COMPARING METRICS FOR LLM MODELS ON AVIATION TEXTS IN SLOVAK LANGUAGE

Marek DOBEŠ
CSPV SAV, Karpatská 5, 04001 Košice, Slovak Republic, dobes@saske.sk

Abstract. In this article, we present a comprehensive evaluation of three large language models (LLMs) – LLaMA3, Gemma7B, and Aya – focusing on their performance in generating aviation-related texts in the Slovak language. The study employs three widely recognized evaluation metrics: METEOR, BLEU, and ROUGE, to systematically assess the quality of the generated texts.

Keywords: large language models, metrics, Slovak language

1. INTRODUCTION

Large Language Models (LLMs) have revolutionised the field of natural language processing (NLP) by enabling the generation of human-like text across various domains [1]. These models, such as GPT-3, BERT, and Claude, leverage deep learning architectures to process and generate text, exhibiting remarkable proficiency in tasks ranging from translation to summarization and question answering [2]. LLMs are pre-trained on a vast corpora of text data and can be fine-tuned for specific applications, making them versatile tools for a multitude of NLP tasks [3].

Evaluating the performance of LLMs is crucial for understanding their capabilities and limitations. Metrics such as METEOR, BLEU, and ROUGE are commonly used for this purpose. METEOR (Metric for Evaluation of Translation with Explicit ORdering) evaluates the alignment between generated and reference texts by considering synonymy, stemming, and word order [4]. BLEU (Bilingual Evaluation Understudy) measures the precision of n-grams in the generated text compared to reference translations, providing a quantitative assessment of textual accuracy [5]. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) primarily evaluates text summarization by examining the overlap of n-grams, word sequences, and the longest common subsequence between the generated and reference texts [6].

The Slovak language presents unique challenges and opportunities for LLMs. As a Slavic language, Slovak features complex grammatical structures, rich morphology, and a relatively free word order, which complicate the tasks of parsing and generating text [7]. Additionally, the availability of Slovak language resources is limited compared to more widely spoken languages, posing further challenges for training robust LLMs [8]. Despite these challenges, evaluating LLM performance on Slovak texts is essential for ensuring the applicability of these models in diverse linguistic contexts.

In this article, we conduct a comparative analysis of three LLMs—LLaMA3, Gemma7B, and Aya—focusing on their ability to generate aviation-related texts in Slovak. We utilise the METEOR, BLEU, and ROUGE metrics to assess the quality of the generated texts, providing insights into the strengths and weaknesses of each model in handling the complexities of the Slovak language.

2. METHODOLOGY

2.1 Models used

We chose three models for our evaluation. AYA (CohereForAI/aya-101 on HuggingFace.co), Gemma (google/gemma-7b) Llama 3 (meta-llama/Meta-Llama-3-8B-Instruct).

We chose them for two reasons - they are open-source models and they have been trained also on Slovak texts.

2.2 Metrics used

We use METEOR, BLEU and ROGUE metrics for evaluating the models. We computed the metrics using "nltk" library in Python.

We run the same prompt three times to get an average score. Reference texts, prompts and responses are provided in supplementary materials.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is an evaluation metric for machine translation that aims to address some of the shortcomings of BLEU. It was designed to improve correlation with human judgments of translation quality. Key aspects of METEOR include:

a) Precision and Recall: Unlike BLEU, which focuses more on precision, METEOR considers both precision and recall. This helps to better balance the evaluation of how much of the generated text matches the reference text.
b) Stem and Synonym Matching: METEOR allows for stemming (matching words with the same root) and synonym matching (using WordNet), which makes it more flexible in recognizing semantically similar phrases.
c) Alignment: It aligns the generated text with the reference text and penalises long chunks of mismatched words, which helps in evaluating the fluency and readability of the text.

BLEU (Bilingual Evaluation Understudy) is one of the most widely used metrics for evaluating machine translation models. It measures the precision of n-grams (contiguous sequences of n items) in the generated text against the reference text. Key aspects of BLEU include:

a) N-gram Precision: BLEU calculates the precision of unigrams, bigrams, trigrams, and up to four-grams by default. This helps in evaluating both word choice and phrase structure.
b) Brevity Penalty: BLEU includes a brevity penalty to avoid favoring shorter translations that might be overly concise and thus miss essential content.
c) Cumulative Score: BLEU combines the precision of different n-gram lengths into a single cumulative score, which provides a broad overview of the model's performance.

Limitations: BLEU can be insensitive to small differences in meaning and may not handle synonyms or paraphrasing well, leading to lower scores for translations that are correct but use different wording than the reference.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics primarily used for evaluating automatic summarization and machine translation. The most commonly used ROUGE metrics are ROUGE-N, ROUGE-L, and ROUGE-W. Key aspects of ROUGE include:

a) ROUGE-N: Measures the overlap of n-grams between the generated text and the reference text. ROUGE-1 measures unigrams, ROUGE-2 measures bigrams, and so on.
b) ROUGE-L: Focuses on the longest common subsequence (LCS) between the generated text and the reference text, which helps in evaluating the fluency and coherence of the generated text.
c) ROUGE-W: A weighted LCS measure that gives more importance to consecutive matches, thus evaluating the text's fluency and coherence in a more nuanced manner.

Strengths: ROUGE is particularly useful for summarization tasks as it emphasizes recall and is sensitive to the overall structure and content coverage of the generated text.

Interpretation of Results:

METEOR Scores reflect the models' ability to generate text that is not only lexically similar to the reference text but also semantically similar, considering synonyms and stemming.

BLEU Scores indicate how well the models can generate text with precise word choices and phrase structures that match the reference text. Extremely low BLEU scores, like those seen for LLaMA, suggest significant issues with grammatical correctness and fluency.

ROUGE Scores measure the models' ability to capture the overall content and structure of the reference text. High ROUGE scores imply that the generated text includes many of the same phrases and sequences as the reference text, while low scores indicate a lack of content coverage and coherence.

3. RESULTS

The results of the METEOR, BLEU, and ROUGE metrics for the three models – Gemma, LLaMA, and Aya – indicate significant challenges in generating high-quality Slovak aviation-related texts (Table 1).

Gemma shows relatively better performance compared to the other models, particularly in the ROUGE-1 metric, which suggests it captures some aspects of the reference texts. However, its BLEU score is low, indicating issues with fluency and precise word choices

LLaMA performs the worst among the three models, especially notable in its BLEU score, which is effectively zero, indicating severe problems with generating coherent and contextually relevant sentences. Its ROUGE scores are also very low, suggesting it fails to match the reference texts adequately.

Aya has the highest METEOR score and performs relatively well in ROUGE-L compared to the other models, indicating it may be better at capturing the longer dependencies in the text. However, its BLEU score remains very low, pointing to substantial issues with grammatical correctness and lexical choice.

The low scores across all metrics for these models highlight several key issues:

- Underrepresentation of Slovak: These results suggest that Slovak language data is underrepresented in the training data of these models, leading to poor performance.
- Grammatical Errors and Repetition: The models frequently produce grammatically incorrect and repetitive texts, as indicated by the low BLEU scores.

Specialized Topic of Aviation: Part of the poor performance may also be due to the specialised nature of the aviation topic and the lower availability of aviation-related texts in Slovak, which could further limit the models' ability to generate high-quality content in this specific domain.

Need for Specialised Training: There is a clear need for the development of open-source models trained on a more extensive and diverse corpus of Slovak texts to improve the quality of language generation for Slovak. This would likely result in better fluency, coherence, and relevance in generated texts, as measured by these evaluation metrics.

Model	METEOR	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Gemma	0,105441	0,01356	0,225609	0,037867	0,132492
Llama	0,025545	1,51E-81	0,094931	0,015495	0,05963
Aya	0,117684	0,000228	0,164556	0,025687	0,137207

Table 1 Results of metrics for respective models

4. CONCLUSIONS

In conclusion, the evaluation of the three large language models – LLaMA3, Gemma7B, and Aya – revealed that their performance in generating aviation-related texts in Slovak was suboptimal. The evaluation metrics, METEOR, BLEU, and ROUGE, all returned very low scores. This underperformance can likely be attributed to the underrepresentation of Slovak in the training data. Additionally, the generated responses were often grammatically incorrect and repetitive, further highlighting the limitations of these models. This study underscores the urgent need for open-source models specifically trained on a larger corpus of Slovak texts to improve the quality and accuracy of language generation in Slovak.

One more aspect to be taken into consideration in future research is that evaluating language models in Slovak, especially for specialised domains such as aviation, presents unique challenges that are not fully addressed by existing metrics like METEOR, BLEU, and ROUGE. To make these metrics more suitable for the Slovak language, several key improvements can be considered. Slovak has a relatively free word order compared to English, which current metrics may not adequately capture. Metrics should be adjusted to better account for valid word order variations. Additionally, Slovak is a highly inflected language with complex usage of case, gender, and number. Incorporating morphological analysis into evaluation metrics could help in assessing the correct use of these inflections. This adjustment would ensure that generated texts are grammatically correct and syntactically appropriate.

To improve metrics like METEOR, it would be beneficial to use an expanded synonym database specifically tailored for Slovak, similar to WordNet but focused on the Slovak language. This would enhance the recognition of synonym variations and improve the evaluation of semantic similarity. Furthermore, developing a paraphrase database for Slovak would help capture the variety of ways the same meaning can be expressed, thus improving the assessment of fluency and expressiveness in generated texts.

To better reflect the syntactic and morphological richness of Slovak, modifications to BLEU could include adjusting the weight of higher-order n-grams. For example, placing more emphasis on bigrams and trigrams could better capture relevant contextual information in inflected forms. Additionally, implementing a version of Levenshtein distance (edit distance) on n-grams, rather than strict n-gram matching, would allow for small variations that are still grammatically and semantically valid. This approach would accommodate the inherent flexibility and richness of the Slovak language.

Enhancing ROUGE-L to consider morphological variants as part of the longest common subsequence would reward the correct use of inflections, even if they are not exact matches. This adjustment would improve the evaluation of grammatical correctness in generated texts. Additionally, implementing character-level ROUGE metrics would capture fine-grained similarities and variations, which can be particularly useful for inflected languages like Slovak. This would allow for a more nuanced evaluation of the generated texts' quality.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).
[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[3] Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1-67.
[4] Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (pp. 65-72).
[5] 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).
[6] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out (pp. 74-81).

[7] Maučec, M. S., and Janez Brest. "Slavic languages in phrase-based statistical machine translation: a survey." Artificial intelligence review 51.1 (2019): 77-117.
[8] Diandaru, R., et al. "What Linguistic Features and Languages are Important in LLM Translation?." arXiv preprint arXiv:2402.13917 (2024).

Acknowledgement

Research results were partially obtained using the computational resources procured in the national project National competence centre for high performance computing (project code: 311070AKF2) funded by European Regional Development Fund, EU Structural Funds Informatization of society, Operational Program Integrated Infrastructure.

Received 5, 2024, accepted 7, 2024