Bilingual Evaluation Understudy (BLEU)

Evaluating Machine Translation

Ramaseshan Ramachandran

1

2

3

DIFFICULTIES WITH HUMAN EVALUATION

- Human evaluations are extensive but expensive
- A need for quick, reusable, inexpensive method that correlates highly with human evaluation
- Many aspects of translation,including adequacy and fluency should be considered during the automatic evaluation
- Automatic evaluation is a boon to developers of LLM
- Two important aspects required for automatic evaluation
  1. A good metric
  2. A good/gold standards as references

THE IDEA

- Many possible sentences for a given sentence
- A good evaluator identifies a good candidate using adequacy and fluency

The main idea is to use a weighted average of variable length phrase matches against the reference translations[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: **It is a guide to action that ensures that the military will for ever heed Party commands**

If many words and phrases are shared between the candidate and the reference translations, then it is a good candidate

Can n-grams help in matching the words and phrases?

UNIGRAM PRECISION

C1: It is a guide to action which ensures that the military always obeys the commands of the party

R1: It is a guide to action that ensures that the military will forever heed Party commands

R2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

R3: It is the practical guide for the army to heed the directions of the party.

$$\text{Unigram precision} = \frac{17}{18}$$

C2: It is to insure the troops forever hearing the activity guidebook that party direct

R1: It is a guide to action that ensures that the military will forever heed Party commands

R2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

R3: It is the practical guide for the army always to heed the directions of the party.

$$\text{Unigram precision} = \frac{8}{14}$$

BLEU EVALUATION

```
from nltk.translate.bleu_score import sentence_bleu
references = [
['It', 'is', 'a', 'guide', 'to', 'action', 'that',
  'ensures', 'that', 'the', 'military', 'will',
  'forever', 'heed', 'Party', 'commands'],
['It', 'is', 'the', 'guiding', 'principle',
 'which', 'guarantees', 'the', 'military',
  'forces', 'always', 'being', 'under',
  'the', 'command', 'of', 'the', 'Party.'],
['It', 'is', 'the', 'practical', 'guide', 'for',
 'the', 'army', 'to', 'heed', 'the',
  'directions', 'of', 'the', 'party.']
]
hypothesis = ['It', 'is', 'a', 'guide', 'to',
 'action', 'which', 'ensures', 'that', 'the',
 'military', 'always', 'obeys', 'the',
 'commands', 'of', 'the', 'party']
score = sentence_bleu(references, hypothesis, weights=(1, 0, 0, 0))
print(score)
```

MODIFIED- N-GRAM PRECISION

Compare the number of n-grams in the candidate and in the reference translation
Penalize models that produces many words of the same type

- Count the number of times a word occurs in any single reference translation
- $\text{Count}_{\text{clip}} = \min(\text{Candidate Count}, \text{Maximum Reference Count})$

Refer the previous slide for the examples

$$\text{Modified unigram precision for C1} = \frac{17}{18} \qquad \text{Modified unigram precision for C2} = \frac{8}{14}$$

C3: the the the the the the the
R4: the cat is on the mat

$$\text{Unigram precision} = \frac{7}{7}$$
$$\text{Modified unigram precision} = \frac{2}{7}$$
$$\text{Modified bigram precision} = 0$$

Modified Unigram precision defines the adequacy of the translation, while modified bigram precision matches the fluency of the translation

MODIFIED BIGRAM PRECISION

(It,is),(is,a),(a,guide),
(guide,to),(to,action),
(action,which),(which,ensures),
(ensures,that),(that,the),
(the,military),(military,always),
(always,obeys),(obeys,the),
(the,commands),(commands,of),
(of,the),(the,party)

Modified bigram precision for C1 $= \frac{10}{17}$

```
(It,is),(is,a),(a,guide),(guide,to),
(to,action),(action,that),(that,ensures),
(ensures,that),(that,the),(the,military),
(military,will),(will,forever),(forever,heed),
(heed,Party),(Party,commands)

(It,is),(is,the),(the,guiding),
(guiding,principle),(principle,which),
(which,guarantees),(guarantees,the),
(the,military),(military,forces),(forces,always),
(always,being),(being,under),(under,the),
(the,command),(command,of),
(of,the),(the,Party)

(It,is),(is,the),(the,practical),(practical,guide),
(guide,for),(for,the),(the,army),
(army,always),(always,to),(to,heed),
(heed,the),(the,directions),
(directions,of),(of,the),(the,party)
^^I^^I
```

MODIFIED BIGRAM PRECISION - CANDIDATE 2

(It,is),(is,to),(to,insure),
(insure,the),(the,troops),
(troops,forever),(forever,hearing),
(hearing,the),(the,activity),
(activity,guidebook),
(guidebook,that),(that,party),
(party,direct)

Modified bigram precision for C2 $= \frac{1}{13}$

```
(It,is),(is,a),(a,guide),(guide,to),
(to,action),(action,that),(that,ensures),
(ensures,that),(that,the),(the,military),
(military,will),(will,forever),(forever,heed),
(heed,Party),(Party,commands)

(It,is),(is,the),(the,guiding),
(guiding,principle),(principle,which),
(which,guarantees),(guarantees,the),
(the,military),(military,forces),(forces,always),
(always,being),(being,under),(under,the),
(the,command),(command,of),
(of,the),(the,Party)

(It,is),(is,the),(the,practical),(practical,guide),
(guide,for),(for,the),(the,army),
(army,always),(always,to),(to,heed),
(heed,the),(the,directions),
(directions,of),(of,the),(the,party)
^^I^^I
```

COMBINING N-GRAM PRECISIONS

- Modified n-gram precisions decay exponentially as n increases[1]
- BLEU averages the log-precisions with uniform weights (the geometric mean of the modified n-gram precisions) to handle this decay
- A short candidate ($c < r$) can inflate precision, so a brevity penalty (BP) is applied when $c \leq r$

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

where r is the effective length of the reference corpus and c is the length of the candidate sentence

BLEU score is obtained by

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

where N is the n-gram size (BLEU uses 4-gram by default),w_n is the weights associated with unigram,bigram,trigram and 4-grams,and p_n is the modified precision score of the test corpus. Here,$\sum_{n=1}^{N} w_n = 1$. One option for $w_n = \frac{1}{N}$

$$p_n = \frac{\sum_{c \in C} \sum_{ngrams \in C} \text{Count}_{clip}(ngrams)}{\sum_{c \in C} \sum_{ngrams \in C} \text{Count}(ngrams)} \tag{2}$$

BLEU - DEMO

BLEU Demo

APPLICATIONS OF BLEU

BLEU is designed as a corpus measure

- Machine translation
- Image labeling
- Text summarization
- Speech recognition

OTHER METRICS

- NIST - National Institute of Standards and Technology - based on BLEU
- METEOR - Metric for Evaluation of Translation with Explicit ORdering
  - Uses stemming and synonymy matching
- WER - Word Error Rate
  - Uses edit distance (Levenshtein distance)
  - Finds minimum number of edit operations such as insertion,deletions or substitutions,needed to change the candidate sentence into the reference sentence
- GLEU - Google BLEU
  - Correlates well with BLEU, and works at the sentence level
- Embedding / model-based metrics (increasingly preferred for fluent LLM output)
  - BERTScore measures token similarity using contextual embeddings
  - BLEURT and COMET are learned metrics trained to match human judgments
  - LLM-as-a-judge prompts a strong LLM to rate or compare outputs

REFERENCES I

[1] Kishore Papineni et al. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://www.aclweb.org/anthology/P02-1040.