BLEU Score Analysis for COSC478 - Assignment 1/Task 1

David Ewing

2024.09.07

Abstract

This report presents the analysis of BLEU scores comparing translations generated by two different LLMs, CoPilot and GPT-4.o, against reference translations provided by an expert. The analysis covers four questions, with detailed results including unigram to 4-gram scores, as well as the final BLEU score for each question.

1 Introduction

In this assignment, we evaluate the performance of two language models (CoPilot and GPT-4.o) against a set of reference translations. The BLEU score is used as a metric to quantify the similarity between the machine-generated translations and the reference. BLEU scores are computed for individual n-grams (unigram to 4-gram) and aggregated into a final score that reflects the overall translation quality.

The reference translations were provided by four individuals, using a dictionary resource for additional reference. The dictionary used in this task was [INSERT DICTIONARY NAME AND URL]. These translations were used to evaluate the accuracy of machine-generated translations from the CoPilot and GPT-4.o LLMs.

2 Methodology

The BLEU score was calculated using a custom Python script, comparing translations generated by CoPilot and GPT-4.o to expert-provided translations. The final BLEU score is a geometric mean of the unigram, bigram, trigram, and 4-gram scores, with a brevity penalty to penalise overly short translations.

The term "brevity penalty" is commonly used in the context of the BLEU metric for evaluating machine translation models. It was introduced in the original BLEU paper (Papineni, Roukos, Ward, & Zhu, 2002) to address the issue that shorter translations tend to achieve artificially high scores by producing fewer but exact words. The brevity penalty penalises translations shorter than the reference, ensuring that translation length and precision are balanced.

2.1 BLEU Score Calculation

The BLEU score is calculated based on the geometric mean of the n-gram precisions (unigram, bigram, trigram, and 4-gram). The formula for the BLEU score is:

$$BLEU = BP \times \left(\prod_{n=1}^{4} p_n \right)^{\frac{1}{4}}$$

Where:

- BP is the **brevity penalty**, which penalises overly short translations. It is defined as:

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

  where:
  c is the length of the candidate translation, and
  r is the length of the reference translation.

- p_n is the precision for the n-gram (unigram, bigram, trigram, 4-gram).

- $\prod_{n=1}^{4} p_n$ represents the product of the 1-gram, 2-gram, 3-gram, and 4-gram precisions.

- $\frac{1}{4}$ represents the 4th root to compute the geometric mean.

3 Discussion

The analysis of the BLEU scores reveals notable patterns in the translation accuracy between the two LLMs, CoPilot and GPT-4.o. While both models exhibit strengths in certain areas, their performance shows clear differences when handling specific types of sentences.

For instance, sentences involving technical or structured language, such as those describing processes or architectures, tend to show a higher BLEU score for GPT-4.o compared to CoPilot. This suggests that GPT-4.o may have been trained more extensively on technical text, allowing it to better capture precise terminology and sentence structure. Conversely, CoPilot's performance on these sentences was weaker, with lower n-gram scores across the board.

On the other hand, both models struggled with more conversational or colloquial translations, such as sentences about pets. The BLEU scores for these sentences were notably lower, indicating that neither model was able to fully capture the natural language flow and colloquial expressions used in the reference translations. Interestingly, CoPilot showed marginally better performance in these cases, possibly due to its stronger handling of everyday language patterns.

Another trend observed was that GPT-4.o tended to produce longer translations compared to CoPilot, often introducing additional context or phrasing that was not present in the reference translations. This occasionally led to a lower BLEU score due to the brevity penalty, despite the translation being accurate. CoPilot, while more concise, often missed nuances that were essential for an accurate translation, which also affected its scores.

Overall, the performance of both models suggests that while they are capable of producing accurate translations, their strengths vary depending on the sentence type. GPT-4.o excels in technical accuracy but is prone to over-explanation, while CoPilot is more suited to conversational language but struggles with precision.

4 Conclusion

Based on the BLEU score analysis, both CoPilot and GPT-4.o performed similarly across all questions, but each had unique strengths and weaknesses. The BLEU scores indicate that neither model consistently outperformed the other across all sentence types, but their performance diverged depending on the complexity and structure of the sentences.

CoPilot demonstrated a strength in translating more conversational language, as evidenced by its higher BLEU scores for sentences related to pets. However, its weakness lay in translating more technical or structured language, where it

often fell short in capturing the necessary precision and structure required for an accurate translation.

Conversely, GPT-4.o performed well on technical translations, achieving higher scores on sentences involving processes or architectures. This indicates that the model is more adept at handling structured and specialised language, likely due to extensive training on technical text. However, its tendency to over-explain and introduce additional context resulted in lower scores for more straightforward translations, particularly when the brevity penalty was applied.

The findings from this analysis highlight the importance of selecting the right model for the task at hand. For conversational translations, a more concise model like CoPilot may be more suitable, whereas for technical or specialised translations, GPT-4.o provides more accurate results. However, neither model was able to consistently match the quality of human translations, as reflected in the generally low BLEU scores.

Further work could explore fine-tuning these models for specific tasks, such as adjusting their verbosity or improving their ability to handle colloquial expressions. Future research could also involve comparing these models against other LLMs or exploring hybrid approaches to translation.

5 Appendix A: BLEU Score Printout (Truncated)

```
PS C:\GitHub\COSC478-OXO> ...
REFERENCE_WHO: ajveendijk@proton.me
Reference Translation: Een transformatie architectuur is een diep leren architectuur wat een zelf attentie ...
LLM_WHO: CoPilot
LLM Translation: Een transformer-architectuur is een deep learning modelarchitectuur die zelfaandachtmechanismen ...
Score Unigram: 0.5
Score Bigram:  0.2308
Score Trigram: 0.0
Score 4-gram:  0.0
Final BLEU Score: 0.0
--------------------------------------------------
Question: 2
REFERENCE_WHO: ajveendijk@proton.me
Reference Translation: In het tokenisatie proces wordt tekst omgezet naar kleinere eenheden die kunnen ...
LLM_WHO: CoPilot
LLM Translation: Tokenisatie is het proces van het omzetten van tekst in kleinere eenheden die door been ...
Score Unigram: 0.5
Score Bigram:  0.1579
Score Trigram: 0.0556
Score 4-gram:  0.0
Final BLEU Score: 0
--------------------------------------------------
...
--------------------------------------------------
Question: 4
REFERENCE_WHO: ajveendijk@proton.me
Reference Translation: Katten houden ervan om dutjes te doen in het huis en honden hollen graag achter stokken aan.
LLM_WHO: GPT-4.o
LLM Translation: Katten houden ervan om door het huis te dutten en honden vinden het leuk om achter stokken aan te rennen.
Score Unigram: 0.6471
Score Bigram:  0.3158
Score Trigram: 0.1111
Score 4-gram:  0.0588
Final BLEU Score: 0.19116681108045488

--------------------------------------------------
PS C:\GitHub\COSC478-OXO>
```

6 Appendix B: Translations Overview

This appendix provides a truncated version of the two LLM translations for readability, followed by a summary table of the BLEU scores.

6.1 Question 1

- **Reference Translation (ajveendijk@proton.me)**:
  Een transformatie architectuur is een diep leren architectuur wat een zelf attentie mechanisme gebruikt om de data te verwerken.
- **LLM Translation 1 (CoPilot)**:
  Een transformer-architectuur is een deep learning modelarchitectuur die zelfaandachtmechanismen gebruikt om invoergegevens te verwerken.
- **LLM Translation 2 (GPT-4.o)**:
  Een transformer-architectuur is een diep leermodelarchitectuur die zelfaandachtsmechanismen gebruikt om invoergegevens te verwerken.

6.2 Question 2

- **Reference Translation (ajveendijk@proton.me)**:
  In het tokenisatie proces wordt tekst omgezet naar kleinere eenheden die kunnen worden verwerkt door grote taalmodellen.
- **LLM Translation 1 (CoPilot)**:
  Tokenisatie is het proces van het omzetten van tekst in kleinere eenheden die door een groot taalmodel kunnen worden verwerkt.
- **LLM Translation 2 (GPT-4.o)**:
  Tokenisatie is het proces van het omzetten van tekst in kleinere eenheden die door een groot taalmodel kunnen worden verwerkt.

6.3 Question 3

- **Reference Translation (ajveendijk@proton.me)**:
  Zowel katten als honden zijn leuke en loyale huisdieren voor het leven.
- **LLM Translation 1 (CoPilot)**:
  Zowel katten als honden zijn geweldige en loyale metgezellen voor het leven.
- **LLM Translation 2 (GPT-4.o)**:
  Zowel katten als honden zijn geweldige en trouwe metgezellen voor het leven.

6.4 Question 4

- **Reference Translation (ajveendijk@proton.me)**:
  Katten houden ervan om dutjes te doen in het huis en honden hollen graag achter stokken aan.
- **LLM Translation 1 (CoPilot)**:
  Katten houden ervan om in huis te dutten en honden vinden het leuk om achter stokken aan te rennen.
- **LLM Translation 2 (GPT-4.o)**:
  Katten houden ervan om door het huis te dutten en honden vinden het leuk om achter stokken aan te rennen.

7 BLEU Score Summary Table

The table below summarises the BLEU scores for each comparison, including the question number, BLEU score, and the individuals who provided the reference and LLM translations.

Question	BLEU Score	Reference WHO	LLM WHO
1	0.0	ajveendijk@proton.me	CoPilot
1	0.0	ajveendijk@proton.me	GPT-4.o
2	0.0	ajveendijk@proton.me	CoPilot
2	0.0	ajveendijk@proton.me	GPT-4.o
3	0.465953841519	ajveendijk@proton.me	CoPilot
3	0.436683544285	ajveendijk@proton.me	GPT-4.o
4	0.200935383892	ajveendijk@proton.me	CoPilot
4	0.191166811080	ajveendijk@proton.me	GPT-4.o

References

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318).