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

Graham Neubig

Carnegie Mellon University
Language Technologies Institute

INSPIRED COGNITION
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, Renqiu 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

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.

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

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?

In percentage terms?

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

Has the US reduced its dependence on Middle Eastern oil since then?

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

By Leonardo De Cosmo on July 12, 2022
Vermont is a state in the northeast US

Simple idea!
But train 500 billion parameters on terabytes of data and…
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.
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.
Evaluating generated text is now about as hard as generating it.
The NLP Development Pipeline

Training Data → System → Testing Data

Ideas → System → Evaluation Result
The Gold-standard?: Manual Evaluation

Source

Hypothesis 1

Hypothesis 2

→ 0.8

→ 0.5
An Alternative: Automatic Evaluation

Source

Reference

Hypothesis 1

Hypothesis 2

0.8

0.5
The Old Reliabiles: 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

Reference $x$
The weather is cold today

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

Contextual Embedding  Pairwise Cosine Similarity  Maximum Similarity  Importance Weighting

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

https://github.com/Tiiiger/bert_score
Learning to Evaluate

Source

Reference

Hypothesis 1

Update

Difference

0.5

0.4

0.1

https://unbabel.github.io/COMET/
Generative Text Evaluation

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

\[ P(\text{Source}) \rightarrow P(\text{Hypothesis}) \rightarrow P(\text{Reference}) \]

[GitHub link](https://github.com/neulab/BARTScore)
How Do We Evaluate Evaluation?

<table>
<thead>
<tr>
<th>Human</th>
<th>Automatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>0.6</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Correlation:
- Pearson = 0.23
- Kendall = 0.33
Meta-Evaluation Results
What’s Next?

- Training Data
- System
- Testing Data
- Ideas
- Evaluation Result
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

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

<table>
<thead>
<tr>
<th>ID</th>
<th>True Label</th>
<th>Predicted Label</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>positive</td>
<td>negative</td>
<td>but he somehow pulls it off .</td>
</tr>
<tr>
<td>15</td>
<td>positive</td>
<td>positive</td>
<td>a thoughtful , provocative , insistently humanizing film .</td>
</tr>
<tr>
<td>133</td>
<td>positive</td>
<td>negative</td>
<td>must be seen to be believed .</td>
</tr>
</tbody>
</table>

https://explainaboard.inspiredco.ai
Using SOTA Metrics and Aggregate Analysis

import os
import explainaboard_client

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

# 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/explainaboard_client
Open-source SDK: https://github.com/neulab/explainaboard
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  gneubig@inspiredco.ai

https://explainaboard.inspiredco.ai