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

Graham Neubig
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

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 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.
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

System
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

Source  Reference  Hypothesis 1  Hypothesis 2

0.8  0.5
The Old Reliabes: 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 $\rightarrow$ high score

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

little overlap $\rightarrow$ 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


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

Source

Reference

Hypothesis 1

0.5

0.4

Update

Difference

0.1


https://unbabel.github.io/COMET/
Learning to Evaluate w/ Pseudo-data

UniEval


https://github.com/maszhongming/UniEval
Generative Text Evaluation

Use the probability of a generative model to evaluate text

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


https://github.com/neulab/BARTScore
Generative Pre-training, Discriminative Fine-tuning


https://github.com/qinyiwei/T5Score
### 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
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

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

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

Summary Quality  Translation Quality  Fluency  Toxicity  Factual Consistency

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

Training Data → System → Testing Data

Ideas ↔ Evaluation Result
Metric-aware Training

Standard MLE Loss

Ranking Loss Based on Metrics


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

<table>
<thead>
<tr>
<th>Error Matrix (e.g. 1-metric)</th>
<th>Probability</th>
<th>Bayes Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>My name is Bob.</td>
<td>0.0 0.9 0.9 0.9</td>
<td>0.2</td>
</tr>
<tr>
<td>This is true.</td>
<td>0.9 0.0 0.5 0.5</td>
<td>0.19</td>
</tr>
<tr>
<td>This isn’t true.</td>
<td>0.9 0.5 0.0 0.1</td>
<td>0.18</td>
</tr>
<tr>
<td>This is not true.</td>
<td>0.9 0.5 0.1 0.0</td>
<td>0.17</td>
</tr>
</tbody>
</table>


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

<table>
<thead>
<tr>
<th>Model</th>
<th>Prompt</th>
<th>UniEval (Consistency)</th>
<th>UniEval (Coherence)</th>
<th>UniEval (Fluency)</th>
<th>UniEval (Relevance)</th>
<th>BartScore (Coverage)</th>
<th>Length Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>cohere_medium</td>
<td>standard</td>
<td>0.7466</td>
<td>0.4006</td>
<td>0.8698</td>
<td>0.3428</td>
<td>-3.4095</td>
<td>2.5533</td>
</tr>
<tr>
<td>cohere_medium</td>
<td>tldr</td>
<td>0.5006</td>
<td>0.2967</td>
<td>0.8539</td>
<td>0.3312</td>
<td>-3.1348</td>
<td>2.5800</td>
</tr>
<tr>
<td>cohere_medium</td>
<td>concise</td>
<td>0.8542</td>
<td>0.6115</td>
<td>0.9140</td>
<td>0.6167</td>
<td>-3.4220</td>
<td>2.4500</td>
</tr>
<tr>
<td>cohere_medium</td>
<td>complete</td>
<td>0.8331</td>
<td>0.4845</td>
<td>0.8825</td>
<td>0.5214</td>
<td>-3.1689</td>
<td>2.6767</td>
</tr>
<tr>
<td>openai_babbage_001</td>
<td>standard</td>
<td>0.9409</td>
<td>0.9036</td>
<td>0.8782</td>
<td>0.7975</td>
<td>-3.4083</td>
<td>2.0800</td>
</tr>
<tr>
<td>openai_babbage_001</td>
<td>tldr</td>
<td>0.8728</td>
<td>0.9072</td>
<td>0.9593</td>
<td>0.8145</td>
<td>-3.5234</td>
<td>1.0200</td>
</tr>
<tr>
<td>openai_babbage_001</td>
<td>concise</td>
<td>0.9483</td>
<td>0.9365</td>
<td>0.8669</td>
<td>0.8431</td>
<td>-3.2528</td>
<td>2.2800</td>
</tr>
<tr>
<td>openai_babbage_001</td>
<td>complete</td>
<td>0.9306</td>
<td>0.8278</td>
<td>0.8634</td>
<td>0.6951</td>
<td>-3.2720</td>
<td>2.2633</td>
</tr>
<tr>
<td>openai_ada_001</td>
<td>standard</td>
<td>0.6750</td>
<td>0.7270</td>
<td>0.8850</td>
<td>0.8174</td>
<td>-3.6719</td>
<td>2.0067</td>
</tr>
<tr>
<td>openai_ada_001</td>
<td>tldr</td>
<td>0.7999</td>
<td>0.7122</td>
<td>0.7973</td>
<td>0.6728</td>
<td>-3.7436</td>
<td>1.5300</td>
</tr>
<tr>
<td>openai_ada_001</td>
<td>concise</td>
<td>0.7776</td>
<td>0.7439</td>
<td>0.8106</td>
<td>0.5852</td>
<td>-3.6096</td>
<td>2.3600</td>
</tr>
<tr>
<td>openai_ada_001</td>
<td>complete</td>
<td>0.7732</td>
<td>0.5008</td>
<td>0.7332</td>
<td>0.3283</td>
<td>-3.5246</td>
<td>2.4567</td>
</tr>
</tbody>
</table>

https://github.com/inspired-cognition/prompt-gym/
Fine-grained Analysis and Understanding of NLP Models
To the next step!

Training Data → System → Testing Data

Ideas → System → 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.
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?
<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**

<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](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**