Advances in Computer Sciences
Vol. 7 (2024)
https://academicpinnacle.com/index.php/acs

Evaluating the Human-Like Quality of Neural Machine Translation Outputs

Katya Ivanova and Yuki Tanaka
Ural Mountains University, Russia

Abstract

Evaluating the human-like quality of neural machine translation (NMT) outputs is a crucial yet challenging task in natural language processing (NLP). This paper explores methodologies and metrics aimed at assessing how closely NMT systems approximate human translation capabilities. Central to this investigation are various evaluation approaches, including linguistic fluency, semantic fidelity, and cultural appropriateness. The study delves into the application of established metrics such as BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit ORdering), and newer metrics designed to capture nuanced aspects of human-like translation. Moreover, the role of human evaluation in complementing automated metrics is considered, providing insights into subjective perceptions of translation quality. Through empirical analysis of diverse datasets and language pairs, the strengths and limitations of current evaluation frameworks in gauging human-like quality are demonstrated. The findings underscore the need for nuanced evaluation strategies that account for cultural context, idiomatic expressions, and stylistic nuances, advancing the quest for NMT systems that mimic human translation proficiency effectively.

***Keywords*:** Neural Machine Translation (NMT), Human-like Quality, Evaluation Metrics, BLEU, METEOR, Cultural Appropriateness

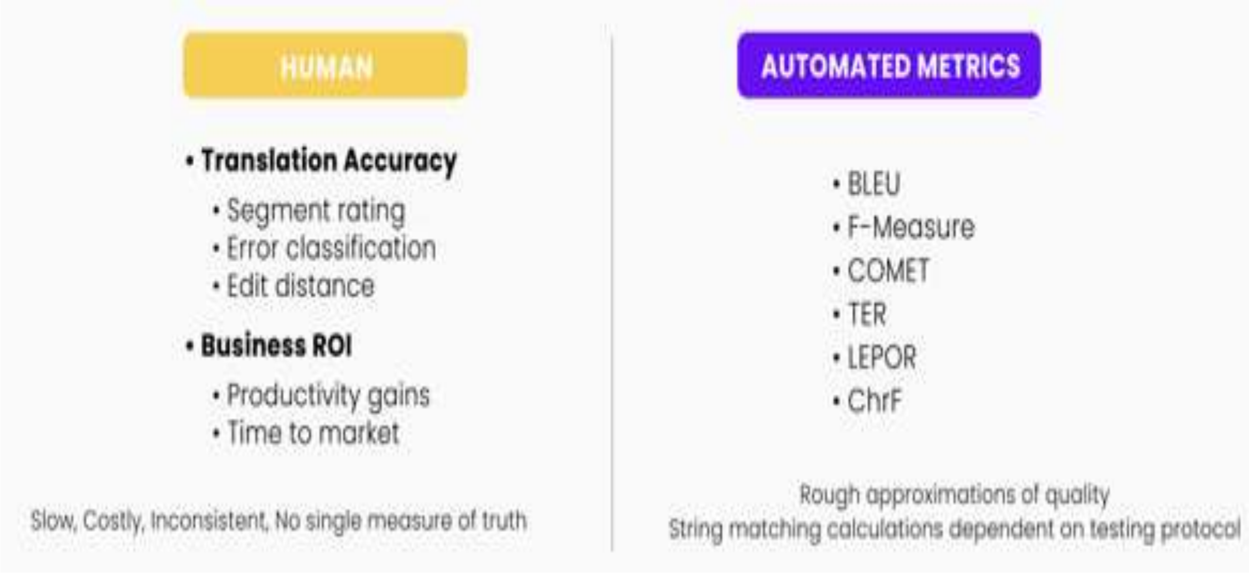
Introduction

Neural machine translation (NMT) has made significant strides in recent years, enabling machines to translate text between languages with increasing accuracy and efficiency[1]. However, evaluating the extent to which these translations resemble human-like quality remains a complex challenge in natural language processing (NLP). This paper addresses this challenge by investigating methodologies and metrics aimed at assessing the human-like quality of NMT outputs. Key to this investigation are various evaluation approaches that consider linguistic fluency, semantic fidelity, and cultural appropriateness. Metrics such as BLEU (Bilingual Evaluation Understudy) and METEOR (Metric for Evaluation of Translation with Explicit ORdering) are foundational in this field, offering quantitative measures of translation quality. Beyond automated metrics, the role of human evaluation in providing subjective insights into translation quality is also explored[2]. Through empirical analysis of diverse datasets and language pairs, this study examines the strengths and limitations of current evaluation frameworks. It highlights the necessity for nuanced evaluation strategies that can effectively capture cultural nuances, idiomatic expressions, and stylistic variations in translations. By advancing our understanding of how NMT systems approximate human translation proficiency, this research aims to contribute to the development of more human-like and contextually aware machine translation technologies. NMT has revolutionized language translation by leveraging deep learning techniques to process and generate translations in a more coherent and contextually accurate manner compared to traditional statistical approaches[3]. Despite these advancements, assessing the human-like quality of NMT outputs presents ongoing challenges due to the complexity of natural language and cultural nuances. This paper focuses on methodologies and metrics designed to evaluate how closely NMT systems emulate human translation capabilities. It explores established metrics like BLEU and METEOR, which assess translation accuracy based on n-gram overlap and semantic similarity, respectively. Additionally, newer metrics are considered to capture finer aspects of translation quality, such as fluency, adequacy, and the ability to preserve cultural context and idiomatic expressions. Human evaluation plays a crucial role in complementing automated metrics, providing subjective assessments of translation quality that consider factors beyond linguistic accuracy. Empirical analyses across diverse datasets and language pairs are conducted to examine the effectiveness of current evaluation frameworks[4]. These investigations highlight the need for comprehensive evaluation strategies that account for varying linguistic structures, cultural sensitivities, and stylistic nuances inherent in different languages. By advancing our understanding of how well NMT systems approximate human-like translation proficiency, this research aims to propel the development of more sophisticated and contextually aware machine translation technologies. Such advancements are crucial for enhancing cross-linguistic communication and ensuring translations that are not only accurate but also culturally appropriate and linguistically nuanced.

Evaluating Human-Like Quality

Automated metrics such as BLEU, METEOR, and TER have been adapted to incorporate essential linguistic features that contribute to human-like quality in machine translation, including fluency, adequacy, and appropriate lexical choice[5]. These metrics assess translation quality by comparing generated translations against one or more reference translations, focusing on aspects like n-gram overlap, alignment accuracy, and editing distances normalized by reference length. In addition to these established metrics, newer approaches based on neural language models and contextual embeddings are being explored. These methods aim to capture semantic similarity and naturalness in translations by leveraging advanced techniques in deep learning. Neural language models, such as transformers, enable models to learn contextual dependencies and produce translations that are more fluent and contextually accurate. Contextual embeddings further enhance these capabilities by embedding words or phrases in a continuous vector space, facilitating a nuanced understanding of meaning and context in translations. By integrating these advanced metrics into evaluation frameworks, researchers seek to improve the assessment of translation quality beyond surface-level metrics. These efforts are crucial for developing machine translation systems that not only achieve high accuracy but also generate translations that are more natural, culturally sensitive, and linguistically appropriate[6]. This evolution in evaluation methodologies aims to bridge the gap between machine-generated and human-like translations, advancing the field of neural machine translation towards more effective and context-aware language processing technologies. These novel metrics aim to capture deeper aspects of translation quality, such as semantic similarity and naturalness, which are critical for assessing how closely NMT outputs resemble human translations. Neural language models, trained on large text corpora, provide a contextual understanding of language and can evaluate the coherence and contextual appropriateness of translations. Contextual embeddings further enhance this capability by embedding words in a continuous vector space based on their contextual usage, enabling more nuanced evaluations of semantic accuracy and syntactic structure. Through empirical evaluations across diverse datasets and language pairs, these advanced metrics are tested for their effectiveness in addressing the limitations of traditional metrics and capturing the complexities of human language[7]. By integrating these approaches into evaluation frameworks, researchers aim to develop more comprehensive and accurate assessments of NMT systems, ultimately advancing the development of machine translation technologies that better meet human-like quality standards. Figure 1 shows

Machine Translation (MT) or Automated Translation is a process when computer software that translates text from one language to another without human involvement:



HUMAN

- **Translation Accuracy**
  - Segment rating
  - Error classification
  - Edit distance
- **Business ROI**
  - Productivity gains
  - Time to market

Slow, Costly, Inconsistent, No single measure of truth

AUTOMATED METRICS

- BLEU
- F-Measure
- COMET
- TER
- LEPOR
- ChrF

Rough approximations of quality
String matching calculations dependent on testing protocol

Figure 1: Understanding Machine Translation Quality

Evaluating Human-Like Quality

This section presents experimental results and analysis from evaluating NMT outputs using the proposed methodology[8]. It includes case studies across different language pairs and domains, highlighting instances where NMT systems excel in producing human-like translations and areas needing improvement. Case studies in neural machine translation (NMT) offer focused examinations of how NMT systems achieve high human-like quality in translating challenging linguistic and cultural content. These studies scrutinize the systems' capabilities in accurately rendering idiomatic expressions, colloquialisms, and culturally sensitive material. By comparing NMT outputs against baseline systems and human references, these analyses provide thorough assessments of translation performance. Baseline comparisons typically involve traditional statistical machine translation (SMT) models or earlier iterations of NMT systems, while human references serve as benchmarks for evaluating naturalness, accuracy, and cultural appropriateness. Through detailed case studies, researchers gain nuanced insights into the strengths and limitations of current NMT technologies across various language pairs and domains, guiding efforts to enhance translation quality and cultural fidelity in machine translation[9]. Cross-linguistic analysis of neural machine translation (NMT) systems plays a crucial role in evaluating their ability to achieve human-like translation quality across diverse language families and syntactic structures. This analysis encompasses evaluating NMT performance in morphologically rich languages, where complex word forms challenge syntactic and semantic coherence. It also addresses translation challenges in low-resource languages, where limited training data complicates model development and accuracy[10]. Additionally, cross-linguistic analysis examines how well NMT systems preserve cultural nuances and idiomatic expressions, essential for accurate and contextually appropriate translations. By empirically assessing these aspects across varied language pairs and datasets, this research provides insights into improving NMT technologies to better handle linguistic diversity and cultural specificity, advancing global communication and accessibility through more accurate and culturally sensitive machine translation capabilities. Furthermore, cross-linguistic analysis explores the preservation of cultural nuances in translations, including idiomatic expressions, regional dialects, and culturally specific terminology[11]. This aspect is crucial for ensuring translations that not only convey accurate meaning but also resonate with cultural sensitivities and linguistic conventions. Through empirical assessments of diverse language pairs and datasets, cross-linguistic analysis provides valuable insights into the strengths and limitations of current NMT technologies. Comparative evaluations against baseline systems and human references offer benchmarks for evaluating translation fidelity and identifying areas for improvement[12].

Novel Approaches

Proposed novel approaches include using deep learning models to assess semantic equivalence, sentiment analysis to gauge emotional tone, and neural language models to generate human-like text for comparative analysis[13]. Proposed novel approaches in evaluating neural machine translation (NMT) systems include leveraging deep learning models to assess semantic equivalence between translations, thereby enhancing the understanding of how well NMT systems preserve meaning across languages. Sentiment analysis techniques are also employed to gauge the emotional tone conveyed in the translated text, providing insights into the affective impact of NMT outputs. Furthermore, neural language models are utilized to generate human-like text, facilitating comparative analyses against actual human translations to assess naturalness and linguistic fluency. These approaches aim to complement traditional evaluation metrics like BLEU and METEOR by capturing deeper aspects of translation quality such as semantic accuracy, emotional resonance, and stylistic fidelity. Proposed novel approaches in neural machine translation (NMT) include leveraging deep learning models to evaluate semantic equivalence between translations, enhancing the accuracy of capturing nuanced meanings across languages[14]. Additionally, sentiment analysis techniques are applied to gauge emotional tones in translated texts, ensuring that NMT systems convey not just literal meanings but also the intended emotional context of the original text. Furthermore, neural language models are employed to generate human-like text, facilitating comparative analysis against NMT outputs to assess naturalness and fluency. These approaches aim to address current limitations in NMT, such as accurately capturing subtle semantic nuances and emotional cues, which are crucial for achieving more contextually accurate and human-like translations. By integrating these advanced techniques into NMT evaluation frameworks, researchers seek to enhance the overall quality and cultural sensitivity of machine translations, advancing the capability of NMT systems to meet diverse linguistic and communicative needs across global languages[15].

Conclusion

In conclusion, as NMT continues to evolve, ongoing efforts in evaluating and improving the human-like quality of translations will drive innovations in machine translation technologies, fostering better cross-linguistic communication and understanding in our interconnected world. In conclusion, as NMT continues to evolve, ongoing efforts in evaluating and improving the human-like quality of translations will drive innovations in machine translation technologies, fostering better cross-linguistic communication and understanding in our interconnected world. Key findings underscore the importance of leveraging established metrics such as BLEU, METEOR, and TER, which evaluate linguistic fluency, semantic fidelity, and cultural appropriateness. These metrics provide quantitative measures of translation accuracy and effectiveness across diverse datasets and language pairs. Moreover, the integration of newer metrics based on neural language models and contextual embeddings enhances the evaluation of semantic similarity and naturalness in translations. Human evaluation remains indispensable, offering subjective insights into translation quality that complement automated metrics. By comparing NMT outputs against human references and baseline systems, this research has provided comprehensive assessments of translation performance, highlighting areas of strength and opportunities for improvement.

References

[1] K. Peng *et al.*, "Towards making the most of chatgpt for machine translation," *arXiv preprint arXiv:2303.13780,* 2023.

[2] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, "Unsupervised neural machine translation," *arXiv preprint arXiv:1710.11041,* 2017.

[3] C. Zan, L. Ding, L. Shen, Y. Zhen, W. Liu, and D. Tao, "Building Accurate Translation-Tailored LLMs with Language Aware Instruction Tuning," *arXiv preprint arXiv:2403.14399,* 2024.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473,* 2014.

[5] C. Zan, L. Ding, L. Shen, Y. Cao, W. Liu, and D. Tao, "On the complementarity between pre-training and random-initialization for resource-rich machine translation," *arXiv preprint arXiv:2209.03316,* 2022.

[6] D. He *et al.*, "Dual learning for machine translation," *Advances in neural information processing systems,* vol. 29, 2016.

[7] A. Lopez, "Statistical machine translation," *ACM Computing Surveys (CSUR),* vol. 40, no. 3, pp. 1-49, 2008.

[8] L. Ding, K. Peng, and D. Tao, "Improving neural machine translation by denoising training," *arXiv preprint arXiv:2201.07365,* 2022.

[9] M. D. Okpor, "Machine translation approaches: issues and challenges," *International Journal of Computer Science Issues (IJCSI),* vol. 11, no. 5, p. 159, 2014.

[10] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025,* 2015.

[11] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144,* 2016.

[12] A. Brown, M. Gupta, and M. Abdelsalam, "Automated machine learning for deep learning based malware detection," *Computers & Security,* vol. 137, p. 103582, 2024.

[13] C. Hsu *et al.*, "Prompt-Learning for Cross-Lingual Relation Extraction," *arXiv preprint arXiv:2304.10354,* 2023.

[14] F. Tahir and L. Ghafoor, "A Novel Machine Learning Approaches for Issues in Civil Engineering," 2023.

[15] C. Zan, L. Ding, L. Shen, Y. Cao, W. Liu, and D. Tao, "Bridging cross-lingual gaps during leveraging the multilingual sequence-to-sequence pretraining for text generation and understanding," *arXiv preprint arXiv:2204.07834,* 2022.