Lengua y Sociedad. Revista de Lingüística Teórica y Aplicada,
Vol. 23, n.º 2, julio-diciembre 2024, pp. 791-805,
ISSN-L 1729-9721; eISSN: 2413-2659
https://doi.org/10.15381/lengsoc.v23i2.26968

Traduttore, traditore: Can machines outperform humans in translation accuracy?

Traduttore, traditore: ¿Puede las máquinas superar a los humanos en precisión de traducción?

Traduttore, traditore: As máquinas podem superar os humanos em precisão de tradução?

Felipe von Hausen
Universidad de Las Américas, Concepción, Chile
fvonhausen@udla.cl
https://orcid.org/0000-0001-7355-7561

Cristóbal Muñoz
Universidad de Concepción, Concepción, Chile
cristourtubia@gmail.com
https://orcid.org/0009-0002-3345-8694

Carlos Contreras Aedo †
Universidad de Concepción, Concepción, Chile
carlcont@udec.cl
https://orcid.org/0009-0006-8506-8654

Abstract

This study examines the efficacy and accuracy of Automatic Translation (AT) compared to Human Translation (HT), employing the Bilingual Evaluation Understudy (BLEU) metric for evaluation. The rapidly evolving field of language translation, especially in the context of machine learning and Artificial Intelligence (AI), necessitates a critical assessment of AT versus HT. We aim to compare the quality of machine-generated translations from Google Translate, DeepL, and ChatGPT 3.5 with HT in the English-to-Spanish language pair. The study employs the BLEU metric, comparing machine and human translations with a professional standard. Data from translation student exams are used for human-generated translations. Our findings indicate a higher structural correlation in machine-generated translations than previously reported, suggesting an increasing proficiency in AT. However, this study emphasises the need for continued evaluation as translation technologies evolve.

Keywords: automatic translation; human translation; BLEU metric; translation accuracy; comparative analysis.

Resumen

Este estudio examina la eficacia y precisión de la Traducción Automática (TA) en comparación con la Traducción Humana (TH), para lo cual utiliza la métrica Bilingual Evaluation Understudy (BLEU) en la evaluación. El campo en rápida evolución de la traducción de idiomas, especialmente en el contexto del aprendizaje automático y la inteligencia artificial, requiere una evaluación crítica de la TA versus TH. Se busca comparar la calidad de las traducciones generadas por máquina de Google Translate, DeepL y ChatGPT 3.5 con traducciones humanas en el par lingüístico inglés-español. El estudio utiliza la métrica BLEU, comparando las traducciones máquina y humanas con un estándar profesional. Se utilizan datos de exámenes de estudiantes de traducción para las traducciones generadas por humanos. Nuestros hallazgos indican una mayor correlación estructural en las traducciones generadas por máquina de lo que se había informado anteriormente, sugiriendo una creciente competencia en la TA. Sin embargo, este estudio subraya la necesidad de una evaluación continua a medida que evolucionan las tecnologías de traducción.

Palabras clave: Traducción Automática; Traducción Humana; Métrica BLEU; Precisión de Traducción; Análisis Comparativo.

Resumo

Este estudo examina a eficácia e precisão da Tradução Automática (TA) em comparação com a Tradução Humana (TH), empregando a métrica de Avaliação Bilíngue Sob Estudo (BLEU) para avaliação. O campo rapidamente evolutivo da tradução de línguas, especialmente no contexto de aprendizado de máquina e Inteligência Artificial (IA), exige uma avaliação crítica da TA versus TH. Nosso objetivo é comparar a qualidade das traduções geradas por máquina do Google Tradutor, DeepL e ChatGPT 3.5 com a TH no par de idiomas inglês-espanhol. O estudo utiliza a métrica BLEU, comparando traduções de máquina e humanas com um padrão profissional. Dados de exames de estudantes de tradução são usados para traduções geradas por humanos. Nossos achados indicam uma correlação estrutural mais alta em traduções geradas por máquina do que o relatado anteriormente, sugerindo uma crescente proficiência em TA. Contudo, este estudo enfatiza a necessidade de avaliação contínua à medida que as tecnologias de tradução evoluem.

Palavras-chave: tradução automática; tradução humana; métrica BLEU; precisão de tradução; análise comparativa.

Received: 12/10/2023 Accepted: 06/17/2024 Published: 12/30/2024

1. Introduction

Neural Machine Translation (NMT) has revolutionised the translation industry since its introduction, significantly impacting translation workflows and quality. The advancements in NMT have resulted in improved fluency and accuracy of machine translation (MT) outputs, prompting integration into various computer-aided translation tools and translation management systems. This technological shift has been embraced by a majority of the translation industry, with over 50 % of industry players adopting MT solutions by 2018 (Loboda & Mastela, 2023).

The study is grounded in the significance of automatic translation as an instrumental skill in the field of translation, as noted by Munday (2011) and Papineni (2002). The importance of professional translators being familiar with and efficiently utilising these tools is highlighted. Various perspectives on the quality of automatic translation are considered. While some authors, such as Hartley (2007), argue that automatic translation surpasses human translation due to its access to a wide range of options and databases, others like Perez (2013) emphasise that sociocultural and pragmatic contexts are better expressed by professional human translators.

Our objective is, therefore, to shed light on the current quality level of automatic translation and its comparison with human translation. Although assessing the quality of automatic translation typically involves human revision, this study explores software evaluations as an alternative, following the approach proposed by Papineni (2002). The focus is on students of Translation and Interpretation in Foreign Languages at a Chilean University. Particularly, our aim is to evaluate the

translations produced by these students and contrast them with the results from Google Translate, DeepL, and ChatGPT 3.5, leaders in automatic translation. This research intends to contribute to the knowledge in the area of translation and technology and is expected to serve as a basis for similar future research. The findings will provide a clearer vision of the effectiveness of automatic translation and its impact on the field.

All in all, this study compares medical texts translated by humans with those translated by ChatGPT 3.5, DeepL, and Google Translate, highlighting differences in quality. It examines both human and automatic translation (HT and AT), focusing on how each handles terminological accuracy and contextual coherence. The functioning of automatic translators, including their natural language processing algorithms and ability to manage specialised medical terminology, is analysed. The study employs the BLEU metric (Papineni, 2002) to evaluate translation quality, using the BLEU+ program (Tantuğ, 2007) for its user-friendly interface and intuitive configuration. The goal is to determine if automatic translation can match or surpass human translation quality in the medical field.

We organised this study as follows: the introduction outlines the background and objectives of the study; the theoretical framework delves into the concepts and previous research related to Human Translation (HT) and Automatic Translation (AT); the methodology details the study design, data collection processes, and analytical tools employed; the analysis presents the results of the BLEU scores and compares the quality of human and automatic translations; and the conclusions summarise the findings, discuss their implications, and suggest directions for future research.

2. Theoretical Framework

Research on machine translation (MT) has highlighted various challenges. Some studies introduce a custom MT model, compared to baselines like Google Translate or improved versions of the same model. Training and validation methods frequently vary, with a focus on single language pairs and specific subfields, such as electronic prescriptions or public health information, due to the need for different MT models for each language pair (Dew *et al.*, 2018; Vieira *et al.*, 2021). Despite extensive MT research, empirical studies are limited. Dew *et al.* (2018) conducted a broad study on health communication from 2006 to 2016, but it excluded neural machine translation (NMT) due to its later development. Recent studies by Vieira *et al.* (2021) examine user perspectives and qualitative analyses in medical and legal contexts. On the other hand, the conferences on machine translation are significant for biomedical MT research, presenting annual tasks that drive advancements in the field (Zappatore & Ruggieri, 2023). Khoong & Rodriguez (2022) suggest focusing on communication scenarios, target populations, MT algorithms, and translation outcomes to improve MT research in clinical care. This section explores the state of the art in Human Translation (HT) and Automatic Translation (AT), focusing on the roles and advancements of both, with an emphasis on leading AT tools like Google Translate, DeepL, and ChatGPT 3.5. It delves into methodologies for evaluating translations, highlighting the role of the BLEU metric in assessing translation quality.

2.1. Human Translation (HT) and Automatic Translation (AT)

Translation, in its modern context, represents a blend of human expertise and computational assistance. Munday (2011) emphasises the significance of computational tools in augmenting the translator's tools, ranging from translation memories to automatic translation aids. While HT is

vital, especially in sensitive or complex contexts such as legal documentation and medical records, AT's role has expanded, offering rapidity and a broad linguistic range. However, as Gouadec (2010) points out, the extent to which human intervention is required in producing a final translated text in today's landscape remains a subject of ongoing debate.

The choice between HT and AT often depends on the context and requirements of the translation task. In scenarios where the stakes are low, such as leisure reading, AT tools offer a convenient alternative. Their benefits, including speed and cost-effectiveness, make them a viable option for many. Yet, when it comes to translations where accuracy and efficacy are indispensable, the irreplaceability of human translators is evident. The terms *accuracy* and *efficacy* in the field of translation studies focus on fidelity to the source text and the effectiveness of the translation in achieving its intended purpose, respectively. Accuracy is closely linked to the linguistic and semantic faithfulness to the original text, while efficacy relates to how well the translation communicates the intended message to its audience, ensuring cultural and contextual appropriateness (Abu-Zahra & Shayeb, 2022; Lee, 2022).

The advancements in AT have brought it closer to the quality of HT. Innovations in neural machine translation, as demonstrated by the developments in Google's GNMT system (Wu *et al.*, 2016), have made translations more fluid and human-like. Google's Neural Machine Translation (GNMT) system uses a sequence-to-sequence framework employing neural networks, particularly a type known as Long Short-Term Memory (LSTM) networks. The GNMT architecture includes an encoder that processes the input text in the source language and transforms it into a high-dimensional vector representation. This vector captures semantic and syntactic properties of the input. The decoder then interprets this vector to generate the output text in the target language, maintaining contextual relevance across the entire sentence. This end-to-end learning process enables more coherent and contextually appropriate translations (Ahammad *et al.*, 2024; Mikros & Boumparis, 2022; Colman *et al.*, 2021). Despite these advancements, as Le (2016) notes, AT is yet to consistently match the nuanced understanding and contextual awareness of a professional human translator.

2.2. Automatic translators

Understanding the functionality of AT is key to grasping its developmental trajectory. The shift to Neural Machine Translation (NMT) marked an important advancement in online translation services. This neural network-based approach, explained by Google's GNMT system as described above, and inspired by biological neural synapses, has improved translation accuracy and context comprehension, as detailed by van Gerven (2017).

Top-tier online translators like Google Translate, DeepL, and Bing Microsoft Translator also utilise NMT to enhance their performance. These platforms have transitioned from Phrase-Based Machine Translation (PBMT) –a method that translates sequences of words or phrases based on statistical models that considers grouping, translating and reordering, without considering full sentence contexts (Dodos, 2017)– to NMT, addressing the complexities and contextual challenges of translation. The continual refinement of these NMT systems, as Brownlee (2017) notes, represents the ongoing quest for greater accuracy and efficiency in AT.

The evolution of AT, especially the move from PBMT to NMT systems like Google's GNMT, has significantly enhanced translation quality. NMT's word encoding and decoding processes offer a deeper understanding of language, addressing the limitations of previous translation models. This shift has been relevant in improving the fluidity and contextuality of translations.

2.3. ChatGPT 3.5 in automatic translation

ChatGPT 3.5, developed by OpenAI, represents a significant advancement in the field of AT. Its sophisticated language model, grounded in deep learning and neural networks, allows for an understanding and generation of human-like text. The integration of ChatGPT 3.5 into the landscape of translation tools brings an approach, leveraging its extensive training data and advanced algorithms to provide translations that aim to balance accuracy with contextual relevance.

The capabilities of ChatGPT 3.5 extend beyond mere word-to-word translation; it encompasses an understanding of idiomatic expressions, cultural differences, and the subtleties of language that are often challenging for traditional AT systems. This makes ChatGPT 3.5 a notable contender in the realm of translation. ChatGPT 3.5's contribution to the field of translation is not just in providing an additional tool for translators but also in enhancing the understanding of how free artificial intelligence can mimic human-like language processing. Its application in translation studies offers valuable insights into the potential and limitations of AI in linguistic tasks, setting the stage for future developments in AT.

2.4 Evaluation of automatic translations

The evaluation of AT's reliability is crucial for its adoption in professional and academic contexts. Papineni's (2002) introduction of the BLEU metric marked a new era in AT assessment, offering a quantifiable method to compare machine-generated translations with human translations. The closer an AT's output is to human translation, the better it is considered to be.

Human evaluators typically focus on aspects such as translation adequacy, which refers to the extent to which the text meets its intended purpose; fidelity, the accuracy with which it conveys the original message; and fluency, the readability and naturalness of the translation in the target language (Ali, 2020; Hovy, 1999; Lee, 2022; Popović, 2020). However, human evaluation, especially for large volumes of text, poses challenges in terms of resource intensity and standardisation. Lavie (2011) discusses the difficulties in maintaining consistency in human evaluations, considering the variance in evaluators' backgrounds and focuses.

Software evaluation presents a scalable and economical alternative to human evaluation, especially in handling extensive translation volumes. While software evaluation initially requires human validation to ensure its correlation with human standards, it simplifies the analysis of a large number of translations. Papineni (2002) and Lavie (2011) acknowledge its limitations, particularly in error detection and consistency in single-phrase analysis, yet its utility in managing large-scale translation evaluations is undeniable.

2.5. BLEU metric

One of the critical strengths of the BLEU metric is its ability to provide a consistent and reproducible measure of translation quality, which is relevant for large-scale evaluations and comparisons. Its standardised approach allows researchers and practitioners to benchmark different translation systems and track improvements over time. As translation technologies continue to evolve, there is ongoing research aimed at enhancing the metric to better reflect the complexities of human language and translation quality.

The BLEU metric, conceptualised by Papineni (2002), is an algorithm designed to measure the quality of translated texts against high-quality human translations. It uses an n-gram precision algorithm, which assesses the overlap of n-gram (a contiguous sequence of n words) matches between the machine-translated text and a reference human translation. This approach quantifies how many n-grams in the translated text are also found in the reference text, thus evaluating the translation's lexical similarity. Additionally, the metric includes penalties for brevity to avoid favouring overly short translations and unigram precision, which specifically measures the accuracy of individual word matches, regardless of their position or context within the sentence.

This metric requires a reference translation, which is either 'perfect' or 'very good' by human standards, against which other translations (candidates) are compared. The n-gram algorithm assigns values between 0 and 1 for each n-gram, with 1 indicating an exact match with the reference. However, even excellent human translations may not achieve a perfect score, illustrating the metric's rigorous evaluation criteria. Determining an ideal BLEU score is vital. A score of 0.50 generally indicates a significant structural correlation with the reference, suggesting legibility. On the other hand, a score of 0.30 or lower suggests poor correlation and, at best, basic comprehensibility. The disparity in BLEU scores between human and automatic translations, even against the same reference, is a critical aspect of this metric's application.

3. Methodology

This empirical study employs a comparative analytical approach, focusing on the evaluation of translation quality from human and automatic sources. The methodology is grounded in the principles outlined by Hernández & Mendoza (2018) and Hernández *et al.* (2014), emphasising empirial data collection, ethical research practices, and thorough analysis processing.

3.1. Study design and data collection

Adopting a comparative and analytical design, this research contrasts the BLEU scores between human translations by university students and automatic translations from Google Translate, DeepL, and ChatGPT 3.5. The corpus selection involved translations from students of a Translation and Interpretation in Foreign Languages program from a Chilean University. A scientific publication was selected for translation because it minimises the variance in expression and style compared to other textual typologies like popular science or journalistic texts. This choice helps standardise evaluations, particularly important when assessing translations by students, as scientific texts typically avoid localisms, jargon, and ambiguities, ensuring clearer and more direct comparisons (De García & Pérez, 2011). An online questionnaire was utilised to determine the suitability of the students' translations for this study, focusing on their coursework completion and consent to provide access to their exams (Hernández-Sampieri *et al.*, 2018).

3.2 Sample selection criteria

The selection process involved a total of 15 students from the course, out of which six met the criteria of completing relevant courses in the selected program without failing any language subjects. These six were selected as they fulfilled our inclusion criteria, ensuring the translations' academic standard and relevance to the study's objectives. Ethical considerations were followed, with informed consent obtained from all participants, in line with the ethical research guidelines (Hernández-Sampieri *et al.*, 2014). There was a case of corruption in the original file of one of the participants, for this reason, only five distinct human candidates were available at the time of analysis.

3.3. BLEU+ and iBLEU software utilisation

For the analysis, we selected the BLEU+ software (Tantuğ, 2007) for its capability to calculate detailed BLEU scores. All texts were adjusted for structure and format consistency, necessary for accurate software analysis. Alongside, the iBLEU software provided a proper data visualisation, enabling a more detailed comparison between the candidates and reference translations (Madnani, 2011). This dual-software approach was instrumental in ensuring a comprehensive evaluation of the translations.

3.4. Analysis procedure

The methodology involved systematic processing of the collected texts. Each student's exam translation was entered into BLEU+ for preliminary analysis. Adjustments were made to align the formats of all texts – both human and automatic translations – for compatibility with the analysis software. The final analysis involved recording BLEU scores and correlational data for each n-gram in Python, ensuring thorough data documentation and analysis (Hernández-Sampieri *et al.*, 2018).

3.5 Hypothesis testing and objective fulfillment

The study's hypothesis was that human translations would achieve higher BLEU scores and n-gram correlations compared to automatic translations. This hypothesis was examined using BLEU+ and iBLEU analyses. BLEU+ is an enhanced version of the BLEU metric that allows for more detailed evaluation metrics, including adjustments for various n-gram lengths and refined penalty factors to assess the fluency and precision of translations more accurately. iBLEU, on the other hand, is an interactive tool that provides a deeper analysis of both the structural and qualitative aspects of translations. It allows for side-by-side comparisons of human and machine translations, highlighting differences and evaluating how closely machine translations mimic human translations in terms of syntax and semantic content. This dual approach ensures a comprehensive analysis of the BLEU scores of translations by translation students against those generated by automatic tools, with specific attention to structural analysis using iBLEU (Madnani, 2011).

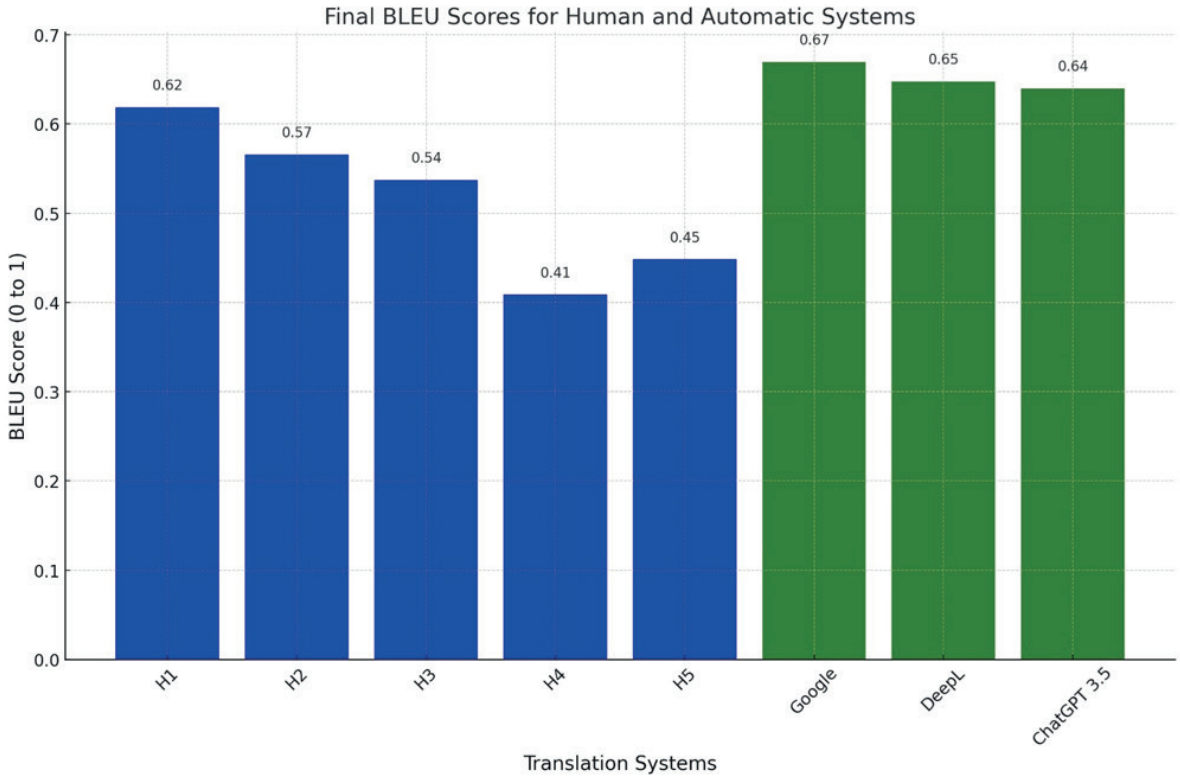
4. Analysis

There were no errors received from BLEU+ in the final analysis, and every n-gram of the candidates was correctly analysed. The final BLEU score of the human and automatic systems was recorded:

A discernible difference in correlation between human and automatic systems is evident in the final BLEU scores, where automatic systems achieved an average score higher than that of human

systems (Figure 1). Initially, the correlation results for each n-gram align with the literature review, showing a significant but expected decrease from n-gram 1 to n-gram 4 in all systems. Comparing the 4-gram results of humans with those in Papineni's (2002) study reveals a higher correlation percentage in all candidates of this research, with each candidate exceeding 30% correlation in this n-gram, unlike Papineni (2002), where correlations below 30% were obtained for the 4-gram. Notably, the correlation percentage of automatic systems is above 50% for both candidates in the 4-gram. The focus here is on the disparity between human and automatic systems' correlation. In Papineni's (2002) research, automatic translators exhibited significantly lower correlation than human systems, especially in the 4-gram, where automatic systems did not reach 10% correlation.

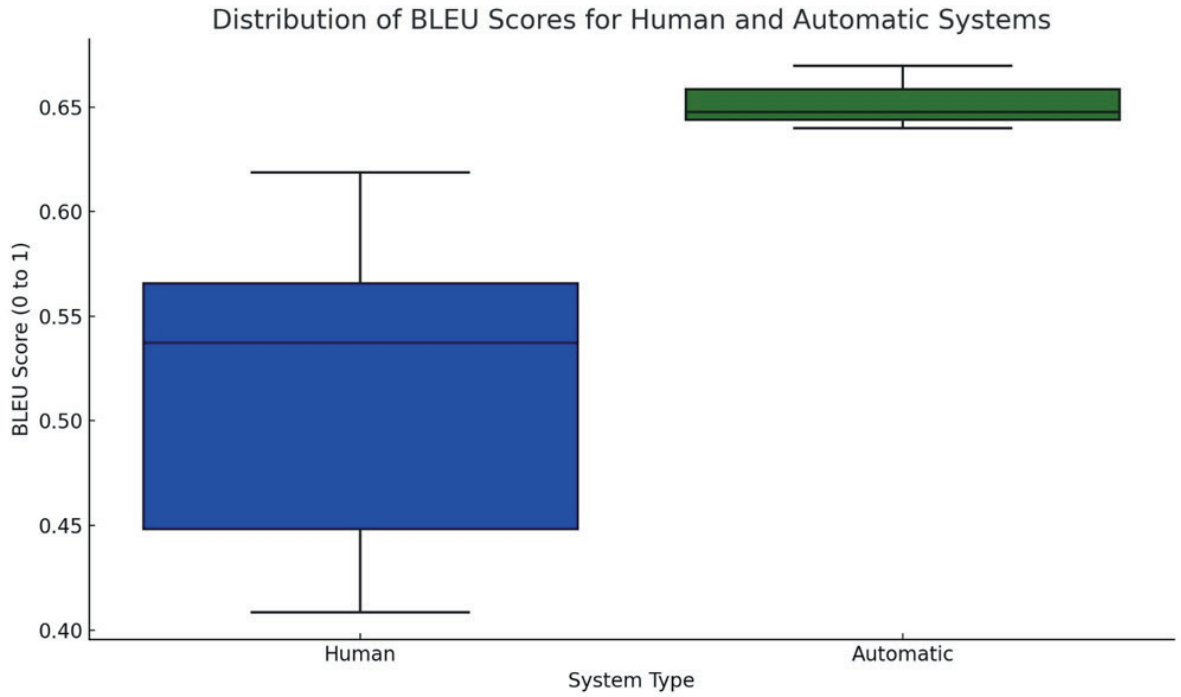
Figure 1
Final BLEU scores for Human and Automatic Translations



When analysing the distribution of BLEU scores (Figure 2), the results are more intriguing: Human systems achieved an average score of approximately 0.534 with the lowest being around 0.42 (candidate H5) and the highest approximately 0.64 (candidate H1). Automatic systems achieved an average BLEU score of about 0.67. These values not only contradict the hypothesis based on the literature review but also diverge from the results in Papineni (2002), where BLEU scores of automatic systems were lower than those of human systems across all presented values. Despite following a similar methodological line, this research presents differences from Papineni (2002). The most notable is the time elapsed between studies; Papineni's research, conducted 16 years prior, does not reflect the capability and modernity of today's automatic translators. The 2002 scores used candidates translated by automatic translators lacking the computational capacity to use NMT as effectively as today. According to both Papineni's (2002) results and this research, there is a BLEU

score discrepancy even when different automatic systems translate the same text. This information could have elucidated the difference in BLEU scores between human and automatic systems in this study.

Figure 2
Distribution of BLEU Scores for Human and Automatic Systems



Methodologically, there are differences too: Papineni (2002) used 500 sentences from 40 general news articles for translation from Chinese to English, highlighting a significant difference in the total number of structures analysed and the language pair. However, the specific use of news articles stands out. Journalistic texts generally present information differently than, for example, a scientific publication. They often reiterate information using lexical or stylistic changes to make the text more attractive and connotative. This variety of expression is typically avoided in texts requiring denotative expression, like scientific texts (De García & Pérez, 2011). Such differences can result in varied expression forms when translating, thus lower BLEU scores without sufficient references when using the BLEU metric. Papineni (2002) addresses this by presenting results of human candidates versus 2 and 4 references. The difference in BLEU scores was less when comparing human candidates with 2 references. Notably, Papineni (2002) does not mention the specific automatic translators used to create the candidates. This information could be useful for understanding how the automatic translator rendered the translation or if it concerns a translator not evaluated in this research.

Another crucial factor is information on who conducted the candidate translations in the case of humans. Papineni (2002) mentions using two individuals with very different backgrounds but does not specify these individuals' backgrounds in detail. Comparing the results of this research with more recent studies, such as those by Google (Wu *et al.*, 2016) and DeepL (DeepL, 2017), is also intriguing. In Google's case, automatic translators from English to French and German were

analysed. Interestingly, human translations were not evaluated with the BLEU metric for comparison. Instead, human presence was via judges assessing both human and automatic sentences (Wu *et al.*, 2016). Furthermore, the research evaluated the improvement of GNMT systems; however, it was human judges who attested to this improvement, not BLEU scores (Wu *et al.*, 2016). DeepL's (2017) published results are not transparent enough for a true comparison with this research. They report not releasing specific details for the time being (DeepL, 2017). Nonetheless, a comparison of BLEU scores among automatic systems is possible. The results obtained do not match the information published by DeepL (2017), where they claimed their research's BLEU scores were higher than other translators like Google Translate. Contrarily, in this research, Google Translate achieved a higher BLEU score than DeepL.

Finally, the analysis of candidates using iBLEU yielded the following result: Candidate H5 had the lowest BLEU score among all human candidates. In this case, the sentences receiving the lowest scores correspond to segments 3 and 14 of the text. In segment 3, the candidate's translation of 'optimum target blood pressure' (term from the original document) to "(la) óptima presión arterial" is noteworthy. Other variations of this term were seen in other candidates, but most did not match this particular referent's variations (Mikros & Boumparis, 2022; Colman *et al.*, 2021). The presence of numerous terms in the original text means that if a candidate's term does not match a referent's term repeatedly, a noticeable decrease in the final BLEU score will be observed. The section 14 of the candidate is also notable as it received the lowest score. In this case, a more literal translated structure was significantly penalised, especially in the translation of 'was safe' (original) to "se considera segura" or, as in other candidates, "fue segura".

Candidate H1 had the highest BLEU score among all human candidates. Here, the segment with the lowest score was segment 3, as with candidate H5. This case also shows no correlation between the term 'optimum target blood pressure' (original) translated as "(el) objetivo óptimo de la presión arterial" in the candidate and the referent's version. Regarding automatic translators, specifically the analysis of Google Translate, the lowest score was not in segment 3 but in segment 10. The omission of 'a' present in the referent, resulting in a notable change in the translation structure, was highlighted. This omission also occurred in the DeepL candidate. Google Translate was the only candidate to achieve a perfect correlation score in a structure of more than one word (segment 20). Additionally, the closeness in correlation of segment 17 compared to the referent is remarkable.

5. Conclusions

There is a clear divergence from the results of similar studies, yet a definitive conclusion on the cause of this difference cannot be reached without more in-depth analysis. Despite the need for further research to elucidate this issue, speculative insights can be drawn. The difference in typologies used in these studies could be an important factor affecting the final BLEU scores. As Munday (2011) and O'Brien (2012) suggest, the use of a typology with a higher prevalence of terminology results in translations with a more uniform lexicon. This uniformity is mirrored in references with less lexical variance, impacting the score if the candidate fails to correlate a given term with a reference and repeats that term throughout the translation. The context in which human candidates produced their translations is crucial. As Chesterman (2009) and Venuti (2017) highlight, while human translators are irreplaceable in high-stakes scenarios, these human candidates are not yet professional translators; they are students. The stress, inexperience, and

time constraints, as identified by Inghilleri (2005), may have influenced their translation quality, potentially leading to errors affecting the final BLEU score.

The disparity in BLEU scores of automatic translators from previous research underlines the need to expand this investigation. Future studies could explore the use of different metrics, as Papineni (2002) and Lavie (2011) have suggested, or include human judges who are professional translators. A deeper comparison between human and automatic systems, as Callison-Burch et al. (2006) argue, could provide more insights into translational differences. Furthermore, employing various typologies in future studies, as per Gouadec (2010), could quantify the potential BLEU score differences based on typological variance. The expansion of such studies using software assistance, as per the suggestions of Papineni (2002) and Tantuğ & Oflazer (2007), will further help determine its viability. The higher BLEU scores in automatic translators, as observed in this study, might be a direct consequence of improvements in systems as suggested by Uszkoreit (2017) and Gouws & Dehghani (2018). However, considering the longstanding relevance of this metric, other metrics, as Banerjee *et al.* (2005) propose, might more accurately represent the differences or similarities between human and automatic translations. Lastly, the identification and transparency of variables are paramount in such studies. AT BLEU scores would have been more insightful had there been more depth to the information about the texts and translators involved, as per Snover *et al.* (2006). This information could significantly contribute to a detailed analysis of the results. Despite the ongoing debate about the use of automatic translators, their significant advancements in quality and precision are undeniable. Yet, as Pym (2010) suggest, there is considerable potential for progress in the methods used to quantify these advancements, whether through metrics or different evaluation types. This will bring us closer to comprehending the full potential of automatic translators in society, the country, and the discipline.

References

Abu-Zahra, M. J., & Shayeb, A. S. (2022). Do mobile translation apps enhance or hinder translation trainees' linguistic competence: The case study of translation students at Birzeit University. *Journal of Language and Linguistic Studies, 18*(4).

Ahammad, S. H., Kalangi, R. R., Nagendram, S., Inthiyaz, S., Priya, P. P., Faragallah, O. S., Mohammad, A., Mahmoud, M. A., & Rashed, A. N. Z. (2024). Improved neural machine translation using Natural Language Processing (NLP). *Multimedia Tools and Applications, 83*(13), 39335-39348.

Ali, M. A. (2020). Quality and machine translation: An evaluation of online machine translation of English into Arabic texts. *Open Journal of Modern Linguistics, 10*(5), 524-548.

Brownlee, J. (2017, november 20). A Gentle Introduction to Calculating the BLEU Score for Text in Python [Blog post]. *Machine Learning Mastery*.

Callison-Burch, C., Koehn, P., & Osborne, M. (2006, june). Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference* (pp. 17-24).

Chesterman, A. (2009). The name and nature of translator studies. *HERMES-Journal of Language and Communication in Business*, (42), 13-22.

Colman, T., Fonteyne, M., Daems, J., & Macken, L. (2021). *It's all in the eyes: An eye tracking experiment to assess the readability of machine translated literature*. 31st Meeting of Computational Linguistics in The Netherlands (CLIN 31).

DeepL Translator. (2017). *DeepL Translator*.

De García, J., & Pérez, R. (2011). Reflexiones y recomendaciones sobre buenas prácticas en la traducción científica y técnica. *Tecnologí@ y Desarrollo, 9*, 1-9.

Dew, K. N., Turner, A. M., Choi, Y. K., Bosold, A., & Kirchhoff, K. (2018). Development of machine translation technology for assisting health communication: A systematic review. *Journal of Biomedical Informatics, 85*, 56-67.

Dodos, A. (2017). *Phrase-based translation model*.

Google Translate. (2018). *GNMT en acción*.

Gouadec, D. (2010). *Quality in translation. In Handbook of Translation Studies* (pp. 270-275). John Benjamins.

Gouws, S., & Dehghani, M. (2018, august 15). Moving Beyond Translation with the Universal Transformer. *Google AI Blog*.

Hartley, T. (2007). *MT Evaluation – challenges and techniques*.

Hernández, R., & Mendoza, C. (2018). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta*. McGraw Hill Education.

Hernández-Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). Metodología de la investigación (6.ª ed.). McGraw Hill.

Inghilleri, M. (2005). The sociology of Bourdieu and the construction of the 'object' in translation and interpreting studies. *The translator*, *11*(2), 125-145.

Khoong, E. C., & Rodriguez, J. A. (2022). A research agenda for using machine translation in clinical medicine. *Journal of General Internal Medicine, 37*(5), 1275-1277.

Lavie, A. (2011). *Statistical MT with Syntax and Morphology: with Syntax and Morphology: Challenges and Some Solutions*.

Lee, J. (2022). Comparing student self-assessment and teacher assessment in Korean-English consecutive interpreting: Focus on fidelity and language. *INContext: Studies in Translation and Interculturalism, 2*(3).

Loboda, K., & Olga Mastela. (2023). Machine translation and culture-bound texts in translator education: a pilot study. In G. Massey, M. Piotrowska and M. Marczak (Eds.), *(Re-)profiling T&I education: meeting evolution with innovation* (pp. 503–525).

Madnani, N. (2011). *iBLEU: Interactively Debugging & Scoring Statistical Machine Translation Systems*. Proceedings of the Fifth IEEE International Conference on Semantic Computing.

Madnani, N. (2011). *Interfaz de la sección de visualización en iBLEU*.

Mikros, G., & Boumparis, D. (2022). *Cross-linguistic authorship attribution and author profiling. Is machine translation the solution?*

Munday, J. (2011). *Introducing translation studies*. Routledge.

O'Brien, S. (2012). Translation as human–computer interaction. *Translation spaces*, *1*(1), 101-122.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). *BLEU: a method for automatic evaluation of machine translation*. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, United States.

Pérez, C. (2013). *Post-edición en el contexto de la Traducción Controlada*.

Popović, M. (2020). Relations between comprehensibility and adequacy errors in machine translation output. *Association for Computational Linguistics (ACL)*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *In Proceedings of Association for Machine Translation in the Americas*.

Tantuğ, A., & Oflazer, K. (2007). *BLEU+: a tool for fine-grained BLEU computation*. Proceedings of the Sixth International Language Resources and Evaluation, Marrakech, Morocco.

Tantuğ, A. (2007). *Interfaz predeterminada de BLEU+*.

Uszkoreit, J. (2017). *Transformer: A Novel Neural Network Architecture for Language Understanding*.

Van Gerven, M. (2017). Computational foundations of natural intelligence. *Frontiers in Computational Neuroscience, 11*.

Venuti, L. (2017). *The translator's invisibility: A history of translation*. Routledge.

Vieira, L. N., O'Hagan, M., & O'Sullivan, C. (2021). Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society, 24*(11), 1515-1532.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cap, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gows, S., Kato, Y., Kudo, T., Kazawa, H.,... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv, 1*.

Zappatore, M., & Ruggieri, G. (2023). Adopting machine translation in the healthcare sector: A methodological multi-criteria review. *Computer Speech & Language*, *84*, 1-46. https://doi.org/10.1016/j.csl.2023.101582

Authors contribution

Felipe von Hausen has participated in the preparation, design of the research, writing, and critical revision of the article, and gives approval to the version published in the journal. Cristóbal Muñoz Urtubia and Carlos Contreras Aedo have participated in the preparation of the study, design of the research, writing, data collection, and analysis.

Acknowledgements

We, Felipe von Hausen and Cristóbal Muñoz, thank Dr. Carlos Contreras Aedo, linguist and academic of the Translation and Interpretation in Foreign Languages programme at the Universidad de Concepción, Chile, for his valuable contributions. We would like to express our profound gratitude to Dr. Carlos Contreras Aedo, who has sadly passed away. His contributions and unwavering support have been invaluable, and he will be greatly missed. We owe him a great deal for his guidance and friendship. Rest in peace, dear professor, dear colleague, and dear friend.

Funding

The research was conducted without funding.

Conflict of interest

The authors declare no conflict of interest.

Correspondence: fvonhausen@udla.cl

Authors information

Felipe von Hausen is a translator of Spanish, German and English from the Universidad de Concepción, Chile. Teacher of German Language from the Universidad de Talca, Chile. He holds a Master's in Higher Education and a Master's in Applied Linguistics, both from the Catholic University of the Most Holy Conception, Chile. Academic at the Faculty of Communications and Arts, Universidad de las Américas, Concepción, Chile. His research area is experimental psycholinguistics, with a special emphasis on the processing of syntax and lexicon in L2.

Cristóbal Muñoz is a translator of Spanish, English and German from the Universidad de Concepción, Chile. Japanese Teacher, Master's in Spanish Language Teaching from the Universidad de Barcelona, Spain, and Software Developer. Independent Translator and Teacher in Australia. His research areas are translation studies and L2 learning.

Carlos Contreras was a distinguished linguist and translator with a diverse academic background. He earned a Bachelor's degree in Education, a degree in English Language Teaching, and a professional degree in Translation from the University of Concepción (UdeC), specialising in German, French, English, and Spanish. Dr. Contreras pursued advanced studies, obtaining a Master's and PhD in Linguistics from UdeC. Additionally, he completed postgraduate studies in Italian Language and Culture, and had proficiency in Swedish. He taught Italian and significantly contributed to the field of translation education, particularly in German and French, shaping the curriculum of the Translation and Interpretation Programme at UdeC.