

Natural Language Generation:

Evaluation

Antoine Bosselut

Today's Outline

- **Lecture:**
 - **Evaluation:** Content overlap metrics, model-based metrics, human evaluations
- **Exercise Session**
 - **Review:** Robustness & Prompting
 - **New:** Text Generation

Midterm Logistics

- **Date:**

- Wednesday, April 9th, 2025
- 11:30 - 13:00

- **Location:** 6 rooms on campus

- INF 1, AAC 2 31, CE 1 3, CM 1 1, CM 1 2, CM 1 4
- Each student will be pre-assigned to one of the rooms.
 - **YOUR EXAM COPY WILL BE IN THE ROOM WHERE YOU WERE ASSIGNED.**
 - **YOU MUST GO TO THE CORRECT ROOM.**
 - **NO EXTRA EXAMS WILL BE AVAILABLE IN OTHER ROOMS.**

- **Schedule**

- **A few days before:** room assignments and seating chart released on Moodle
- **11:00** on the day of: Exam rooms set up
- **11:20** on the day of: Exam hall doors open and students take their seat.
- **11:30** on the day of: Exam begins
- **During exam:** EPFL ID Cards checked (**don't forget CAMIPRO!**)
- **13:00** on the day of: Exams collected

Midterm Format

- 30 Multiple Choice QA
 - From **lectures** AND **exercise sessions** (up to and including **Week 7**)
 - **No** negative points on MCQA
- Materials
 - 1 double-sided A4 crib sheet allowed
 - Non-programmable calculators allowed
 - CAMIPRO card
- No phones or other electronic devices allowed

Midterm Office Hour

- Class period (13h15 - 14h) on **April 3rd**
- Come with questions!

Reward Estimation

- How should we define a reward function? Just use your **evaluation metric**!
 - **BLEU** (machine translation; Ranzato et al., ICLR 2016; Wu et al., 2016)
 - **ROUGE** (summarization; Paulus et al., ICLR 2018; Celikyilmaz et al., NAACL 2018)
 - CIDEr (image captioning; Rennie et al., CVPR 2017)
 - SPIDEr (image captioning; Liu et al., ICCV 2017)
- Be careful about **optimizing for the task** as opposed to **“gaming” the reward**!
 - Evaluation metrics are merely proxies for generation quality!
 - “even though RL refinement can achieve better BLEU scores, it barely improves the human impression of the translation quality” – Wu et al., 2016

Reward Estimation

- How should we define a reward function? Just use your **evaluation metric**!

**Why would optimising for the evaluation metric
"game" the reward ?**

**Shouldn't the evaluation metric measure the
target behaviour we want (e.g., quality) ?**

the translation quality" – Wu et al., 2016

Judging the quality of generations

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Continuation: The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the **Universidad Nacional Autónoma de México (UNAM)** and **the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/ Universidad Nacional Autónoma de México/ Universidad Nacional Autónoma de México/ Universidad Nacional Autónoma de México...**

**How should we evaluate the
quality of this sequence?**

Perplexity: A first try

- Evaluate quality of the model based on the perplexity of the model on reference sentences

Perplexity: A first try

- Evaluate quality of the model based on the perplexity of the model on reference sentences
- **Why can't we use perplexity of our generated sentences?**

Perplexity: A first try

- Evaluate quality of the model based on the perplexity of the model on reference sentences
- **Why can't we use perplexity of our generated sentences?**
- Decoding algorithms that minimise perplexity (i.e., argmax, beam search) would be advantaged even if they don't produce the best text

Case in point

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Continuation: The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the **Universidad Nacional Autónoma de México (UNAM)** and **the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/ Universidad Nacional Autónoma de México/ Universidad Nacional Autónoma de México/ Universidad Nacional Autónoma de México...**

Perplexity: A first try

- Evaluate quality of the model based on the perplexity of the model on reference sentences
- **Why can't we use perplexity of our generated sentences?**
- Decoding algorithms that minimise perplexity (i.e., argmax, beam search) would be advantaged even if they don't produce the best text
- Perplexity of reference sequences tell us how calibrated our model is to real sequences, but doesn't say much about the generations it produces

How do you think text generation evaluation differs compared to classification evaluation?

A simple dialogue



Are you going to Prof.
Bosselut's CS552 lecture?

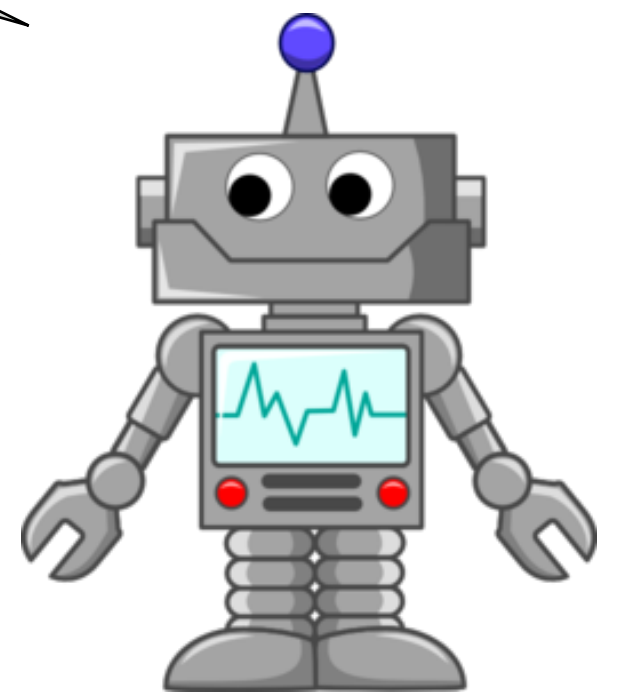
Heck yes !



Yes !

You know it !

Yup .




Any "right" answer you know could be one of many!

Outline

Ref: They walked **to the** grocery **store** .

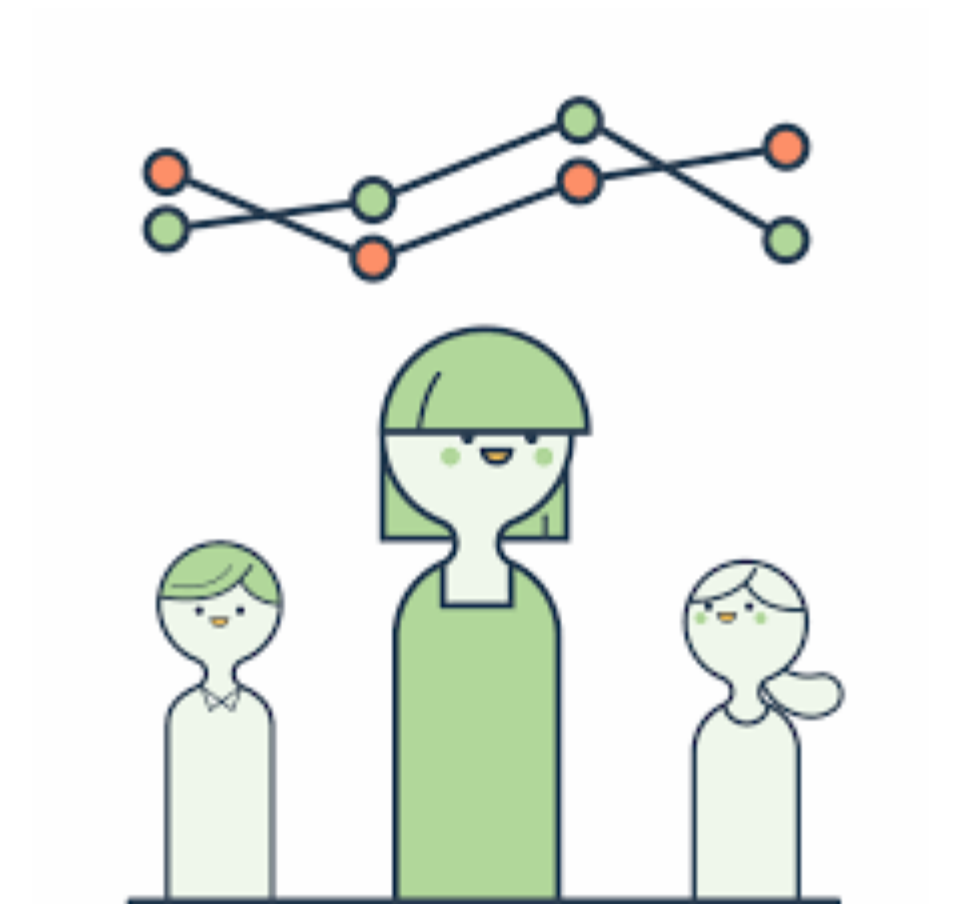
Gen: **The woman** went **to the** **hardware** **store** .



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** .



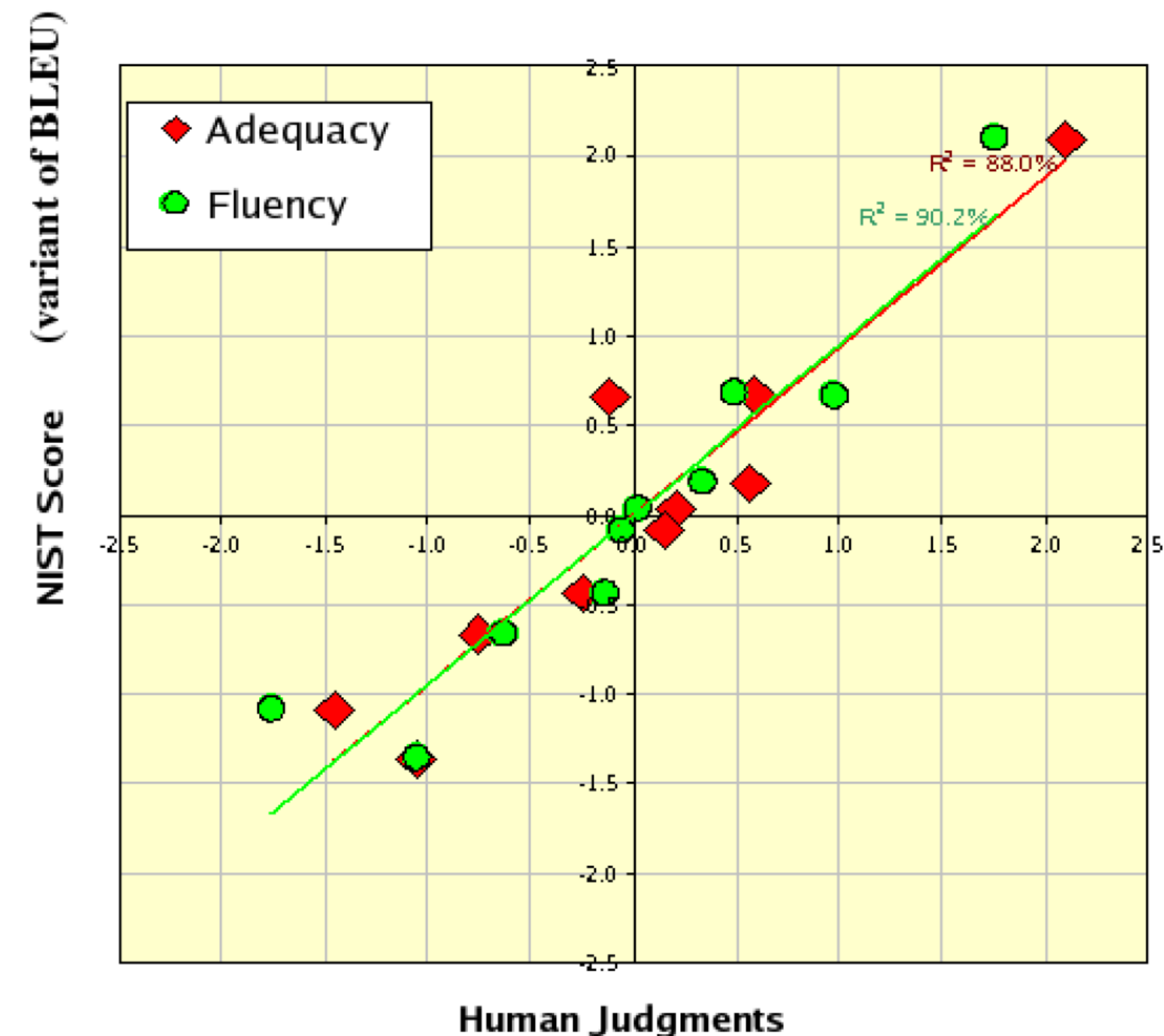
- Compute a score that indicates the similarity between *generated* and *gold-standard (human-written)* text
- Fast and efficient and widely used
- Two broad categories:
 - N-gram overlap metrics (e.g., **BLEU**, ROUGE, METEOR, CIDEr, etc.)
 - Semantic overlap metrics (e.g., PYRAMID, SPICE, SPIDEr, etc.)

N-gram overlap metrics

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

- They're **not ideal for machine translation**, but are correlated with human judgments of quality

Typically, to evaluate the “quality” of an *evaluation metric*, we measure its correlation with human judgments (requires reliable human scale)



A simple failure case



Are you going to Prof.
Bosselut's CS552 lecture?

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

Score:

0.61

0.25

False negative 0

False positive 0.67

Heck yes !

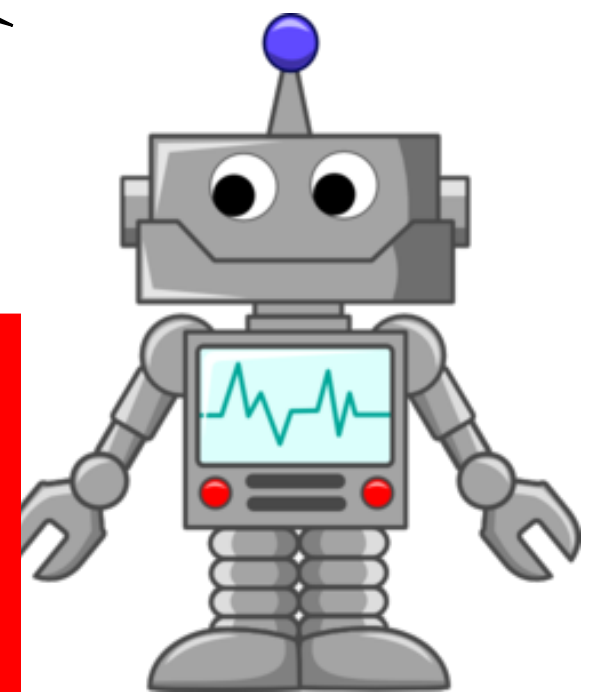


Yes !

You know it !

Yup .

Heck no !



A more comprehensive failure analysis

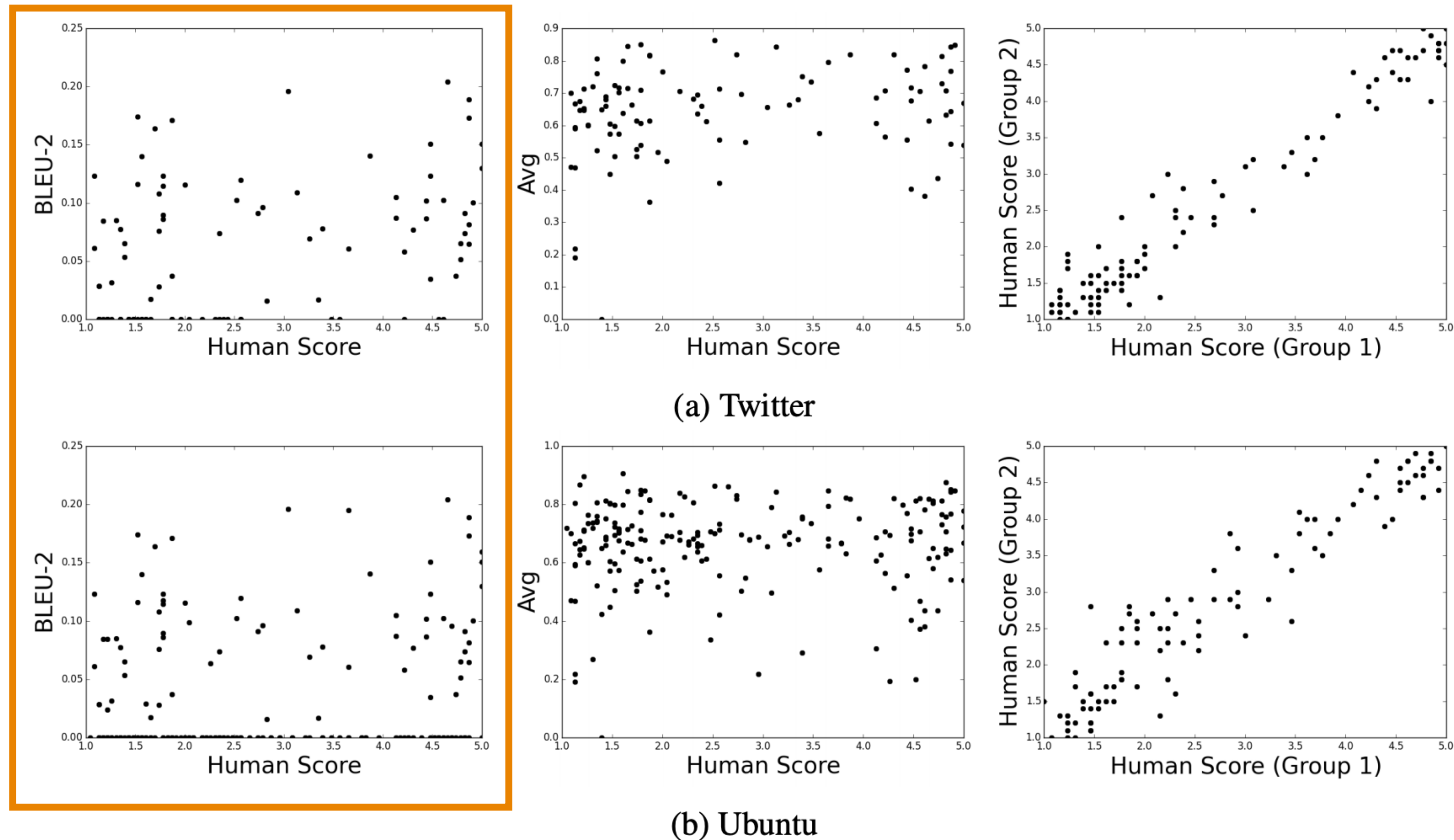


Figure 1: Scatter plots showing the correlation between metrics and human judgements on the Twitter corpus (a) and Ubuntu Dialogue Corpus (b). The plots represent BLEU-2 (left), embedding average (center), and correlation between two randomly selected halves of human respondents (right).

N-gram overlap metrics

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

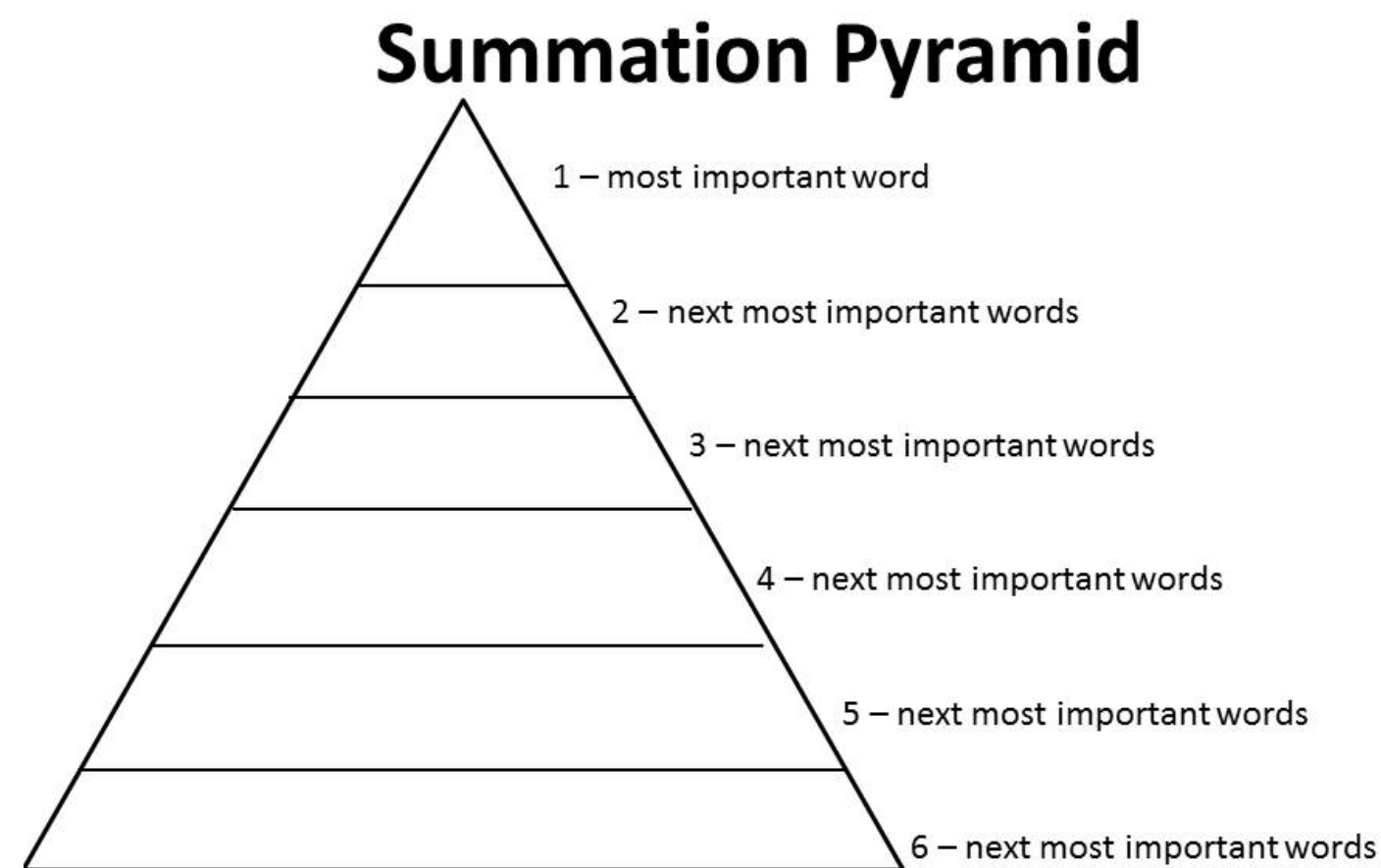
- They're **not ideal for machine translation**
- They get progressively **much worse** for tasks that are more open-ended than machine translation
 - Worse for **summarization**, as longer output texts are harder to measure
 - Much worse for **dialogue**, which is more open-ended than summarization

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 **story generation**, which is also open-ended, but whose sequence length can make it seem you're getting decent scores!

Semantic overlap metrics



PYRAMID:

- Incorporates human content selection variation in summarization evaluation.
- Identifies Summarization Content Units (SCU)s to compare information content in summaries.

(Nenkova, et al., 2007)



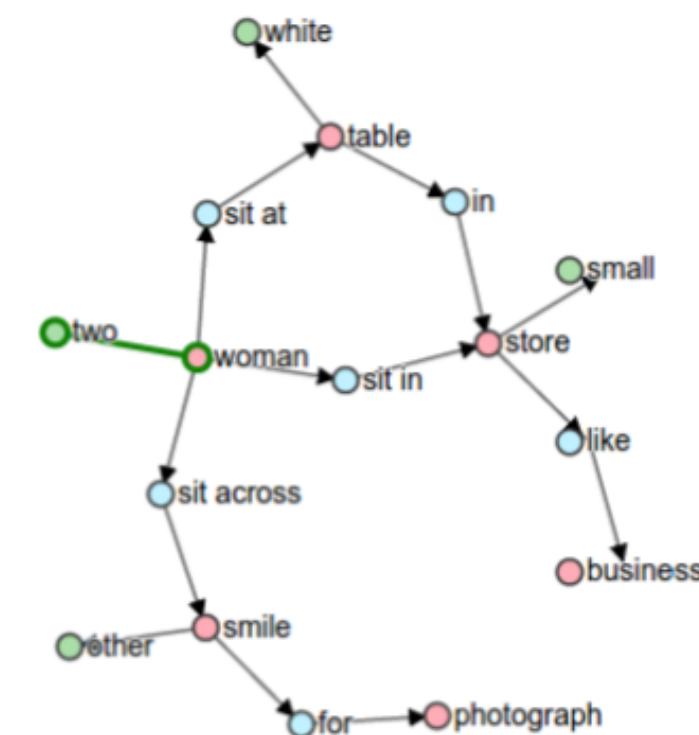
"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

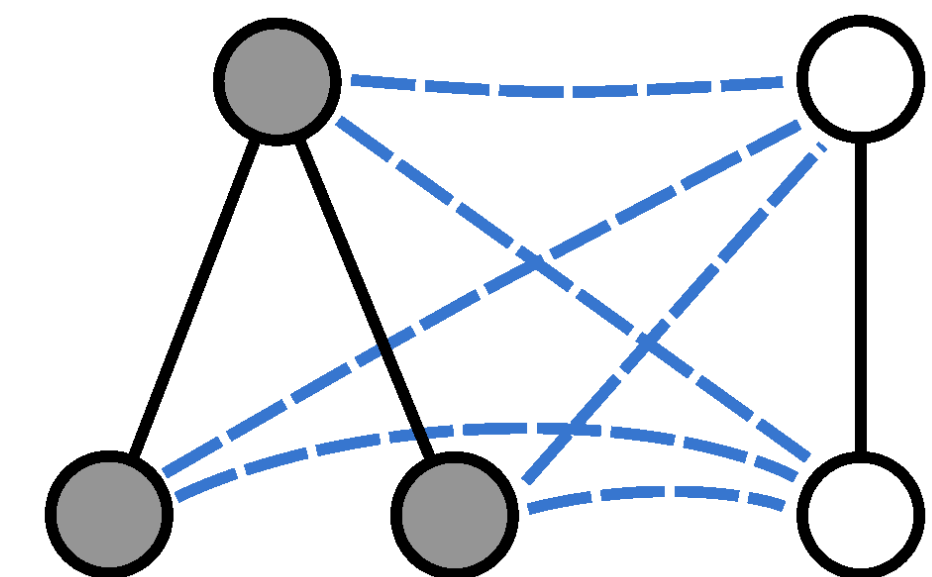
"two woman are sitting at a table"



SPICE:

Semantic propositional image caption evaluation is an image captioning metric that initially parses the reference text to derive an abstract scene graph representation.

(Anderson et al., 2016)



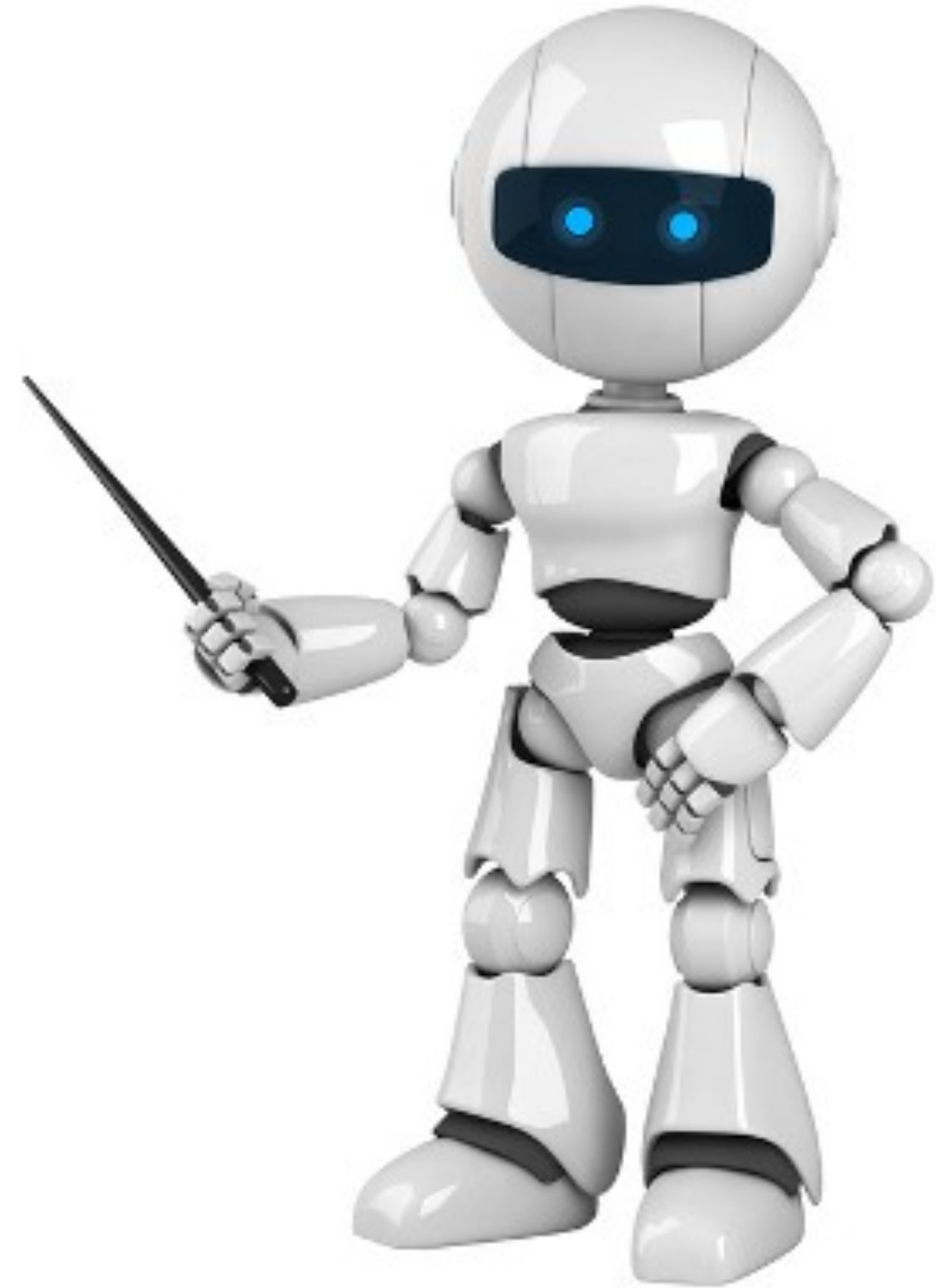
SPIDER:

A combination of semantic graph similarity (**SPICE**) and n -gram similarity measure (**CIDER**), the SPICE metric yields a more complete quality evaluation metric.

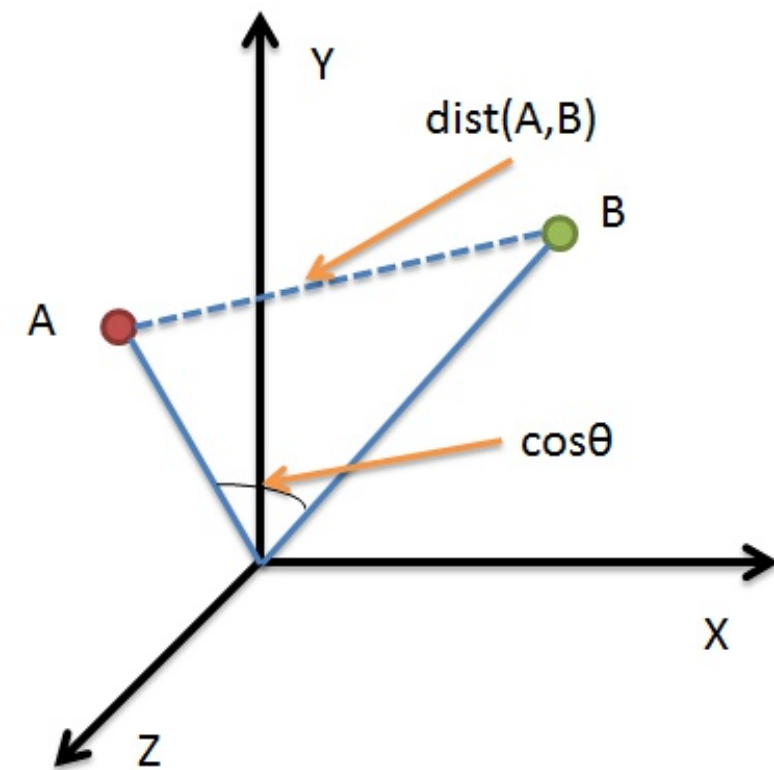
(Liu et al., 2017)

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**



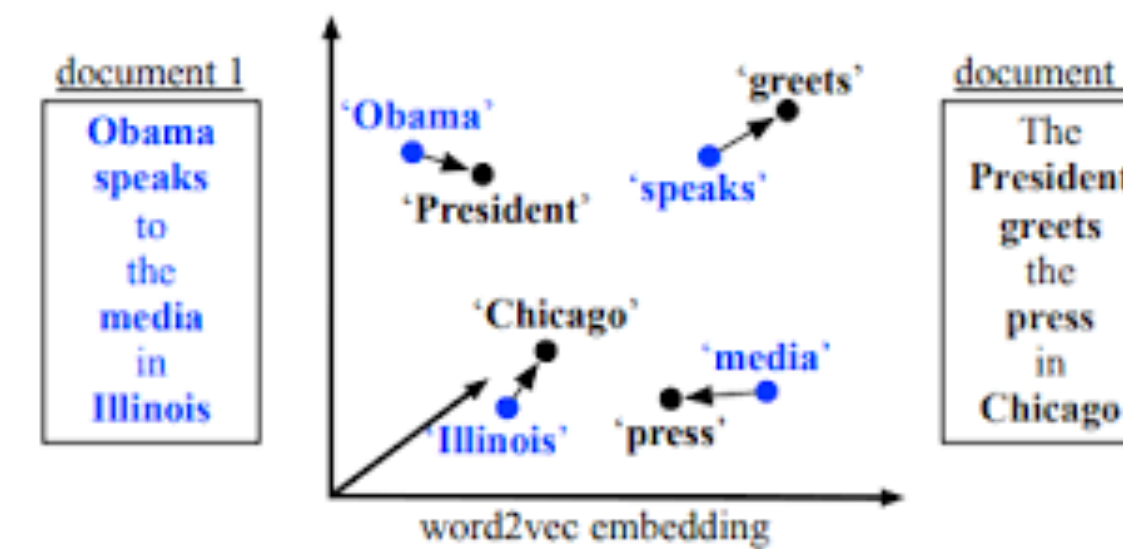
Model-based metrics: Word distance functions



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:

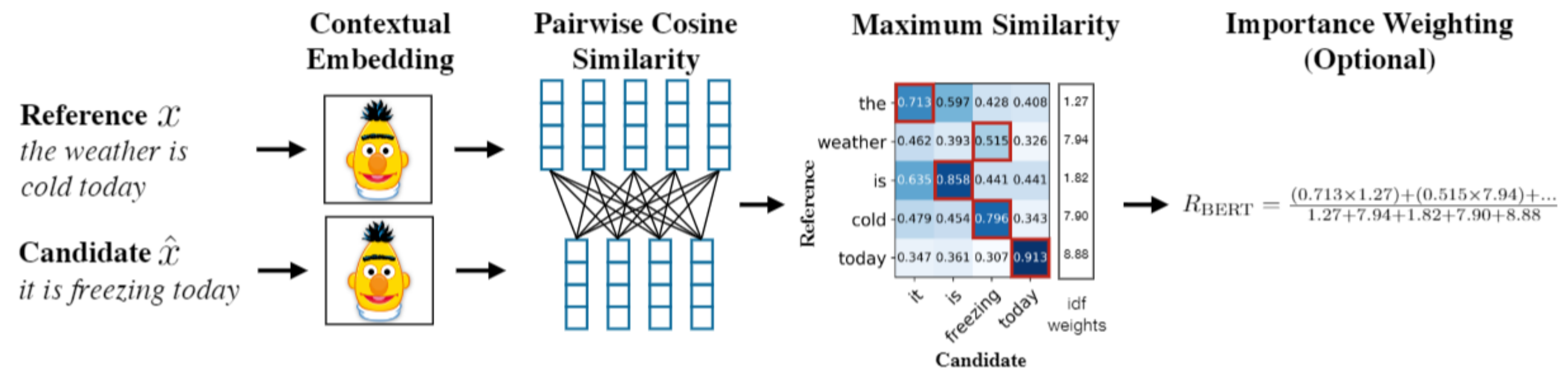
Measures the distance between two sequences (e.g., sentences, paragraphs, etc.), using word embedding similarity matching.

(Kusner et al., 2015; Zhao et al., 2019)

BERTScore:

Use pre-trained contextual embeddings from BERT and match words in candidate and reference sentences by cosine similarity

(Zhang et al., 2020)

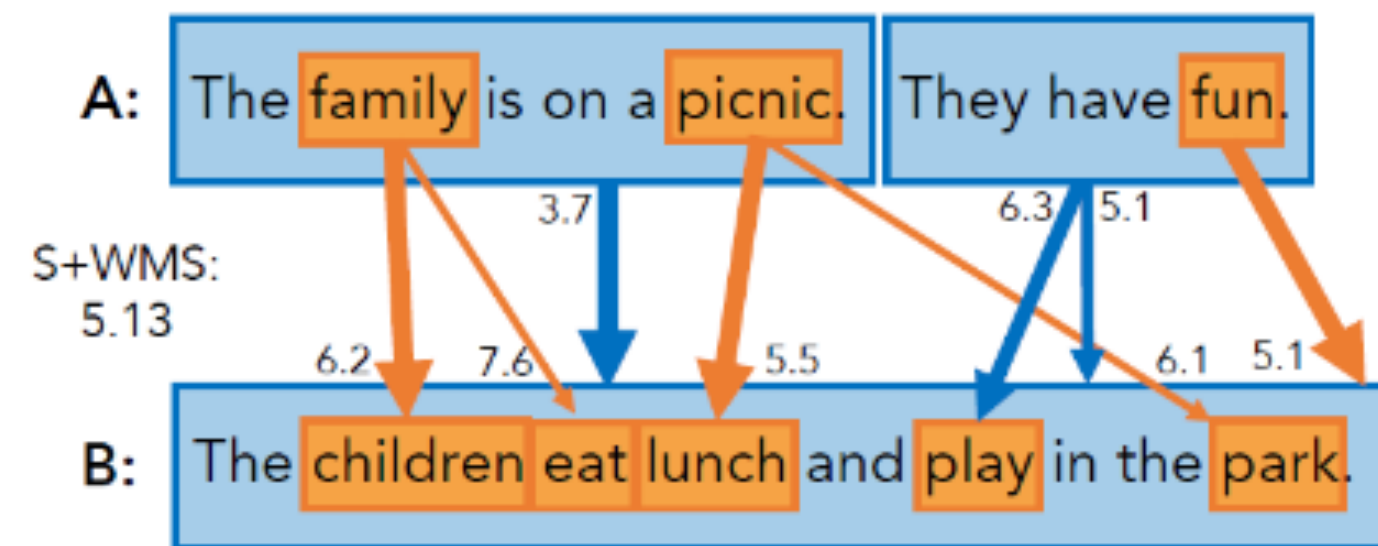


Model-based metrics: Beyond word matching

Sentence Movers Similarity :

Based on Word Movers Distance to evaluate text in a continuous space using sentence embeddings from recurrent neural network representations.

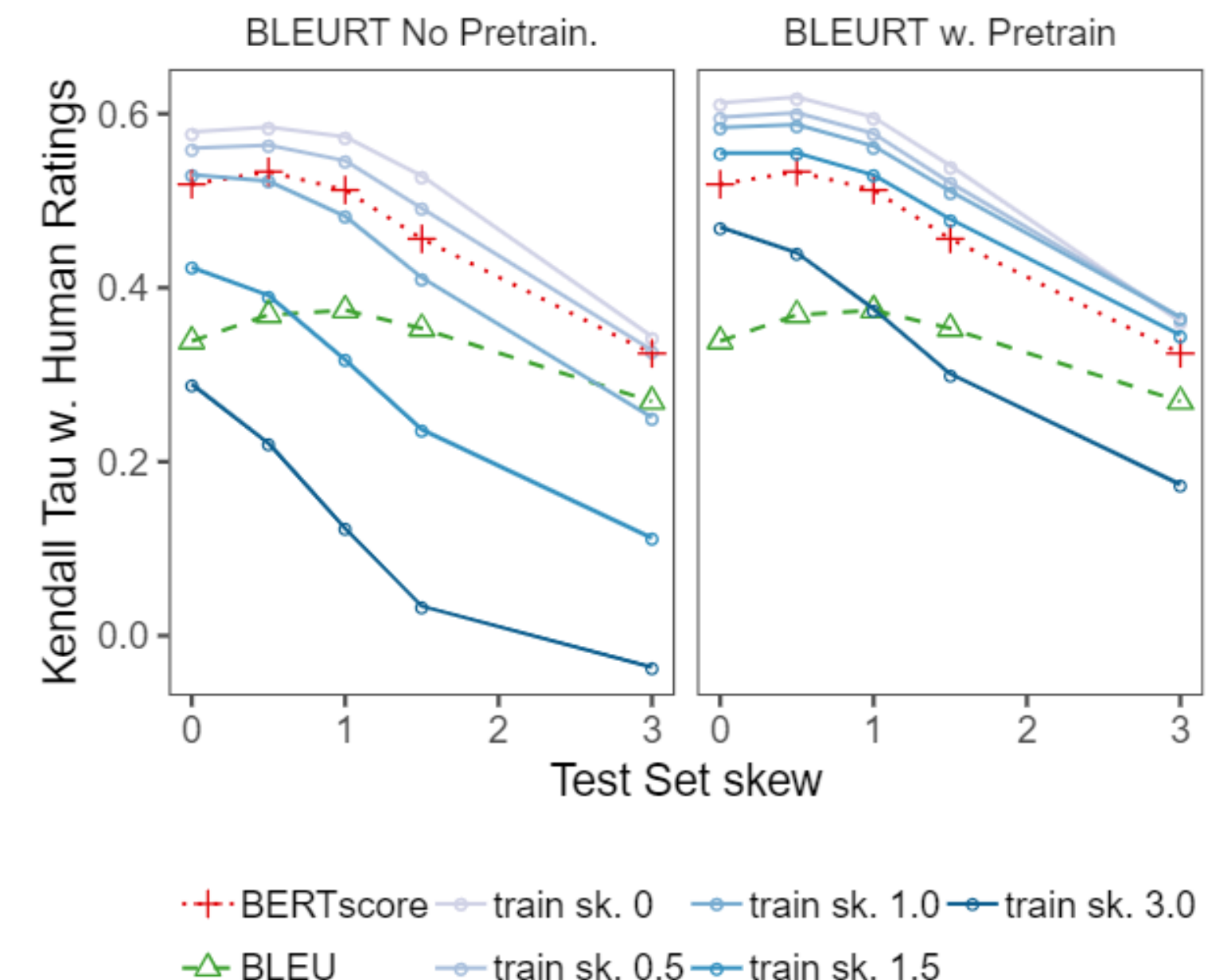
(Clark et.al., 2019)



BLEURT:

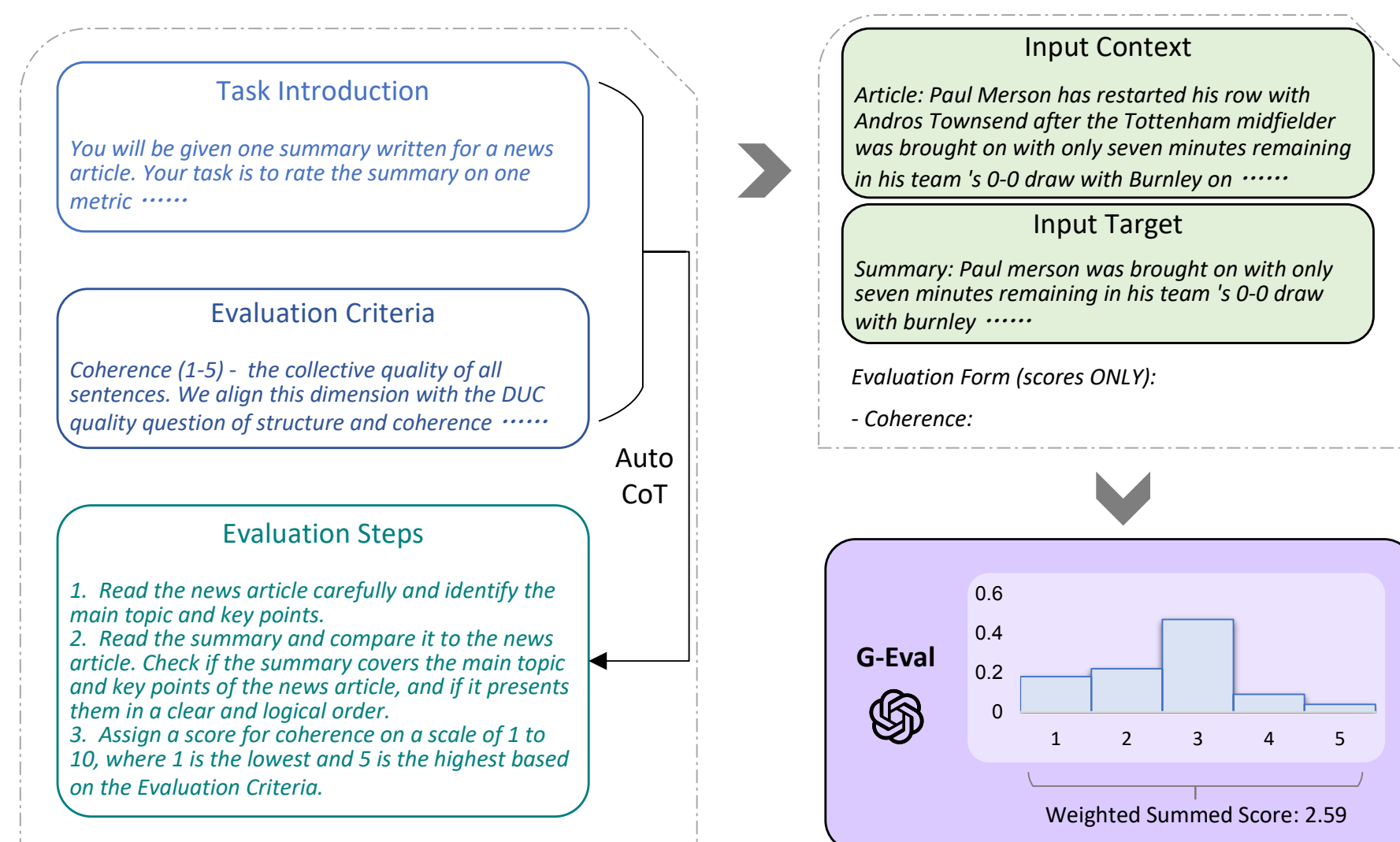
A regression model based on BERT returns a score that indicates to what extent the candidate text is grammatical and conveys the meaning of the reference text.

(Sellam et.al. 2020)



Model-based metrics: LLMs

- Use LLMs to evaluate generation outputs according to clearly defined rubric
 - **G-Eval** (Liu et al., 2023)
 - **LLM-as-a-judge** (Zheng et al., 2023)



[System]
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

[User Question]
{question}

[The Start of Assistant A's Answer]
{answer_a}
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
{answer_b}
[The End of Assistant B's Answer]

[System]
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, please rate the response on a scale of 1 to 10 by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".

[Question]
{question}

[The Start of Assistant's Answer]
{answer}
[The End of Assistant's Answer]

What might be a benefit of model-based metrics compared to overlap metrics?

Human evaluations

- Automatic metrics fall short of matching human decisions
- Most important form of evaluation for text generation systems
 - >75% generation papers at ACL 2019 include 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

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: Never compare human evaluation scores across differently-conducted studies

Even if they claim to evaluate the same dimensions!

Human evaluations: case study



200 Human-written Stories

PROMPT: *You, a video game character, just died. You had no spare lives left.*

STORY: *Link woke in a panic, gasping and flailing wildly at the air above him. The panic quickly turned to confusion, wasn't he just...? He heard footsteps approaching from behind. Link leapt up from the floor and drew his sword, ready to strike. The sight of a slight man in a suit added to Link's already mounting confusion. Is this another trap set by Ganondorf? (...)*

~150 words

200 GPT-2 Generated Stories

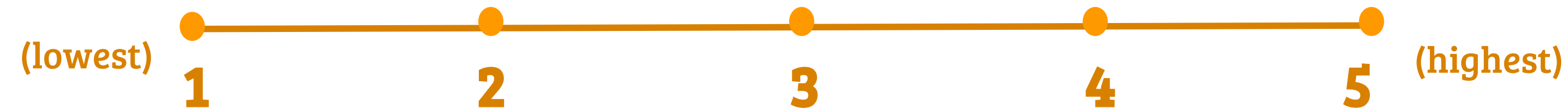


PROMPT: *You, a video game character, just died. You had no spare lives left.*

STORY: **The US Space Station exploded with automatic fire in preparation for tonight's nuclear strike on Earth. This is video from the control center inside the Space Station**
**A figure wearing a manilla envelope falls off of the back of the space station, detonating a nuclear device.*
***An audio file is made of the sirens, and some screams (...)*

Human evaluations: case study

AMAZON MECHANICAL TURK



GRAMMAR

How grammatically correct is the text of the story fragment?

COHERENCE

How well do the sentences in the story fragment fit together?

LIKABILITY

How enjoyable do you find the story fragment?

RELEVANCE

How relevant is the story fragment to the prompt?

Human evaluations: case study

AMAZON MECHANICAL TURK

Evaluating Machine-Generated Text



GPT-2

1. Rating Only GPT-2 Generated Stories



Type of text	Grammar		Coherence		Relevance		Likability	
	Mean _{STD}	IAA%	Mean _{STD}	IAA%	Mean _{STD}	IAA%	Mean _{STD}	IAA%
<i>AMT workers fail to effectively distinguish between human written and GPT-2 generated stories</i>								
Ref. (Day 1)	4.00 _{0.92}	0.21 _{15.5}	4.11 _{0.96}	0.14 _{16.5}	3.71 _{1.26}	0.27 ₁₀	3.37 _{1.18}	0.11 _{7.5}
Ref. (Day 2)	3.86 _{0.92}	-0.03 _{10.5}	3.92 _{0.98}	-0.03 _{6.5}	3.71 _{1.08}	0.02 ₁₁	3.73 _{0.97}	-0.04 _{8.5}
Ref. (Day 3)	3.98 _{0.96}	0.18 ₁₁	4.05 _{0.94}	0.13 _{10.5}	3.46 _{1.29}	0.26 ₈	3.42 _{1.16}	0.07 _{4.5}
GPT-2	3.94 _{0.93}	0.11 _{17.5}	3.82 _{1.12}	0.05 _{7.5}	3.44 _{1.41}	0.10 ₇	3.42 _{1.25}	0.02 _{4.5}

Human evaluation: Issues

- Human judgments are regarded as the **gold standard**
- Human evaluation is **slow** and **expensive**

Suppose you can run a human evaluation

Do we have anything to worry about?

Human evaluation: Issues

AMAZON MECHANICAL TURK

Time Spent on the Task



HUMAN



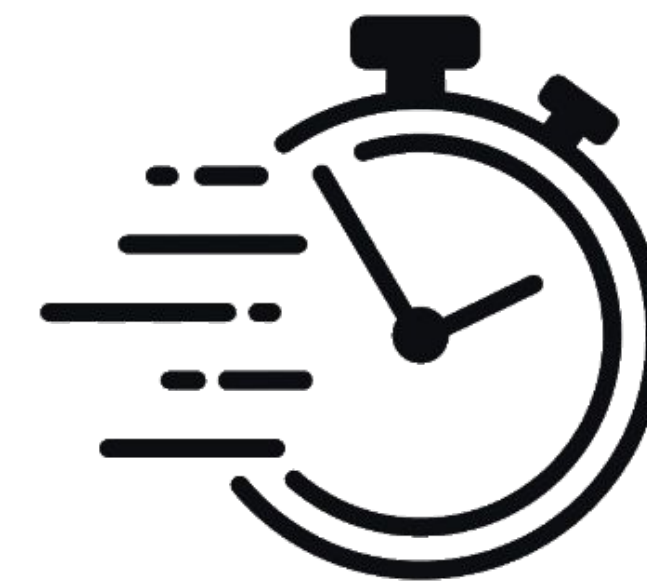
360 sec

WorkTimeInSeconds



22 sec

Mean



13 sec

Median

Human evaluation: Issues

ENGLISH TEACHERS

Post-Task Interviews

- Need **10–20 examples** to calibrate ratings
- **Coherence** was the easiest to rate for human-written stories
- **Coherence** was also the most challenging to rate for GPT-2 stories
- **Relevance** was the easiest to rate for GPT-2 stories (clearly not following the prompt)
- Overall **GPT-2** generated stories were **difficult to rate** (average time per story raised from **69.8s** → **87.3s**)
- Preferred to rate **GPT-2** and **human-written** stories **together** (better calibration)
- Suggested to employ a **rubric**



GPT-2+HUM



Human evaluation: Issues

- Human judgments are regarded as the **gold standard**
- Human evaluation is **slow** and **expensive** (compared to automatic evaluation), even if your humans try to speed it up!
- Conducting effective human evaluations is difficult

Humans:

- are inconsistent
- can be illogical
- lose concentration
- misinterpret your question
- can't always explain why they feel the way they do
- May try to speed through your evaluation

Evaluation: Takeaways

- *Content overlap metrics* provide a good starting point for evaluating the quality of generated text, 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 ones that can directly evaluate *factuality* – is the model saying correct things?
 - **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! — “Vibes-based evaluation”**

Concluding Thoughts

- 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
- With the advent of large-scale language models, deep NLG research has been reset
 - it's **never been easier to jump in the space!**
- One of the **most exciting areas** of NLP to work in!