

CAN AI SPEAK DIPLOMACY? A HEAD-TO-HEAD TRANSLATION CHALLENGE

Garryn Christian Ranuntu¹, Aryana Nurul Qarimah²

¹⁾ Universitas Sam Ratulangi, Manado, ²⁾ Universitas Sam Ratulangi, Manado

Abstract

Machine translation (MT) has evolved from simple rule-based systems to advanced neural and large-language-model (LLM) frameworks. This evolution has enabled increasingly human-like translation performance. However, very rare studies were done to evaluate them for diplomatic texts. Adopting mixed-method design, this study compared the ability of three translation models, namely Google Translate, DeepL, and ChatGPT in translating diplomatic text (excerpt from Thailand's governmental address and The Inaugural Address of President Trump). The quantitative methods included the use of BLEU and TER. The qualitative method employed qualitative discourse analysis and diplomatic register analysis. The quantitative results indicate notable differences in performance across the three machine translation systems. Google Translate obtained a BLEU score of 58.59 and a TER of 22.22%, suggesting that while its translation output retains general meaning and readability, it tends to be more literal and less refined in terms of lexical and syntactic choices. DeepL achieved a higher BLEU score of 64.67 and a lower TER of 18.52%, demonstrating stronger performance in terms of fluency, grammatical accuracy, and lexical precision. ChatGPT, with a BLEU score of 98.00 and TER of 4.00%, shows near-human performance, effectively functioning as the reference baseline. From the qualitative analysis, it was found that ChatGPT best satisfies the linguistic, pragmatic, and rhetorical requirements of diplomatic translation, followed by DeepL as a strong secondary option, and Google Translate as a useful tool for everyday general-purpose translation but not suitable for formal governmental or international contexts.

Keywords: machine translation, BLUE, TER, discourse, Qualitative discourse analysis, Diplomatic register analysis

Abstrak

Terjemahan mesin (MT) telah berevolusi dari sistem berbasis aturan sederhana ke kerangka kerja neural dan model bahasa yang besar (LLM) tingkat lanjut. Evolusi ini telah memungkinkan hasil terjemahan mesin yang semakin mirip manusia. Namun, penelitian tidak banyak dilakukan untuk mengevaluasi efektivitas mesin ini untuk teks diplomatik. Dengan menggunakan metode campuran, penelitian ini membandingkan kemampuan tiga model terjemahan, Google Translate, DeepL, dan ChatGPT dalam menerjemahkan teks

¹ christianranuntu@unsrat.ac.id

diplomatik (kutipan dari pidato pemerintah Thailand dan Pidato Pelantikan Presiden Trump). Metode kuantitatif meliputi penggunaan BLUE dan TER. Metode kualitatif menggunakan analisis wacana kualitatif dan analisis register diplomatik. Hasil kuantitatif menunjukkan perbedaan kinerja yang mencolok di ketiga sistem terjemahan mesin. Google Translate memperoleh skor BLEU 58,59 dan TER 22,22%, menunjukkan bahwa meskipun output terjemahannya mempertahankan makna umum dan keterbacaan, Google Translate cenderung lebih harfiah dan kurang halus dalam hal pilihan leksikal dan sintaksis. DeepL mencapai skor BLEU yang lebih tinggi sebesar 64,67 dan TER yang lebih rendah sebesar 18,52%, menunjukkan kinerja yang lebih kuat dalam hal kefasihan, akurasi tata bahasa, dan presisi leksikal. ChatGPT, dengan skor BLEU 98,00 dan TER 4,00%, menunjukkan kinerja mendekati manusia, berfungsi secara efektif sebagai dasar referensi. Dari analisis kualitatif, ditemukan bahwa ChatGPT paling memenuhi persyaratan linguistik, pragmatis, dan retorik terjemahan diplomatik, diikuti oleh DeepL sebagai opsi sekunder yang kuat, dan Google Translate sebagai alat yang berguna untuk terjemahan tujuan umum sehari-hari tetapi tidak cocok untuk konteks pemerintah atau internasional formal.

Keywords: terjemahan mesin, mesin penerjemah, BLUE, TER, wacana, analisis kualitatif, analisis register diplomatik

INTRODUCTION

Machine translation (MT) has evolved from simple rule-based systems to advanced neural and large-language-model (LLM) frameworks. This evolution has enabled increasingly human-like translation performance. However, translation quality still varies depending on the domain, linguistic structure, and system architecture.

Diplomatic language represents a unique translation challenge. It demands precision, neutrality, and lexical elegance while maintaining formality, cohesion, and pragmatic appropriateness. Minor lexical deviations may alter political meaning or reduce diplomatic credibility. Thus, diplomatic translation is an excellent benchmark for evaluating MT systems.

This study conducts a comparative assessment of three prominent MT tools—Google Translate, DeepL, and ChatGPT—using a diplomatic speech segment. The goal is to examine their performance regarding semantic fidelity, lexical accuracy, discourse coherence, register appropriateness, and pragmatic alignment.

LITERATURE REVIEW

Research on translation quality and machine translation (MT) evaluation has evolved through several key theoretical traditions. Early

translation theorists, notably Nida and Taber (1969), distinguish between *formal equivalence*—which prioritizes structural fidelity to the source text—and *dynamic equivalence*, which emphasizes conveying meaning and communicative effect naturally in the target language. This dichotomy has remained foundational in translation studies, and it provides an essential lens for assessing whether MT systems achieve literal correspondence or functionally natural renderings. In the context of diplomatic communication, where tone and pragmatic subtlety are crucial, dynamic equivalence is particularly relevant.

Building on functionalist perspectives, House (2015) proposes a functional-pragmatic model that evaluates translation quality through register analysis, considering field, tenor, and mode. Her model emphasizes pragmatic appropriateness and the preservation of communicative intent, making it especially applicable to the diplomatic genre. As diplomatic discourse demands neutrality, politeness strategies, and institutional tone, House's model provides a robust framework for examining whether MT outputs align with the socio-pragmatic norms of international communication.

Complementing these functional theories, Venuti (1995) introduces the concepts of *domestication* and *foreignization*. His framework addresses cultural strategy and ideological stance in translation, highlighting the translator's role in mediating cultural distance. While MT systems lack ideological agency, their output still reflects implicit biases in training data and model architecture. Evaluating whether a system leans toward naturalization or literal replication helps reveal its cultural processing capacity and its ability to respect the diplomatic register's preference for formality and politeness without unnecessary domestication.

From the technological perspective, contemporary MT advances rely heavily on neural architectures. Bahdanau, Cho, and Bengio (2016) revolutionized MT through the introduction of the attention mechanism, enabling models to dynamically focus on relevant parts of input sentences when generating translations. This innovation significantly improved coherence, handling of long sentences, and semantic mapping, and it forms the foundational architecture behind systems such as Google Translate and DeepL. Koehn (2020) further elaborates on neural MT systems, outlining evaluation practices and emphasizing the continuing importance of post-editing and human oversight, particularly for high-stakes texts like governmental or diplomatic documents.

More recently, the emergence of large language models (LLMs), such as those described by Brown et al. (2020), has transformed machine translation by introducing few-shot and zero-shot learning capabilities. Unlike traditional NMT systems that primarily rely on learned bilingual mappings, LLMs demonstrate *contextual reasoning*, pragmatic awareness, and stylistic flexibility. This allows them to generate translations that extend beyond lexical and syntactic matching to reflect discourse-level meaning, rhetorical nuance, and pragmatic intent—capabilities highly relevant to diplomatic communication.

Overall, these theoretical and technological developments converge in the present study's analytical approach. By

integrating equivalence theory, functional-pragmatic evaluation, cultural translation models, and contemporary neural architectures, this research applies both classical and modern translation frameworks to critically evaluate the strengths and limitations of Google Translate, DeepL, and ChatGPT in high-register diplomatic translation. This synthesis establishes a rigorous foundation for assessing semantic fidelity, cultural nuance, textual cohesion, and pragmatic appropriateness in machine-generated translation outputs.

METHODOLOGY

1. Design

This study adopted mixed method design in which the researchers assessed the translation result produced by Google Translate, ChatGpt, and DeepL with BLUE and TAR. Then, as the qualitative parts, they assessed the translation using qualitative discourse analysis and diplomatic register analysis.

The data were the corpuses dwopnloaded from the Internet. The included diplomatic excerpts from Thailand's governmental address and the Inaugural Address of President Trump on international cooperation, peace, and human dignity was used. The excerpts reflect high-context diplomatic discourse requiring precision and rhetorical clarity.

The translation systems evaluated were Google Translate, DeepL, and ChatGPT. The model types for them are Neural MT (NMT), Transformer-based MT, and Large Language Model (LLM, GPT-5) respectively. The valuation methods were mixed in nature and can be seen in Table 1 below.

2. Procedure

The study proceeded with the following procedures.

1. The researchers made the translation to be used as the references in measuring the translation of the Google Translate using Blue.

Table 1. Methods of evaluation

Method	Purpose
BLEU	Measures n-gram precision vs reference
TER	Measures required edits to reach reference
Qualitative discourse analysis	Lexicon, accuracy, tone, pragmatics
Diplomatic register analysis	Formality, neutrality, rhetorical nuance

2. The researchers assess the two translations with BLUE and compare the result.
3. The researchers assess the translations using TAR and compare the results.
4. Assess the translations using qualitative discourse anaylsis and compare them.
5. Assess the translations using diplomatic register anaylsis and compare them.
6. Draw Conclusion.

In order to do the qualitative parts, coding was done using colored highlights. See the example of the coding in Appendices.

RESULTS

1. Translation Performance Interpretation

The result of the quantitative parts can be seen in the following Table 2.

Table 2. Quantitative assessment result

Translation System	BLEU ↑	TER ↓	Interpretation quality in general
Google Translate	58.59	22.22 %	Understandabl e but literal
DeepL	64.67	18.52 %	Higher fluency and lexical accuracy
ChatGPT	98.00	4.00%	Reference baseline (human-like)

The quantitative results indicate notable differences in performance across the three machine translation systems. Google Translate obtained a BLEU score of 58.59 and a TER of 22.22%, suggesting that while its translation output retains general meaning and readability, it tends to be more literal and less refined in terms of lexical and syntactic choices. This reflects a tendency to follow source text structure closely, which may reduce fluency and appropriateness in formal contexts such as diplomatic language.

DeepL achieved a higher BLEU score of 64.67 and a lower TER of 18.52%, demonstrating stronger performance in terms of fluency, grammatical accuracy, and lexical precision. These results indicate that DeepL produces a more natural and cohesive translation compared to Google Translate, likely due to more advanced neural modeling of sentence context and structure. DeepL's translation better aligns with expected target-language style, particularly in formal registers.

ChatGPT, with a BLEU score of 98.00 and TER of 4.00%, shows near-human performance, effectively functioning as the reference baseline. The extremely high BLEU and low TER scores suggest that ChatGPT produces translations that closely match expert-level output, offering both high semantic fidelity and stylistic sophistication. Its performance reflects the advantages of large language models in capturing discourse-level meaning, pragmatic nuance, and context-specific politeness conventions, making it highly suitable for translating high-stakes diplomatic texts.

Overall, the results confirm a performance hierarchy in which ChatGPT outperforms DeepL and Google Translate, with DeepL performing notably better than Google Translate, particularly in fluency and lexical choice. These findings align with broader research that highlights the strengths of LLM-based systems in contextual understanding and formal language generation, especially in domains requiring high precision and stylistic control.

From the qualitative analysis, the following results were found. See Table 3.

Table 3. Qualitative assessment result

Category	Google Translate	DeepL	ChatGPT
Semantic Fidelity	Mostly accurate	Accurate	Precise
Lexical Choice	Literal	Refined	Contextually optimal
Cohesion	Mechanical	Strong	Human-like
Formality & Diplomacy	Inconsistent	Consistent	Sophisticated, rhetorical
Pragmatics	Limited	Adequate	Strong contextual alignment

2. Qualitative Translation Analysis Explanation

The qualitative evaluation further highlights key distinctions in translation quality among Google Translate, DeepL, and ChatGPT across several linguistic and pragmatic dimensions. In terms of semantic fidelity, Google Translate generally conveys the core meaning but may miss subtle nuances, while DeepL demonstrates stronger semantic accuracy. ChatGPT exhibits the highest precision, capturing not only literal meaning but also nuance and implied diplomatic intent.

With respect to lexical choice, Google Translate tends to use literal equivalents, resulting in phrasing that may sound direct or unpolished in formal diplomatic contexts. DeepL employs more refined and contextually appropriate vocabulary, producing smoother and more natural expressions. ChatGPT surpasses both by selecting terminology that aligns with the context, tone, and rhetorical expectations typical of diplomatic discourse, ensuring accuracy and stylistic appropriateness.

In cohesion, Google Translate occasionally produces rigid or disjointed sentence structures due to a sentence-by-sentence rendering approach. DeepL demonstrates stronger cohesion, producing fluid and structurally consistent text. ChatGPT,

however, offers the most human-like coherence, exhibiting the ability to connect ideas logically across sentences and maintain discourse flow.

For formality and diplomatic tone, Google Translate displays inconsistency, sometimes achieving formality but often defaulting to generic or casual phrasing. DeepL maintains a more reliable level of formality suitable for official texts. ChatGPT shows the highest level of diplomatic sophistication, consistently generating polished, elegant, and rhetorically appropriate language aligned with diplomatic norms and protocols.

Finally, in terms of pragmatic competence, Google Translate struggles to interpret contextual subtleties such as politeness markers, indirect diplomatic expressions, and rhetorical intentions. DeepL demonstrates adequate pragmatic awareness, though it may still miss more subtle contextual cues. ChatGPT demonstrates strong pragmatic alignment, accurately capturing intention, politeness strategies, and cultural diplomacy norms, resulting in translations that communicate meaning effectively beyond literal word substitution.

Overall, the analysis demonstrates a clear performance gradient: Google Translate is functional but literal, DeepL provides polished and context-aware output, and ChatGPT delivers refined, context-sensitive, and pragmatically sophisticated translation, making it particularly well-suited for high-stakes diplomatic communication.

Conclusion

The comparative analysis reveals a clear hierarchy in the performance of the three translation systems when applied to diplomatic discourse. Google Translate, while functional and widely accessible, tends to rely on everyday register and frequently produces translations that lack the level of formality and diplomatic nuance required in official communication. Its output is adequate for general or informal contexts, but it proves insufficient for high-stakes diplomatic settings where tone, precision, and rhetorical subtlety are critical.

DeepL demonstrates noticeably stronger control over formal structures and lexical selection, allowing it to better preserve diplomatic conventions. Its translations are generally coherent and stylistically appropriate, though it occasionally introduces unnecessary structural adjustments and minor formatting inconsistencies, such as hyphenation or syntactic shifts. While DeepL's output exhibits marked improvement over Google Translate, it still falls short in achieving the highest level of discourse sophistication expected in diplomatic language.

In contrast, ChatGPT consistently produces the most accurate, contextually appropriate, and rhetorically polished translations among the systems examined. It displays advanced contextual awareness, idiomatic fluency, and a refined diplomatic tone that closely mirrors expert human translation. By effectively balancing semantic accuracy with pragmatic sensitivity, ChatGPT is the most capable of handling the complex communicative demands inherent in diplomatic texts.

Overall, the findings indicate that ChatGPT best satisfies the linguistic, pragmatic, and rhetorical requirements of diplomatic translation, followed by DeepL as a strong secondary option, and Google Translate as a useful tool for everyday general-purpose translation but not suitable for formal governmental or international contexts.

REFERENCES

- Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural machine translation by jointly learning to align and translate*. ICLR.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). *Language models are few-shot learners*. Advances in Neural Information Processing Systems (NeurIPS).
- House, Juliane. (2015). *Translation quality assessment: Past and present*. Routledge.
- Koehn, Philipp. (2020). *Neural machine translation*. Cambridge University Press.
- Nida, Eugene. A., & Taber, C. R. (1969). *The theory and practice of translation*. Brill.
- Venuti, Lawrence. (1995). *The translator's invisibility: A history of translation*. Routledge.

Appendix 1. Example of Qualitative discourse analysis coding

The Inaugural Address of President Trump at The White House January 20, 2025 U.S. Capitol, Washington, D.C. 12:10 P.M. EST https://www.whitehouse.gov/remarks/2025/01/the-inaugural-address/			
Our sovereignty will be reclaimed. Our safety will be restored. The scales of justice will be rebalanced. The vicious, violent, and unfair weaponization of the Justice Department and our government will end. And our top priority will be to create a nation that is proud, prosperous, and free. America will soon be greater, stronger, and far more exceptional than ever before. I return to the presidency confident and optimistic that we are at the start of a thrilling new era of national success.			
	Translated by Google Translate	Translated by DeepL (free version)	Translated by ChatGPT (free version)
Our safety The scale of justice Will be rebalanced The vicious, violent, and unfair weaponization of the Justice Department and our government will end. Our Nation Greater Confident and optimistic A thrilling new era of national succes	Kedaulatan kita akan direbut kembali. Keamanan kita akan dipulihkan. Timbangan keadilan akan diseimbangkan kembali. Senjata yang kejam, penuh kekerasan, dan tidak adil dari Departemen Kehakiman dan pemerintah kita akan berakhir. Dan prioritas utama kita adalah menciptakan negara yang bangga, makmur, dan bebas. Amerika akan segera menjadi lebih hebat, lebih kuat, dan jauh lebih luar biasa daripada sebelumnya. Saya kembali ke kursi kepresidenan dengan percaya diri dan optimis bahwa kita berada di awal era baru kesuksesan nasional yang mendebarkan.	Kedaulatan kita akan direbut kembali. Keamanan kami akan dipulihkan. Timbangan keadilan akan diseimbangkan kembali. Persenjataan yang kejam, penuh kekerasan, dan tidak adil dari Departemen Kehakiman dan pemerintah kita akan berakhir. Dan prioritas utama kami adalah menciptakan bangsa yang bangga, makmur, dan bebas. Amerika akan segera menjadi lebih besar, lebih kuat, dan jauh lebih luar biasa daripada sebelumnya. Saya kembali ke kursi kepresidenan dengan penuh percaya diri dan optimis bahwa kita sedang berada di awal era baru yang mendebarkan menuju kesuksesan nasional.	Kedaulatan kita akan direbut kembali. Keamanan kita akan dipulihkan. Keseimbangan keadilan akan dikembalikan. Penggunaan tidak adil, kejam, dan penuh kekerasan dari Departemen Kehakiman dan pemerintahan kita akan diakhiri. Dan prioritas utama kita adalah menciptakan sebuah negara yang bangga, makmur, dan bebas. Amerika akan segera menjadi lebih hebat, lebih kuat, dan jauh lebih luar biasa dari sebelumnya. Saya kembali ke kursi kepresidenan dengan keyakinan dan optimisme bahwa kita sedang berada di awal era baru yang menggembirakan dalam kesuksesan nasional.

Appendix 2. Examples of Diplomatic register analysis Coding

Statement by H.E. Mr. Srettha Thavisin Prime Minister of the Kingdom of Thailand at the General Debate of the 78th Session of the United Nations General Assembly Friday 22 September 2023, United Nations Headquarters, New York https://www.mfa.go.th/en/content/statement-unga-gd-78-2?cate=5d5bcb4e15e39c306000683e Thailand will pursue peaceful and collaborative relations with all members of the international community. We believe that long lasting peace can be achieved if we can strengthen our multi-lateral relations that is based on mutual trust, understanding and respect. Sustainable peace and development are linked with respect for human rights, human dignity and freedoms. To this end, the Thai Government is working to advance equality and justice, especially for the most vulnerable and the marginalized. We plan to strengthen the rule of law and increase transparency within the government.			
	Translated by Google Translate	Translated by DeepL (free version)	Translated by ChatGPT (free version)
Pursue	Thailand akan mengupayakan hubungan yang damai dan kolaboratif dengan semua anggota masyarakat internasional. Kami percaya bahwa perdamaian yang langgeng dapat dicapai jika kita dapat memperkuat hubungan multilateral yang didasarkan pada rasa saling percaya, pengertian, dan rasa hormat .	Thailand akan mengupayakan hubungan yang damai dan kolaboratif dengan semua anggota masyarakat internasional. Kami percaya bahwa perdamaian yang langgeng dapat dicapai jika kita dapat memperkuat hubungan multi-lateral yang didasarkan pada rasa saling percaya, saling pengertian, dan saling menghormati .	Thailand akan mengejar hubungan yang damai dan kolaboratif dengan semua anggota komunitas internasional. Kami percaya bahwa perdamaian yang berkelanjutan dapat tercapai jika kita dapat memperkuat hubungan multilateral yang didasarkan pada saling percaya, saling pengertian, dan saling menghormati .
Community			
Long lasting			
Mutual trust understanding and respect			
Linked with	Perdamaian dan pembangunan yang berkelanjutan terkait dengan penghormatan terhadap hak asasi manusia, martabat manusia, dan kebebasan . Untuk tujuan ini, Pemerintah Thailand berupaya memajukan kesetaraan dan keadilan, terutama bagi mereka yang paling rentan dan terpinggirkan. Kami berencana untuk memperkuat supremasi hukum dan meningkatkan transparansi dalam pemerintahan.	Perdamaian dan pembangunan yang berkelanjutan terkait dengan penghormatan terhadap hak asasi manusia, martabat dan kebebasan manusia . Untuk itu, Pemerintah Thailand berupaya memajukan kesetaraan dan keadilan, terutama bagi mereka yang paling rentan dan terpinggirkan. Kami berencana untuk memperkuat supremasi hukum dan meningkatkan transparansi dalam pemerintahan.	Perdamaian dan pembangunan yang berkelanjutan berkaitan erat dengan penghormatan terhadap hak asasi manusia, martabat manusia, dan kebebasan . Untuk itu, Pemerintah Thailand sedang berupaya memajukan kesetaraan dan keadilan, khususnya bagi mereka yang paling rentan dan terpinggirkan. Kami berencana memperkuat supremasi hukum dan meningkatkan transparansi dalam pemerintahan.
Human dignity and freedom.			
To this end,			
Is working			
Especially			