MeTRICS



Tier 2 Advance

ConceptBoard

Created by Amardeep Singh Sidhu

[Root 2](#_Toc42457393)

[Metrics 2](#_Toc42457394)

[BLEU 2](#_Toc42457395)

[Resources Considered 2](#_Toc42457396)

[BLEU Paper 3](#_Toc42457397)

[How do humans Evaluate Machine Translation(MT)?? 3](#_Toc42457398)

[Criteria Set 3](#_Toc42457399)

[Understanding Terms 3](#_Toc42457400)

[BLEU 7](#_Toc42457401)

[Numerical Metric 7](#_Toc42457402)

[Smooth Belu[3] 9](#_Toc42457403)

[Learn from NLTK, Pytorch & Tensorflow Implementation 9](#_Toc42457404)

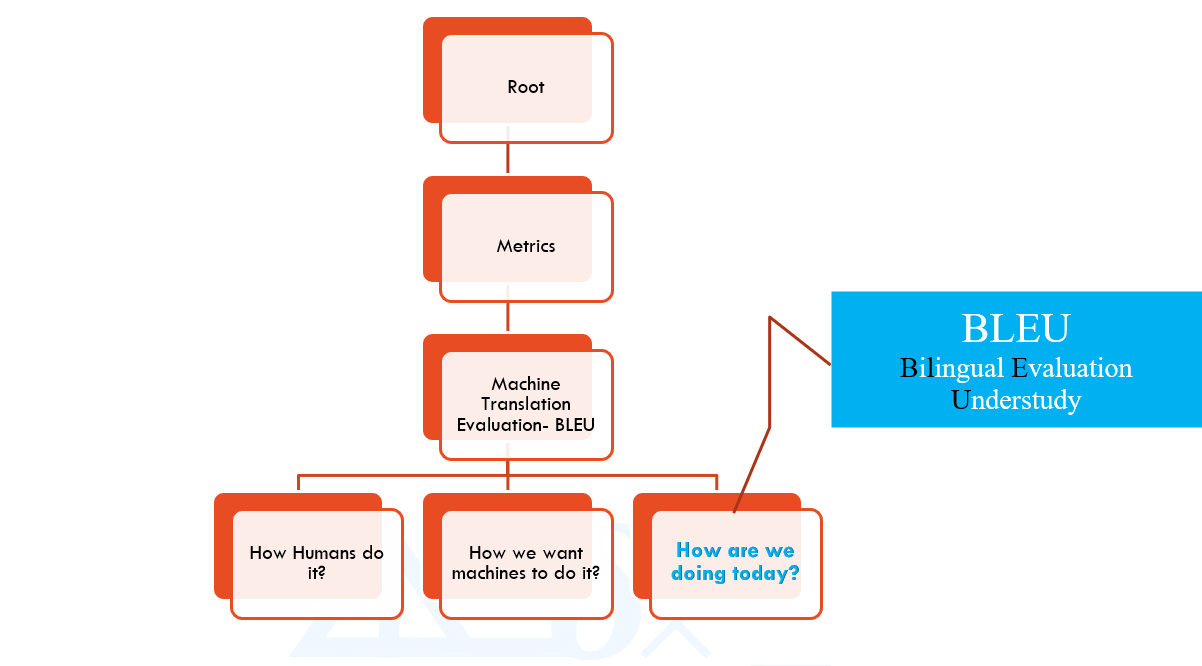
[Does BLEU take care of Coherence?? 9](#_Toc42457405)

[Bibliography 9](#_Toc42457406)

[INDEX 10](#_Toc42457407)

[Table of figures 11](#_Toc42457408)

# Root



# Metrics

## BLEU

### Resources Considered

* BLUE PAPER: Papineni et al, BLEU: a Method for Automatic Evaluation of Machine Translation (2002) [1]
* NLTK[[1]](#footnote-1)

### BLEU Paper

Papineni et al, BLEU: A Method for Automatic Evaluation of Machine Translation (2002)

Objective: Find out method of automatic evaluation of machine translation which is quick, inexpensive, language independent & correlates highly with **human evaluation**.

#### How do humans Evaluate Machine Translation(MT)??

Sources

* (Eduard Hovy, 1999) Principles of Context-Based Machine Translation Evaluation.[4]
* (Guzman) How do Humans Evaluate Machine Translation[3]

##### Criteria Set

* Adequacy(meaning), Fidelity (exactness or accuracy) & Fluency[8]
* Rank a set of hypotheses in order of preference

##### Understanding Terms

* **Fluency**:

1. Fluency is an abstract concept[8], so we assign observable variables. Two of the most reliable factors are “speech rate” and “utterance length”. Speech rate can be defined as how much (effective) language you’re producing over time, for example how many syllables per minute. Utterance length is, as an average, how much you can produce between disfluencies (e.g. a pause or hesitation). You could look at accuracy as being subsumed into fluency, in terms of grammatical accuracy, lexical choice, pronunciation, and precision.
2. Fluency is ability to produce translation which are perceived natural by human addressee[5]. Fluency is naturalness, grammaticality or readability. (Lau, 2017)

* **Discourse Management[2]**: Discourse management refers to the ability to produce extended written and spoken texts, for example conversations. We use a range of devices to produce effective discourse, including cohesion and coherence[6], paralinguistic tools, different communicative functions, and conversational principles.
* **Cohesion and Coherence**: [6]

Coherence is defined as the quality of being logical, consistent and able to be understood. Imagine coherence as a building (It’s an analogy, go with it).

Cohesion on the other hand refers to the act of forming a whole unit. It is effectively a subset of coherence. Picture cohesion as the bricks and cement which make up the building.

You can have cohesion without coherence but you cannot have coherence without cohesion. The picture does not make sense unless the correct pieces are placed in the correct order, even if certain pieces may be the same size and shape.

“I bought some hummus to eat with celery. Green vegetables can boost your metabolism. The Australian Greens is a political party. I couldn’t decide what to wear to the new year’s party.”

In the example above, there are lexical links from one sentence to the next; cohesive ties are used to join the sentences. There is evidence of lexical repetition, ‘green’ ‘party’ and collocations, ‘new years’.

However, this string of sentences does not make any sense; there is no binding semantic link. This is an example of cohesion without coherence.

Cohesive devices effectively help the discourse flow. They include **collocations**, **lexical repetition, linking adverbials**, **substitution**, **ellipsis**, **conjunctions**, **synonymy/antonymy**, **hypernyms/hyponyms** and **referencing (anaphoric, cataphoric, deictic)**. These devices create physical links between the words in a discourse.

Coherence which we previously defined as understanding can be achieved through devices such as **cohesive ties**, **formatting techniques**, **inference, logical ordering of information**, **semantic patterning** and **consistency**

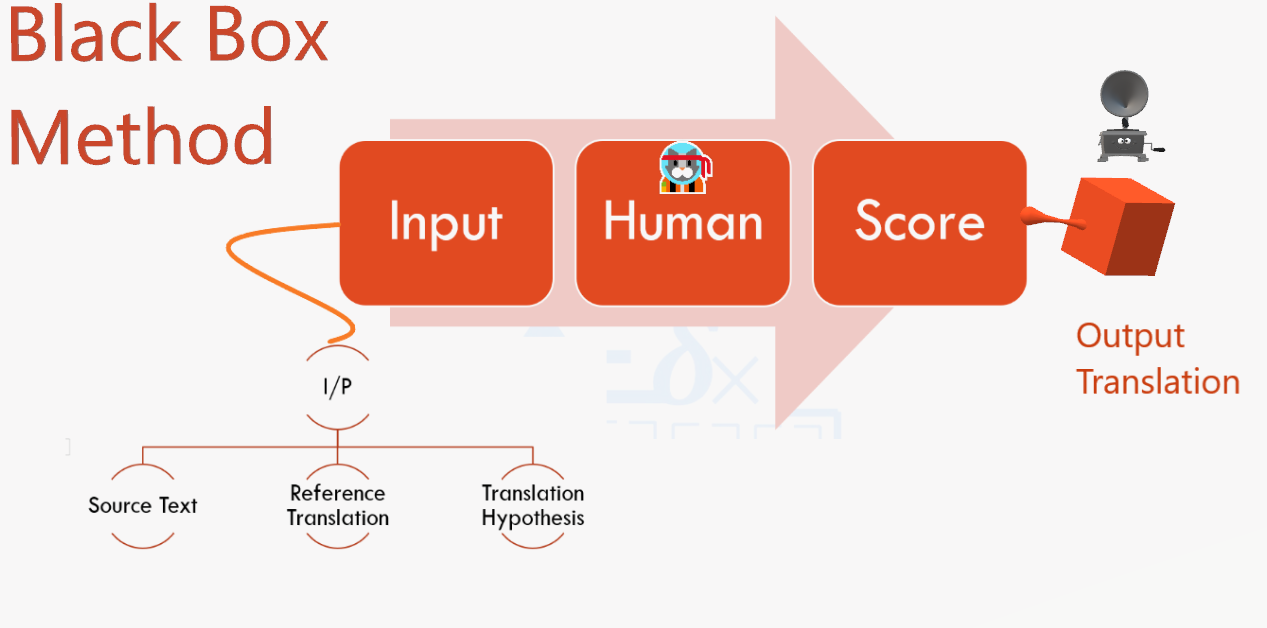
These all enhance the ability of a text to be successfully interpreted and understood. Recipes, terms and condition documents, informative brochures all make use of formatting in the form of headings, bolding, underlines etc. to emphasis certain aspects of the text and draw audience attention to the most important elements.

Focus and use of coherence factors depends on purpose and context.

Types of Evaluation

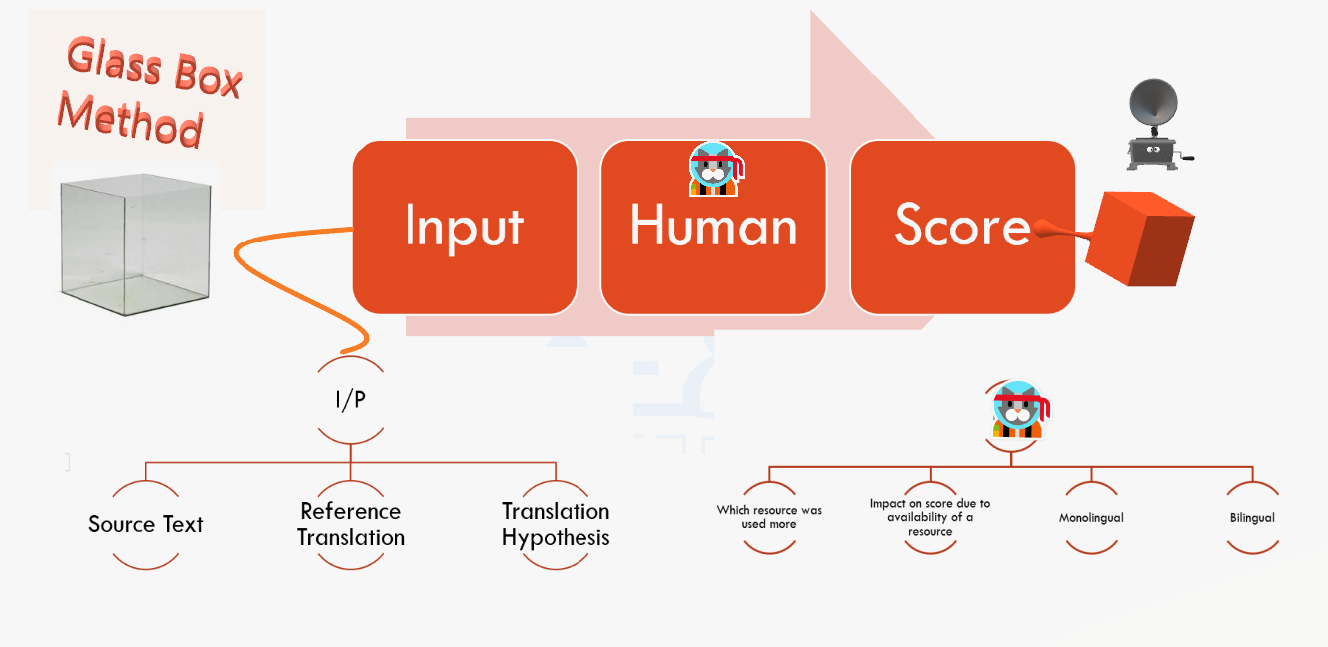
* Black Box Evaluation: Inputs are sources of information (Source Text, Reference Translation, and translation hypothesis) & Output is a score.

Figure 1: How do Humans Evaluate Machine Translation-Black Box (Guzman et al)



* Glass Box Method

Figure 2:How do Humans Evaluate Machine Translation-Glass Box (Guzman et al)



### BLEU

BLUE judges’ quality of machine translation according to numerical metric which measures its closeness to one or more human translation. Intention is to select machine translation which are closer to human translation.

BLUE has 2 main ingredients

* Corpus of good quality human translation.
* Numerical Metric

#### Numerical Metric

Word error rate metric modified for:

* Multiple reference translation
* Allow for legitimate difference in word order.
* Weighted average variable length phrase matches against reference translation.

**Compare & Count matches**

**Modified Unigram Precision**

“Clipping Factor”: Find out how many times a word appears in reference translations. In case below ‘the’ appears a maximum of 2 times (See Reference 1). So CountClip= 2

In more general form CountClip = min(Count;Max Ref Count)

In Case below

CountClip=min(7,2)=2

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 Unigram Precision =

Algorithm to compute modified precision score pn

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

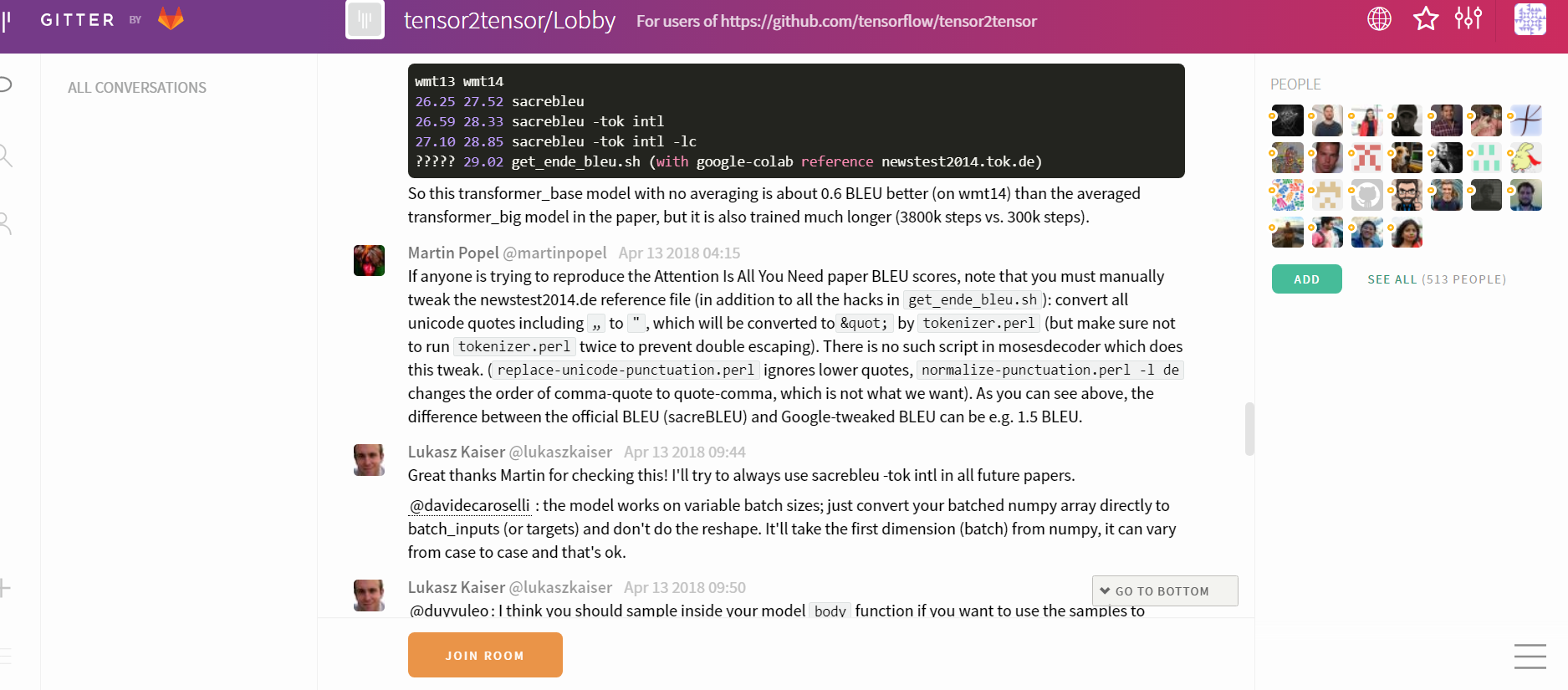
As per Paper

A translation using the same words (1-grams) as in the references tends to satisfy adequacy. The longer n-gram matches account for fluency.

#### Smooth Belu[3]

### BLEU in WMT

<https://gitter.im/tensor2tensor/Lobby?at=5acfe16c7c3a01610dd81b46>



### Learn from NLTK, Pytorch & Tensorflow Implementation

### Does BLEU take care of Coherence??

# Bibliography

1. al, P. e. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. Retrieved from https://www.aclweb.org/anthology/P02-1040.pdf
2. *British Council*. (n.d.). Retrieved from https://www.teachingenglish.org.uk/article/discourse-management#:~:text=Discourse%20management%20refers%20to%20the,of%20spoken%20and%20written%20language.
3. Chin-Yew Lin, F. J. (2004). ORANGE: a method for evaluating automatic. *COLING*. Retrieved from https://www.aclweb.org/anthology/C04-1072.pdf
4. Guzman, F. (n.d.). How do Humans Evaluate Machine Translation. Retrieved from https://www.aclweb.org/anthology/W15-3059.pdf
5. Hovy, E. (1999). Principles of Context-Based Machine Translation. Retrieved from http://www.cs.cmu.edu/~./hovy/papers/03MT-FEMTI-MT-eval.pdf
6. Katharina Kann, S. R. (2018). Sentence-Level Fluency Evaluation:References Help, But Can Be Spared! Retrieved from https://www.aclweb.org/anthology/K18-1031.pdf
7. Kaur, H. (2018, April). *Coherence and Cohesion*. Retrieved from https://www.vcestudyguides.com/blog/coherence-and-cohesion
8. Lau, L. e. (2017). Grammaticality, Acceptability, and Probability:A Probabilistic View of Linguistic Knowledge. *Cognitive Science*. Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1111/cogs.12414
9. O'Connel, W. a. (1994). The ARPA MT Evaluation Methodologies:Evolution, Lessons, and Future Approaches. Retrieved from https://pdfs.semanticscholar.org/bf45/f9e578cb4b43a2604d6149553ae8cfee3016.pdf?\_ga=2.3307190.78158728.1591129098-1678786335.1585380834
10. Sandoval, E. (2019). *BBC*. Retrieved from https://www.bbc.com/future/article/20190903-linguistic-fluency-proficiency-second-language-learning
11. Sennrich, R. (2016). Neural Machine Translation of Rare Words with Subword Units. Retrieved from https://www.aclweb.org/anthology/P16-1162.pdf

# INDEX

A

Adequacy · 3

**anaphoric** · 4

**antonymy** · 4

B

BLEU Paper · 2

C

**cataphoric** · 4

Coherence · 3

Cohesion · 3

**cohesive ties** · 4

**collocations** · 4

**conjunctions** · 4

D

**deictic** · 4

**Discourse Management** · 3

E

**ellipsis** · 4

F

Fidelity · 3

Fluency · 3

**formatting techniques** · 4

H

**hypernyms** · 4

**hyponyms** · 4

I

**inference** · 4

L

**lexical** · 4

**linking adverbials** · 4

**logical ordering of information** · 4

R

**referencing** · 4

**repetition** · 4

S

**semantic patterning** · 4

**substitution** · 4

**synonymy** · 4

# Table of figures

[Figure 1: How do Humans Evaluate Machine Translation-Black Box (Guzman et al) 6](#_Toc42329547)

[Figure 2:How do Humans Evaluate Machine Translation-Glass Box (Guzman et al) 6](#_Toc42329548)

1. nltk.translate.bleu\_score (<https://www.nltk.org/_modules/nltk/translate/bleu_score.html>) [↑](#footnote-ref-1)