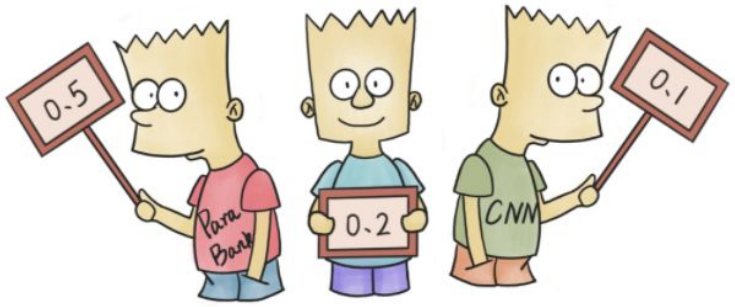
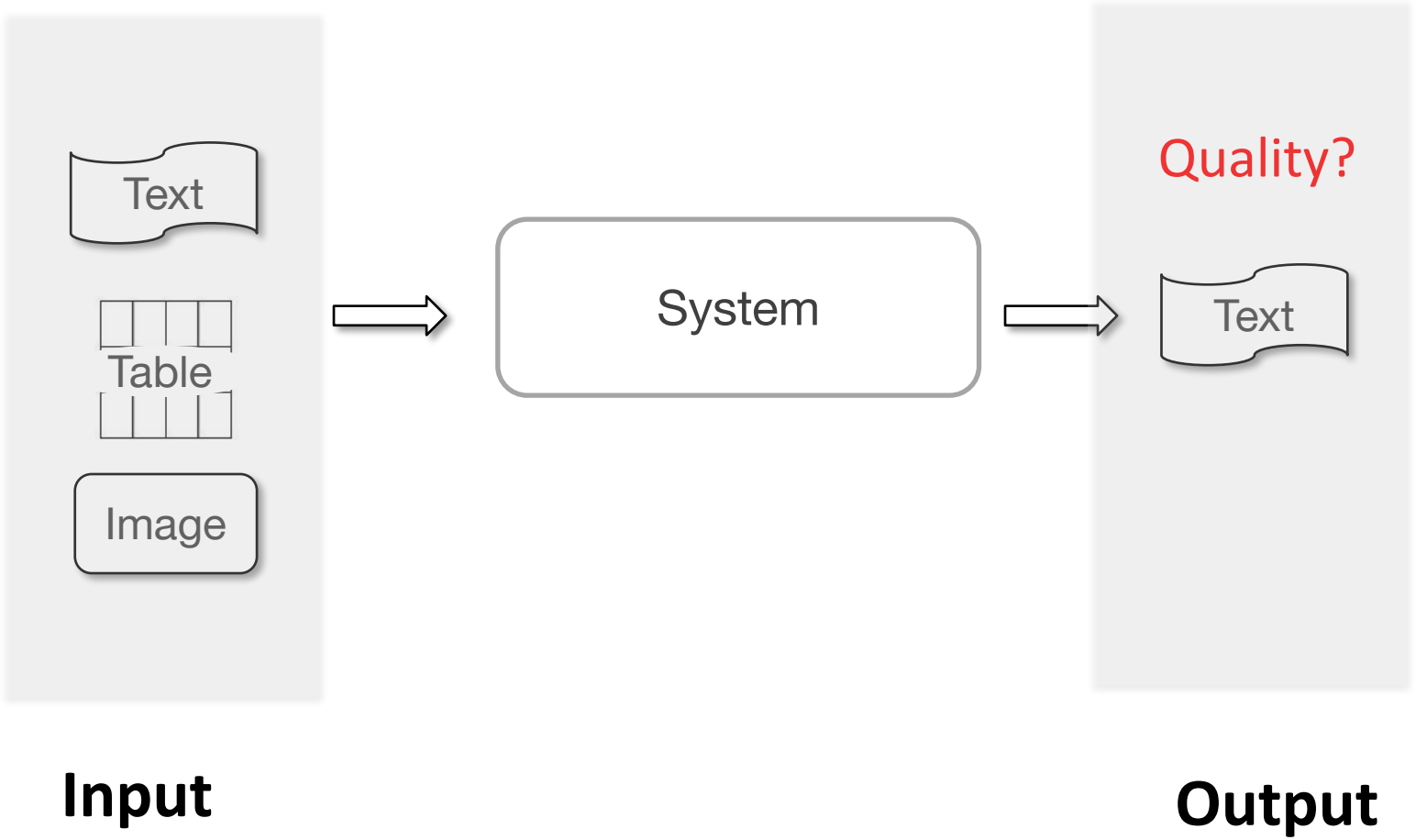


BARTScore: Evaluating Generated Text as Text Generation

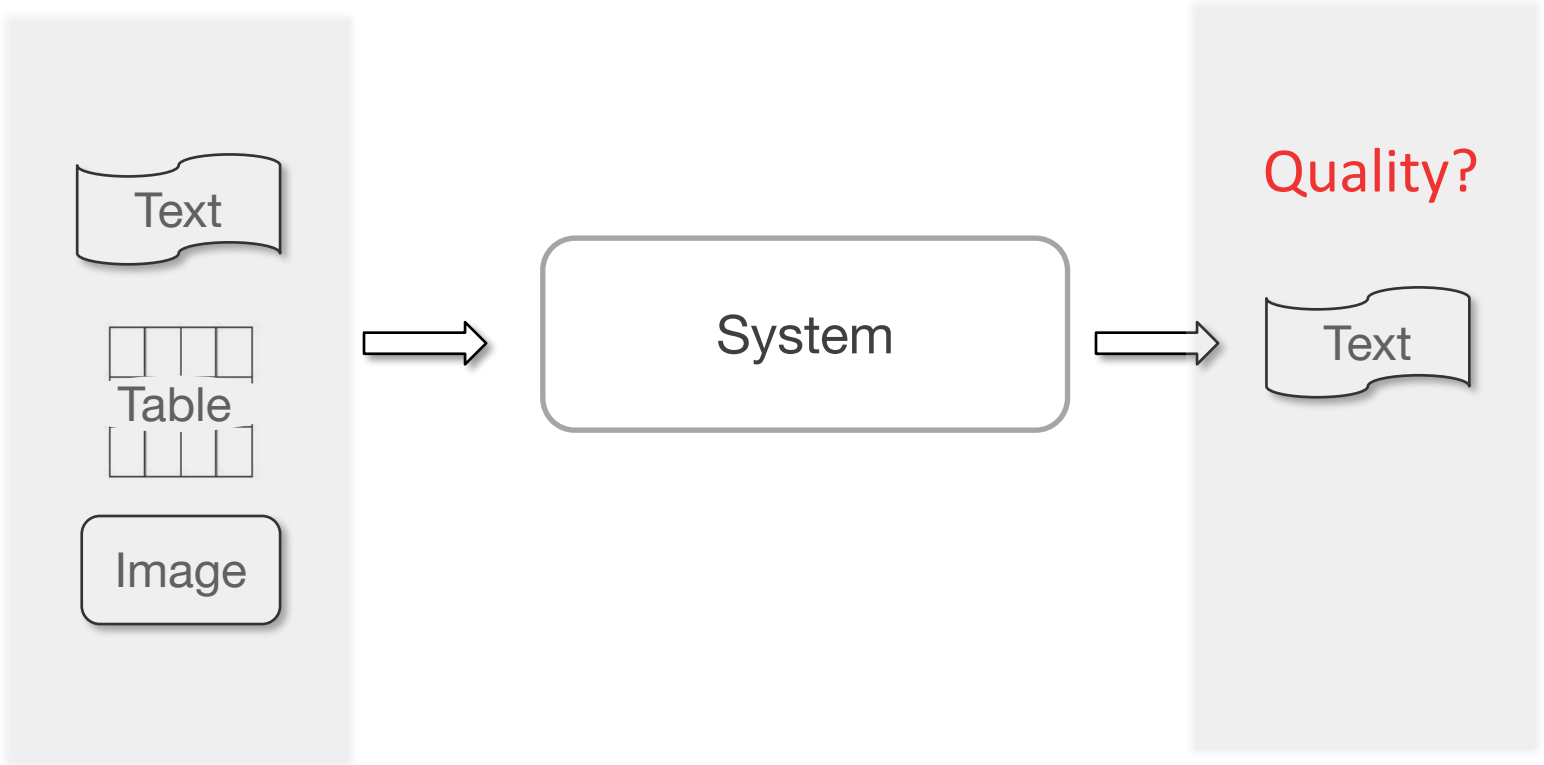
Weizhe Yuan, Graham Neubig, Pengfei Liu

Carnegie Mellon University
Language Technologies Institute

What's the goal of this task?



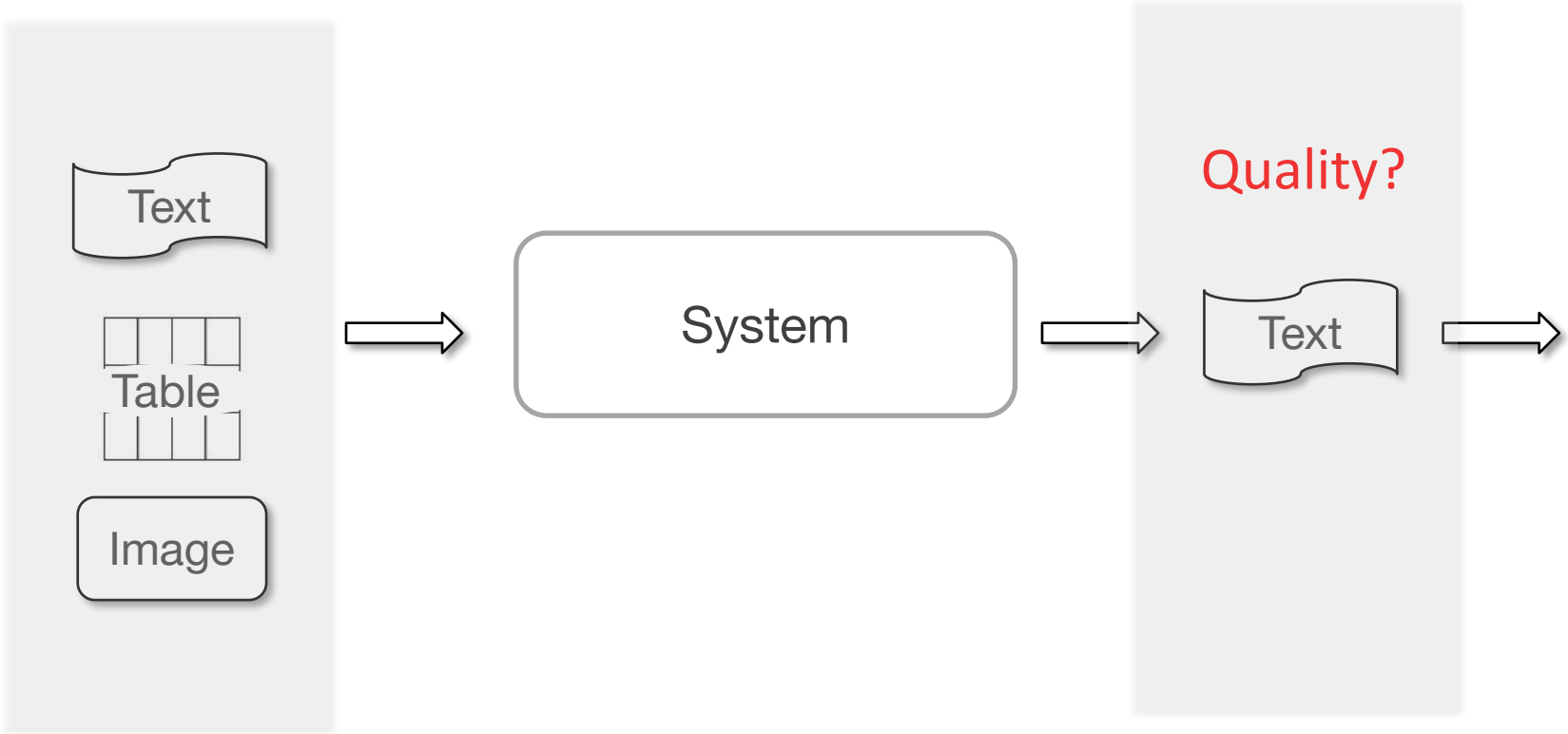
What's the goal of this task?



- ☐ Fluency
- ☐ Relevance
- ☐ Coherence
- ☐ Informativeness
- ☐ Factuality
- ☐ Semantic Coverage
- ☐ Adequacy

Input **Output** **Perspectives**

What's the goal of this task?



- Human
- Automated Metric

How does this field progress so far?

Let’s try to summarize
characteristics of existing research

Background

- Most of the current metrics mainly can only evaluate text from a limited number of perspectives.

  - ☐ Fluency
  - ☐ Relevance
  - ☐ Coherence
  - ☐ Informativeness
  - ☐ Factuality
  - ☐ Semantic Coverage
  - ☐ Adequacy

Perspectives

Background

- Most of the current metrics mainly can only evaluate text from a limited number of perspectives.
  - In practice, we need to use multiple metrics to evaluate different perspectives

- ☐ Fluency
- ☐ Relevance
- ☐ Coherence
- ☐ Informativeness
- ☐ Factuality
- ☐ Semantic Coverage
- ☐ Adequacy

ROUGE

Perspectives

Background

- Most of the current metrics mainly can only evaluate text from a limited number of perspectives.
  - In practice, we need to use multiple metrics to evaluate different perspectives

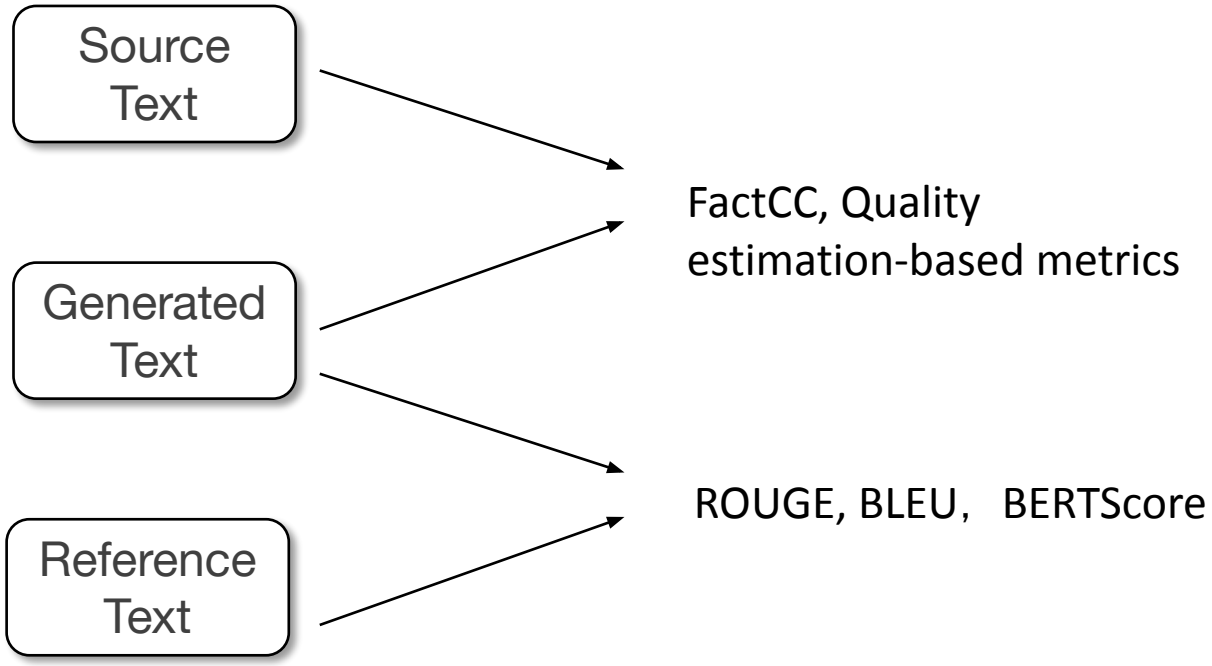
- ☐ Fluency
- ☐ Relevance
- ☐ Coherence
- ☐ Informativeness
- ☐ Factuality
- ☐ Semantic Coverage
- ☐ Adequacy

FactCC

Perspectives

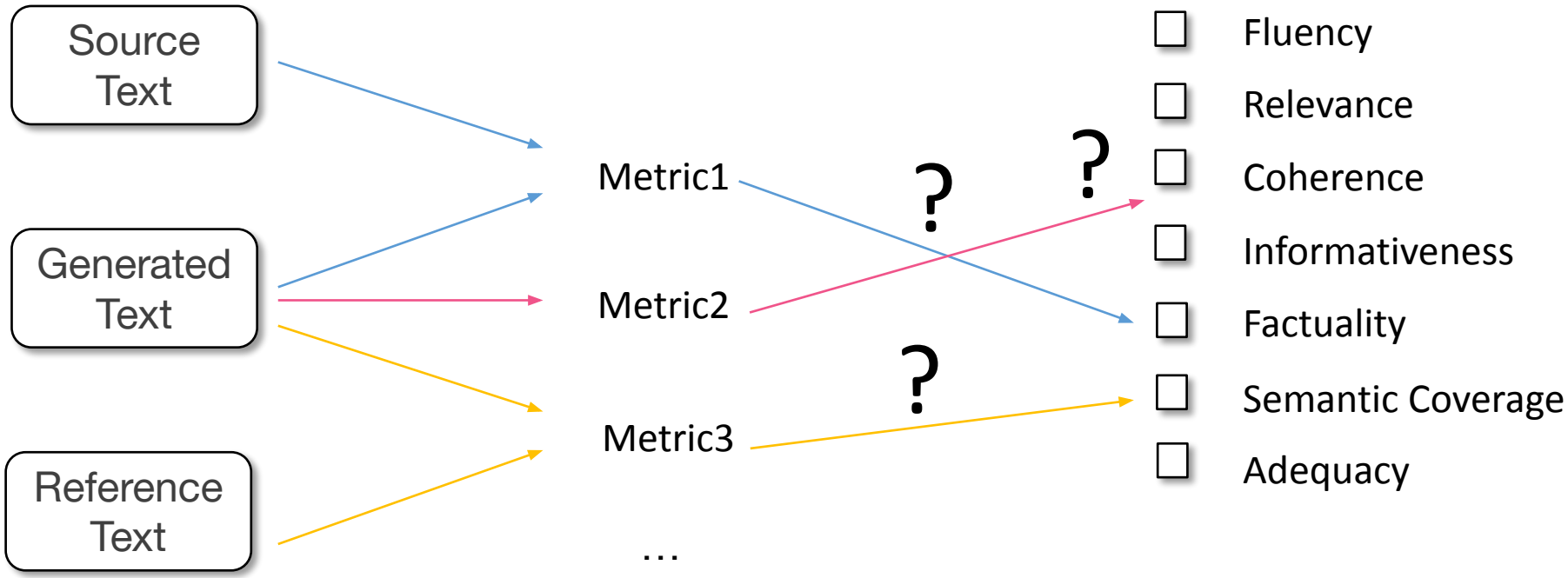
Background

- Most existing metrics only consider the relationship between
  - Generated text <-> reference text OR
  - Generated text <-> source text



Background

- Existing metrics only consider the relationship between (src, gen) or (src, ref)
  - Unclear: how different choices of text combination influence different evaluation perspectives?



Background

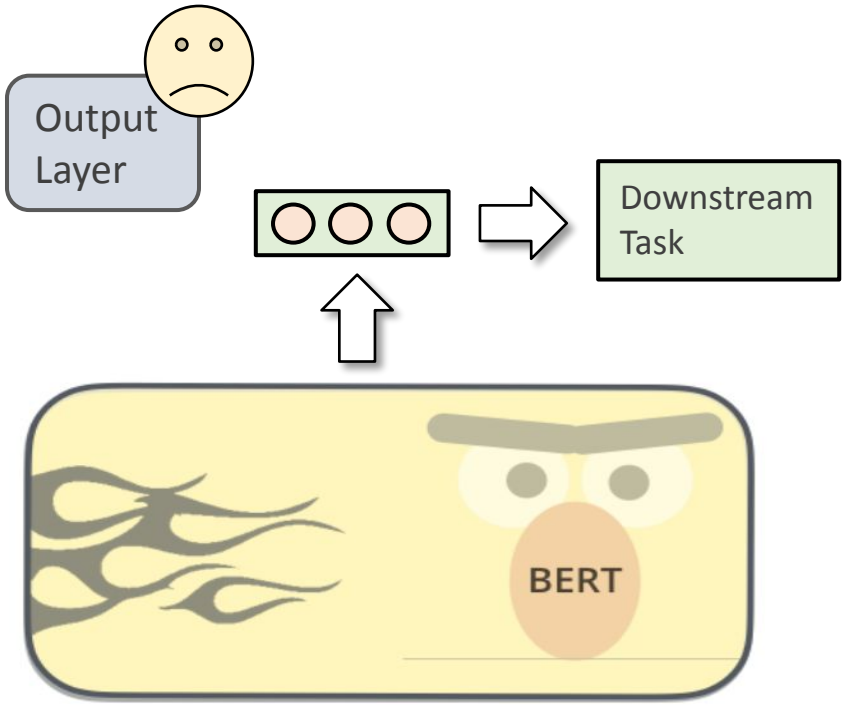
- More and more metrics seek to take the advantage of pre-trained language models.

BERTScore
MoverScore
BLEURT
COMET
…

Background

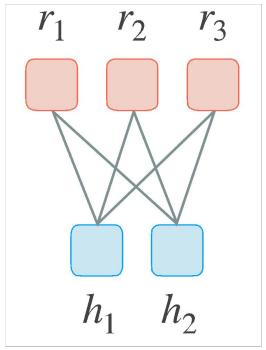
- More and more metrics seek to take the advantage of pre-trained language models.
  - However, the PLMs' parameters may not be fully utilized.

BERTScore
MoverScore
BLEURT
COMET
…



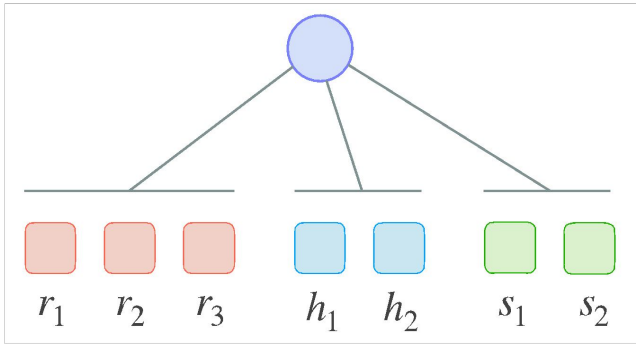
Background

- Most of the metrics take evaluation as unsupervised matching, supervised regression, or supervised ranking problems.



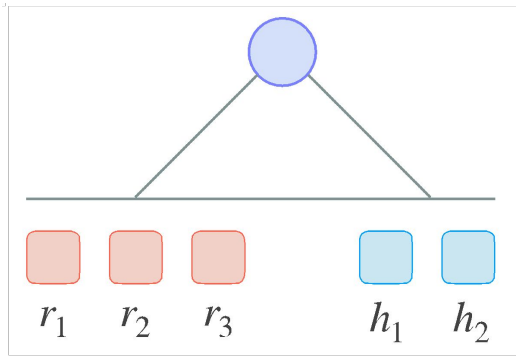
Match

ROUGE, BLUE



Ranking

COMET, BEER

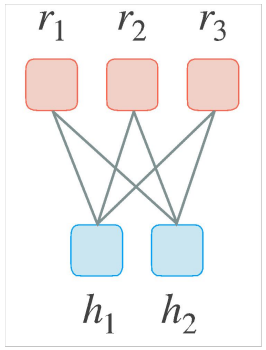


Regression

BLEURT COMET

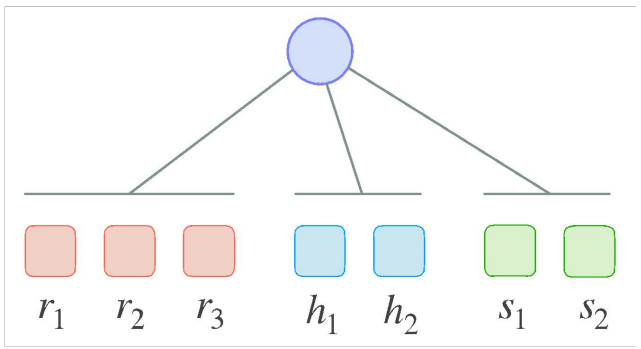
Background

- Most of the metrics take evaluation as unsupervised matching, supervised regression, or supervised ranking problems.
  - SOTA generation systems are Seq2Seq models, why not using them?



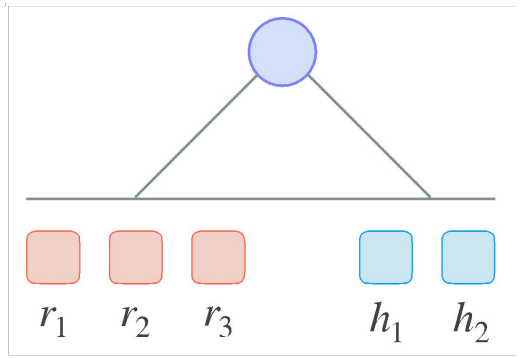
Match

ROUGE, BLUE



Ranking

COMET, BEER

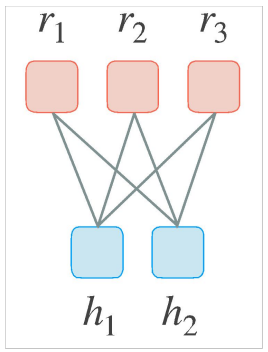


Regression

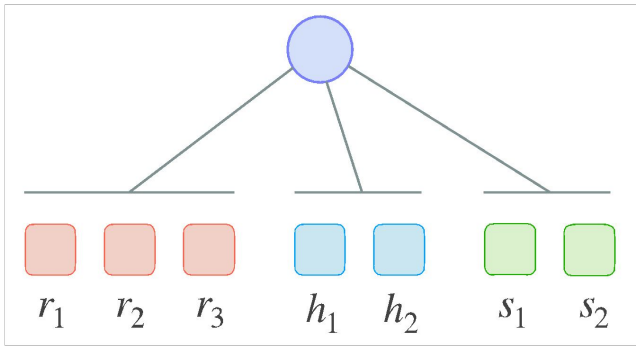
BLEURT COMET

Background

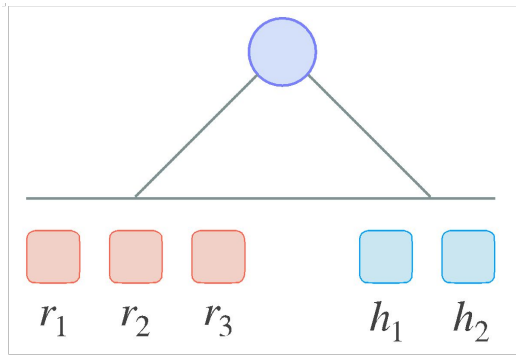
- Is there a metric that can
  - flexibly model different relationships among (source, generated, reference) texts
  - support evaluation from multiple perspectives
  - make full use of pre-trained models?



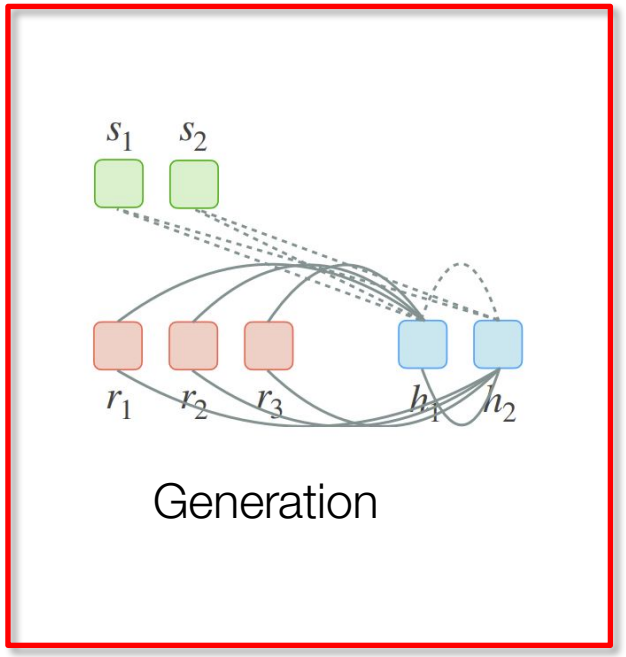
Match



Ranking



Regression

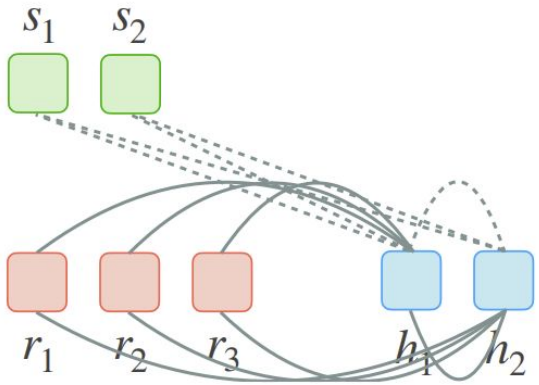


Generation

Text Generation Evaluation as Text Generation

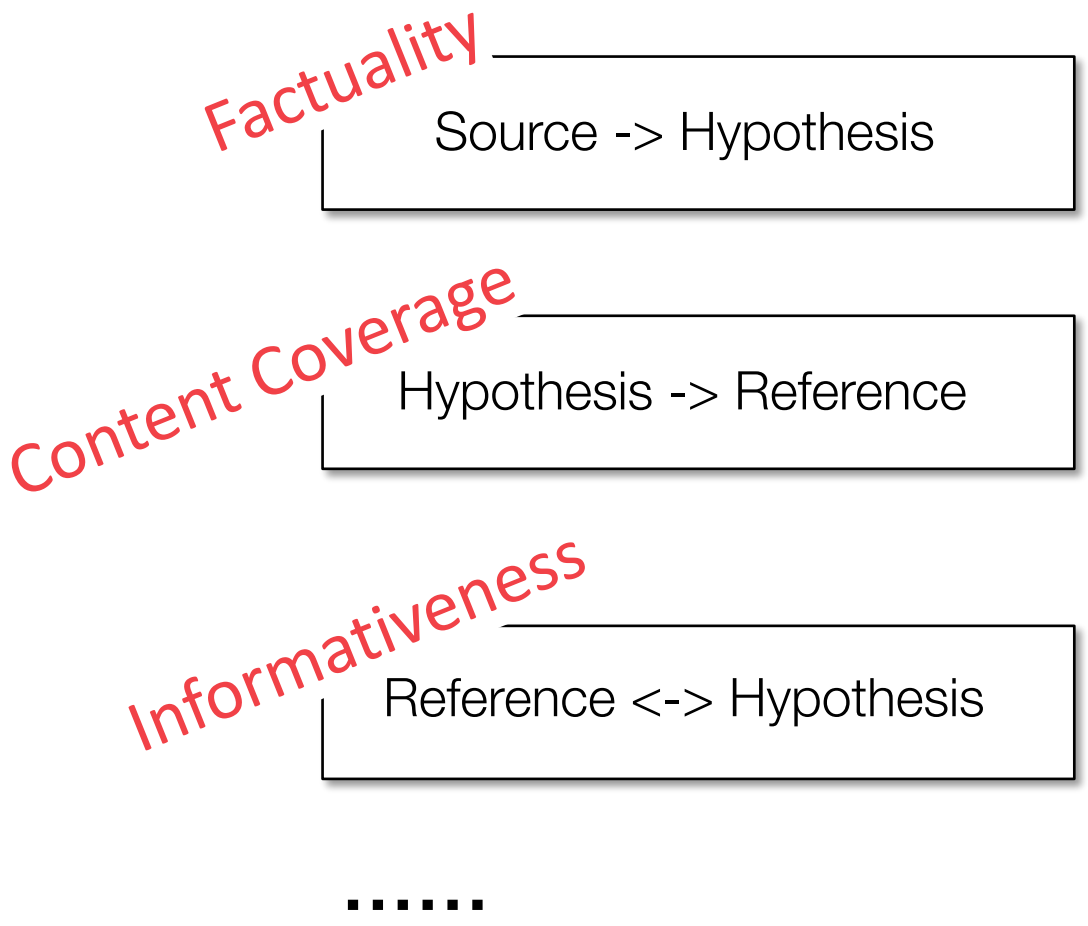
General Idea:

- models trained to convert the generated text to/from a reference output or the source text will achieve **higher scores** when **the generated text is better**



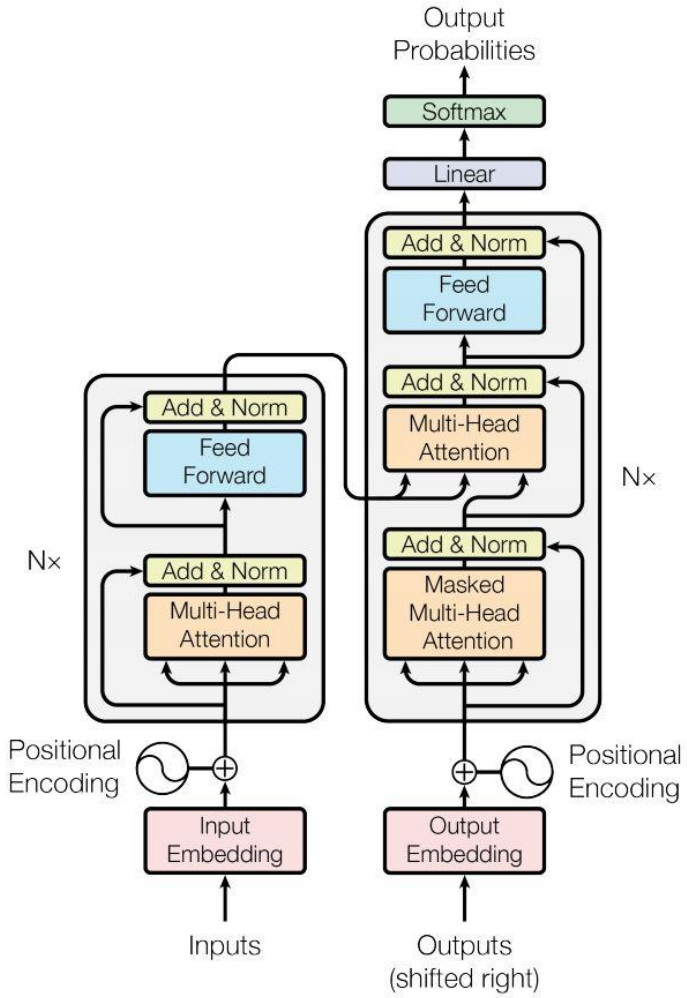
Benefits

- Benefit 1: The different evaluation perspectives can be naturally supported.



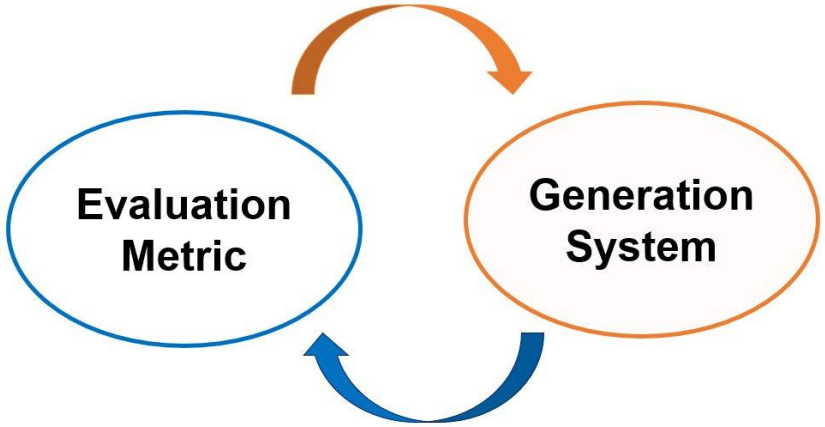
Benefits

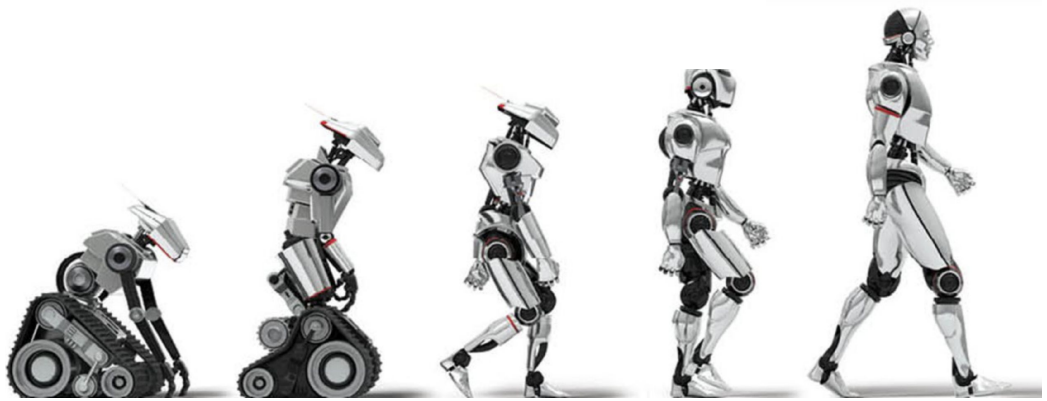
- Benefit 2: This new formulation can make **full** use of the parameters of PLMs.



Benefits

- Benefit 3: Co-evolving of generation systems and evaluation metrics.





- Better systems will result in better evaluation metrics.
- Better evaluation metrics will guide the systems to become better.

BARTScore Basics

BARTScore is used to get the generation probability from a source text x to a target text y (Note: the calculated scores are negative numbers)

$$BARTScore = \sum_{t=1}^{m} w_t \log p(y_t | y_{<t}, x, \theta)$$

BARTScore Basics

- We consider BARTScore variants from two dimensions:
  - **Fine-tuning**: Change the parameters θ of PLM by considering different fine-tuning tasks to make the pre-training domain closer to the evaluation domain.
  - **Prompting**: Prompt the source text x or target text y to better elicit knowledge from PLMs.

$$BARTScore = \sum_{t=1}^{m} w_t \log p(y_t | y_{<t}, x, \theta)$$

Prompting

- Instead of adapting pre-trained LMs to downstream tasks via objective engineering, downstream tasks are **reformulated** to look more like those solved during the original **LM pre-training** with the help of **a textual prompt**.

Prompting

- Instead of adapting pre-trained LMs to downstream tasks via objective engineering, downstream tasks are reformulated to look more like those solved during the original LM training with the help of a textual prompt.
  - E.g. Sentiment Analysis
    <A movie review> The review is __ .

Prompting

- Instead of adapting pre-trained LMs to downstream tasks via objective engineering, downstream tasks are reformulated to look more like those solved during the original LM training with the help of a textual prompt.
  - E.g. Sentiment Analysis
    <A movie review> The review is __ .
  - E.g. MT
    English: I missed the bus today. French: __

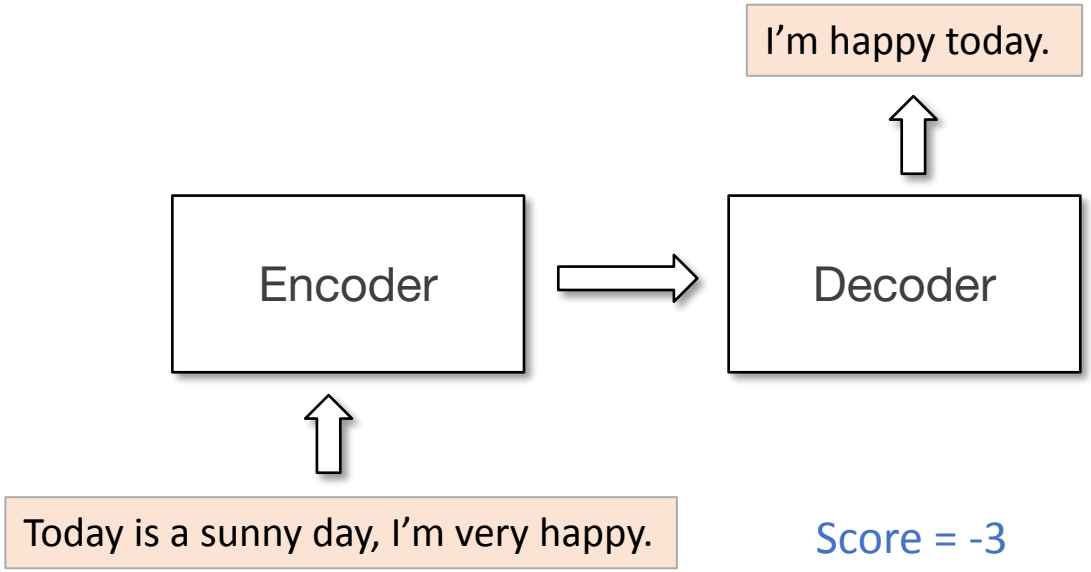
Prompting

- Instead of adapting pre-trained LMs to downstream tasks via objective engineering, downstream tasks are **reformulated** to look more like those solved during the original **LM pre-training** with the help of **a textual prompt**.
  - E.g. Sentiment Analysis
    <A movie review> The review is __ .
  - E.g. MT
    English: I missed the bus today. French: __

- Better elicit knowledge from PLMs

Prompting

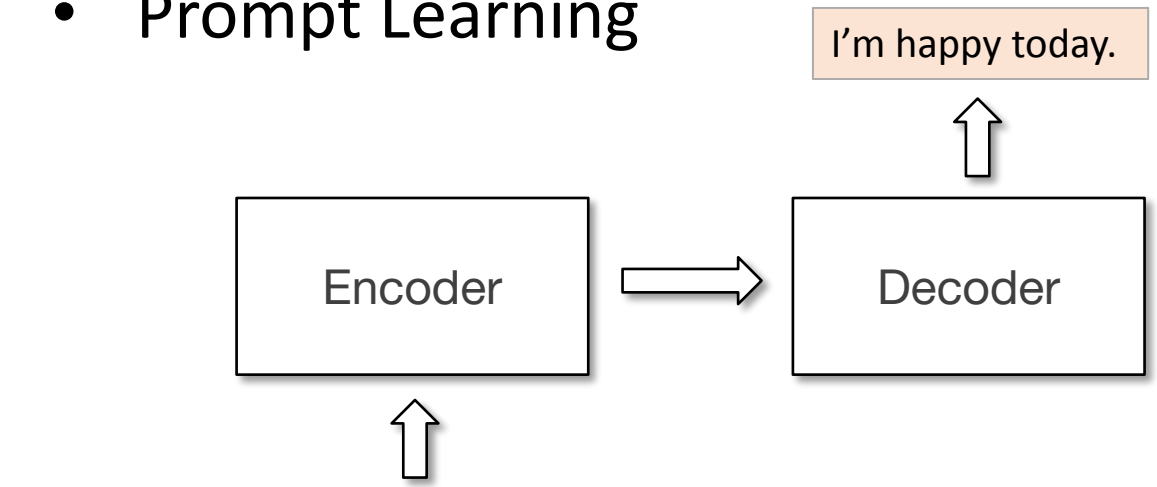
- Original



Score = -3

Final score = -3

- Prompt Learning



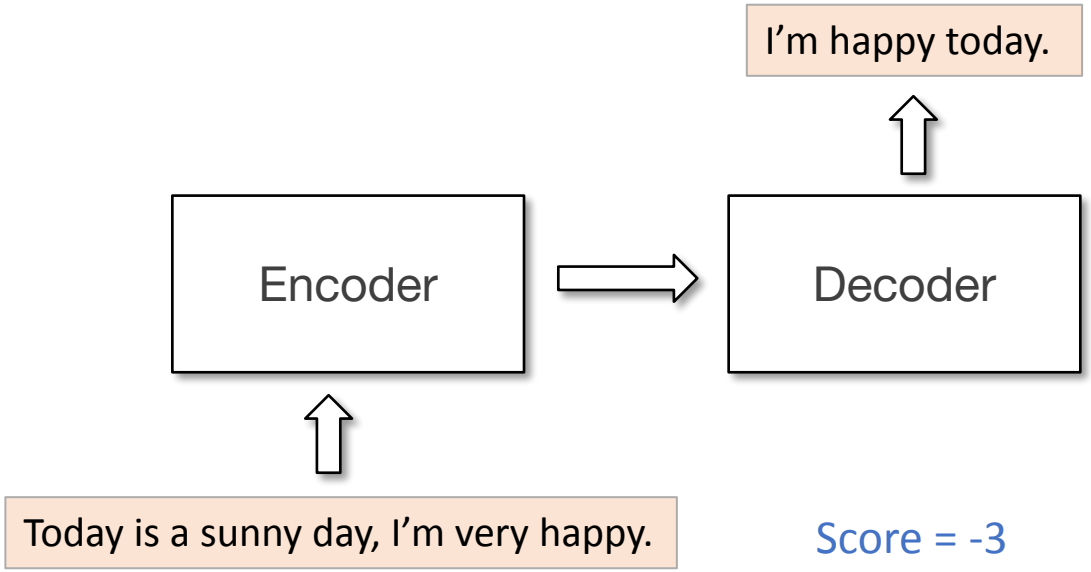
Score = -3

Score = -2

Final score = $\frac{(-3)+(-2)}{2} = -2.5$

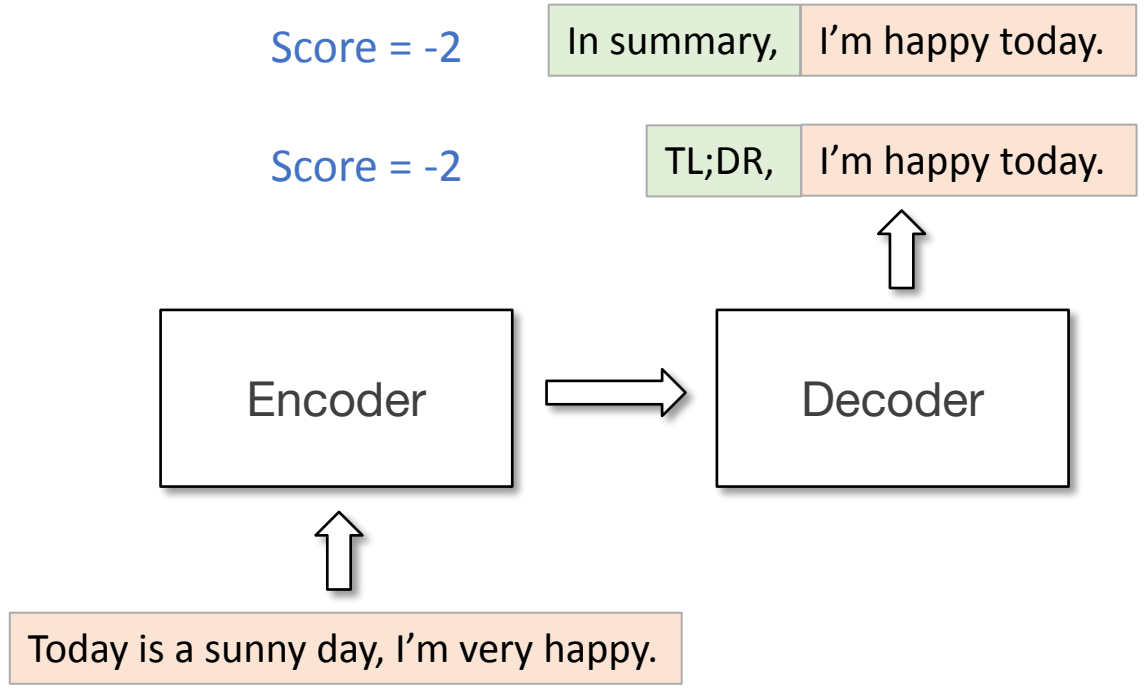
Prompting

- Original



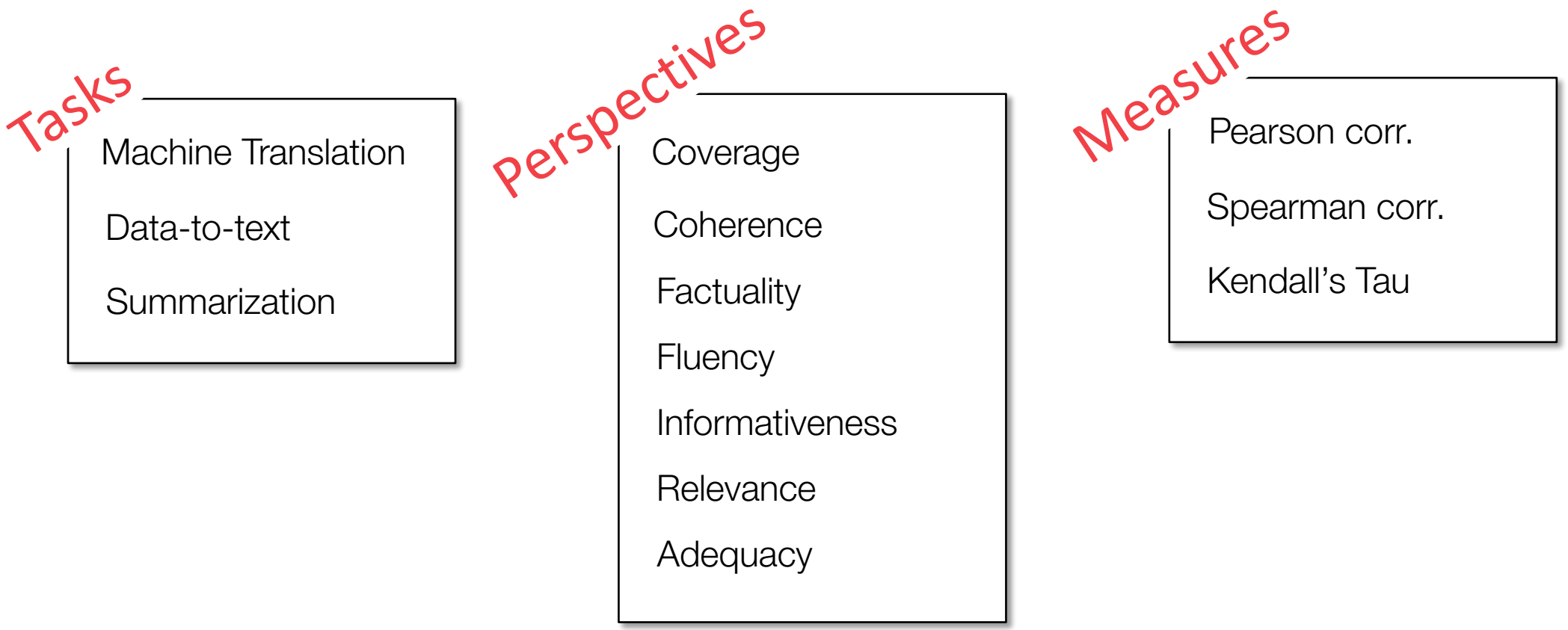
Final score = -3

- Prompt Learning



Final score = $\frac{(-2)+(-2)}{2} = -2$

Experiments



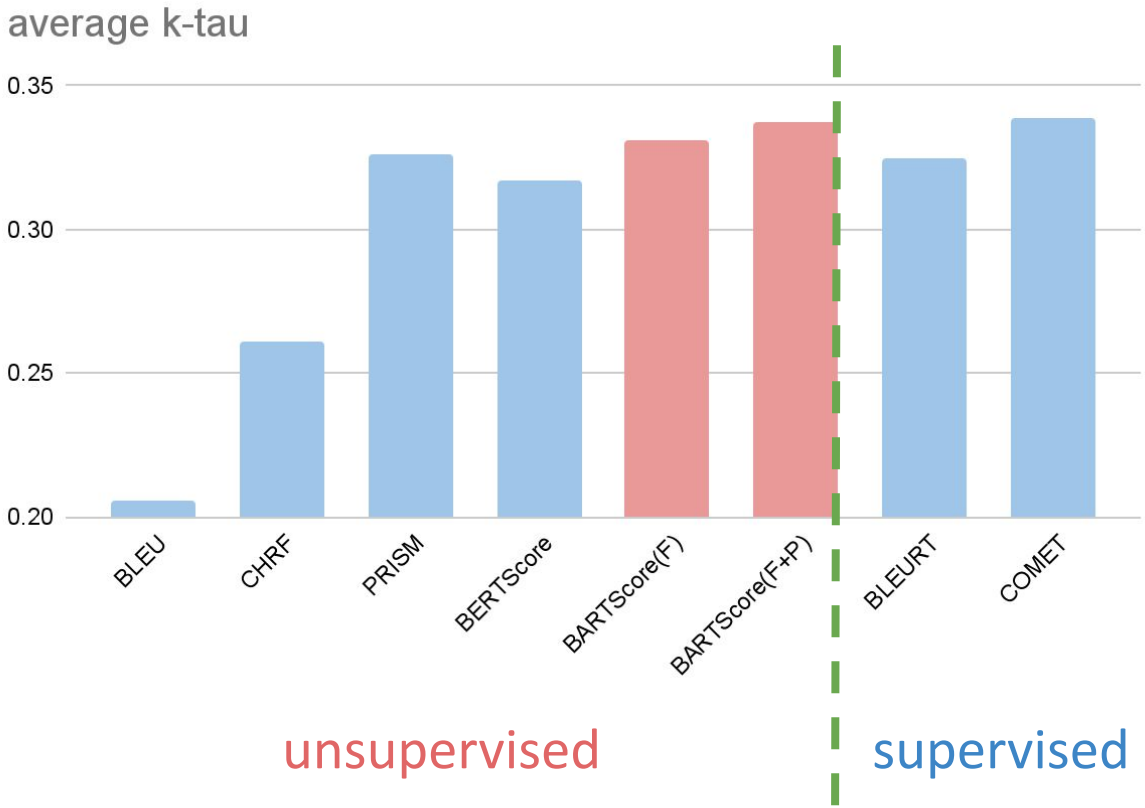
3 tasks, 16 datasets, 7 perspectives

Baseline Metrics

- We consider the following baseline metrics in our experiments.

  - ☐ ROUGE (1, 2, L)
  - ☐ BLEU
  - ☐ CHRF
  - ☐ BERTScore
  - ☐ MoverScore
  - ☐ PRISM
  - ☐ BLEURT
  - ☐ COMET

Results: Machine Translation

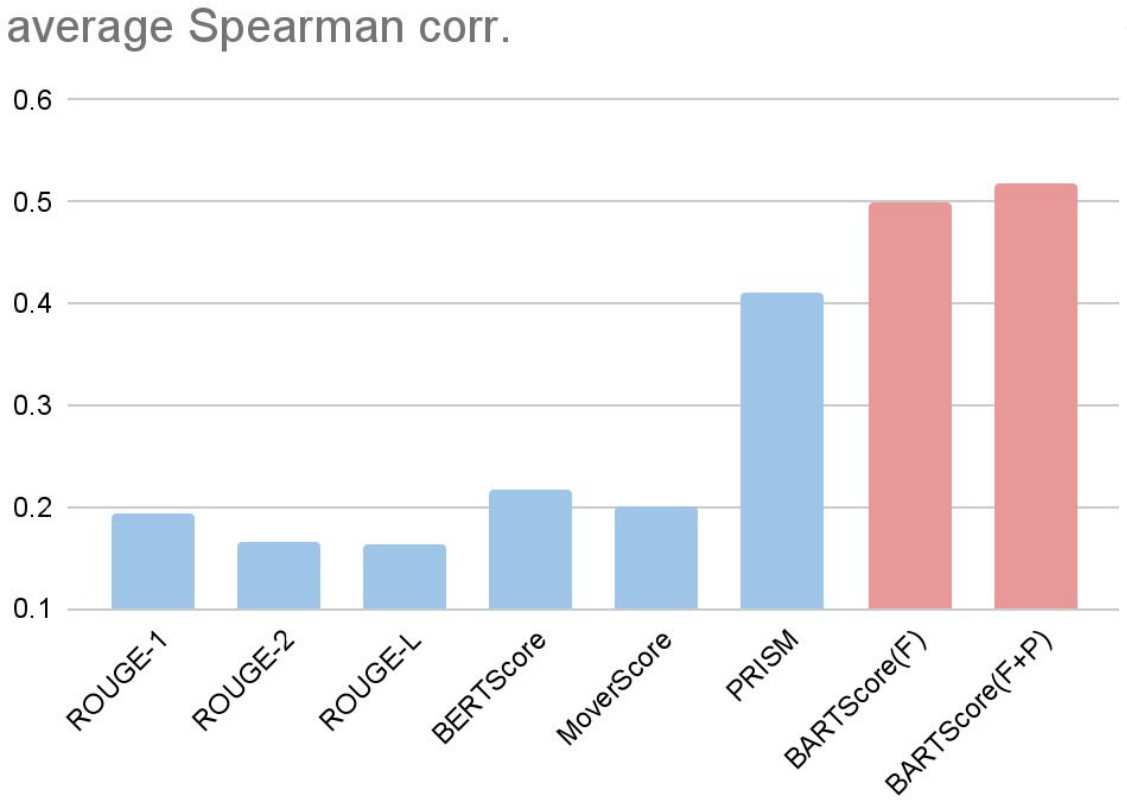


- F: Fine-tuning
- P: Prompting

➢ Unsupervised SOTA

➢ Improvements through prompting

Results: Summarization



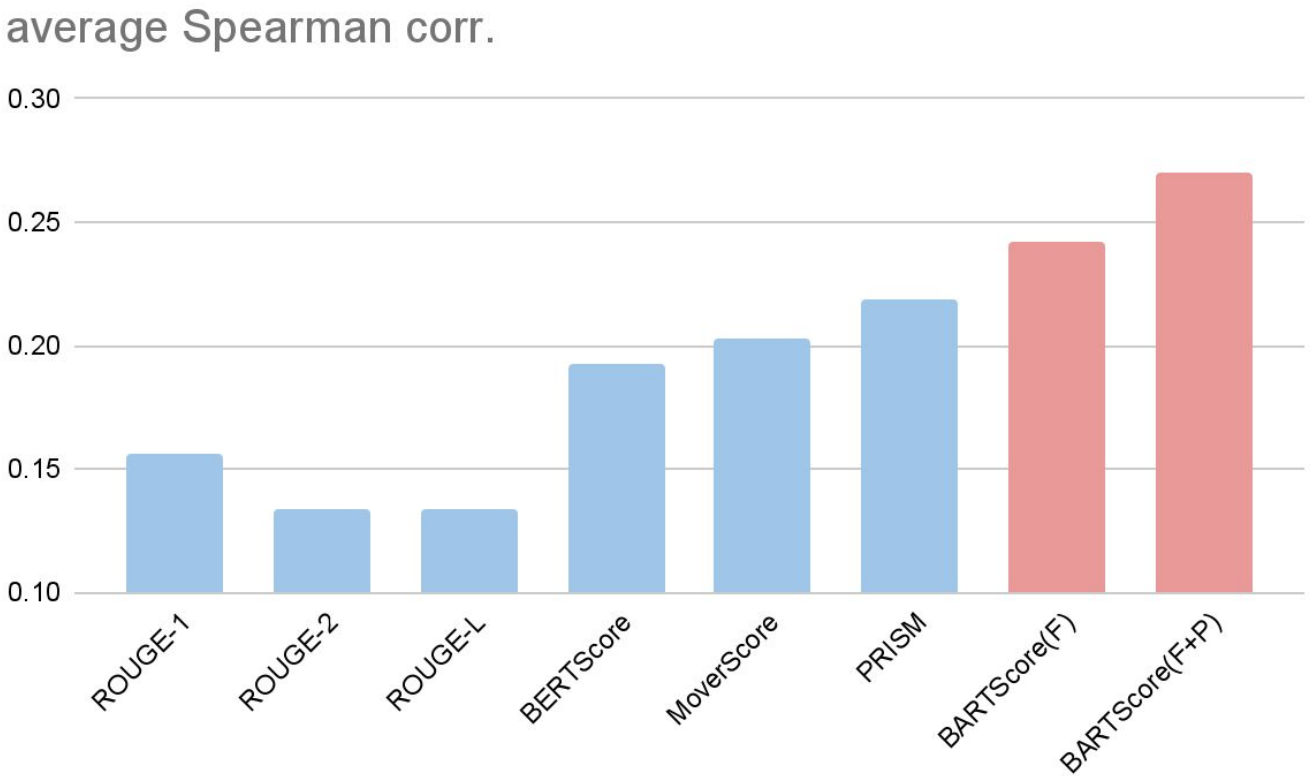
- F: Fine-tuning
- P: Prompting

➢ Unsupervised SOTA

➢ Outperform others by large margin

➢ Prompting brings improvements

Results: Data-to-text



- F: Fine-tuning
- P: Prompting

➢ SOTA

➢ Prompt helps informativeness

Fine-grained Analysis

- Prompt Analysis (Summarization & Data-to-text)

We first group all the evaluation perspectives into three categories:

1) *semantic overlap* (informativeness, pyramid score, and relevance)
2) *linguistic quality* (fluency, coherence)
3) *factual correctness* (factuality).

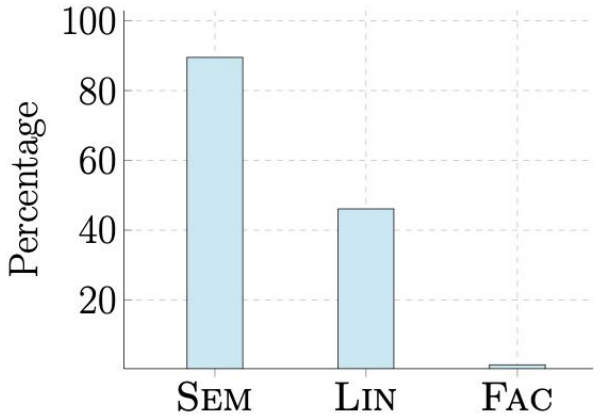
Fine-grained Analysis

- Prompt Analysis (Summarization & Data-to-text)

SEM: Semantic Overlap

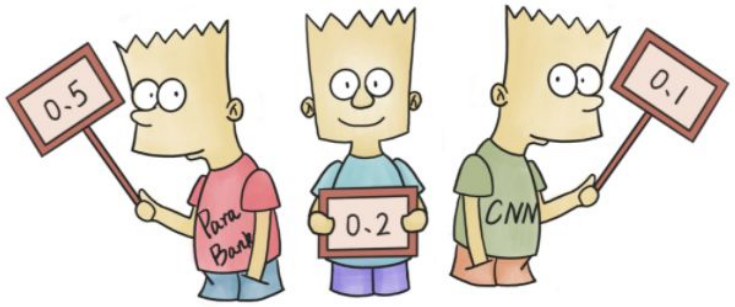
LIN: Linguistic Quality

FAC: Factual Correctness



(c) Evaluation perspective

- Prompt helps semantic overlap
- Prompt effect on linguistic quality unclear
- Prompt does not help factual correctness

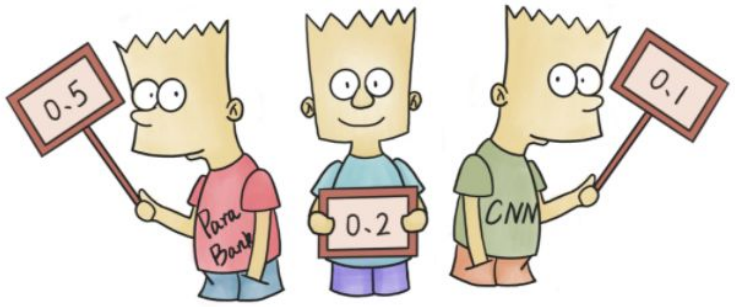


BARTScore: Evaluating Generated Text as Text Generation

Demo: http://bartscore.sh/

Leaderboard: *http://explainaboard.nlpedia.ai/leaderboard/task-meval/*

Code: https://github.com/neulab/BARTScore



BARTScore: Evaluating Generated Text as Text Generation

Thank you