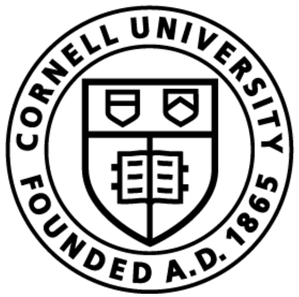
Lecture 10: Training + Evaluation



Cornell Bowers CIS
Computer Science

Claire Cardie, Tanya Goyal

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

Announcements

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

Recap: Pre-training vs Fine-tuning

- Pre-training:
  - Train on a large corpus of text (e.g. internet data).
  - Decoder-only models (GPT2, GPT3, all modern LLMs) trained with the next-token prediction task, encode-only models (BERT, RoBERTA) trained with the masked language modeling task.
  - Models learn general language capabilities, but cannot perform any task.
-

Recap: Pre-training vs Fine-tuning

- Fine-tuning
  - Identify downstream task(s) that are important/relevant. These could be QA, Sentiment Classification, Story Generation, Summarization, etc.
  - Collect (x,y) pairs for this downstream task.
  - Train!!

Recap: How do we evaluate a fine-tuned model?

Basic Premise: Divide our data into train/validation/test corpus.

1. Train the **parameters** of the model on the train set/corpus.
2. Choose model + hyperparameters that perform the best on the **validation** set.
3. Test the model's performance on the unseen **test** set.

What metric do we use for evaluation in step 2 and 3?

Recap: Evaluation for Classification Tasks

- Output y belongs to a fixed set of labels.
- Compute metrics Acc/Prec/Recall/etc.

Today

- How do we evaluate a pre-trained model?
  - We'll focus on Decoder-only models. How do we know pre-trained checkpoint A is better than pre-trained checkpoint B.
- How do we evaluation a model for generation tasks? (Either pre-trained or fine-tuned)

How do we evaluate a pre-trained model?

- What is the task that the model is trained on? Next-token prediction.
- The downstream use cases are different — e.g. qa, generation, etc.
- Our goal: **Intrinsic Evaluation:** how to we evaluate the the model's performance on the objective it is trained for.
- Downstream evaluations (e.g. evaluating QA performance, Summarization performance, etc.) are **Extrinsic Evaluations.**

How do we evaluate a pre-trained model?

- We pre-train a model on the language modeling (LM) task.
- **Intrinsic Evaluation:** Measure quality of the model independent of the downstream application.
- Standard measure: **Perplexity**
  - Intuition: the better model is the one that has a tighter fit to the test data....one that predicts the test data

Intuition behind perplexity.

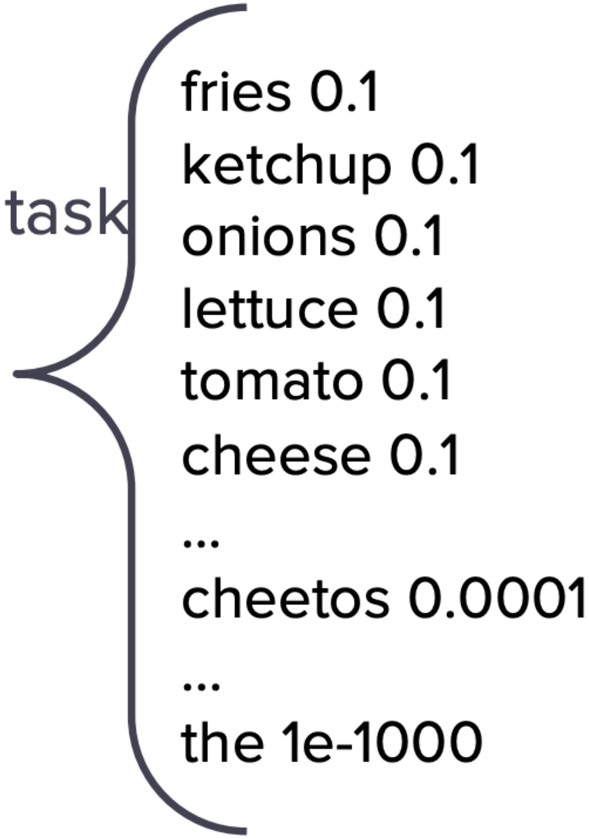
Shannon game

- Similar to the predict next word task

I always order burgers with _____

The cat sat on the _____

NLP is _____



- A better model of a text is one which assigns a higher probability to the word that actually occurs

Perplexity Definition.

- For a test set $W = w_1 w_2 \ldots w_N$

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-1/N}$$

$$= \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i \mid w_1 \ldots w_{i-1})}}$$

- Perplexity is the **inverse probability of the test set, normalized by the #words.**
- **Minimizing perplexity is the same as maximizing probability.**

Perplexity as branching factor

- **How to think about perplexity intuitively?**
- Let's suppose a sentence consists of random digits.
- What is the probability of this sequence according to a model that assigns a uniform prob to all digits?
  - $PP(W) = P(w_1 w_2 \ldots w_N)^{-1/N} = ((1/10)^N)^{-1/N} = 10$
  - A better model reduces this branching factor!

Perplexity

- **In practice, we use logprobs**:

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-1/N}$$

$$= \exp(-\frac{1}{N} \sum_{i=1}^{N} \log P(w_i \mid w_1 \ldots w_{i-1}))$$

- Perplexity is the exponentiated average (token-level) negative log likelihood.

Perplexity

Lower perplexity = better model

- **Training 38 million words, test 1.5 million words, Wall Street Journal**

N-gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109

- Neural LMs will be much lower (=better!)

Today

- How do we evaluate a pre-trained model?
  - We'll focus on Decoder-only models. How do we know pre-trained checkpoint A is better than pre-trained checkpoint B.
- How do we evaluation a model for generation tasks? (Either pre-trained or fine-tuned)

What are generation tasks?

- Story generation

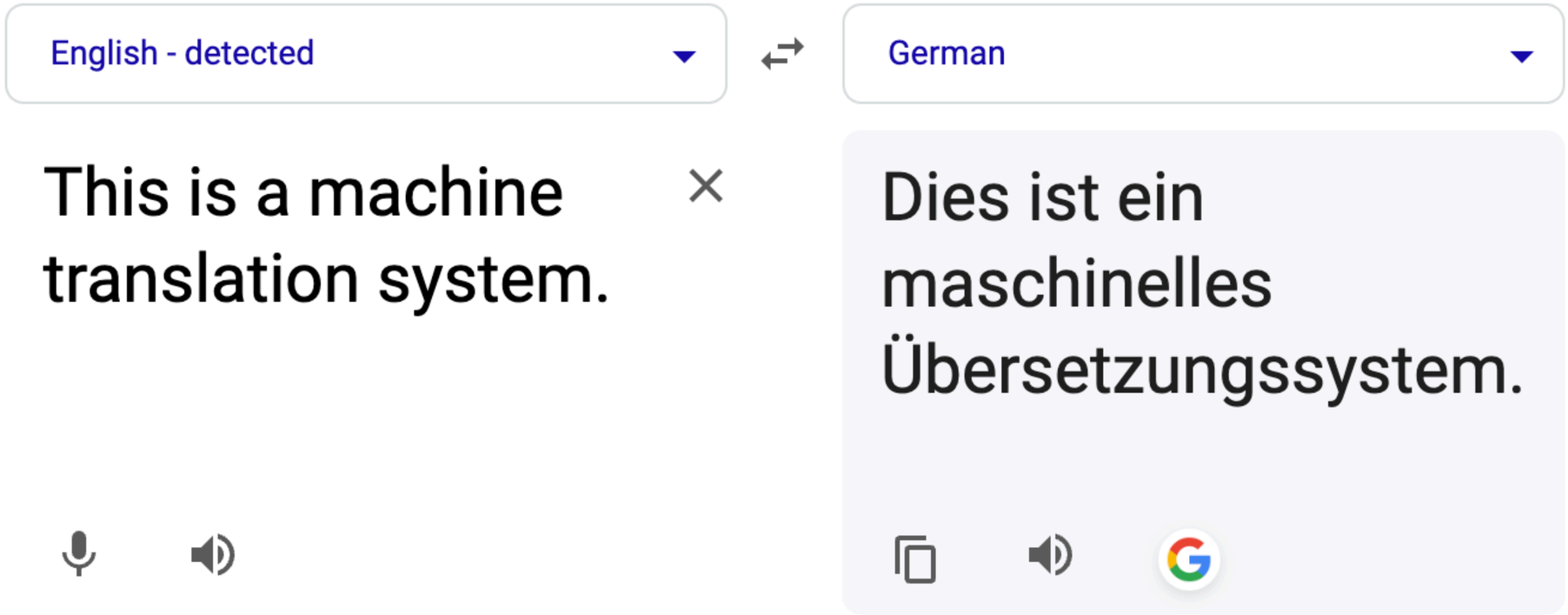
Write a story about an alien who wants to return to his home planet.

OpenAI's ChatGPT

Once upon a time, in a galaxy far, far away, there was an alien named Zor. Zor was a curious and adventurous creature […] As the years passed, Zor grew more and more homesick […]

What are generation tasks?

- Machine Translation



What are generation tasks?

- Paraphrase generation:

> Why are so many Quora users posting questions that are readily answered on Google?

> Why do people ask Quora questions which can be answered easily by Google?

- Summarization:

> 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 […]

> A Chicago woman is searching for her cat after it went missing while being transported on an Etihad flight.

+ many more

- Question Answering

Why don't adults roll off the bed?

Great question! Adults usually *don't* roll off the bed for a few reasons—most of them are automatic and subconscious:

1. **Body Awareness (Proprioception):** Adults develop a strong sense of their body position in space. Even while asleep, your brain keeps some awareness of where you are on the bed.
2. **Sleep Habits:** Over time, adults get used to sleeping in a particular way. They typically move less during sleep compared to kids, which lowers the chances of rolling around too much.

- Anything that can be cast as a seq-to-seq task is a generation task!

Recap: Decoding / generation using a LM

- Given an input prefix x and a language model, how do we "generate"?
- Remember: LM places a probability distribution over next tokens:

$$P(w \mid x)$$

- How do we generate using this language model?

while stopping_condition

sample $\hat{w} \sim P(w \mid x)$

append prefix $x \leftarrow x \oplus \hat{w}$

What can be our stopping condition?

Recap: Decoding / generation using a LM

$\hat{w} = None$

while $\hat{w} \neq < /s >$

sample $\hat{w} \sim P(w \mid x)$

append prefix $x \leftarrow x \oplus \hat{w}$

- Which of the following models can we use for generation?

  A) Encoder-decoder

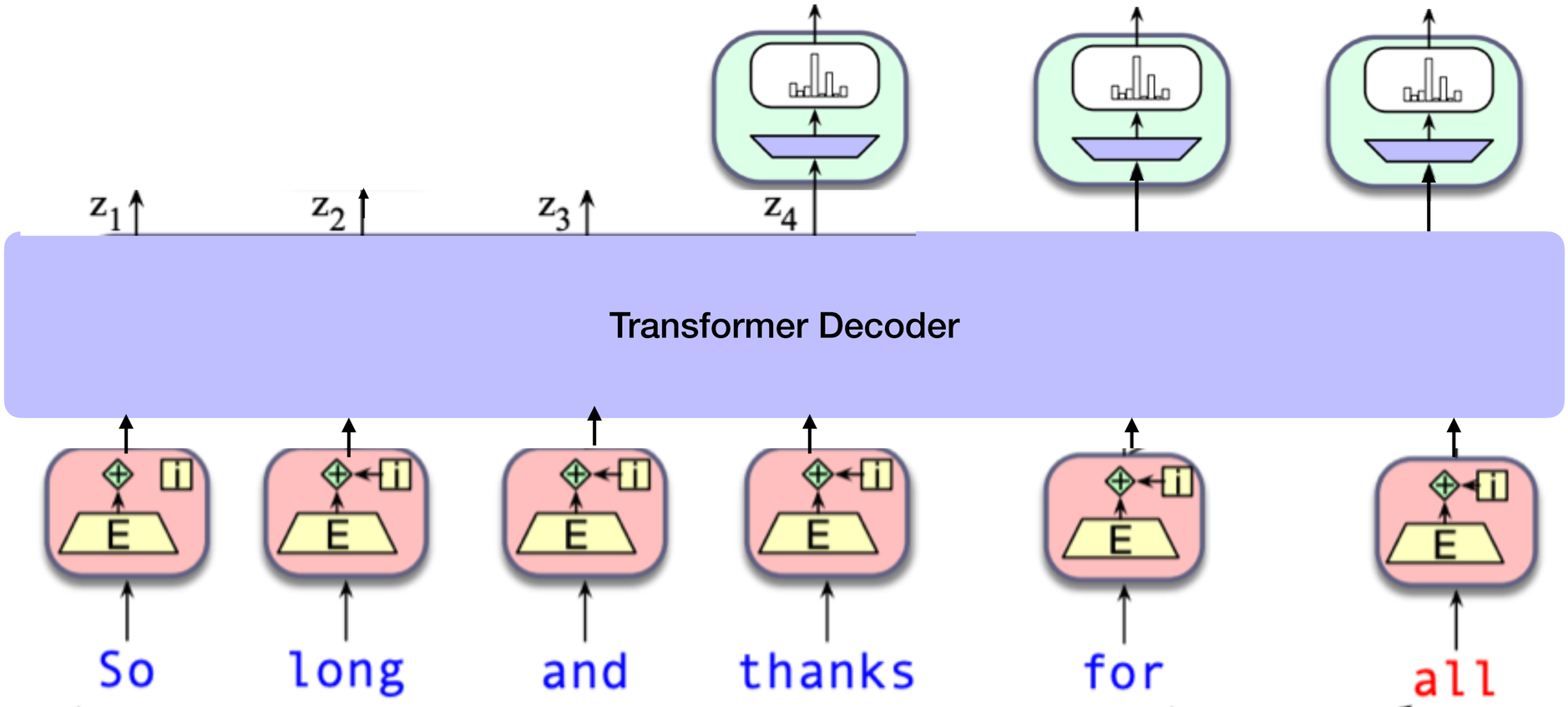
  B) Encoder

  C) Decoder-only

Recap: Decoding / generation using a LM

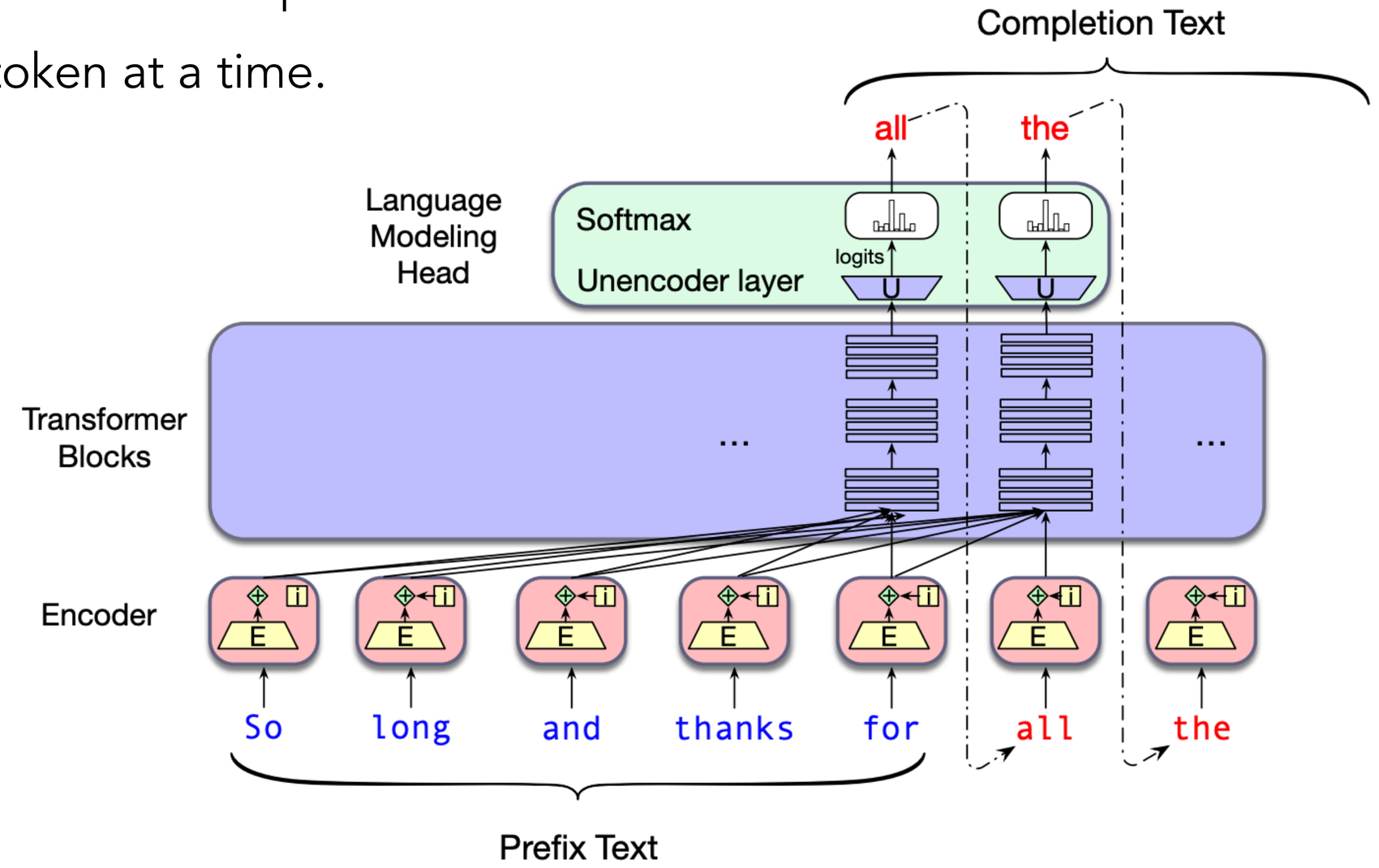
- Given input x, do a forward pass.
- Generate one token at a time.

$$\hat{w} \sim P(w \mid \textit{So long and thanks})$$



Recap: Decoding / generation using a LM

- Given input x, do a forward pass.
- Generate one token at a time.



Decoding / generation using a LM

$\hat{w} = None$

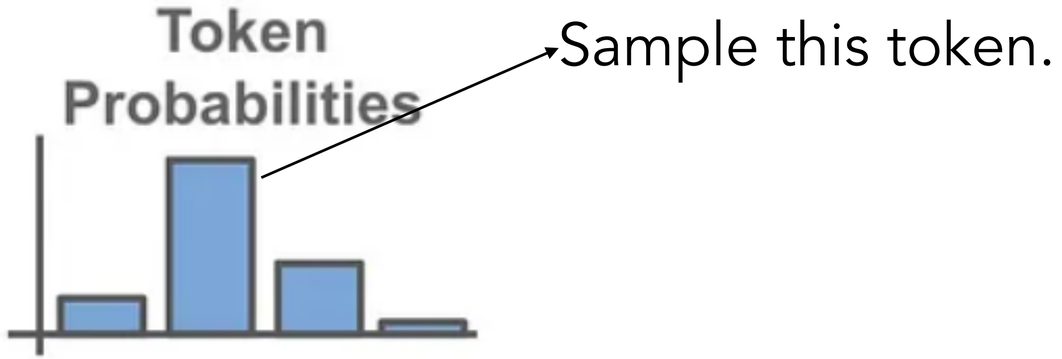
while $\hat{w} \neq < /s >$

sample $\hat{w} \sim P(w \mid x)$

append prefix $x \leftarrow x \oplus \hat{w}$

- Variations: Greedy decoding

$$\hat{w} = \arg\max_{w} P(w \mid x)$$



Decoding / generation using a LM

$\hat{w} = None$

while $\hat{w} \neq < /s >$

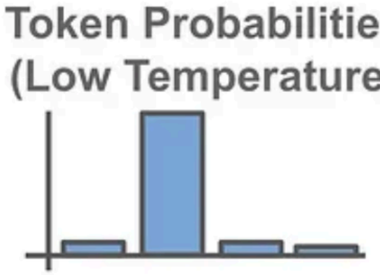
sample $\hat{w} \sim P(w|x)$

append prefix $x \leftarrow x \oplus \hat{w}$

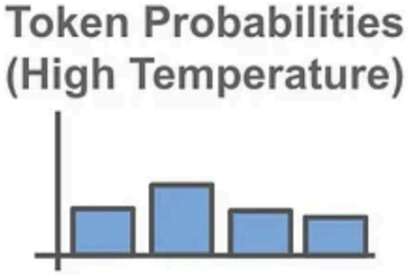
- Variations: Temperature scaling

$$\text{softmax}(\mathbf{o})_i = \frac{e^{o_i}}{\sum_{k=1}^{V} e^{o_k}}$$

$$\text{softmax}(\mathbf{o})_i = \frac{e^{o_i/T}}{\sum_{k=1}^{V} e^{o_k/T}}$$



"sharper" distribution



"flatter" distribution

Decoding / generation using a LM

$\hat{w} = None$

while $\hat{w} \neq < /s >$

sample $\hat{w} \sim P(w \mid x)$

append prefix $x \leftarrow x \oplus \hat{w}$

- Variations: Top-k Sampling
  - Retain only the top-k tokens
  - Rescale probabilities so that they add up to 1.

Original

The sky is

blue 0.4
overcast 0.1
limit 0.1
clear 0.4

Top-k(=2)

The sky is

blue 0.5
overcast 0
limit 0
clear 0.5

Decoding / generation using a LM

$\hat{w} = None$

while $\hat{w} \neq < /s >$

sample $\hat{w} \sim P(w \mid x)$

append prefix $x \leftarrow x \oplus \hat{w}$

- Variations: Top-p Sampling / nucleus sampling
  - Sort the distribution from most probable.
  - Retain smallest set of words $V(p)$ such that:

$$\sum_{w \in V(p)} P(w \mid w_1 \ldots w_{i-1}) \geq p$$

  - Rescale so probabilities add up to 1.

Original

The sky is
- blue 0.3
- overcast 0.15
- limit 0.25
- clear 0.3

Top-p(=0.7)

The sky is
- blue ≈0.3529
- overcast 0
- limit ≈0.2941
- clear ≈0.3529

The sky is
- blue 0.3
- overcast 0
- limit 0.25
- clear 0.3

Decoc

- THE CURIOUS C
  TEXT DEGENER
  HOLTZMANN ET

WebText

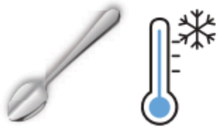
An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

Beam Search, *b*=16

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

Pure Sampling

The Australian Food Safety Authority has warned Australia's beaches may be revitalised this year because healthy seabirds and seals have been on the move. More than 50,000 seabirds, sea mammals and seahorses have been swept into the sea by the Holden CS118 and Adelaide Airport CS300 from 2013. A major white-bat and umidauda migration across Australia is under way in Australia for the first time, with numbers reaching an estimated 50,000.



Sampling, *t*=0.9

Last week's intense storms and a series of powerful cyclones have been officially blamed for the deaths of at least nine large fin whales near Whitsundays - the largest loss of any species globally. The fin whales: packed in the belly of one killer whale thrashing madly in fear as another tries to bring it to safety. When the colossal animal breached the waters of Whitsundays, he'd been seen tagged for a decade.

Top-*k*, *k*=640

Pumping Station #3 shut down due to construction damage Find more at:
www.abc.net.au/environment/species-worry/
in-the-top-10-killer-whale-catastrophes-in-history.html
"In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

Top-*k*, *k*=40, *t*=0.7

The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.

Nucleus, *p*=0.95

There has been an unprecedented number of calves caught in the nets of whaling stations that operate in WA. Pilot whales continue to migrate to feeding grounds to feed their calves. They are now vulnerable due to the decline of wild populations; they are restricted to one breeding site each year. Image copyright Yoon Bo Kim But, with sharp decline in wild populations the size of the Petrels are shrinking and dwindling population means there will only be room for a few new fowl.

How good is a generated output?

Why is perplexity not a good measure for generation performance?

- There is no generation involved! Teacher forcing is used! Issues?

Perplexity

- Assume test set = <s> So long and thanks for all the fish </s>
- Perplexity:

- Generation:

We need to evaluate actual outputs!

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.

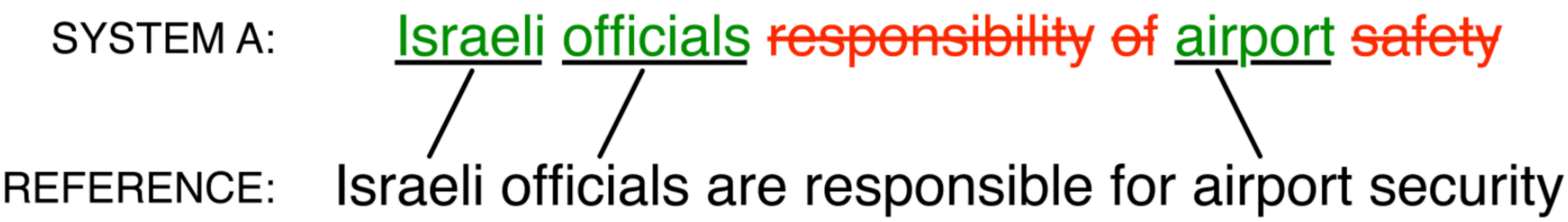
Precision and Recall of Words

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

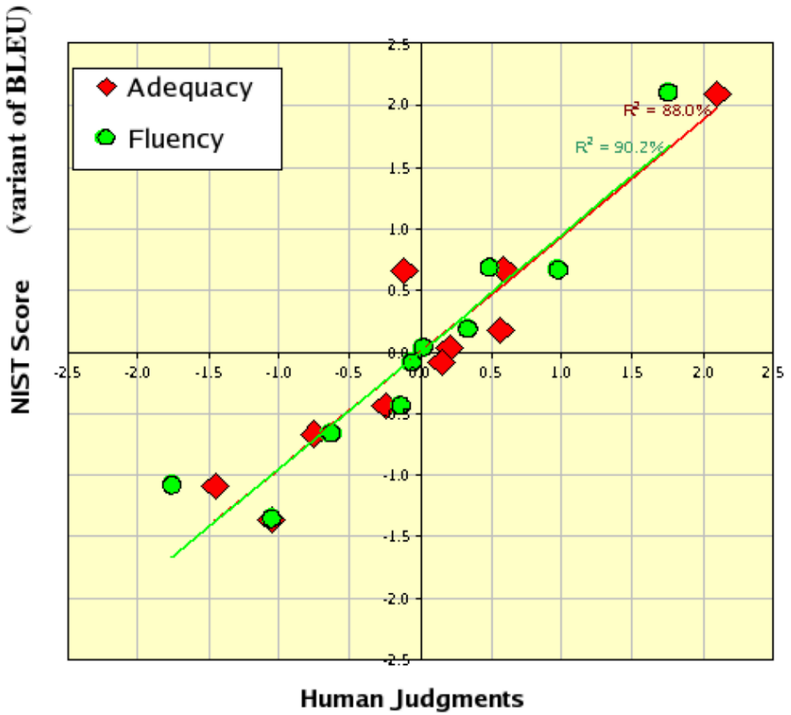
Example

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.

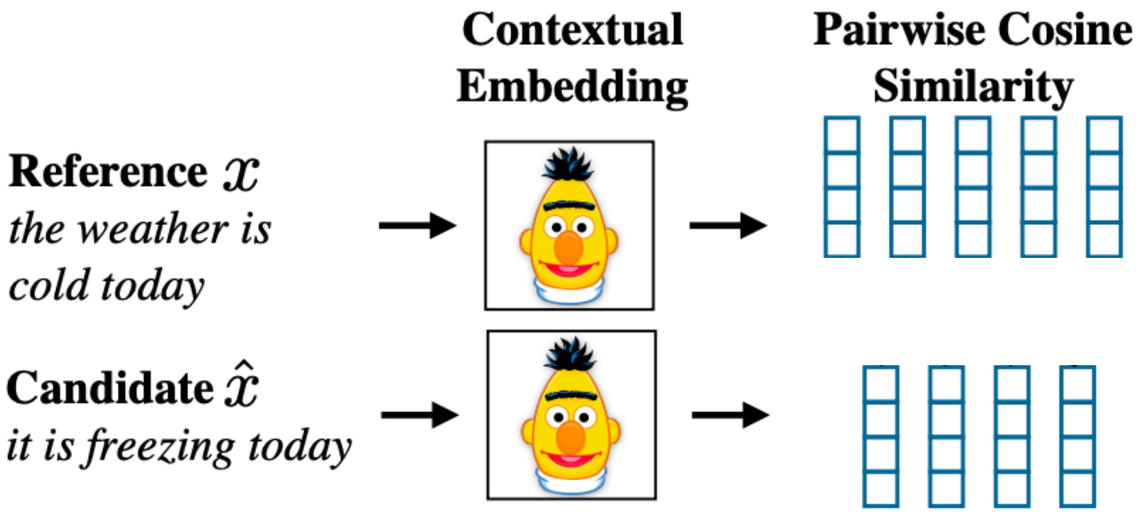
    Generated Output: The film was great.

  - 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

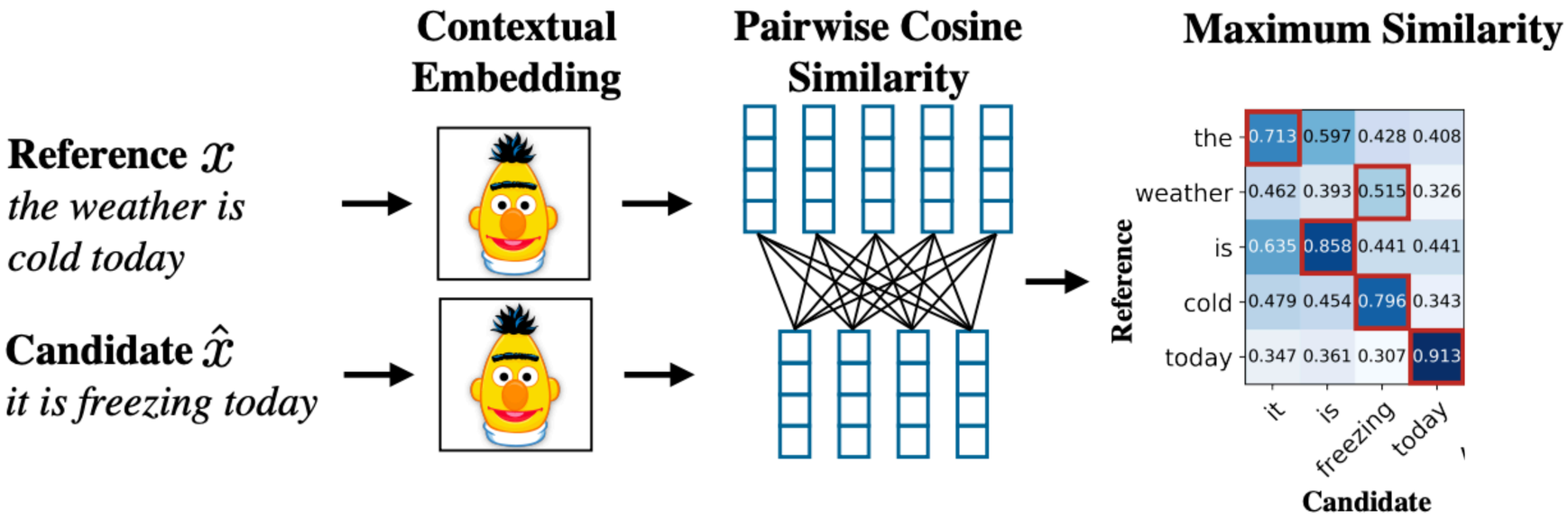
- BERTScore (Zhang et al, ICLR 2020) **Cornell authors!!



- 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!!

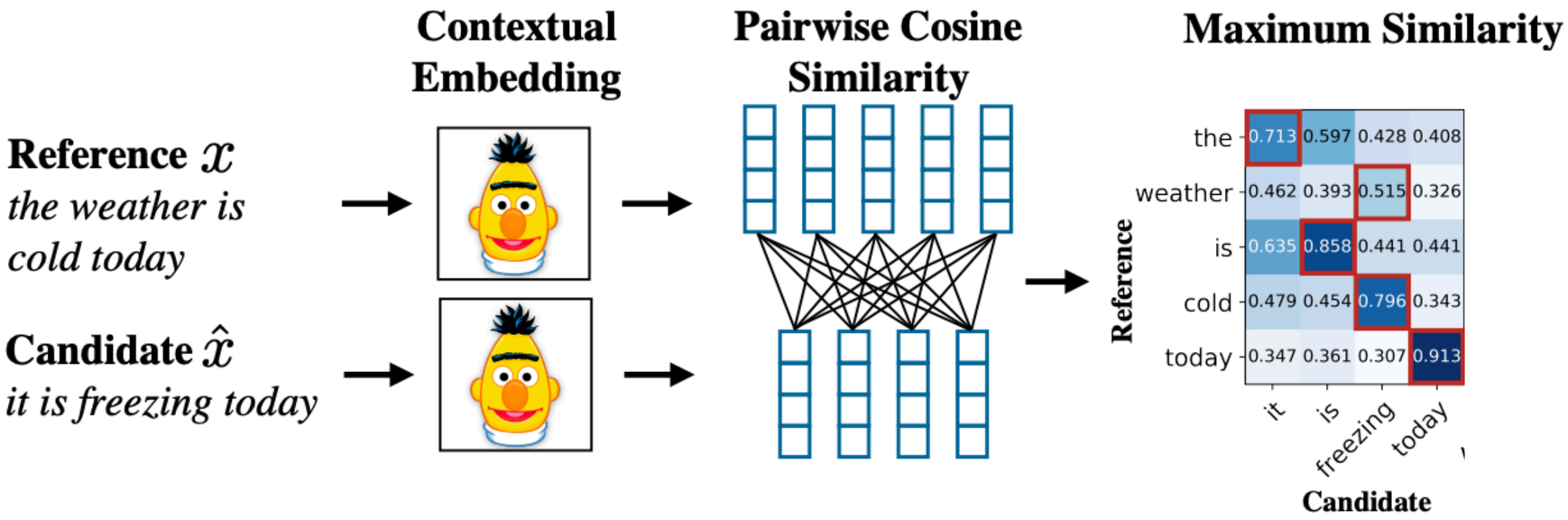


$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

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

Solution: Distributional similarity-based metrics

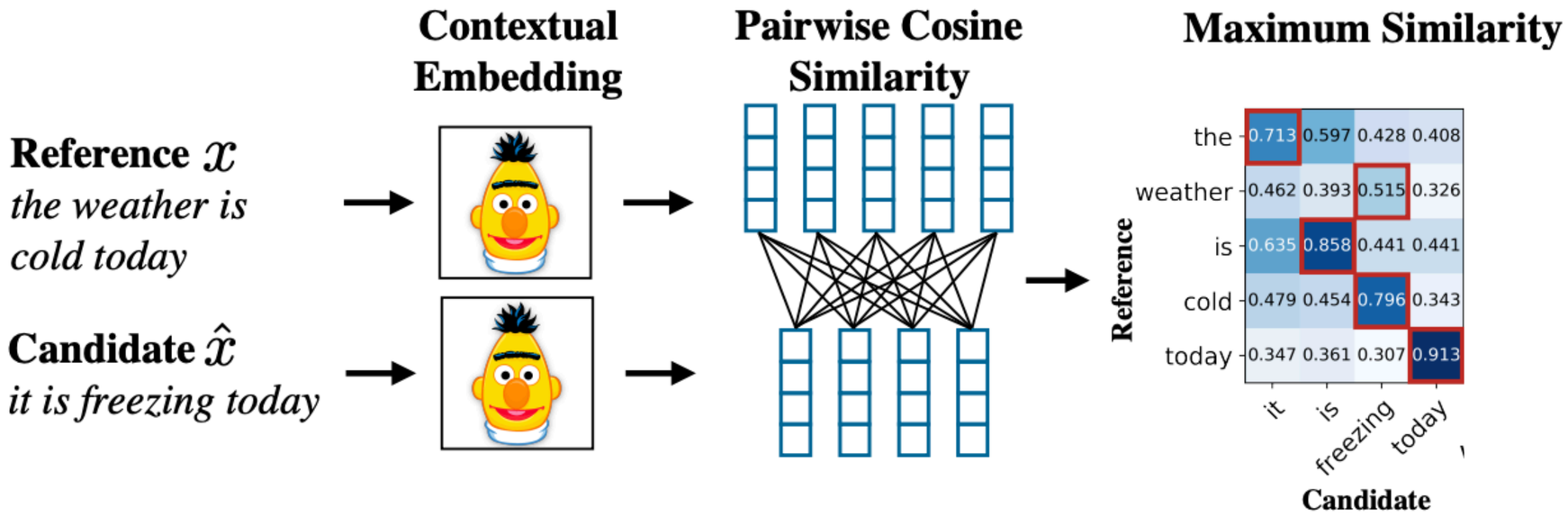
- BERTScore (Zhang et al, ICLR 2020) **Cornell authors!!



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

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

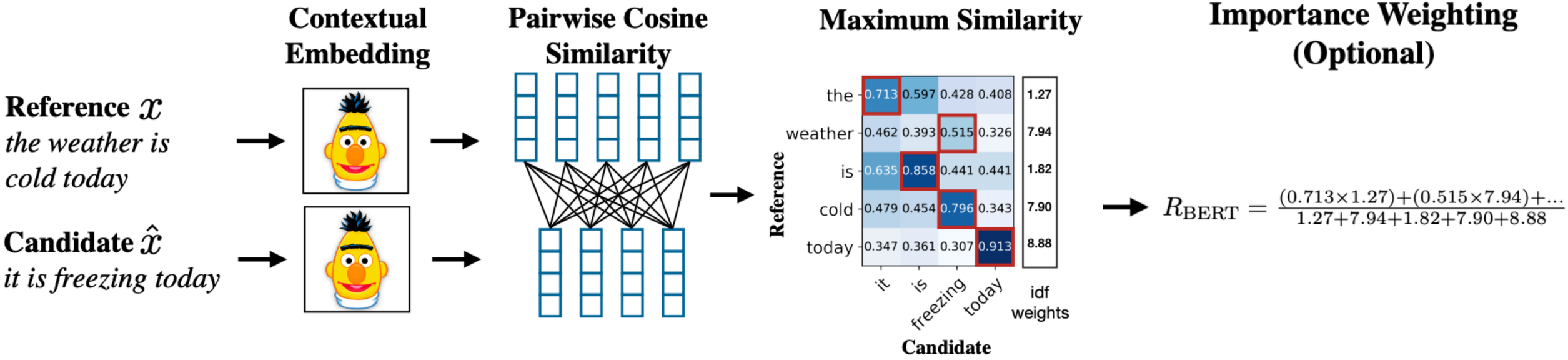
Solution: Distributional similarity-based metrics



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

Solution: Distributional similarity-based metrics

- Optional importance weighting of each reference token to compute recall



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

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

Evaluation Today

- LLM Judges!!

Slide Acknowledgements

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