Automatic Evaluation

Ehud Reiter

Automatic Evaluation

- Use an algorithm or model to evaluate a generated text
  - Sometimes generic
  - Sometimes based on quality criteria
- Sometimes called “metrics”
- Focus on basics here
  - New metrics constantly being introduced
- Huggingface good source for code
  - https://huggingface.co/docs/evaluate/en/index

Reference-based Evaluation

- Compare generated text to high-quality “reference” text
  - Example: edit distance between NLG text and reference text
- Similarity to reference text is used as a proxy for text quality
  - Older techniques are generic
  - Some newer techniques can be tuned for a specific quality criteria
- Need reference texts to do this!
  - Creating high-quality reference texts is expensive
  - Low-quality ref texts (crowdworkers, internet) not useful in assessing high-quality LLM output

Referenceless Evaluation

- Evaluate quality criteria without a reference text
  - Flesch-Kincaid grade level sort of assesses reasability
- In 2024 usually done with an LLM
- Readability and other “linguistic” quality criteria
  - Just need output text (*source-free*)
- Other criteria (accuracy, content, etc)
  - Provide input data to the metric

Contents

- Example techniques
- Validation
- Experimental design and what goes wrong

Simplest metric: Edit distance

- Reference-based metric
- How many words (characters) need to be changed in the generated text in order to match the reference text?

- *Transparent*: Can understand why text scored badly
  - Much harder with trained or LLM metric

Example: Weather

Reference text:

SSW 16-20 GRADUALLY BACKING SSE THEN BECOMING VARIABLE 10 OR LESS BY MIDNIGHT

Generated text:

SSW'LY 16-20 GRADUALLY BACKING SSE'LY THEN DECREASING VARIABLE 4-8 BY LATE EVENING

Differences:

SSW'~~LY~~ 16-20 GRADUALLY BACKING SSE'~~LY~~ THEN ~~DECREASING~~ *BECOMING* VARIABLE ~~4-8~~ *10 OR LESS* BY ~~LATE EVENING~~ *MIDNIGHT*.

Edit count:

- Two deletions of 'LY (one token deleted, twice)
- DECREASING changed to *BECOMING* (one token changed)
- 4-8 changed to *10 OR LESS* (three tokens changed)
- LATE EVENING changed to *MIDNIGHT* (two tokens changed)
- No tokens added

Token-level edit distance is *8* tokens deleted, changed, or modified

Character-level (Levenshtein) edit distance is 27

Many variations

- Word or character edit-distance
- Measure similarity at ngram level
  - Word level: BLEU, ROUGE
  - Character level: chrF
- Many enhancements proposed

- Character level seems better?
  - Studies show chrF better than BLEU/ROUGE
  - My student found character edit-distance very effective

Trained metric: BLEURT

- Train a model to predict the quality of a text
- BLEURT
  - Reference-based metric
  - Fine-tuned BERT to predict quality score
  - Heavy use of synthetic data to supplement genuine training data (human ratings of generated texts)
- Can be used “off the shelf”, or further fine-tuned to a domain and/or quality criteria
- Mostly (not always) better than edit-distance metrics

Other trained metrics

- *Many* trained metrics proposed
- COMET is one of the best trained metrics for evaluating MT
  - No explicit quality criteria, optimized for MT quality
  - Referenceless as well as reference-based versions
- BERTScore is older, probably less effective, still used

LLM Evaluator (LLM as Judge)

- Currently there is a lot of interest in using LLMs to evaluate texts
  - Ie, just ask GPT4 how good a text is, either in general or for a quality criteria
- LLMs should not be used to evaluate their own output
  - GPT4 is biased and “likes” its own output texts
- Otherwise seems to work well, but limitations still being explored

Example: GEMBA-MQM prompt

(System) You are an annotator for the quality of machine translation. Your task is to identify errors and assess the quality of the translation.
(user) {source_language} source:\n
‘‘‘{source_segment}‘‘‘\n
{target_language} translation:\n
‘‘‘{target_segment}‘‘‘\n
\n
Based on the source segment and machine translation surrounded with triple backticks, identify error types in the translation and classify them. The categories of errors are: accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation, register, spelling),
[*etc*]

Contents

- Example techniques
- *Validation*
- Experimental design and what goes wrong

Validation

- How well do metric results agree with real quality criteria?
- Accuracy: does metric agree with actual error counts?
  - Real measure: number of errors (perhaps weighted by severity)
  - If text 1 gets metric score of 0.8 and text 2 gets a metric score of 0.4, does text 2 actually have twice as many errors as text 1?
- Utility: does metric predict actual utility of texts
  - Real measure: time required by a person to post-edit a text
  - If text 1 gets metric score of 0.8 and text 2 gets a metric score of 0.4, will it take a human twice as long to post-edit text 2 compared to text 1?
- Etc

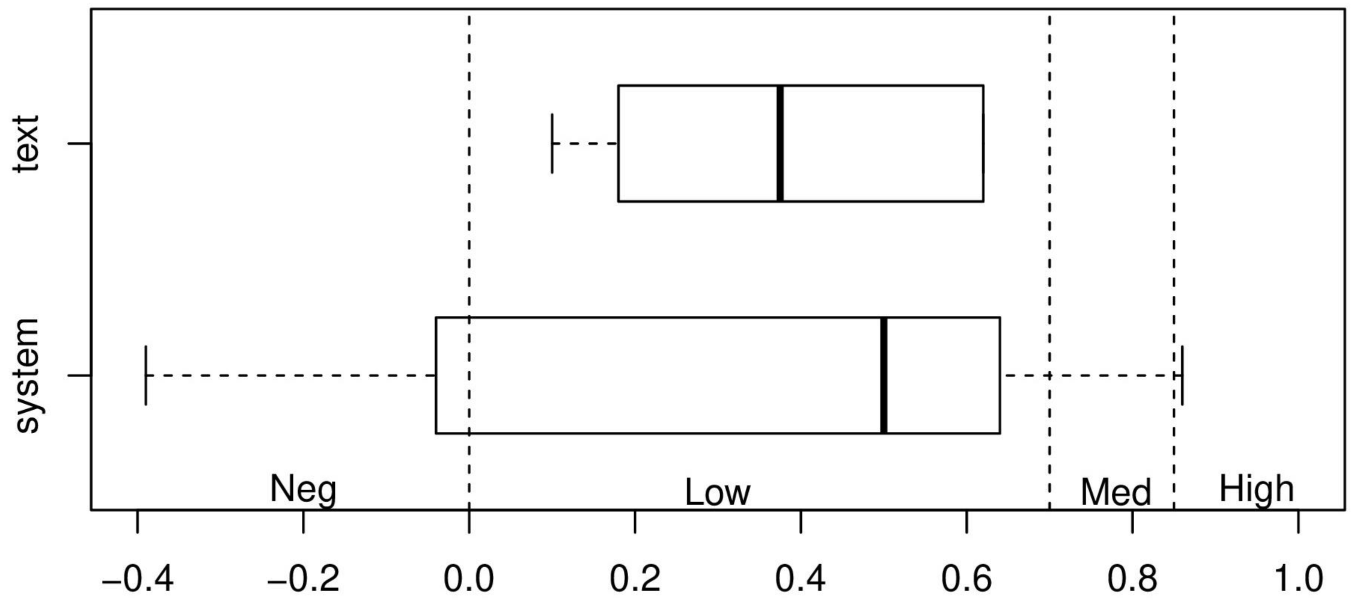
Validation

- *Validation* is process of measuring how well a metric agrees with actual measurements of quality criteria
  - Or perhaps with high-quality human assessments of criteria
- Usually done experimentally

Validation experiments

- Get N generated texts
  - Representative, good coverage
- Carefully measure actual criteria on these texts
  - Must be done well for experiment to be valid!
- Run metric on these texts
- Measure correlation between metric and actual measurement
  - Want to see correlation coefficient of at least 0.85
  - Usually a lot lower…

Correlations of BLEU to human evals in NLG



- Plot of reported correlations and BLEU and human evaluations in NLG, in papers published in ACL Anthology up to 2017
- https://doi.org/10.1162/coli_a_00322

Correlations with post-edit time

Criterion:	Post-edit times				
Reference:	human	edited	eval	avg	max
ROUGE-1-F1*	0.334	0.627	0.160	0.443	0.550
ROUGE-2-F1*	0.384	0.653	0.166	0.551	0.570
ROUGE-3-F1*	0.366	0.645	0.117	**0.576**	0.565
ROUGE-4-F1*	0.342	0.632	0.076	0.575	0.557
ROUGE-L-Pr*	0.348	0.471	0.169	0.366	0.427
ROUGE-L-Re*	0.409	0.614	**0.300**	0.520	0.551
ROUGE-L-F1*	0.384	0.646	0.285	0.538	0.564
CHRF*	0.341	0.460	-0.075	0.463	0.438
METEOR*	**0.415**	**0.667**	0.203	0.529	**0.581**
BLEU*	0.382	0.642	0.098	0.557	0.565
Levenshtein dist.	**0.547**	**0.780**	**0.453**	**0.600**	**0.654**
WER	0.239	0.629	0.059	0.326	0.550
MER	0.392	0.635	0.156	0.565	0.557
WIL	0.394	0.649	0.117	**0.590**	0.566
ROUGE-WE*	0.402	0.624	0.165	0.496	0.549
SkipThoughts*	0.298	0.403	-0.067	0.229	0.375
Embedding Avg*	0.266	0.375	-0.209	0.064	0.412
VectorExtrema*	0.409	0.553	0.127	0.424	0.500
GreedyMatching*	0.308	0.577	-0.041	0.295	0.520
USE*	0.339	0.522	0.201	0.366	0.476
WMD	0.354	0.594	0.154	0.421	0.529
BertScore*	**0.497**	**0.688**	**0.340**	0.571	**0.590**
MoverScore*	0.360	0.640	0.246	0.570	0.559
Stanza+Snomed*	0.334	0.508	0.118	0.354	0.460

Character edit distance (Levenshtein) has best correlation with human eval!

- 1960s tech beats 2020s tech…

Correlations between metric scores and time required by a human expert to post-edit a generated text https://aclanthology.org/2022.acl-long.394/

Weak correlations

- Many other papers have reported that even the “best” metrics are poor predictors of quality, especially when
  - Evaluating accuracy or utility (metrics do better at readability)
  - Texts are long and complicated (metrics do better with simple texts)
  - Evaluating quality of a single text (metrics do better at assessing the average quality of texts produced by a system)
- Caveat
  - Also get worse correlation when the “ground truth” quality assessments are sloppy, poorly done, noisy
  - These need to be done well!

Metrics must be valid

- Only use metrics when correlate well with “ground truth” assess
  - Don’t use unvalidated (or poorly validated) metrics!
- Note correlation depends on context!
  - Which quality criteria (readability vs utility)
  - Genre (sports story vs medical report)
  - Text quality (terrible vs near-human)
- Some metrics are well validated, others are not
  - Check the validation evidence, esp for unusual metrics!

Advice

- When using metrics, I personally try to also do a good human evaluation, even if on a small scale
- If human eval agrees with metrics, I have more confidence in metrics
- If human eval disagrees with metrics, I try to find out why

Contents

- Example techniques
- Validation
- *Experimental design and what goes wrong*

Experimental design for automatic eval

- *Research question*: Which quality criteria are we trying to estimate, which baseline are we comparing to
  - Depends on use case, research goals
- *Metric:* Which metric(s) will we use
  - Depends on validation evidence and practicalities
- *Material*: Which texts will we assess
  - Representative, good coverage
  - Unseen (avoid data contamination)

Experimental design

- *Experimental procedure*: Details of metrics
  - Implementation, parameters, fine tuning (needed for replication)
  - Supplementary human evaluation?
- *Analysis*: How do we analyse the data
  - Statistics – can use t-test, sometimes different tests recommended
    - See best practice for specific metrics
  - Qualitative error analysis – Good practice to qualitatively analyse texts with poor scores (are they bad, or did metric make mistake)
  - Correlation with human eval (if done)

What goes wrong

- See previous lecture
- Use metric which is not valid for key quality criteria
  - If it doesn’t measure what we care about, its not useful!
  - If we care about accuracy, don’t use readability metric…
  - Unfortunately I see a lot of this
- (Worse) use metric that is not validated
  - “I am convinced GPT4 does a great job on eval, so we use it”
  - Need evidence, not “gut feeling”
  - Especially when danger of conflict-of-interest
    - Above said by someone from OpenAI

What goes wrong

- Use synthetic test data
  - “We didn’t have much real test data, so we asked GPT4 to create more”
  - Very dangerous, don’t do this unless really know what you are doing
  - Training on synthetic data can be OK, testing on it is dubious
- Low quality reference texts and/or test data
  - Reference texts (if used) must be high-quality!
  - Test data must be accurate!
  - Quality check for errors, noise, etc
  - Big problem in many current test sets
  - “Academics terrible at creating test data because they ignore quality”??

What goes wrong

- Repeatedly rerun experiment until get desired result
  - Different random seeds
  - Tweak parameters on NLG system
  - Doing this invalidates results!
  - Test data should ideally only be used once
- Calculate dozens of metrics, just report "best" ones
  - Invalidates results
  - Best to just use a few metrics
  - Essential to report every result

Conclusion

- Automatic (metric) evaluation is very popular in NLG
- However results can be meaningless
  - Not good predictors of actual quality/utility
- Only use well-validated metrics
- Careful experimental design and reporting essential

- Proper evaluation requires **meaningful** numbers!

Discussion: Experience with metrics?