CSCI 5541: Natural Language Processing

Lecture 9: Language Models: Evaluations

computer science
& engineering

MINNESOTA · NLP · EST. 2021 ·

UNIVERSITY OF MINNESOTA
Driven to Discover®

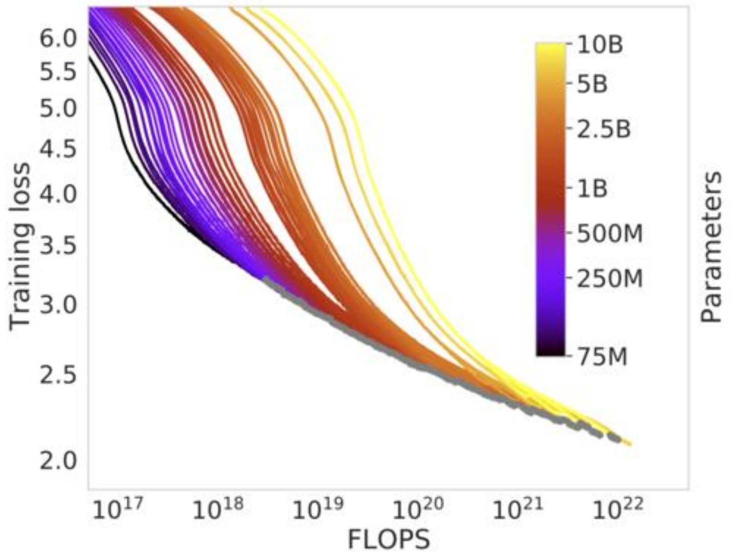
Announcement (0304)

- ❑ HW3 → Due Today (Mar 4)
- ❑ Proposal Report → Due Thursday (Mar 6)
- ❑ HW4 Out

Evaluation methods on generated text

When a language model outputs text, how do we determine if the text it creates is ‘good’?

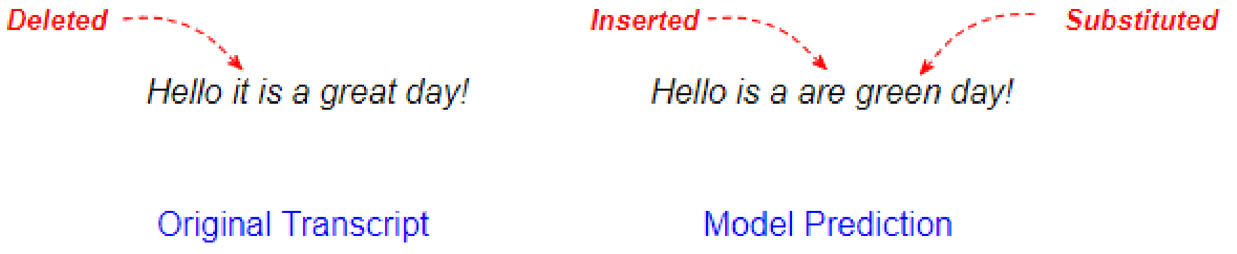
Intrinsic vs. Extrinsic Evaluation



Intrinsic

(perplexity)

Task Agnostic



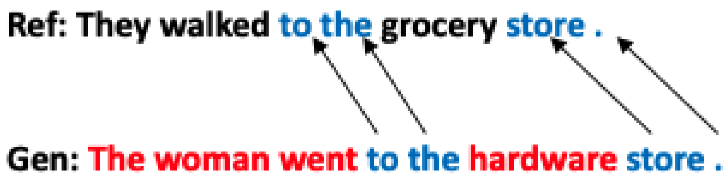
GSM8K

Extrinsic

(BLEU/ROUGE/BERTScore/etc.)

Task Specific

Types of evaluation methods in NLG



Ref: They walked to the grocery store .

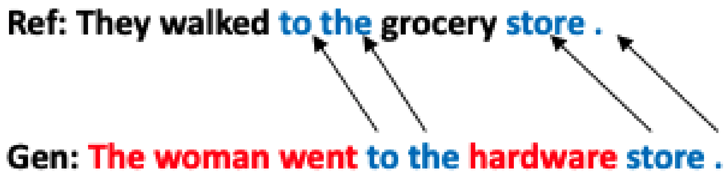
Gen: The woman went to the hardware store .

Content overlap metrics

Model-based metrics

Human evaluations

Types of evaluation methods in NLG



Ref: They walked to the grocery store .

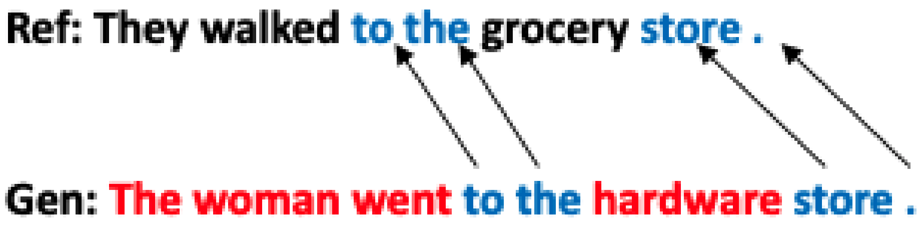
Gen: The woman went to the hardware store .

Content overlap metrics

Model-based metrics

Human evaluations

Content overlap metrics



Ref: They walked to the grocery store .

Gen: The woman went to the hardware store .

- ❑ Compute a score that indicates the similarity between **generated** and **gold-standard** (human-written) text
- ❑ Fast, efficient and widely used
- ❑ Hard to capture context with this method
- ❑ Two broad categories:
  - ○ **N-gram overlap metrics** (e.g., BLEU, ROUGE, METEOR)
  - ○ **Semantic overlap metrics** (e.g., PYRAMID, SPICE)

N-gram overlap metrics

Word overlap–based metrics (BLEU, ROUGE, METEOR, CIDEr, etc.)

❑ They're not ideal for machine translation

❑ They get progressively much worse for tasks that are more open-ended than machine translation

- Worse for summarization, as longer output texts are harder to measure
- Much worse for dialogue, which is more open-ended than summarization
- Much, much worse for story generation, which is also open-ended, but whose sequence length can make it seem you're getting decent scores!

Bilingual Evaluation Understudy (BLEU)

- N-gram overlap between generated text and reference text
- Compute precision for n-grams of size 1 to 4
- Add brevity penalty (for too short translations)
- Typically computed over the entire corpus, not single sentences

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

Bilingual Evaluation Understudy (BLEU)

BLEU (Papineni et al. 2002): what fraction of {1-4}-grams in the system translation appear in the reference translations?

Precision

$$P_n = \frac{\text{Number of ngrams in system and reference translations}}{\text{Number of ngrams in system translation}}$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{1-r/c} & if\ c \leq r \end{cases}$$

c = length of hypothesis translation
r = length of closest reference translation

$$BLEU = \quad \text{BP} \quad \exp \frac{1}{N} \sum_{n=1}^{N} \log p_n$$

brevity penalty

Hypothesis/system translation

Appeared calm when he was taken to the American plane, which will Miami Florida, USA.

Appeared
calm
when
he
was
taken
to
the
American
plane
,
which
will
to
Miami
Florida
USA
.

$$P_1 = \frac{15}{18} = 0.833$$

Reference translation

Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.

Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.

Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.

Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Ngrams appearing >1 time in the hypothesis can match up to the max number of times they appear in a single reference e.g., two commas in hypothesis but one max in any single reference.

Hypothesis/system translation

Reference translation

Appeared calm when he was taken to the American plane, which will to Miami, Florida.

Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.

Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.

Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.

Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

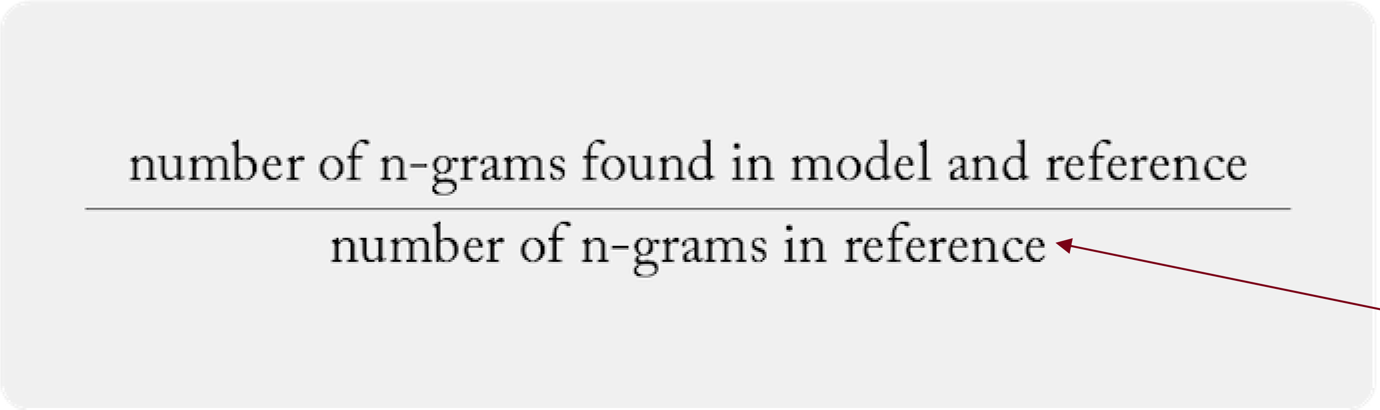
Appeared calm
calm when
when he
he was
was taken
taken to
to the
the American
American plane

plane ,
, which
which will
will to
to Miami
Miami ,
, Florida
Florida .

$$P_2 = \frac{10}{17} = 0.588$$

Recall Oriented Understudy for Gisting Evaluation (ROUGE)

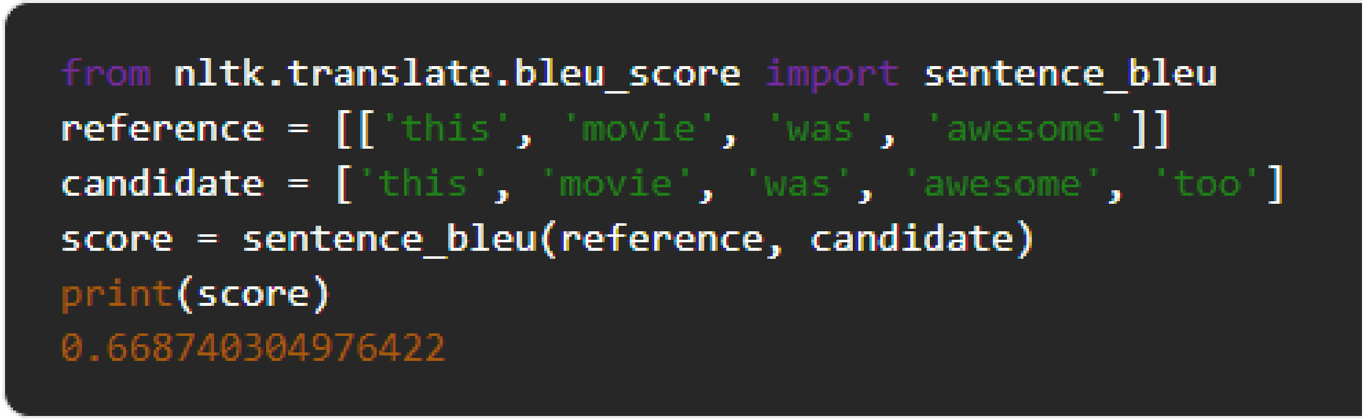
- Overlap between generated text and reference text in terms of **recall**.
- Three types:
  - Rouge-N: the most prevalent form that detects n-gram overlap;
  - Rouge-L: identifies the Longest Common Subsequence
  - Rouge-S: concentrates on skip grams.



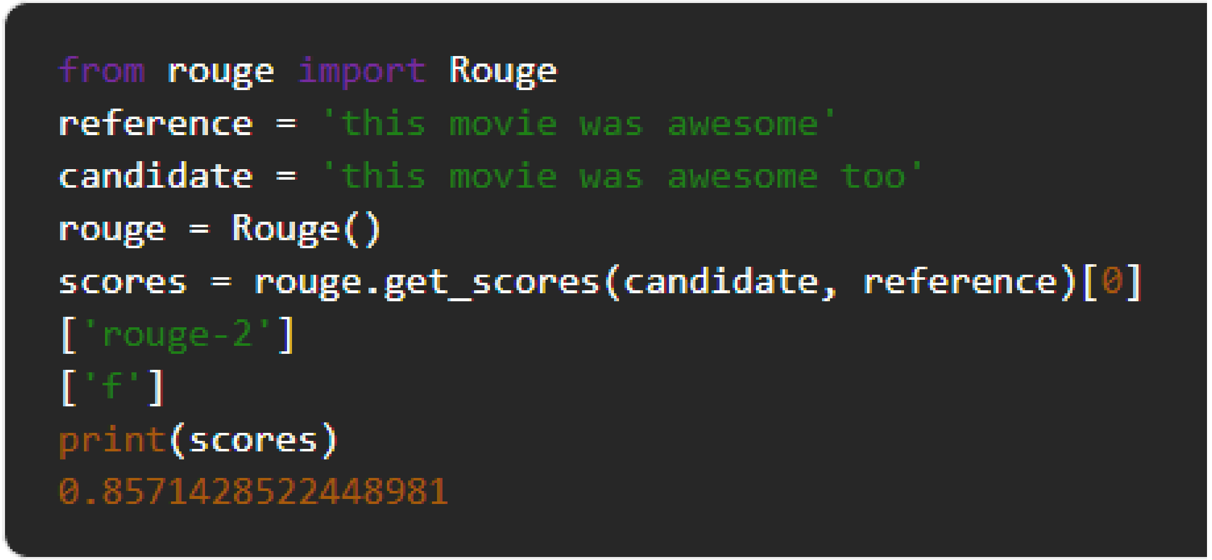
$$\frac{\text{number of n-grams found in model and reference}}{\text{number of n-grams in reference}}$$

The main difference between rouge and bleu is that bleu score is precision-focused whereas rouge score focuses on recall.

BLEU and ROUGE Examples



```
from nltk.translate.bleu_score import sentence_bleu
reference = [['this', 'movie', 'was', 'awesome']]
candidate = ['this', 'movie', 'was', 'awesome', 'too']
score = sentence_bleu(reference, candidate)
print(score)
0.668740304976422
```

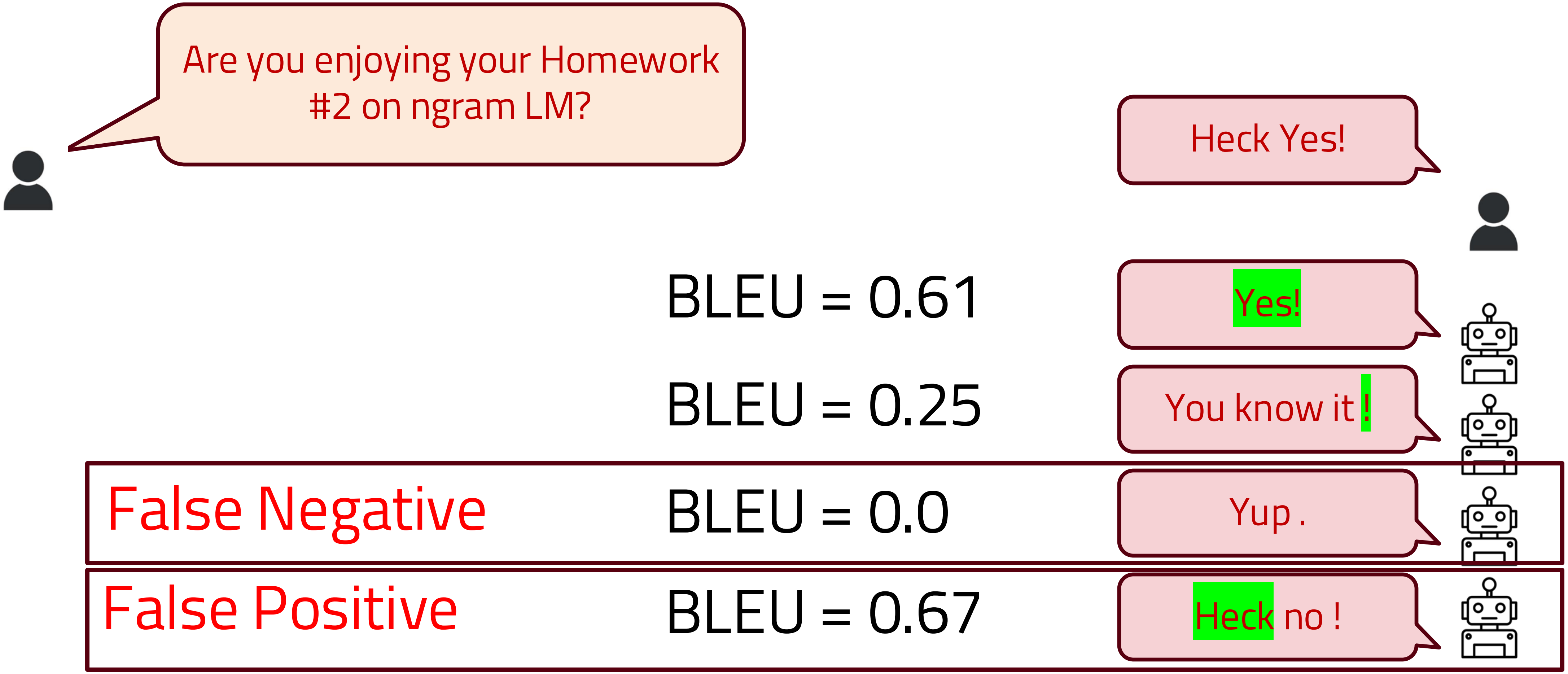


```
from rouge import Rouge
reference = 'this movie was awesome'
candidate = 'this movie was awesome too'
rouge = Rouge()
scores = rouge.get_scores(candidate, reference)[0]
['rouge-2']
['f']
print(scores)
0.8571428522448981
```

https://arize.com/blog-course/generative-ai-metrics-bleu-score/

A simple failure case of BLEU

n-gram overlap metrics have no concept of semantic relatedness!



Are you enjoying your Homework #2 on ngram LM?

Heck Yes!

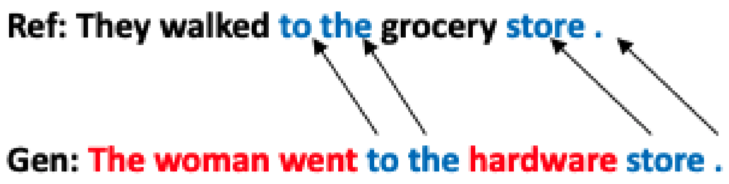
BLEU = 0.61 Yes!

BLEU = 0.25 You know it !

False Negative BLEU = 0.0 Yup .

False Positive BLEU = 0.67 Heck no !

Types of evaluation methods in NLG



Ref: They walked to the grocery store .

Gen: The woman went to the hardware store .

Content overlap metrics

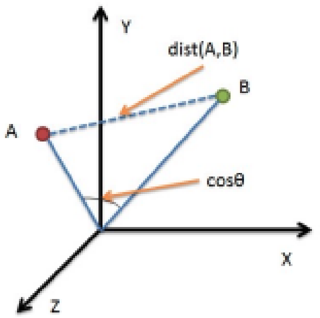
Model-based metrics

Human evaluations

Model-based metrics

- Use learned representations of words and sentences to compute semantic similarity between generated and reference texts
- No more n-gram bottleneck because text units are represented as embeddings
- Even though embeddings are pretrained, distance metrics used to measure the similarity can be fixed

Model-based metrics: Word distance functions

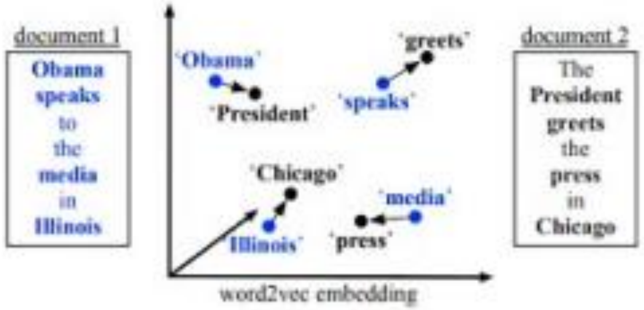


Vector Similarity

Embedding based similarity for semantic distance between text.

- ❑ Embedding Average (Liu et al., 2016)
- ❑ Vector Extrema (Liu et al., 2016)
- ❑ MEANT (Lo, 2017)
- ❑ YISI (Lo, 2019)

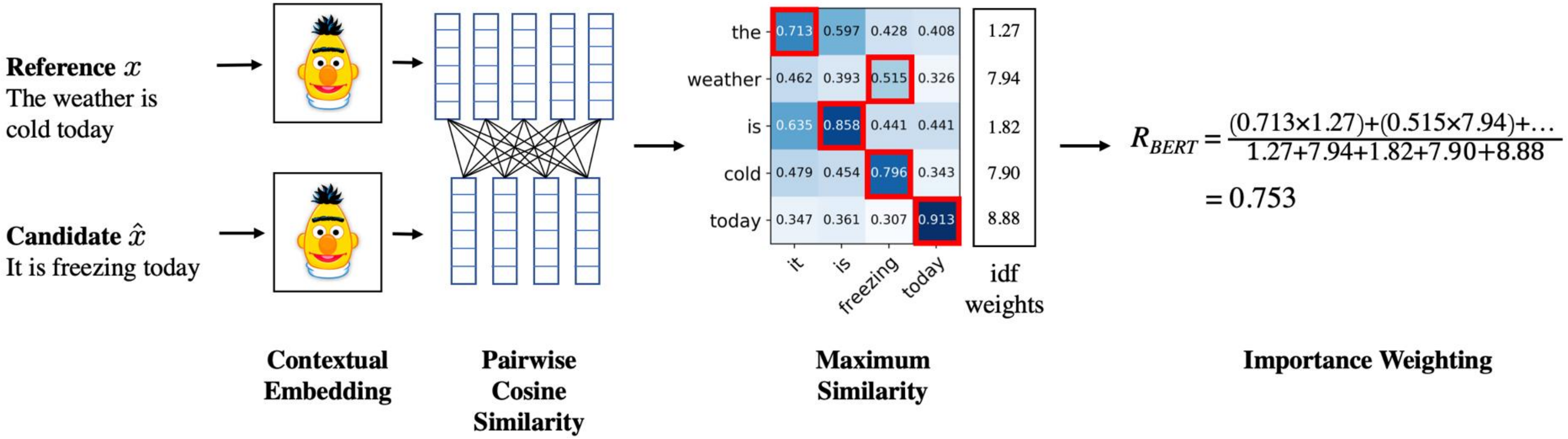
Word Mover's Distance



Measures the distance between two sequences (e.g., sentences, paragraphs, etc.), using word embedding similarity matching. (Kusner et.al., 2015; Zhao et al., 2019)

BERTScore

Uses pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity. (Zhang et.al. 2020)



	it	is	freezing	today	idf weights
the	0.713	0.597	0.428	0.408	1.27
weather	0.462	0.393	0.515	0.326	7.94
is	0.635	0.858	0.441	0.441	1.82
cold	0.479	0.454	0.796	0.343	7.90
today	0.347	0.361	0.307	0.913	8.88

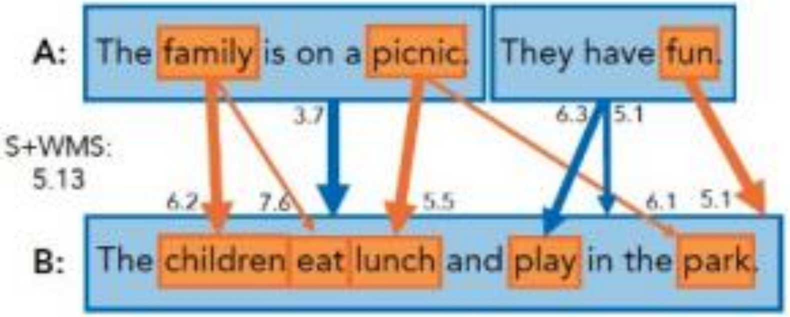
$$R_{BERT} = \frac{(0.713\times1.27)+(0.515\times7.94)+\ldots}{1.27+7.94+1.82+7.90+8.88}$$

$$= 0.753$$

Model-based metrics: Beyond word matching

Sentence Movers Similarity

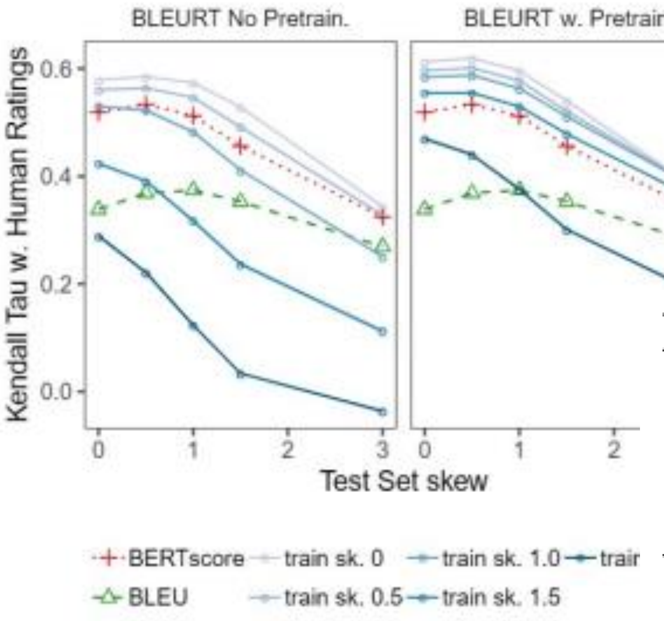
Based on Word Movers Distance to evaluate text in a continuous space using sentence embeddings from recurrent neural network representations. (Clark et.al., 2019)

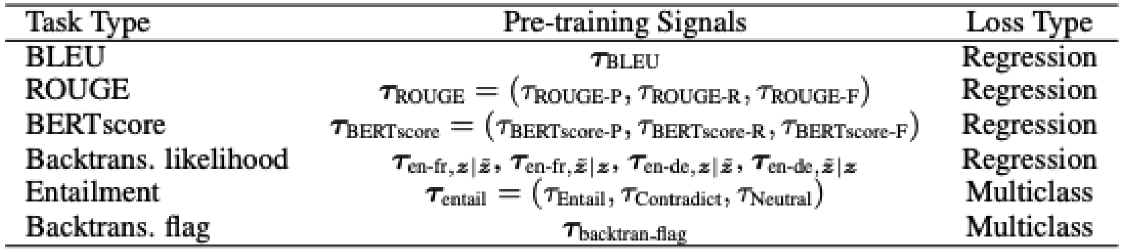


A:
The family is on a picnic.
They have fun.
S+WMS:
5.13
3.7
6.3
5.1
6.2
7.6
5.5
6.1
5.1
B:
The children eat lunch and play in the park.

BLEURT

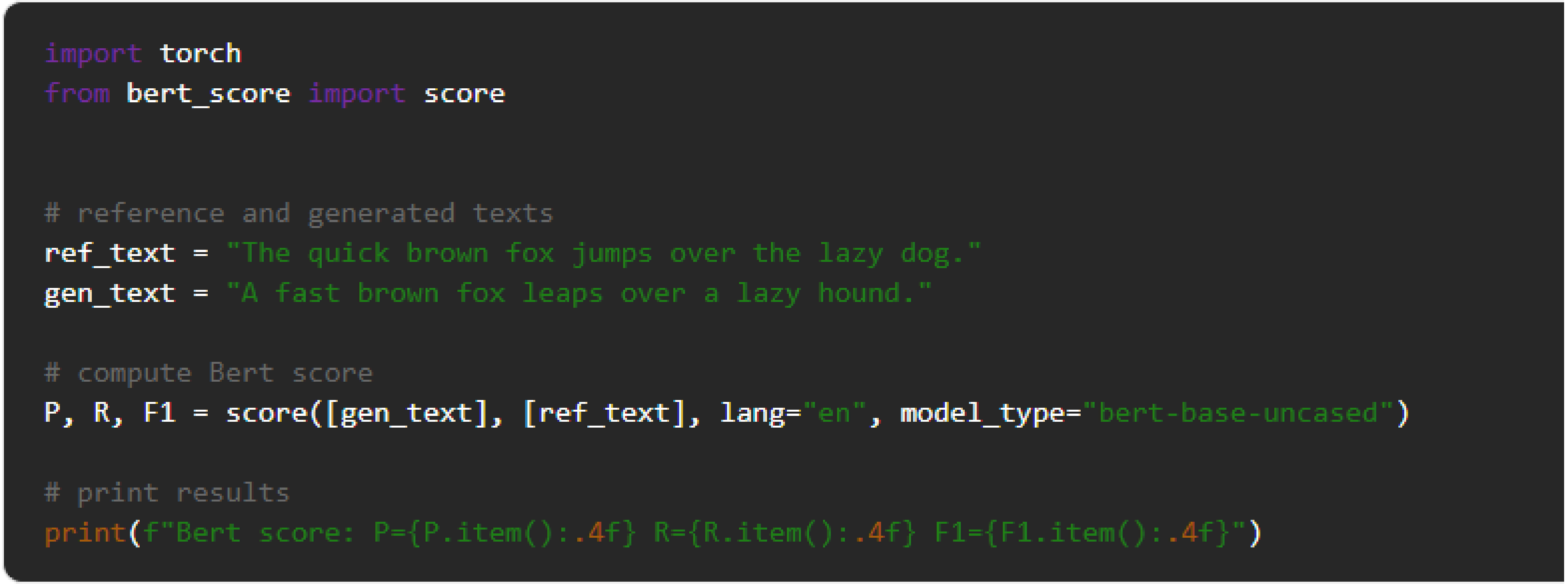
A regression model based on BERT returns a score that indicates to what extent the candidate text is grammatical and conveys the meaning of the reference text. (Sellam et.al. 2020)





Task Type	Pre-training Signals	Loss Type
BLEU	τ_{BLEU}	Regression
ROUGE	$\tau_{\text{ROUGE}} = (\tau_{\text{ROUGE-P}}, \tau_{\text{ROUGE-R}}, \tau_{\text{ROUGE-F}})$	Regression
BERTscore	$\tau_{\text{BERTscore}} = (\tau_{\text{BERTscore-P}}, \tau_{\text{BERTscore-R}}, \tau_{\text{BERTscore-F}})$	Regression
Backtrans. likelihood	$\tau_{\text{en-fr},z\|\tilde{z}}, \tau_{\text{en-fr},\tilde{z}\|z}, \tau_{\text{en-de},z\|\tilde{z}}, \tau_{\text{en-de},\tilde{z}\|z}$	Regression
Entailment	$\tau_{\text{entail}} = (\tau_{\text{Entail}}, \tau_{\text{Contradict}}, \tau_{\text{Neutral}})$	Multiclass
Backtrans. flag	$\tau_{\text{backtran\_flag}}$	Multiclass

Table 1: Our pre-training signals.



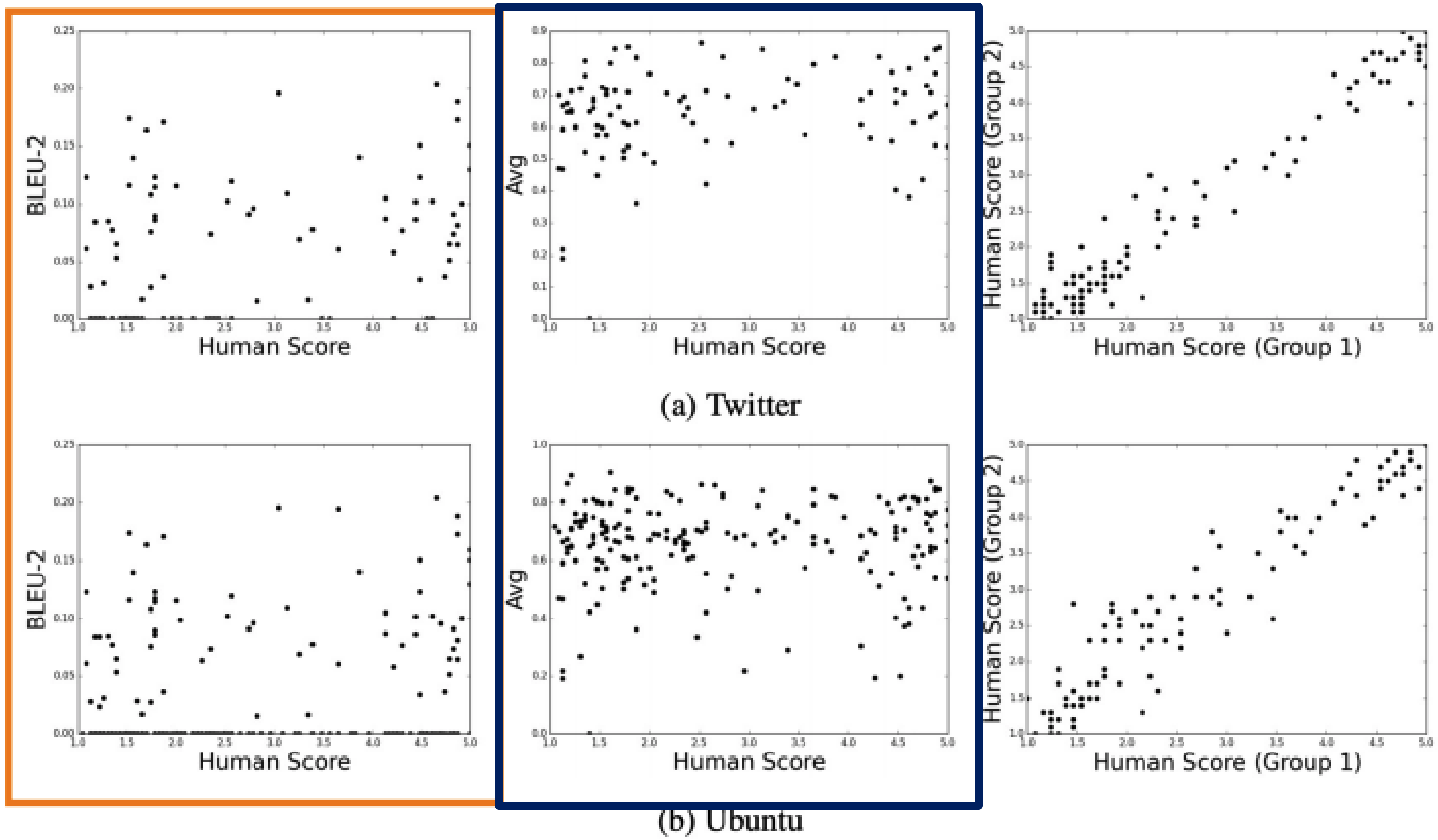
```python
import torch
from bert_score import score


# reference and generated texts
ref_text = "The quick brown fox jumps over the lazy dog."
gen_text = "A fast brown fox leaps over a lazy hound."

# compute Bert score
P, R, F1 = score([gen_text], [ref_text], lang="en", model_type="bert-base-uncased")

# print results
print(f"Bert score: P={P.item():.4f} R={R.item():.4f} F1={F1.item():.4f}")
```

Automatic metrics in general don’t really work



(a) Twitter

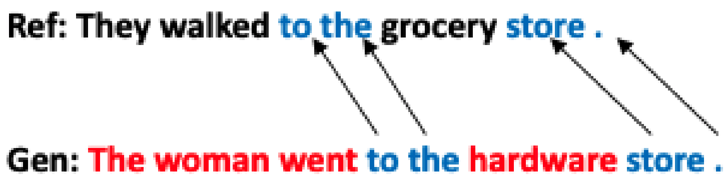
(b) Ubuntu

Figure 1: Scatter plots showing the correlation between metrics and human judgements on the Twitter corpus (a) and Ubuntu Dialogue Corpus (b). The plots represent BLEU-2 (left), embedding average (center), and correlation between two randomly selected halves of human respondents (right).

(Liu et.al., 2016)

What if there is no reference text?

Types of evaluation methods in NLG



Ref: They walked to the grocery store .

Gen: The woman went to the hardware store .

Content overlap metrics

Model-based metrics

Human evaluations

Human Evaluations

- ❑ Automatic metrics fall short of matching human decisions

- ❑ Human evaluation is most important form of evaluation for text generation systems
  - ○ >75% generation papers at ACL 2019 included human evaluations

- ❑ Gold standard in developing new automatic metrics
  - ○ New automated metrics must correlate well with human evaluations!

Human Evaluations

❑ Ask humans to evaluate the quality of generated text

❑ Overall or along some specific dimension:

- fluency
- coherence / consistency
- factuality and correctness
- commonsense
- style / formality
- grammaticality
- typicality
- redundancy

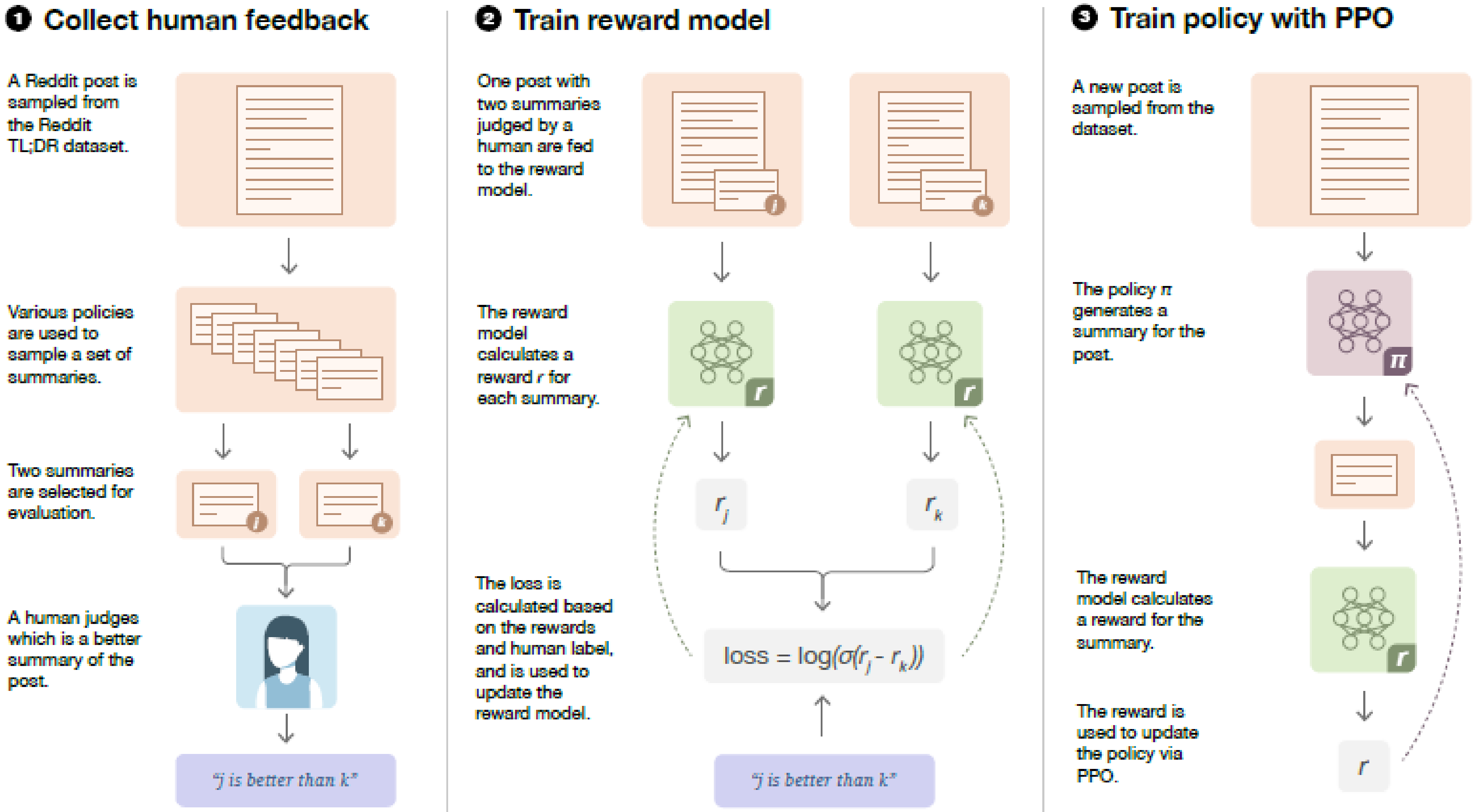
Note: Don’t compare human evaluation scores across differently conducted studies
Even if they claim to evaluate the same dimensions!

Human evaluation: Issues

- ❑ Human judgments are regarded as the gold standard

- ❑ Of course, we know that human eval is slow and expensive

- ❑ Conducting human evaluation effectively is very difficult
  - ○ Humans are *are inconsistent*
    can be illogical
    lose concentration
    misinterpret your question
    can't always explain why they feel the way they do



❶ Collect human feedback

A Reddit post is sampled from the Reddit TL;DR dataset.

Various policies are used to sample a set of summaries.

Two summaries are selected for evaluation.

A human judges which is a better summary of the post.

"j is better than k"

❷ Train reward model

One post with two summaries judged by a human are fed to the reward model.

The reward model calculates a reward *r* for each summary.

The loss is calculated based on the rewards and human label, and is used to update the reward model.

$\text{loss} = \log(\sigma(r_j - r_k))$

"j is better than k"

❸ Train policy with PPO

A new post is sampled from the dataset.

The policy *π* generates a summary for the post.

The reward model calculates a reward for the summary.

The reward is used to update the policy via PPO.

[2009.01325] Learning to summarize from human feedback

Evaluation: Takeaways

- ❑ Content overlap metrics provide a good starting point for evaluating the quality of generated text. You will need to use one but they're not good enough on their own.
- ❑ Model-based metrics can be more correlated with human judgment, but behavior is not interpretable
- ❑ Human judgments are critical
  - ○ Only thing that can directly evaluate factuality, but humans are inconsistent!
- ❑ In many cases, the best judge of output quality is YOU!
  - ○ Look at your model generations. Don't just rely on numbers!
  - ○ Don't cherry pick! Publicly release large samples of the output of systems that you create!

Conclusion

- ❑ Interacting with natural language generation systems quickly shows their limitations
- ❑ Even in tasks with more progress, there are still many improvements ahead
- ❑ Evaluation remains a huge challenge.
  - ○ We need better ways of automatically evaluating performance of NLG systems
- ❑ One of the most exciting and fun areas of NLP to work in!