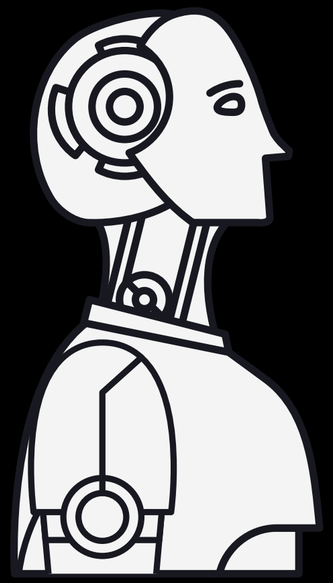
Artificial Intelligence 101

BLEU: a Method for Automatic Evaluation of Machine Translation

Watt Wizards



Machine Translation (MT) is the automated process of converting text or speech from one language to another.

Types of MT:-

- Rule-Based : Based on linguistic rules
- Statistical (SMT): Uses probability and data patterns.
- Neural (NMT): Employs deep learning for more natural translations.

Why evaluate machine translation?

MT systems need constant evaluation for meaningful translations to happen

Manual evaluation is slow and expensive and MT makes it more easier with automated evaluation .

Machine Translation quality was traditionally assessed through subjective human judgment.

MT systems were frequently updated, requiring rapid evaluation methods over human judgment.

Different evaluators may disagree on quality.

The BLEU metric reduces human
labour in evaluating
machine translation system.

The human evaluations
were time-consuming and costly.

BLEU offers an automated, objective,
and reproducible method for the evaluation

Problems with Human Evaluation

Extensive but time-consuming (weeks/months).

Involves non-reusable human labor.

Cannot be scaled for frequent translations.

Proposed Solution

BLEU (Bilingual Evaluation Understudy): A quick, inexpensive, and language-independent evaluation method.

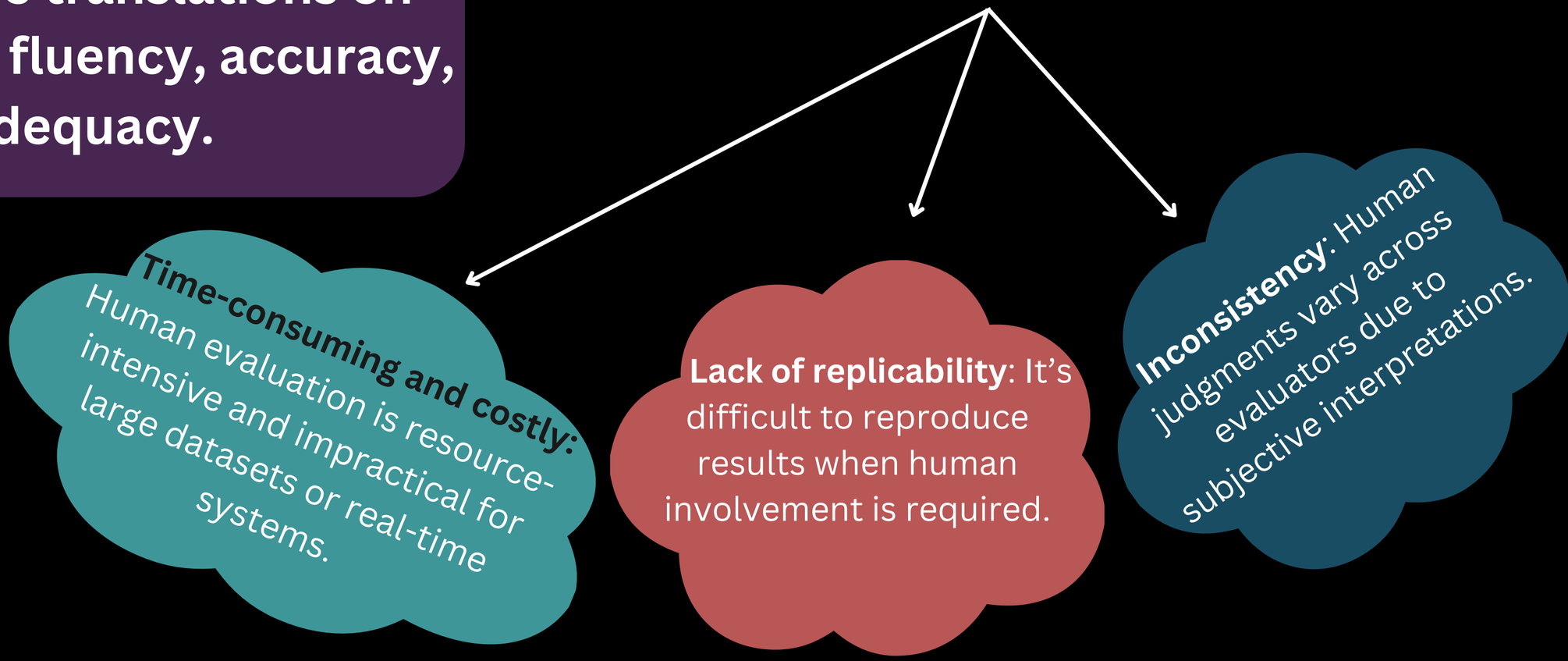
- High correlation with human evaluation
- Low marginal cost per run
- Suitable for frequent, rapid evaluations

Related Approaches and Why they have failed:-

1.Human Judgments:-

Experts or native speakers would manually score translations on dimensions like fluency, accuracy, and adequacy.

Why was it not adequate?



2.WER(Word Error Rate):-

Counts the number of insertions, deletions, and substitutions needed to transform a candidate translation into a reference translation.

WER = (Substitutions + Insertions + Deletions) / Number of Words Spoken

Formula for Word Error Rate (WER)

Well It’s Simple to compute and offers clear measure of how close two texts are.

Why This!!

Why Failed Then??

WER harshly penalizes linguistic diversity, ignoring valid translation variations. Additionally, it lacks fluency assessment, relying only on word-by-word matching to calculate substitutions, deletions, or insertions needed to align output with the reference.

3.Precision and Recall in Translation Evaluation

What They Measure?

False Positives

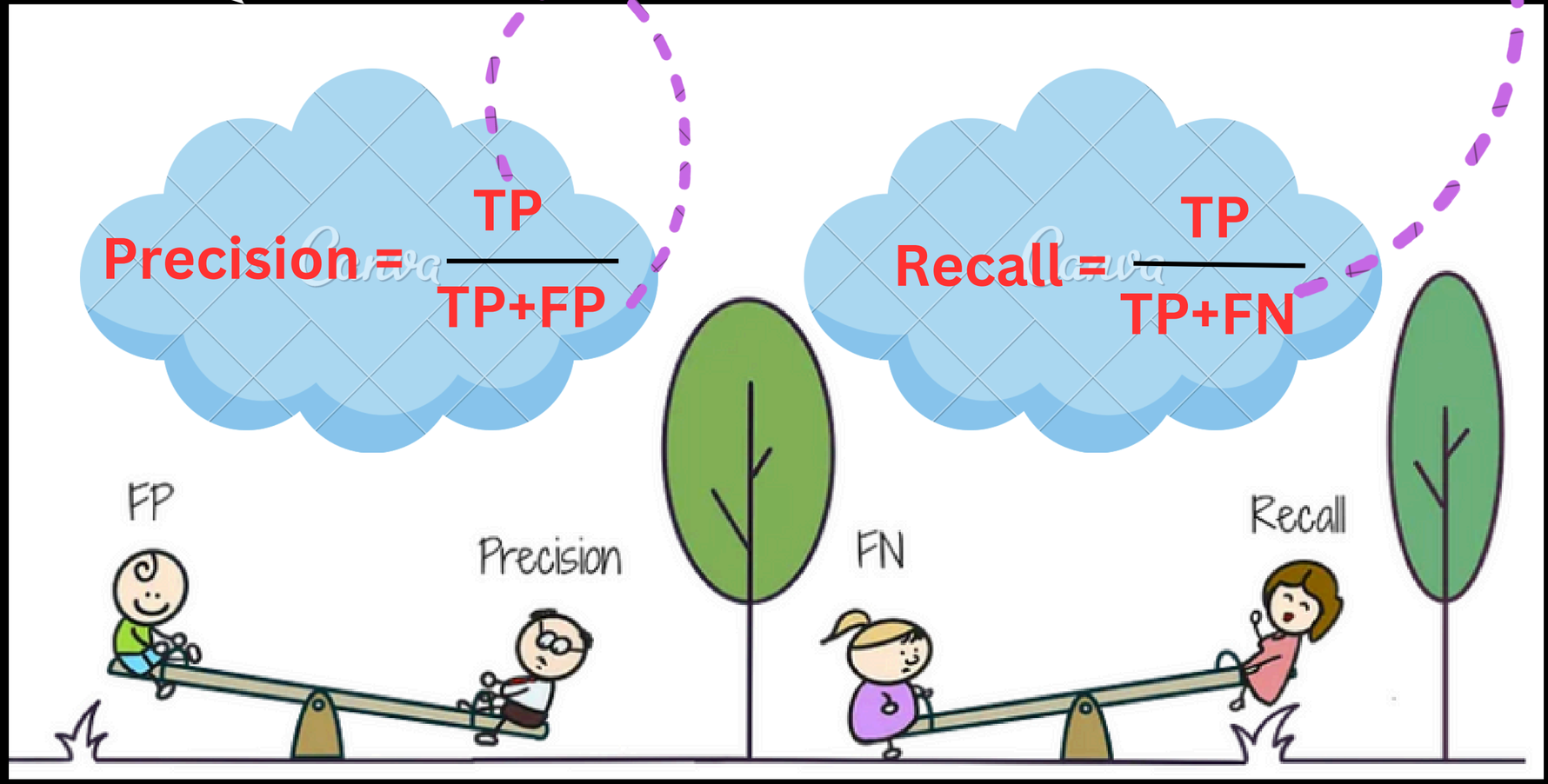
True Positives

False Negatives

overlap between candidate translations (hypotheses) and reference translations.

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$



Failures

Precision-Related Issues:
Encourages brevity, omitting key details
ex. : High precision may result from using only a few relevant words.

Recall-Related Issues:
Rewards verbosity, adding irrelevant content.
ex. : Extra words inflate recall.

Neglect of Fluency:
Ignores grammar and naturalness, focusing on word overlap.

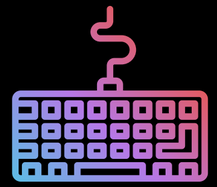
Directions Explored

Sentence-Level Evaluation:
Incorporates grammar to enhance word matching

F1-Score:
Combines Precision and Recall to balance brevity and coverage, though fluency remains unaddressed.

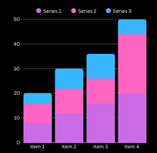
BERTScore:
Uses contextual embeddings (e.g., BERT) to evaluate semantic similarity instead of exact matches, addressing fluency and content overlap.

4.Simple N-Gram Matching:-



Definition

Measures the overlap of contiguous word sequences (n-grams) between the candidate and reference translations.



Scoring Mechanism

n-gram matching aligns score based on degree of overlap in the source and target texts,assisting in objective evaluations of translation fidelity.

Types

1-gram corresponds to single words,while 2-grams captures pair of adjacent words and N-gram can represent any count of consecutive words.

Limitations of Simple N-gram Matching:-

Overly long translations can score artificially high.

N-gram matching favors high-frequency words, as their frequent overlap skews scores, overshadowing meaningful, rare words.

Ignores context and fluency, penalizing word order changes, synonyms, and grammar, focusing only on exact matches.

The Core Idea Behind BLEU

Objective: Provide an automated metric for machine translation evaluation.

Features

- Measures similarity between candidate translations and reference translations.
- Uses n-gram precision and brevity penalty to balance accuracy and translation length.
- Brevity Penalty ensures translations aren't artificially short to inflate accuracy.

How BLEU Works

Steps

Tokenization: Split candidate and reference translations into tokens (words or subwords).

Modified N-Gram Precision: Penalize overuse of words. Focus on the maximum number of matches allowable per reference.

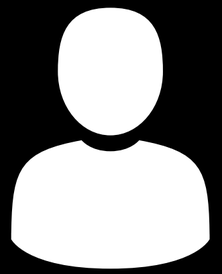
Brevity Penalty (BP): Apply a penalty if the candidate translation is shorter than the reference to discourage under-translation.

N-Gram Matching: Count matches for 1-gram, 2-gram, and higher n-grams between candidate and reference translations.

Geometric Mean: Combine n-gram precisions across all levels into a single score.

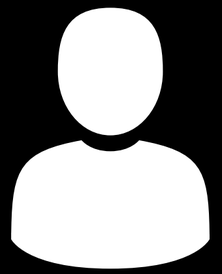
The Score is...

N-Gram Precision Example



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.

N-Gram Precision Example

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

Clearly , Candidate 1 is better.

But, How to quantify that?

N-Gram Precision Example

Candidate 1

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

References

1. It is a guide to action that ensures that the military will forever heed Party commands.
2. It is the guiding principle which guarantees the military forces always being under the command of the Party.
3. It is the practical guide for the army always to heed the directions of the party

N-gram Precision : 17/18

N-Gram Precision Example

Candidate 2

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

References

1. It is a guide to action that ensures that the military will forever heed Party commands.
2. It is the guiding principle which guarantees the military forces always being under the command of the Party.
3. It is the practical guide for the army always to heed the directions of the party

N-gram Precision : 8/14

Issues with N-gram Precision

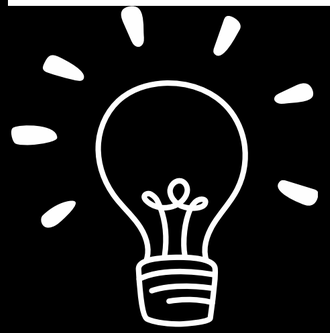
Candidate

The the the the the the the.

N-gram Precision : 7/7

References

1. The cat is on the mat.
2. There is a cat on the mat



Reference word should be exhausted after it is matched

Modified N-gram Precision

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 = Clipped Count / Total no. of candidate word

Modified n-gram precision on blocks of text

Compute the N-gram matches sentence by sentence

Add clipped count for all candidate sentences

Divide the sum by total N-gram in test corpus

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

Modified N-gram Precision

Candidate

The the the the the the the.

- Unigram Count = 7
- Clipped Unigram = 2
- Total count = 2/7

References

1. The cat is on the mat.
2. There is a cat on the mat

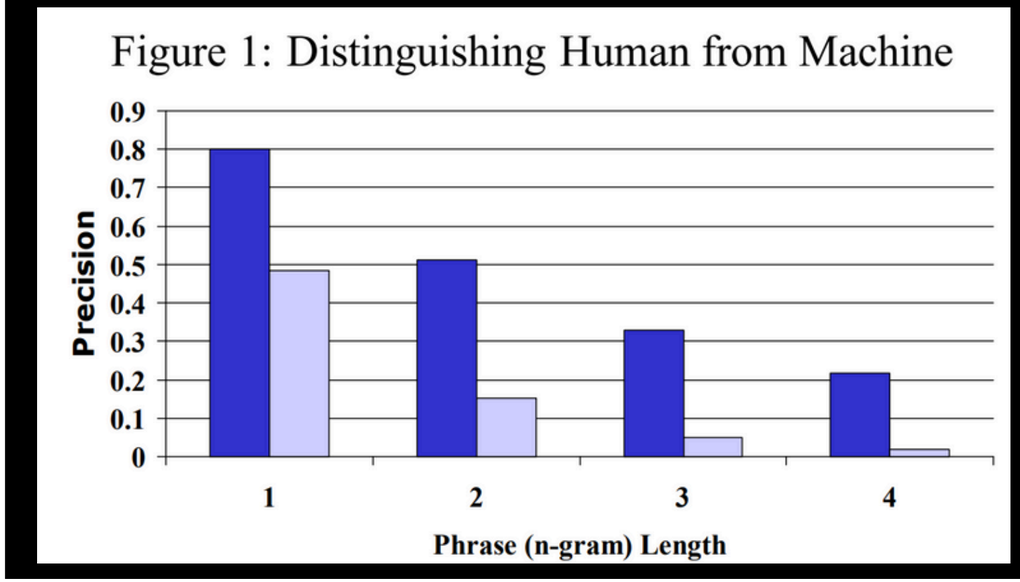
Modified N-gram Precision : 2/7

Modified N-gram Precision

A translation using the same words (1-grams) as in the references tends to satisfy adequacy.

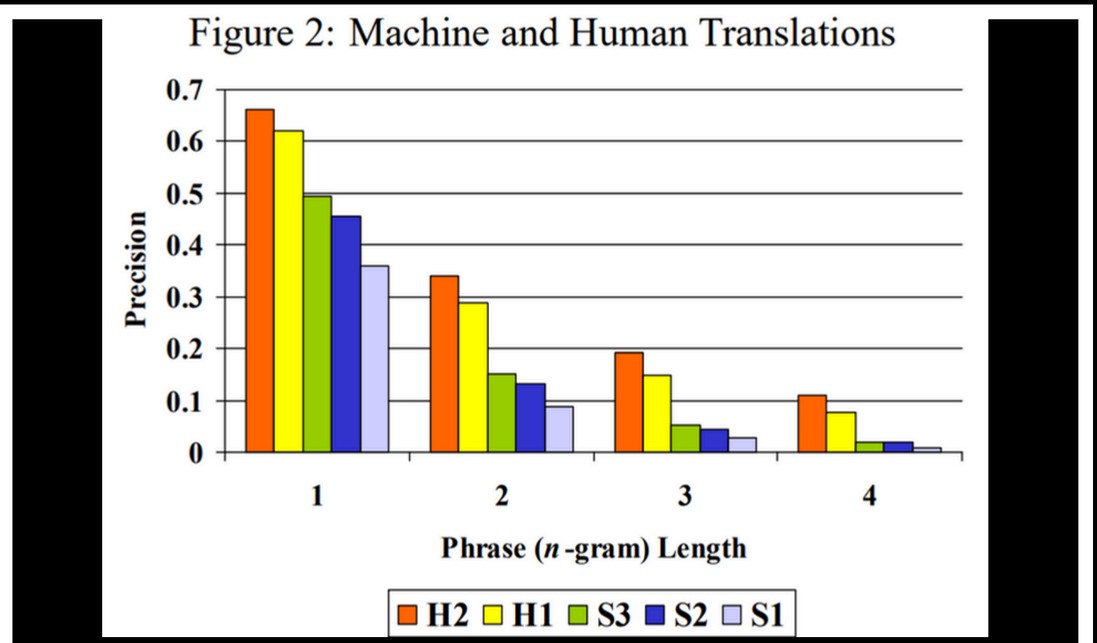
When Longer N-gram precision is high, the reference tends to account for fluency.

Ranking systems using only modified n-gram precision



Findings :

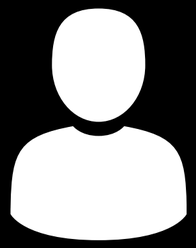
- **Human translations show high precision, while machine translations show low precision.**
- **Precision difference becomes stronger from 1-gram to 4-gram.**



Evaluated translations from:
Two human translators (H1 and H2).
Three machine translation systems (S1, S2, S3).
Scored against two professional reference translations.
Results:
Ranking**: H2 > H1 >> S3 > S2 > S1.**
Alignment**: Same rank order as human judges.**

Combining all n-gram precision scores creates a **more robust metric** than using individual scores alone.

Issues of Modified N-gram Precision : Sentence Length



Candidate 3

of the

Modified Unigram Precision : 2/2

1. It is a guide to action that ensures that the military will forever heed Party commands.
2. It is the guiding principle which guarantees the military forces always being under the command of the Party.
3. It is the practical guide for the army always to heed the directions of the party

Issues of Modified N-gram Precision : Trouble with recalls

Candidate 1
(Bad Translation)

I always invariably perpetually do.

Candidate 2
(A complete Match)

I always do.

1. I always do.
2. I invariably do.
3. I perpetually do.

Geometric Mean in BLEU

Why Geometric Mean?

Balances evaluation across n-grams of varying lengths

Avoids dominance of frequent unigrams over rare higher-order n-grams.

Penalizes poor performance on any n-gram level.

Formula

$$=\exp\left(\frac{1}{N}\sum_{i=1}^{N}\log P_n\right)$$

Impact:

A robust, aggregated measure reflecting the alignment between candidate and reference translations across multiple n-gram levels.

Brevity Penalty

Consider two reference sentences and consider a candidate sentence
R1 : The cat is on the mat .
R2 : There is a cat on the mat.
C : There is a cat.
BLEU (n gram) = 1 for n = 1 to 4.
MEAN_BLEU** = 1

We can see the MEAN_BLEU** = 1 for the candidate translation even though it is missing a piece of text (“on the mat”) with respect to reference translations. To avoid this the BLEU metric introduces Brevity penalty. Brevity penalty penalizes translations that do not contain relevent text from the reference translations.

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

- c is the length of candidate translation
- r is the length of reference translatiom
- BP is a decaying exponential
- BP can attain a maximum value of 1 (when candidate translation is longer than or equal to reference translation)

The BLEU Metric

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

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

- We use N = 4 and uniform weights(wn) = 1/N
- wn are positive weights summing to 1
- pn is modified n-gram precision
- c is the length of candidate translation
- r is the length of the reference translation

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

The ranking behaviour is more visible in log domain

The BLEU metric ranges from 0 to 1. Few translations will attain a score of 1 unless they are identical to a reference translation. For this reason, even a human translator will not necessarily score 1. It is important to note that the more reference translations per sentence there are, the higher the score is.

Experimental Results

BLEU Scores

Table 1: BLEU on 500 sentences

S1	S2	S3	H1	H2
0.0527	0.0829	0.0930	0.1934	0.2571

Mean across 500 sentences

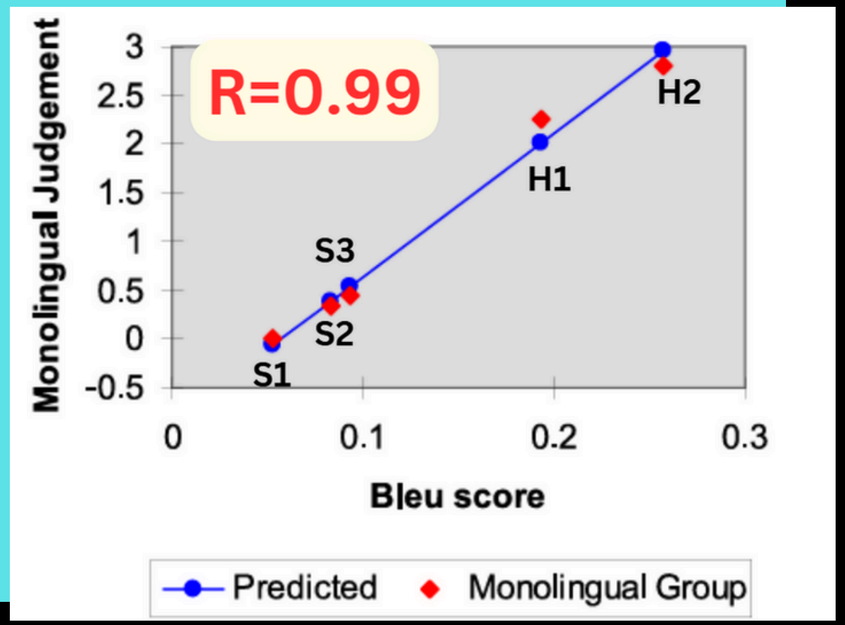
Table 2: Paired t-statistics on 20 blocks

	S1	S2	S3	H1	H2
Mean	0.051	0.081	0.090	0.192	0.256
StdDev	0.017	0.025	0.020	0.030	0.039
t	—	6	3.4	24	11

Mean and StdDev across 20 blocks of 25 sentences each

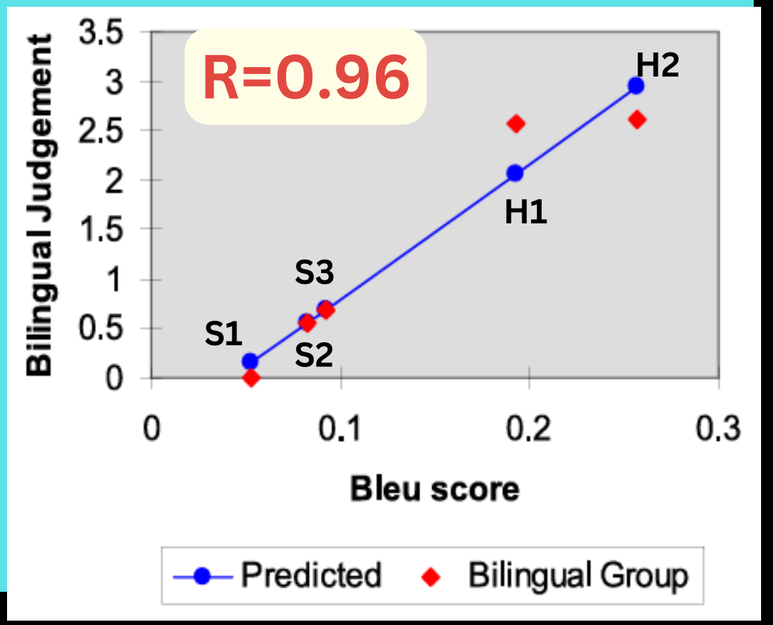
S1, S2, S3- Commercial Machine Translation Systems
H1, H2 - Human Translations

BLEU and Human Judges



Monolingual- English Speakers

Bilingual- Both Chinese and English Speakers



BLEU Score is calculated using 2 references from professional human translators

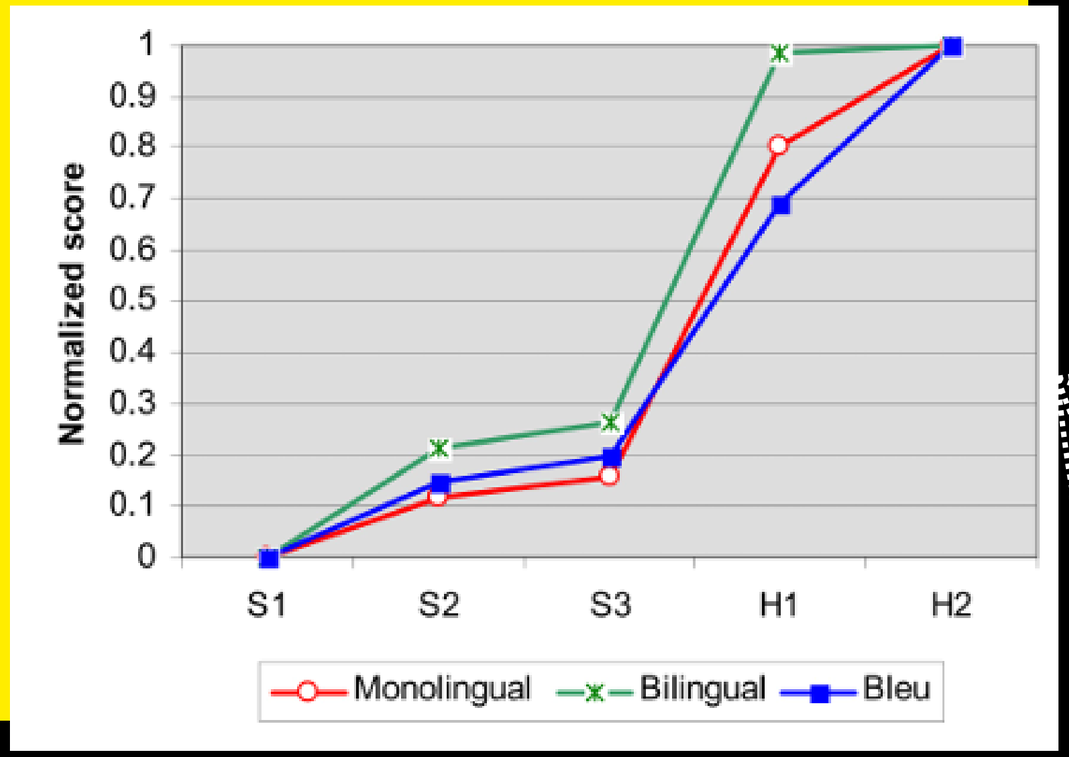
Monolingual Judgments-

Purpose: To evaluate fluency and readability of the translated sentences without considering the source text.

Bilingual Judgments-

Purpose: To evaluate both the accuracy (faithfulness to the source text) and fluency of translations.

Comparing BLEU and Human Judgement Scores



BLEU vs Bilingual and Monolingual Judgements

Used Worst System as Reference

Normalised Scores

BLEU and monolingual group scores align closely, reflecting strong agreement on fluency.

Small differences in bilingual group judgments show sensitivity to semantic accuracy and leniency toward human translators.

Disadvantages of Using BLEU:

1.Doesn't understand meaning:

BLEU only looks for exact word matches, not the meaning. It might give a low score to translations that use synonyms or slightly different words, even if the meaning is correct.

2.Too strict with word choices:

BLEU can punish translations that use different words with the same meaning, even if people would find them accurate.

3.Scores are hard to understand:

Small changes in BLEU scores don't clearly show if the translation quality improved in a way that matters.

4.Doesn't care about word order:

BLEU doesn't check if the words are in the right order. A sentence with the same words in a jumbled order can get the same score as the correct sentence.

5.Can be tricked:

A system can be tuned to score high on BLEU by focusing on matching words, even if the overall translation quality doesn't really improve.

6.Bad with complex language:

BLEU struggles with idioms, cultural meanings, and tricky language where word-for-word translation doesn't work.

Thank You