Evaluation for Natural Language Generation

Guest lecture

Natural Language Processing Group

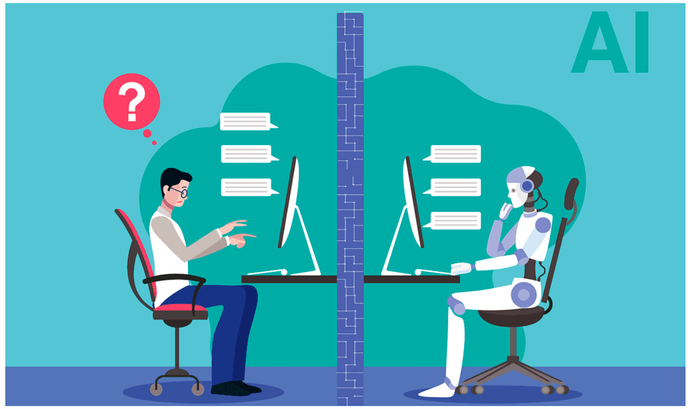
Dr. Wei Zhao 30/01/2024

Motivation

Why are we learning this

Evaluation is the center of mass in science and life

Homework Review

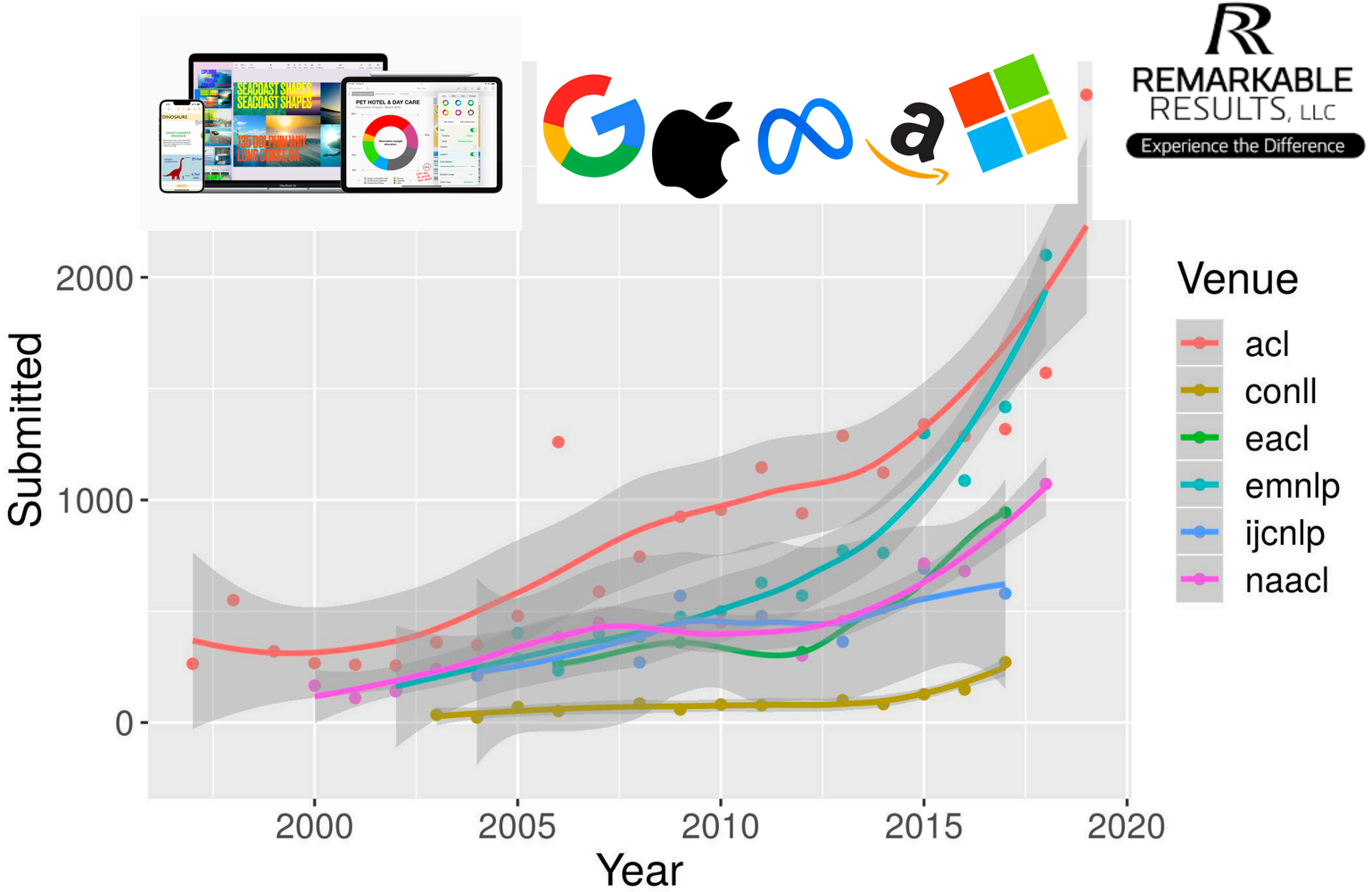


Turing Test

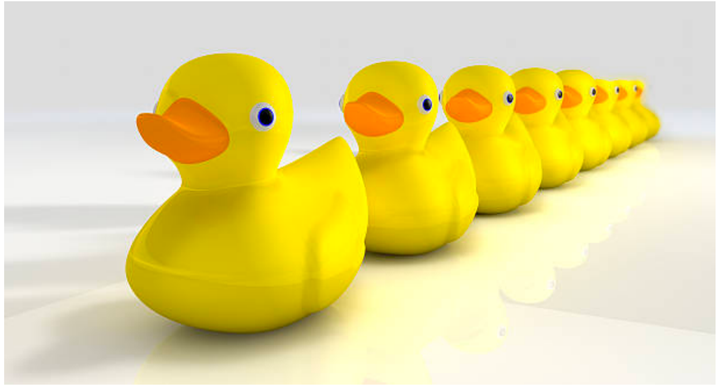
Job Application

Motivation

Why are we learning this



Evaluation is a goalkeeper



NLP systems in a row to-be-evaluated

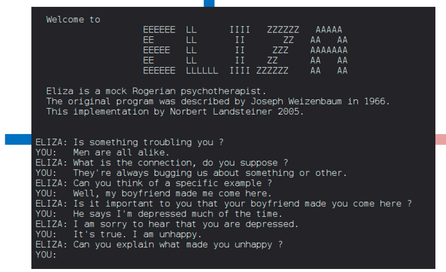
History of NLP Evaluation

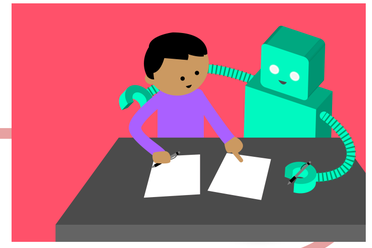
What it happened before 2000

1970's

William A Woods

- Model size
- Running time





Chatbot

1995

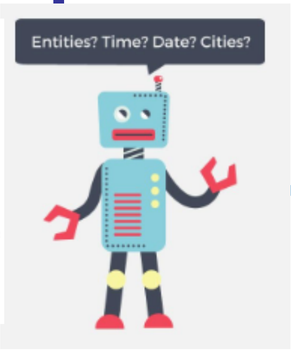
Jones and Galliers

Intrinsic and extrinsic evaluation

Guide and Mauri

Quality assurance:
- Quantitative evaluation
- Troubleshoot bad cases

1986



EAGLES for NLP

Evaluation principles:
- Evaluation Concerns
- Evaluation Criteria
- Evaluation Measures

1996

Now

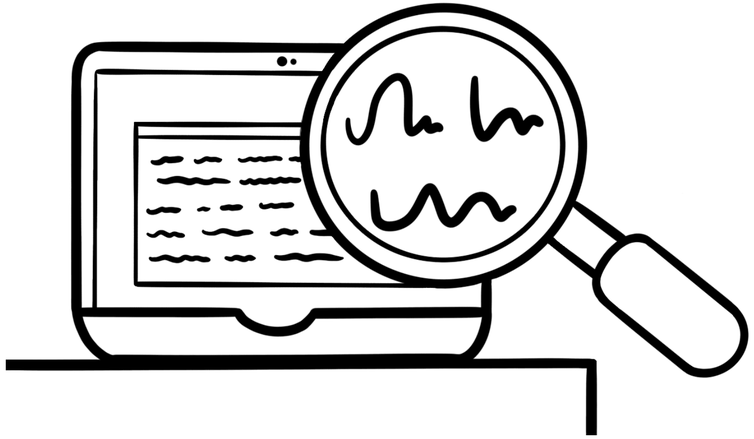
ChatGPT

History of NLP Evaluation

EAGLES for NLP in 1996

Adequacy Evaluation

Progress Evaluation

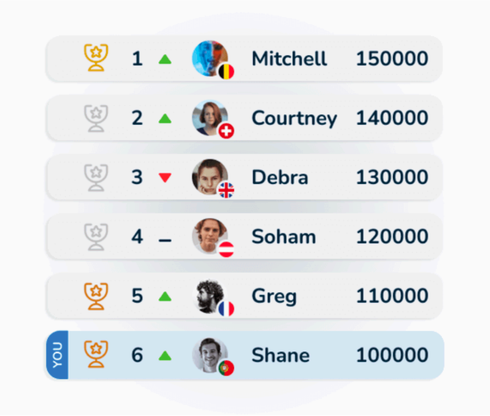


Diagnostic Evaluation

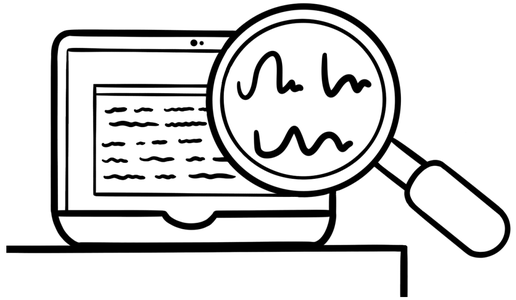
Today's NLP Evaluation

New terminologies

Evaluation Metrics



System Comparison



Explainability

Today's NLP Evaluation

They are not mutually exclusive

NLP Evaluation = Metrics X System Comparison X Explainability

Reliable instructions

Misuse instructions

Explainable

Outline

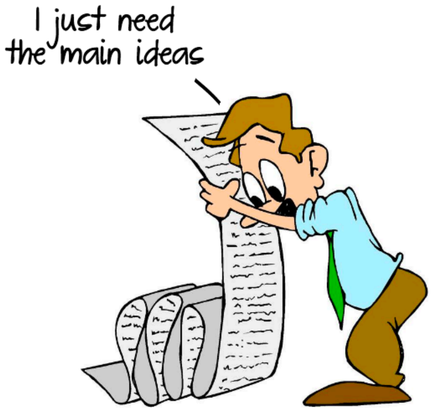
About this lecture

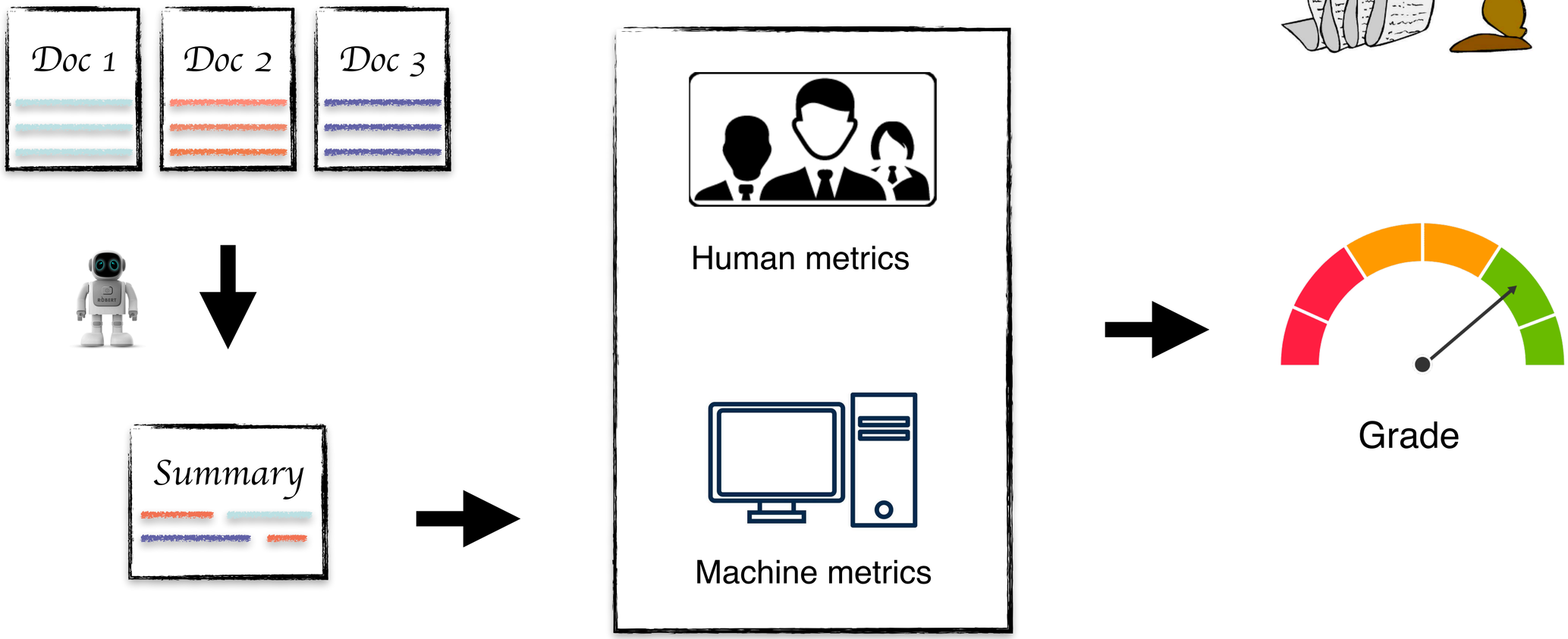
- Evaluation Metrics
- System Comparison
- Explainability

Natural Language Generation

Evaluation Metrics

Framework





Evaluation Metrics

Human metrics

Candidate: Which language model uses a sequence-to-sequence architecture?

Reference: Which language model is based on a sequence-to-sequence architecture?

To what extent does the black text **adequately** express the meaning of the grey text?

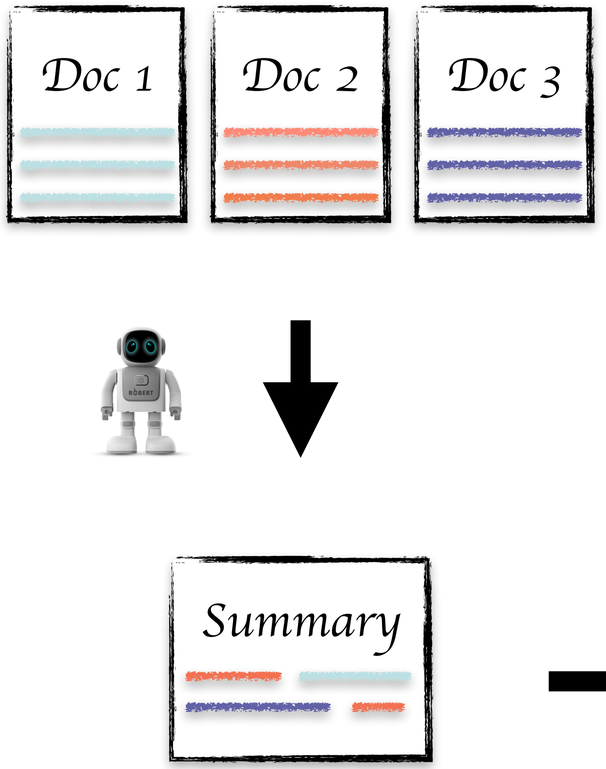
Strongly Disagree Strongly Agree

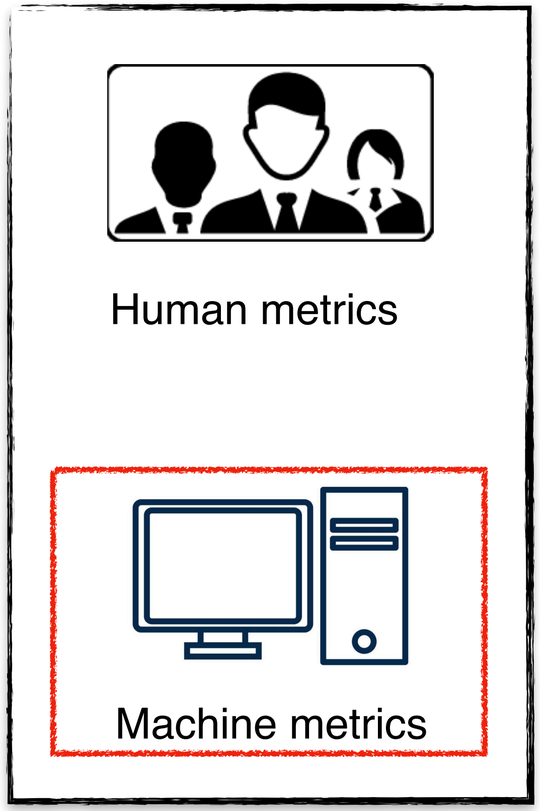
How much does the black text is **fluent** in English?

Strongly Disagree Strongly Agree

Evaluation Metrics

Framework



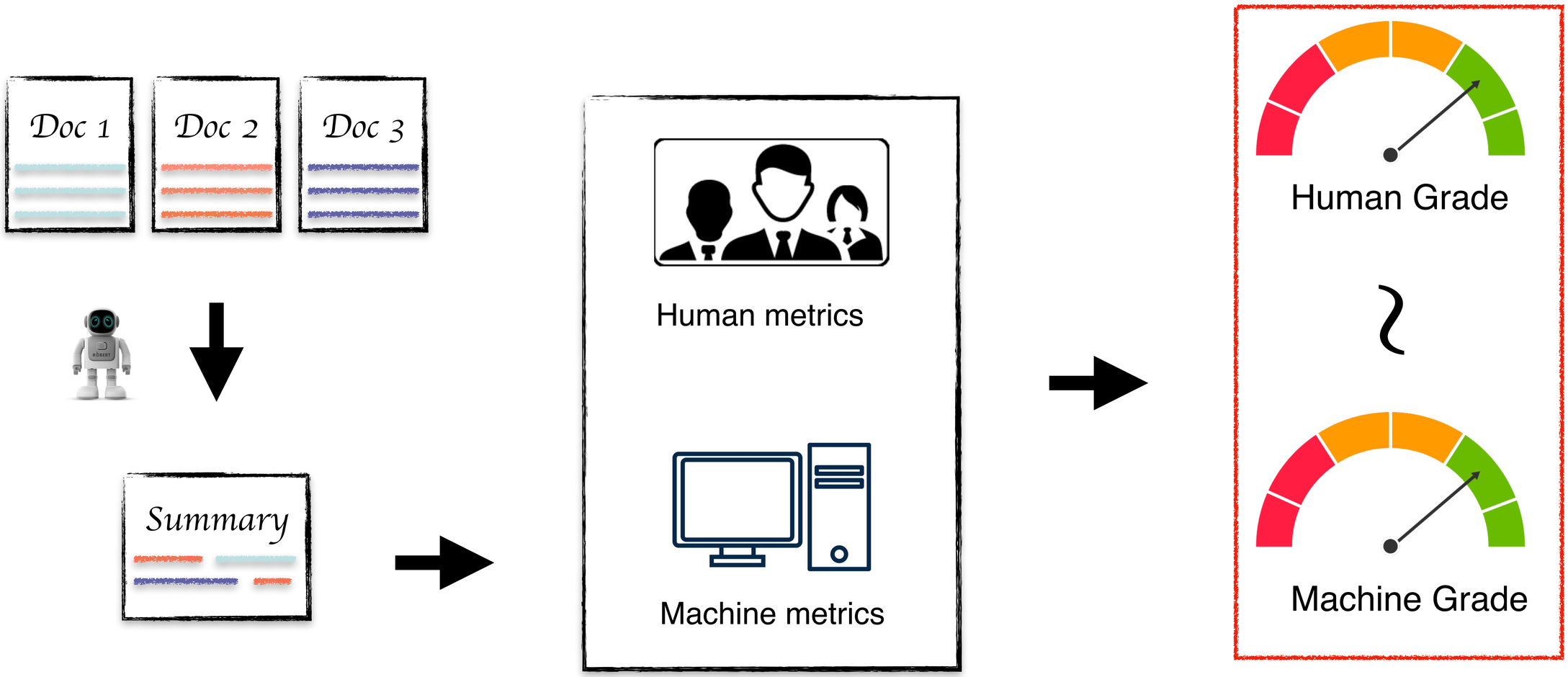


Grade

Evaluation done in minutes or seconds

Evaluation Metrics

When machine metrics are considered good



Machine Metrics

Ref-based evaluation

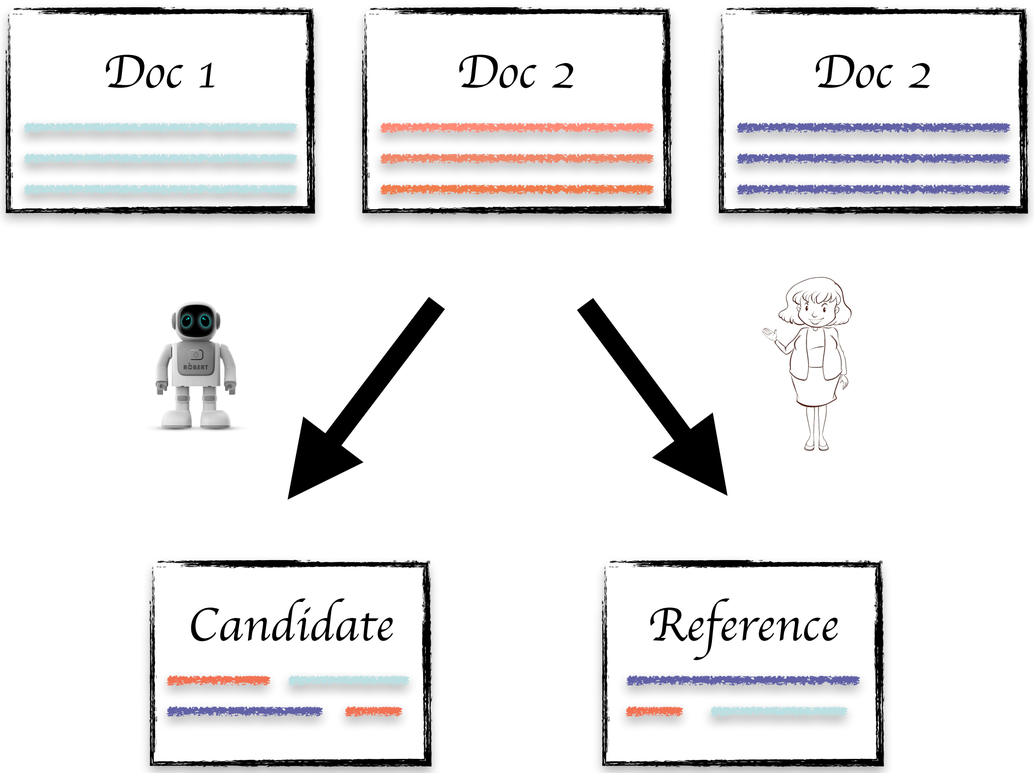
Reference-based Evaluation

A. Costly human references needed

B. Limited evaluation

C. Homogeneous (length, language)

D. Accurate



Machine Metrics

Ref-free evaluation

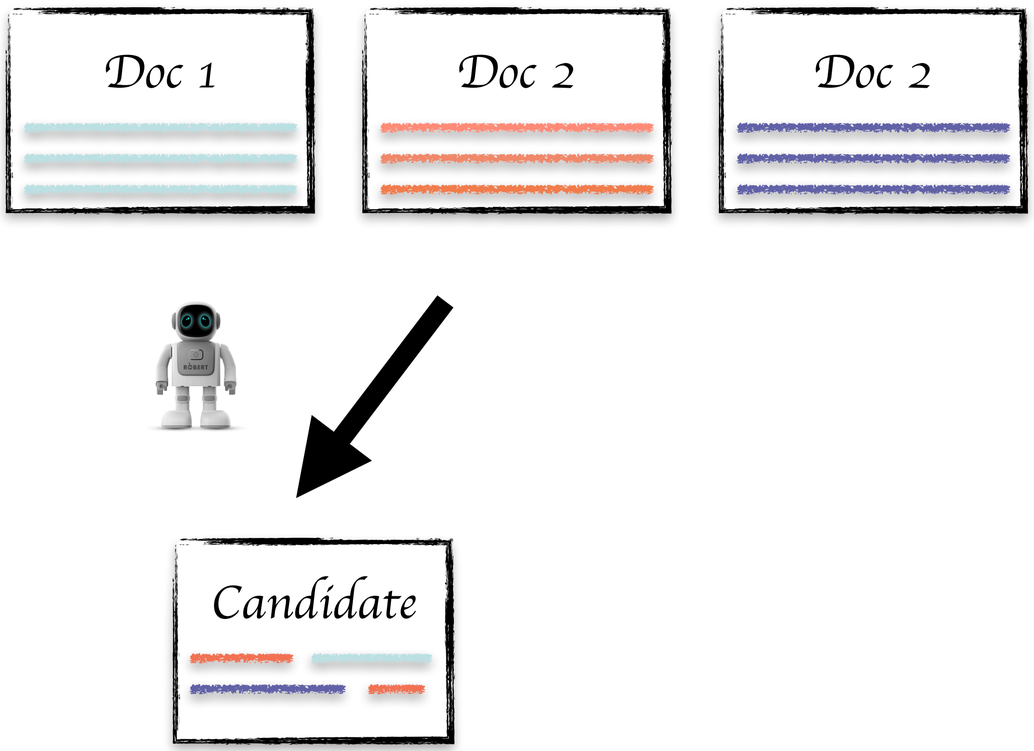
Reference-free Evaluation

A. No human references needed

B. Web-scale evaluation

C. Heterogeneous (length, language)

D. Less accurate



Machine Metrics

Taxonomy

Lexicality

Lexical metrics compare candidates with human references on lexical level, often termed retrieval-based metrics, which treat

Semantics

Semantic metrics focus on soft lexical matching, which address lexical similarity with synonym matching and broader word

Discourse

Discourse metrics target the assessment of text coherence, which particularly capture coherence

Factuality

Factual metrics are concerned with the factual consistency between candidates and human references (or source texts). Those metrics rely on external systems, e.g.,

Likelihood

Generation based metrics judge the text quality of candidates in the form of text generation, based on

Machine Metrics

Lexicality

Candidate 1: The technique has many limitations.

Candidate 2: The approach shows numerous merits.

Reference: The technique has many advantages.

Lexical metrics do not identify synonyms and are fooled by candidate 1.

Machine Metrics

Semantics

Candidate 1: The approach shows numerous merits.

Reference: The technique has many advantages.

```
def is_synonym(word_a, word_b, epsilon=0.1):
    # word_a from candidate
    # word_b from reference
    # return matched or not
    if wordnet.is_synset(word_a, word_b):
        return 1
    if d(emb_word_a, emb_word_b) < epsilon:
        return 1
    return 0
```

Synonym Matching

A. WordNet

B. Word embeddings

Machine Metrics

Semantics

Candidate 1: The approach shows numerous merits.

Reference: The technique has many advantages.

- WordNet: limited coverage of words
- Word embeddings: relying on whether two words have similar companies, not on their word meanings

```
if wordnet.is_synset(word_a, word_b):
    return 1
if d(emb_word_a, emb_word_b) < epsilon:
    return 1
return 0
```

Machine Metrics

Discourse coherence

Coherence (linguistics)

20 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

This article's **tone or style may not reflect the encyclopedic tone used on Wikipedia**. The reason given is: **reasoned, academic style.** See Wikipedia's guide to writing better articles for suggestions. *(February 2021)* (*Learn how and when to remove this template message*)

Coherence in linguistics is what makes a text semantically meaningful. It is especially dealt with in text linguistics. Coherence is achieved through syntactical features such as the use of deictic, anaphoric and cataphoric elements or a logical tense structure, as well as presuppositions and implications connected to general world knowledge. The purely linguistic elements that make a text coherent are subsumed under the term cohesion.

Machine Metrics

Discourse coherence

Tom loves reading books.

He prefers reading books at the library.

Thus, he always goes to the library.

More coherent

Tom loves reading books.

He missed his lunch today.

Thus, he always goes to the library.

Less coherent

Coreference (anaphora)

Repetition

How linguists Interpret Text Coherence?

1. Mike: What kind of food do you like to eat?
2. Anne: I like ice-cream
3. Tom: I'm a vegetarian
4. Jane: Fruit is my favorite
5. Anne: Fruit's cool.
6. Anne: I had a banana for breakfast
7. Joe: I like ice cream too.
8. Mike: My favorite ice-cream flavor is chocolate.

Focus	Focus Transitions
1. food	
	Shift
2. Anne	
	Shift
3. Tom	
	Shift
4. Fruit	
	Continuation
5. Fruit	
	Shift
6. Anne	
	Shift
7. Joe	
	Shift
8. Ice-cream flavor	

Machine Metrics

Discourse coherence

Candidate

Chelsea have made an offer for FC Tokyo forward Yoshinori Muto. The 22-year-old will join Chelsea 's Dutch partner club Vitesse Arnhem on loan next season if he completes a move to Stamford Bridge. Chelsea signed a £200million sponsorship deal with Japanese company Yokohama Rubber in February.

Reference

Naoki Ogane says that Chelsea have made an offer for Yoshinori Muto. The 22-year-old forward has one goal in 11 games for Japan. Muto admits that it is an 'honor' to receive an offer from the Blues. Chelsea have signed a £200m sponsorship deal with Yokohama Rubber. Muto graduated from university with an economics degree two weeks ago. He would become the first Japanese player to sign for Chelsea.

- What are foci and focus transition?
- Are these similar in candidate and ref?

https://aclanthology.org/2023.eacl-main.278.pdf

Machine Metrics

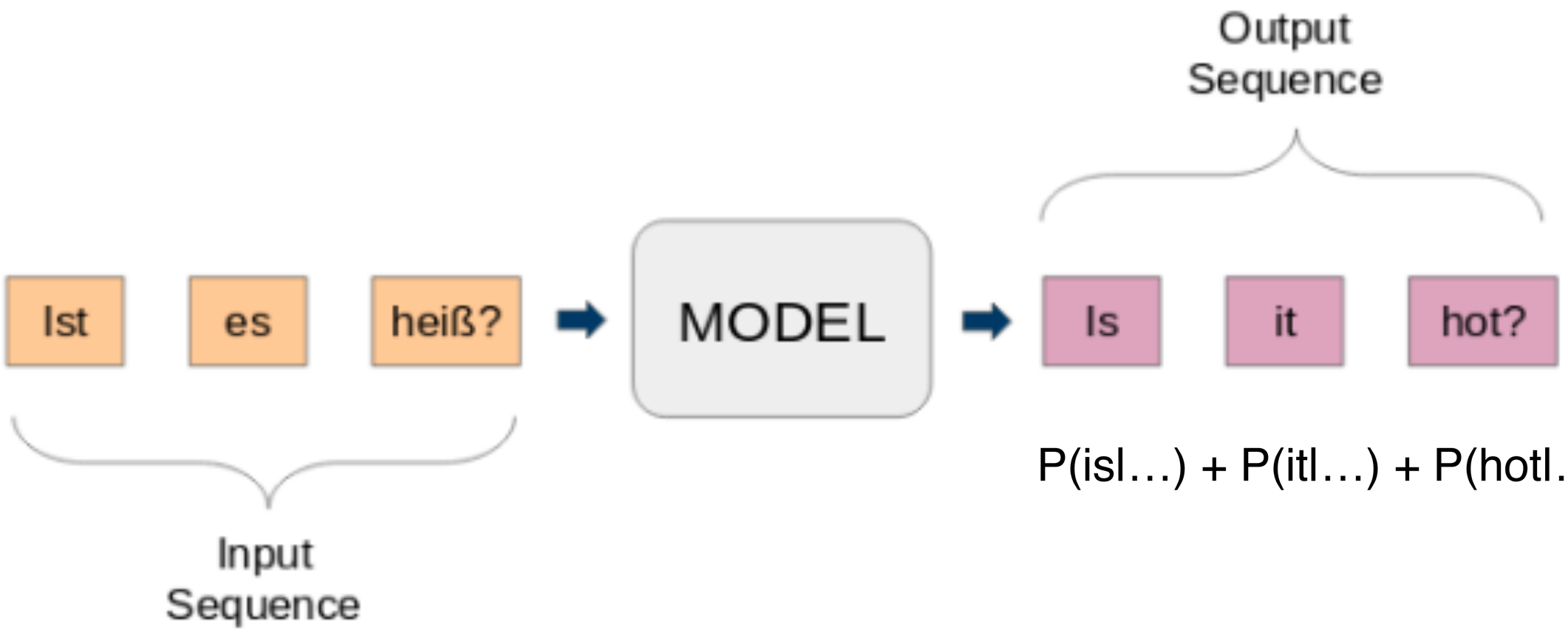
Factuality

Document	
The Knicks beat the Rockets. The fans were excited.	
Summary	
The Knicks beat the Bucks.	
Entailment Matrix [Contra, Neutral, Support] $\begin{bmatrix} 0.90 & 0.07 & 0.03 \\ 0.02 & 0.90 & 0.08 \end{bmatrix}$	**Selected Answer** the Bucks
	Generated Question Who did the Knicks beat?
	QA Output the Rockets
Max Support Score 0.08	**Answer Overlap Score** 0.20

https://aclanthology.org/2022.naacl-main.187.pdf

Machine Metrics

Likelihood



P(isl…) + P(itl…) + P(hotl…)

https://openreview.net/pdf?id=5Ya8PbvpZ9

Outline

About this lecture

- Evaluation Metrics
- System Comparison
- Explainability

System Comparison

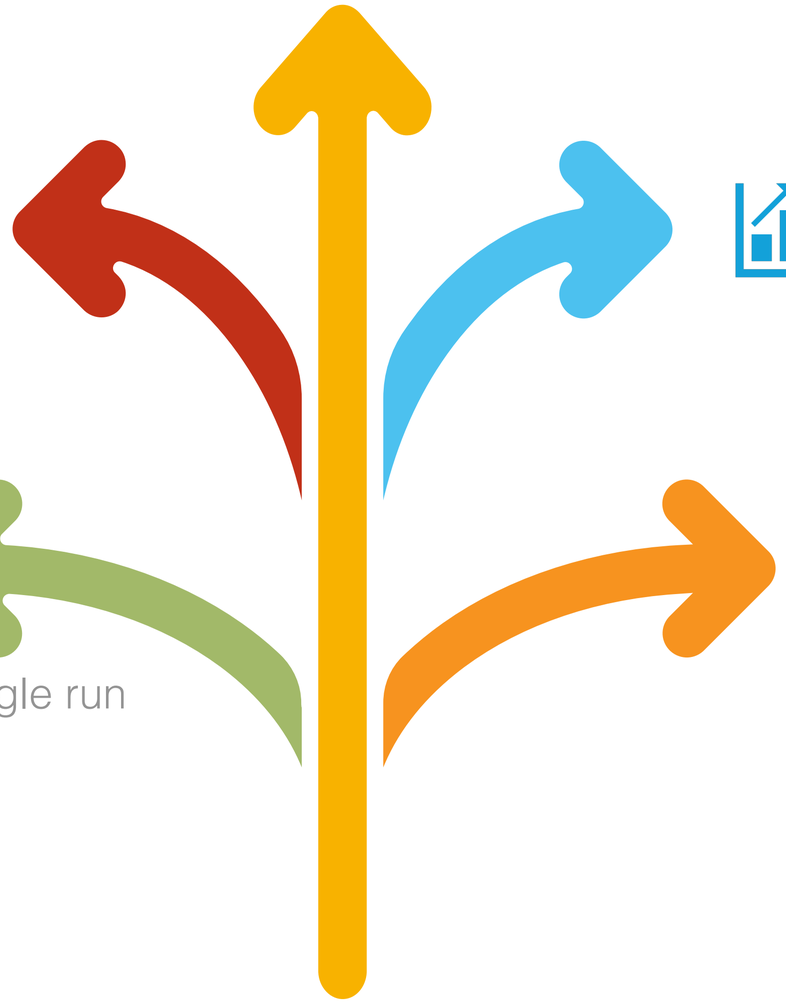
Overview

Significance Test

When the size of test set is small

Computational Budget

Linear regressions over neural systems



Multi-Run Results

Reporting score distribution over over single run

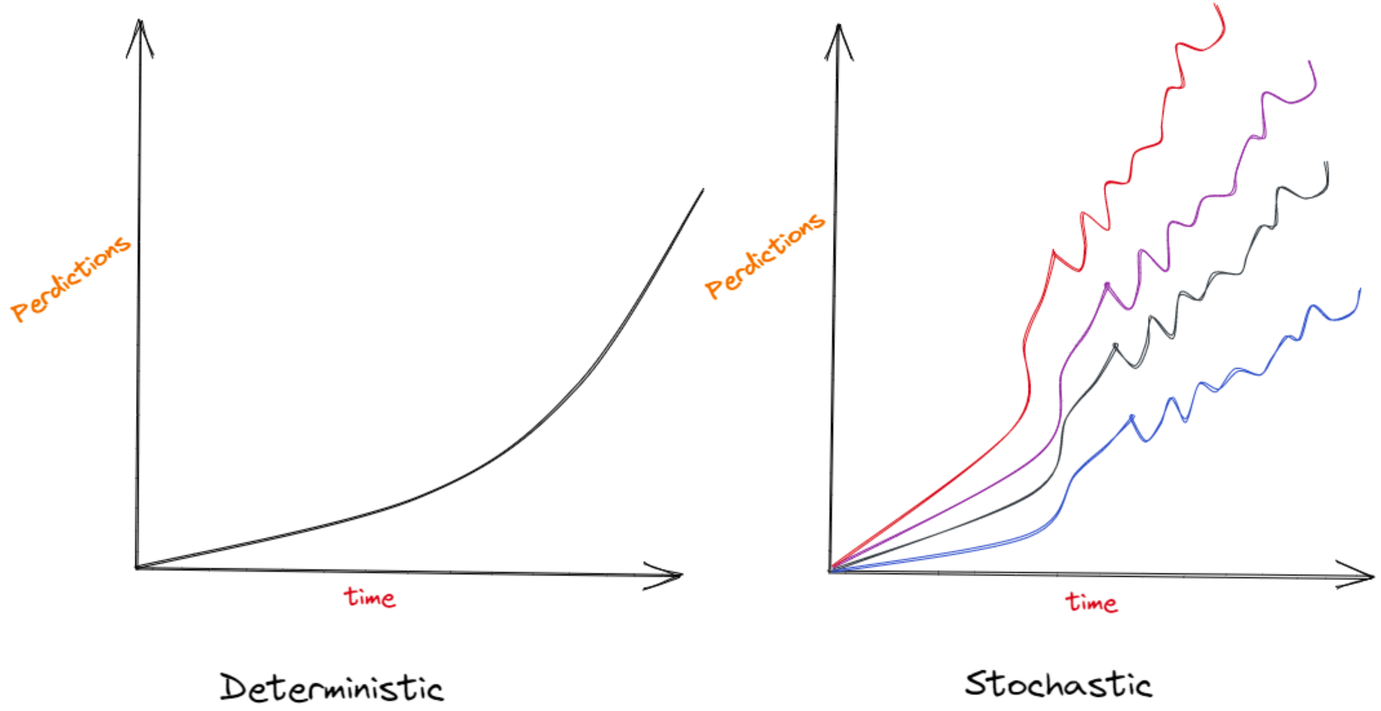
Fair Configuration & Simpson's paradox

Consistent scheme of data processing

Misuse of statistics

System Comparison

Multiple runs



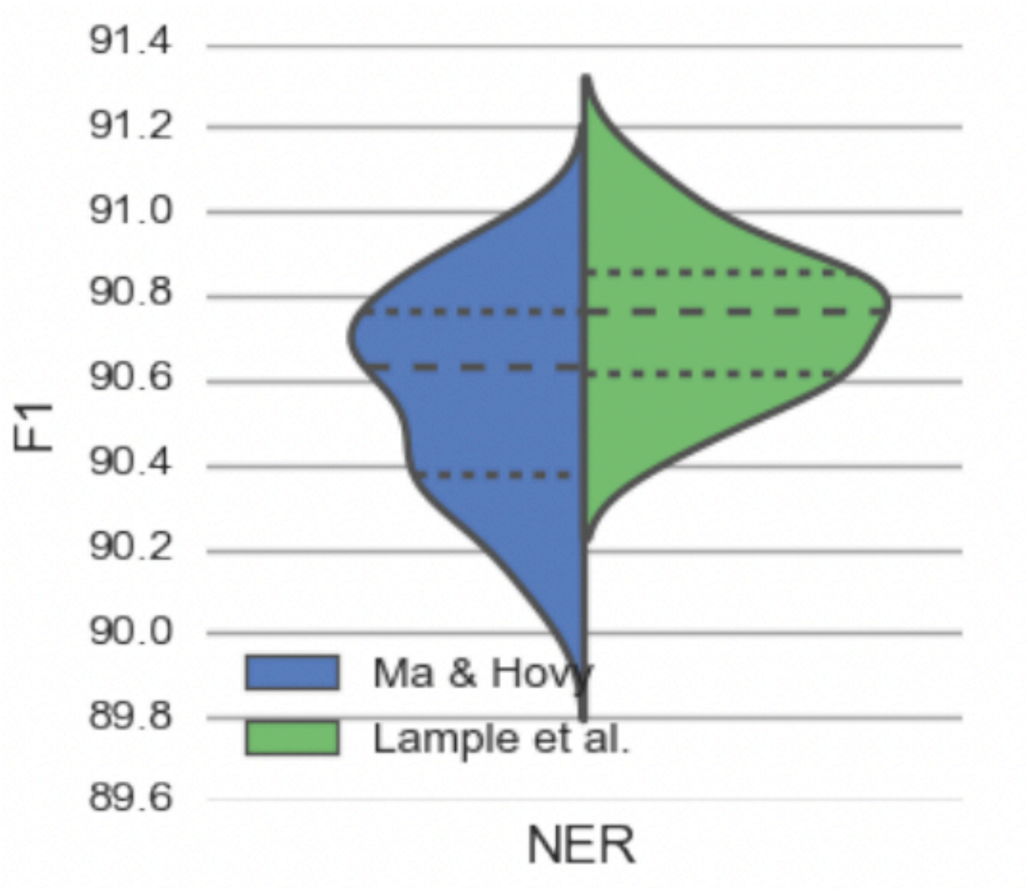
Randomness

1. Weight initialization
2. Data shuffling
3. Dropout techniques

Why?

System Comparison

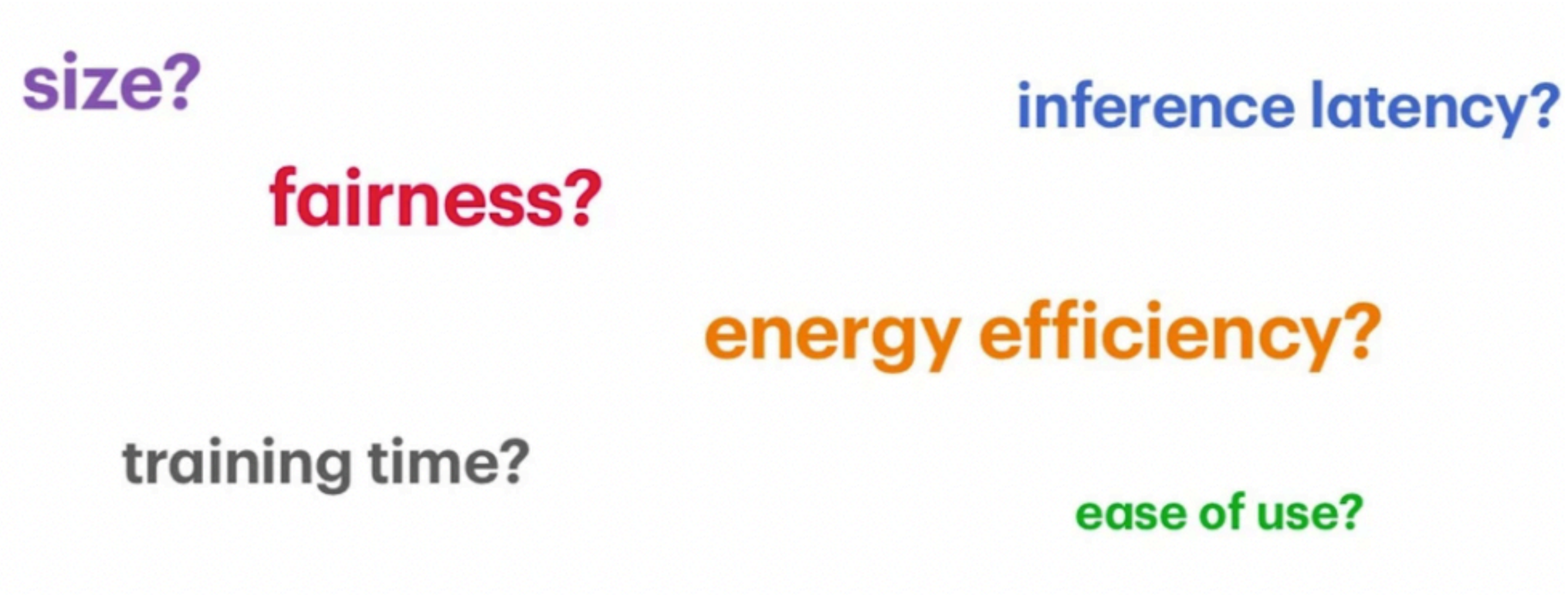
Multiple runs - the results are not reproducible



https://aclanthology.org/D17-1035.pdf

System Comparison

Computational budget

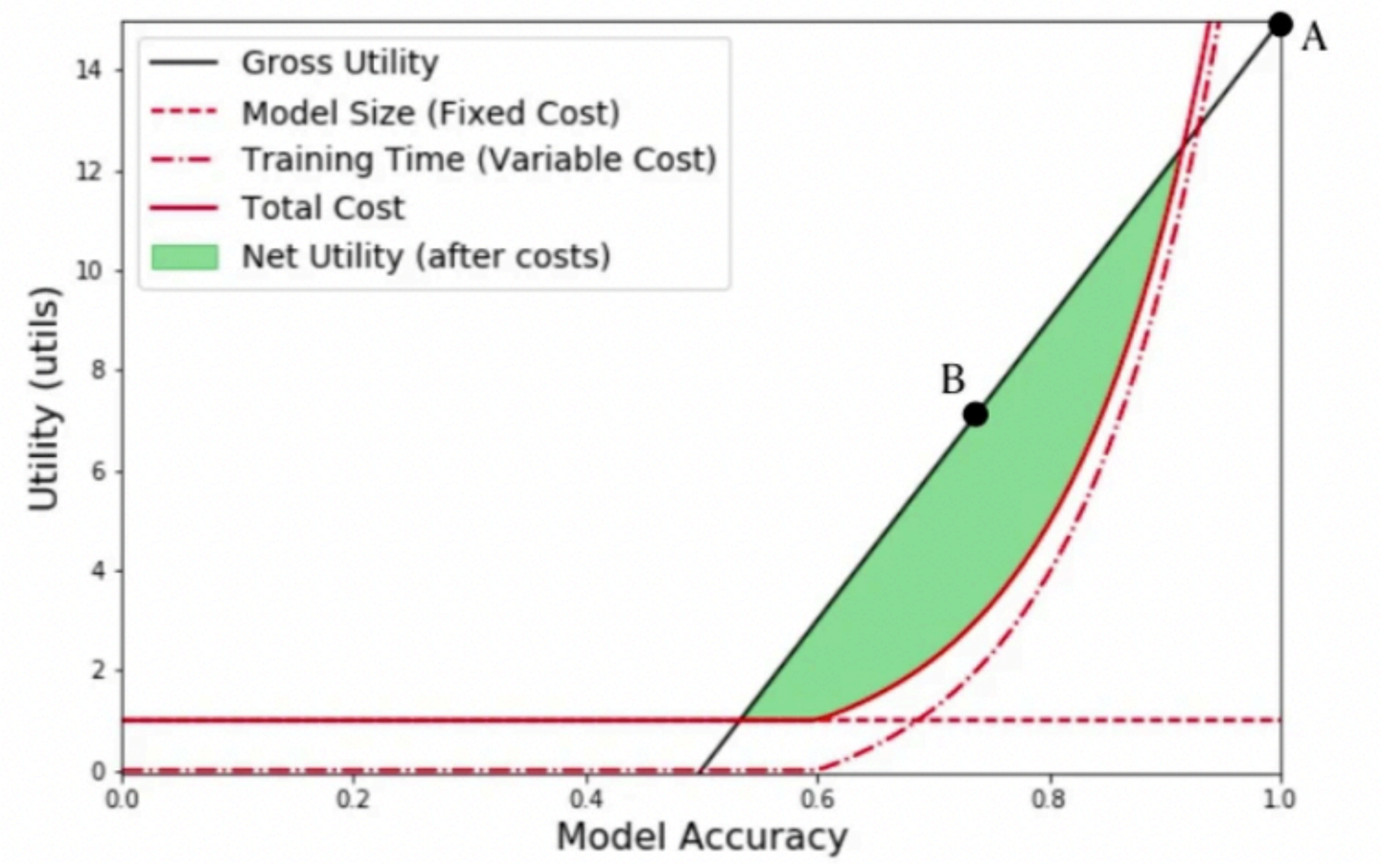


Accuracy does not always take the priority, sometimes one wants to make an economic choice.

System Comparison

Computational budget

Utility = the benefit that NLP Practitioners gain from using a model for a task



https://aclanthology.org/2020.emnlp-main.393.pdf

Outline

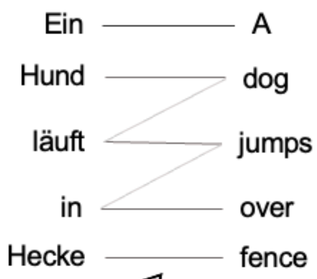
About this lecture

- Evaluation Metrics
- System Comparison
- Explainability

Explainability

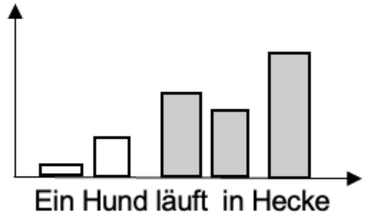
Overview in evaluation metrics

Visual Explanation



Visualizing alignment weights obtained from MoverScore

Feature Importance Explanation



A distribution of word-level regarding translation quality

Local Explanation

X: Ein Hund läuft in die Hecke

Y: A dog jumps over the fence

Gold Explanation Source: [0, 0, 1, 1, 1]

Gold Explanation Translation: [0, 0, 1, 1, 1]

Identification of translation errors to explain sentence-level translation quality

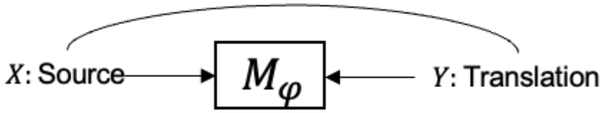
Text Explanation

#1 Generated rationale: the translation receives a rating score of 0.7 because 'läuft', 'in' and 'Hecke' are mistranslated.

Providing justifications in the form of generated text

Concepts

$m(X,Y)$: Evaluation Metric



Explanation By Simplification

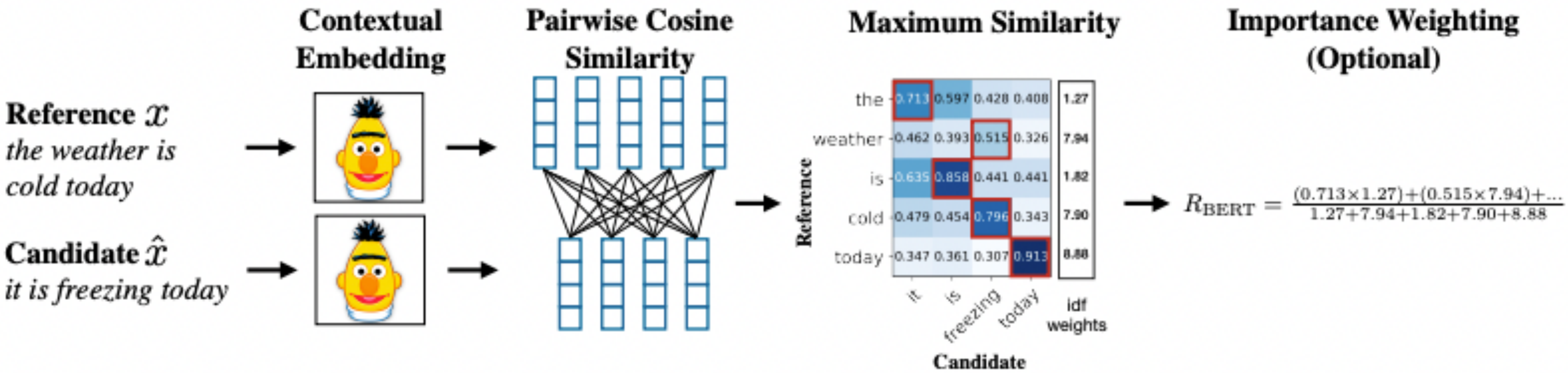
$$m(X,Y) = a \cdot sem(X,Y) + b \cdot syn(X,Y) + c \cdot lex\,(X,Y) + d \cdot morph\,(X,Y) + \varepsilon$$

$$m(X,Y) = \mathcal{R}(x) - \mathcal{R}(y) + \varepsilon$$

Using a simpler regression to mimic the judgments of non-transparent metrics

Explainability

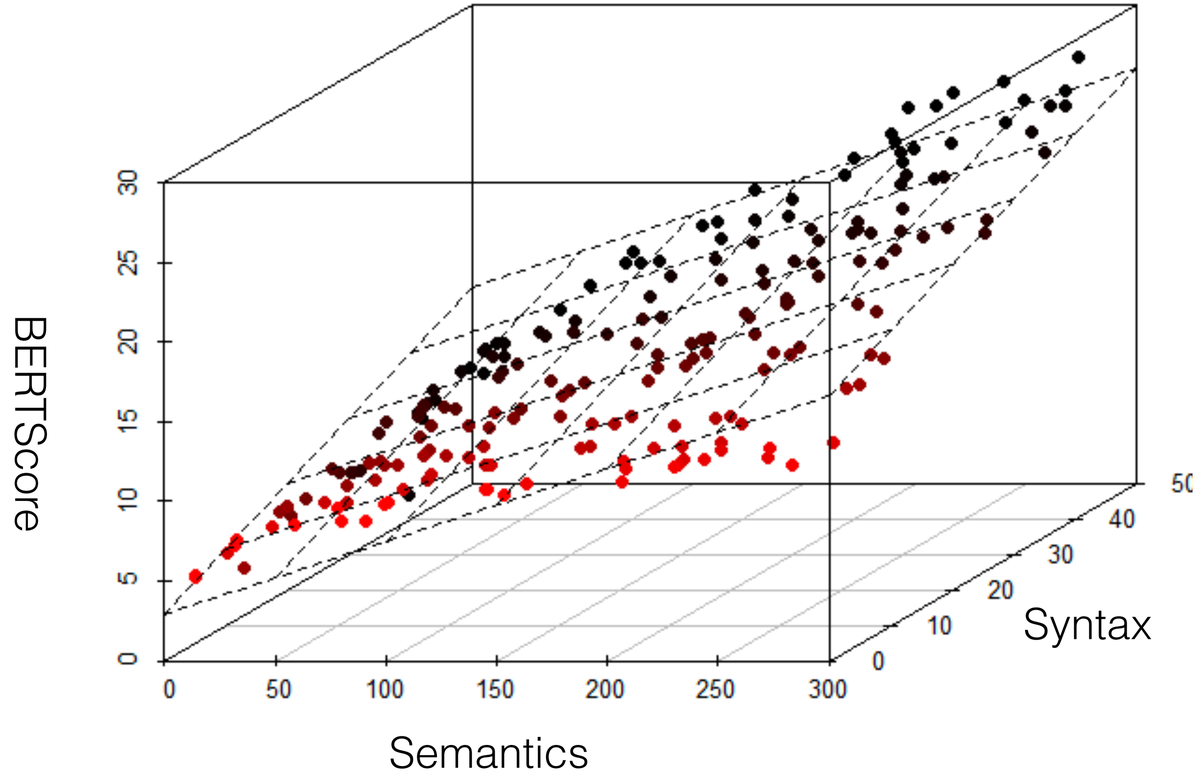
Visual explanation in metrics



https://openreview.net/pdf?id=SkeHuCVFDr

Explainability

Explanation by simplification



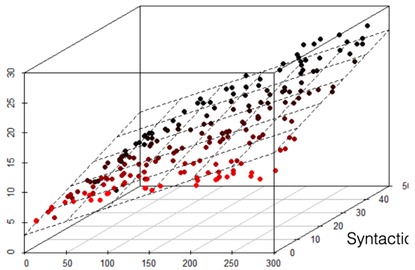
$$\mathscr{L} = \{p \mid p^{\intercal}\beta + \beta_0 = 0\}$$

$$p_{\text{BERTScore(x,y)}} = \beta_1 \cdot p_{\text{SEM(x,y)}} + \beta_2 \cdot p_{\text{SYN(x,y)}} + \beta_0$$

https://aclanthology.org/2021.emnlp-main.701.pdf

Explainability

Explanation by simplification



$$p_{\text{BERTScore(x,y)}} = \boxed{\beta_1} \cdot p_{\text{SEM(x,y)}} + \boxed{\beta_2} \cdot p_{\text{SYN(x,y)}} + \boxed{\beta_3} \cdot p_{\text{LEX(x,y)}} + \boxed{\beta_4} \cdot p_{\text{MOR(x,y)}} + \beta_0$$

Candidate x	Source/Reference y	SEM	SYN	LEX	MOR
It is a boy , likes to sport , but it cannot do it because of their very.	He is a boy, he likes sports but he can't take part because of his knee.	-1.57	0.98	-0.59	-0.87
Zwei Besatzungsmitglieder galten als vermisst.	Two crew members were regarded as missing.	0.83	0.99	0.46	-2.40

Take-aways

Things to remember

- Evaluation is the goalkeeper, keeping track of progress
- Evaluation: metrics x system comparison x explainability
- Metrics: they should agree with human judgments
- System comparison: how to use metrics properly
- Explainability: metric judgments should be transparent