BLEU Scores for Machine Translation

November 9, 2024

Introduction to BLEU

- **BLEU (Bilingual Evaluation Understudy)** is a metric for evaluating the quality of text that has been machine-translated from one language to another.
- Proposed by Papineni et al. (2002) to provide an automated and quantitative measure of translation quality.
- Compares n-grams of the candidate translation with n-grams of reference translations.

How BLEU Works

- **N-gram Precision**: Measures the proportion of n-grams in the candidate translation that match n-grams in the reference translations.
- **Brevity Penalty (BP)**: Penalizes translations that are shorter than the reference.
- **Geometric Mean**: BLEU score is the geometric mean of n-gram precisions, multiplied by the brevity penalty.

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

N-gram Precision

- **Unigram Precision** evaluates word choices.
- **Bigram, Trigram, and Higher-Order Precisions** evaluate fluency.
- Example: For the candidate sentence "The cat is on the mat" and reference "The cat is sitting on the mat":
  - Unigrams: "The", "cat", "is", "on", "the", "mat"
  - Bigrams: "The cat", "cat is", "is on", "on the", "the mat"

Brevity Penalty (BP)

- Ensures that shorter candidate translations are penalized.
- If the candidate length c is shorter than the reference length r:

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp(1 - \frac{r}{c}) & \text{if } c \leq r \end{cases} \tag{2}$$

- Encourages translations that are similar in length to reference translations.

Example Calculation of BLEU

- Consider a candidate translation and two reference translations.
- Calculate unigram, bigram, trigram, and 4-gram precisions.
- Apply brevity penalty and compute BLEU score.

Example Calculation of BLEU

- **Candidate Translation**: "The cat is on the mat"
- **Reference Translations**:
  1. "The cat is on the mat"
  2. "There is a cat on the mat"

Contd...

- **Step-by-Step Calculation**:
  - **Unigram Precision** (1-gram):

$$\text{Precision}_1 = \frac{6}{6} = 1.0 \tag{3}$$

(Matches: "The", "cat", "is", "on", "the", "mat")

  - **Bigram Precision** (2-gram):

$$\text{Precision}_2 = \frac{4}{5} = 0.8 \tag{4}$$

(Matches: "The cat", "is on", "on the", "the mat")

  - **Trigram Precision** (3-gram):

$$\text{Precision}_3 = \frac{2}{4} = 0.5 \tag{5}$$

(Matches: "The cat is", "on the mat")

  - **4-gram Precision** (4-gram):

$$\text{Precision}_4 = \frac{1}{3} = 0.33 \tag{6}$$

(Matches: "The cat is on")

Contd...

- **Brevity Penalty (BP)**:
  - Candidate length $c = 6$, Closest reference length $r = 6$
  - $$BP = \exp\left(1 - \frac{r}{c}\right) = \exp(0) = 1$$
- **Final BLEU Score**:

$$\text{BLEU} = BP \times \exp\left(\frac{1}{4}\sum_{n=1}^{4}\log(\text{Precision}_n)\right) \tag{7}$$

$$\text{BLEU} = 1 \times \exp\left(\frac{1}{4}\left(\log(1.0) + \log(0.8) + \log(0.5) + \log(0.33)\right)\right) \tag{8}$$

$$\text{BLEU} \approx 0.594 \tag{9}$$

Strengths and Limitations of BLEU

- **Strengths**:
  - Automated, reproducible, and language-agnostic.
  - Correlates well with human judgment at corpus level.
- **Limitations**:
  - Does not account for semantic meaning or context.
  - Sensitive to exact n-gram matches, which can be overly strict.
  - Less effective at the sentence level.

Applications in Machine Translation

- Widely used to evaluate machine translation models like **Google Translate**, **Bing Translator**, etc.
- Serves as a benchmark in many NLP competitions and research.
- Used to compare performance across different models and algorithms.

Conclusion

- BLEU score remains a popular choice for evaluating machine translations despite its limitations.
- Future improvements may focus on incorporating semantic understanding and context awareness.
- For more details, check the original paper: Papineni et al. (2002) BLEU: a Method for Automatic Evaluation of Machine Translation.

References

- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation*. ACL '02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.