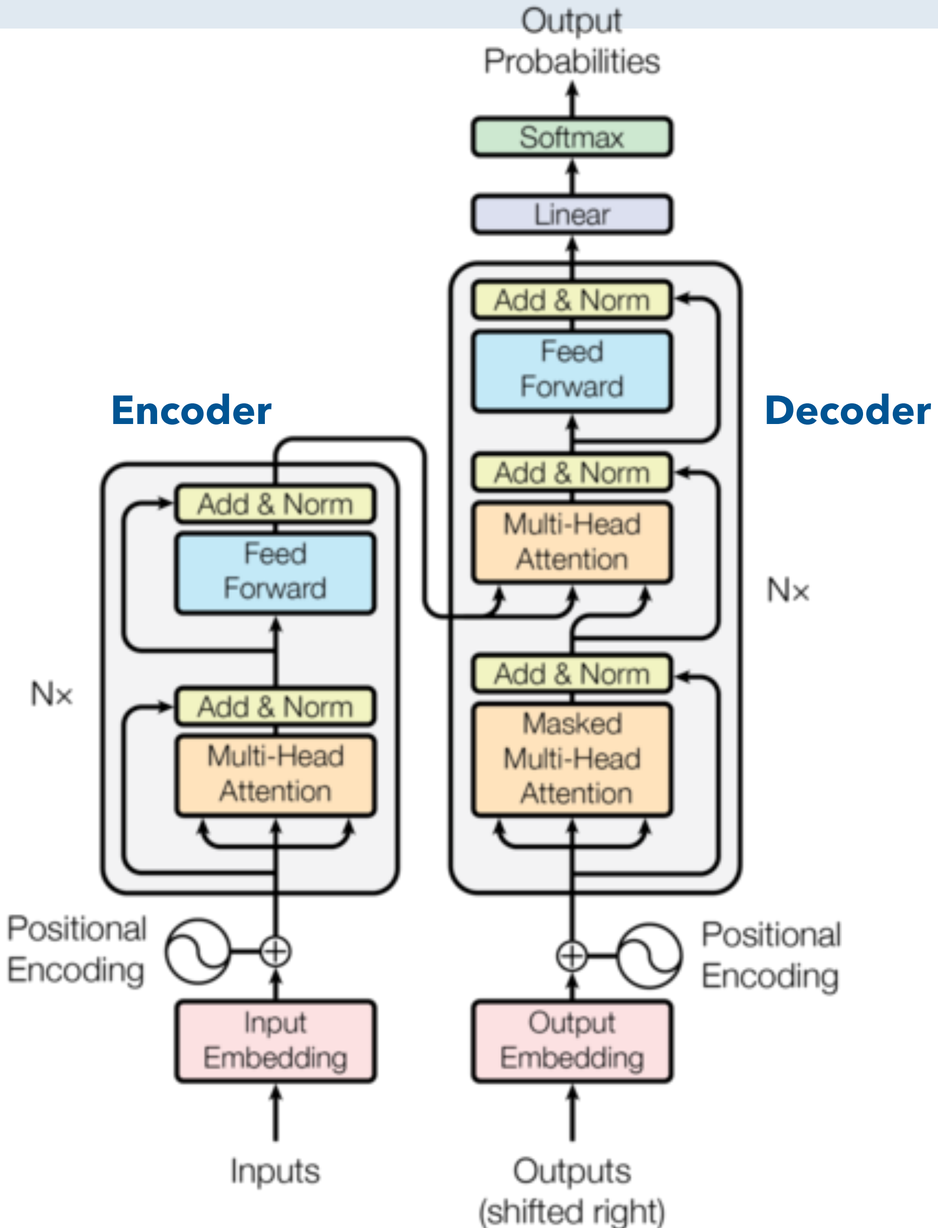
Encoder-Decoder Architecture

- Encoder architecture is similar to decoder.
- Encoder attention is **not** causal.
  - All tokens attend to all other tokens.

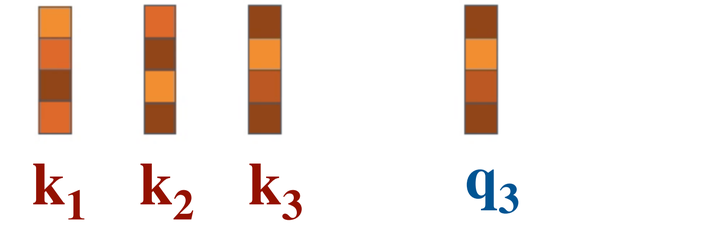




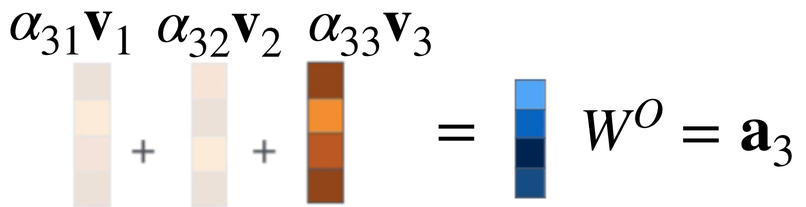
Encoder-Decoder Architecture



- Recall:



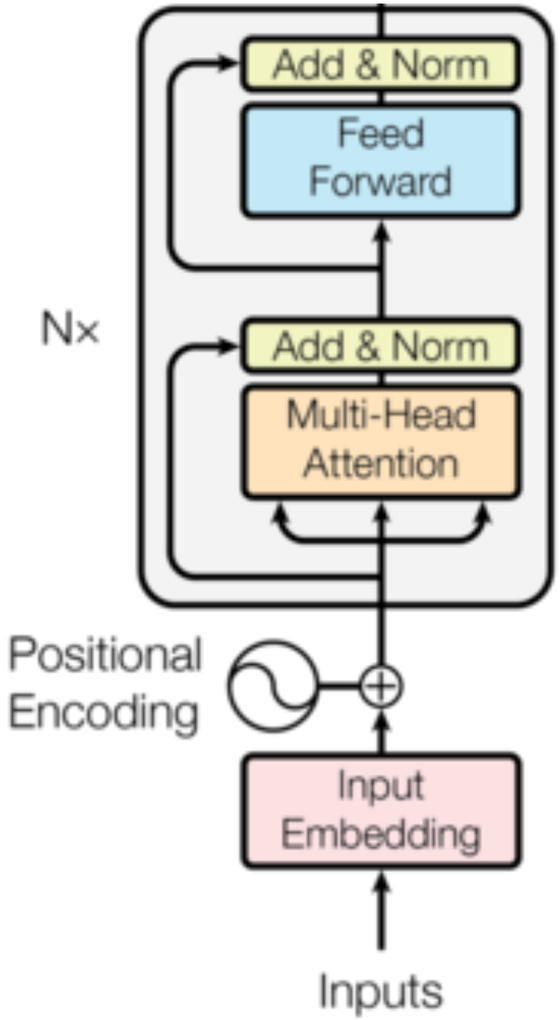
0.02	0.02	0.96



- What is the source of keys, queries, values in attention from decoder to encoder?
  - Queries <— derived from the decoder input.
  - Keys, Values <— derived from encoder outputs.

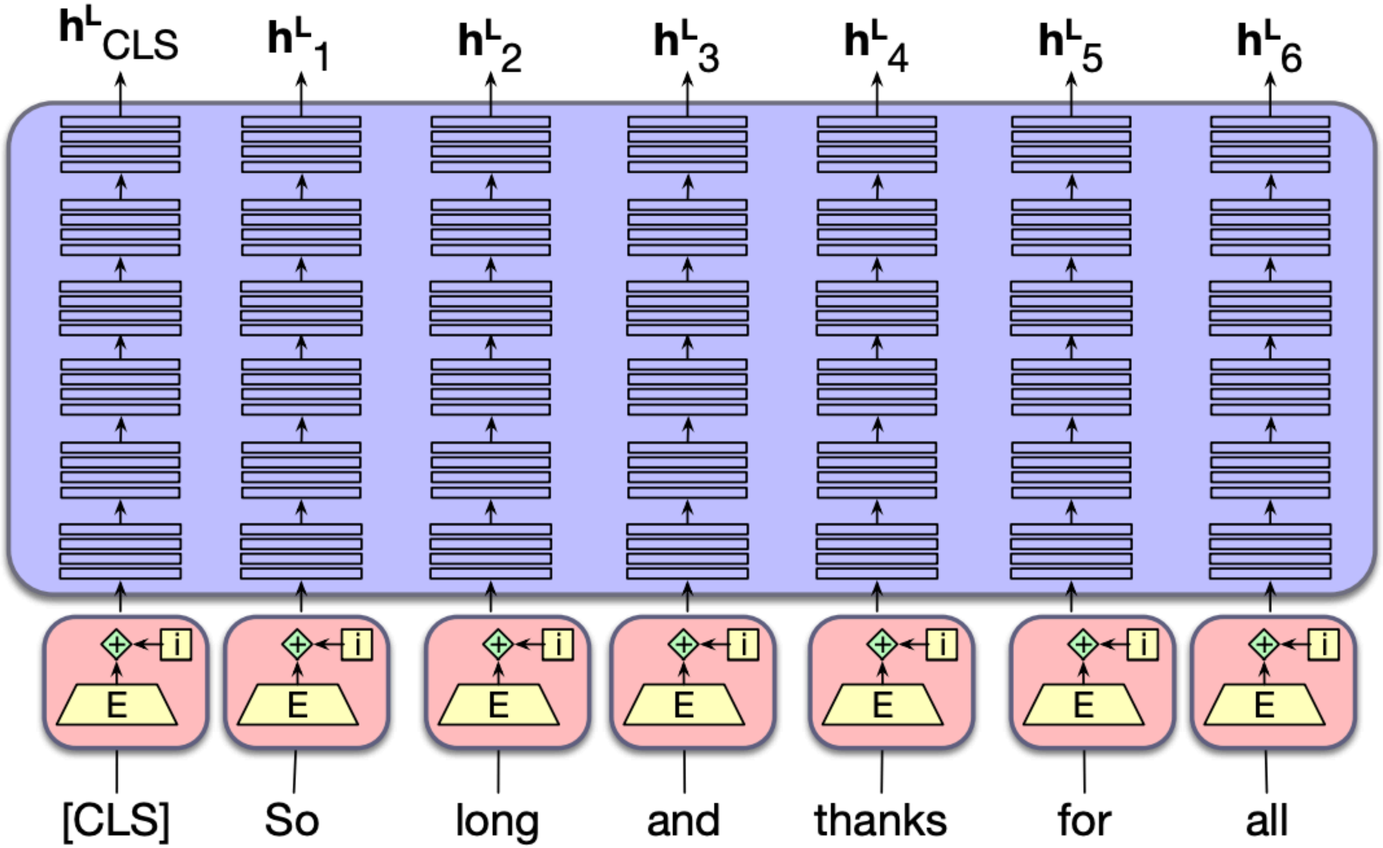
Encoder-Decoder Architecture

- At each time step, the encoder output h_t can be viewed as a "contextual" representation of the input word w_t.
- The inputs are raw word vectors (non-contextual), outputs as contextual word vectors.



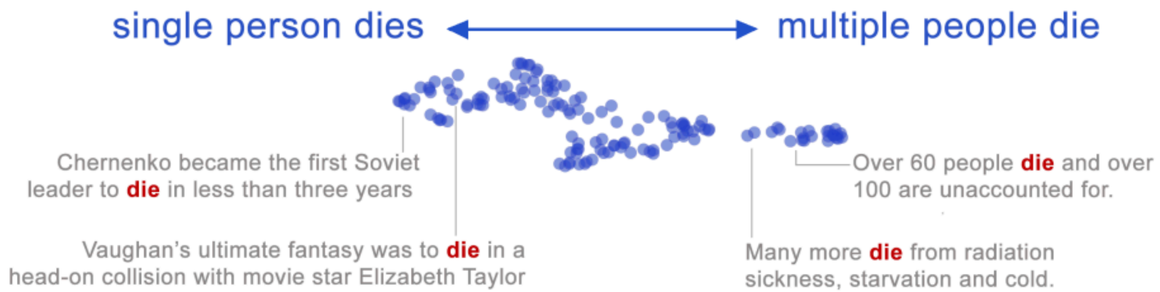
Encoder-only Transformers

- What is this useful for?



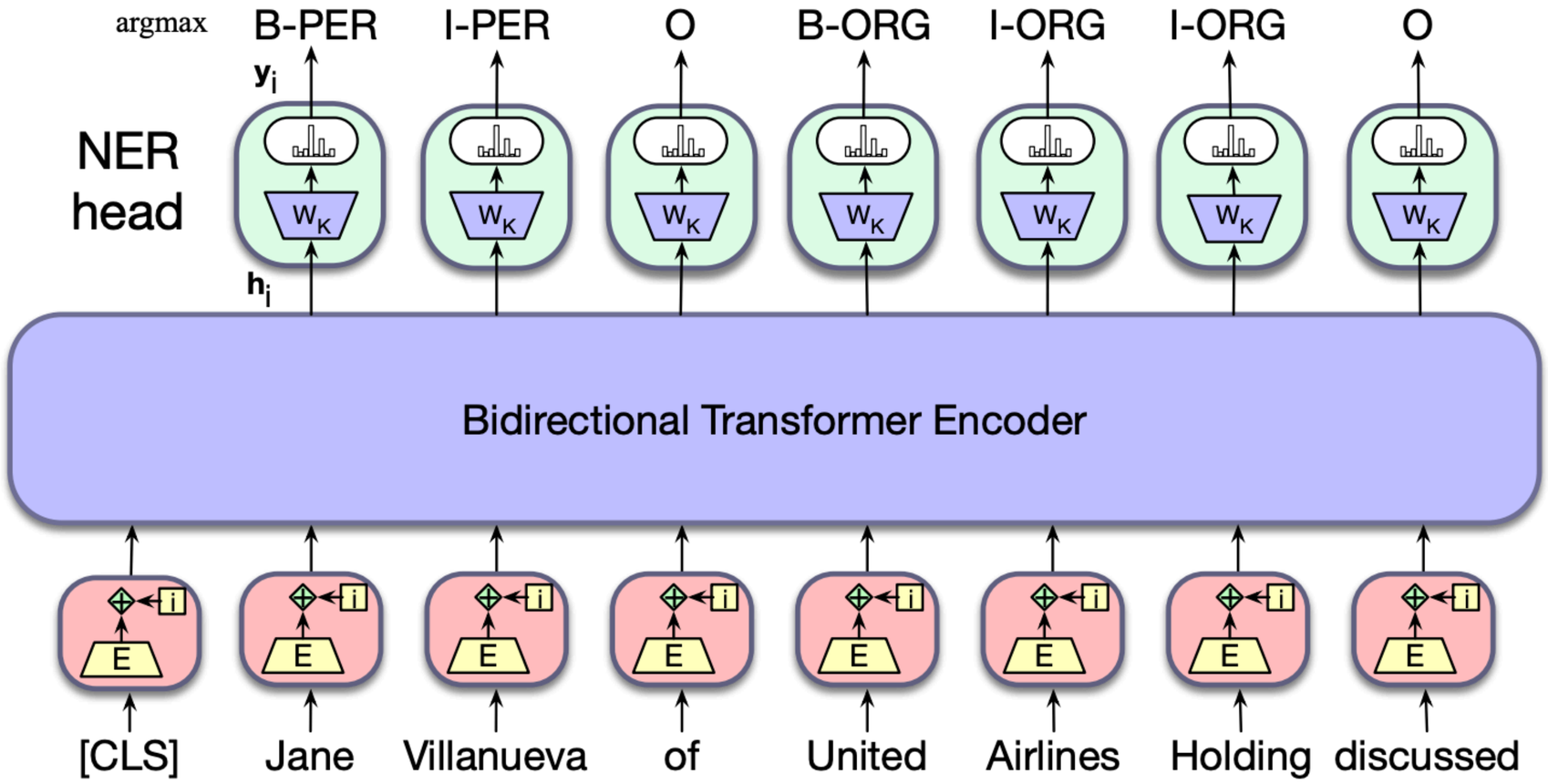
Word sense disambiguation

- A sense (or word sense) is a discrete representation of one aspect of the meaning of a word



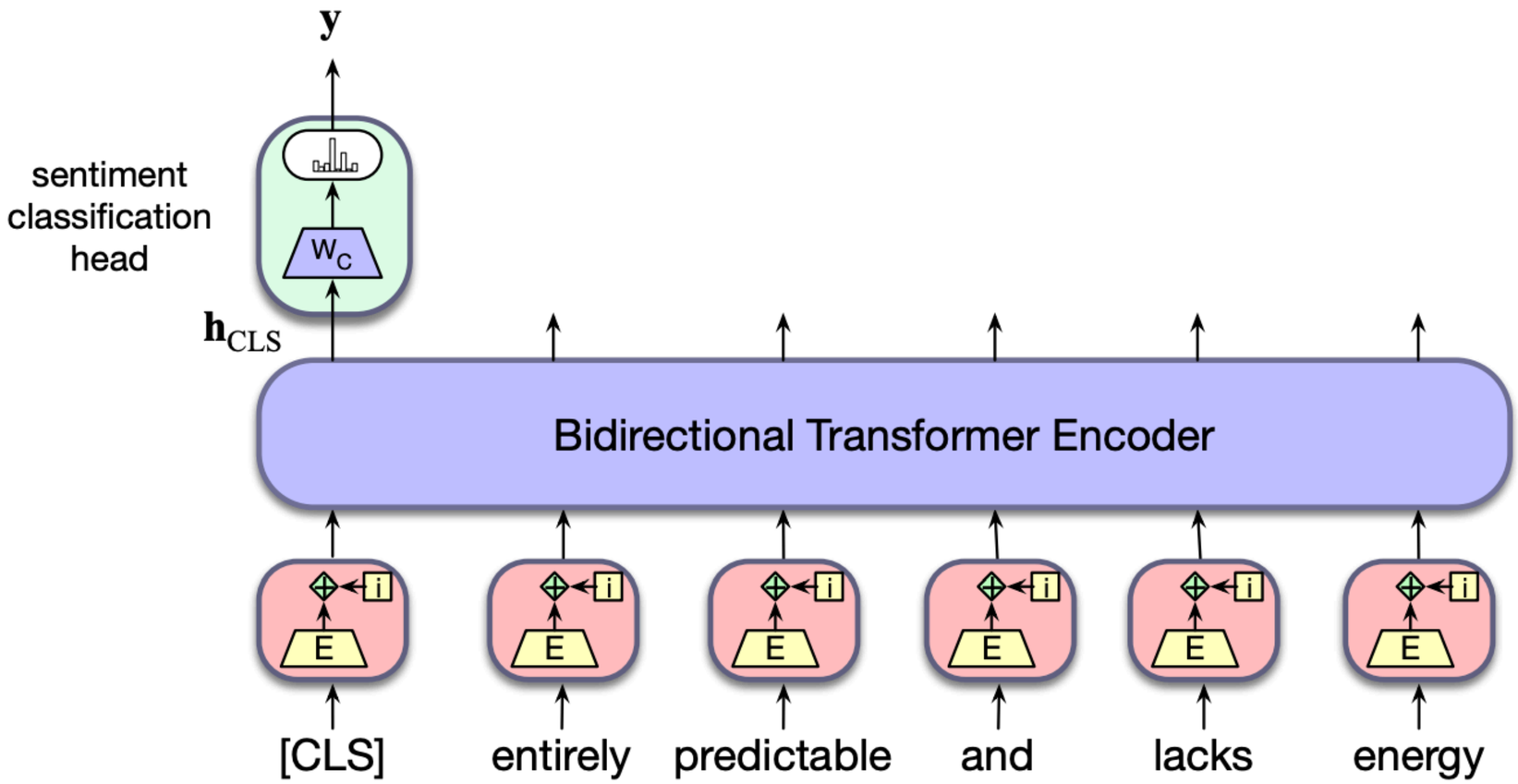
Encoder-only Architecture

- Classification?
  - Sequence Tagging



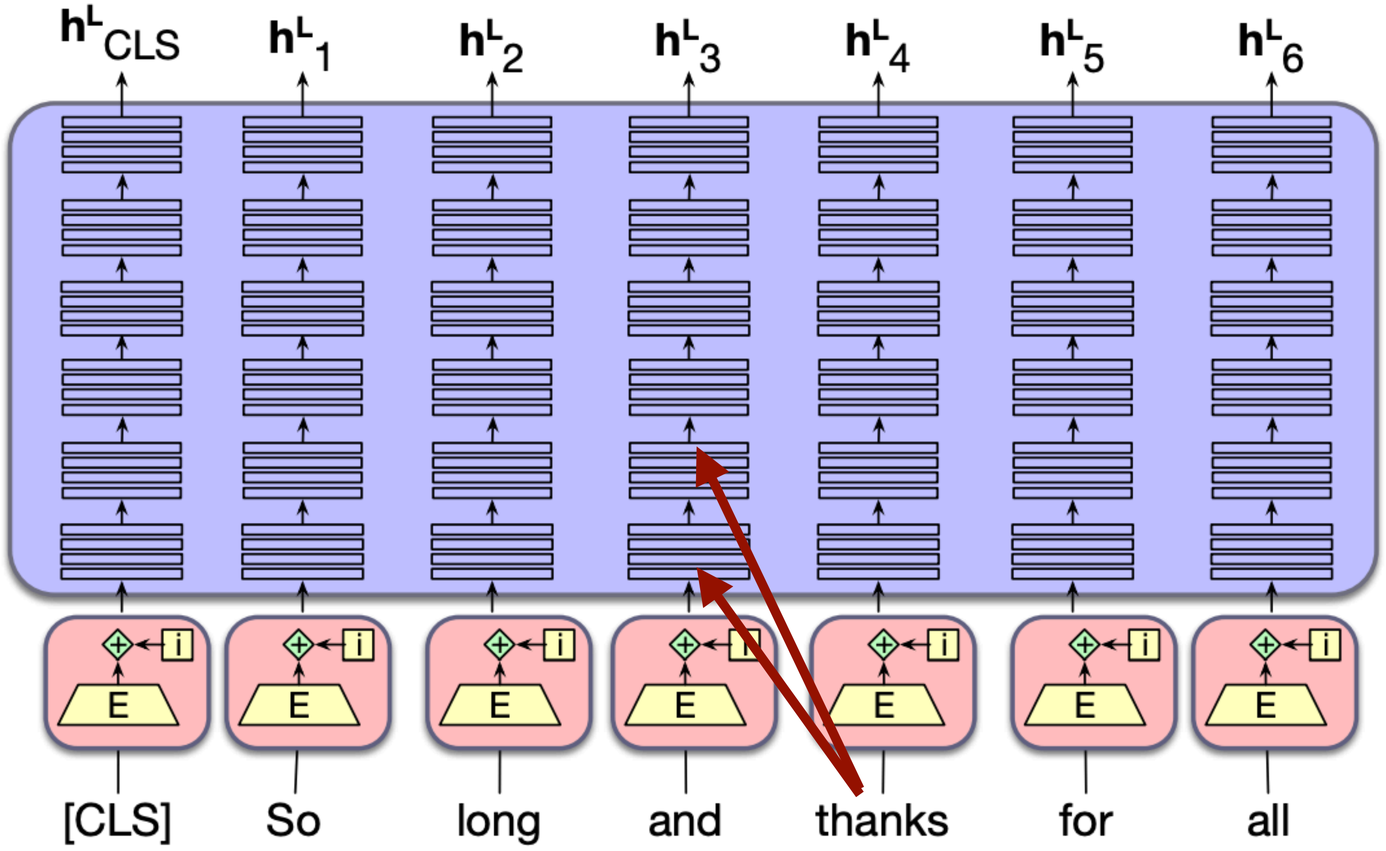
Encoder-only Architecture

- Classification?
  - Sequence-level classification



Encoder-only architectures

- How do we train encoder only architectures?
- Can we use the same next-token prediction task to train encoder models?

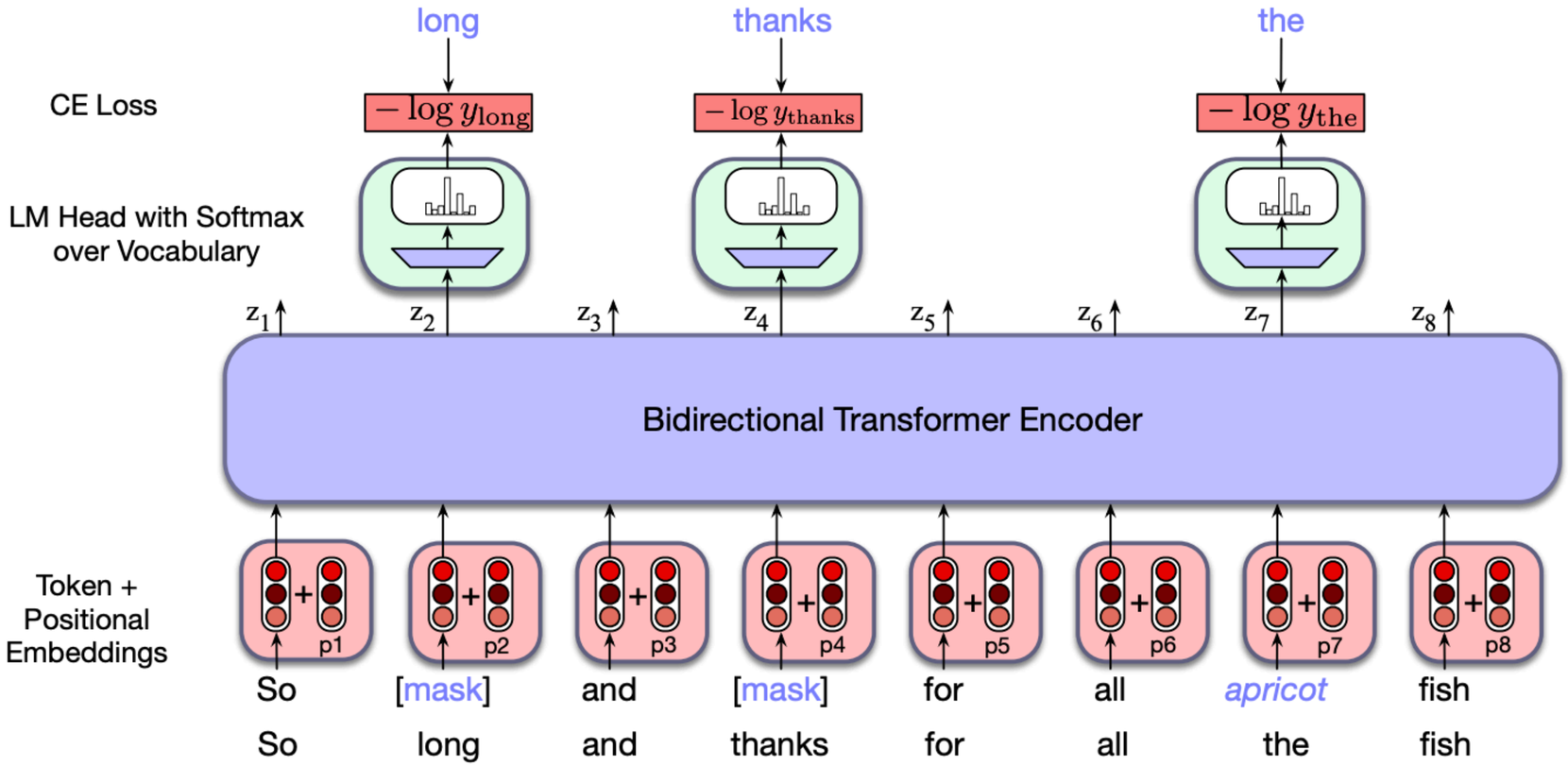


Training objective - Encoder-only models

- BERT (Devlin et al. 2018) introduced the idea of **masked** language modeling.
  - **B**idirectional **E**ncoder **R**epresentations from **T**ransformer
- Key idea:
  - Randomly replace some tokens (15% in the original paper) with [MASK] tokens.
  - Learn to predict these words.

Masked Language Modeling Objective

- Randomly mask 15% of the input tokens.
- Use the output at the masked token's position to predict the actual token.
- $L_{\text{MLM}} = -\log P(x_i \mid \mathbf{h}_i^L)$



Additional Objective

- **Bert** model uses another loss term (in addition to MLM)

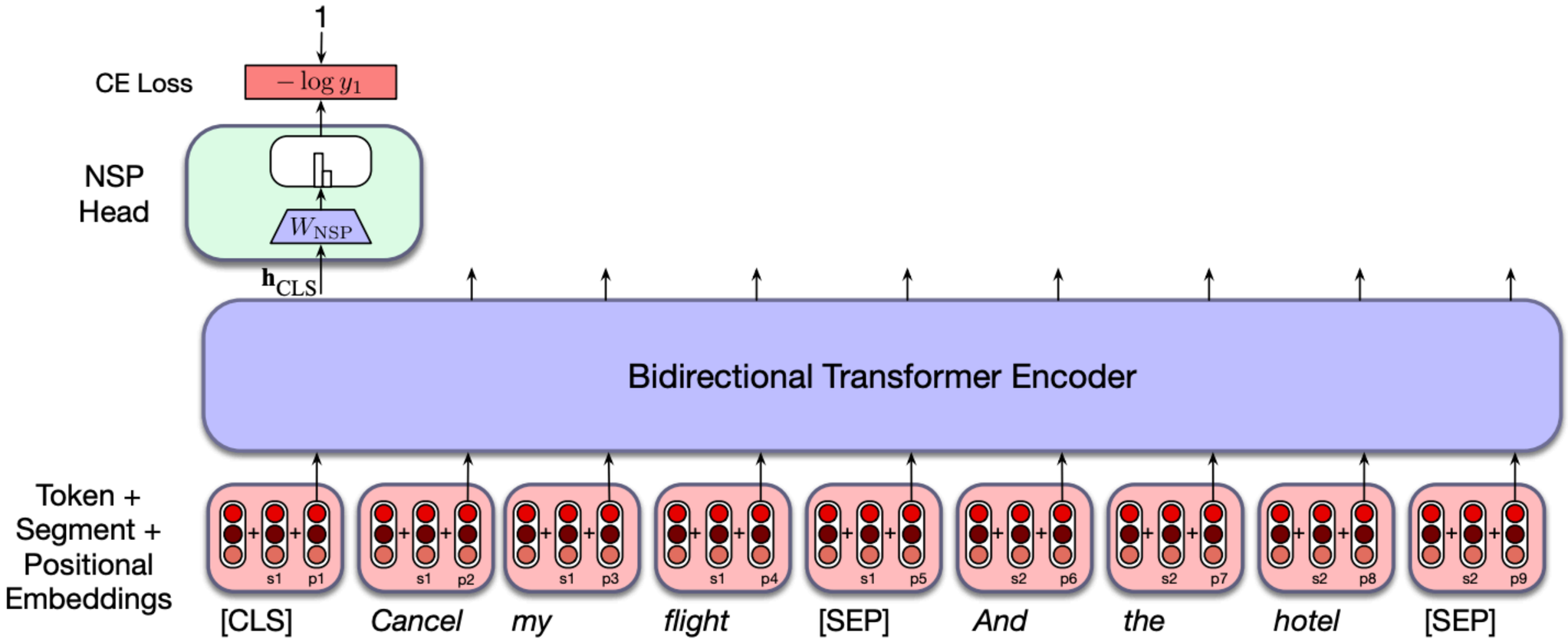
- Next sentence prediction ~~recognition~~ (NSP) (like a binarization of the generation version of long-span language modeling)
  - Given 2 sentences, predict whether Sentence B is the actual sentence that follows Sentence A, or is a random sentence

Sentence A = `The man went to the store.`
Sentence B = `He bought a gallon of milk.`
Label = `IsNextSentence`

Sentence A = `The man went to the store.`
Sentence B = `Penguins are flightless.`
Label = `NotNextSentence`

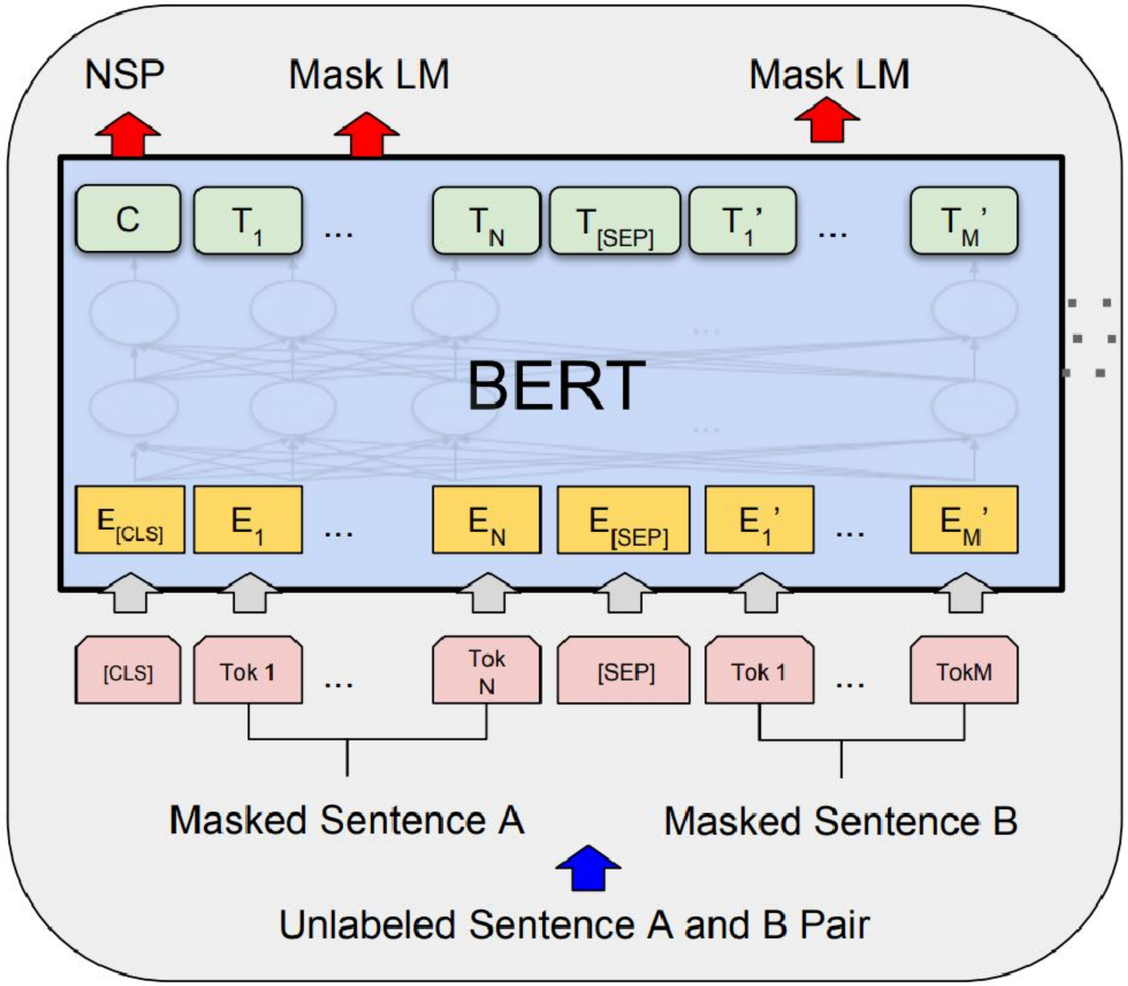
Training Encoder-only models

- Additional “sentence embedding” to distinguish between 1st and 2nd sentence.
- Special token [CLS] and [SEP] added to the vocabulary.



Pre-Training a BERT model

- Training combines the two objectives: Masked Language Modeling + Next-sentence prediction..
- Training on raw text corpora is called "**pre-training**".
  - Model learns general language capabilities but not specialized to any task.



Pre-training

BERT training regime

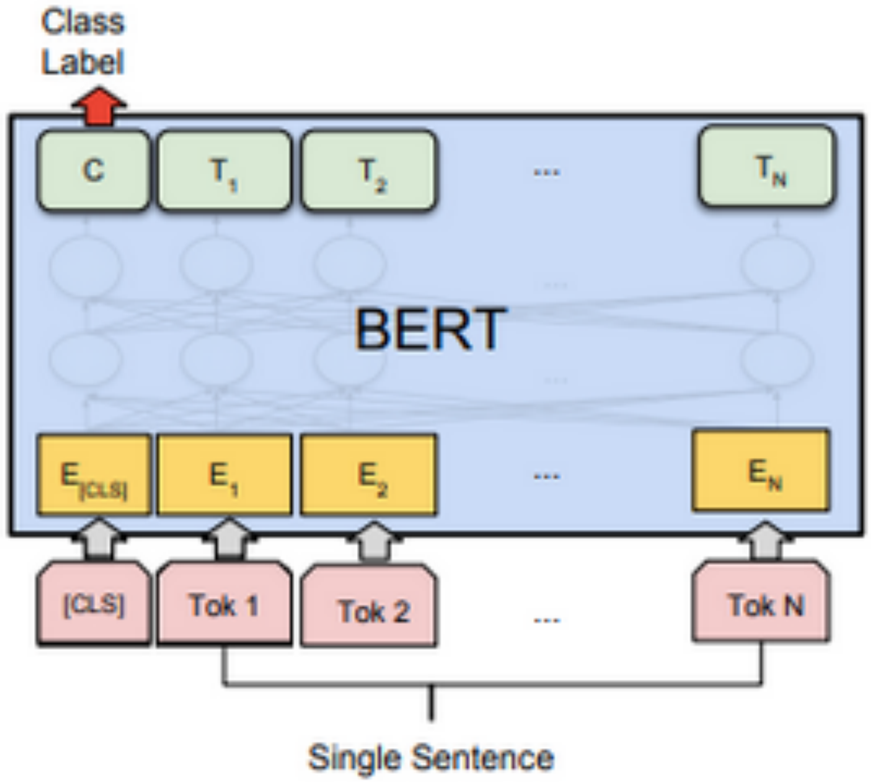
- BERTBASE (L=12, H=768, A=12, Total Parameters=110M) and BERTLARGE (L=24, H=1024, A=16, Total Parameters=340M)
- 15% of the tokens were replaced by the [MASK] token for MLM.
- 50/50 split between the two labels Y = {0,1} for next sentence pred task.
- 512 sequence length

Fine-tuning a BERT model

- BERT model provides contextual representations of words. How do we use it for any task? e.g. sequence tagging, sentiment classification, etc.?
- We can **fine-tune** BERT for different downstream tasks.
    - Suppose we have some labeled data for our downstream task, e.g. NER tagging.
    - Starting from the pre-trained model, we can re-train BERT on these data samples.

Fine-tuning BERT for different tasks

- Single sentence classification e.g. sentiment classification

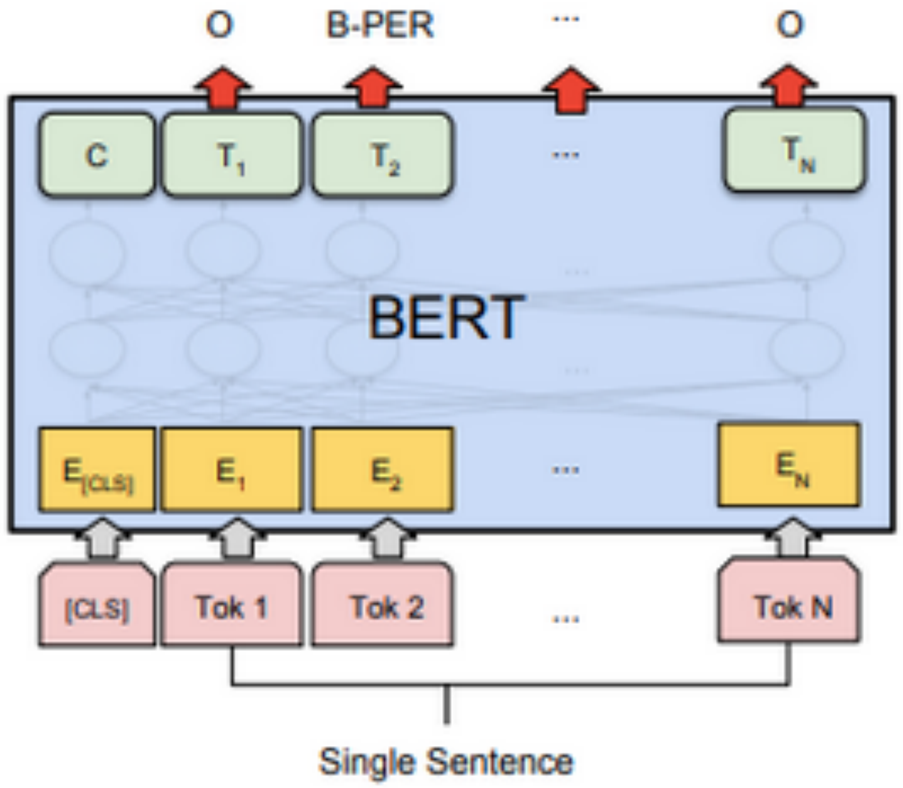


(b) Single Sentence Classification Tasks: SST-2, CoLA

- Add a classification head (linear layer) on the [CLS] token's output representation.
- Hidden outputs for the rest of the tokens discarded.
- We can train the whole model (gradient updates for ALL model parameters), or just the classification head, or some other subset.
- What is the size of the classification head?

Fine-tuning BERT for different tasks

- Token-level classification e.g. NER

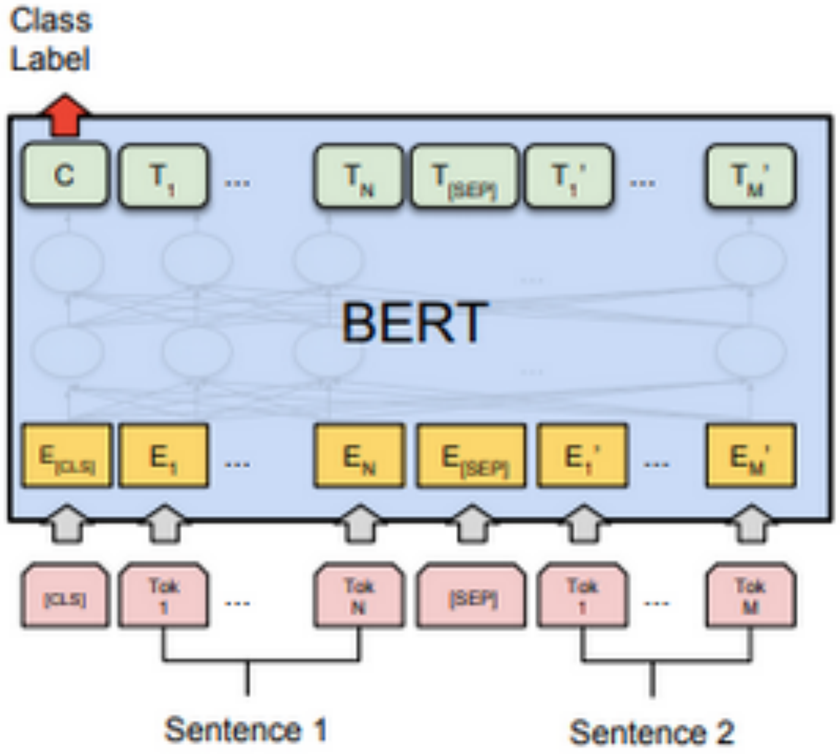


(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

- Discard the [CLS] token's output representation.
- Classification head on the rest of tokens.
- We can train the whole model (gradient updates for ALL model parameters), or just the classification head, or some other subset.

Fine-tuning BERT for different tasks

- Sentence-pair classification tasks, e.g. paraphrase identification.



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

- Add a classification head (linear layer) on the [CLS] token's output representation.
- We can train the whole model (gradient updates for ALL model parameters), or just the classification head, or some other subset.
- What is the size of the classification head?

BERT performance

- Original paper fine-tuned the pre-trained BERT model on various tasks.

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Table 2: SQuAD 1.1 results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.

Takeaways

- Encoders provide “contextual” embeddings for words.
- Masked Language Modeling objective allows us to train encoders without leaking the output.
- Pre-training using self-supervised data, followed by fine-tuning was a powerful idea.

Mini-break + Questions

- Which architectures (encoder-only, encoder-decoder, decoder-only) are appropriate for the following tasks. Choose all that can be used.
  - Machine Translation.
  - Language Modeling
  - Sentiment Classification for a paragraph.
  - Question-Answering

Evaluation of Text Classification

- This is straightforward.

Evaluation of Text Generation

- How do we evaluate a summarization/machine translation model?
  - Given two trained models, A and B, how do we determine which is better?

- Solution 1: **Extrinsic Evaluation —** Evaluate on the downstream application that the model is trained for.
  - For each question/input in the test set, **generate** outputs using A and B. Score the outputs for each input, compute aggregate scores.

- Solution 2: **Intrinsic Evaluation** — **perplexity** which evaluates how well the model fits the (test) data distribution. (next class)

How good is a generated output?

- American Jennifer Stewart says she was devastated to learn that Etihad Airways lost her most important baggage: her 2-year-old pet cat, Felix. Stewart said that she booked Felix on their Etihad Airways flight from the United Arab Emirates to Chicago's O'Hare Airport on April 1. […]
- Generated Output: A Chicago woman is searching for her cat after it went missing while being transported on an Etihad flight.

- Problem: no single answer!
  - An Etihad Airways passenger was devastated after the airline lost her cat Felix.
  - Etihad Airlines loses a passenger's 2-year old pet enroute to Chicago from UAE.

How good is a generated summary?

- How good is a summary?
- Many possible summaries are acceptable.

- Evaluation metrics:
  - Subjective evaluation by humans.
    - Costly, slow, inconsistent.
  - Automatic evaluation using models

Automatic Evaluation Metrics

- Goal: a model / computer program that computes the quality of generations. Rankings based on this should agree with expert humans.
- Advantages:
  - Low cost
  - Optimizable (easy to evaluate intermediate models, hill-climb)
  - Consistent

Reference-based automatic metrics

- Lots of options of automatic evaluation metrics in literature.
- We will focus on **reference-based metrics** in this lecture. Setting:
    - For each input x
    - Given: **Output Summary**
    - Given: **Human reference / gold summary(s)**
- **Metric:** Compute similarity between them!

- Reference-based metrics can be used for all generation-based tasks.
- Need a test set with (input, gold output) pairs.
- Examples of metrics: ROUGE, BLEU, BertScore, MoverScore, etc.

Lexical Overlap based metrics

- ROUGE: Recall-Oriented Understudy for Gisting Evaluation
- BLEU: Bilingual Evaluation Understudy

- Both follow the same basic idea of using n-gram overlap between generated and reference outputs to compute similarity.

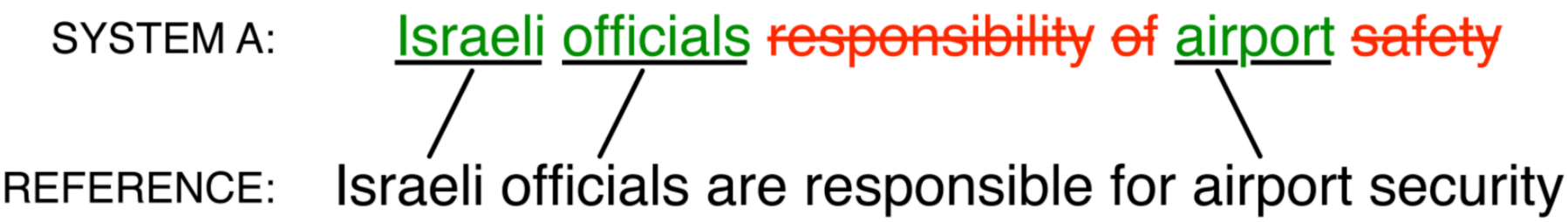
Precision and Recall of Words

Precision

Recall

F-measure

Precision and Recall of Words



Precision $\frac{correct}{output\text{-}length}$

Recall $\frac{correct}{reference\text{-}length}$

F-measure $\frac{precision \times recall}{(precision + recall)/2} =$

Precision and Recall of Words

SYSTEM A: Israeli officials ~~responsibility~~ ~~of~~ airport ~~safety~~

REFERENCE: Israeli officials are responsible for airport security

Precision

$$\frac{\textit{correct}}{\textit{output-length}} = \frac{3}{6} = 50\%$$

Recall

$$\frac{\textit{correct}}{\textit{reference-length}} = \frac{3}{7} = 43\%$$

F-measure

$$\frac{\textit{precision} \times \textit{recall}}{(\textit{precision} + \textit{recall})/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$$

Precision and Recall of Words

Metric	System A	System B
precision	50%	100%
recall	43%	100%
f-measure	46%	100%

- Issue: no penalty for re-ordering of words

BLEU: Bilingual Evaluation Understudy

- N-gram overlap between generated output and reference output
  - Originally proposed for machine translation, also used extensively for paraphrasing, etc.
- Computes precision for n-grams of size 1-4 in the generated output, adds a brevity penalty (for very short outputs)

$$\text{BLEU} = \min(1, \frac{\text{output-length}}{\text{reference-length}}) \prod_{i=1}^{4} (\text{precision}_i)^{1/4}$$

- Computed over the entire test corpus.

BLEU examples

SYSTEM A: Israeli officials (2-GRAM MATCH) responsibility of airport (1-GRAM MATCH) safety

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security (2-GRAM MATCH) Israeli officials are responsible (4-GRAM MATCH)

Metric	System A	System B
precision (1gram)		
precision (2gram)		
precision (3gram)		
precision (4gram)		
brevity penalty		
BLEU		

BLEU examples

SYSTEM A: Israeli officials responsibility of airport safety
2-GRAM MATCH 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible
2-GRAM MATCH 4-GRAM MATCH

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

BLEU: Bilingual Evaluation Understudy

- Ideally, multiple reference outputs to account for variability.
  - N-grams may match in any of the references
  - Closest reference output length is used

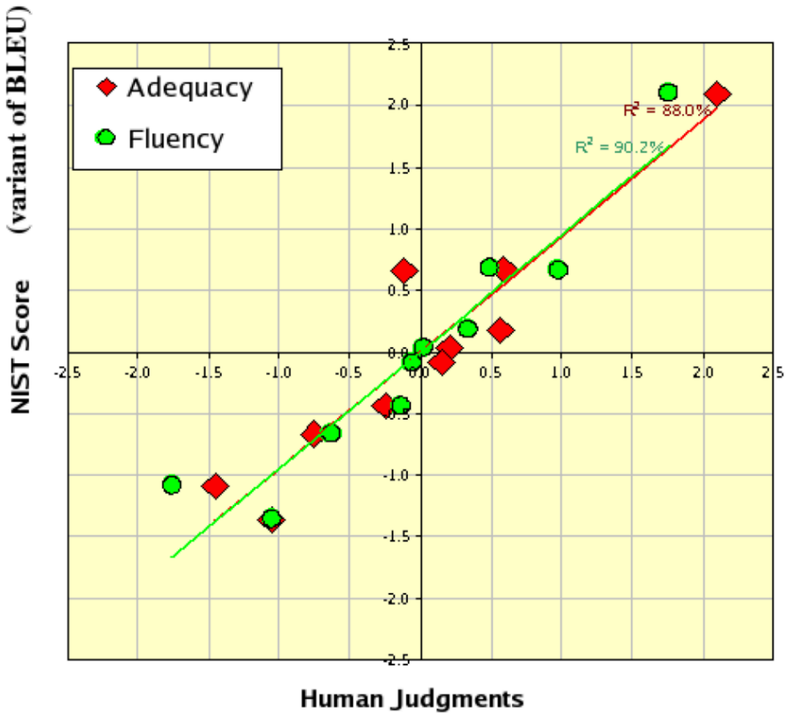
Example

SYSTEM: Israeli officials (2-GRAM MATCH) responsibility of (2-GRAM MATCH) airport (1-GRAM) safety

REFERENCES:
Israeli officials are responsible for airport security
Israel is in charge of the security at this airport
The security work for this airport is the responsibility of the Israel government
Israeli side was in charge of the security of this airport

Is an automatic metric good?

- What we want:
  - The rankings between systems given by a metric should match the ranking given by expert humans
  - BLEU shows high correlation with human judgments (for old systems)



Lexical Overlap-based metrics

- Limitations of lexical overlap based metrics?
  - Depend on strict overlap. Do not account for say synonym replacements.

    Consider : Reference: The movie was amazing.

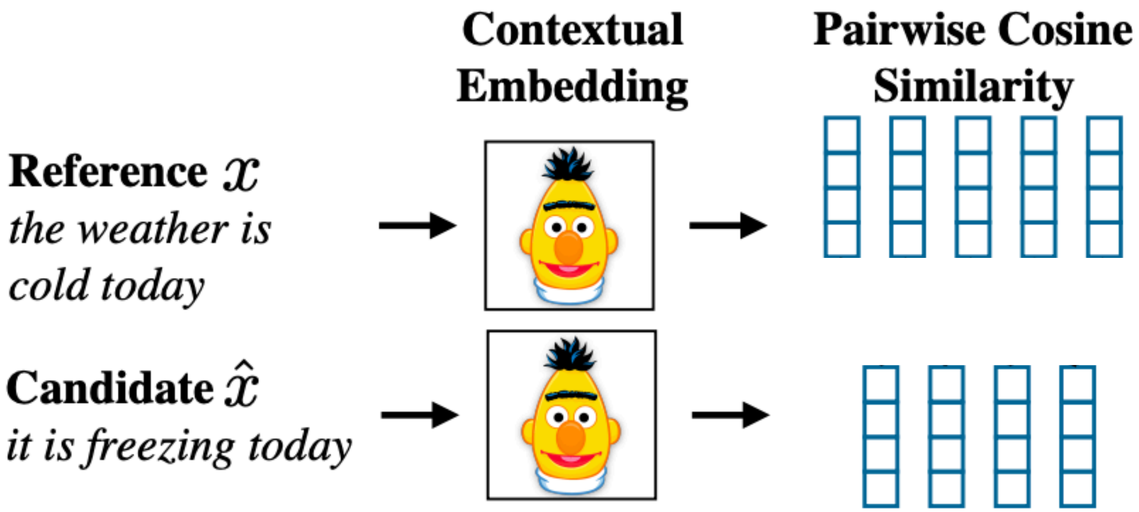
    Generated Output: The film was great.

  - All words are treated equally
  - Need (input, gold output) pairs. This data is difficult to get!

- Lots of variants with the same basic idea (n-gram overlap) proposed: METEOR (uses stemming, lemmatization, and identifies paraphrastic matches), CIDEr (down-weights common n-grams), etc.

Solution: Distributional similarity-based metrics

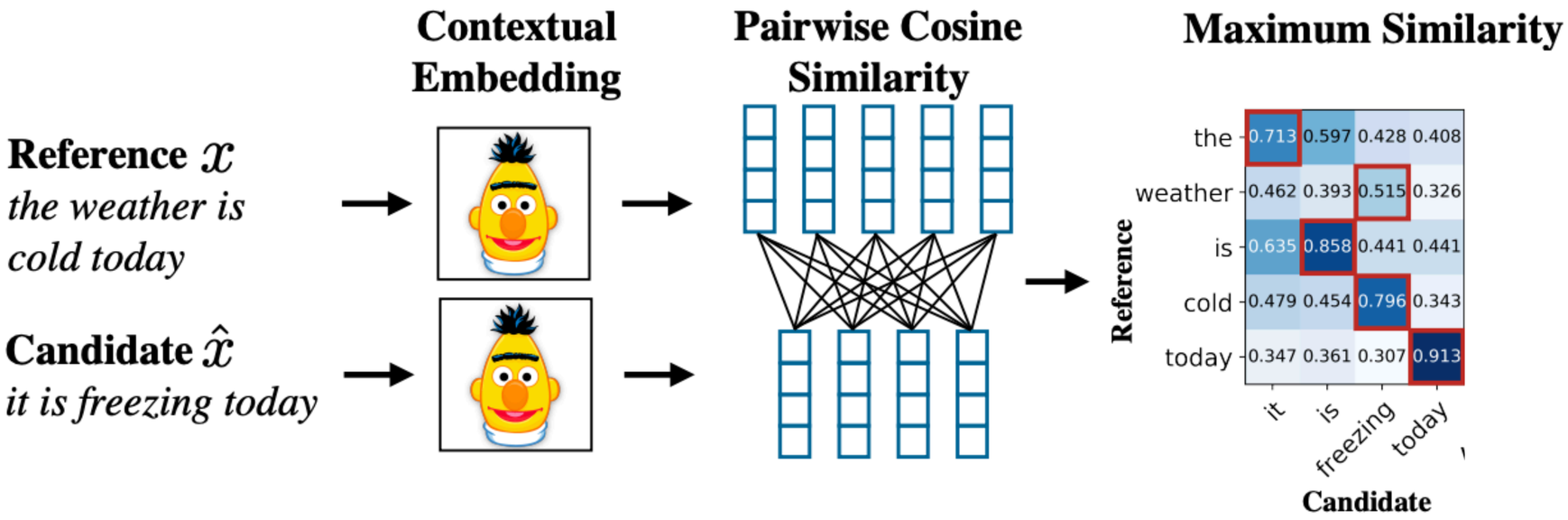
- BERTScore (Zhang et al, ICLR 2020) **Cornell authors!!



- Use BERT (what was this architecture?) to get representations for each word for both the reference and the output candidate

Solution: Distributional similarity-based metrics

- BERTScore (Zhang et al, ICLR 2020) **Cornell authors!!

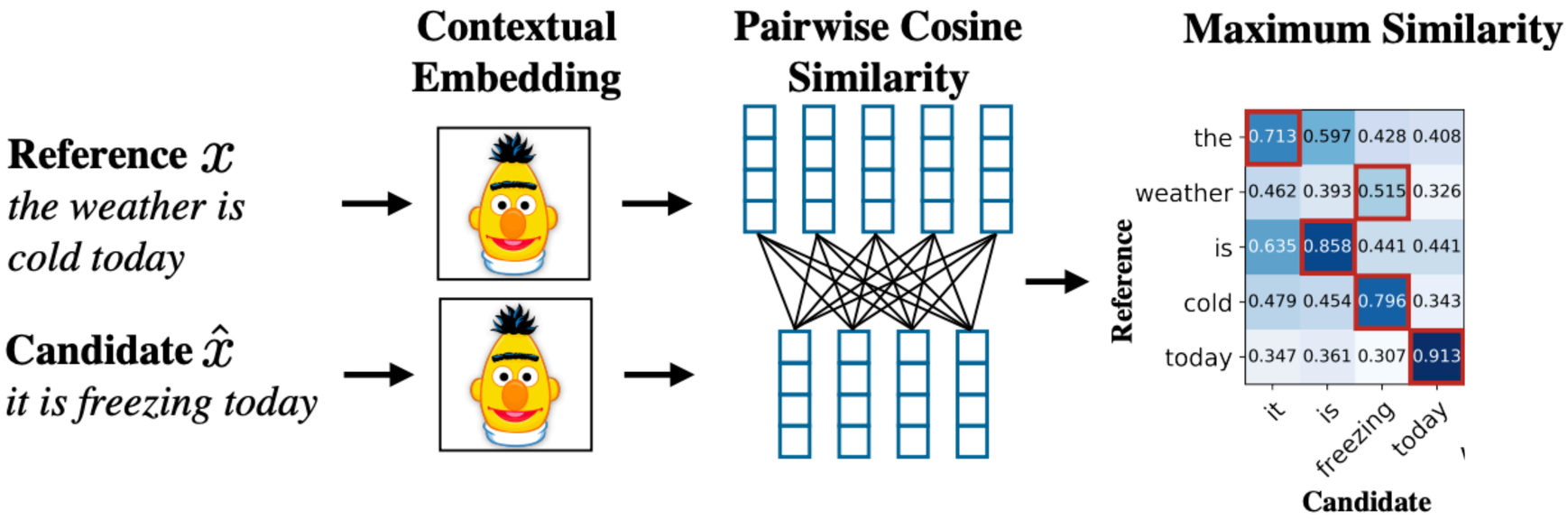


$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

- For each word in the reference, find the closest match in the generated output

Solution: Distributional similarity-based metrics

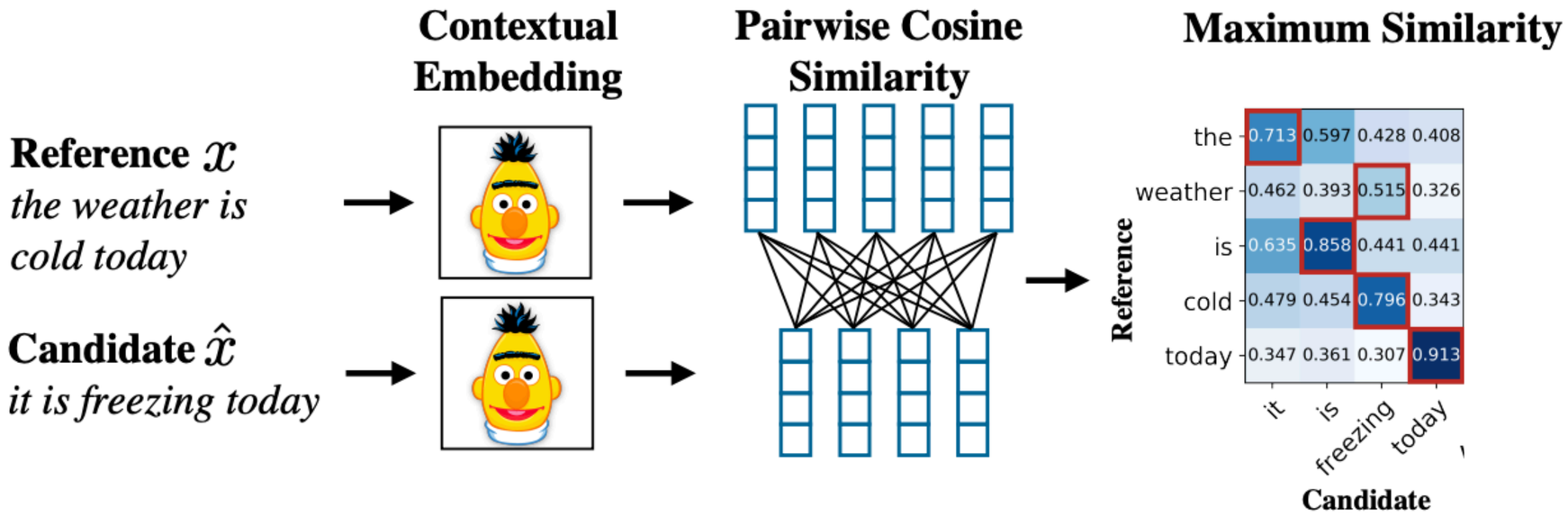
- BERTScore (Zhang et al, ICLR 2020) **Cornell authors!!



$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

- For each word in the generated output, find the closest match in the reference.

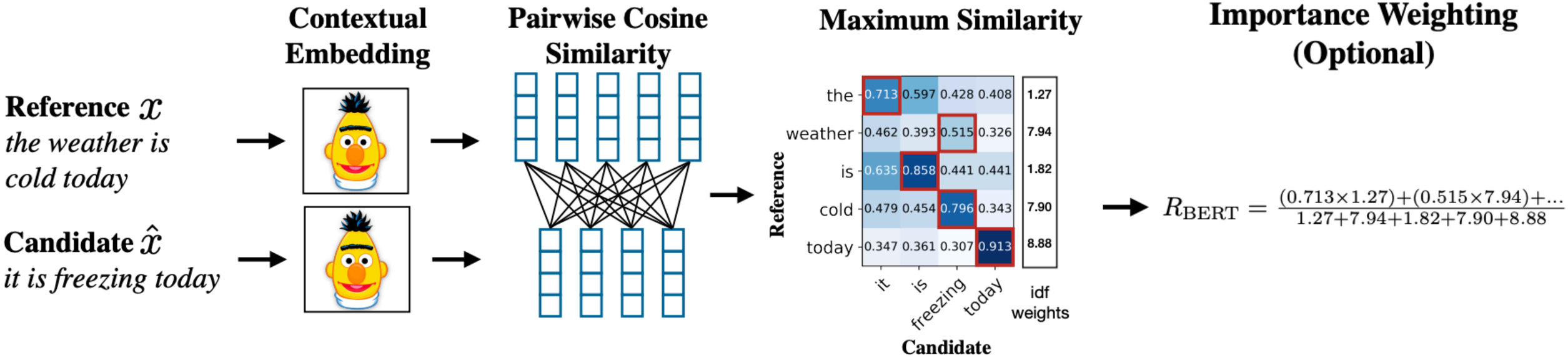
Solution: Distributional similarity-based metrics



$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \qquad P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \qquad F1 = \frac{2 * P * R}{(P + R)}$$

Solution: Distributional similarity-based metrics

- Optional importance weighting of each reference token to compute recall



$$\text{idf}(w) = -\log\left(\frac{1}{M}\sum_{i=1}^{M}\mathbb{I}[w \in x^{(i)}]\right)$$

$$R_{\text{BERT}} = \frac{\sum_{x_i \in x} \text{idf}(x_i)\max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j}{\sum_{x_i \in x} idf(x_i)}$$

Distributional similarity-based metrics

- Q: Why is only “1-gram overlap” used in BERTScore computation?
  - Encoder representations are contextual. Representation of the same token within different n-grams will be different.

Slide Acknowledgements

- Earlier versions of this course offerings including materials from Claire Cardie, Marten van Schijndel, Lillian Lee.