

From Scratch to Fine-Tuned: A Comparative Study of Transformer Training Strategies for Legal Machine Translation

Amit Barman, Atanu Mandal, Sudip Kumar Naskar

Jadavpur University, Kolkata, INDIA,

Correspondence: amitbarman811@gmail.com

Abstract

In multilingual nations like India, access to legal information is often hindered by language barriers, as much of the legal and judicial documentation remains in English. Legal Machine Translation (L-MT) offers a scalable solution to this challenge by enabling accurate and accessible translations of legal documents. This paper presents our work for the JUST-NLP 2025 Legal MT shared task, focusing on English–Hindi translation using Transformer-based approaches. We experiment with 2 complementary strategies, fine-tuning a pre-trained OPUS-MT model for domain-specific adaptation and training a Transformer model from scratch using the provided legal corpus. Performance is evaluated using standard MT metrics, including SacreBLEU, chrF++, TER, ROUGE, BERTScore, METEOR, and COMET. Our fine-tuned OPUS-MT model achieves a SacreBLEU score of 46.03, significantly outperforming both baseline and from-scratch models. The results highlight the effectiveness of domain adaptation in enhancing translation quality and demonstrate the potential of L-MT systems to improve access to justice and legal transparency in multilingual contexts.

1 Introduction

Since India’s independence in 1947, language has remained one of the defining features and challenges of its democracy. The Constitution recognizes 22 scheduled languages, but much of the country’s legal, administrative, and judicial work continues to be conducted in English. This linguistic imbalance often leaves citizens dependent on translations to understand laws, judgments, or government notifications that affect their rights. There have been documented instances where individuals have misunderstood court proceedings or official orders simply because they were not available in their native language, an obstacle that runs counter to the ideal of “equal access to justice”.

In a multilingual democracy, ensuring that legal information is accessible to all citizens is not only a linguistic challenge but also a civic necessity. Legal texts are particularly complex, they demand precision, consistency, and adherence to jurisdiction-specific terminology. Even small translation errors can lead to misinterpretations, contractual disputes, or procedural delays. As legal materials increasingly move to online platforms, the need for accurate, scalable translation tools has become even more urgent.

Advances in Neural Machine Translation (NMT) have transformed the field of translation, enabling systems to model intricate linguistic relationships and long-range dependencies through attention mechanisms (Vaswani et al., 2017). The rise of Large Language Models (LLMs) trained on vast multilingual data has further improved translation fluency and generalization. Yet, these models often struggle in highly specialized domains like law, where vocabulary, syntax, and semantics diverge significantly from general text. Domain-specific adaptation remains essential for achieving accurate and trustworthy translations.

This paper focuses on developing Legal Machine Translation (L-MT) systems that bridge the linguistic divide in the Indian legal context. As part of the JUST-NLP 2025 Legal MT shared task¹, we investigate how Transformer-based models can be adapted for English–Hindi legal translation. We explore two strategies, training a Transformer model from scratch and fine-tuning the OPUS-MT model, to assess how domain-focused training influences translation quality.

Through this work, we aim to advance the development of reliable and inclusive Legal MT systems that make legal information accessible across languages, supporting transparency, participation, and justice in multilingual societies.

¹<https://exploration-lab.github.io/JUST-NLP/>

The key contributions of this paper are summarized as follows:

- We trained and evaluated a Transformer model from scratch on legal-domain data.
- We fine-tuned the Helsinki Opus MT for legal-domain adaptation.
- We analyzed translation robustness and domain adaptability across evaluation datasets.

2 Related Works

Machine Translation (MT) has long been one of the most prominent applications of Natural Language Processing (NLP). Early MT systems were primarily built upon sequence-to-sequence architectures using encoder–decoder frameworks. However, due to their sequential nature and reliance on recurrent neural networks, these models often struggled to capture long-range contextual dependencies effectively.

The introduction of the self-attention mechanism revolutionized MT by enabling models to capture global dependencies among tokens more efficiently. Transformer-based architectures have since become the foundation of modern MT systems, demonstrating exceptional generalization across languages and domains through large-scale multilingual pretraining. This paradigm shift has significantly improved translation fluency, adequacy, and semantic consistency.

Recent advancements in LLMs have further enhanced multilingual translation capabilities through zero-shot and few-shot learning. These pretrained multilingual models can produce reasonable translations even without explicit task-specific finetuning. However, their performance tends to degrade substantially for low-resource language pairs, where limited data hampers generalization. To address this, research has increasingly focused on fine-tuning and transfer learning strategies that enable domain and language adaptation. Techniques such as multilingual continued pretraining, cross-lingual embeddings, and parameter-efficient finetuning (e.g., adapters like LoRA (Hu et al., 2021) and QLoRA (Dettmers et al., 2023)) have proven effective in improving translation quality for low-resource scenarios. These methods balance computational efficiency with adaptability, allowing pretrained multilingual models to specialize in specific linguistic domains such as legal, medical, or conversational text.

In the Indian context, legal translation has emerged as a crucial area of research due to the nation’s linguistic diversity and the absence of a single national language. The growing need to make legal documents accessible across India’s many official languages highlights the importance of domain-specific MT systems. However, Indian languages often lack large, high-quality parallel corpora, posing challenges for training robust legal MT models (Joshi et al., 2024).

Over the past decade, several multilingual parallel corpora have been developed for Indian languages. Notable examples include Samanantar (Ramesh et al., 2022), corpus for 11 Indian languages and the corpus by Siripragada et al. (2020), which covers 10 Indian languages. Broader evaluation was also enabled by the FLORES-200 benchmark (Team et al., 2022). Other valuable resources include IndoWordNet (Kunchukuttan, 2020), PMIndia (Haddow and Kirefu, 2020), and datasets such as IITB English-Hindi (Kunchukuttan et al., 2018), BUET English-Bangla (Hasan et al., 2020), English-Tamil (Ramasamy et al., 2012), English-Odia (Parida et al., 2020), and the Mizo-English corpus (Haulai and Hussain, 2023). However, these datasets generally pertain to general-domain translation and are not tailored to the legal domain.

In contrast, the legal domain has seen relatively limited multilingual MT resources. International initiatives such as the Europarl corpus (Koehn, 2005), EUR-Lex (Baisa et al., 2016), and the UN Parallel Corpus (Ziemski et al., 2016), the Bilingwits Swiss Law Text collection (Höfler and Sugisaki, 2014) have provided valuable multilingual datasets for legal proceedings in European languages. However, these resources are largely tailored to European legal systems, linguistic structures, and translation conventions, which differ substantially from the Indian legal and linguistic context. Consequently, such corpora cannot be directly leveraged for Indian-language MT tasks, where distinct terminologies, legal frameworks, and multilingual diversity necessitate domain-specific datasets and adaptation strategies.

Within India, only a handful of initiatives have attempted to build legal-domain corpora. The Hindi–Telugu legal dataset from LTRC (Mujadia and Sharma, 2022) and the Anuvaad corpus² repre-

²<https://github.com/project-anuvaad/anuvaad-parallel-corpus>

sents early efforts, however, they lack expert validation. The recently introduced MILPaC corpus (Majhapatra et al., 2025) marks a significant advancement, offering a well-curated, expert-validated, and multilingual benchmark for legal MT. Additionally, the WMT25 Legal Domain Test Suite (Singh et al., 2025) provides a robust evaluation framework for assessing MT capabilities in English–Hindi legal translation. Together, these initiatives represent an emerging but still underdeveloped ecosystem for legal-domain MT in Indian languages.

3 Dataset Description

Table 1 summarizes the dataset used in this study. Provided by the task organizers, it consists of English–Hindi parallel sentence pairs from the legal domain. Only the training pairs were initially released, while validation and test references were withheld. Participants generated translations for these sets during the evaluation and final phases, with the reference translations revealed after the leaderboard announcement.

Table 1: Dataset statistics for Legal Machine Translation (L-MT) Shared Task

Language Pair	Train	Validation	Test
English-Hindi	50,000	5,000	5,000

The dataset contains 60,000 English–Hindi parallel sentences from the legal domain, divided into 50,000 for training and 5,000 each for validation and testing. The wide variation in sentence length suggests diverse syntactic structures typical of legal text. While the dataset is well-balanced and cleanly split, the absence of metadata on text type or source (e.g., statutes, judgments, or general documents) limits fine-grained domain analysis.

4 Experiments

Transformer-based architectures have become the foundation of modern NMT due to their ability to model complex contextual relationships through self-attention. They outperform traditional sequence-to-sequence models, particularly in tasks requiring structural precision and contextual awareness, which are vital in legal translation. However, large Transformer models are computationally expensive. For this study, we employed two complementary training strategies suitable for constrained resources:

- Fine-tuning a pre-trained OPUS-MT model to adapt general translation knowledge to the legal domain, and
- Training a Transformer from scratch to evaluate its capability to learn domain-specific patterns directly from legal text.

Our experimental setup is available in the following Link³.

4.1 Opus Fine-Tune

We fine-tuned the Helsinki Opus-MT model⁴ using the provided training corpus. Since validation references were initially withheld, evaluation was based on interim submissions. The model was optimized using the AdamW optimizer with a learning rate of 2×10^{-5} , weight decay 0.01, and batch size 32. Both input and target sequences were limited to 128 tokens to ensure computational efficiency without excessive truncation. As previous research (Cho et al., 2014) suggests, excessively long inputs degrade model performance due to weakened attention over long dependencies, hence, this cap provides an effective trade-off between fidelity and efficiency.

4.2 Transformer Training

To evaluate the impact of learning solely from domain-specific data, we trained a compact Transformer model from scratch. The configuration included 4 encoder-decoder layers, 8 attention heads, model dimension of 128, dropout of 0.1, and token length of 256, and a vocabulary size derived from a SentencePiece tokenizer of 32,000. The model was trained with a batch size of 32, using the Adam optimizer. Despite limited data, this model demonstrated strong convergence, underscoring the ability of smaller Transformers to effectively learn domain-specific translation patterns when carefully optimized.

4.3 Evaluation Setup

Model outputs were assessed using multiple metrics capturing lexical, syntactic, and semantic correspondence: SacreBLEU (Papineni et al., 2002; Post, 2018), chrF++ (Popović, 2015), TER (Snover et al., 2006), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), METEOR (Lavie and Agarwal, 2007; Banerjee and Lavie, 2005),

³<https://github.com/atanumandal0491/Legal-Translation>

⁴Helsinki Opus-MT

Table 2: Final leaderboard results for the JUST-NLP 2025 Shared Task on Legal Machine Translation (English-Hindi). The best scores for each metric are highlighted. Our system (JUNLP) achieved Rank 4 with competitive performance across lexical and semantic metrics.

Rank	Team Name	Country	BLEU ↑	chrF++ ↑	TER ↓	BERTScore (F1) ↑	METEOR ↑	COMET ↑	AutoRank ↑
1	Team-SVNIT	India	51.61	73.29	37.09	92.61	75.80	76.36	61.62
2	FourCorners	Thailand	50.19	73.67	42.32	92.70	69.54	75.74	60.31
3	goodmen	India	48.56	73.07	41.63	92.38	67.15	75.16	59.39
4	JUNLP	India	46.03 ⁶	70.59 ⁴	42.08 ³	91.19 ⁴	71.84 ³	73.72 ⁴	58.90
5	JUST-MEI	India	46.67	70.03	44.63	90.86	72.86	72.12	58.79
6	Lawgorithms	India	46.27	68.32	43.06	91.03	71.80	72.14	58.26
7	Tokenizers	India	34.08	56.75	55.25	87.39	61.78	65.20	50.87

Table 3: Comparison of translation performance across different models on the English-Hindi legal dataset.

Model	Fine-Tuned	BLEU ↑	chrF++ ↑	TER ↓	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	BERTScore (F1) ↑	METEOR ↑	COMET ↑
OPUS-MT (fine-tuned)	✓	46.03	70.59	42.08	72.42	52.63	69.05	91.19	71.84	73.72
OPUS-MT (baseline)	✗	9.39	27.66	83.40	36.30	13.38	32.93	76.91	30.25	50.80
Transformer (trained from scratch)	✗	37.77	60.88	59.72	35.98	13.62	35.69	88.37	65.58	64.29
NLLB (3.3B distilled)	✗	23.72	47.50	63.29	49.00	26.31	45.78	85.14	45.32	67.25
IndicTrans2	✗	10.87	42.36	81.25	37.89	11.07	37.10	81.21	41.78	60.38

and COMET (Rei et al., 2020). These complementary measures ensure robust evaluation across the dimensions of precision, recall, fluency, and semantic alignment.

5 Results and Analysis

Table 2 presents the final leaderboard results from the JUST-NLP 2025 Shared Task on Legal Machine Translation, comparing the performance of participating systems across a range of lexical, semantic, and edit-based evaluation metrics. Our system, JUNLP, achieved an overall Rank 4, with a SacreBLEU score of 46.03, chrF++ of 70.59, and TER of 42.08, demonstrating strong translation accuracy, requiring relatively low post-editing effort. The model also performed competitively in semantic evaluation, achieving a BERTScore (F1) of 91.19, METEOR of 71.84, and COMET of 73.72, indicating high alignment with human reference translations. While the best-performing team attained marginally higher results across several metrics, our system performed in the mid-range compared to the other participating systems (cf. Table 3), underscoring the effectiveness of domain-focused fine-tuning for legal translation.

Table 3 summarizes our experimental outcomes, comparing the fine-tuned OPUS-MT models with baseline multilingual models. The baseline OPUS-MT (without fine-tuning) performed poorly, with a

SacreBLEU of 9.39 and chrF++ of 27.66, revealing significant deviation from reference translations. BERTScore F1 of 76.91 and a COMET score of 50.8 further indicate weak semantic alignment and limited adaptability of the baseline OPUS-MT to the legal domain.

The fine-tuned OPUS-MT markedly improved translation quality, achieving a SacreBLEU of 46.03, chrF++ of 70.59, and TER of 42.08, demonstrating high lexical accuracy and fluency. The BERTScore (91.19) and COMET (73.72) show strong semantic alignment with human references, while METEOR (71.84) and ROUGE scores confirm consistent n-gram and paraphrase correspondence. These results suggest that fine-tuning effectively transfers linguistic and contextual knowledge from general corpora to specialized legal data without overfitting. This performance reinforces the viability of fine-tuning for domain-specific translation and motivates further exploration of scalable approaches such as parameter-efficient tuning and extension to additional Indian languages.

The Transformer model trained from scratch performed competitively, achieving a SacreBLEU of 37.77 and COMET of 64.29. Despite lacking pre-trained initialization, it captured domain patterns effectively, although the fine-tuned OPUS-MT maintained an edge in fluency and semantic coherence. Multilingual baselines, such as NLLB and

IndicTrans2, performed moderately, underscoring that general-purpose models struggle with domain-specific precision.

Overall, the fine-tuned OPUS-MT model produced fluent, accurate, and contextually faithful translations, confirming its effectiveness for English–Hindi legal MT in real-world settings.

6 Conclusion and Future Work

This work explored domain adaptation strategies for Legal Machine Translation (L-MT) in the English-Hindi context, highlighting how fine-tuning enhances translation quality for specialized text. Among all systems tested, the fine-tuned OPUS-MT model achieved the highest performance, demonstrating superior lexical accuracy and semantic consistency. Training a Transformer model from scratch also yielded promising results, showing that domain-specific supervision alone can produce competitive results under constrained resources.

Future work will extend these experiments to other Indian languages and evaluate parameter-efficient fine-tuning techniques such as LoRA and QLoRA to scale Legal MT further. Ultimately, such systems can play a transformative role in democratizing access to legal knowledge, ensuring that linguistic diversity does not become a barrier to justice.

Limitations

While the proposed approach demonstrates strong empirical performance, several limitations constrain the generalizability and scope of the current study:

- **Restricted training corpus:** The model was trained exclusively on the dataset released by the shared task organizers, without augmentation from external legal or general-domain corpora. Consequently, the system’s exposure to broader linguistic variability and complex domain phenomena remains limited.
- **Lack of comprehensive validation data:** Complete source-target validation pairs were unavailable during training, which hindered reliable monitoring of model behavior (e.g., overfitting or underfitting) and constrained opportunities for principled hyperparameter optimization.
- **Sequence length constraints:** Input-output sequences were truncated to a maximum of 128 to-

kens due to computational limitations. Although suitable for most sentence-level examples, this restriction may adversely affect the processing of lengthy statutory clauses, compound sentences, or cross-referential structures.

- **Absence of human evaluation:** The assessment relies primarily on automated metrics (SacreBLEU, chrF++, BERTScore, COMET), and does not incorporate expert human judgement, limiting deeper qualitative insights into adequacy, legal constancy, and pragmatic interpretability.
- **Resource constraints:** Due to time and computational constraints, broader experimental exploration, such as parameter-efficient tuning, multilingual transfer, or alternative architectures-was not undertaken.

Acknowledgments

This research was funded by the ‘VIDYAAPATI: Bidirectional Machine Translation Involving Bengali, Konkani, Maithili, Marathi, and Hindi’ under the Project titled ‘National Language Translation Mission (NLTM): BHASHINI’.

References

- Vít Baisa, Jan Michelfeit, Marek Medved’, and Miloš Jakubíček. 2016. European Union language resources in Sketch Engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 2799–2803, Portorož, Slovenia. European Language Resources Association (ELRA).
- Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *Preprint*, arXiv:1409.1259.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.
- Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india. *Preprint*, arXiv:2001.09907.

- Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masmus Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. *Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.
- Thangkhanhau Haulai and Jamal Hussain. 2023. *Construction of mizo: English parallel corpus for machine translation*. 22(8).
- Stefan Höfler and Kyoko Sugisaki. 2014. *Constructing and exploiting an automatically annotated resource of legislative texts*. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 175–180, Reykjavik, Iceland. European Language Resources Association (ELRA).
- J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. *Lora: Low-rank adaptation of large language models*. ArXiv, abs/2106.09685.
- Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. *IL-TUR: Benchmark for Indian legal text understanding and reasoning*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11460–11499, Bangkok, Thailand. Association for Computational Linguistics.
- Philipp Koehn. 2005. *Europarl: A parallel corpus for statistical machine translation*. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.
- Anoop Kunchukuttan. 2020. Indowordnet parallel corpus. https://github.com/anoopkunchukuttan/indowordnet_parallel.
- Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. *The IIT Bombay English-Hindi parallel corpus*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. pages 228–231.
- Chin-Yew Lin. 2004. *ROUGE: A package for automatic evaluation of summaries*. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Sayan Mahapatra, Debnan Datta, Shubham Soni, Adrijit Goswami, and Saptarshi Ghosh. 2025. *Milpac: A novel benchmark for evaluating translation of legal text to Indian languages*. 24(8).
- Vandan Mujadia and Dipti Sharma. 2022. *The LTRC Hindi-Telugu parallel corpus*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3417–3424, Marseille, France. European Language Resources Association.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *Bleu: a method for automatic evaluation of machine translation*. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Shantipriya Parida, Satya Ranjan Dash, Ondřej Bojar, Petr Motlicek, Priyanka Pattnaik, and Debasish Kumar Mallick. 2020. *OdiEnCorp 2.0: Odia-English parallel corpus for machine translation*. In *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*, pages 14–19, Marseille, France. European Language Resources Association (ELRA).
- Maja Popović. 2015. *chrF: character n-gram F-score for automatic MT evaluation*. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Matt Post. 2018. *A call for clarity in reporting BLEU scores*. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Loganathan Ramasamy, Ondrej Bojar, and Z. Žabokrtský. 2012. *Morphological processing for english-tamil statistical machine translation*.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravindh Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. *Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages*. *Transactions of the Association for Computational Linguistics*, 10:145–162.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. *COMET: A neural framework for MT evaluation*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Kshetrimayum Boynao Singh, Deepak Kumar, and Asif Ekbal. 2025. *Evaluation of LLM for English to Hindi legal domain machine translation systems*. In *Proceedings of the Tenth Conference on Machine Translation*, pages 823–833, Suzhou, China. Association for Computational Linguistics.

Shashank Siripragada, Jerin Philip, Vinay P. Namboodiri, and C V Jawahar. 2020. [A multilingual parallel corpora collection effort for Indian languages](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3743–3751, Marseille, France. European Language Resources Association.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. [Bertscore: Evaluating text generation with bert](#). *ArXiv*, abs/1904.09675.

Michał Ziemska, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. [The United Nations parallel corpus v1.0](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).