

# BLEU – Its Motivation

- Central Idea:
  - “The closer a machine translation is to a professional human translation, the better it is.”
- Implication
  - A evaluation metric could be evaluated
    - If it correlates with human evaluation, it would be a useful metric
- BLEU was proposed
  - as an *aid*
  - as a *quick substitute* of humans when *needed*

# What is BLEU? A Big Picture

- Requires multiple good reference translations
- Depends on modified n-gram precision (or co-occurrence)
  - Co-occurrence: if translated sentence hit n-gram in any reference sentences
- Computes Per-corpus n-gram co-occurrence
  - n can have several values and a weighted sum is computed
- Penalizes very brief translation

# N-gram Precision: an Example

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

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

**Clearly Candidate 1 is better**

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 directions of the party*

# N-gram Precision

- To rank Candidate 1 higher than 2
  - Just count the number of N-gram matches
  - The match could be position-independent
  - Reference could be matched multiple times
  - No need to be linguistically-motivated

# BLEU – Example : Unigram Precision

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

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 directions of the party.*

N-gram Precision : 17

## Example : Unigram Precision (cont.)

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 directions of the party.*

N-gram Precision : 8

# Issue of N-gram Precision

- What if some words are over-generated?
  - e.g. “the”
  - An extreme example

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.*

- N-gram Precision: 7 (Something wrong)
- **Intuitively : reference word should be exhausted after it is matched.**

# Modified N-gram Precision :

## Procedure

- Procedure
  - Count the max number of times a word occurs in any single reference
  - Clip the total count of each candidate word
  - Modified N-gram Precision equal to
    - $\text{Clipped count} / \text{Total no. of candidate word}$

### ➤ Example:

Ref 1: *The cat is on the mat.*

Ref 2: *There is a cat on the mat.*

“the” has max count 2

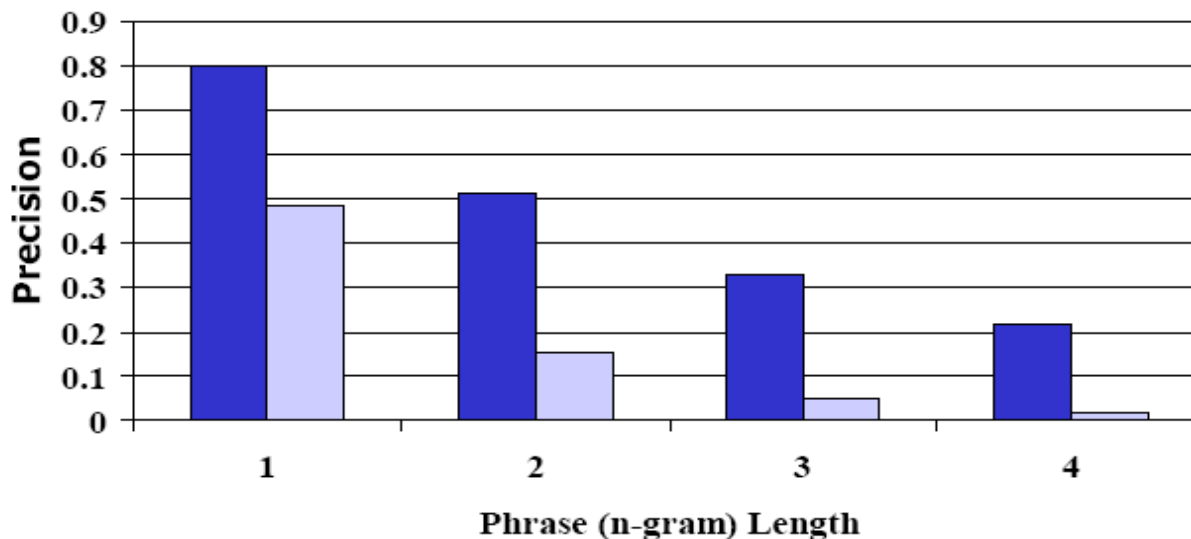
- Unigram count = 7  
Clipped unigram count = 2  
Total no. of counts = 7
- Modified-ngram precision:
  - Clipped count = 2
  - Total no. of counts = 7
  - Modified-ngram precision =  $2/7$

# Different N in Modified N-gram Precision

- $N > 1$  is computed in a similar way
  - When 1-gram precision is high, the reference tends to satisfy *adequacy*
  - When longer n-gram precision is high, the reference tends to account for *fluency*

# Experiment 1 of N-gram Precision: Can it differentiate good and bad translation?

Figure 1: Distinguishing Human from Machine



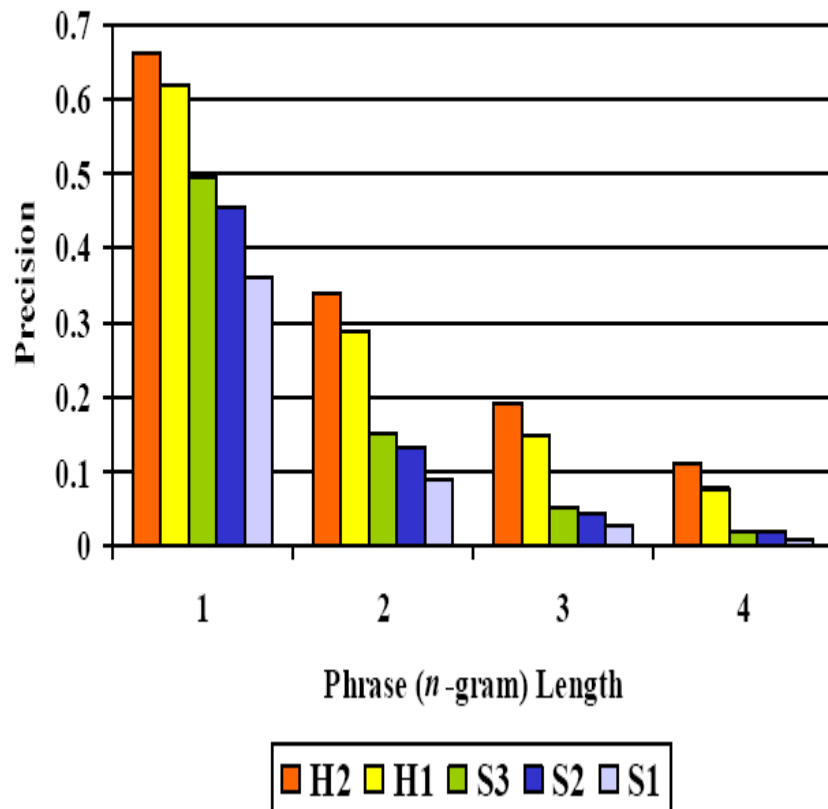
- Source : Chinese, Target: English
- Human (Blue) vs (Machine) Light Blue

Observation: Human scores much better than Machine

Conclusion: BLEU is useful for translation with great difference in quality.

# Experiment 2 of N-gram Precision: Can it differentiate with very close quality?

Figure 2: Machine and Human Translations



- From BLEU:  $H2 > H1 > S3 > S2 > S1$
- Same as human judgment
- Not shown in paper
- Conclusion: It is still quite useful when quality is similar

# Combining modified n-gram precision

- The measure becomes more robust
- Precision has exponential decay
  - => Geometric mean is used
  - => sensitive to higher n-gram
- 4-gram was shown to be the best among (3,4,5)-gram
- Arithmetic means was also tried
  - Underweighting of unigram found to be a good match with human.

# Issues of Modified N-gram Precision : Sentence Length

Candidate 3: *of the*

Modified Unigram Precision : 2/2

Modified Bigram Precision : 1/1

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 directions *of the* party.*

# Issues of Modified N-gram Precision : Trouble with Recalls

- Good candidate should only use (recall) one possible word choices
- Example:
  - Candidate 1: *I always invariably perpetually do.* (Bad Translation)
  - Candidate 2: *I always do.* (A complete Match)
  - Reference 1: *I always do.*
  - Reference 2: *I invariably do.*
  - Reference 3: *I perpetually do.*

# Solution: Brevity Penalty

- When a translation matches a reference
  - $BP = 1$
- When a translation is shorter than the reference
  - $BP < 1$

# Brevity Penalty Computation

- IBM's BP –corpus-based
  - best match lengths
    - The closest reference sentence length
      - E.g. If references have 12, 15, 17 words and candidate has 12
  - Exponential decay in  $r/c$  if  $c < r$ 
    - $r$  is the sum of the best match lengths of the candidate sentence in the test corpus
    - $c$  is the total length of the candidate translation corpus (?)
      - (?) is  $c$  the candidate sentence?
- (?) BP shouldn't be computed by averaging sentence penalties in sentence-by-sentence basis
  - => That will punish length deviation of short sentence very harshly.

# Formulae of BLEU Computation

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

Then,

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

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

# Experimental Evidence of BLEU

- 500 sentences (40 general news stories)
- 4 references for each sentence

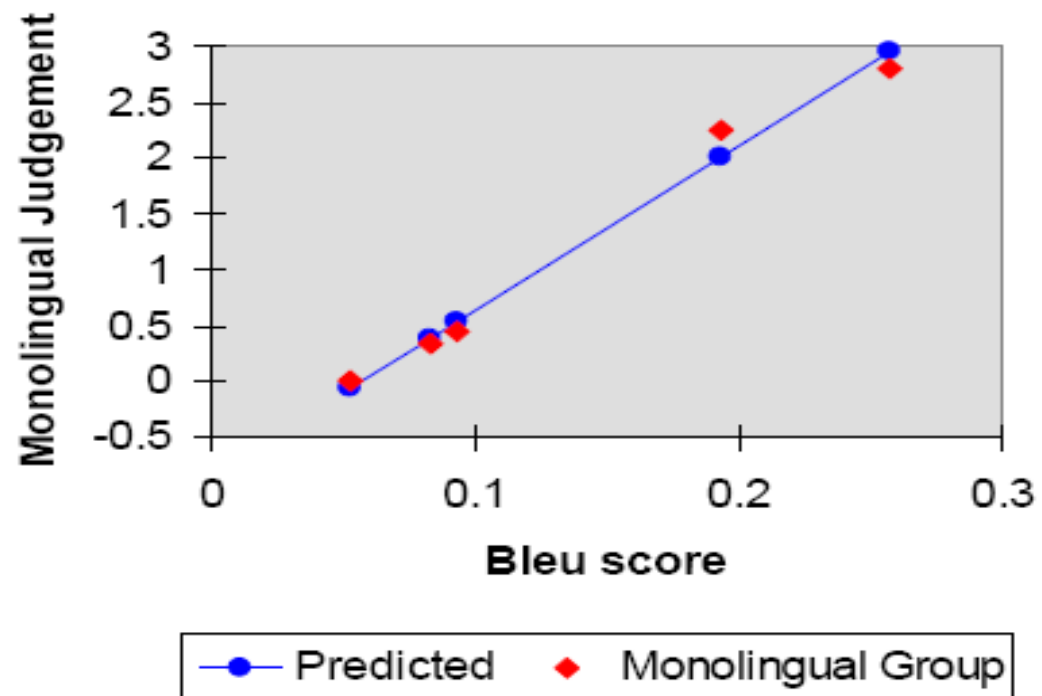
Table 1: BLEU on 500 sentences

S1	S2	S3	H1	H2
0.0527	0.0829	0.0930	0.1934	0.2571

# Human vs. BLEU

- BLEU shows high correlation with both monolingual (0.99) and bilingual group (0.96)

Figure 5: BLEU predicts Monolingual Judgments



# Human vs. BLEU (cont.)

Figure 7: BLEU vs Bilingual and Monolingual Judgments

