

Methods and metrics: Natural language generation metrics

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding



Challenges

1. There is more than one effective way to say most things.
2. What are we measuring?
 - ▶ Fluency?
 - ▶ Truthfulness?
 - ▶ Communicative effectiveness?



Perplexity of a probability distribution

Perplexity

For a sequence $\mathbf{x} = [x_1, \dots, x_n]$ and probability distribution p :

$$\mathbf{PP}(p, \mathbf{x}) = \prod_{i=1}^n \left(\frac{1}{p(x_i)} \right)^{\frac{1}{n}}$$

Token-level perplexity

$$\mathbf{token-PP}(p, \mathbf{x}) = \exp\left(\frac{\log \mathbf{PP}(p, \mathbf{x})}{n}\right)$$

Mean perplexity

For a corpus X of m examples:

$$\mathbf{mean-PP}(p, X) = \exp\left(\frac{1}{m} \sum_{\mathbf{x} \in X} \log \mathbf{token-PP}(p, \mathbf{x})\right)$$



Properties

- Bounds: $[1, \infty]$, with 1 best.
- Equivalent to the exponentiation of the cross-entropy loss.
- Value encoded: does the model assign high probability to the input sequence?
- Weaknesses:
 - ▶ Heavily dependent on the underlying vocabulary.
 - ▶ Doesn't allow comparisons between datasets.
 - ▶ Even comparisons between models are tricky.

Word-error rate

Edit distance

A measure of distance between strings. Word-error rate can be seen as a family of measures depending on the choice of distance measure.

Word-error rate

$$\mathbf{wer}(\mathbf{x}, \mathbf{pred}) = \frac{\mathbf{distance}(\mathbf{x}, \mathbf{pred})}{\mathbf{length}(\mathbf{x})}$$

Corpus word-error rate

For a corpus X :

$$\frac{\sum_{\mathbf{x} \in X} \mathbf{distance}(\mathbf{x}, \mathbf{pred})}{\sum_{\mathbf{x} \in X} \mathbf{length}(\mathbf{x})}$$

Properties

- Bounds: $[0, \infty]$, with 0 the best.
- Value encoded: how aligned is the predicted sequence with the actual sequence – similar to F scores.
- Weaknesses:
 - ▶ Just one reference text.
 - ▶ A very syntactic notion – consider *It was good* vs. *It was not good.* vs. *It was great*

BLEU scores

BLEU scores

Modified n-gram precision

Candidate: the the the the the the the

Ref 1: the cat is on the mat

Ref 2: there is a cat on the mat

Score: 2 / 7

BLEU scores

Modified n-gram precision

Candidate: the the the the the the the

Ref 1: the cat is on the mat

Ref 2: there is a cat on the mat

Score: 2 / 7

Brevity penalty

- r : sum of all minimal absolute length differences between candidates and referents.
- c : total length of all candidates
- BP: 1 if $c > r$ else $e^{1-\frac{r}{c}}$

BLEU scores

Modified n-gram precision

Candidate: the the the the the the the

Ref 1: the cat is on the mat

Ref 2: there is a cat on the mat

Score: 2 / 7

Brevity penalty

- r : sum of all minimal absolute length differences between candidates and referents.
- c : total length of all candidates
- BP: 1 if $c > r$ else $e^{1-\frac{r}{c}}$

BLEU

BP · the sum of weighted modified n -gram precision values for each n considered

Properties

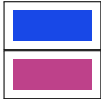







- Bounds: $[0, 1]$, with 1 the best, though with no expectation that any system will achieve 1.
- Value encoded:
 - ▶ Appropriate balance of (modified) precision and “recall” (BP).
 - ▶ Similar to word-error rate, but seeks to accommodate the fact that there are typically multiple suitable outputs for a given input.
- Weaknesses:
 - ▶ Callison-Burch et al. (2006) argue that BLEU fails to correlate with human scoring of translations.
 - ▶ Very sensitive to n-gram order.
 - ▶ Insensitive to n-gram types (*that dog* vs. *the dog* vs. *that toaster*).
 - ▶ Liu et al. (2016) specifically argue against BLEU as a metric for assessing dialogue systems.

Other n-gram-based metrics

Word-error rate	Edit-distance from a single reference text
BLEU	Modified precision and brevity penalty, against many reference texts
ROUGE	Recall-focused variant of BLEU, focused on assessing summarization systems
METEOR	Unigram-based alignments using exact match, stemming, synonyms
CIDEr	Weighted cosine similarity between TF-IDF vectors

Communication-based metrics

For NLU, it's worth asking whether you can evaluate your system based on how well it actually communicates in the context of a real-world goal.

	Context		Utterance
			The darker blue one
			dull pink not the super bright one
			not any of the regular greens

References I

- Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. [Re-evaluating the role of Bleu in machine translation research](#). In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. [How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.
- Benjamin Newman, Reuben Cohn-Gordon, and Christopher Potts. 2020. Communication-based evaluation for natural language generation. In *Proceedings of the Society for Computation in Linguistics*, pages 234–244, Washington, D.C. Linguistic Society of America.