FOLT Lecture 8, NLG Evaluation

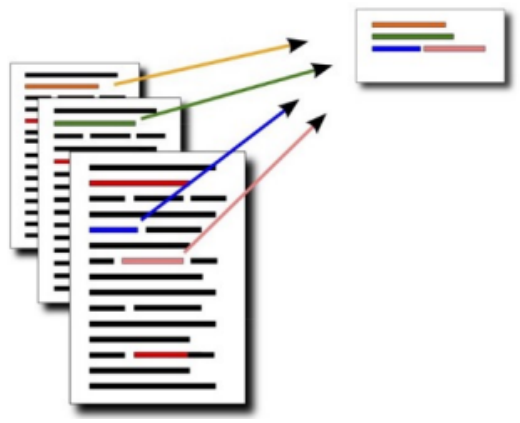
PART1: Intro

Definition: Natural Language Generation (NLG) refers to the process of automatically generating human-understandable text in one or more natural languages.

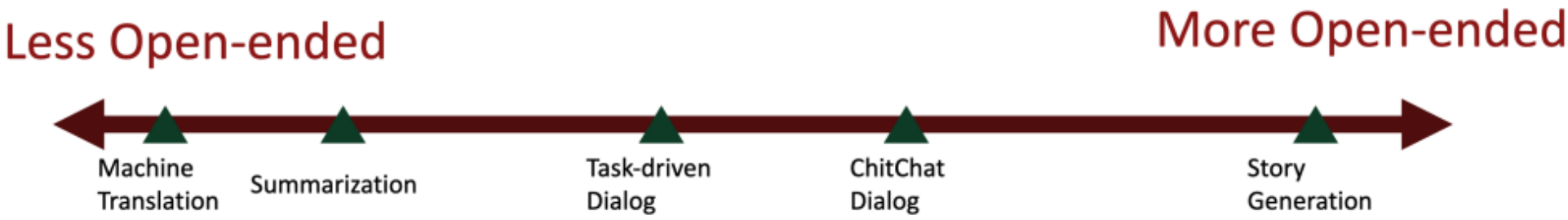
Machine Translation

Dialogue Systems

Multi-document summarization



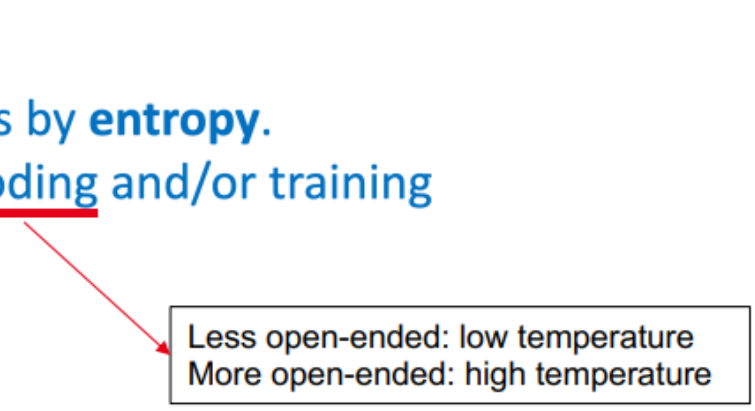
1)Categorization of NLG Tasks



Open-ended generation: the output distribution still has high freedom

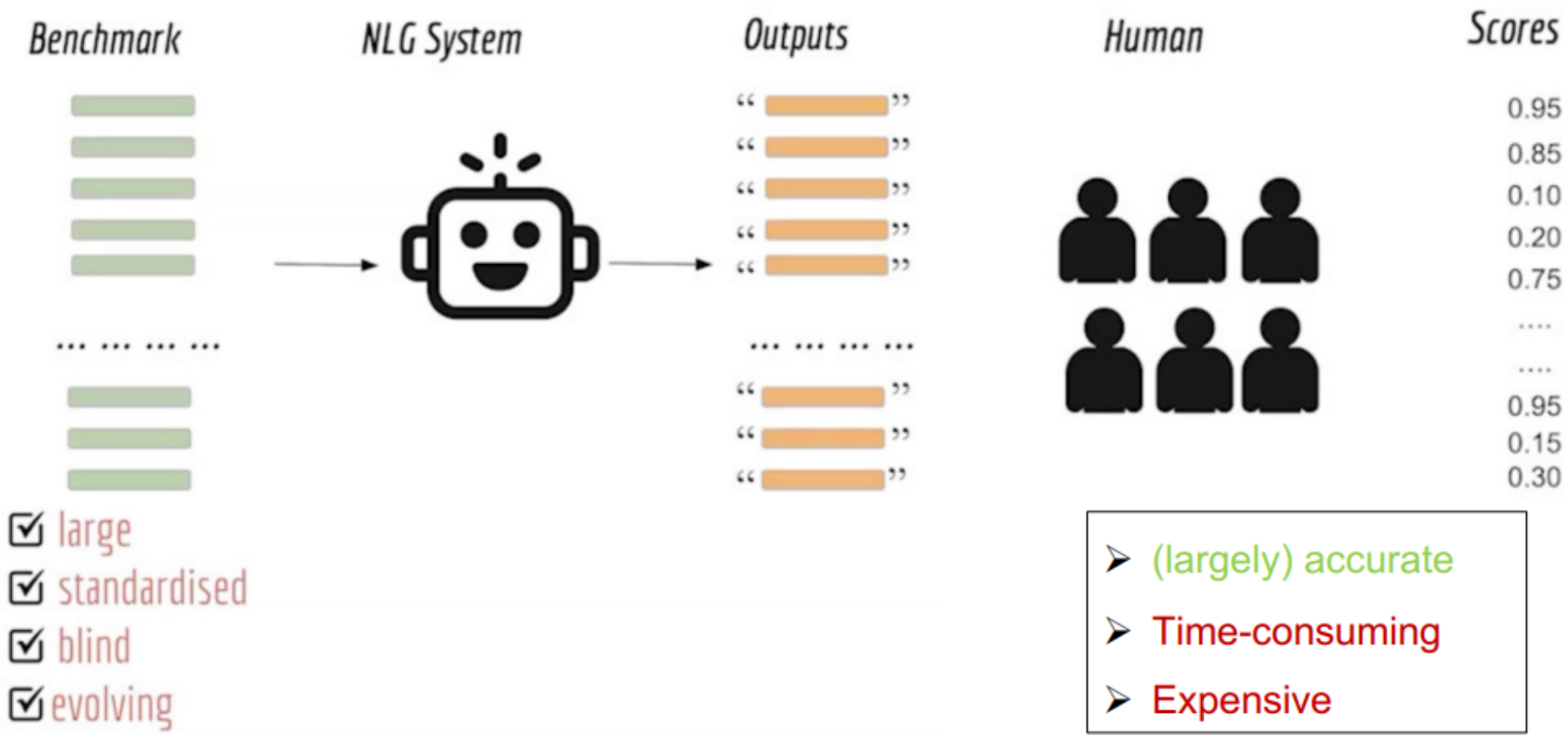
Non-open-ended generation: the input mostly determines the output generation.

Remark: One way of formalizing categorization this is by **entropy**.
These two classes of NLG tasks require different decoding and/or training approaches!

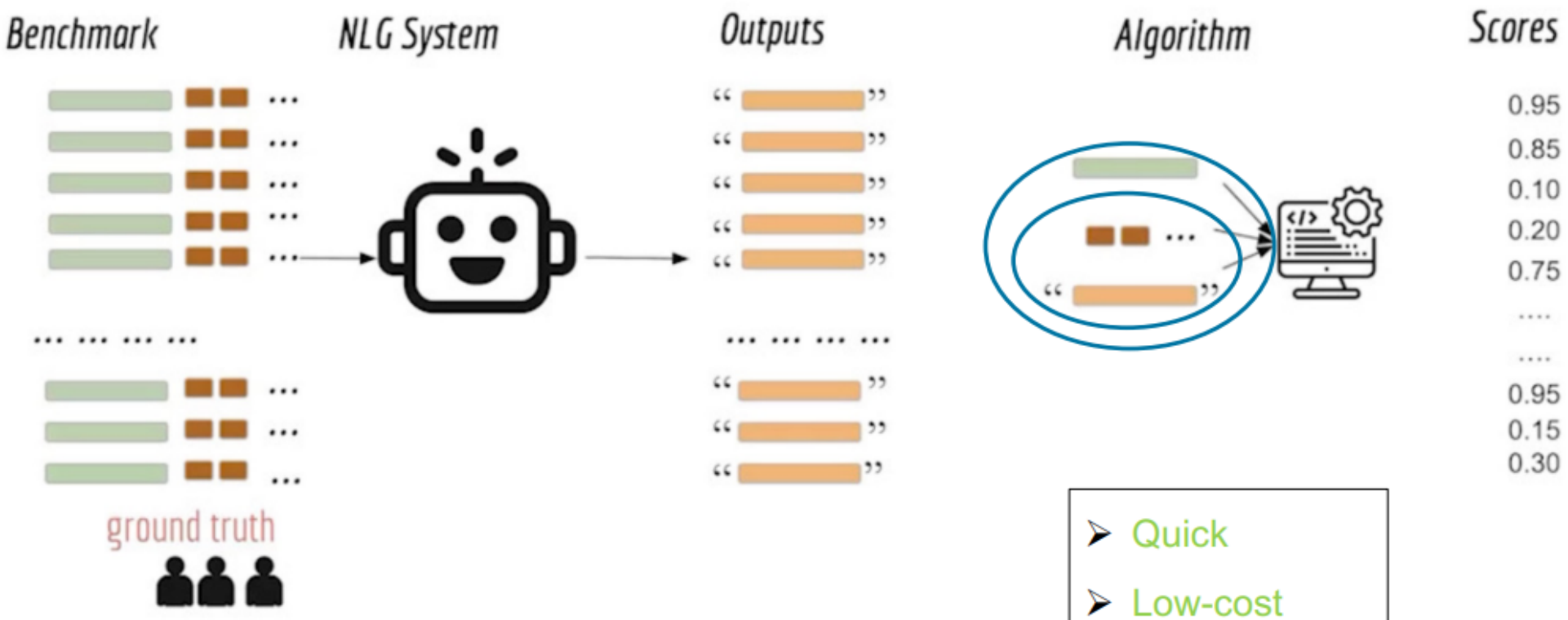


2) Quantifying NLG Progress

The Ideal way: human evaluation



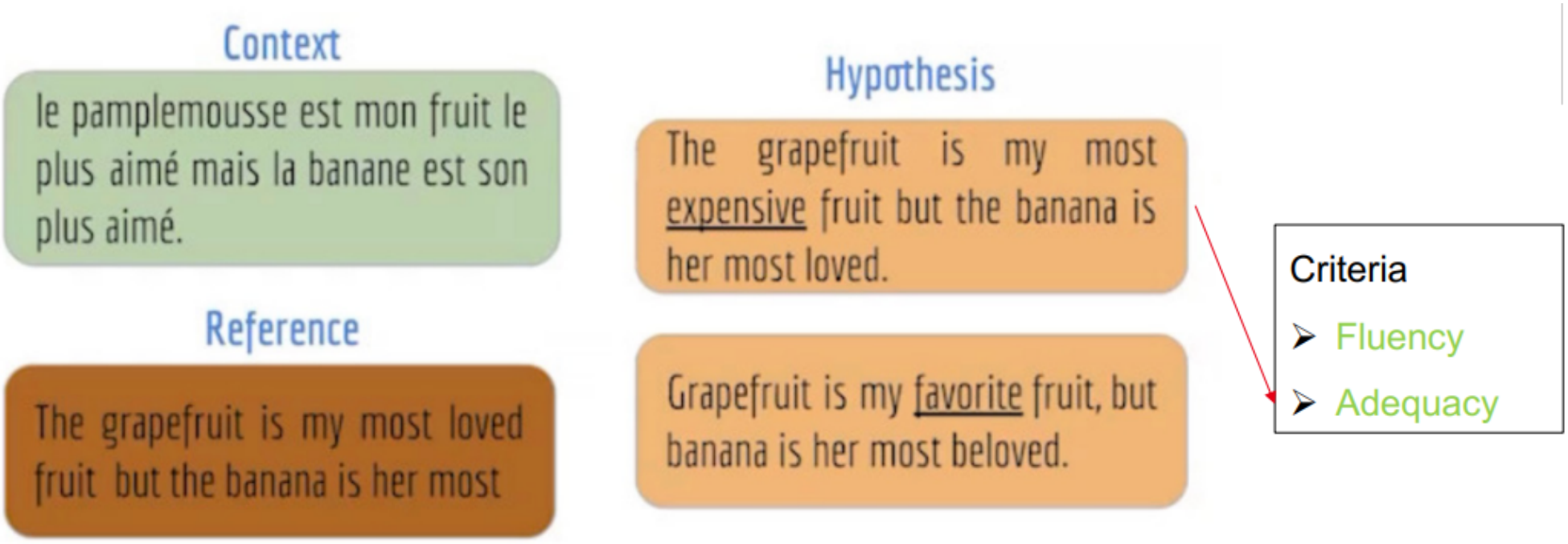
The practical way : Automatic evaluations



choosing between the two ways is trade-off

PART2:Challenges in evaluating NLG systems

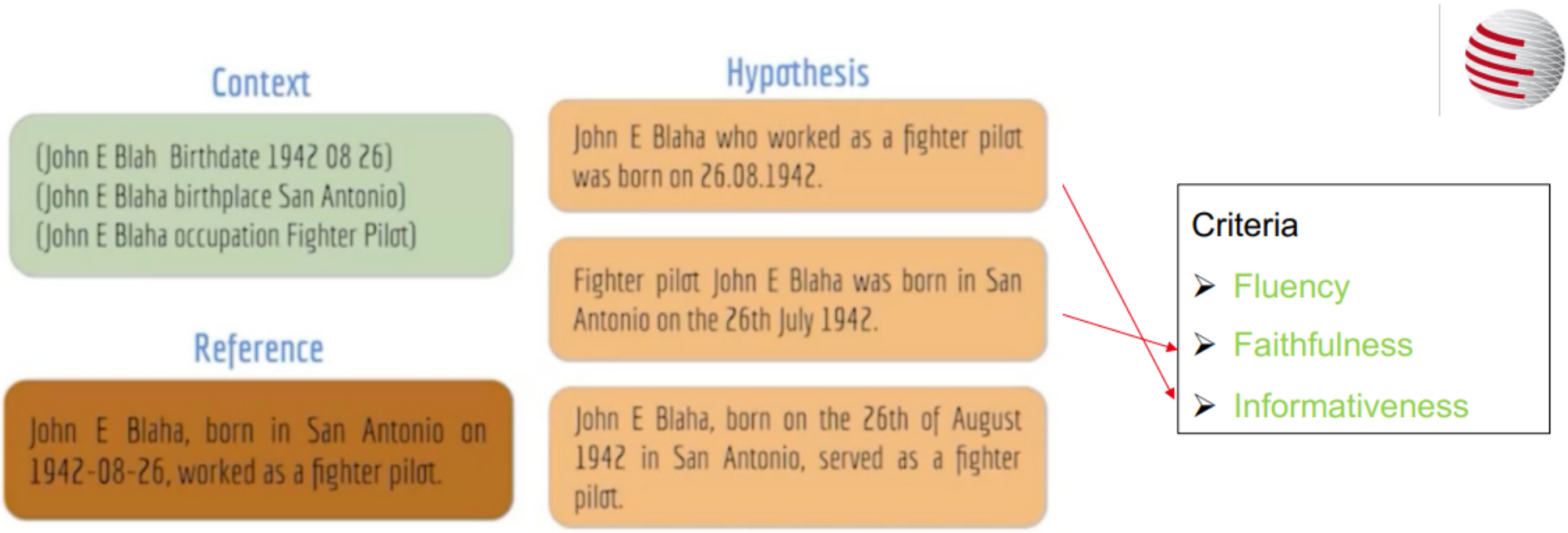
1) The Challenges - Machine Translation



2) The Challenges - Automatic Summarization

3) The Challenges – Dialogue Generation

4)The Challenges – Data to Text Generation



The Challenges

- Multiple correct answers
- Word overlap is neither necessary nor sufficient
- Very diverse criteria within and across tasks
- Require domain and common-sense knowledge
- Require linguistic knowledge (syntax, semantics, co-reference, etc.)
- Require parsing the context (coherence, consistency, fact checking)
- The score must be interpretable (poor fluency vs. poor coherency?)

PART3: NLG evaluation methods

- Human evaluation
- Untrained Automatic Evaluation Metrics
- Machine-learnt Evaluation Metrics
- Factual Consistency Metrics

1) Human Evaluations

- Most important form of evaluation for NLG systems
- Automatic metrics fall short of replicating human decisions
- Gold standard in developing new automatic metrics
  Issues: Expensive, Time-Consuming, Quality Control, Challenging criteria, Inconsistency in Evaluations, Inconsistency in reporting

a) Intrinsic Human evaluation

- Ask humans to evaluate the quality of generated text
- Overall or along some specific dimension:
  o fluency
  o coherence
  o factuality and correctness
  o adequacy o common sense
  o style / formality
  o grammaticality
  o typicality
  o redundancy

b) Human Evaluations: Extrinsic Human Evaluations

→ Humans evaluate a system’s performance on the task for which it was designed
→ For instance, dialogue systems are typically evaluated extrinsically!

Turn Level	Dialog Level
▪ Interesting ▪ Engaging ▪ Generic/Specific ▪ Relevant ▪ Semantically appropriate ▪ Understandable ▪ Fluently Written ▪ Correct vs. Misunderstanding ▪ Overall Impression	• Coherent • Recovers from errors • Consistent • Diversity in its responses • Topic Depth • Likable (empathy, personality) • Understanding • Flexible and adaptable • Informative • Inquisitive • Overall Impression

c) Other Aspects

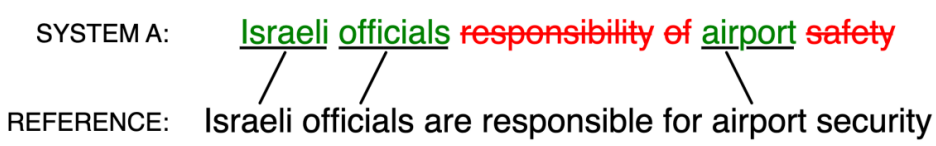
- Evaluators
- Inter-Annotator Agreement
- Evaluation experiment design
  - o Percent agreement, Cohen's kappa (Lecture 2), Fleiss's kappa, Krippendorff's alpha
  - o Side by side or singleton?
  - o The amount context (e.g., dialog or summarization)
  - o How many models to compare at a given time?

2) Untrained Automatic Evaluation Metrics

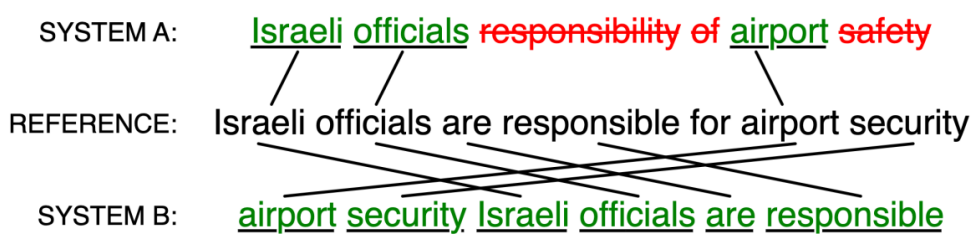
→ Measure the effectiveness of the models that generate text
→Compute a score that indicates the similarity between generated and gold standard (human written) text
→Fast and efficient and widely used

a)N-gram overlap metrics

a1) F-score



- Precision $\frac{correct}{output\text{-}length} = \frac{3}{6} = 50\%$
- Recall $\frac{correct}{reference\text{-}length} = \frac{3}{7} = 43\%$
- F-measure $\frac{precision \times recall}{(precision + recall)/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$



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

flaw: no penalty for reordering

a2) BLEU

→ The Bilingual Evaluation Understudy (BLEU) is one of the first metrics used to measure the similarity between two sentences
→ It compares a candidate translation of text to one or more reference translations

→ It is a weighted geometric mean of n-gram (1-4) precision scores
→ Difficult to interpret absolute scores, useful to compare two systems: typically computed over the entire corpus, not single sentences
→ Add brevity penalty (for too short translations)

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

Cumulative BLEU-4

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)	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%

→ To account for variability, use multiple reference translations
o n-grams may match in any of the references
o closest reference length used

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

a3) ROUGE

→ Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004)
→ Compared to BLEU, ROUGE focuses on recall rather than precision and is more interpretable than BLEU
→ A set of metrics for evaluating automatic summarization of long texts consisting of multiple sentences or paragraph
o ROUGE-{1/2/3/4} measures the overlap of unigrams/bigrams/trigrams/four-grams (single tokens) between the reference and hypothesis text (e.g., summaries)
o ROUGE-L measures the longest matching sequence of words using longest common sub-sequence (LCS)

$$R_{lcs} = \frac{LCS(X,Y)}{m} \quad (2)$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \quad (3)$$

$$F_{lcs} = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \quad (4)$$

- LCS(X,Y) is the length of a longest common subsequence of X and Y
- One advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order as n-grams.

- Ref: The cat is on the mat.
- System output: The cat and the dog.

	Score	Overlap n-grams	N-gram list in Ref or System output
ROUGE-1 Precision	3/5 = 0.6	{the, cat, the}	{the, cat, and, the, dog}
ROUGE-1 Recall	3/6 = 0.5	{the, cat, the}	{the, cat, is, on, the, mat}
ROUGE-1 F-score	2*(p*r)/(p+r) = 0.55		
ROUGE-2 Precision	1/4= 0.25	{the cat}	{the cat, cat and, and the, the dog}
ROUGE-2 Recall	1/5 = 0.2	{the cat}	{the cat, cat is, is on, on the, the mat}
ROUGE-2 F-score	2*(p*r)/(p+r) = 0.22		
ROUGE-L Precision	3/5 = 0.6	**{the cat the}**	{the, cat, and, the, dog}
ROUGE-L Recall	3/6 = 0.5	**{the cat the}**	{the, cat, is, on, the, mat}
ROUGE-L F-score	2*(p*r)/(p+r) = 0.55		

b) Problem with N-gram overlap metrics

- **Lack of semantic understanding:** N-gram metrics primarily focus on surface-level text features (like word overlap) and do not capture the deeper semantic meaning of the text. -> Two sentences can have a high n-gram overlap but differ significantly in meaning.
- **Poor correlation with human judgement:** these metrics often do not correlate well with human judgments of text quality, especially in tasks that require a high degree of creativity or subjectivity.
- **Sensitivity to Paraphrasing:** N-gram overlap metrics might underrate text that conveys the same meaning but uses different wording or phrasing.
- **Inability to Evaluate Novel Text:** In creative text generation tasks, these metrics may penalize novel or creative outputs that do not closely align with reference texts, even if these outputs are valid and high-quality.

c) General notes

- 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!
- The embeddings are **pretrained**, distance metrics used to measure the similarity can be **fixed**

3) Machine-learnt Evaluation Metrics

a) BERTSCORE (Zhang et al., 2020)

- Leverages the pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity.
- Computes precision, recall, and F-score, which are useful for evaluating a range of NLG tasks.
- It has been shown to correlate well with human judgments on sentence-level and system-level evaluations.

evaluations.

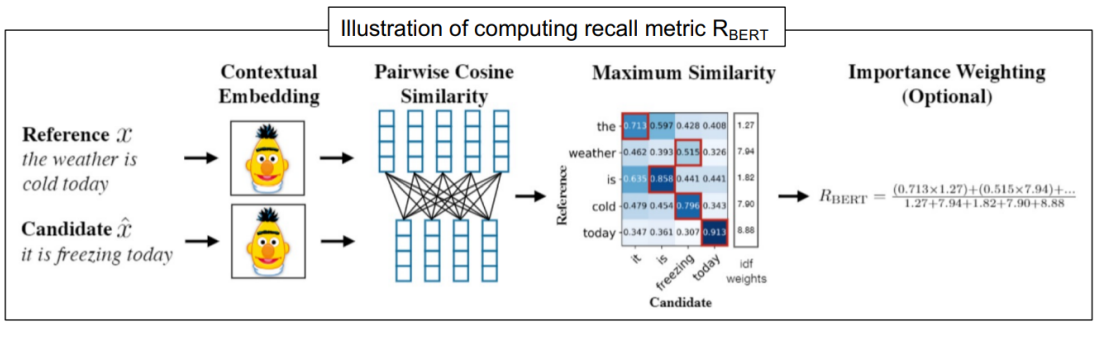
$$R_{\text{BERT}} = \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \quad , \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \quad , \quad F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

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

M ref sentences $\{x^{(i)}\}_{i=1}^{M}$

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

Illustration of computing recall metric R_{BERT}

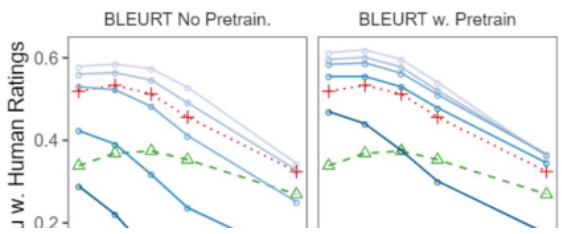


b) BLEURT

- 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.

- A checkpoint from BERT is taken and **fine-tuned on synthetically generated sentence pairs** using **automatic evaluation scores such as BLEU or ROUGE**, and then further fine-tuned on system generated outputs and human-written references using human ratings and automatic metrics as labels.
  - **Masking filling with BERT**: inserting masks at random positions in the Wikipedia sentences, and fill them with the BERT language model
  - **Backtranslation**: generate paraphrases and perturbations with backtranslation, that is, round trips from English to another language and then back to English with a translation model
  - **Dropping words**: randomly drop words from the synthetic examples above to create other examples

- The fine-tuning of BLEURT on synthetic pairs is an important step because it improves the robustness to quality drifts of generation systems.

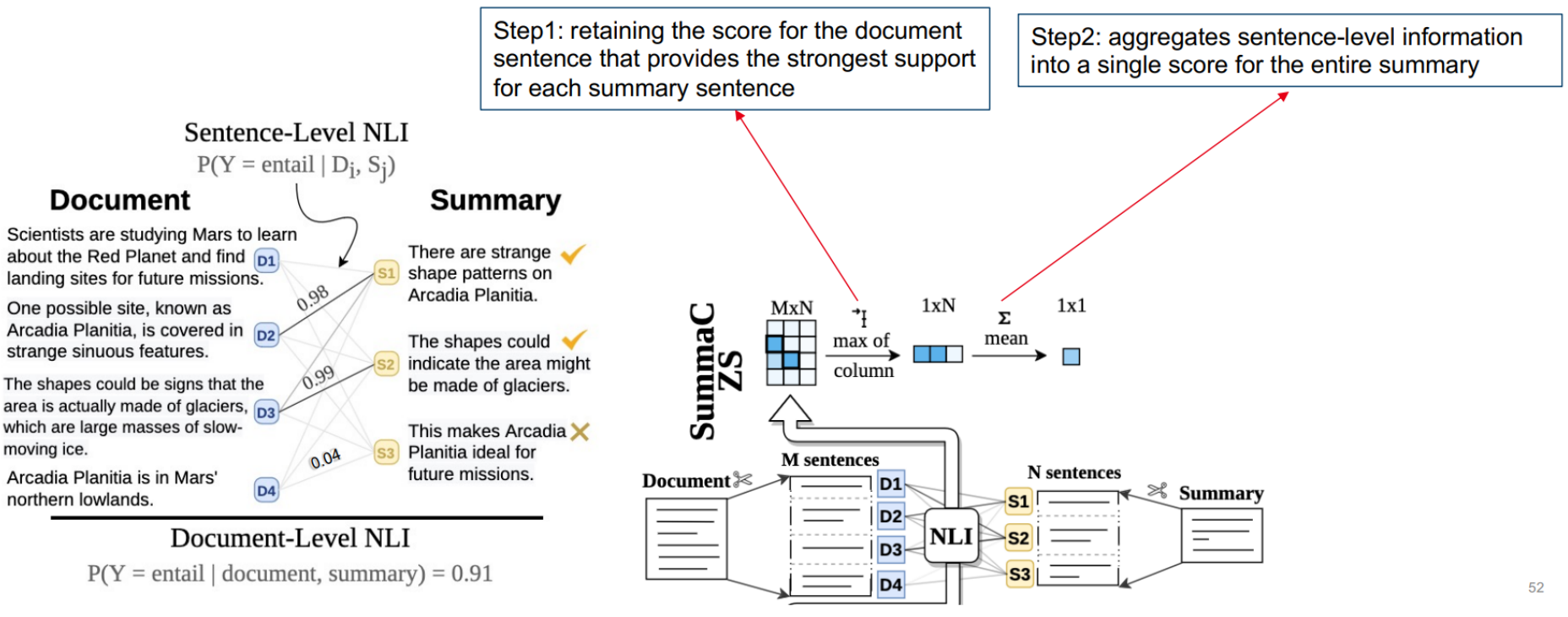


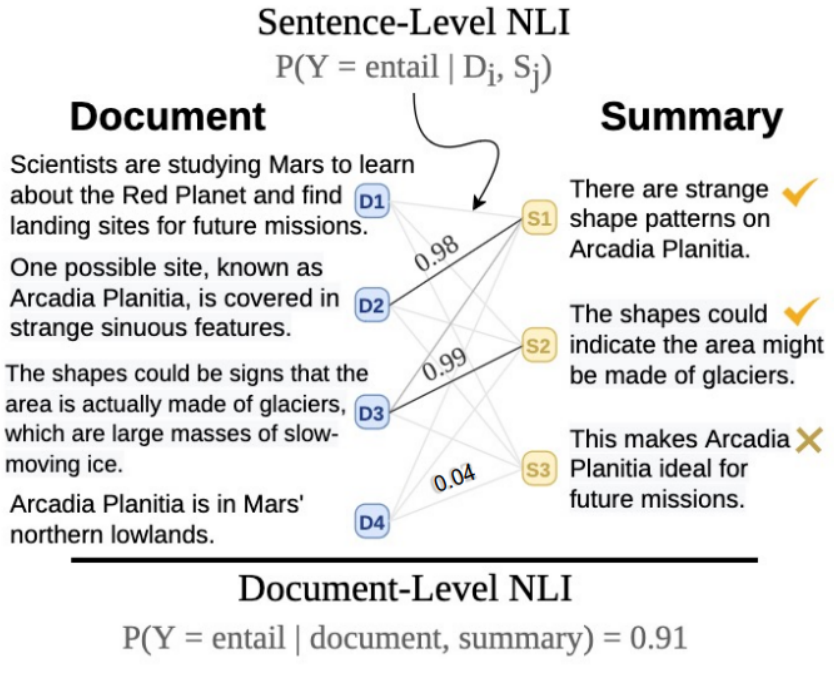
4) Factual Consistency Metrics

- Current pretrained language models are generating increasingly convincing text .However, the generated text is often factually incorrect

a) SummaC

- An approach for inconsistency detection based on the aggregation of sentence-level entailment scores for each pair of input document and summary sentences.

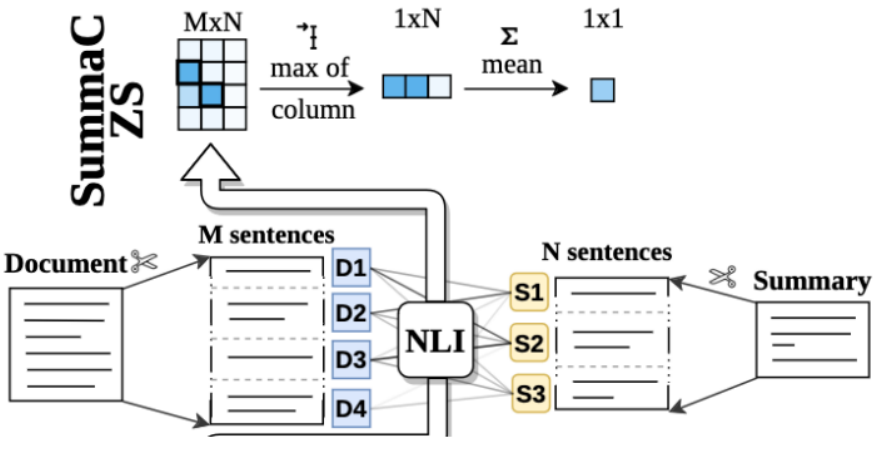




Quiz

$$X_{pair} = \begin{bmatrix} 0.02 & 0.02 & 0.04 \\ 0.98 & 0.00 & 0.00 \\ 0.43 & 0.99 & 0.00 \\ 0.00 & 0.00 & 0.01 \end{bmatrix}$$

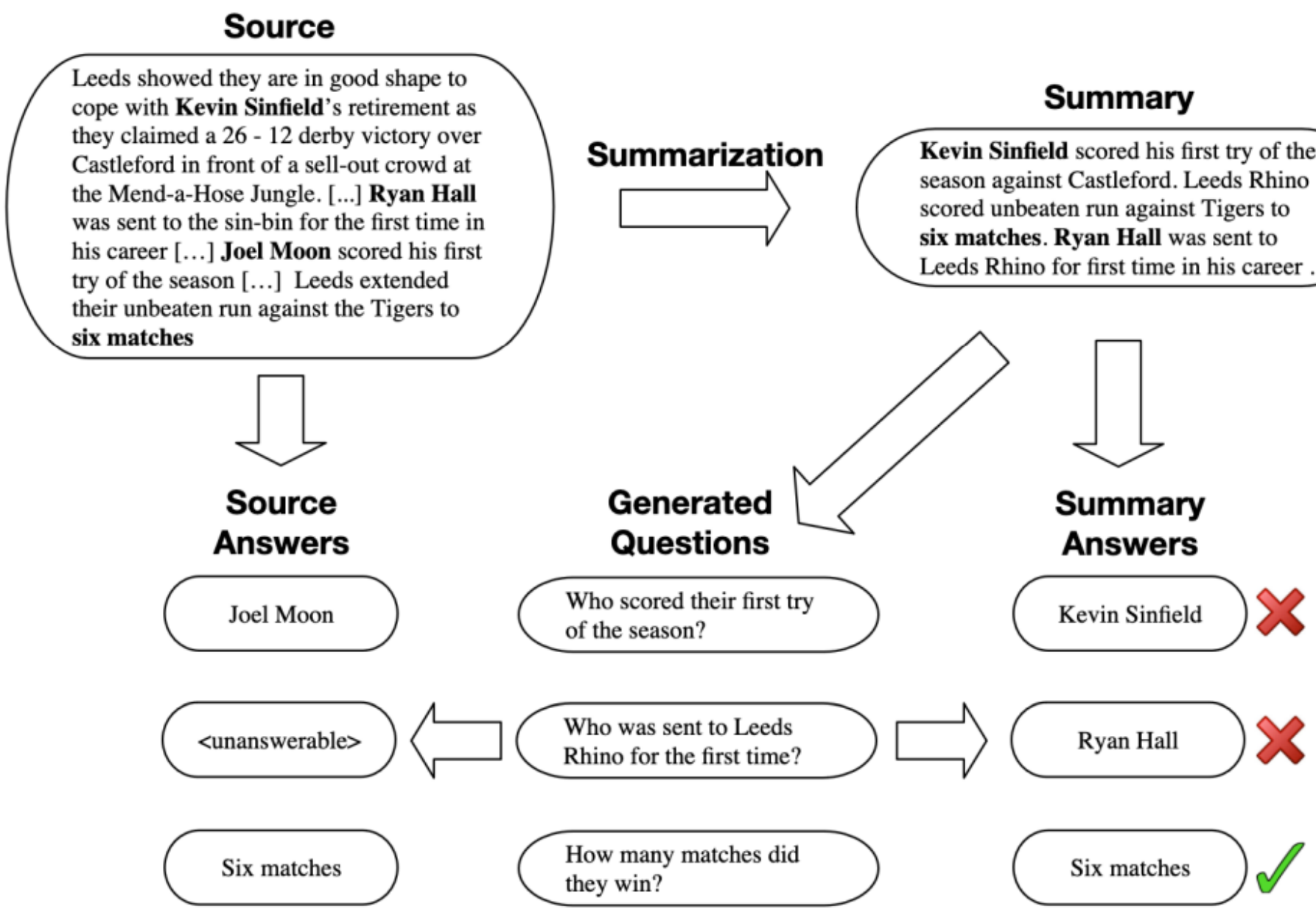
SUMMACZS =?



Answer the quiz as practice

b) QAGS

A question-answering and generation based automatic evaluation protocol that is designed to identify factual inconsistencies in a generated summary. They use fairseq for generation and BERT for QA model as a backbone.



Part 4 : Limitations and emergent directions of the current NLG evaluation metrics

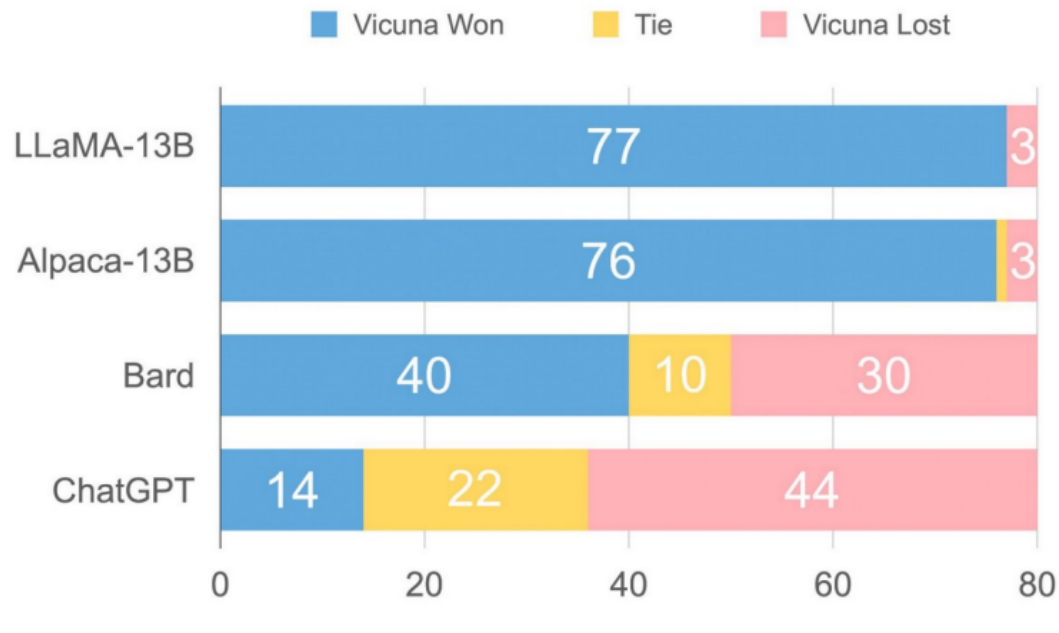
Note : I think we can read this part and understand it no need to memorize everything

- **N-gram content overlap metrics** provide a good starting point for evaluating the quality of generated text, but they're **not good enough on their own**
- **Machine-learnt metrics** can be more correlated with human judgment, but behavior is **not interpretable**
- **Human judgement** is critical, but **humans are not consistent**

Some Recent Trend on NLG Evaluation

1) Automatic evaluation by GTP-4

- Vicuna (Chiang et al., 2023): LLaMA+70K user-shared conversational data
- 8 question categories; 10 questions per category
- Ask GPT-4 to rate the quality of their answers based on helpfulness, relevance, accuracy, and detail
- GPT-4 offers consistent scores and detailed explanations, but struggles with coding/math tasks



Vicuna vs. **different chatbots**

❑ It is still not clear whether we can use GPT-4 to replace human evaluation

[Submitted on 29 May 2023 (v1), last revised 30 Aug 2023 (this version, v2)]

Large Language Models are not Fair Evaluators

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, Zhifang Sui

In this paper, we uncover a systematic bias in the evaluation paradigm of adopting large language models~(LLMs), e.g., GPT-4, as a referee to score and compare the quality of responses generated by candidate models. We find that the quality ranking of candidate responses can be easily hacked by simply altering their order of appearance in the context. This manipulation allows us to skew the evaluation result, making one model appear considerably superior to the other, e.g., Vicuna-13B could beat ChatGPT on 66 over 80 tested queries with ChatGPT as an evaluator. To address this issue, we propose a calibration framework with three simple yet effective strategies: 1) Multiple Evidence Calibration, which requires the evaluator model to generate multiple evaluation evidence before assigning ratings; 2) Balanced Position Calibration, which aggregates results across various orders to determine the final score; 3) Human-in-the-Loop Calibration, which introduces a balanced position diversity entropy to measure the difficulty of each example and seeks human assistance when needed. We also manually annotate the "win/tie/lose" outcomes of responses from ChatGPT and Vicuna-13B in the Vicuna Benchmark's question prompt, and extensive experiments demonstrate that our approach successfully mitigates evaluation bias, resulting in closer alignment with human judgments. We release our code and human annotation at \url{this https URL} to facilitate future research.

[Submitted on 12 Oct 2023]

Prometheus: Inducing Fine-grained Evaluation Capability in Language Models

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, Minjoon Seo

Recently, using a powerful proprietary Large Language Model (LLM) (e.g., GPT-4) as an evaluator for long-form responses has become the de facto standard. However, for practitioners with large-scale evaluation tasks and custom criteria in consideration (e.g., child-readability), using proprietary LLMs as an evaluator is unreliable due to the closed-source nature, uncontrolled versioning, and prohibitive costs. In this work, we propose Prometheus, a fully open-source LLM that is on par with GPT-4's evaluation capabilities when the appropriate reference materials (reference answer, score rubric) are accompanied. We first construct the Feedback Collection, a new dataset that consists of 1K fine-grained score rubrics, 20K instructions, and 100K responses and language feedback generated by GPT-4. Using the Feedback Collection, we train Prometheus, a 13B evaluator LLM that can assess any given long-form text based on customized score rubric provided by the user. Experimental results show that Prometheus scores a Pearson correlation of 0.897 with human evaluators when evaluating with 45 customized score rubrics, which is on par with GPT-4 (0.882), and greatly outperforms ChatGPT (0.392). Furthermore, measuring correlation with GPT-4 with 1222 customized score rubrics across four benchmarks (MT Bench, Vicuna Bench, Feedback Bench, Flask Eval) shows similar trends, bolstering Prometheus's capability as an evaluator LLM. Lastly, Prometheus achieves the highest accuracy on two human preference benchmarks (HHH Alignment & MT Bench Human Judgment) compared to open-sourced reward models explicitly trained on human preference datasets, highlighting its potential as an universal reward model. We open-source our code, dataset, and model at this https URL.

Comments: Work in Progress