Is My NLP Model Working?
The Answer is Harder Than You Think

Graham Neubig

Carnegie Mellon University
Language Technologies Institute

INSPIRED COGNITION

Microsoft

Search Microsoft Research

Research

Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan Awadalla, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao C
Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, Ming Zhou
March 2018
arXiv:1803.05567

Flo Crivello
@Altimor

GPT3 has already replaced much of my Google usage, and almost all my Wikipedia usage. (Forgive the naive questions!)

The CIA orchestrated a coup in Iran in 1953 in order to overthrow the democratically elected government of Prime Minister Mohammad Mossadegh. The CIA did this because Mossadegh had nationalized Iran's oil industry, which was previously controlled by a British company. The CIA's intervention restored control of Iran's oil industry to Western companies.

producer of oil at the time.

in percentage terms?

Iran produced about 10% of the world's oil at the time.

that seems minor, why would the US be so worried about it?

the time, the US was heavily dependent on oil imports from the Middle East. If Iran's oil industry had remained nationalized, it could have reduced the amount of oil available to the US and driven up oil prices.

What percent of oil production at the time came from Iran?

The US was heavily dependent on oil imports from the Middle East at the time. So even a small reduction in the amount of oil available could have had a major impact on the US economy.

Has the US reduced its dependence on middle eastern oil since then?

Yes, the US has reduced its dependence on

ARTIFICIAL INTELLIGENCE

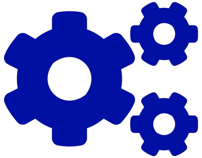
Google Engineer Claims AI Chatbot Is Sentient: Why That Matters

Is it possible for an artificial intelligence to be sentient?

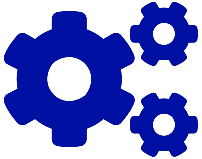
By Leonardo De Cosmo on July 12, 2022

A New Paradigm in NLP: Text Generation through LMs+Prompting

Q: What is the capital of Pennsylvania?

A:  Harrisburg

It’s for real this time. After months of legal drama, bad memes and will-they-or-won’t-they-chaos to put your favorite rom-com to shame, Elon Musk has closed his $44 billion acquisition of Twitter. Musk sealed the deal Thursday night, taking Twitter private and ousting a handful of top executives — CEO Parag Agrawal included — in the process.

TL;DR:  Elon Musk has bought Twitter.

How Good is Resulting Text?

How do the muscles in your arm work?

The muscles in your arm work by contracting and pulling on the bones in your arm. This movement is possible because the muscles are attached to the bones by tendons.

Why do we sleep? Why do we dream?

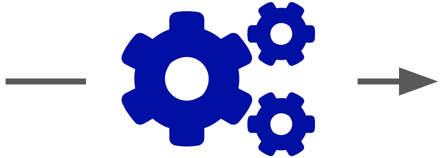
We sleep to give our bodies and brains a chance to rest and repair themselves. Dreams may help us process and store memories, or they may be a way for our brains to work through problems or anxiety.

From GPT-3 versus Google Search by David E. Weekly

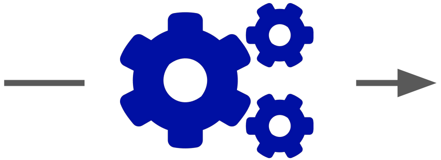
https://twitter.com/dweekly/status/1580676295444217857

How Bad is Generated Text? - Factuality

Q: What are the largest states in the US by population and surface area?

A:  Alaska and Texas

Q: What is the net worth of the CEO of Twitter?

A:  Jack Dorsey is the CEO of Twitter.

He is worth $2.5 billion.

How Bad is Generated Text? - Coherence / Plausibility

This is a story about a programmer on her first day of work.

She was given a task to write a program that would print the numbers from 1 to 100.

But for multiples of three print “Fizz” instead of the number and for the multiples of five print “Buzz”.

For numbers which are multiples of both three and five print “FizzBuzz”.

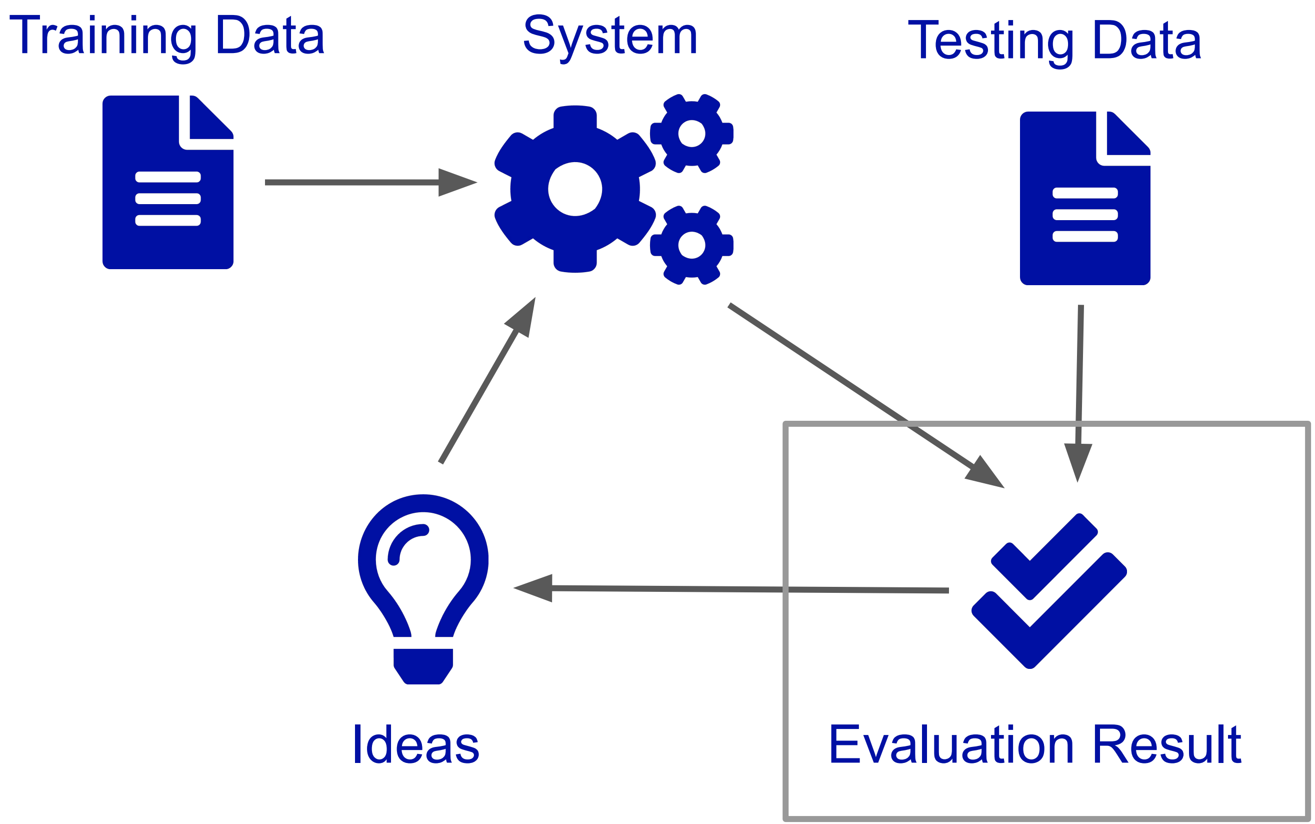
She was given a week to complete the task.

She was given a week to complete the task.

She was given a week to complete the task.

Evaluating generated text is now about as hard as generating it.

The NLP Development Pipeline



How Hard is Evaluation?

Classification → easy, measure **exact match**

Translation → more difficult, many different good translations with the **same semantics**

Dialog → even more difficult, many different good answers with **different semantics**

Quiz: What about *question answering*? *Summarization*?

The Gold-standard?: Manual Evaluation

Source

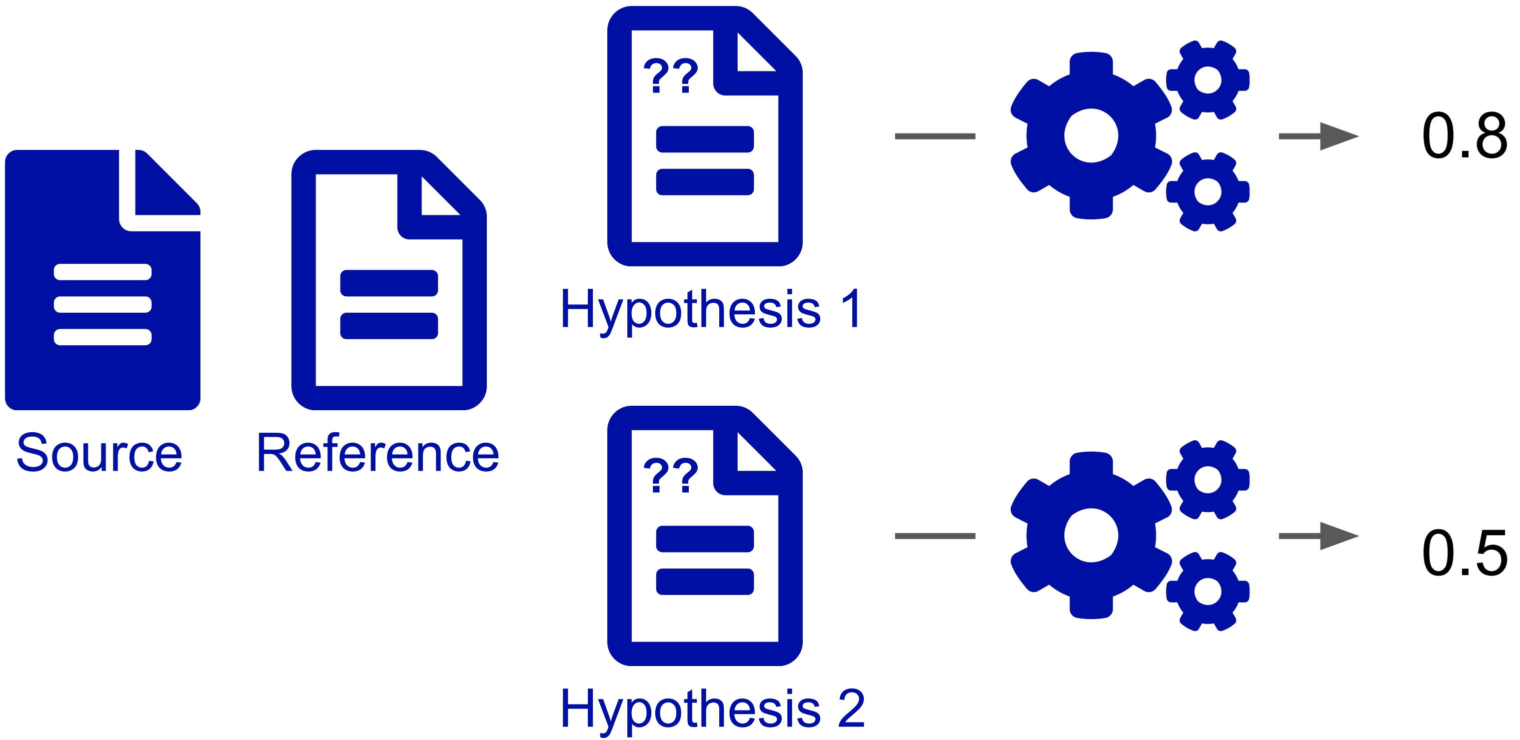
?? Hypothesis 1

?? Hypothesis 2

→ 0.8

→ 0.5

An Alternative: Automatic Evaluation



The Old Reliables: BLEU/ROUGE Score

Reference: I am giving a talk at a data science conference

Hyp 1: I am giving a talk at a conference about data science

lots of overlap → high score

Hyp 2: This talk is about recent advances in medical imaging

little overlap → low score

Why is Evaluation Hard?

Reference: I am giving a talk at a data science conference

Hyp 1: I am giving a talk at a political science conference

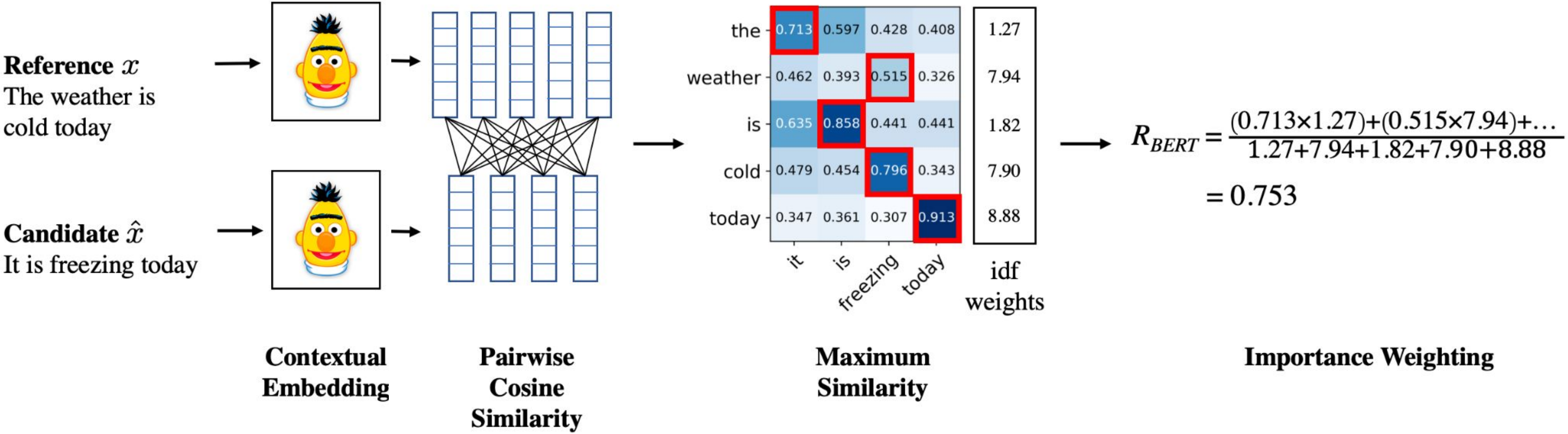
lots of overlap but bad output

Hyp 2: My lecture will be given to the meeting on data analytics

little overlap but good output
(particularly difficult for open-ended problems)

Embedding-based Evaluation

BERTScore

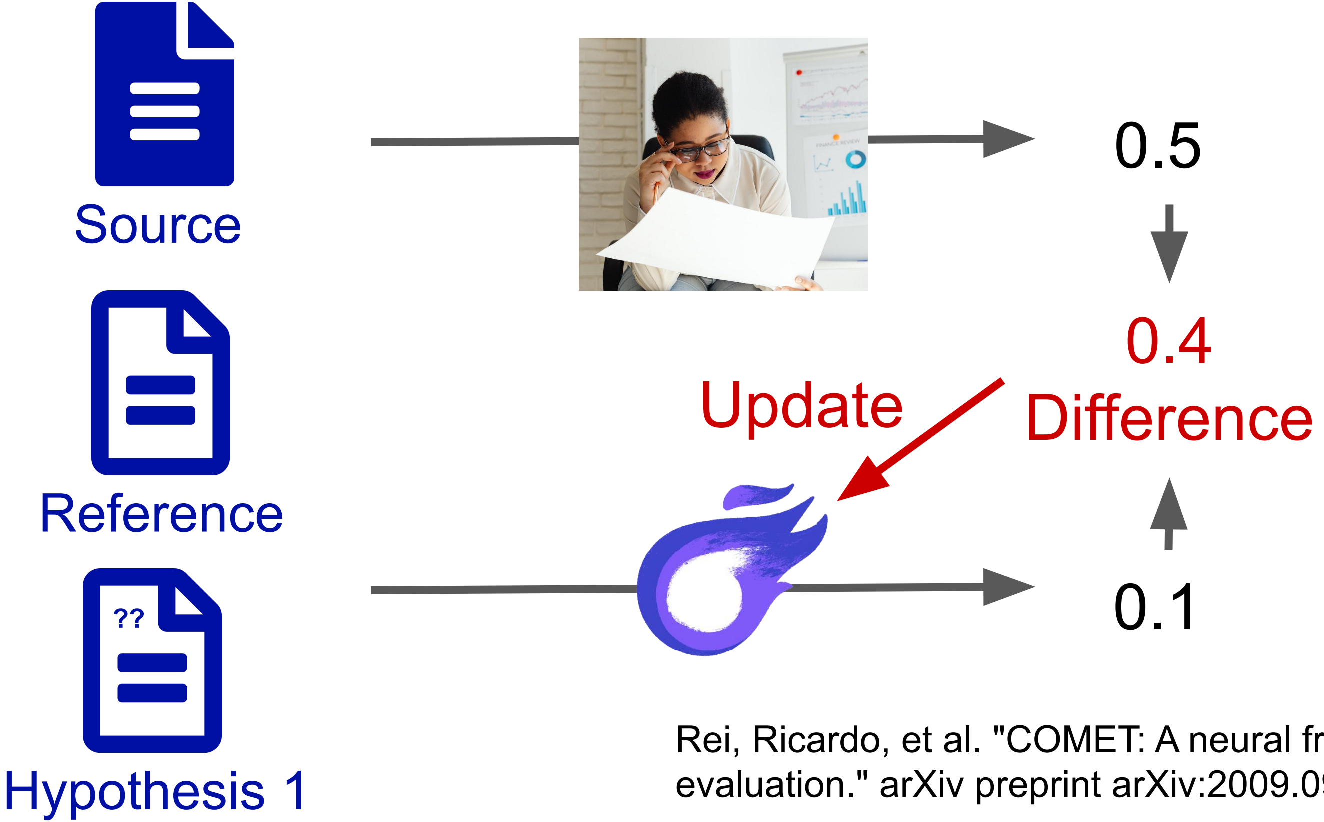


Zhang, Tianyi, et al. "Bertscore: Evaluating text generation with bert." arXiv preprint arXiv:1904.09675 (2019).

https://github.com/Tiiiger/bert_score

Learning to Evaluate

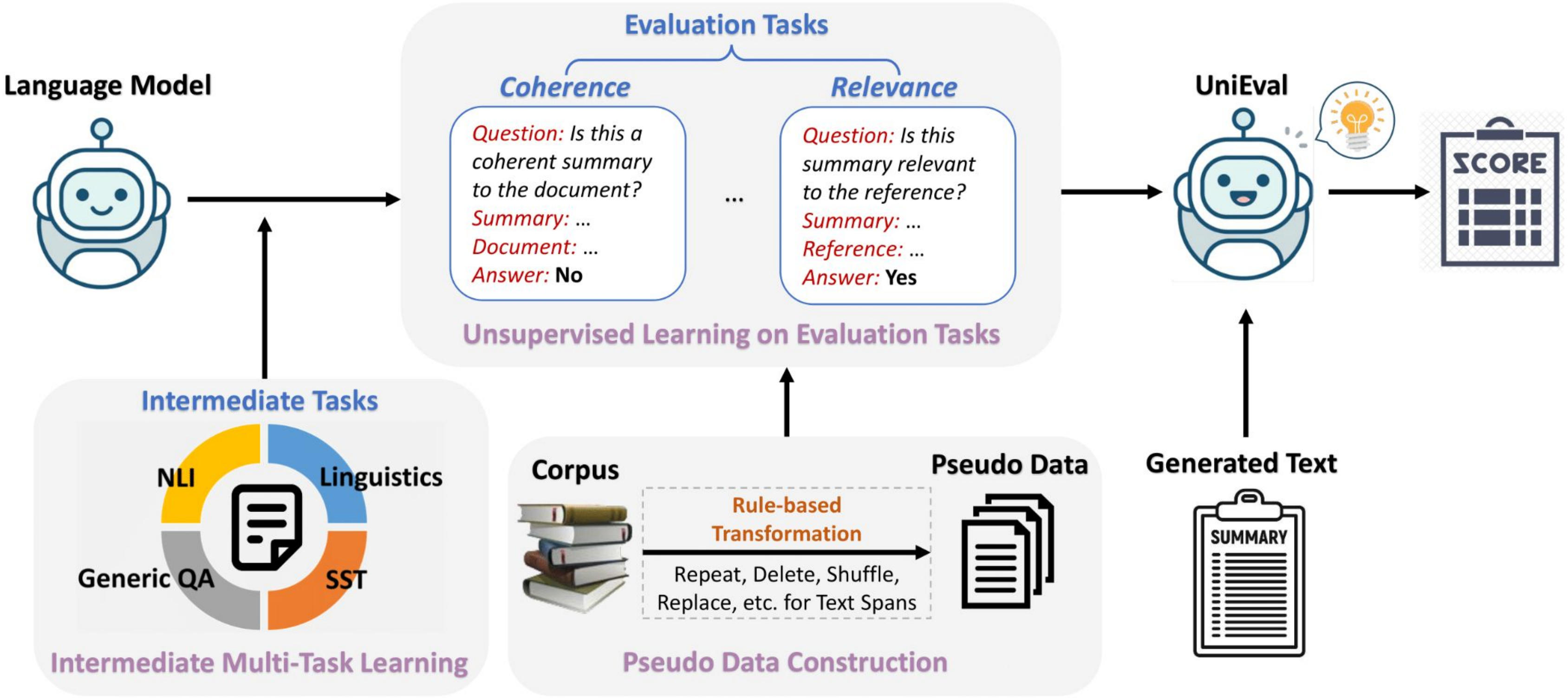
COMET
by Unbabel



Rei, Ricardo, et al. "COMET: A neural framework for MT evaluation." arXiv preprint arXiv:2009.09025 (2020).

https://unbabel.github.io/COMET/

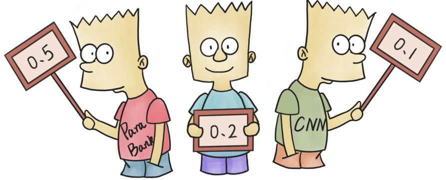
Learning to Evaluate w/ Pseudo-data UniEval



Zhong, Ming, et al. "Towards a Unified Multi-Dimensional Evaluator for Text Generation." (2022).

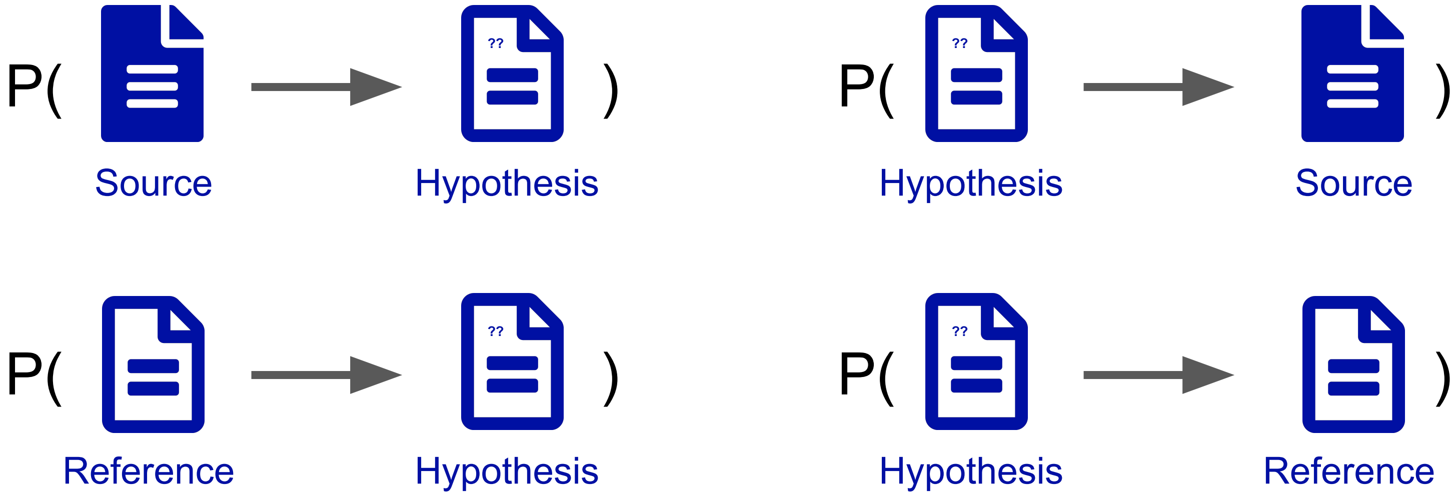
https://github.com/maszhongming/UniEval

Generative Text Evaluation



BARTScore

Use the probability of a *generative* model to evaluate text

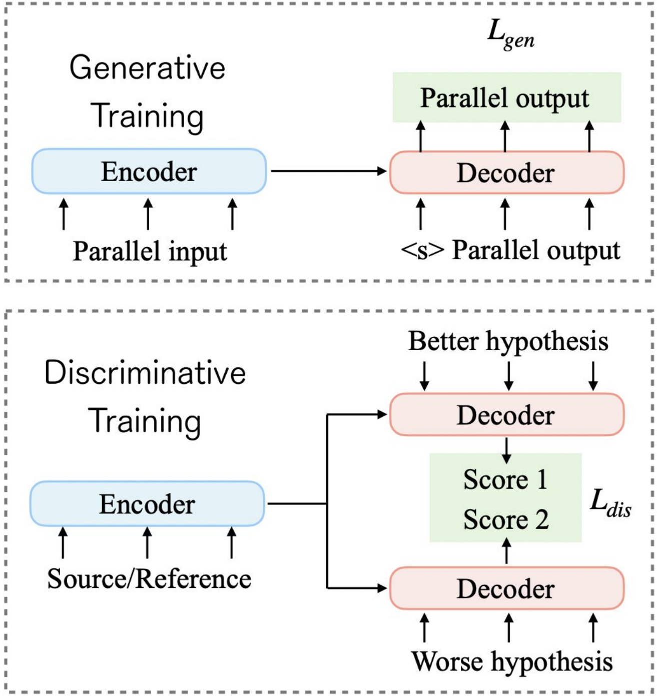


Yuan, Weizhe, Graham Neubig, and Pengfei Liu. "Bartscore: Evaluating generated text as text generation." Advances in Neural Information Processing Systems 34 (2021): 27263-27277.

https://github.com/neulab/BARTScore

Generative Pre-training, **Discriminative** Fine-tuning

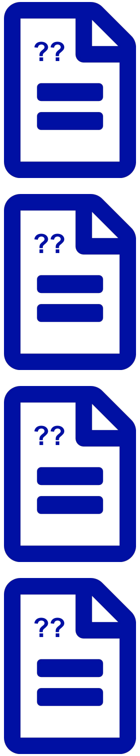
T5Score



Qin, Yiwei, et al. "T5Score: Discriminative Fine-tuning of Generative Evaluation Metrics." (2022).

https://github.com/qinyiwei/T5Score

How Do We **Evaluate Evaluation**?



Human	Automatic
0.8	0.7
0.5	0.1
0.1	0.5
0.6	0.4

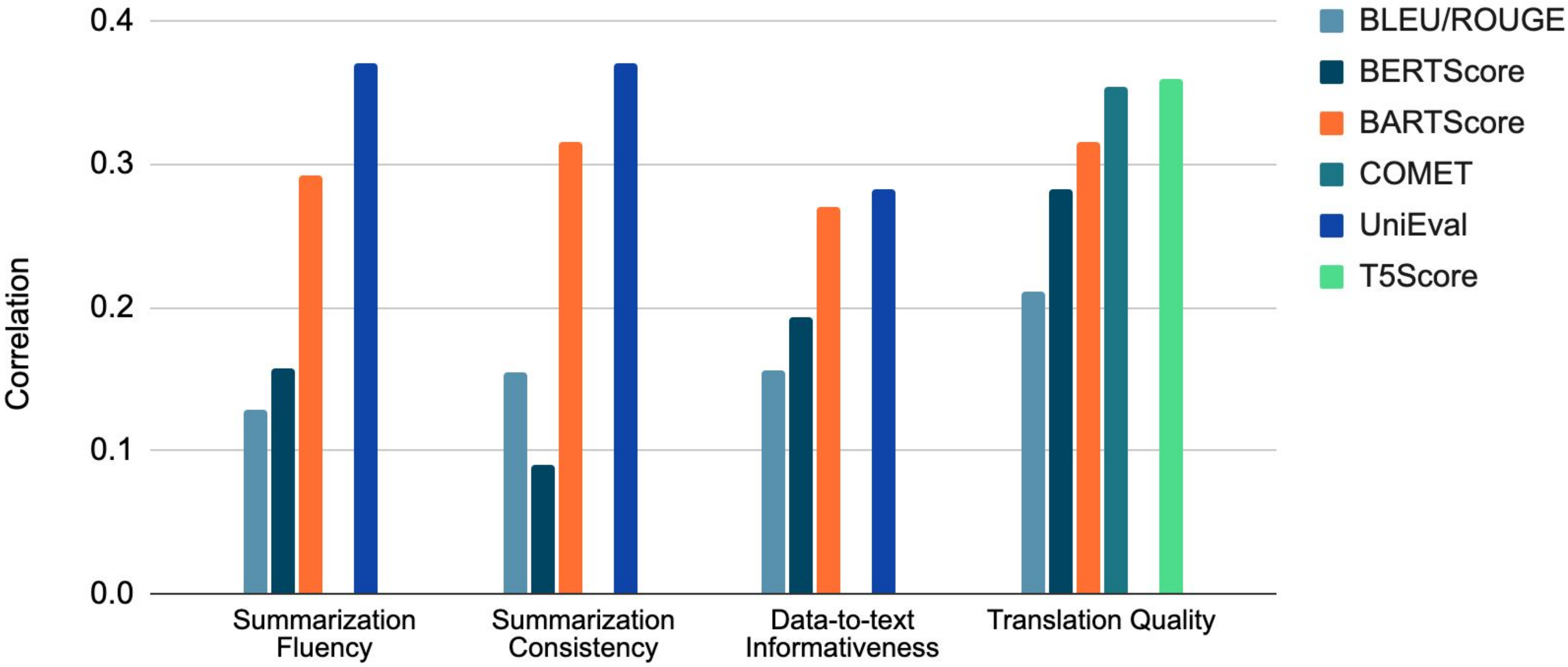
→

Correlation

Pearson = 0.23

Kendall = 0.33

Meta-Evaluation Results



So many evaluation metrics! What to do next?

- Multi-metric evaluation
- Metric-aware training/inference
- Fine-grained analysis

Multi-dimensional evaluation of text-generation tasks.

Current Text Generation Evaluation Standard

System	R-1	R-2	R-L
	CNNDM		
BART*	44.16	21.28	40.90
PEGASUS*	44.17	21.47	41.11
GSum*	45.94	22.32	42.48
ConSum*	44.53	21.54	41.57
SeqCo*	45.02	21.80	41.75
GOLD-*p**	45.40	22.01	42.25
GOLD-*s**	44.82	22.09	41.81
SimCLS*	46.67	22.15	43.54
BART‡	44.29	21.17	41.09
BRIO-Ctr	47.28†	22.93†	44.15†
BRIO-Mul	**47.78**†	**23.55**†	**44.57**†

https://arxiv.org/abs/2203.16804

Why are we stuck?

- Running evaluation is slow, requires software install, GPU
- There's always other things to do!

Critique: A Simple Evaluation API for Text

Summaries Translations Dialogs Question Answers

INSPIRED CRITIQUE

Summary Quality Translation Quality Fluency Toxicity Factual Consistency

https://docs.inspiredco.ai/critique/

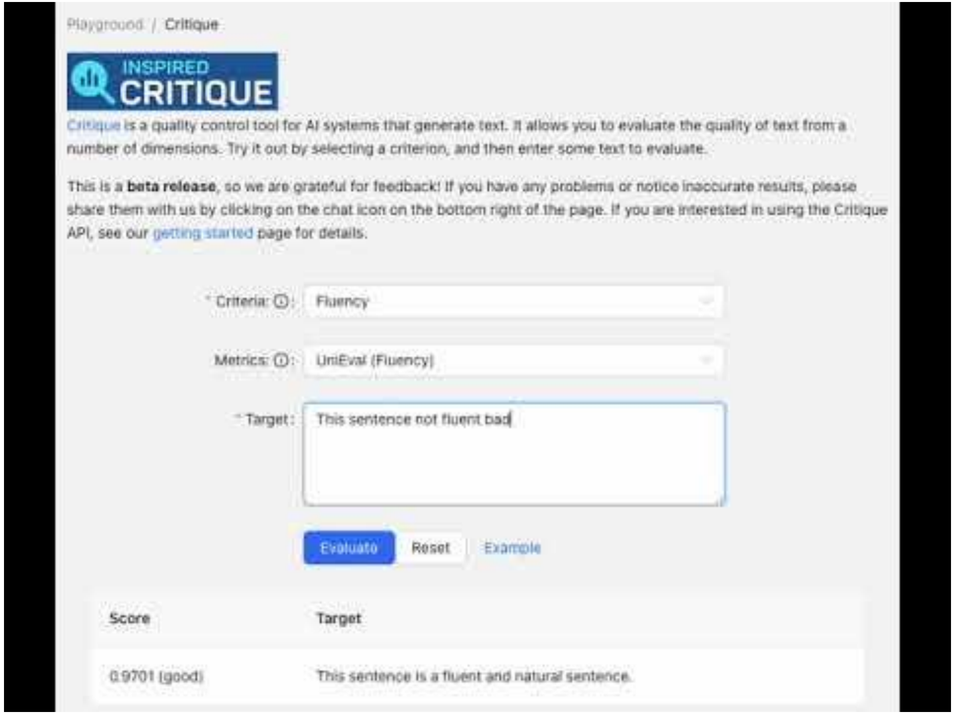
API-based Usage

```
import os
from inspiredco.critique import Critique

client = Critique(api_key=os.environ["INSPIREDCO_API_KEY"])

dataset = [
    {"target": "This is a really nice test sentence."},
    {"target": "This sentence not so good."},
]

results = client.evaluate(
    metric="uni eval",
    config={"task": "summarization", "evaluation_aspect": "fluency"},
    dataset=dataset,
)

for datapoint, result in zip(dataset, results["examples"]):

    print(f"Text: {datapoint['target']}, Fluency: {result['value']}")
```

- Several lines of code, trivial installation, no GPUs required

https://docs.inspiredco.ai/critique/

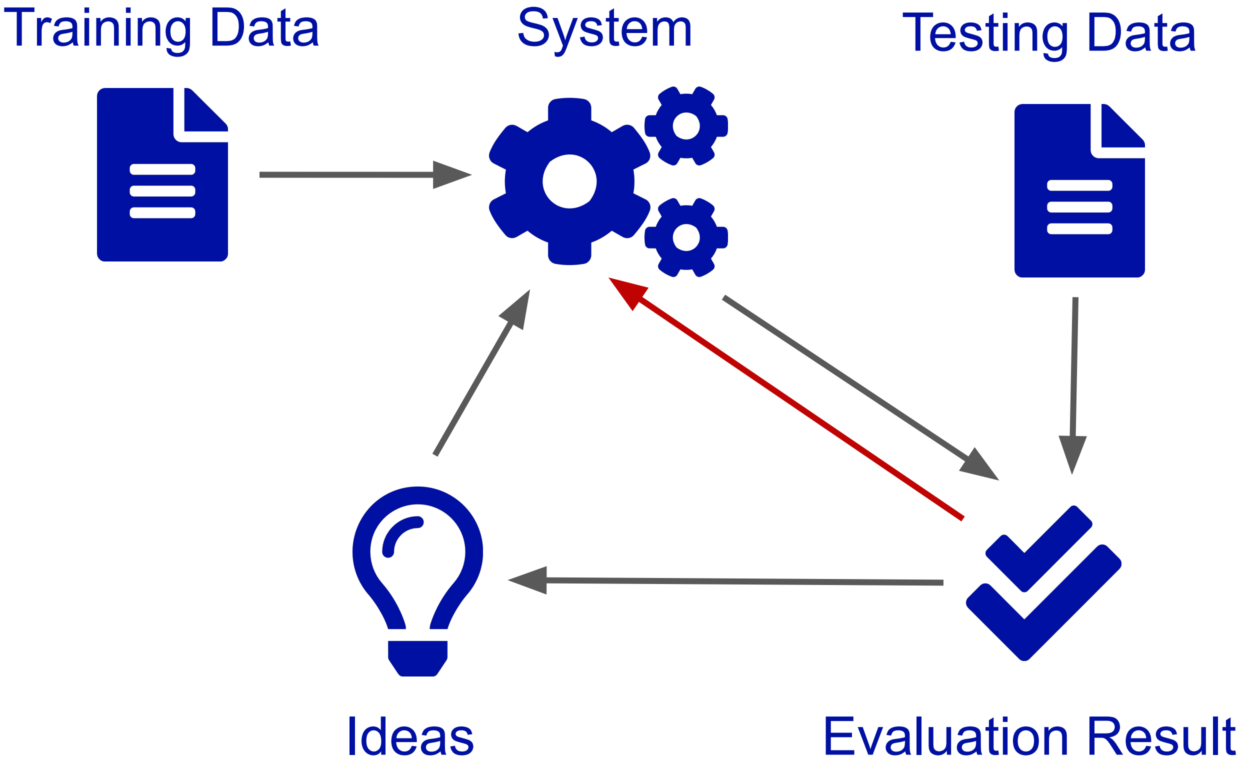
Online Interface



https://dashboard.inspiredco.ai/

Metric-aware Training/Inference

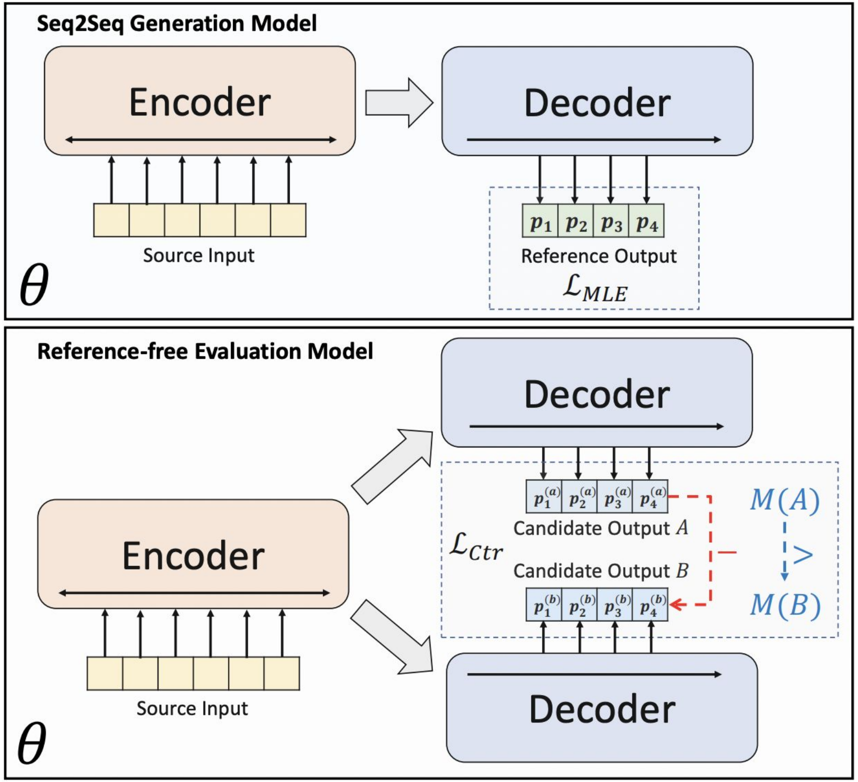
To the next step!



Metric-aware Training

Standard MLE Loss

Ranking Loss Based on Metrics



Liu, Yixin, et al. "BRIO: Bringing order to abstractive summarization." (2022).

https://arxiv.org/abs/2203.16804

Metric-aware Reranking

- Sample a bunch of outputs and rerank according to metrics
- Reference-free metrics, just rerank according to metric
- Reference-using metrics, use **minimum Bayes risk**

	Error Matrix (e.g. 1-metric)	Probability		Bayes Risk
My name is Bob.	0.0 0.9 0.9 0.9	0.2		0.486
This is true.	0.9 0.0 0.5 0.5	0.19	=	0.355
This isn't true.	0.9 0.5 0.0 0.1	0.18		0.292
This is not true.	0.9 0.5 0.1 0.0	0.17		0.293

Fernandes, Patrick, et al. "Quality-aware decoding for neural machine translation." (2022).

https://arxiv.org/abs/2205.00978

Metric-aware Prompt Optimization

- Prompting methods are hard to train can benefit from systematic analysis

Prompt Gym

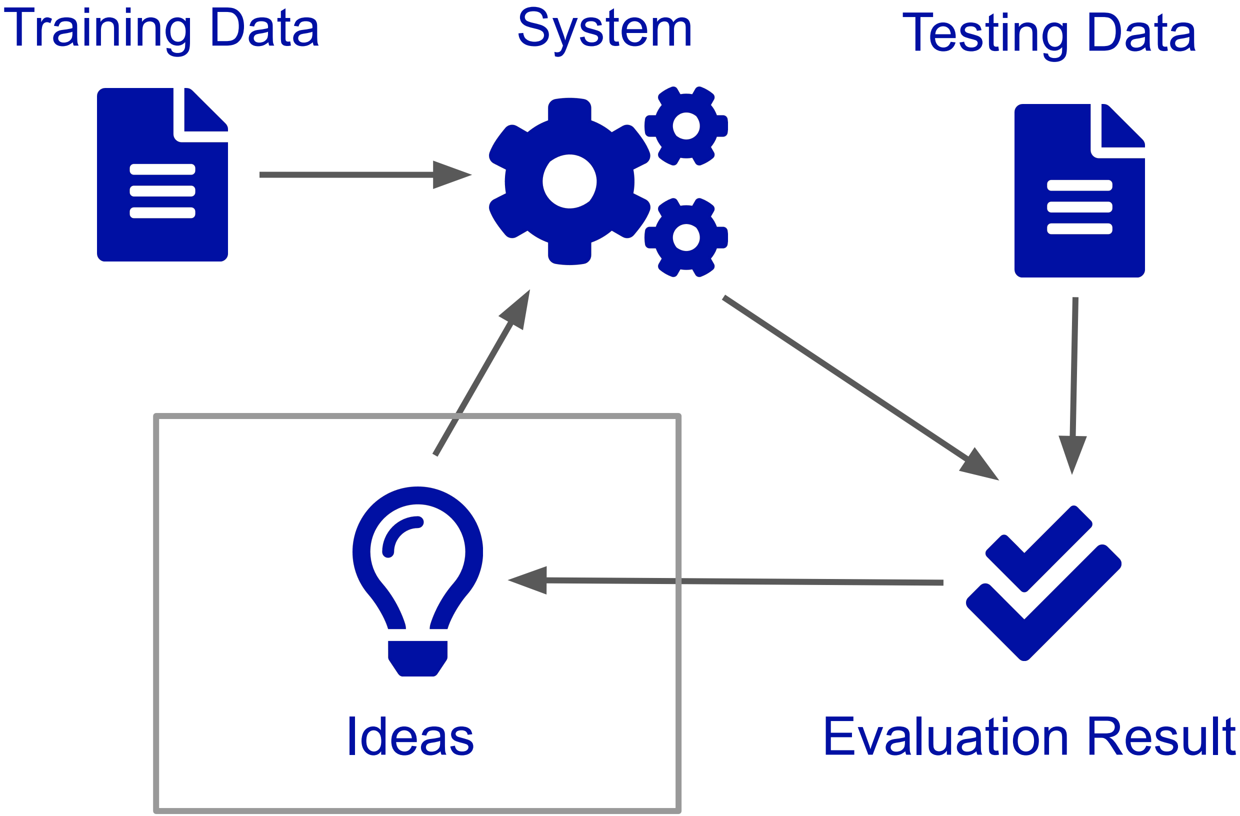
Evaluate different models, different prompts, different metrics

Model	Prompt	UniEval (Consistency)	UniEval (Coherence)	UniEval (Fluency)	UniEval (Relevance)	BartScore (Coverage)	Length Ratio
cohere_medium	standard	0.7466	0.4006	0.8869	0.3438	-3.4095	2.5533
cohere_medium	tldr	0.5006	0.2967	0.8539	0.3312	-3.1348	2.5800
cohere_medium	concise	0.8542	0.6115	0.9140	0.6167	-3.4220	2.4500
cohere_medium	complete	0.8331	0.4845	0.8825	0.5214	-3.1689	2.6767
openai_babbage_001	standard	0.9409	0.9036	0.8782	0.7975	-3.4083	2.0800
openai_babbage_001	tldr	0.8728	0.9072	0.9593	0.8145	-3.5234	1.0200
openai_babbage_001	concise	0.9483	0.9365	0.8669	0.8431	-3.2528	2.2800
openai_babbage_001	complete	0.9306	0.8278	0.8634	0.6951	-3.2720	2.2633
openai_ada_001	standard	0.6750	0.7270	0.8850	0.8174	-3.6719	2.0067
openai_ada_001	tldr	0.7999	0.7122	0.7973	0.6728	-3.7436	1.5300
openai_ada_001	concise	0.7776	0.7439	0.8106	0.5852	-3.6096	2.3600
openai_ada_001	complete	0.7732	0.5008	0.7332	0.3283	-3.5246	2.4567

https://github.com/inspired-cognition/prompt-gym/

Fine-grained Analysis and Understanding of NLP Models

To the next step!



NLP Debugging: Understanding the Flaws in Our Systems

- We have a number, but where do we go next?
- **Fine-grained aggregate analysis**

“Your model is under-performing on short sentences.”

- **Case studies**

“Caution, potentially incorrect sentence:”

Source: Voda byla skvělá.

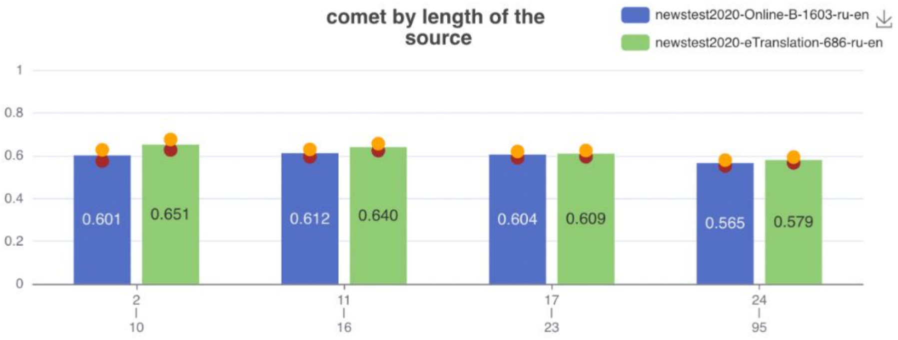
Reference: The water was great.

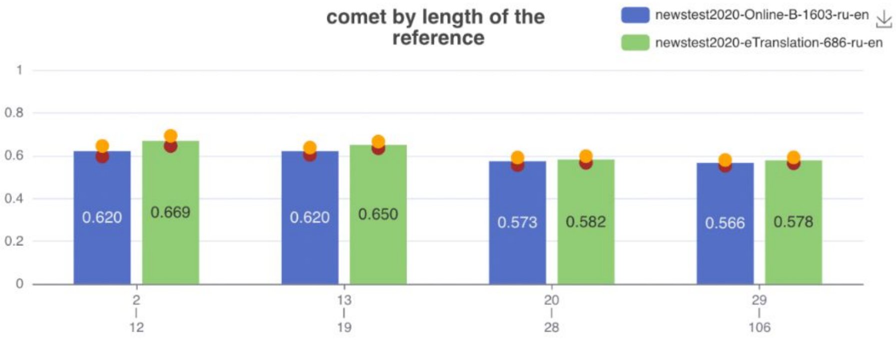
Hypothesis: The water was.

A Case Study: Russian-English Translation

Overall performance: Similar by lexical metrics, but green system better in COMET.

Example-based Aggregate Analysis

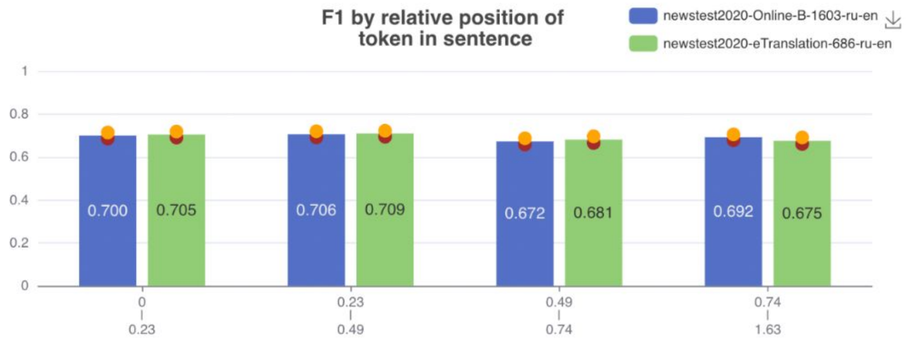




Green system better at short sentences:

-> Green system might be better at resolving cross-sentence ambiguity.

Token-based Aggregate Analysis

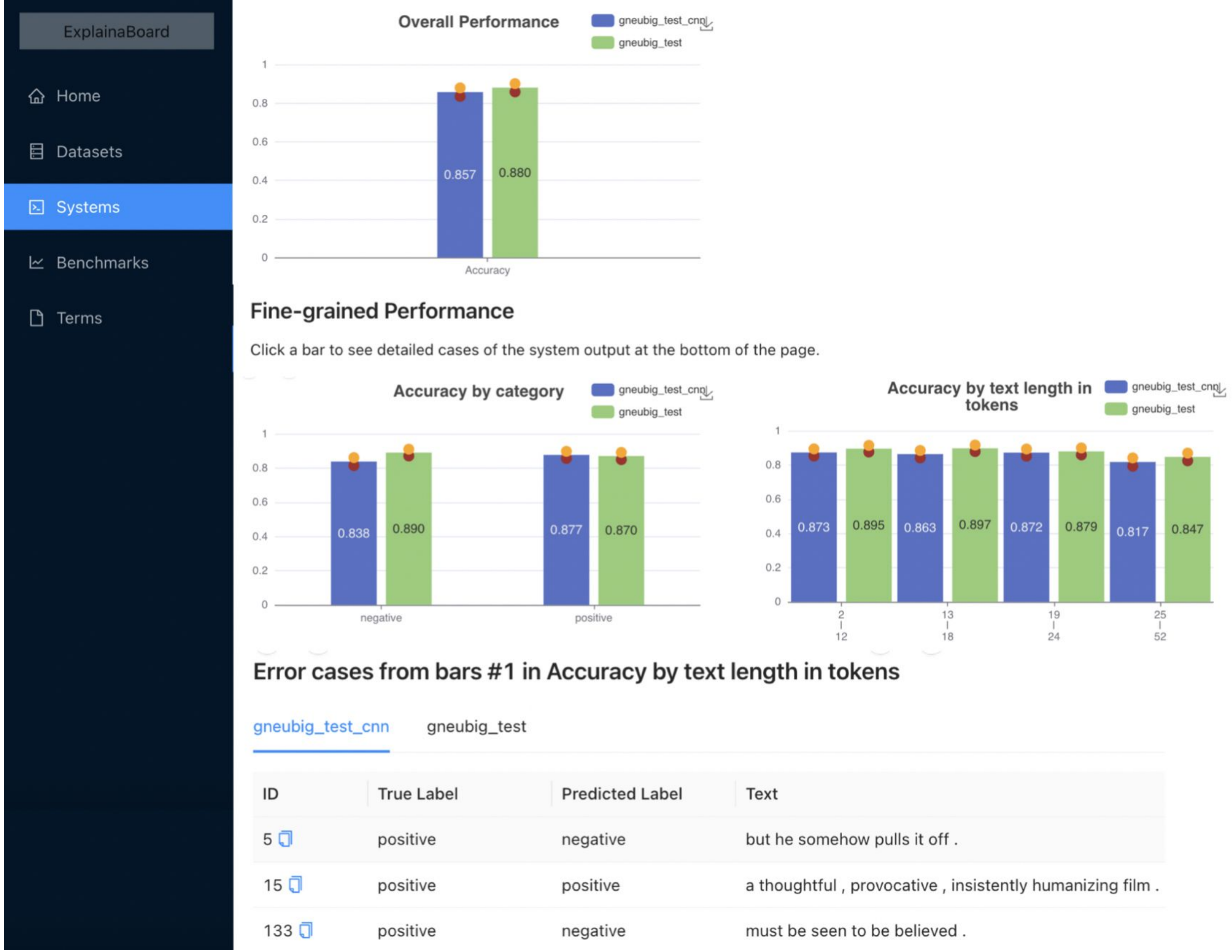


Green system better at short words, blue system better at long words.

-> Green system needs work on technical terms?

Looking Through Examples

#	Source	Reference	Hyp1	Hyp2
335	Также в зоне отчуждения снят запрет на съемку.	The ban on photography in the exclusion zone has also been lifted.	Also in the exclusion zone, a ban on shooting was lifted.	A ban on filming has also been lifted in the exclusion zone.
358	Кто же мог оказаться лучше Гуфа?	Who could be better than Guf?	Who could be better than Guf?	Who could have been better than Goof?
364	У него пушечные удары.	His strikes are like cannon blows.	He has cannon strikes.	He has cannonballs.

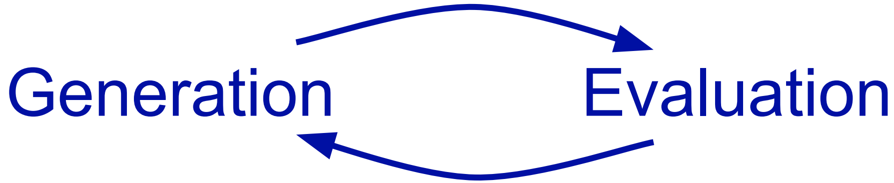


https://explainaboard.inspiredco.ai

What’s next?

Still Challenges!

- **Evaluation:** “arms race” of evaluation, generation, and human standard



- **Automating Fine-grained Analysis:** how to discover interesting behaviors automatically?

“Your model is under-performing on sentences
with numerals greater than 5000.”

- **Few-shot Evaluation/Improvement for New Tasks**