

# Is My NLP Model Working?

## The Answer is Harder Than You Think

Graham Neubig



**Carnegie Mellon University**  
Language Technologies Institute





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](#), Frank Seide, [Xu Tan](#), Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, [Dongdong Zhang](#), Zhirui Zhang, Ming Zhou  
March 2018  
arXiv:1803.05567

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



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.

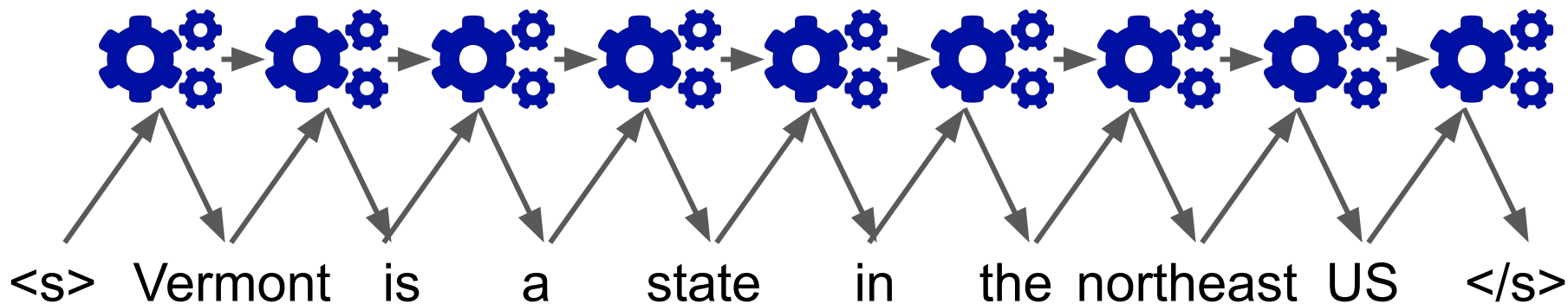
What percent of oil production at the time came from Iran?

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

# Language Models

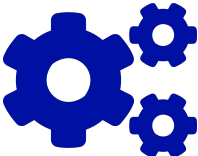


Simple idea!

But train 500 billion parameters on terabytes of data and...

# 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 Generated 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?

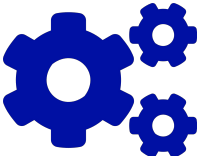
We sleep to give our bodies and brains a chance to rest and repair themselves. Dreams may help us process and store memories, or they may be a way for our brains to work through problems or anxiety.

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

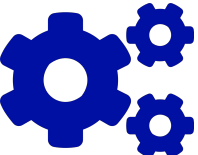
<https://twitter.com/dweekly/status/1580676295444217857>

# 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.

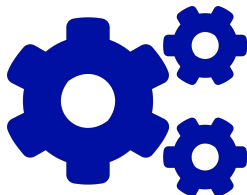


# The NLP Development Pipeline

Training Data



System



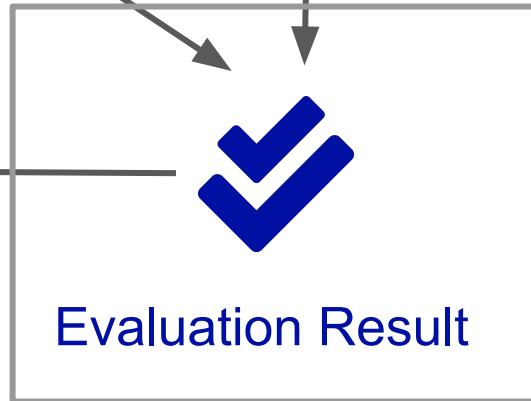
Testing Data



Ideas



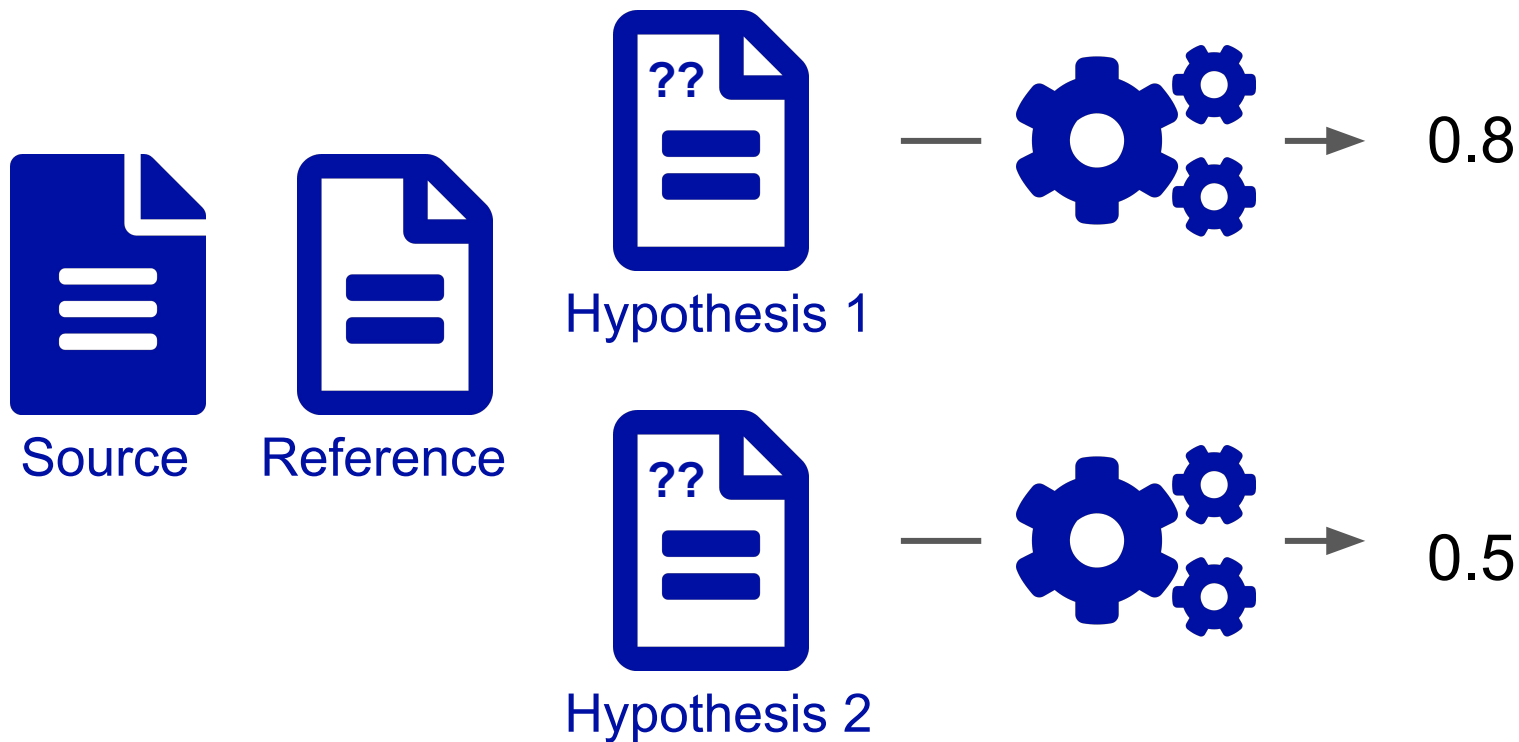
Evaluation Result



# The Gold-standard?: Manual Evaluation



# 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

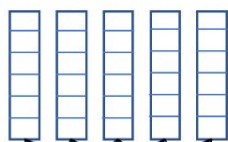
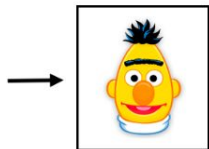
little overlap but good output  
(particularly difficult for open-ended problems)

# Embedding-based Evaluation

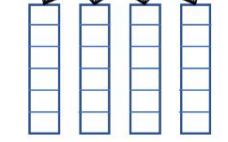
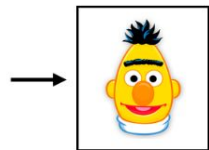


## BERTScore

**Reference  $x$**   
The weather is cold today



**Candidate  $\hat{x}$**   
It is freezing today



**Contextual  
Embedding**

**Pairwise  
Cosine  
Similarity**

the	0.713	0.597	0.428	0.408	1.27
weather	0.462	0.393	0.515	0.326	7.94
is	0.635	0.858	0.441	0.441	1.82
cold	0.479	0.454	0.796	0.343	7.90
today	0.347	0.361	0.307	0.913	8.88
	it	is	freezing	today	idf weights

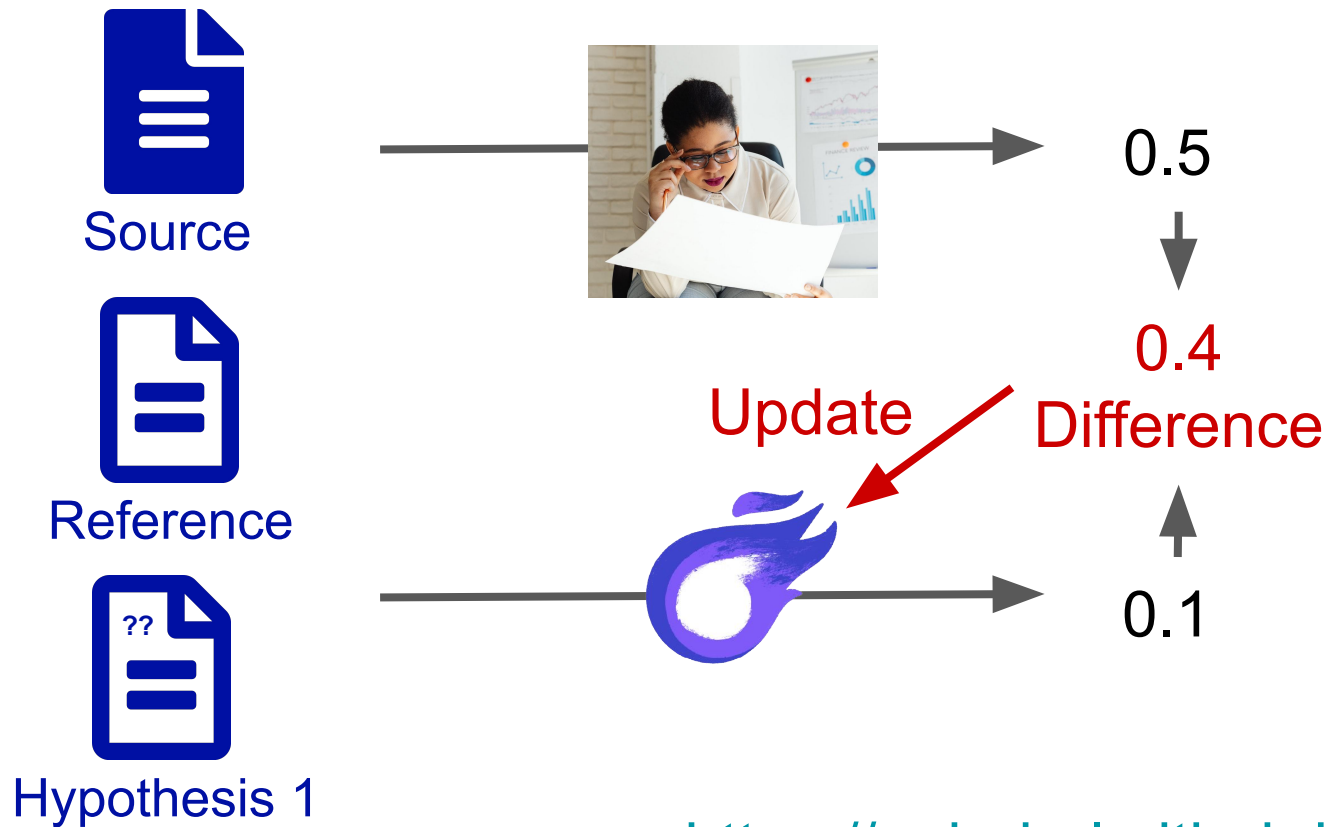
**Maximum  
Similarity**

$$R_{BERT} = \frac{(0.713 \times 1.27) + (0.515 \times 7.94) + \dots}{1.27 + 7.94 + 1.82 + 7.90 + 8.88}$$
$$= 0.753$$

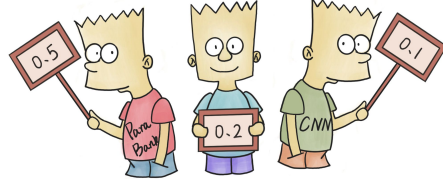
**Importance Weighting**

[https://github.com/Tiiiger/bert\\_score](https://github.com/Tiiiger/bert_score)

# Learning to Evaluate

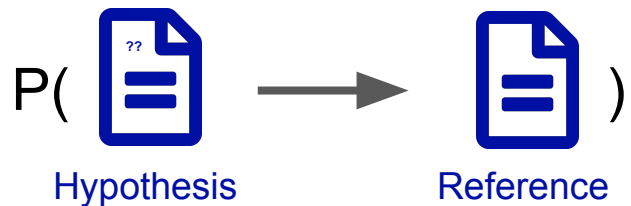
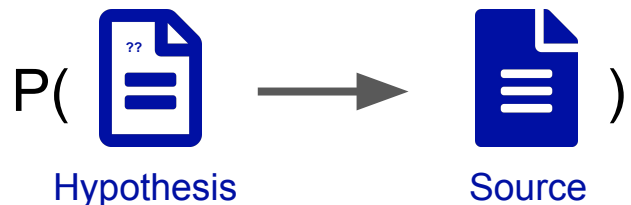
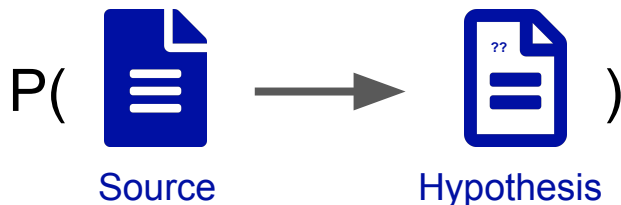


# Generative Text Evaluation



BARTScore





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



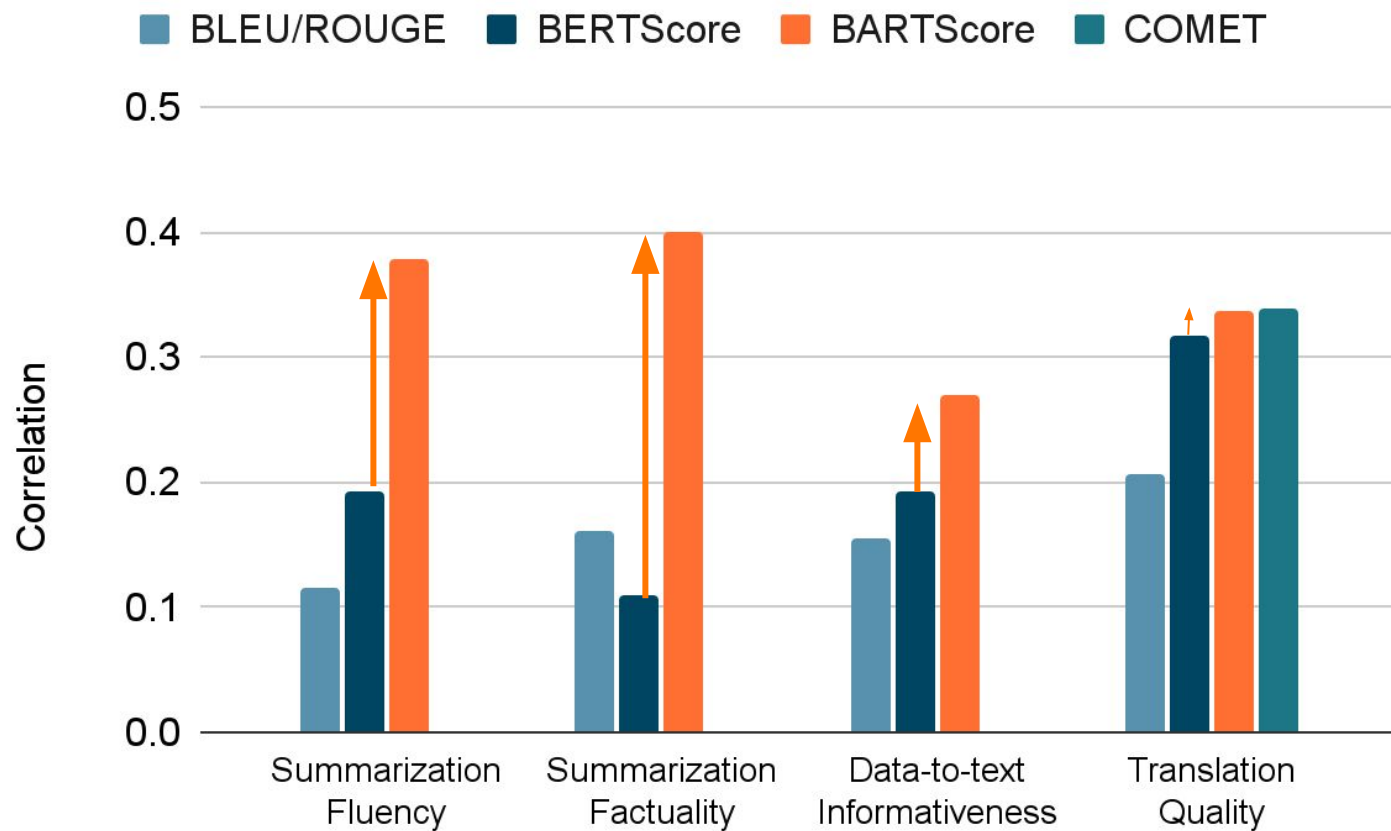
<https://github.com/neulab/BARTScore>



# How Do We Evaluate Evaluation?

	<u>Human</u>	<u>Automatic</u>	
	0.8	0.7	
	0.5	0.1	
	0.1	0.5	
	0.6	0.4	
			<u>Correlation</u>
			Pearson = 0.23
			Kendall = 0.33

# Meta-Evaluation Results

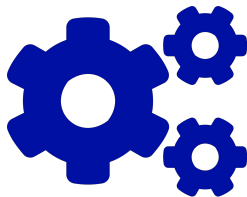


# What's Next?

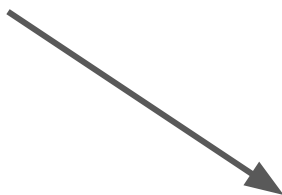
Training Data



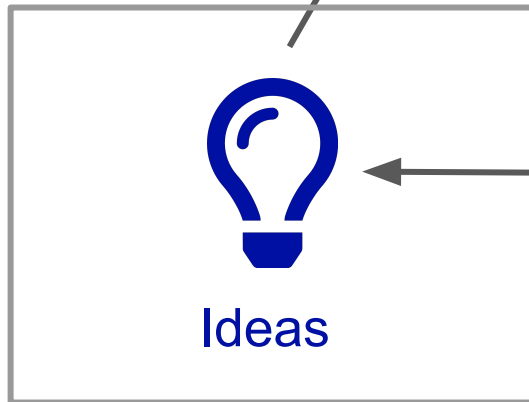
System



Testing Data



Evaluation Result



Ideas



# 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:”

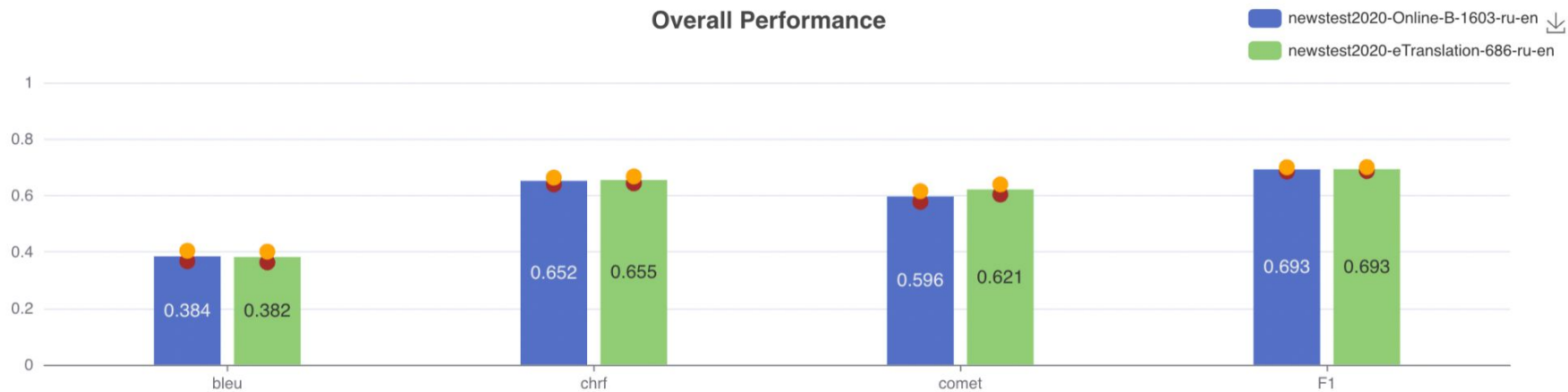
Source: Voda byla skvělá.

Reference: The water was great.

Hypothesis: The water was.

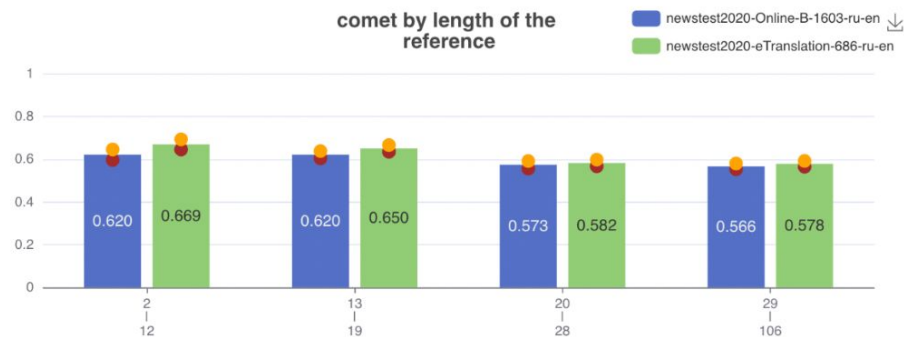
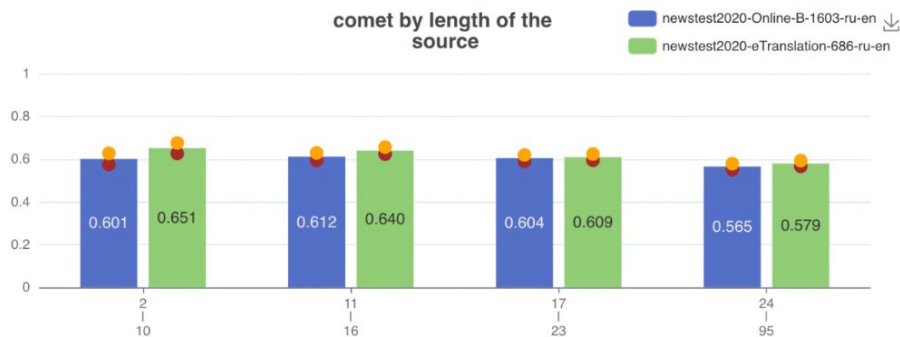
# A Case Study: Russian-English Translation

## Overall Performance



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

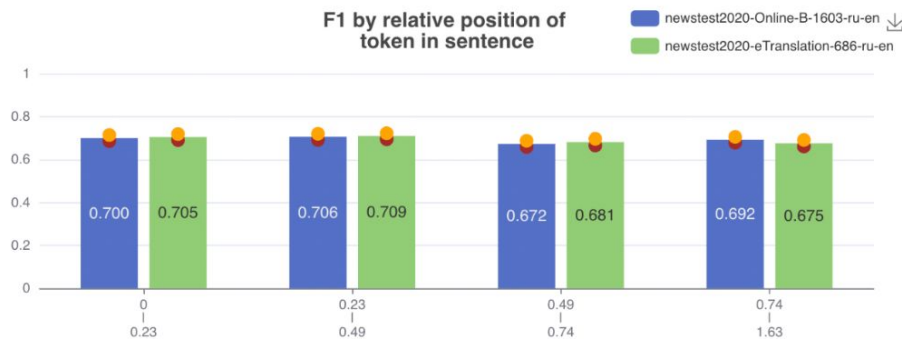
# Example-based Aggregate Analysis



Green system better at short sentences:

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

# Token-based Aggregate Analysis



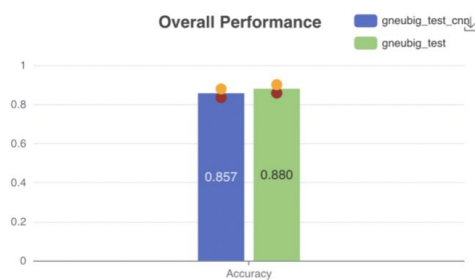
Green system better at short words, blue system better at long words.

-> Green system needs work on technical terms?

# Looking Through Examples

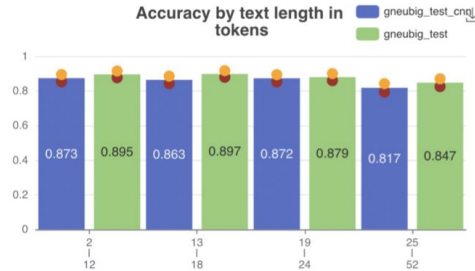
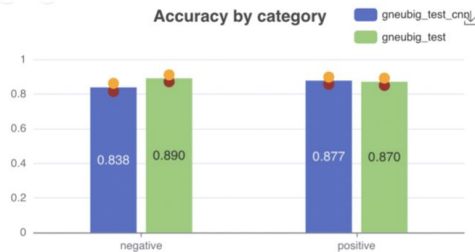
#	Source	Reference	Hyp1	Hyp2
<b>335</b>	Также в зоне отчуждения снят запрет на съемку.	The ban on photography in the exclusion zone has also been lifted.	Also in the exclusion zone, a ban on shooting was lifted.	A ban on filming has also been lifted in the exclusion zone.
<b>358</b>	Кто же мог оказаться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





### Fine-grained Performance

Click a bar to see detailed cases of the system output at the bottom of the page.



### Error cases from bars #1 in Accuracy by text length in tokens

[gneubig\\_test\\_cnn](#)   gneubig\_test

ID	True Label	Predicted Label	Text
5	positive	negative	but he somehow pulls it off .
15	positive	positive	a thoughtful , provocative , insistently humanizing film .
133	positive	negative	must be seen to be believed .

# Using SOTA Metrics and Aggregate Analysis

```
import os
import explainboard_client

# Set up your environment
explainboard_client.username = os.environ[EB_USERNAME']
explainboard_client.api_key = os.environ[EB_API_KEY']
client = explainboard_client.ExplainboardClient()

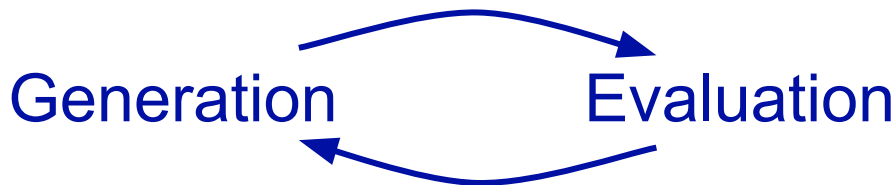
# Do the evaluation
evaluation_result = client.evaluate_system_file(
    task='machine-translation',
    system_name='machine-translation-test',
    system_output_file='my-system.txt',
    system_output_file_type='text',
    dataset='wmt20',
    sub_dataset='ruen',
    split='test',
    source_language='rus',
    target_language='eng',
    metric_names=['bleu', 'chrf', 'comet'],
)
```

Web Client: [https://github.com/neulab/explainboard\\_client](https://github.com/neulab/explainboard_client)

Open-source SDK: <https://github.com/neulab/explainboard>

# 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.”

# A Recipe for Modern NLP Evaluation

## State-of-the-art Metrics

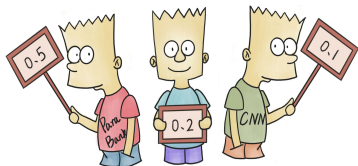


**BERTScore**



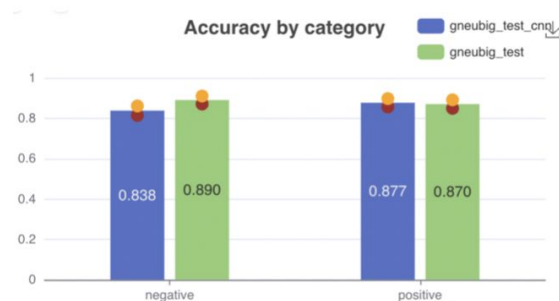
**COMET**

by Unbabel



**BARTScore**

## Fine-grained Analysis



Eager to hear about your NLP problems!

[gneubig@cs.cmu.edu](mailto:gneubig@cs.cmu.edu)

[gneubig@inspiredco.ai](mailto:gneubig@inspiredco.ai)

<https://explainboard.inspiredco.ai>