COLX-531: Neural Machine Translation

Muhammad Abdul-Mageed

muhammad.mageed@ubc.ca

Deep Learning & NLP Lab

The University of British Columbia

Table of Contents

BLEU Paper

Proceedings of the 40th Annual Meeting of the Association for
Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

BLEU: a Method for Automatic Evaluation of Machine Translation

Kishore Papineni, Salim Roukos, Todd Ward, and **Wei-Jing Zhu**
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{papineni,roukos,toddward,weijing}@us.ibm.com

Abstract

Human evaluations of machine translation are extensive but expensive. Human evaluations can take months to finish and involve human labor that can not be reused. We propose a method of automatic machine translation evaluation that is quick, inexpensive, and language-independent,

the evaluation bottleneck. Developers would benefit from an inexpensive automatic evaluation that is quick, language-independent, and correlates highly with human evaluation. We propose such an evaluation method in this paper.

1.2 Viewpoint

How does one measure translation performance? *The closer a machine translation is to a professional*

MT Eval Rationale (Papineni et al., 2002)

Why MT Eval Needed?

- Evaluating translation is hard. Why?
- Human evaluation is costly
- An automatic method promises to accelerate MT progress
- Measure MT based on **numerical closeness** to a human reference
- The metric can be used to **optimize** MT systems
- **BLEU Main Idea:** Use a weighted avg of variable-length phrase matches against reference translations
- i.e., compare **n-grams of the candidate** with the **n-grams of the reference** translation, and count the **number of matches**
- Matches are ***position-independent***.
- The more **the matches,** the better **the candidate translation** is.

Uni-gram Precision

Precision

- **Precision** counts up the number of candidate translation words (unigrams) which occur in any reference translation and then divides by the total number of words in the candidate translation.

1: Precision

$$\frac{\color{red}{count}\_candidate\_trans\_words\_in\_ref}{total\_words\_in\_candidate\_trans}$$

Candidate: the the the the the the the.

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

Modified Uni-gram Precision

Modified Precision

- A ref word should be considered *exhausted* after a matching candidate word is identified
- **Modified unigram precision:**
  1. Count the **maximum** number of times a word occurs **in any single reference translation**.
  2. **Clip** the **total count** of each candidate word **by its maximum reference** count. $count_{\text{clip}} = \mathbf{min}(count, max\_ref\_count)$
  3. **Add clipped counts up**, and divide by the **total (unclipped) number of candidate** words

2: Modified Precision

$$\frac{count_{\text{clip}}}{total\_words\_in\_candidate\_trans}$$

3: Modified Precision

$$count_{\text{clip}} = \mathbf{min}(count, max\_ref\_count)$$

$$\frac{count_{\text{clip}}}{total\_words\_in\_candidate\_trans}$$

Candidate: `the the the the the the the.`

Reference 1: `The cat is on the mat.`

Reference 2: `There is a cat on the mat.`

Modified Bi-gram Precision

Exercise

Calculate modified bi-gram precision for Candidate 1.

Candidate 1: `It is a guide to action which ensures that the military always obeys the commands of the party.`

Candidate 2: `It is to insure the troops forever hearing the activity guidebook that party direct.`

Reference 1: `It is a guide to action that ensures that the military will forever heed Party commands.`

Reference 2: `It is the guiding principle which guarantees the military forces always being under the command of the Party.`

Reference 3: `It is the practical guide for the army always to heed the directions of the party.`

Candidate Bi-grams $\text{Count}_{\text{Clip}}$

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

1 It is   2 is a
3 a guide   4 guide to
5 to action   6 action which
7 which ensures   8 ensures that
9 that the   10 the military
11 military always   12 always obeys
13 obeys the   14 the commands
15 commands of   16 of the
17 the party

Figure: There are 17 bi-grams in Candidate 1. ($\text{Count}_{\text{clip}}$)

Calculating Modified Bi-gram Precision

1 2 3 4 5

6 7 8
9 10

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Shared with Ref 1

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

shared with Ref 2

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Figure: **Modified precision= 10/17**. Count$_{\text{clip}}$=10. Total bi-gram count in Candidate 1 = 17. No need to consider bi-grams in Ref 2 and Ref 3 that already occur in Ref 1.

Modified N-Gram Precision on **Blocks of Text**

Blocks of Text

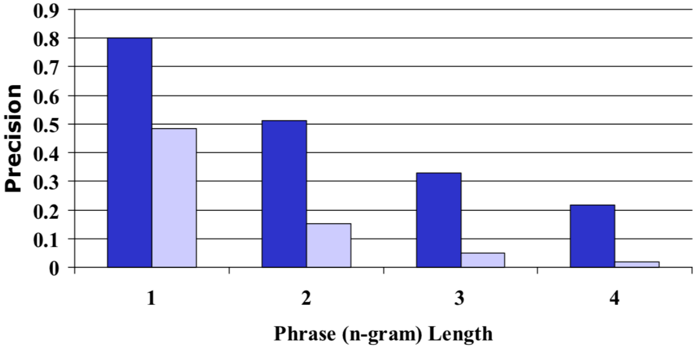
1. Compute the n-gram matches sentence by sentence.
2. Add the clipped n-gram counts for all the candidate sentences
3. Divide by the number of candidate n-grams in the test corpus

4: P_n

$$P_n = \frac{\sum_{C \in Candidates} \sum_{n-gram \in C} count_{\text{clip}}(n-gram)}{\sum_{C' \in Candidates} \sum_{n-gram' \in C'} Count(n-gram')}$$

BLEU on Human vs. Machine Translation

Figure 1: Distinguishing Human from Machine



Combining the modified n-gram Precisions

How should we combine the modified precisions for various n-gram sizes?

- **Modified n-gram precision decays roughly exponentially with n**: the modified un-igram precision is much larger than the modified bi-gram precision which in turn is much bigger than the modified tri-gram precision.
- A **weighted average of the logarithm** of modified precisions take this exponential decay into account.
- BLEU **uses the average logarithm with uniform weights**, which is equivalent to using the geometric mean of the modified n-gram precisions.

Sentence Length

Ensuring Suitable Length

- **N-gram precision penalizes spurious words** in candidate that do not appear in any ref
- **Modified precision is penalized if a word occurs more frequently in a candidate** than its max ref count
- **However**, modified n-gram precision *alone* fails to enforce the proper translation length
- See paper for illustrating examples

Sentence Brevity

Brevity

- Goal: **Make the brevity penalty 1.0** when the **candidate's length is the same as any reference** translation's **length**
- Example: If there are 3 refs with lengths 12, 15, and 17 words and the candidate is 12 words, **we want the brevity penalty to be 1** and call the closest ref sent length the "best match length."
- If we compute brevity penalty sentence by sentence and averaged the penalties, then length deviations on short sentences would be punished harshly.
- Instead, **compute the brevity penalty over the entire corpus** to allow some freedom at the sentence level.
- **First**, compute the test corpus' effective reference length, r, by summing "best match length." for each candidate sent in corpus.
- **Then**, choose penalty to be a decaying exponential in r/c, where c is total length of the candidate corpus.

BLEU Details

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

Then,

$$\textsc{Bleu} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

The ranking behavior is more immediately apparent in the log domain,

$$\log \textsc{Bleu} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n.$$

In our baseline, we use $N = 4$ and uniform weights $w_n = 1/N$.