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 \( \mathbf{x} = [x_1, \ldots, x_n] \) and probability distribution \( p \):

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

Token-level perplexity

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

Mean perplexity

For a corpus \( \mathcal{X} \) of \( m \) examples:

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

\[
wer(x, \text{pred}) = \frac{\text{distance}(x, \text{pred})}{\text{length}(x)}
\]

**Corpus word-error rate**
For a corpus \(X\):

\[
\frac{\sum_{x \in X} \text{distance}(x, \text{pred})}{\sum_{x \in X} \text{length}(x)}
\]
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*
<table>
<thead>
<tr>
<th>Challenges</th>
<th>Perplexity</th>
<th>Word-error rate</th>
<th>BLEU scores</th>
<th>Other n-gram-based metrics</th>
<th>Communication-based metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td><img src="score2_7" alt="Score: 2 / 7" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Brevity penalty**

- $r$: sum of all minimal absolute length differences between candidates and referents.
- $c$: total length of all candidates

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{else} \end{cases}$$

$$\text{BLEU} \cdot BP = \text{the sum of weighted modified n-gram precision values for each n considered}$$
BLEU scores

Modified n-gram precision

Candidate: the the 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 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 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**

$BP \cdot$ 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

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word-error rate</td>
<td>Edit-distance from a single reference text</td>
</tr>
<tr>
<td>BLEU</td>
<td>Modified precision and brevity penalty, against many reference texts</td>
</tr>
<tr>
<td>ROUGE</td>
<td>Recall-focused variant of BLEU, focused on assessing summarization systems</td>
</tr>
<tr>
<td>METEOR</td>
<td>Unigram-based alignments using exact match, stemming, synonyms</td>
</tr>
<tr>
<td>CIDEr</td>
<td>Weighted cosine similarity between TF-IDF vectors</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Context</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>The darker blue one</td>
<td>dull pink not the super bright one</td>
</tr>
<tr>
<td>not any of the regular greens</td>
<td></td>
</tr>
</tbody>
</table>

Newman et al. 2020
References

