Evaluating text generation

CS685 Spring 2023
Advanced Natural Language Processing

Mohit Iyyer
College of Information and Computer Sciences
University of Massachusetts Amherst

some slides from Marine Carpuat & Marzena Karpinska
So far...

- We’ve seen *perplexity* as an automatic measure to evaluate language models.

- However, perplexity alone is insufficient to tell us about how well a model is solving some downstream task (e.g., translation or summarization).

- Today: BLEU score for MT, ROUGE for summarization, BERT-based improvements, and human evaluation.
How Good is Machine Translation?
Chinese > English

Reporters learned from the Ministry of Environmental Protection, "Water 10" requirements before the end of this year before the municipality, the provincial capital city, plans to build a separate city to solve the basic black and black water. Up to now, the country's 224 prefecture-level and above cities were identified to confirm the black and white water 2082, of which 34.9% to complete the renovation, 28.4% is remediation, 22.8% is carrying out the project early.
How Good is Machine Translation?
French > English

A l'orée de ce débat télévisé inédit dans l'histoire de la Ve République, on attendait une forme de «Tous sur Macron» mais c'est la candidate du Front national qui s'est retrouvée au cœur des premières attaques de ses quatre adversaires d'un soir, favorisées par le premier thème abordé, les questions de société et donc de sécurité, d'immigration et de laïcité.

At the beginning of this televised debate, which was unheard of in the history of the Fifth Republic, a "Tous sur Macron" was expected, but it was the candidate of the National Front who found itself at the heart of the first attacks of its four Opponents of one evening, favored by the first theme tackled, the issues of society and thus security, immigration and secularism.
What is MT good (enough) for?

• **Assimilation:** reader initiates translation, wants to know content
  • User is tolerant of inferior quality
  • Focus of majority of research

• **Communication:** participants in conversation don’t speak same language
  • Users can ask questions when something is unclear
  • Chat room translations, hand-held devices
  • Often combined with speech recognition

• **Dissemination:** publisher wants to make content available in other languages
  • High quality required
  • Almost exclusively done by human translators
How good is a translation? Problem: no single right answer

这个 机场 的 安全 工作 由 以色列 方面 负责．

Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport’s security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport’s security is the responsibility of the Israeli security officials.
Evaluation

• How good is a given machine translation system?

• Many different translations acceptable

• Evaluation metrics
  • Subjective judgments by human evaluators
  • Automatic evaluation metrics
  • Task-based evaluation
Adequacy and Fluency

• Human judgment
  • Given: machine translation output
  • Given: input and/or reference translation
  • Task: assess quality of MT output

• Metrics
  • **Adequacy**: does the output convey the meaning of the input sentence? Is part of the message lost, added, or distorted?
  • **Fluency**: is the output fluent? Involves both grammatical correctness and idiomatic word choices.
### Fluency and Adequacy: Scales

<table>
<thead>
<tr>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>flawless English</td>
</tr>
<tr>
<td>4</td>
<td>good English</td>
</tr>
<tr>
<td>3</td>
<td>non-native English</td>
</tr>
<tr>
<td>2</td>
<td>disfluent English</td>
</tr>
<tr>
<td>1</td>
<td>incomprehensible</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Adequacy</th>
<th>Adequacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>all meaning</td>
</tr>
<tr>
<td>4</td>
<td>most meaning</td>
</tr>
<tr>
<td>3</td>
<td>much meaning</td>
</tr>
<tr>
<td>2</td>
<td>little meaning</td>
</tr>
<tr>
<td>1</td>
<td>none</td>
</tr>
</tbody>
</table>
Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l'ue.

Reference: rather, the two countries form a laboratory needed for the internal working of the eu.

<table>
<thead>
<tr>
<th>Translation</th>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>both countries are rather a necessary laboratory the internal operation of the eu.</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>both countries are a necessary laboratory at internal functioning of the eu.</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>the two countries are rather a laboratory necessary for the internal workings of the eu.</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>the two countries are rather a laboratory for the internal workings of the eu.</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>the two countries are rather a necessary laboratory internal workings of the eu.</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
</tbody>
</table>

Annotator: Philipp Koehn Task: WMT06 French-English

Instructions

5= All Meaning
4= Most Meaning
3= Much Meaning
2= Little Meaning
1= None

5= Flawless English
4= Good English
3= Non-native English
2= Disfluent English
1= Incomprehensible
Let’s try:
rate fluency & adequacy on 1-5 scale

– Source:
  N’y aurait-il pas comme une vague hypocrisie de votre part ?

– Reference:
  Is there not an element of hypocrisy on your part?

– System1:
  Would it not as a wave of hypocrisy on your part?

– System2:
  Is there would be no hypocrisy like a wave of your hand?

– System3:
  Is there not as a wave of hypocrisy from you?
what are some issues with human evaluation?
Automatic Evaluation Metrics

• Goal: computer program that computes quality of translations

• Advantages: low cost, optimizable, consistent

• Basic strategy
  • Given: MT output
  • Given: human reference translation
  • Task: compute similarity between them
Precision and Recall of Words

SYSTEM A:  
[Highlighted text: Israeli officials responsibility of airport safety]

REFERENCE:  
[Highlighted text: Israeli officials are responsible for airport security]

Precision  
\[ \frac{\text{correct}}{\text{output-length}} = \frac{3}{6} = 50\% \]

Recall  
\[ \frac{\text{correct}}{\text{reference-length}} = \frac{3}{7} = 43\% \]

F-measure  
\[ \frac{\text{precision} \times \text{recall}}{(\text{precision} + \text{recall})/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\% \]
Precision and Recall of Words

**SYSTEM A:** Israeli officials responsibility of airport safety

**REFERENCE:** Israeli officials are responsible for airport security

**SYSTEM B:** airport security Israeli officials are responsible

<table>
<thead>
<tr>
<th>Metric</th>
<th>System A</th>
<th>System B</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>50%</td>
<td>100%</td>
</tr>
<tr>
<td>recall</td>
<td>43%</td>
<td>100%</td>
</tr>
<tr>
<td>f-measure</td>
<td>46%</td>
<td>100%</td>
</tr>
</tbody>
</table>

flaw: no penalty for reordering
BLEU
Bilingual Evaluation Understudy

N-gram overlap between machine translation output and reference translation

Compute precision for n-grams of size 1 to 4

Add brevity penalty (for too short translations)

\[
\text{BLEU} = \min \left( 1, \frac{\text{output-length}}{\text{reference-length}} \right) \left( \prod_{i=1}^{4} \text{precision}_i \right)^{\frac{1}{4}}
\]

Typically computed over the entire corpus, not single sentences
Multiple Reference Translations

To account for variability, use multiple reference translations

– n-grams may match in any of the references
– closest reference length used

Example

SYSTEM:

 Israeli officials | responsibility of | airport | safety
2-GRAM MATCH     2-GRAM MATCH  1-GRAM

 Israeli officials are responsible for airport security
Israel is in charge of the security at this airport

REFERENCES:
The security work for this airport is the responsibility of the Israel government
Israeli side was in charge of the security of this airport
BLEU examples

**SYSTEM A:**
- Israeli officials  
- responsibility of airport safety

**REFERENCE:**
Israeli officials are responsible for airport security

**SYSTEM B:**
- airport security  
- Israeli officials are responsible

<table>
<thead>
<tr>
<th>Metric</th>
<th>System A</th>
<th>System B</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision (1gram)</td>
<td>3/6</td>
<td>6/6</td>
</tr>
<tr>
<td>precision (2gram)</td>
<td>1/5</td>
<td>4/5</td>
</tr>
<tr>
<td>precision (3gram)</td>
<td>0/4</td>
<td>2/4</td>
</tr>
<tr>
<td>precision (4gram)</td>
<td>0/3</td>
<td>1/3</td>
</tr>
<tr>
<td>brevity penalty</td>
<td>6/7</td>
<td>6/7</td>
</tr>
<tr>
<td>BLEU</td>
<td>0%</td>
<td>52%</td>
</tr>
</tbody>
</table>
BLEU examples

why does BLEU not account for recall?

<table>
<thead>
<tr>
<th>Metric</th>
<th>System A</th>
<th>System B</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision (1gram)</td>
<td>3/6</td>
<td>6/6</td>
</tr>
<tr>
<td>precision (2gram)</td>
<td>1/5</td>
<td>4/5</td>
</tr>
<tr>
<td>precision (3gram)</td>
<td>0/4</td>
<td>2/4</td>
</tr>
<tr>
<td>precision (4gram)</td>
<td>0/3</td>
<td>1/3</td>
</tr>
<tr>
<td>brevity penalty</td>
<td>6/7</td>
<td>6/7</td>
</tr>
<tr>
<td>BLEU</td>
<td>0%</td>
<td>52%</td>
</tr>
</tbody>
</table>
what are some drawbacks of BLEU?

- all words/n-grams treated as equally relevant
- operates on local level
- scores are meaningless (absolute value not informative)
- human translators also score low on BLEU
Yet automatic metrics such as BLEU correlate with human judgement.
ROUGE - a recall-based counterpart to BLEU

• Idea: what % of the words or n-grams in the reference occur in the generated output?

• ROUGE and its variants are often used to evaluate text summarization systems
Can we include learned components in our evaluation metrics?
BLEURT (BLEU + BERT)

• Take a pretrained BERT, and fine-tune it on a variety of synthetic tasks with perturbed data
  • Synthetic data involves a sentence $z$ and “perturbed” version $z'$
  • Objectives include many regression tasks (e.g., predict BLEU, ROUGE, backtranslation likelihood)

• Then, fine-tune the resulting model on small supervised datasets of human quality judgments
BLEURT (BLEU + BERT)

• Take a pretrained BERT, and fine-tune it on a variety of synthetic tasks with perturbed data
  • Synthetic data involves a sentence $z$ and “perturbed” version $z'$
  • Objectives include many regression tasks (e.g., predict BLEU, ROUGE, backtranslation likelihood)
• Then, fine-tune the resulting model on small supervised datasets of human quality judgments

Higher correlation with human judgments than just BLEU, but has limitations…
COMET (now recommended for MT evaluation over BLEU)
How do we evaluate open-ended text generation?
**PROMPT:** You, a video game character, just died. You had no spare lives left.

**STORY:** Link woke in a panic, gasping and flailing wildly at the air above him. The panic quickly turned to confusion, wasn't he just...? He heard footsteps approaching from behind. Link leapt up from the floor and drew his sword, ready to strike. The sight of a slight man in a suit added to Link's already mounting confusion. Is this another trap set by Ganondorf?
**200 Human-written Stories**

**PROMPT:** You, a video game character, just died. You had no spare lives left.

**STORY:** Link woke in a panic, gasping and flailing wildly at the air above him. The panic quickly turned to confusion, wasn't he just...? He heard footsteps approaching from behind. Link leapt up from the floor and drew his sword, ready to strike. The sight of a slight man in a suit added to Link's already mounting confusion. Is this another trap set by Ganondorf? (…)

**~150 words**

**200 GPT-2 Generated Stories**

**PROMPT:** You, a video game character, just died. You had no spare lives left.

**STORY:** *The US Space Station exploded with automatic fire in preparation for tonight's nuclear strike on Earth. This is video from the control center inside the Space Station*

*A figure wearing a manilla envelope falls off of the back of the space station, detonating a nuclear device.

**An audio file is made of the sirens, and some screams (…)*
AMAZON MECHANICAL TURK

GRAMMAR
How grammatically correct is the text of the story fragment?

COHERENCE
How well do the sentences in the story fragment fit together?

LIKABILITY
How enjoyable do you find the story fragment?

RELEVANCE
How relevant is the story fragment to the prompt?
## Evaluating Machine-Generated Text

### 1. Rating Only GPT-2 Generated Stories

<table>
<thead>
<tr>
<th>Type of text</th>
<th>Grammar</th>
<th>Coherence</th>
<th>Relevance</th>
<th>Likability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean&lt;sub&gt;STD&lt;/sub&gt;</td>
<td>IAA%</td>
<td>Mean&lt;sub&gt;STD&lt;/sub&gt;</td>
<td>IAA%</td>
</tr>
<tr>
<td>AMT workers fail to effectively distinguish between human written and GPT-2 generated stories</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ref. (Day 1)</td>
<td>4.00&lt;sub&gt;0.92&lt;/sub&gt;</td>
<td>0.21&lt;sub&gt;15.5&lt;/sub&gt;</td>
<td>4.11&lt;sub&gt;0.96&lt;/sub&gt;</td>
<td>0.14&lt;sub&gt;16.5&lt;/sub&gt;</td>
</tr>
<tr>
<td>Ref. (Day 2)</td>
<td>3.86&lt;sub&gt;0.92&lt;/sub&gt;</td>
<td>-0.03&lt;sub&gt;10.5&lt;/sub&gt;</td>
<td>3.92&lt;sub&gt;0.98&lt;/sub&gt;</td>
<td>-0.03&lt;sub&gt;6.5&lt;/sub&gt;</td>
</tr>
<tr>
<td>Ref. (Day 3)</td>
<td>3.98&lt;sub&gt;0.96&lt;/sub&gt;</td>
<td>0.18&lt;sub&gt;11&lt;/sub&gt;</td>
<td>4.05&lt;sub&gt;0.94&lt;/sub&gt;</td>
<td>0.13&lt;sub&gt;10.5&lt;/sub&gt;</td>
</tr>
<tr>
<td>GPT-2</td>
<td>3.94&lt;sub&gt;0.93&lt;/sub&gt;</td>
<td>0.11&lt;sub&gt;17.5&lt;/sub&gt;</td>
<td>3.82&lt;sub&gt;1.12&lt;/sub&gt;</td>
<td>0.05&lt;sub&gt;7.5&lt;/sub&gt;</td>
</tr>
</tbody>
</table>
Time Spent on the Task

360 sec
WorkTimeInSeconds

22 sec
Mean

13 sec
Median
ENGLISH TEACHERS

3 Certified English Teachers
ENGLISH TEACHERS: RESULTS

English Teachers Rated Human-written Stories significantly higher than GPT-2 Generated Stories (unlike Turkers)

- Grammar: Equal
- Relevance: Human > GPT-2
- Coherence: Human > GPT-2
- Likability: Human > GPT-2
Post-Task Interviews

- Need **10–20 examples** to calibrate ratings
- *Coherence* was the easiest to rate for human-written stories
- *Coherence* was also the most challenging to rate for GPT-2 stories
- *Relevance* was the easiest to rate for GPT-2 stories (clearly not following the prompt)
- Overall GPT-2 generated stories were **difficult to rate** (average time per story raised from 69.8s → 87.3s)
- Preferred to rate GPT-2 and human-written stories **together** (better calibration)
- Suggested to employ a **rubric**
TAKEAWAYS

- Evaluation of open-ended generated text is... **DIFFICULT!** (even for expert raters)

- High variance between workers, poor calibration, and cognitively-demanding tasks can lead researchers to draw misleading scientific conclusions.

- Possible solutions include:

  1. time-filtering,
  2. specifying min/max number of items per worker,
  3. employing a pre-task language proficiency test,
  4. providing training HITs to allow workers to calibrate their ratings,
  5. showing model-generated text along with human-written text,
  6. if possible, employing raters who were already trained to evaluate written text.
MQM: a typology of fine-grained errors for MT
Can we use LLMs to evaluate generated text?

**Task Introduction**
You will be given one summary written for a news article. Your task is to rate the summary on one metric ……

**Evaluation Criteria**
Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence ……

**Evaluation Steps**
1. Read the news article carefully and identify the main topic and key points.
2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.
3. Assign a score for coherence on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

**Input Context**
Article: Paul Merson has restarted his row with Andros Townsend after the Tottenham midfielder was brought on with only seven minutes remaining in his team’s 0-0 draw with Burnley ……

**Input Target**
Summary: Paul Merson was brought on with only seven minutes remaining in his team’s 0-0 draw with Burnley ……

**Evaluation Form (scores ONLY):**
- Coherence:

**GPTEval**

![GPTEval Diagram]

Weighted Summed Score: 2.59

*GPTEval, Liu et al., 2023*