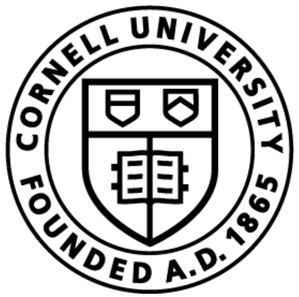
Lecture 19: Evaluation of generated text



Cornell Bowers C·IS
Computer Science

Claire Cardie, Tanya Goyal

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

Announcements

- HW2 written grades released.
  - As with HW1, HW2-programming regrades through a dummy question in HW2-written.
  - Regrades will close on Sunday.
- HW3 due on Monday, April 21.

Recap: Text generation using LMs

- Given an input prefix $\mathbf{x} = x_1 x_2 \cdots x_n$, a LM places a probability distribution over the next token: $P(w \mid x_1 x_2 \cdots x_n)$
- We can generate text using this language model:

$\hat{w} = None$

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

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

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

- Might be called generation / decoding / inference …

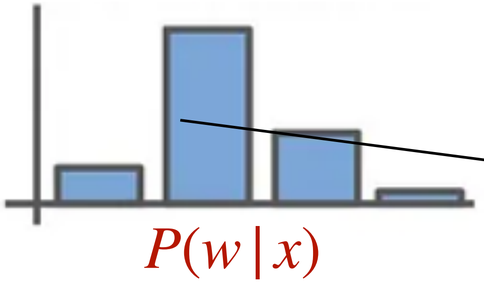
Recap: Decoding / generation using a LM

$\hat{w} = None$

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

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

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



- **<u>Greedy</u>**

Sample this token.

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

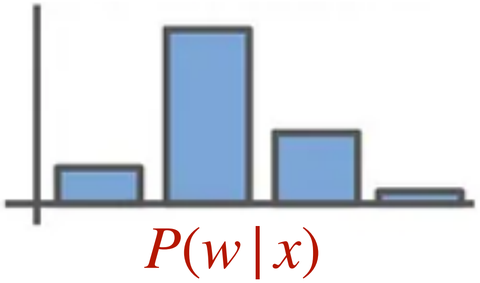
Recap: Decoding / generation using a LM

$\hat{w} = None$

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

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

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



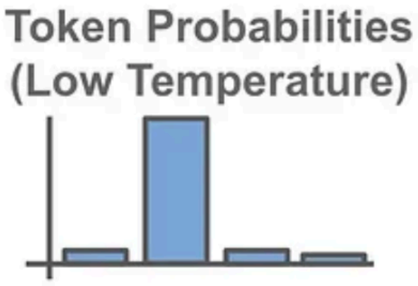
$P(w \mid x)$

- **Temperature scaling**

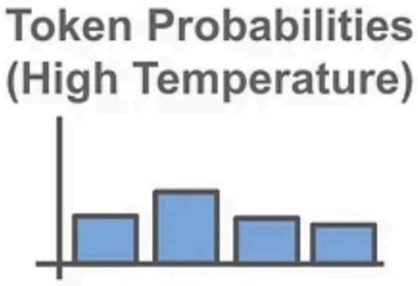
$$\text{softmax}(\mathbf{o})_i = \frac{e^{o_i/T}}{\sum_{k=1}^{V} e^{o_k/T}}$$

- **Greedy**

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



"sharper" distribution



"flatter" distribution

Recap: Decoding / generation using a LM

$\hat{w} = None$
while $\hat{w} \neq < /s >$
  sample $\hat{w} \sim P(w|x)$
  append prefix $x \leftarrow x \oplus \hat{w}$

- **Top-k sampling**
  - Retain only the top-k tokens
  - Rescale probabilities so that they add up to 1.

- **Greedy**

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

- **Temperature scaling**

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

Original	Top-k(=2)
The sky is	The sky is
blue 0.4	blue 0.5
overcast 0.1	overcast 0
limit 0.1	limit 0
clear 0.4	clear 0.5

Recap: Decoding / generation using a LM

$\hat{w} = None$

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

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

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

- **Greedy**

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

- **Temperature scaling**

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

- **Top-k sampling**

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

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

  - 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

Recap: Text generation using LMs

- Suppose we sample multiple outputs from a language model using this algorithm. Which of the following decoding strategies will give the same output each time?

$\hat{w} = None$

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

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

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

A) Regular sampling

B) Top-p sampling

C) Top-k sampling

D) Greedy

E) Temperature scaling

Extrinsic Evaluation of text generation

Input: 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.

- How good is this summary? Hard to quantify.
- Evaluation metrics:
  - Subjective evaluation by humans: Costly, slow, inconsistent.
  - Automatic evaluation using models

Reference-based evaluation metrics

- Key idea:
  - For each input x (say a news article if the task is summarization, or text in the source language if the task is translation), assume access to a **gold output.**
  - **Evaluation metric:** Compute similarity between the model **generated output** and this **gold output**

- 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.

Reference-based evaluation metrics

Generated: It is freezing today.

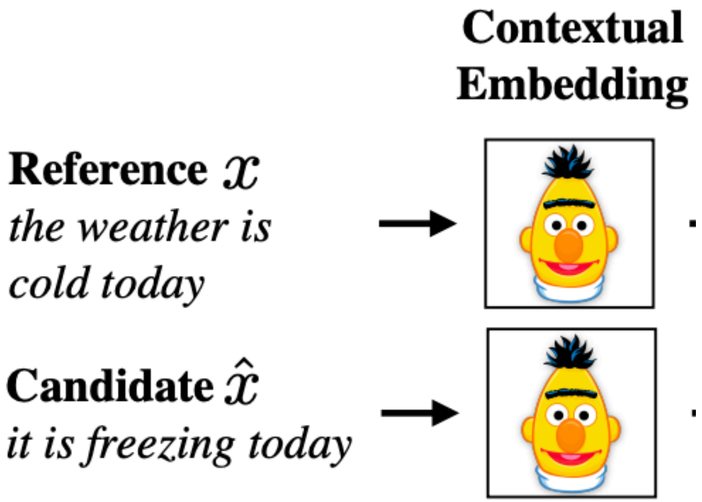
Reference: The weather is cold today.

- Limitations of lexical overlap based metrics?
  - Depend on strict overlap. Do not account for synonym replacements, reordering, etc.
  - All words are treated equally
  - Need (input, gold output) pairs. This data is difficult to get!

- Variants of 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

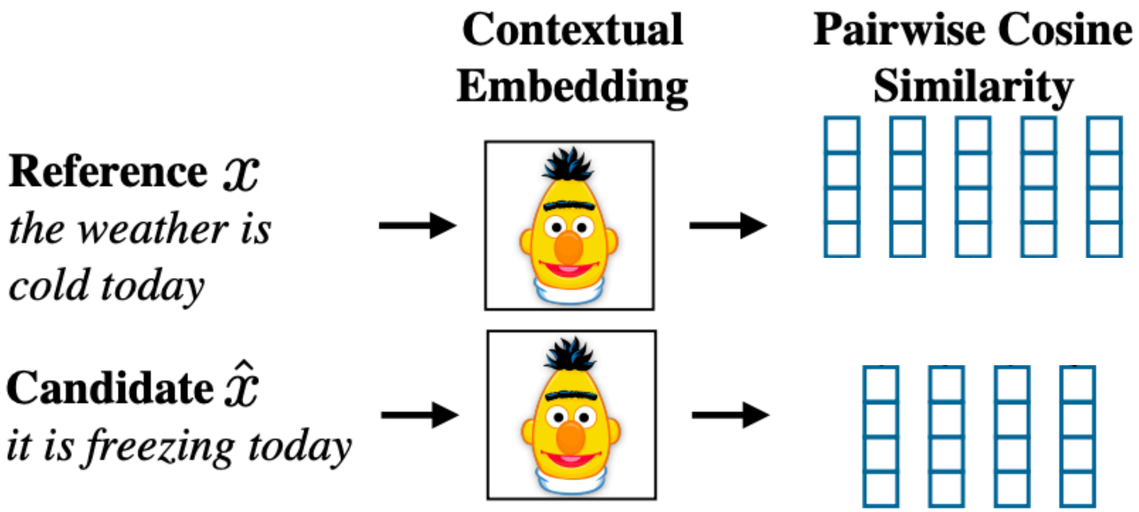
- BERTScore (Zhang et al, ICLR 2020) **Cornell authors!!



- 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

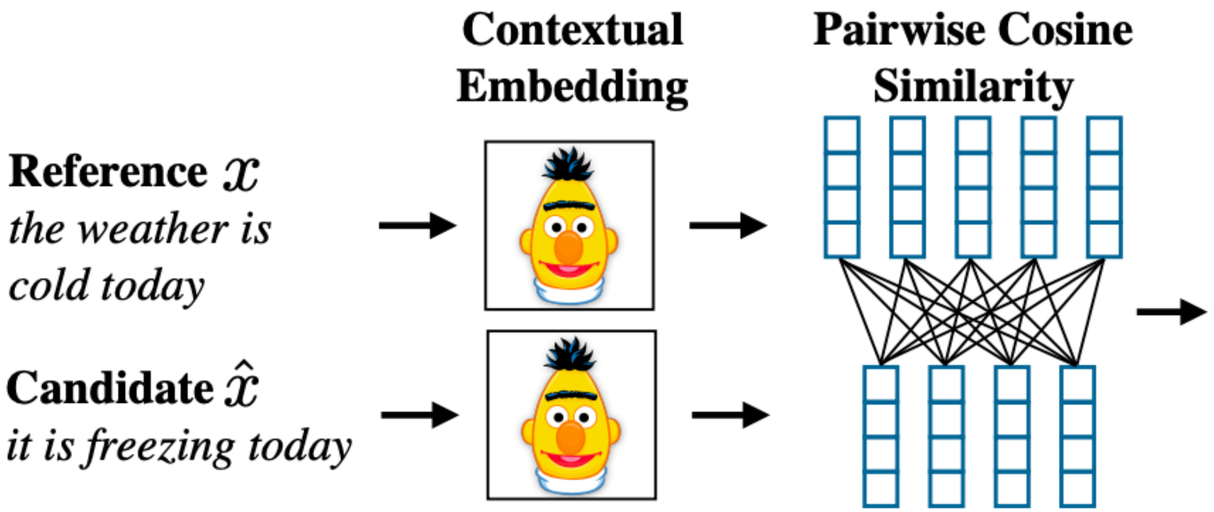
- BERTScore (Zhang et al, ICLR 2020) **Cornell authors!!



- 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!!

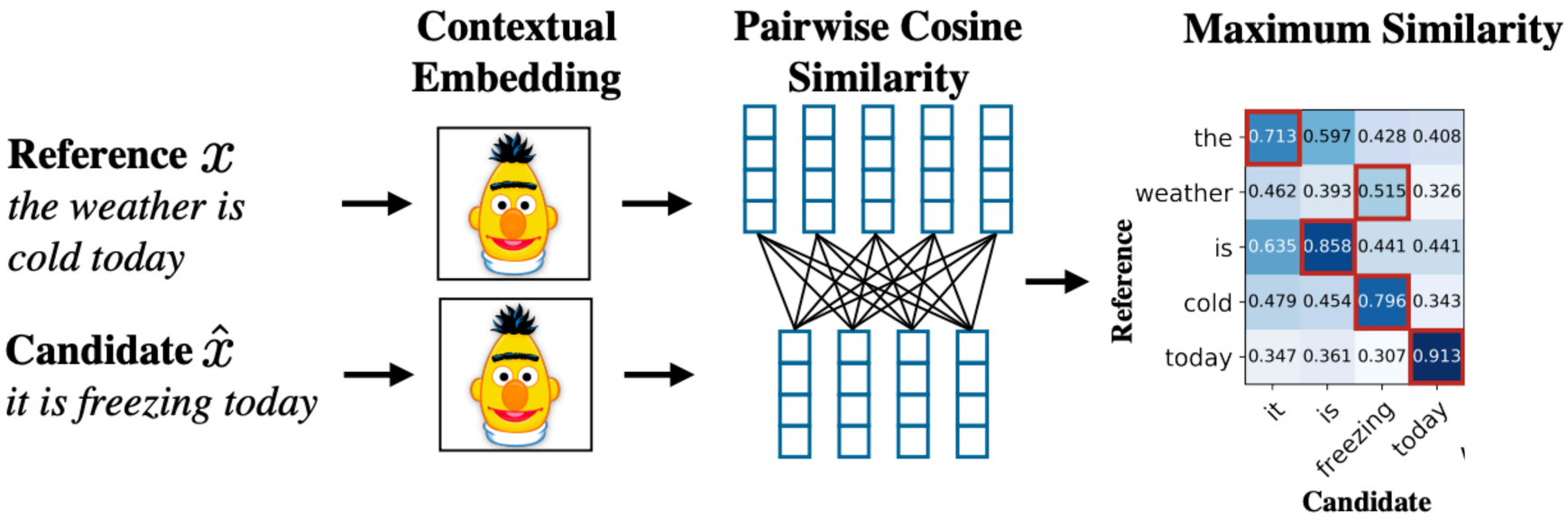


$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

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

Solution: Distributional similarity-based metrics

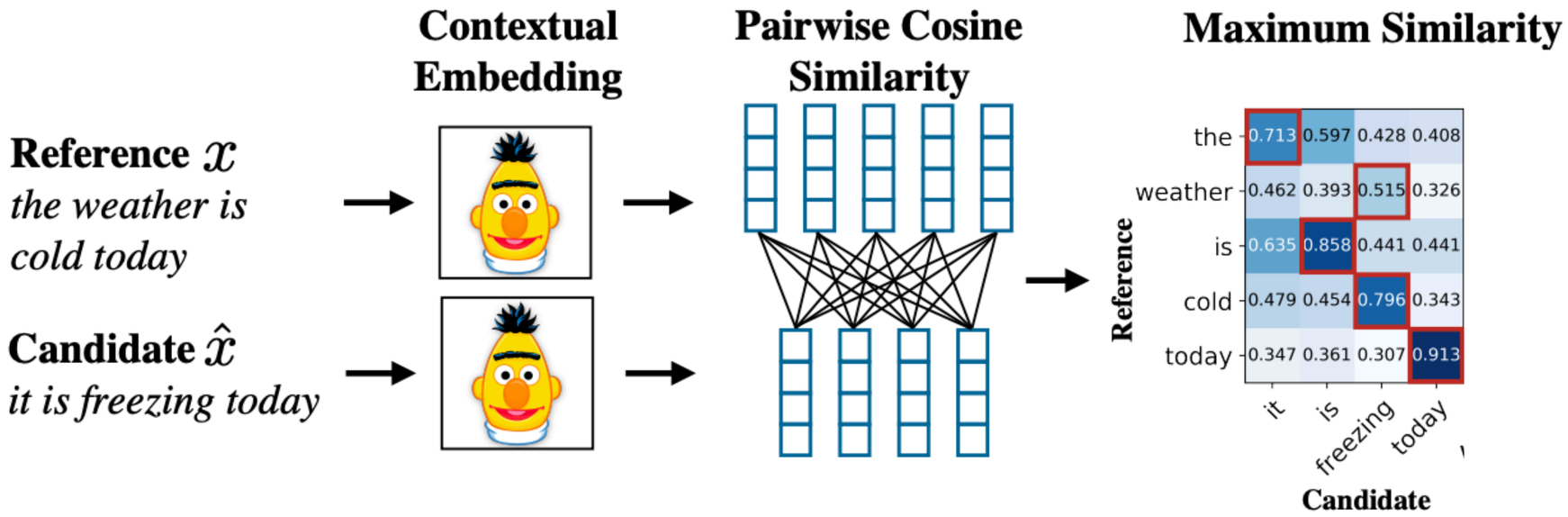
- BERTScore (Zhang et al, ICLR 2020) **Cornell authors!!



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

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

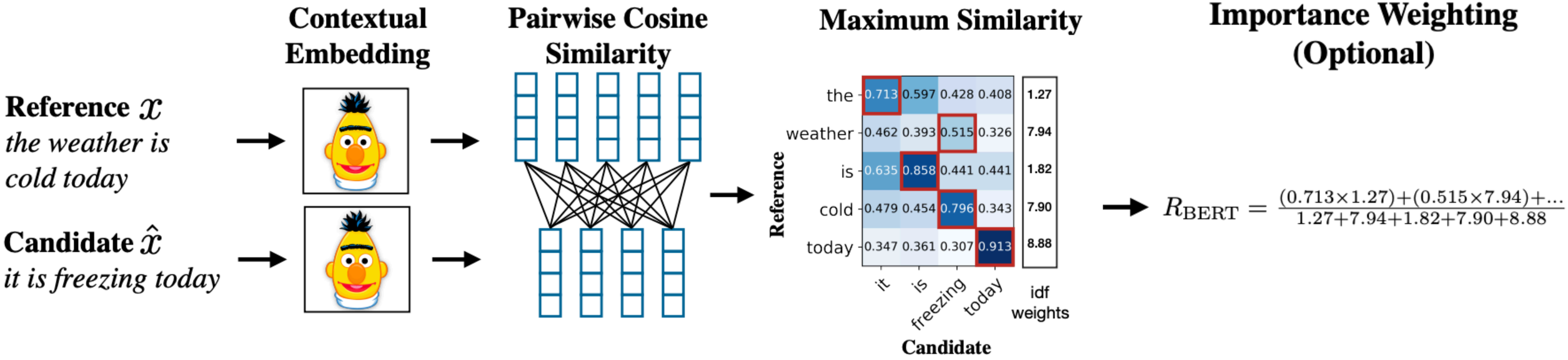
Solution: Distributional similarity-based metrics



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

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.

Any remaining issues?

Input: 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.

- Multiple "good" summaries:
  - 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.

- Comparing against a single **gold summary** will unfairly penalize these other summaries.

Any remaining issues?

Input: 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.

- Output1: An Etihad Airways passenger was devastated after the airline lost her cat Felix.
- Output2: A Chicago woman is searching for her **dog** after it went missing while being transported on an Etihad flight.

- Word overlap cannot account for factuality errors, esp. if minimally lexically different!

Evaluating Factuality of Generated Text

- How can we evaluate if a generated output is factual / non-factual?
- Let's focus on summarization:
  - Given (input, generated output) pair.
  - We want a function $f(\text{input, generated output}) \rightarrow \{0,1\}$, where 0 means that the output is non-factual / has errors, 1 means that the output is factual.

Q1: Why $f(\text{input, generated output})$ and not $f(\text{gold output, generated output})$?

Q2: How can we parameterize this function f?

Evaluating Factuality of Generated Text

- This is a binary text classification task!
- Training data?
  - (input, output, y= {0, 1}) tuples.
  - Train a binary classification model.

$$f\left(\text{  }, \textit{A Chicago woman is searching for her cat after it went missing while being transported on an Etihad flight.}\right) \rightarrow \textbf{\textit{Factual / Non-factual}}$$

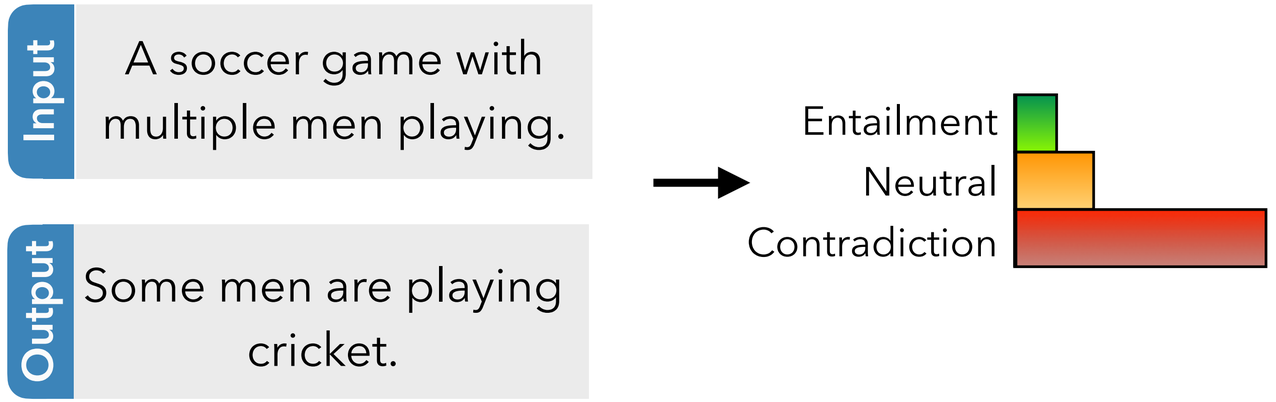
Q: Which of the following models can be used?

A) BERT (encoder-only model)

B) GPT-2 (decoder-only model)

Evaluating Factuality using entailment

- Text entailment is a long-standing task in NLP. Given a premise and a hypothesis, the task is to determine whether the hypothesis is entailed by the premise.



- Very popular to use off-the-shelf entailment models to evaluate factuality of summarization. Scialom et al., EMNLP2021, Fabbri et al., NAACL2022, Laban et al. TACL 2021, etc.

Evaluating Factuality using QA

- Question answering systems to evaluate factuality.
- Idea: Check if, for relevant questions, does the input and the generated output give the same answers?

Generated: A Chicago woman is searching for her dog after ….

Step1: Generated questions

Who is searching?

Who is the Chicago woman searching for?

Step2: Answer questions using both the input and generated summary

✓ Input: Chicago woman Jennifer S. | Generated: Jennifer S.

✗ Input: her cat | her dog

Step2: Compare answers and determine factuality

Non-factual

[1] Falke et al. ACL2019
[2] Wang et al. ACL2020
Etc.

Evaluating Factuality using QA

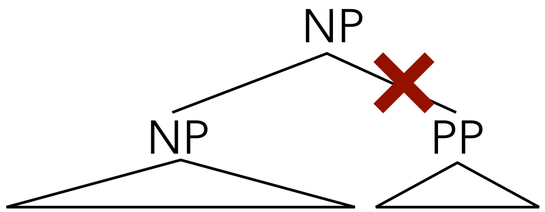
- Question answering systems to evaluate factuality.
- Idea: Check if, for relevant questions, does the input and the generated output give the same answers?
- Requires:
  - A question generation model
  - A question answering model
  - A answer matching model / strategy.
- But
  - Localizes errors to sub-spans

Zooming out

- Which evaluation metrics are appropriate will always depend on the models being evaluated.

Summ. Systems

Extractive Summaries



… is a claims adjuster with Aetna

Extraction + Compression

End-to-End Neural Models

Pre-Trained LMs

Models become strong enough that minimal factuality errors for news summarization. Shift to more difficult summarization settings.

1950s

Mostly extractive systems that copy spans from the input. Non-factuality was not an issue. BLEU/ ROUGE metrics sufficed!

2016 onwards

End-to-end models without explicit copying, but still exhibited a lot of copying behavior

2019

Models become more abstractive. Kick-starts factuality research (entailment models, QA models, etc.)

2023

2025

Slide Acknowledgements

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