Evaluating text generation

CS685 Spring 2022

Advanced Natural Language Processing

Mohit Iyyer

College of Information and Computer Sciences

University of Massachusetts Amherst

some slides from Marine Carpuat & Marzena Karpinska

Stuff from last time

- Quiz 3 due this Friday (3/11)!
- Start your dataset collection soon for HW1
- Project proposal feedback by end of the week

So far…

a low perplexity, a model is doing a good job at modeling the data you put it. higher perplexity is an inverse of the probability, model is not going a good job.

- 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

rouge => variance of bleu score, use for summarization.
human evaluation=> human rate the output

How Good is Machine Translation? Chinese > English

记者从环保部了解到，《水十条》要求今年年底前直辖市、省会城市、计划单列市建成区基本解决黑臭水体。截至目前，全国224个地级及以上城市共排查确认黑臭水体2082个，其中34.9%完成整治，28.4%正在整治，22.8%正在开展项目前期。

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.

better because have more parallel data to train. Perform better than Chinese - English translation.

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

simultaneous translation is much more difficult => enable the real time communications

convey a roughly the same meaning.

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 score goes down, if semantics is loss
  - **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.

MT interested in both of these aspect

Fluency and Adequacy: Scales

Adequacy	
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

Fluency	
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible

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 .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English	Annotate	
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?

metrics they don't consider the whole document - evaluate on sentence level.
Still figure out what is the best way to train this system with more context

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

BLEU is the function of the precision of words

Precision and Recall of Words

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

REFERENCE: Israeli officials are responsible for airport security

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

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

any kind of information is missed

F-measure $$\frac{precision \times recall}{(precision + recall)/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$$

major issue. 1) one reference could not be a reference
2). what if the system produce the negate of the reference (get high score) recall = 100%
compute at word level get these issue.

no penalty of the reordering, the sentences is not grammatical

Precision and Recall of Words

Metric	System A	System B
precision	50%	100%
recall	43%	100%
f-measure	46%	100%

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)

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

BLEU score extends the precision of words to the phrase level. compute the overlap of bigram, trigram and take the geometric mean ;)

Typically computed over the entire corpus, not single sentences

much more useful when average the entire corpus - test set

why BLEU score not use recall?? no single translation will include all the variance (Israeli, Israel/ responsibility, responsible)
if you have multiple translations, we would like our metrics to be able to use all the references.

precision is useful if you have multiple references. any bigram that matches, will get score

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 (2-GRAM MATCH) responsibility of (2-GRAM MATCH) airport (1-GRAM) safety

REFERENCES:
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

BLEU can measure the different in quality.

BLEU examples

SYSTEM A: Israeli officials responsibility of airport safety
(Israeli officials: 2-GRAM MATCH; airport: 1-GRAM MATCH)

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible
(airport security: 2-GRAM MATCH; Israeli officials are responsible: 4-GRAM MATCH)

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

BLEU examples

SYSTEM A: Israeli officials responsibility of airport safety
2-GRAM MATCH 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

can get some grammatically correct with 4 grams

SYSTEM B: airport security Israeli officials are responsible
2-GRAM MATCH 4-GRAM MATCH

why does BLEU not account for recall?

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

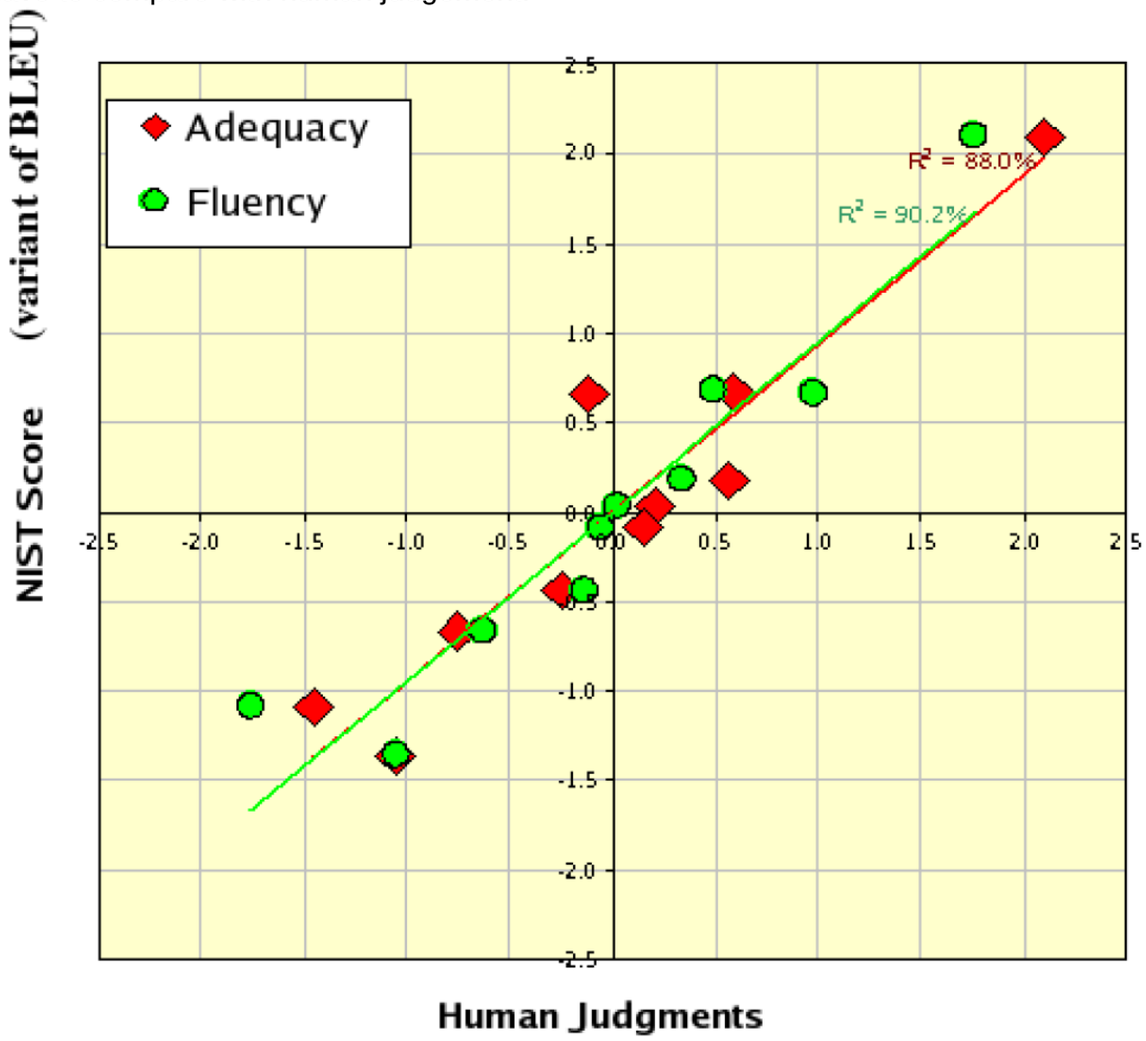
system B is worst than system A. This is an example why you don't want to interpret BLEU score on a sentence level

what are some drawbacks of BLEU?

- all words/n-grams treated as equally relevant
- operates on local level we don't consider any other context in the sentence
- scores are meaningless (absolute value not informative)
- human translators also score low on BLEU

Yet automatic metrics such as BLEU correlate with human judgement

if we want to make a better metric than BLEU, we have to compare with human judgement.



use in summarization

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

primary using the recall.
why is this useful for text summarization?? summary should be as short as possible

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 all the human judgement pg)35. sentence pairs. we train a model to predict a human score given a reference. want a model to predict a human judgement. and test time, pass it to the model and use that for the evaluation score. Not limited to the local matching. given a higher weights to an important word than BLEU

- 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

most of the human judgement are not collected from experts. evaluation is done by human so they end up asking if you submit a system, you must also rate other system

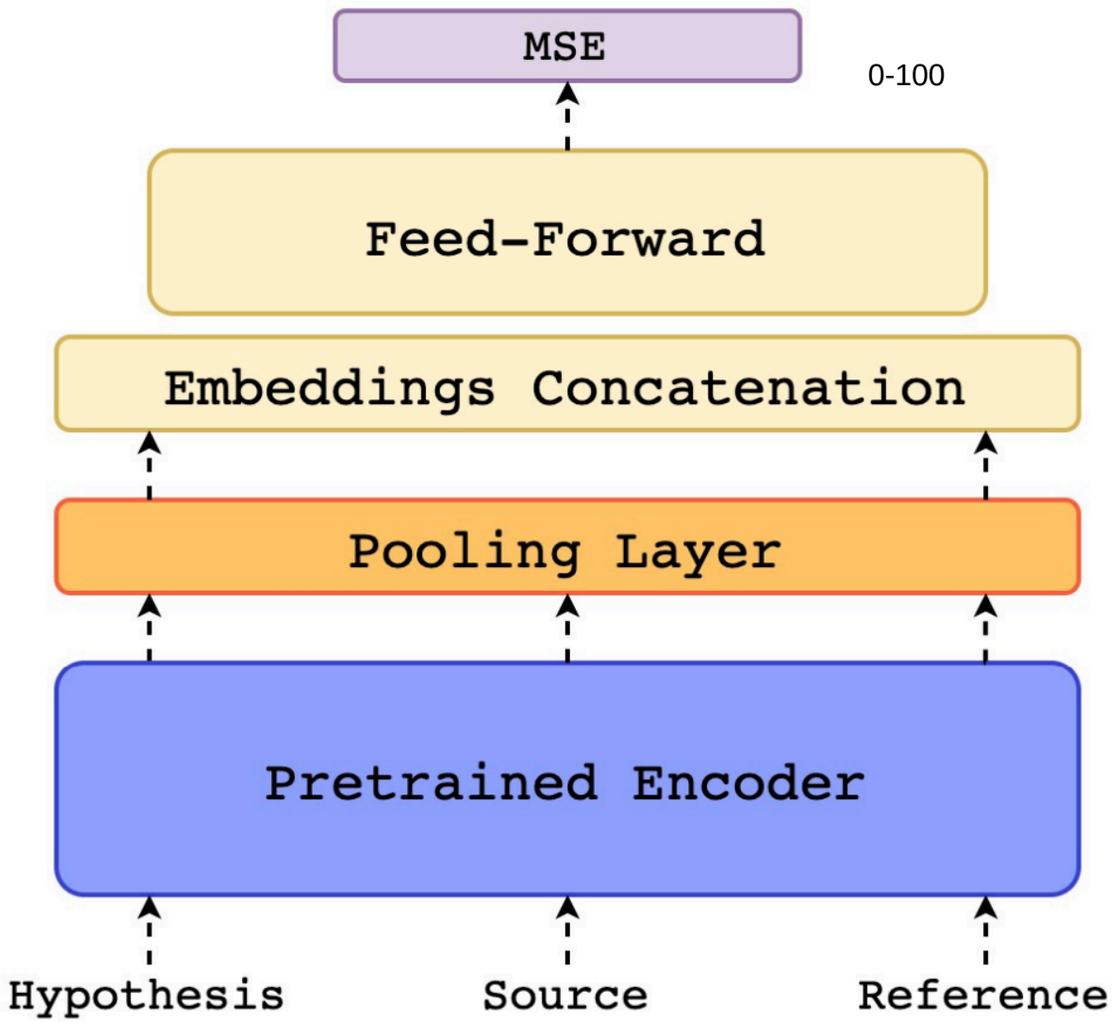
LIMITATION: taking a pretrained and fine-tune it to human judgement.
1. you need a large dataset for human judgement (good for english)
2. harder to get a pre-trained language model, take compute resources, time, not available for every languages

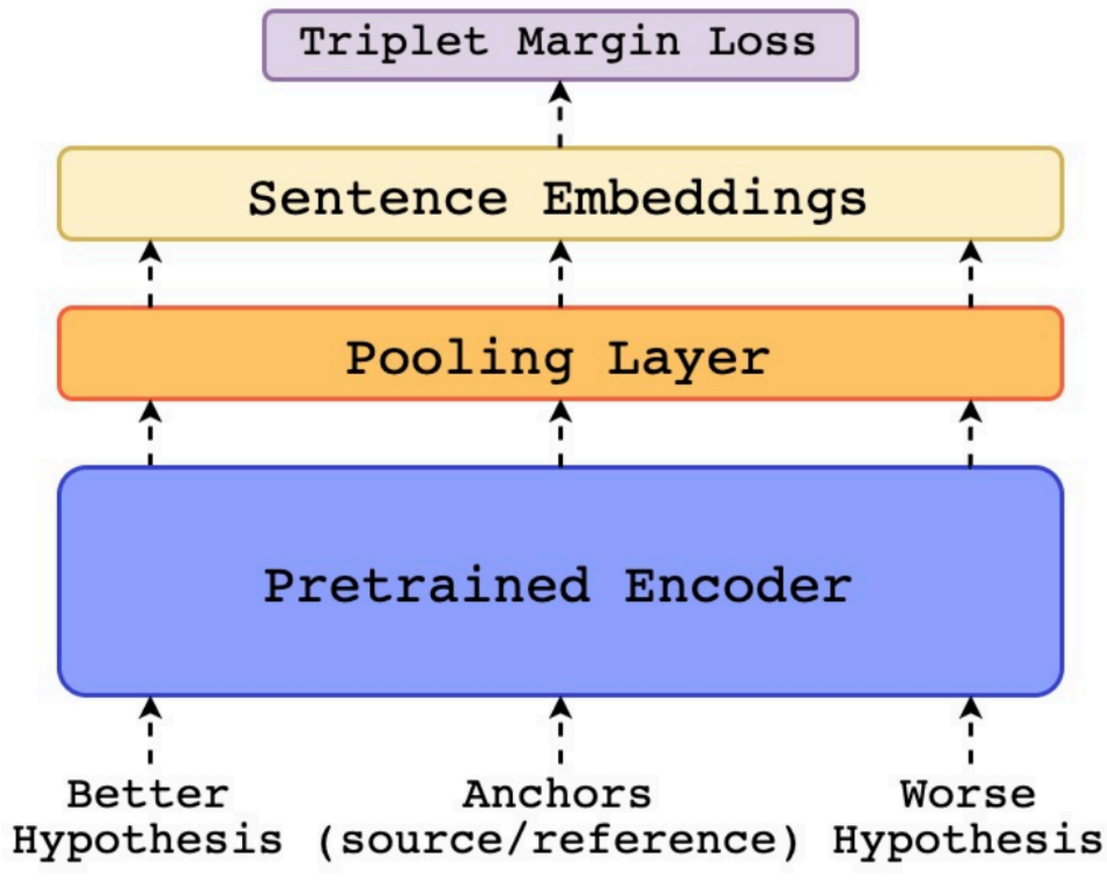
Higher correlation with human judgments than just BLEU, but has limitations…

XLM (multilingual)

COMET (now recommended for MT evaluation over BLEU)

always superior than BLEU score in term of human evaluation





need the encoder to be multilingual

Rei et al., EMNLP 2020

How do we evaluate **open-ended** text generation?

WritingPrompts (Fan et al., 2019)

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?

any language model would work

how reliable of the crowd source of task?

train a simple language model to generate a prompt

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? (...)

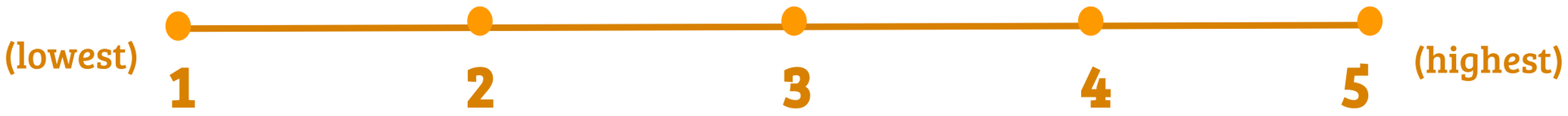
200 GPT-2 Generated Stories

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 (...)*

~150 words

AMAZON MECHANICAL TURK



GRAMMAR

How grammatically correct is the text of the story fragment?

COHERENCE

multiple sentences fit together

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

LIKABILITY

ver subjective, extremely high variance.

How enjoyable do you find the story fragment?

RELEVANCE

How relevant is the story fragment to the prompt?

were not able to detect human vs GPT

AMAZON MECHANICAL TURK

Evaluating Machine-Generated Text

GPT-2

1. Rating Only GPT-2 Generated Stories

Type of text	Grammar		Coherence		Relevance		Likability	
	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$
AMT workers fail to effectively distinguish between human written and GPT-2 generated stories								
Ref. (Day 1)	$4.00_{0.92}$	$0.21_{15.5}$	$4.11_{0.96}$	$0.14_{16.5}$	$3.71_{1.26}$	0.27_{10}	$3.37_{1.18}$	$0.11_{7.5}$
Ref. (Day 2)	$3.86_{0.92}$	$-0.03_{10.5}$	$3.92_{0.98}$	$-0.03_{6.5}$	$3.71_{1.08}$	0.02_{11}	$3.73_{0.97}$	$-0.04_{8.5}$
Ref. (Day 3)	$3.98_{0.96}$	0.18_{11}	$4.05_{0.94}$	$0.13_{10.5}$	$3.46_{1.29}$	0.26_{8}	$3.42_{1.16}$	$0.07_{4.5}$
GPT-2	$3.94_{0.93}$	$0.11_{17.5}$	$3.82_{1.12}$	$0.05_{7.5}$	$3.44_{1.41}$	0.10_{7}	$3.42_{1.25}$	$0.02_{4.5}$

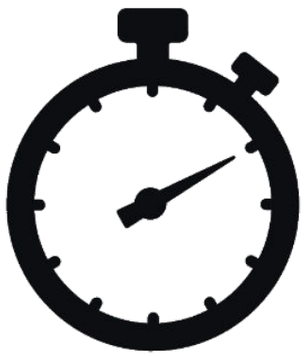
roughly the same.

the workers don't care about your task. :(

AMAZON MECHANICAL TURK

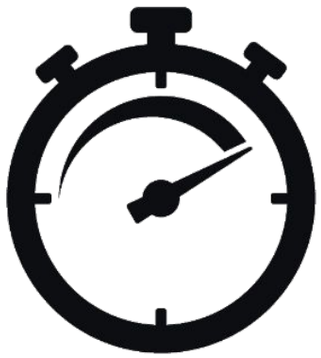
HUMAN

Time Spent on the Task



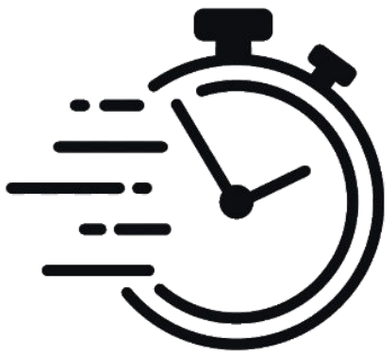
360 sec

WorkTimeInSeconds



22 sec

Mean



13 sec

Median

hire experts :)

ENGLISH TEACHERS

GPT-2+HUM

3 Certified English Teachers

ENGLISH TEACHERS: RESULTS

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

ENGLISH TEACHERS

GPT-2+HUM

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.