Announcements

- If you are using late days, please fill out the Google Form linked on the website (link will be updated shortly).
- HW2 will come out by Friday, but you might not have AWS credit until late next week.
- Chenyan’s office hours for Thursday are canceled.

QUIZ

Carnegie Mellon University

Follow-up on the decline of internet-based training data

Large Language Models: Methods and Applications

Daphne Ippolito and Chenyan Xiong

Pre-training Data Reality

High quality data eventually runs out.

In practice, the web is the most viable option for data collection.

- In the digital era, this is the go-to place for general domain human knowledge.
- It is massive and unlikely to grow slower than computing resources*
- Publicly available*

* More on how true these points are later in the class.

The availability of internet-sourced training data is in decline.

How does web data become **unavailable**?

- Introduction of paywalls

How does web data become **unavailable**?

- Introduction of paywalls
- Restrictive terms of service

How does web data become **unavailable**?

- Introduction of paywalls
- Restrictive terms of service
- Implementation of Robots Exclusion Protocol

Reddit to update web standard to block automated website scraping

By **Reuters**

June 25, 2024 4:45 PM EDT · Updated 2 months ago

Aa

Reddit's logo is displayed, at the New York Stock Exchange (NYSE) in New York City, U.S., March 21, 2024. REUTERS/Brendan McDermid/File Photo *Purchase Licensing Rights*

Companies

Reddit Inc — Follow

June 25 (Reuters) - Social media platform Reddit (RDDT.N) said on Tuesday it will update a web standard used by the platform to block automated data scraping from its website, following reports that AI startups were bypassing the rule to gather content for their systems.

The move comes at a time when artificial intelligence firms have been accused of plagiarizing content from publishers to create AI-generated summaries without giving credit or asking for permission.

Reddit’s robot.txt file in 2019

```
# 80legs
User-agent: 008
Disallow: /

# 80legs' new crawler
User-agent: voltron
Disallow: /

User-Agent: bender
Disallow: /my_shiny_metal_ass

User-Agent: Gort
Disallow: /earth

User-agent: MJ12bot
Disallow: /

User-agent: PiplBot
Disallow: /

User-Agent: *
Disallow: /*.json
Disallow: /*.json-compact
Disallow: /*.json-html
Disallow: /*.xml
Disallow: /*.rss
Disallow: /*.i
Disallow: /*.embed
Disallow: /*/comments/*?*sort=
Disallow: /r/*/comments/*/*/c*
```

Disallow spam bots

Disallow parts of the site that aren’t interesting or will break webcrawlers.

Allow most scraping

Reddit’s robot.txt file in September 2024

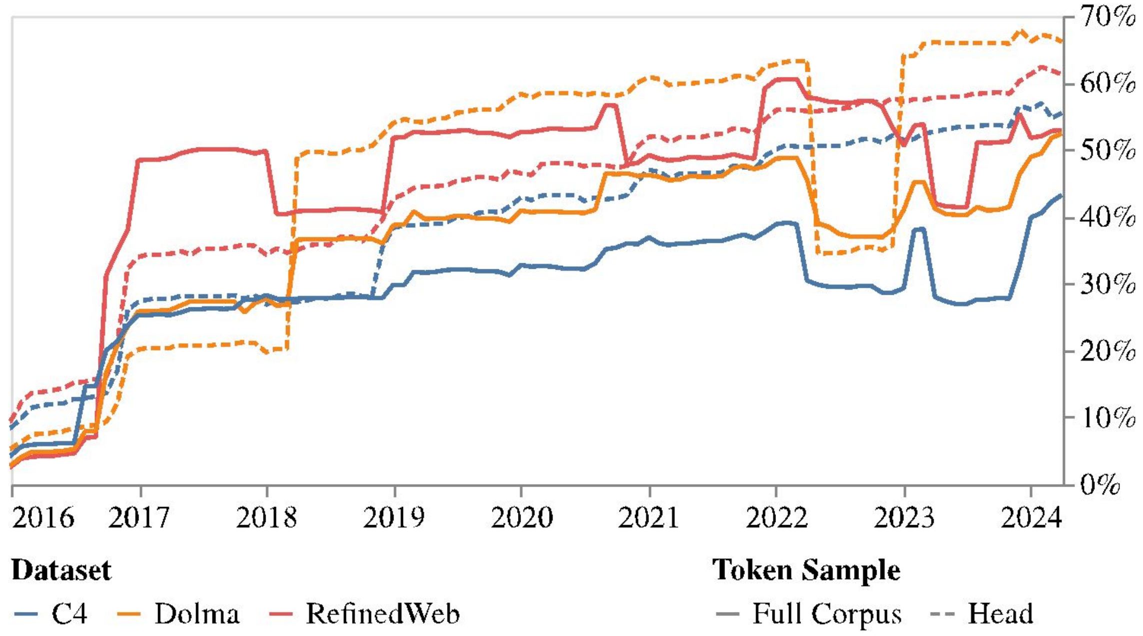
```
# Welcome to Reddit's robots.txt
# Reddit believes in an open internet, but not the misuse of public content.
# See https://support.reddithelp.com/hc/en-us/articles/26410290525844-Public-Content-Policy Reddit's Public Content Policy for access and use restrictions to Reddit content.
# See https://www.reddit.com/r/reddit4researchers/ for details on how Reddit continues to support research and non-commercial use.
# policy: https://support.reddithelp.com/hc/en-us/articles/26410290525844-Public-Content-Policy

User-agent: *
Disallow: /
```

How does web data become **unavailable**?

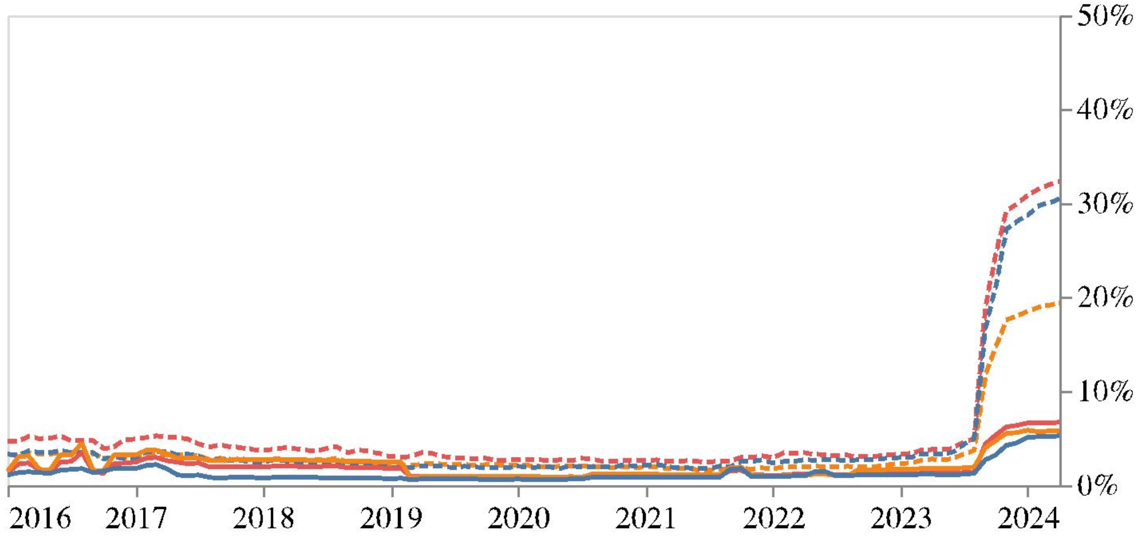
- Introduction of paywalls
- Restrictive terms of service
- Implementation of Robots Exclusion Protocol
- Increased enforcement of copyright law

Datasets created in the past may not be creatable today.

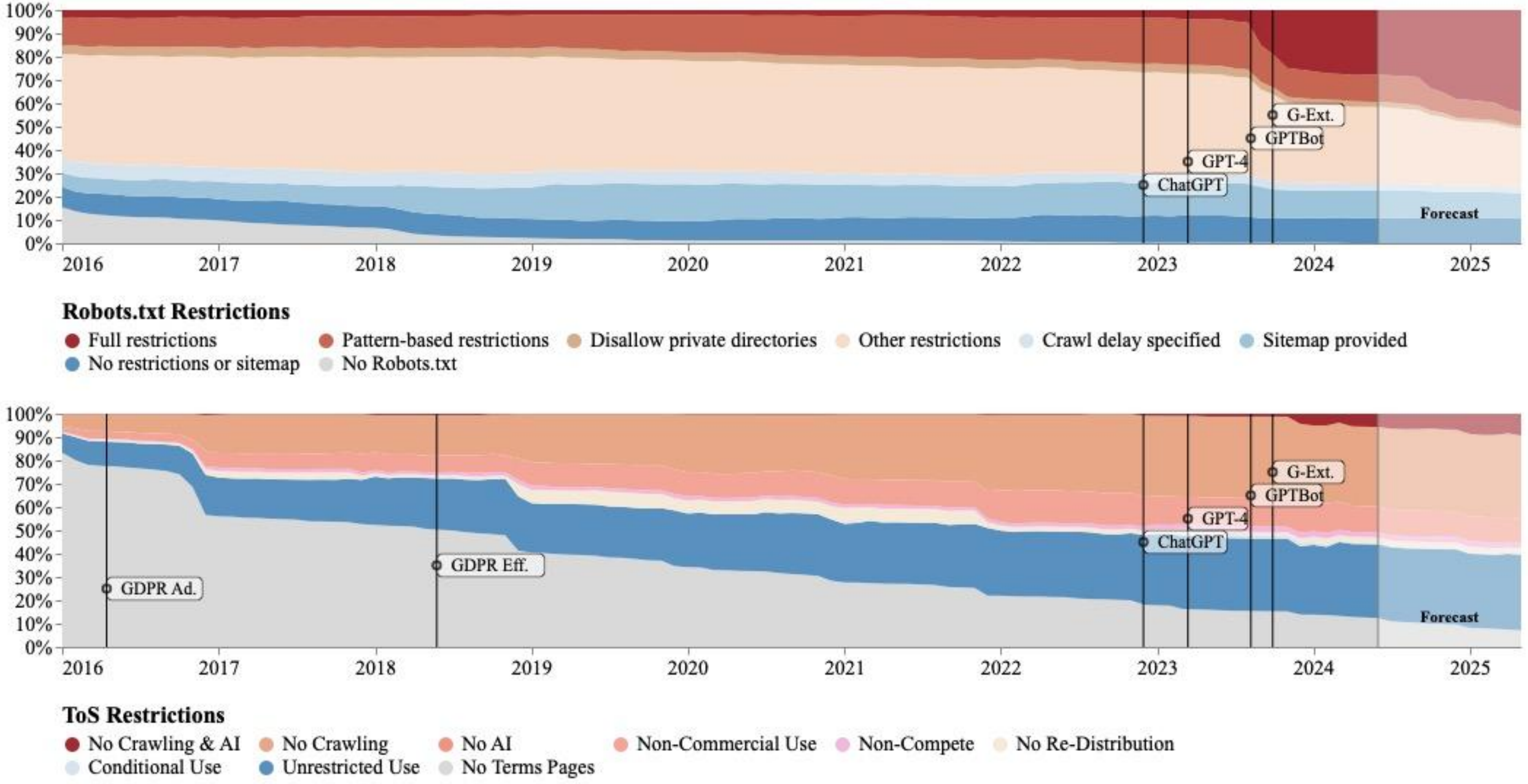
Percentage of examples in dataset restricted by Terms of Service



Percentage of examples in dataset restricted by robots.txt



Datasets created in the past may not be creatable today.



Automatic Evaluation of LLMs

Large Language Models: Methods and Applications

Daphne Ippolito and Chenyan Xiong

How do we identify when one model is better than another?

What is automatic evaluation?

- **Construct**: the property of a system that we want to measure
  - Quality
  - Informativeness
  - Toxicity
  - Interestingness

What is automatic evaluation?

- **Construct**: the property of a system that we want to measure
  - Quality
  - Informativeness
  - Toxicity
  - Interestingness
- **Operationalization:** the measure we use to quantify the construct
  - Perplexity
  - Automatic toxicity score
  - Accuracy at some task
  - Lexical diversity

What is automatic evaluation?

- **Construct**: the property of a system that we want to measure
    - Quality
    - Informativeness
    - Toxicity
    - Interestingness
- **Operationalization:** the measure we use to quantify the construct
    - Perplexity
    - Automatic toxicity score
    - Accuracy at some task
    - Lexical diversity
- **Measurement:** scalar values that we expect to be monotonically related with the construct of interest
    - Humans often understand the construct and can provide accurate ratings or labels.

Intrinsic vs. extrinsic evaluation

- Think of technology as an intervention into a broader process or task.
- **Extrinsic evaluation:** end-to-end evaluation of the entire process or task
- **Intrinsic evaluation:** evaluation of specific components
    - correlated with downstream construct
    - correlated with multiple downstream constructs
    - correlated with important subtask
- Understanding the relationship between different metrics is a fundamental problem in evaluation.

For many constructs, human evaluation is ideal.

Example: evaluating text quality

Please Rate the Story Fragment

The goal of this task is to rate story fragments on four criteria.

NOTE: Please take the time to **fully read** and **understand** the story fragment. **We will reject** submissions from workers that are clearly spamming the task.

Story Fragment

The night before came as a shock for Oren, he was always a conscientious child. It was a necessary skill of a new master, an inherent capability to make the world a better place. But no, today, the day he brought his sister to his cooking school was the first time Oren had been shocked out of a small calm. He looked over at his sister in the small room, who was idly flipping through the magazine he had brought with him, and then back to the breakfast. It took all his willpower to stay calm, he could tell from the way the noodles he was looking at were slathered in gherkin and he felt the freshness of the rice. He shook his head in disbelief, his stomach began to churn and he was too exhausted to react, he was just preparing to go to bed.

1. How **grammatically correct** is the text of the story fragment? (on a scale of 1-5, with 1 *being the lowest*)

   (lowest) ○ **1** ○ **2** ○ **3** ○ **4** ○ **5** (highest)

2. How well do **the sentences** in the story fragment ***fit together***? (on a scale of 1-5, with 1 *being the lowest*)

   (lowest) ○ **1** ○ **2** ○ **3** ○ **4** ○ **5** (highest)

3. How **enjoyable** do you find the story fragment? (on a scale of 1-5, with 1 *being the lowest*)

   (lowest) ○ **1** ○ **2** ○ **3** ○ **4** ○ **5** (highest)

4. Now read the **PROMPT** based on which the story fragment was written.

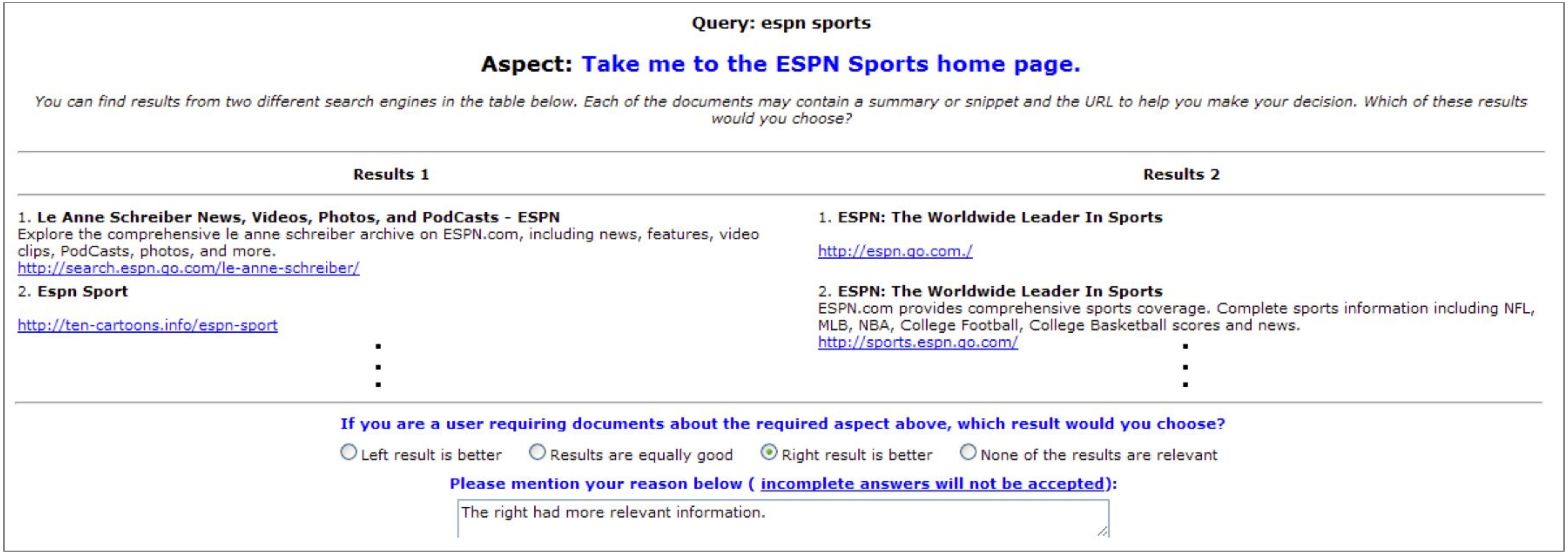
   PROMPT: After brushing your teeth in the morning you go downstairs to fry an egg, but when you try the frying pan buzzes at you and text appears reading, ``level 18 cooking required to use object''.

   How **relevant** is the **story fragment** to the **prompt**? (on a scale of 1-5, with 1 *being the lowest*)

   (lowest) ○ **1** ○ **2** ○ **3** ○ **4** ○ **5** (highest)

Submit

Example: evaluating usefulness


Query: espn sports

Aspect: Take me to the ESPN Sports home page.

You can find results from two different search engines in the table below. Each of the documents may contain a summary or snippet and the URL to help you make your decision. Which of these results would you choose?

Results 1	**Results 2**
1. **Le Anne Schreiber News, Videos, Photos, and PodCasts - ESPN** Explore the comprehensive le anne schreiber archive on ESPN.com, including news, features, video clips, PodCasts, photos, and more. http://search.espn.go.com/le-anne-schreiber/	1. **ESPN: The Worldwide Leader In Sports** http://espn.go.com./
2. **Espn Sport** http://ten-cartoons.info/espn-sport	2. **ESPN: The Worldwide Leader In Sports** ESPN.com provides comprehensive sports coverage. Complete sports information including NFL, MLB, NBA, College Football, College Basketball scores and news. http://sports.espn.go.com/
⋮	⋮

If you are a user requiring documents about the required aspect above, which result would you choose?

○ Left result is better ○ Results are equally good ◉ Right result is better ○ None of the results are relevant

Please mention your reason below (incomplete answers will not be accepted):

The right had more relevant information.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using Mechanical Turk to evaluate open-ended text generation. EMNLP. 2021.

Why do automatic evaluation over human evaluation?

- Human evaluation is expensive.
  - Time: recruiting, training, rating
  - Cost: money to raters
- Human evaluation often does not scale.
  - New systems need a new evaluation
  - Side-by-side comparisons require $O(n^2)$ comparisons for n systems
- Automatic evaluation is sufficient
  - In many cases, there are automatic metrics which highly correlate with the construct of interest.
  - Can you think of any?

Goal when designing automatic evaluation

A resuable, offline metric that either

- Directly models a construct of interest.
- Models reliable human labels of that construct.

General form of an evaluation metric

$$\mu(x, \tilde{y}, \mathcal{D}_x)$$

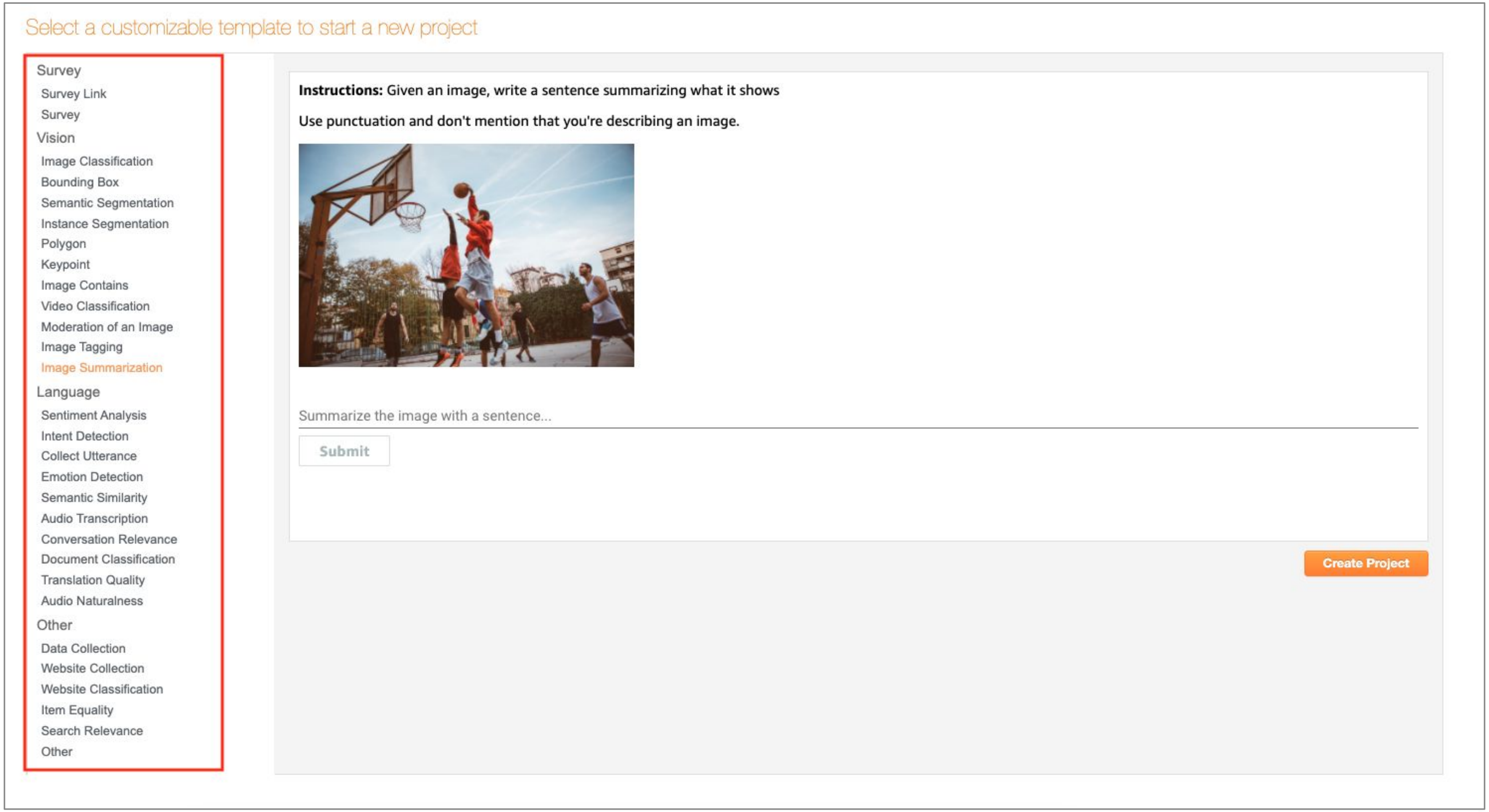
x instance

$\tilde{y}$ system prediction

$\mathcal{D}_x$ test information about x

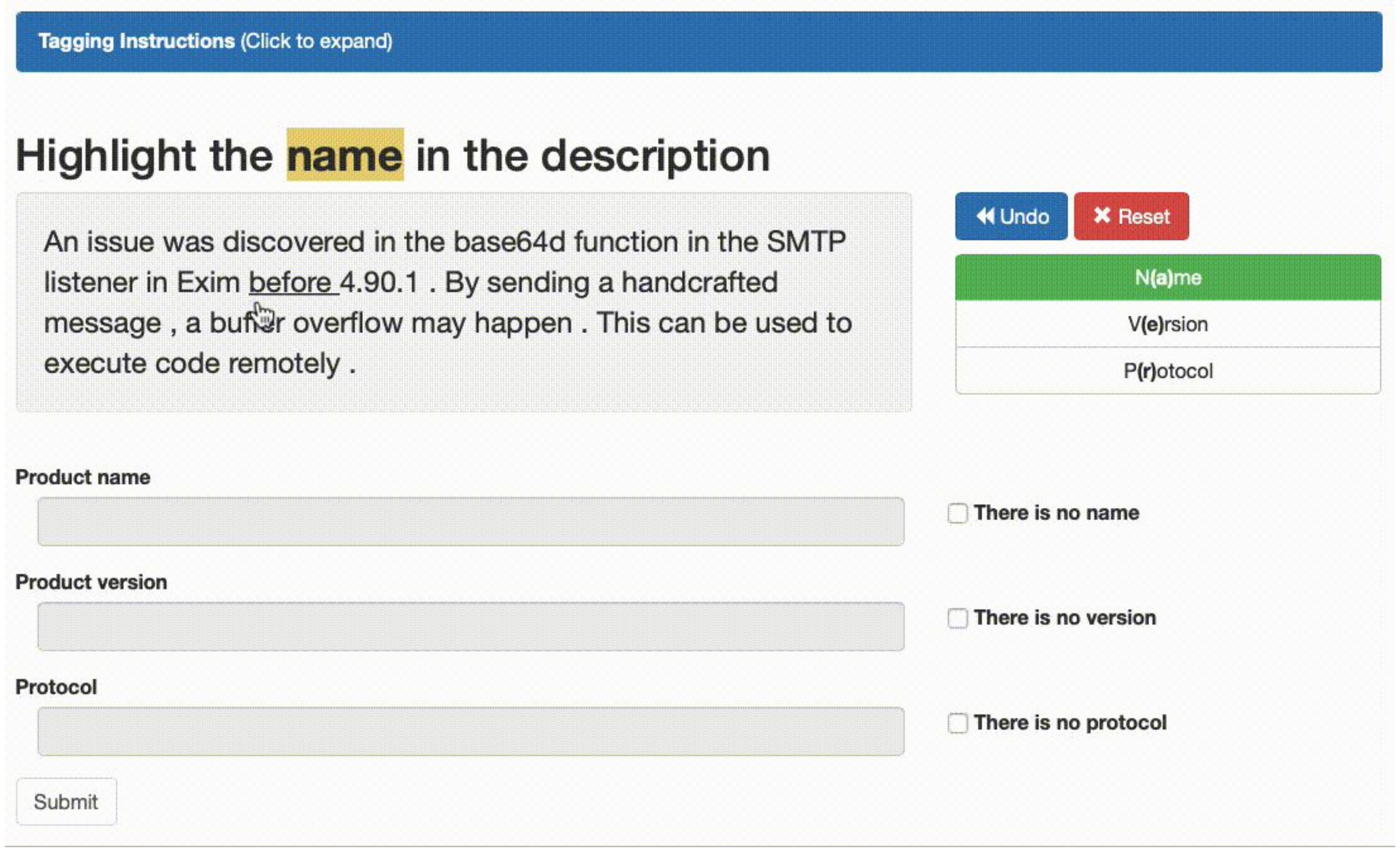
x	$\tilde{y}$	$\mathcal{D}_x$
word prefix	next word	true next word
document	summary	gold summary
question	answer	correct answer
question	ranked answers	correct answer
query	ranked items	relevant items
query	ranked items	logged clicks

Acquiring $\mathcal{D}_x$ through human annotation



Create an Amazon Mechanical Turk project, https://docs.aws.amazon.com/AWSMechTurk/latest/RequesterUI/CreatingYourBatchofHITs.html, 2024.

Acquiring $\mathcal{D}_x$ through human annotation



Goals for this lecture

- Review a catalog of metrics for NLP tasks.
  - *All* of these metrics are useful for language model development, *depending on the context*.
- Review cases where metrics are inconsistent with human raters or constructs.
  - This is to emphasize the importance of understanding metrics, not to dismiss them altogether!

Task templates common for automatic evaluation

- **classification:** given a context *x*, generate a single decision
    - *x*: question, document
    - *y*: label
- **sequence generation:** given a context *x*, generate a sequence of decisions.
    - *x*: prefix, question, document
    - *y*: next word(s), answer string, document summary
- **ranking:** given a context *x*, generate a ranking of items.
    - *x*: prefix, question, document, query
    - *y*: list of next words, answer strings, document summaries, documents
- **multi-task:** support multiple tasks
    - *x*: {prefix, question, document, query}
    - *y*: {list of next words, answer strings, document summaries, documents}

Terminology for evaluating sequence generation

Eval metric: $\mu(y, \tilde{y})$

y target sequence (reference)

$\tilde{y}$ predicted sequence (hypothesis)

Note: this is slightly different from the terminology we saw earlier in class, where I was using $\hat{\boldsymbol{y}}$ to refer to the model output, and $\boldsymbol{y}$ to refer to the target.

Evaluating sequence generation: exact match

$$\mu(y, \tilde{y}) = \mathrm{I}(y = \tilde{y})$$

- advantages
  - high precision: if metric is 1, then we have a good sequence
- disadvantages
  - low recall: in many situations, if the metric is not 1, then we still may have a good hypothesis.
- uses
  - question answering
  - numerical reasoning

Evaluating sequence generation: word error rate

$$\mu(y, \tilde{y}) = \frac{\delta(y, \tilde{y})}{|y|}$$

$$\begin{array}{rl} \delta(y, \tilde{y}) & \text{word edit distance} \\ & \text{between } y \text{ and } \tilde{y} \\ |y| & \text{length of } y \end{array}$$

- advantages
  - relaxes exact match
- disadvantages
  - uniform weight on all transformations
  - semantically similar words ignored
  - questionable correlation with understanding
- uses
  - speech recognition
  - machine translation

Challenge: these metrics may not be correlated with task performance.

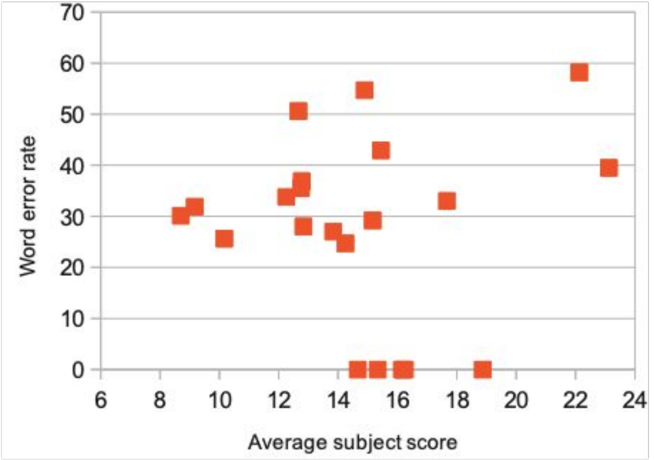


Figure 1: Meeting-level word error rate vs average H-score for all transcript conditions.

https://www.speechmatics.com/company/articles-and-news/the-future-of-word-error-rate
Benoit Favre, et al. Automatic human utility evaluation of ASR systems: does WER really predict performance?. In Proc. Interspeech 2013.

Evaluating sequence generation: perplexity

How surprised is the LM by the text sequence *y*?

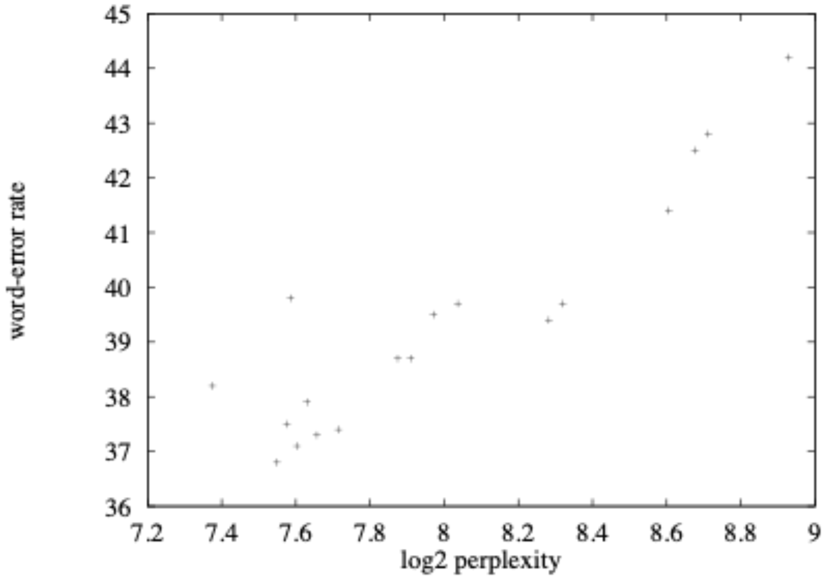
$$\mu(y, \theta) = \exp\left(-\frac{1}{|y|}\sum_{i=1}^{|y|} \log p_\theta(y_i | y_{1:i-1})\right)$$

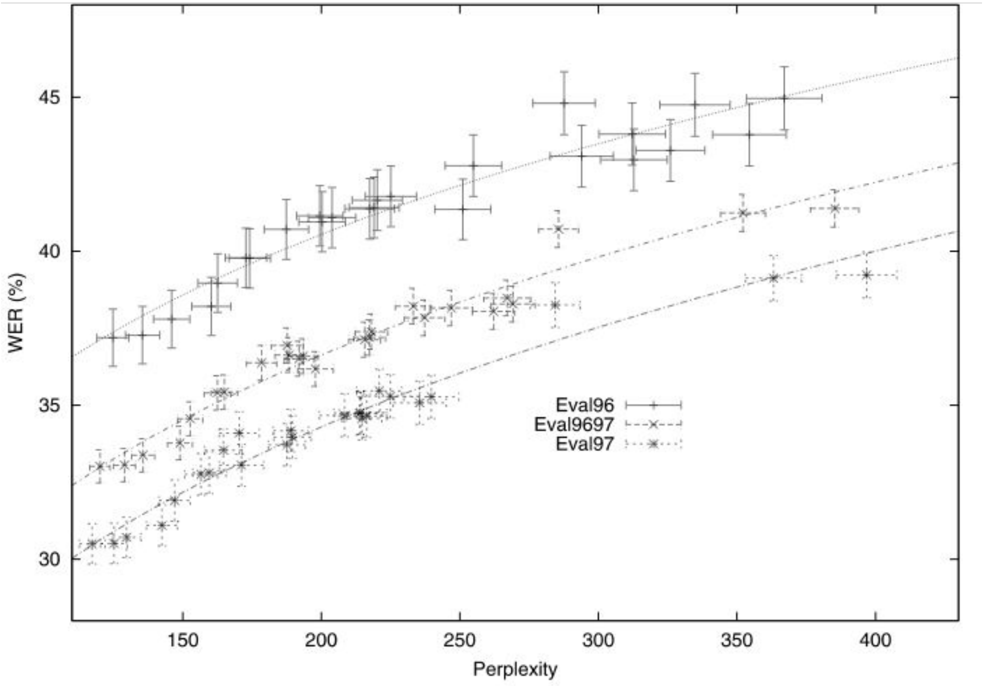
θ language model

- advantages
  - relaxes exact match
- disadvantages
  - per-token decisions
  - vocabulary/model-dependent
- uses
  - language modeling

Evaluating sequence generation: perplexity: Perplexity

Note: intrinsic metrics can be correlated with each other





Chen, S., Beeferman, D., Rosenfeld, R., . Evaluation metrics for language models. In: DARPA Broadcast News Transcription and Understanding Workshop. 1998.
Dietrich Klakow and Jochen Peters. Testing the correlation of word error rate and perplexity. Speech Communication, 38(1):19-28, 2002.

Typical uses of perplexity

- Intrinsic eval: Measure how well an LM models human language.
  - Common use case: during training, plot perplexity of a withheld validation set every k steps
- Extrinsic eval: Given we are confident that our LM reasonably models human language, use it in tasks that require measuring how "human-like" a piece of text is.
  - Common use case: filtering out garbage text
  - Common use case: detection of LM-generated text

Evaluating sequence generation: BLEU score

Let's see an example:

- Target "correct" responses:
  - ***Target 1: He picked up the ball from the ground .***
  - ***Target 2: He took the book off the floor .***
- Model generation:
  - ***He picked the He sphere off the the the floor .***

Word	Freq. in gen	Max. freq in any target	Clipped count
He	2	1	1
picked	1	1	1
the	4	2	2
sphere	1	0	0
off	1	1	1
floor	1	1	1
.	1	1	1

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. ACL 2002, 311—318.

Evaluating sequence generation: BLEU score

Let's see an example:

- Target "correct" responses:
  - ***Target 1: He picked up the ball from the ground .***
  - ***Target 2: He took the book off the floor .***
- Model generation:
  - ***He picked the He sphere off the the the floor .***

Word	Freq. in gen	Max. freq in any target	Clipped count
He	2	1	1
picked	1	1	1
the	4	2	2
sphere	1	0	0
off	1	1	1
floor	1	1	1
.	1	1	1

The total number of words in the target is 11.

The total clipped count is 7.

So the clipped 1-gram precision is 7/11 = 0.64

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. ACL 2002, 311—318.

Basic Implementation of BLEU Score

1. Compute the clipped 1-gram precision.

   Target Sentence: The guard arrived late because ~~it was raining~~

   Predicted Sentence: The guard arrived late because of the rain

2. Compute the clipped 2-gram precision.

   Target Sentence: The guard arrived late because it was raining

   Predicted Sentence: The guard arrived late because of the rain

3. Compute the clipped 3-gram precision.

   Target Sentence: The guard arrived late because it was raining

   Predicted Sentence: The guard arrived late because of the rain

4. Compute the clipped 4-gram precision.

   Target Sentence: The guard arrived late because it was raining

   Predicted Sentence: The guard arrived late because of the rain

5. Take the geometric mean of all of the above.

$$\prod_{i=1}^{k} \left(\frac{|\mathcal{G}_i(y) \cap \mathcal{G}_i(\tilde{y})|}{|\mathcal{G}_i(\tilde{y})|} \right)^{1/k}$$

$\mathcal{G}_n(s)$ n-gram multiset in s

Sequences: BLEU

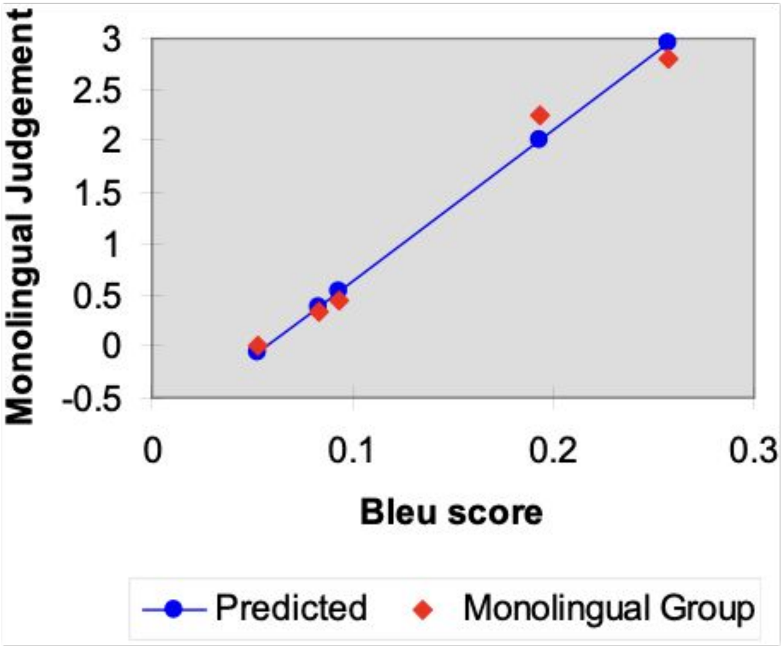
- advantages
  - relaxes exact match
  - Handles tasks with multiple target sequences
  - correlation with human preferences (MT)
- disadvantages
  - semantically similar words ignored
- uses
  - machine translation

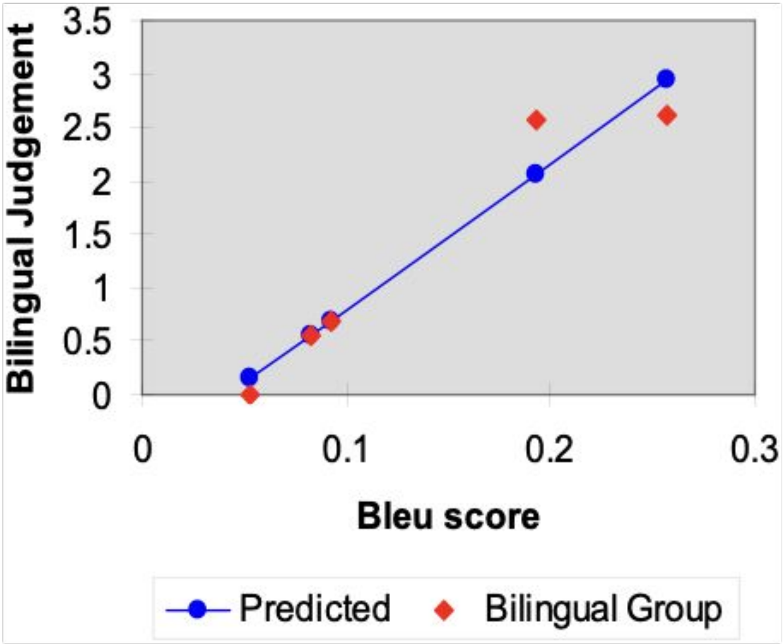
$$\mu(y, \tilde{y}, k) = \prod_{i=1}^{k} \left(\frac{|\mathcal{G}_i(y) \cap \mathcal{G}_i(\tilde{y})|}{|\mathcal{G}_i(\tilde{y})|} \right)^{1/k}$$

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. ACL 2002.

Advantages of BLEU

Measure correlates with human preferences.



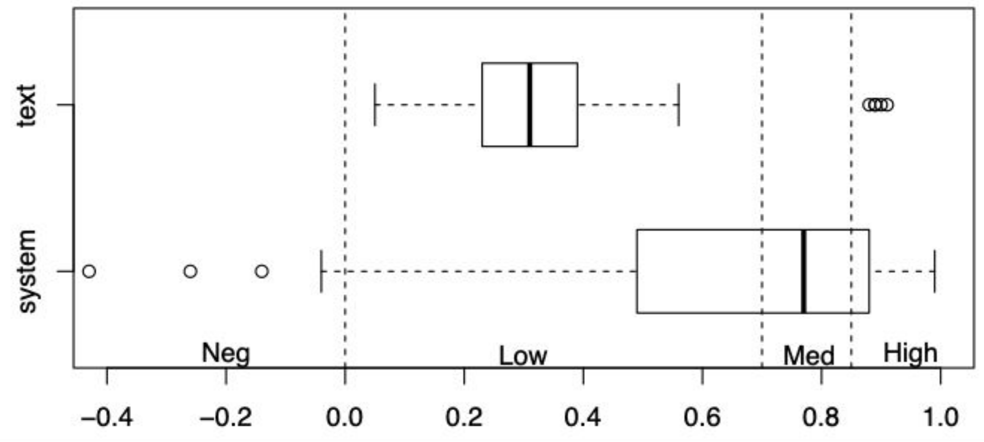


Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, ACL 2002.

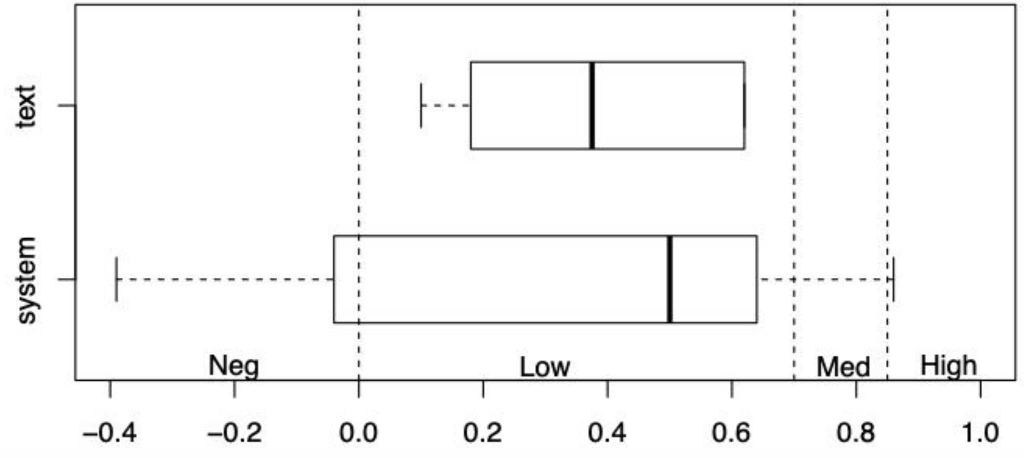
Advantages of BLEU

Measure correlates with human preferences **on SOME tasks more than others**.

machine translation



natural language generation



Ehud Reiter. A structured review of the validity of BLEU. Computational Linguistics, 44(3):393--401, September 2018.

Evaluating sequence generation: $ROUGE_k$

- BLEU measures precision: how many of the generated words are in the references.
- ROUGE is a complimentary to BLEU.
- It measures recall: how many of the words in the references are found in the generation.

Sequences: ROUGE$_k$

- advantages
  - relaxes exact match
  - correlation with human preferences (MDS)
- disadvantages
  - semantically similar words ignored
- uses
  - multidocument summarization (MDS)

$$\mu(y, \tilde{y}, k) = \frac{|\mathcal{G}_k(y) \cap \mathcal{G}_k(\tilde{y})|}{|\mathcal{G}_k(y)|}$$

in practice…
- k={1,2}
- fixed length hypothesis
- extended for multiple targets

Chin-Yew Lin. Rouge: a package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editors, Text summarization branches out. ACL 2004.

Sequences: ROUGE_k

	DUC 2001 100 WORDS SINGLE DOC						DUC 2002 100 WORDS SINGLE DOC					
	1 REF			3 REFS			1 REF			2 REFS		
Method	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP
R-1	0.76	0.76	0.84	0.80	0.78	0.84	0.98	0.98	0.99	0.98	0.98	0.99
R-2	0.84	0.84	0.83	0.87	0.87	0.86	0.99	0.99	0.99	0.99	0.99	0.99
R-3	0.82	0.83	0.80	0.86	0.86	0.85	0.99	0.99	0.99	0.99	0.99	0.99
R-4	0.81	0.81	0.77	0.84	0.84	0.83	0.99	0.99	0.98	0.99	0.99	0.99
R-5	0.79	0.79	0.75	0.83	0.83	0.81	0.99	0.99	0.98	0.99	0.99	0.98
R-6	0.76	0.77	0.71	0.81	0.81	0.79	0.98	0.99	0.97	0.99	0.99	0.98
R-7	0.73	0.74	0.65	0.79	0.80	0.76	0.98	0.98	0.97	0.99	0.99	0.97
R-8	0.69	0.71	0.61	0.78	0.78	0.72	0.98	0.98	0.96	0.99	0.99	0.97
R-9	0.65	0.67	0.59	0.76	0.76	0.69	0.97	0.97	0.95	0.98	0.98	0.96
R-L	0.83	0.83	0.83	0.86	0.86	0.86	0.99	0.99	0.99	0.99	0.99	0.99
R-S*	0.74	0.74	0.80	0.78	0.77	0.82	0.98	0.98	0.98	0.98	0.97	0.98
R-S4	0.84	0.85	0.84	0.87	0.88	0.87	0.99	0.99	0.99	0.99	0.99	0.99
R-S9	0.84	0.85	0.84	0.87	0.88	0.87	0.99	0.99	0.99	0.99	0.99	0.99
R-SU*	0.74	0.74	0.81	0.78	0.77	0.83	0.98	0.98	0.98	0.98	0.98	0.98
R-SU4	0.84	0.84	0.85	0.87	0.87	0.87	0.99	0.99	0.99	0.99	0.99	0.99
R-SU9	0.84	0.84	0.85	0.87	0.87	0.87	0.99	0.99	0.99	0.99	0.99	0.99
R-W-1.2	0.85	0.85	0.85	0.87	0.87	0.87	0.99	0.99	0.99	0.99	0.99	0.99

Table 1: Pearson's correlations of 17 ROUGE measure scores vs. human judgments for the DUC 2001 and 2002 100 words single document summarization tasks

correlation with human preferences depends on systems!

Surrogate	P = 1	P = 2	P = 4
HEAD (RP)	0.1270	0.1943	0.3140
HUM (RP)	0.0632	0.1096	0.1391
HEAD (LDC)	-0.0968	-0.0660	-0.0099
HUM (LDC)	-0.0395	-0.0236	-0.0187

Table 5: Pearson Correlations with ROUGE-1 for Relevance-Prediction (RP) and LDC-Agreement (LDC), where Partition size (P) = 1, 2, and 4

HEAD: "headline" system
HUM: human summary

Chin-Yew Lin. Rouge: a package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editors, Text summarization branches out: proceedings of the acl-04 workshop, 74--81, Barcelona, Spain, July 2004. , Association for Computational Linguistics.
Bonnie Dorr, Christof Monz, Stacy President, Richard Schwartz, and David Zajic. A methodology for extrinsic evaluation of text summarization: does ROUGE correlate?. In Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005.

Sequences: addressing semantically similar words

> Based on this experiment, we conjecture that ROUGE may not be a good method for measuring the usefulness of summaries when the summaries are not extractive. That is, if someone intentionally writes summaries that contain different words than the story, the summaries will also likely contain different words than a reference summary, resulting in low ROUGE scores.

- All metrics so far only consider exact token matches.
- Penalize models that include synonyms.

Bonnie Dorr, Christof Monz, Stacy President, Richard Schwartz, and David Zajic. A methodology for extrinsic evaluation of text summarization: does ROUGE correlate?. In Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005.

Sequences: character n-gram precision (chrP)

$$\mu_{\mathrm{P}}(y, \tilde{y}, k) = \frac{1}{k} \sum_{i=1}^{k} \frac{|\Gamma_i(y) \cap \Gamma_i(\tilde{y})|}{|\Gamma_i(\tilde{y})|}$$

$$\Gamma_n(s) \quad \text{character } n\text{-gram multiset in } s$$

Maja Popović. ChrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the tenth workshop on statistical machine translation, 392--395, Lisbon, Portugal, September 2015. , Association for Computational Linguistics.

Sequences: character n-gram recall (chrR)

$$\mu_{\mathrm{R}}(y, \tilde{y}, k) = \frac{1}{k} \sum_{i=1}^{k} \frac{|\Gamma_i(y) \cap \Gamma_i(\tilde{y})|}{|\Gamma_i(y)|}$$

$$\Gamma_n(s) \quad \text{character } n\text{-gram multiset in } s$$

Maja Popović. ChrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the tenth workshop on statistical machine translation, 392--395, Lisbon, Portugal, September 2015. , Association for Computational Linguistics.

Sequences: character n-gram F-score (chrF)

$$\mu(y, \tilde{y}, k, \beta) = (1 - \beta^2) \frac{\mu_{\mathrm{P}}(y, \tilde{y}, k) \times \mu_{\mathrm{R}}(y, \tilde{y}, k)}{\beta^2 \times \mu_{\mathrm{P}}(y, \tilde{y}, k) + \mu_{\mathrm{R}}(y, \tilde{y}, k)}$$

Maja Popović. ChrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the tenth workshop on statistical machine translation, 392--395, Lisbon, Portugal, September 2015. , Association for Computational Linguistics.

Sequences: character n-gram F-score (chrF)

year	WORDF	CHRF	CHRF3	BLEU	TER	METEOR
2014 (r)	0.810	0.805	0.857	0.845	0.814	0.822
2013 (ρ)	0.874	0.873	/	0.835	0.791	0.876
2012 (ρ)	0.659	0.696	/	0.671	0.682	0.690

Table 2: Average system-level correlations on WMT14 (Pearson's r), WMT13 and WMT12 data (Spearman's ρ) for word 4-gram F1 score, character 6-gram F1 score and character 6-gram F3 score together with the three mostly used metrics BLEU, TER and METEOR.

Maja Popović. ChrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the tenth workshop on statistical machine translation, 392--395, Lisbon, Portugal, September 2015. , Association for Computational Linguistics.

Sequences: character n-gram F-score (chrF)

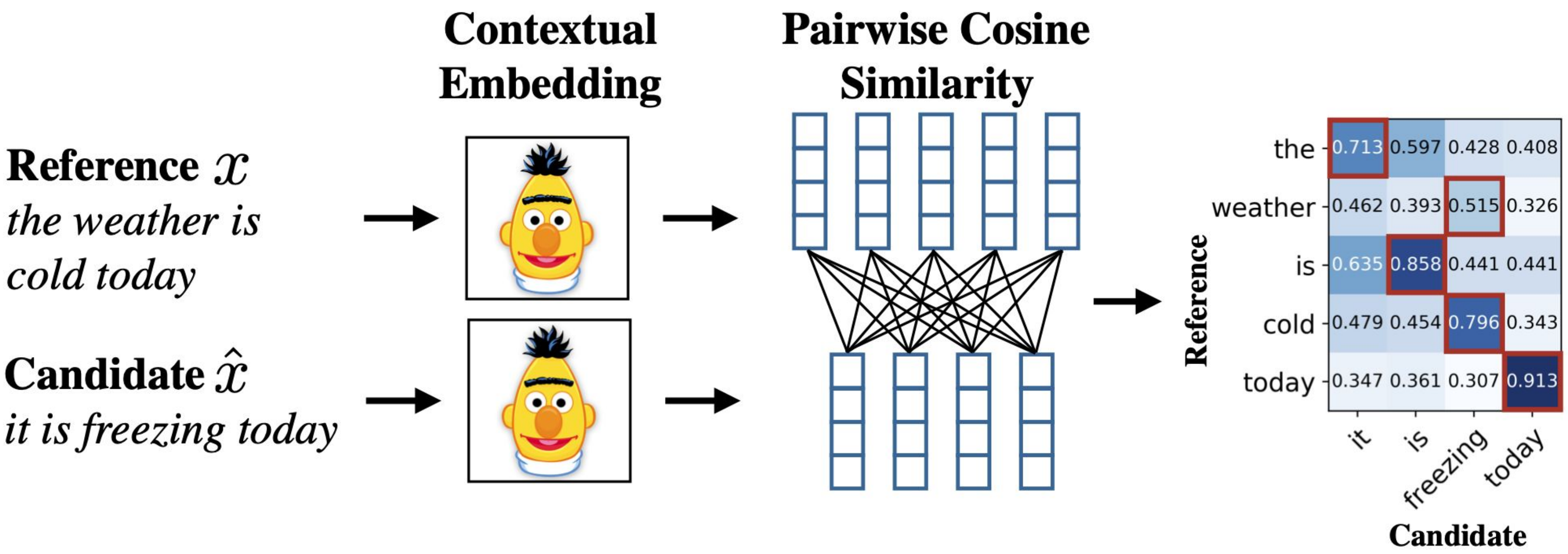
- advantages
  - relaxes exact match and captures (some) morphological similarity
- disadvantages
  - does not capture similarity when there is no character overlap
- uses
  - machine translation
  - summarization

Maja Popović. ChrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the tenth workshop on statistical machine translation, 392--395, Lisbon, Portugal, September 2015. , Association for Computational Linguistics.

From overlap-based metrics to evaluating semantic similarity.

- All the metrics we've discussed so far fail when an answer is correct, but word overlap with the groundtruth answer is low.
- Can we leverage advances in NLP to address lack of non-lexical similarity in metrics?
    - Let's we have access to a model that provides word similarity.

Evaluating sequence generation: BERT-based similarity



Tianyi Zhang et al. BertScore: evaluating text generation with Bert. In International conference on learning representations, 2020.

Evaluating sequence generation: BERT-based prevision and recall

$$\mu_{\mathrm{P}}(y, \tilde{y}) = \frac{1}{|\tilde{y}|} \sum_{\tilde{y}_i \in \tilde{y}} \max_{y_i \in y} \phi_i^\top \tilde{\phi}_i$$

$$\mu_{\mathrm{R}}(y, \tilde{y}) = \frac{1}{|y|} \sum_{y_i \in y} \max_{\tilde{y}_i \in \tilde{y}} \phi_i^\top \tilde{\phi}_i$$

$$\phi_i \quad \text{Bert embedding of } y_i$$

in practice…

- can combine P and R into an F-score
- weigh terms by discrimination power (idf)

Tianyi Zhang et al. BertScore: evaluating text generation with Bert. In International conference on learning representations, 2020.

BertScore correlated with human judgement.

Metric	en↔cs (5/5)	en↔de (16/16)	en↔et (14/14)	en↔fi (9/12)	en↔ru (8/9)	en↔tr (5/8)	en↔zh (14/14)
BLEU	.970/**.995**	.971/**.981**	**.986/.975**	.973/**.962**	.979/**.983**	**.657**/.826	.978/.947
ITER	.975/.915	.990/**.984**	.975/**.981**	**.996/.973**	.937/.975	**.861**/.865	.980/ –
RUSE	.981/ –	.997/ –	**.990/** –	.991/ –	**.988/** –	**.853/** –	**.981/** –
YiSi-1	.950/**.987**	.992/**.985**	.979/**.979**	.973/.940	**.991/.992**	**.958/.976**	.951/**.963**
P_{BERT}	.980/**.994**	**.998/.988**	**.990/.981**	.995/.957	.982/**.990**	**.791/.935**	.981/.954
R_{BERT}	**.998/.997**	.997/**.990**	.986/**.980**	**.997/.980**	**.995/.989**	.054/.879	**.990/.976**
F_{BERT}	**.990/.997**	**.999/.989**	.990/**.982**	**.998/.972**	**.990**/.990	**.499**/.908	**.988**/.967
F_{BERT} (idf)	.985/**.995**	**.999/.990**	**.992/.981**	.992/**.972**	**.991/.991**	**.826/.941**	**.989/.973**

Table 1: Absolute Pearson correlations with system-level human judgments on WMT18. For each language pair, the left number is the to-English correlation, and the right is the from-English. We bold correlations of metrics not significantly outperformed by any other metric under Williams Test for that language pair and direction. The numbers in parenthesis are the number of systems used for each language pair and direction.

Tianyi Zhang et al. BertScore: evaluating text generation with Bert. In International conference on learning representations, 2020.

Sequences: BERTScore

- advantages
  - relaxes exact match
  - incorporates semantic similarity
- disadvantages
  - dependent on embedding model
- uses
  - machine translation
  - image captioning systems

$$\mu_{\mathrm{P}}(y, \tilde{y}) = \frac{1}{|\tilde{y}|} \sum_{y_i \in y} \max_{\tilde{y}_i \in \tilde{y}} \phi_i^\top \tilde{\phi}_i$$

$$\mu_{\mathrm{R}}(y, \tilde{y}) = \frac{1}{|y|} \sum_{\tilde{y}_i \in \tilde{y}} \max_{y_i \in y} \phi_i^\top \tilde{\phi}_i$$

ϕ_i Bert embedding of y_i

Tianyi Zhang et al. BertScore: evaluating text generation with Bert. In International conference on learning representations, 2020.

What we've covered so far

- metrics are models of...
    - ...unobserved constructs
    - ...human preferences
- none of the metrics we have studied so far directly model these things
- given a collection of human judgments,

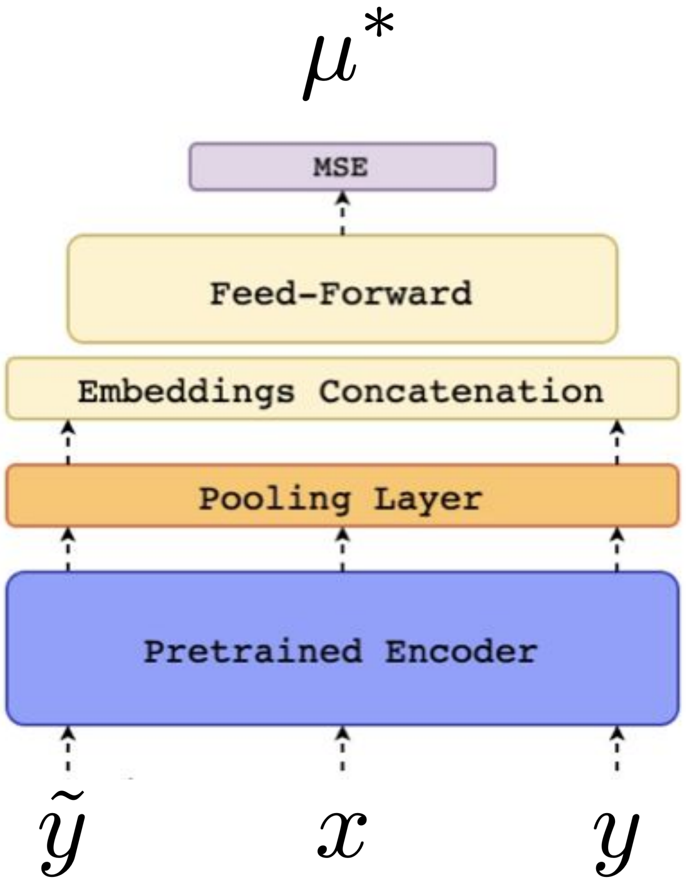
$$\{\langle x, y, \tilde{y}, \mu^* \rangle\}$$

can we directly model constructs or preferences?

Evaluating Sequence Generation: COMET

Main idea: train models to predict human preferences.

Method 1: train a regression model to predict the ratings a human annotator would give.



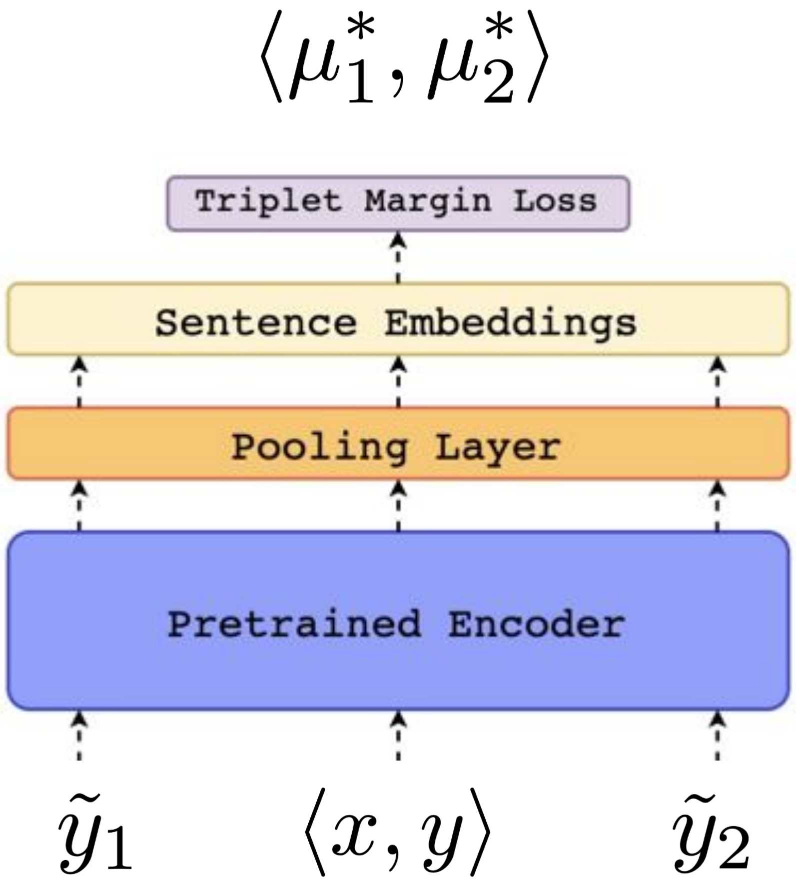
regress against the human ratings

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: a neural framework for MT evaluation. ENLP 2020

Evaluating Sequence Generation: COMET

Main idea: train models to predict human preferences.

Method 2: train a ranking model to give higher rankings to examples a human annotator would rank higher.



learn to rank better hypothesis

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: a neural framework for MT evaluation. ENLP 2020

Evaluating Sequence Generation: COMET

As you'd expect, COMET correlates highly with human preferences.

Table 1: Kendall's Tau (τ) correlations on language pairs with English as source for the WMT19 Metrics DARR corpus. For BERTSCORE we report results with the default encoder model for a complete comparison, but also with XLM-RoBERTa (base) for fairness with our models. The values reported for YiSi-1 are taken directly from the shared task paper (Ma et al., 2019).

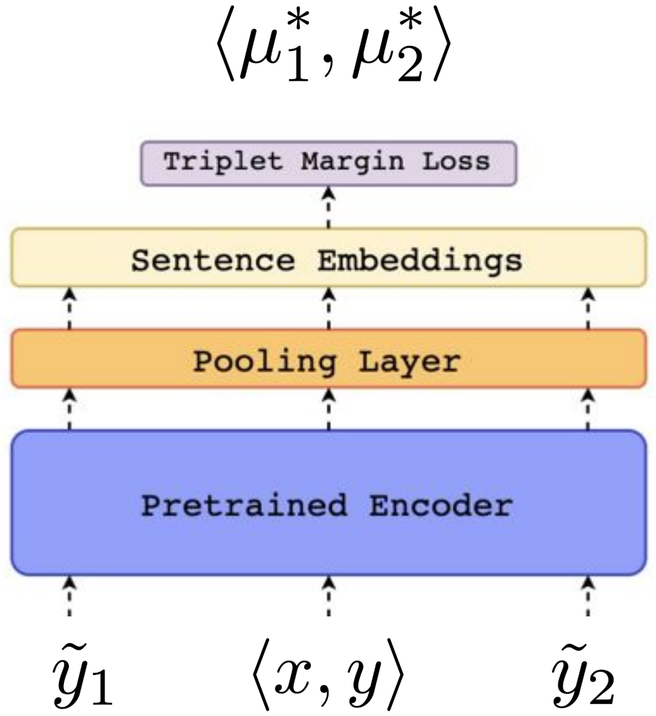
Metric	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
BLEU	0.364	0.248	0.395	0.463	0.363	0.333	0.469	0.235
CHRF	0.444	0.321	0.518	0.548	0.510	0.438	0.548	0.241
YISI-1	0.475	0.351	0.537	0.551	0.546	0.470	0.585	0.355
BERTSCORE (default)	0.500	0.363	0.527	0.568	0.540	0.464	0.585	0.356
BERTSCORE (xlmr-base)	0.503	0.369	0.553	0.584	0.536	0.514	0.599	0.317
COMET-HTER	0.524	0.383	0.560	0.552	0.508	0.577	0.539	0.380
COMET-MQM	0.537	0.398	0.567	0.564	0.534	0.574	**0.615**	0.378
COMET-RANK	**0.603**	**0.427**	**0.664**	**0.611**	**0.693**	**0.665**	0.580	**0.449**

directly model human ratings works

modeling human *preferences* tends to work better

Sequences: COMET

- advantages
  - relaxes exact match
  - incorporates semantic similarity
  - directly modeling human
- disadvantages
  - dependent on embedding model
  - dependent on task-specific annotations
- uses
  - machine translation
  - direct modeling applicable to other tasks



Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: a neural framework for MT evaluation. ENLP 2020

Sequences: constructs

- so far, we have focused on “quality”
- sequences have a diverse set of properties we can measure
- need to be precise in what we are measuring, in designing a metric and eliciting human ratings

Criterion Paraphrase	Count
usefulness for task/information need	39
grammaticality	39
quality of outputs	35
understandability	30
correctness of outputs relative to input (content)	29
goodness of outputs relative to input (content)	27
clarity	17
fluency	17
goodness of outputs in their own right	14
readability	14
information content of outputs	14
goodness of outputs in their own right (both form and content)	13
referent resolvability	11
usefulness (nonspecific)	11
appropriateness (content)	10
naturalness	10
user satisfaction	10
wellorderedness	10
correctness of outputs in their own right (form)	9
correctness of outputs relative to external frame of reference (content)	8
ease of communication	7
humanlikeness	7
appropriateness	6
understandability	6
nonredundancy (content)	6
goodness of outputs relative to system use	5
appropriateness (both form and content)	5

David M. Howcroft et al.. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In Proceedings of the 13th international conference on natural language generation, 169--182, Dublin, Ireland, December 2020.

questions?

Ranking

- in many language tasks, users are presented with a list of predictions, not just one,
  - **search**: list of documents
  - **question answering**: list of answers
  - **autocomplete**: list of suggestions
- an LLM can either select the items in the list from a catalog (e.g., search) or generate the items (e.g., QA, autocomplete).
- formally,

$$\pi \quad \text{system ranking}$$

$$\mathcal{Y}^{+} \quad \text{relevant answer set}$$

Ranking

browsing

$$\pi(1)$$

$$\pi(2)$$

$$\pi(3)$$

$$\vdots$$

$$\pi(n-2)$$

$$\pi(n-1)$$

$$\pi(n)$$

Ranking: expected search length

user model: in-order traversal of a ranked list, collecting up to k items.

metric: number of nonrelevant documents skipped before reaching k relevant items.

uses: interpretable metric but not used often

$$\mathrm{ESL}(\mathcal{Y}^+, \pi, k) = \text{min-}k_{i \in \mathcal{Y}^+} \overline{\pi}(i)$$

$$\begin{array}{rl} \text{min-}k & k\text{th smallest value} \\ \overline{\pi}(i) & \text{rank position of item } i \end{array}$$

William S. Cooper. Expected search length,. American Documentation, 19(1):30–41, 1968.

Ranking: reciprocal rank

user model: in-order traversal of a ranked list, satisfied by one item.

metric: inverse of the number of documents skipped before reaching the relevant item.

uses: one relevant answer; impatient user

$$\mathrm{RR}(\mathcal{Y}^+, \pi) = \frac{1}{\mathrm{ESL}(\mathcal{Y}^+, \pi, 1)}$$

E. Voorhees and D. Tice. The trec-8 question answering track evaluation. TREC, 1999.

Ranking: R-precision

user model: in-order traversal of a ranked list, collecting all relevant items.

metric: precision when recall is 1.

uses: multiple relevant answers; user interested in many answers; more patient

$$\mathrm{RPrec}(\mathcal{Y}^+, \pi) = \mathrm{Prec}(\mathcal{Y}^+, \pi_{1:k^*})$$

Ranking: average precision

user model: in-order traversal of a ranked list, collecting all relevant items.

metric: precision averaged over all recall levels.

uses: multiple relevant answers; user interested in many answers; more patient; average quality across all recall requirements.

$$\mathrm{AP}(\mathcal{Y}^+, \pi) = \frac{1}{|\mathcal{Y}^+|} \sum_{i \in \mathcal{Y}^+} \mathrm{Prec}(\mathcal{Y}^+, \pi_{1:\overline{\pi}(i)})$$

$$= \frac{1}{|\mathcal{Y}^+|} \sum_{r=1}^{|\mathcal{Y}^+|} \frac{r}{\mathrm{ESL}(\mathcal{Y}^+, \pi, r)}$$

Ranking: average precision

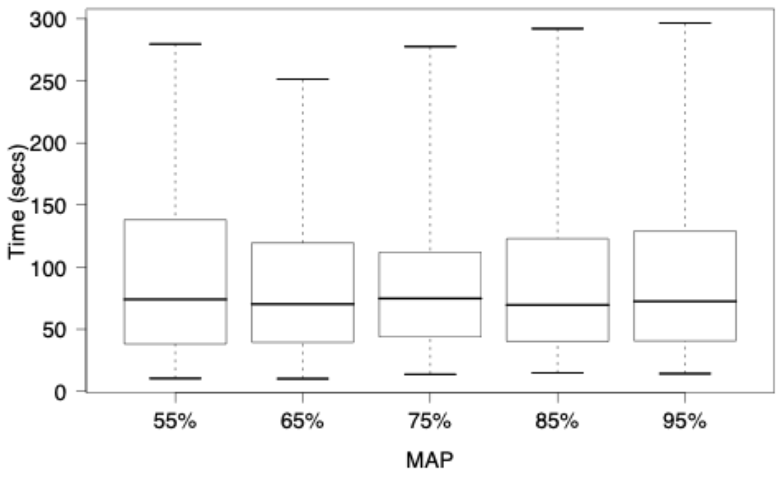


Figure 3: Time taken to find the first relevant document versus the mean average precision of the system used.

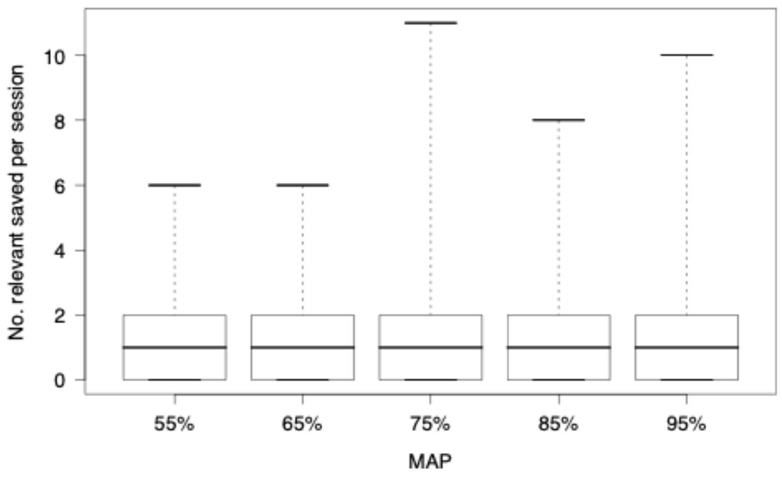


Figure 7: Number of relevant documents found by users within five minutes for systems with differing MAP.

Andrew Turpin and Falk Scholer. User performance versus precision measures for simple search tasks. In Proceedings of the 29th annual international acm sigir conference on research and development in information retrieval, SIGIR '06, 11--18, New York, NY, USA, 2006. , Association for Computing Machinery.

Ranking: normalized discounted cumulative gain

user model: in-order traversal of a ranked list, collecting all relevant items.

metric: accumulated position-discounted utility—proportional to rating—over traversal.

uses: web search.

$$\mathrm{DCG}(\mathcal{Y}^{+}, \pi) = \frac{1}{\mathcal{Z}} \sum_{i \in \mathcal{Y}^{+}} \frac{g(i)}{\log_2(\overline{\pi}_i + 1)}$$

$$\begin{array}{rl} g(i) & \text{rating of document } i \\ \mathcal{Z} & \text{DCG of ideal ranking} \end{array}$$

Kalervo Jarvelin and Jaana Kekalainen. Cumulated gain-based evaluation of ir techniques. TOIS, 20(4):422--446, 2002.

Ranking: normalized discounted cumulative gain

lab experiments

Users	nDCG		MRR	
Agree	160	65%	159	67%
Rnk eql	21	9%	21	9%
Disgree	66	27%	57	24%
	247		**237**	

Table 5: Comparison of Pearson Correlations / Concordance between Satisfaction and Offline Metrics (* indicates t-test statistical significance at $p < 0.01$ level)

	Pearson Correlation	Concordance
CG	0.354*	45.8%
DCG@3	0.356*	61.6%*
DCG@5	0.411*	65.7%*
DCG@10	0.421*	65.3%*
AP	0.396*	60.2%*

online experiments

Table 1: Correlation between IR metrics and interleaving experiments.

Inter'l Scoring	IR Metric	Correlation	p-value
Per impression	NDCG@5	0.882	0.048
	MAP@10	0.689	0.198
	P@5	0.662	0.223
Per query	NDCG@5	0.910	0.032
	MAP@10	0.776	0.122
	P@5	0.733	0.159

Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. Do user preferences and evaluation measures line up?. SIGIR. 2010.
Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. Meta-evaluation of online and offline web search evaluation metrics. SIGIR 2017.
Filip Radlinski and Nick Craswell. Comparing the sensitivity of information retrieval metrics. SIGIR 2010.

Why use just one metric?

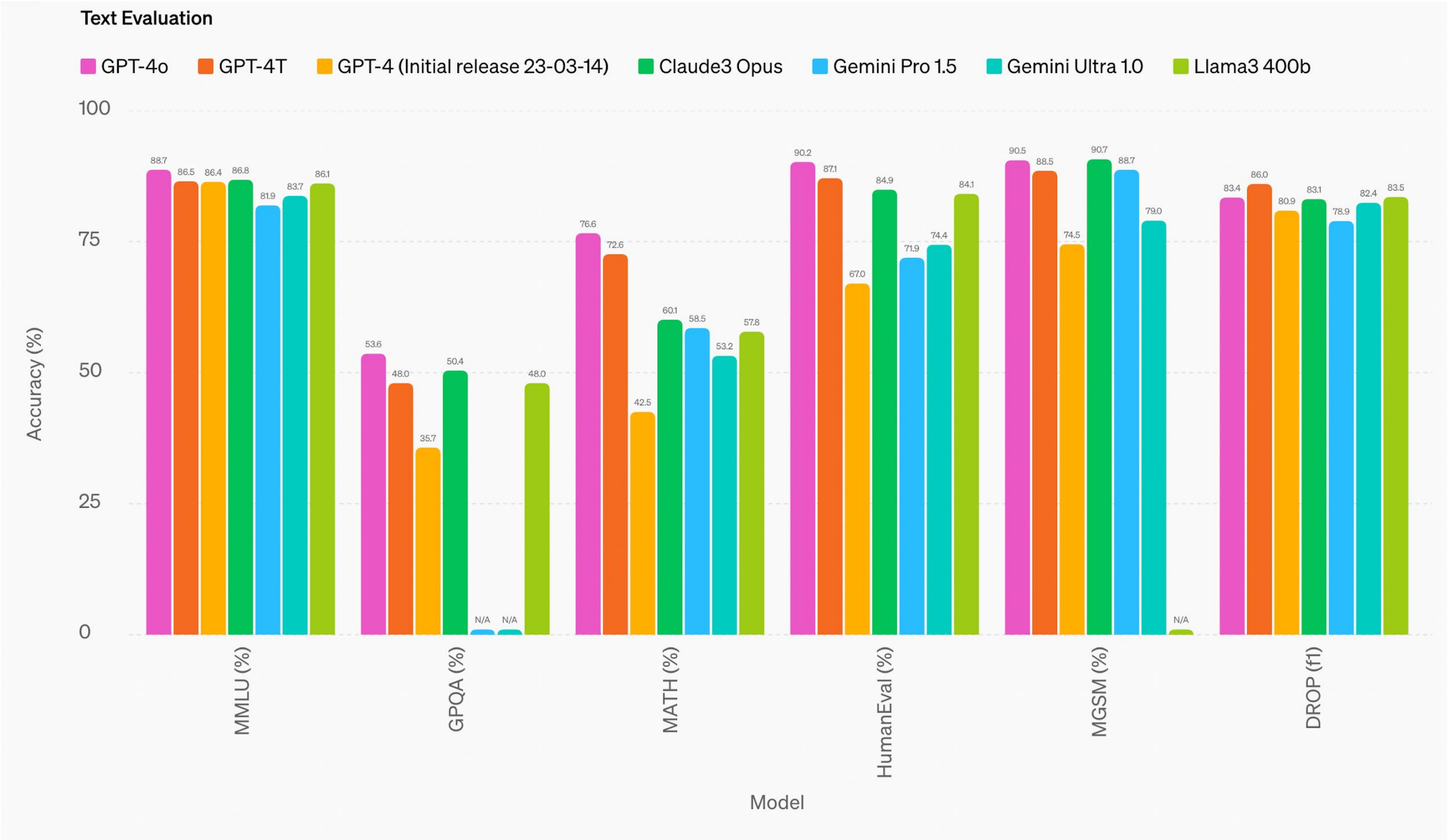
- Modern LLMs can support multiple tasks.
  - MT, summarization, search, dialog
- Even within a specific task, there are multiple subtasks
  - information-retrieval, text-generation
- For decades, production software systems has employed multidimensional scorecards of metrics
  - number of visitors, clicks, clickthrough rate, subscriptions, etc.
  - Increasingly, LLMs are doing the same.

Google’s Gemini release:

TEXT

Capability	Benchmark Higher is better	Description	Gemini Ultra	GPT-4 API numbers calculated where reported numbers were missing
General	MMLU	Representation of questions in 57 subjects (incl. STEM, humanities, and others)	90.0% CoT@32*	86.4% 5-shot** (reported)
Reasoning	Big-Bench Hard	Diverse set of challenging tasks requiring multi-step reasoning	83.6% 3-shot	83.1% 3-shot (API)
	DROP	Reading comprehension (F1 Score)	82.4 Variable shots	80.9 3-shot (reported)
	HellaSwag	Commonsense reasoning for everyday tasks	87.8% 10-shot*	95.3% 10-shot* (reported)
Math	GSM8K	Basic arithmetic manipulations (incl. Grade School math problems)	94.4% maj1@32	92.0% 5-shot CoT (reported)
	MATH	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	53.2% 4-shot	52.9% 4-shot (API)
Code	HumanEval	Python code generation	74.4% 0-shot (IT)*	67.0% 0-shot* (reported)
	Natural2Code	Python code generation. New held out dataset HumanEval-like, not leaked on the web	74.9% 0-shot	73.9% 0-shot (API)

OpenAI’s GPT-4o release:



Multi-task evaluation: GLUE

Corpus	**|Train|**	**|Test|**	**Task**	**Metrics**	**Domain**
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

Alex Wangeet al. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP workshop BlackboxNLP.

Multi-task evaluation: GLUE

		Single Sentence		Similarity and Paraphrase			Natural Language Inference			
Model	**Avg**	**CoLA**	**SST-2**	**MRPC**	**QQP**	**STS-B**	**MNLI**	**QNLI**	**RTE**	**WNLI**
Single-Task Training										
BiLSTM	63.9	15.7	85.9	69.3/79.4	81.7/61.4	66.0/62.8	70.3/70.8	75.7	52.8	**65.1**
+ELMo	66.4	**35.0**	90.2	69.0/80.8	85.7/65.6	64.0/60.2	72.9/73.4	71.7	50.1	**65.1**
+CoVe	64.0	14.5	88.5	73.4/81.4	83.3/59.4	67.2/64.1	64.5/64.8	75.4	53.5	**65.1**
+Attn	63.9	15.7	85.9	68.5/80.3	83.5/62.9	59.3/55.8	74.2/73.8	77.2	51.9	**65.1**
+Attn, ELMo	66.5	**35.0**	90.2	68.8/80.2	**86.5/66.1**	55.5/52.5	**76.9/76.7**	76.7	50.4	**65.1**
+Attn, CoVe	63.2	14.5	88.5	68.6/79.7	84.1/60.1	57.2/53.6	71.6/71.5	74.5	52.7	**65.1**
Multi-Task Training										
BiLSTM	64.2	11.6	82.8	74.3/81.8	84.2/62.5	70.3/67.8	65.4/66.1	74.6	57.4	**65.1**
+ELMo	67.7	32.1	89.3	**78.0/84.7**	82.6/61.1	67.2/67.9	70.3/67.8	75.5	57.4	**65.1**
+CoVe	62.9	18.5	81.9	71.5/78.7	84.9/60.6	64.4/62.7	65.4/65.7	70.8	52.7	**65.1**
+Attn	65.6	18.6	83.0	76.2/83.9	82.4/60.1	72.8/70.5	67.6/68.3	74.3	58.4	**65.1**
+Attn, ELMo	**70.0**	33.6	**90.4**	**78.0**/84.4	84.3/63.1	74.2/72.3	74.1/74.5	**79.8**	58.9	**65.1**
+Attn, CoVe	63.1	8.3	80.7	71.8/80.0	83.4/60.5	69.8/68.4	68.1/68.6	72.9	56.0	**65.1**
Pre-Trained Sentence Representation Models										
CBoW	58.9	0.0	80.0	73.4/81.5	79.1/51.4	61.2/58.7	56.0/56.4	72.1	54.1	**65.1**
Skip-Thought	61.3	0.0	81.8	71.7/80.8	82.2/56.4	71.8/69.7	62.9/62.8	72.9	53.1	**65.1**
InferSent	63.9	4.5	85.1	74.1/81.2	81.7/59.1	75.9/75.3	66.1/65.7	72.7	58.0	**65.1**
DisSent	62.0	4.9	83.7	74.1/81.7	82.6/59.5	66.1/64.8	58.7/59.1	73.9	56.4	**65.1**
GenSen	66.2	7.7	83.1	76.6/83.0	82.9/59.8	**79.3/79.2**	71.4/71.3	78.6	**59.2**	**65.1**

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: analyzing and interpreting neural networks for NLP, 353--355, Brussels, Belgium, November 2018. , Association for Computational Linguistics.

Multi-task evaluation: GLUE

Benchmarks such as GLUE have helped drive advances in NLP by incentivizing the creation of more accurate models. While this leaderboard paradigm has been remarkably successful, a historical focus on performance-based evaluation has been at the expense of other qualities that the NLP community values in models, such as compactness, fairness, and energy efficiency.

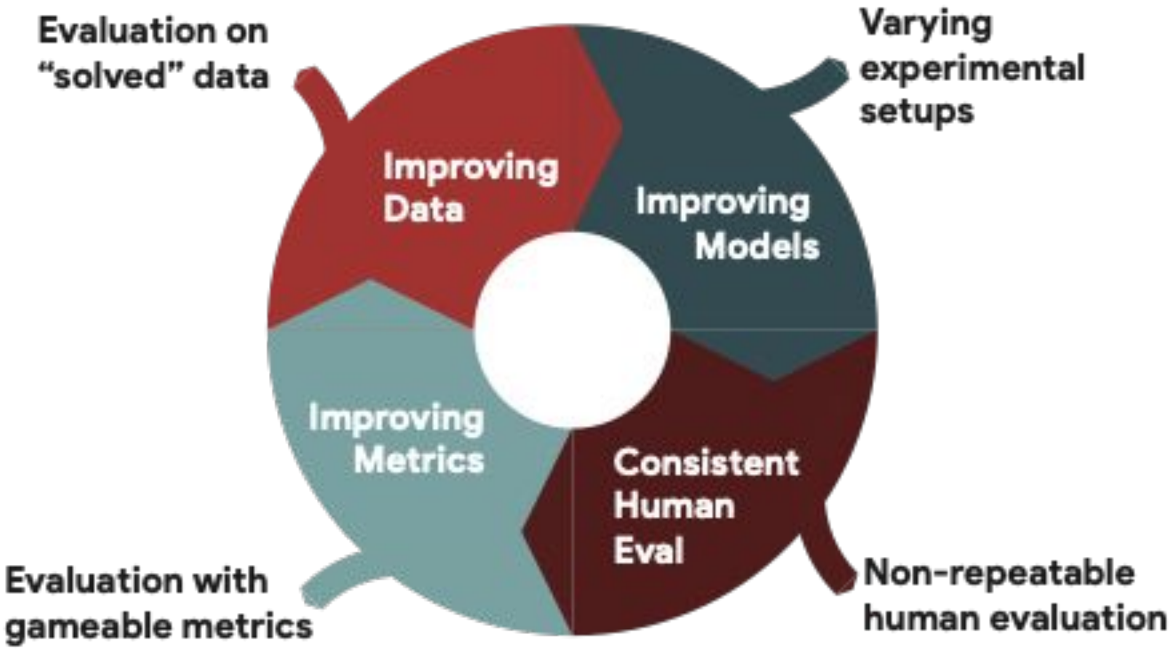
Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: a critique of NLP leaderboards. In EMNLP 2020.

Multitask evaluation: GEM

Dataset	Communicative Goal	Language(s)	Size	Input Type
CommonGEN (Lin et al., 2020)	Produce a likely sentence which mentions all of the source concepts.	en	67k	Concept Set
Czech Restaurant (Dušek and Jurčíček, 2019)	Produce a text expressing the given intent and covering the specified attributes.	cs	5k	Meaning Representation
DART (Radev et al., 2020)	Describe cells in a table, covering all information provided in triples.	en	82k	Triple Set
E2E clean (Novikova et al., 2017) (Dušek et al., 2019)	Describe a restaurant, given all and only the attributes specified on the input.	en	42k	Meaning Representation
MLSum (Scialom et al., 2020)	Summarize relevant points within a news article	*de/es	*520k	Articles
Schema-Guided Dialog (Rastogi et al., 2020)	Provide the surface realization for a virtual assistant	en	*165k	Dialog Act
ToTTo (Parikh et al., 2020)	Produce an English sentence that describes the highlighted cells in the context of the given table.	en	136k	Highlighted Table
XSum (Narayan et al., 2018)	Highlight relevant points in a news article	en	*25k	Articles
WebNLG (Gardent et al., 2017)	Produce a text that verbalises the input triples in a grammatical and natural way.	en/ru	50k	RDF triple
WikiAuto + Turk/ASSET (Jiang et al., 2020) (Xu et al., 2016) (Alva-Manchego et al., 2020)	Communicate the same information as the source sentence using simpler words and grammar.	en	594k	Sentence
WikiLingua (Ladhak et al., 2020)	Produce high quality summaries of an instructional article.	*ar/cs/de/en es/fr/hi/id/it ja/ko/nl/pt/ru th/tr/vi/zh	*550k	Article

Sebastian Gehrmann, et al.. The GEM benchmark: natural language generation, its evaluation and metrics. In Proceedings of the 1st workshop on natural language generation, evaluation, and metrics. ACL 2021.

Multitask evaluation: GEM



Sebastian Gehrmann, et al.. The GEM benchmark: natural language generation, its evaluation and metrics. In Proceedings of the 1st workshop on natural language generation, evaluation, and metrics. ACL 2021.

Multitask evaluation: BigBench

Main idea:

- “Quantity has a quality all its own”
- Anyone was allowed to contribute a task to the evaluation suite.

Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models

Alphabetic author list:*

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima (Shammie) Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, Ziyi Wu

Moving beyond automatic metrics

- Need to understand the precarity of automatic evaluation metrics
    - incompatibility
    - nonstationarity
    - dependence on engineering pipelines
    - variation across subtasks
    - social life of metrics
- Automatic metrics should be complemented with other traditions
    - qualitative evaluation
    - understanding of social context of technology

Summary

- Many, many ways to automatically evaluate performance, each with its own advantages and disadvantages.
    - "All models are wrong but some are useful."
- Important to understand how to interrogate metrics, compare them, and iterate on them.
- LLM community is movingaway from computing a single number to optimize toward
    - Ideally, evaluation should help us to develop a nuanced understanding of the new technology.

In class activity:

Suppose your team has used an LLM to build MovieBot, a chatbot which can give movie recommendations.

You would like to do automatic evaluation MovieBot's capabilities.

You have access to 1,000 ``test set'' conversations, in which a user conversed with a professional movie critic about what kinds of movies they liked, and the critic gave recommendations.

1. Describe an intrinsic automatic evaluation you could do for a component of MovieBot.
2. Describe an extrinsic **human** evaluation you could do of the entire MovieBot system?