Challenges in Urdu Machine Translation

Abdul Basit and **Abdul Hameed Azeemi** and **Agha Ali Raza**
Lahore University of Management Sciences
{a_basit, abdul.azeemi, agha.ali.raza}@lums.edu.pk

Abstract

Recent advancements in Neural Machine Translation (NMT) systems have significantly improved model performance on various translation benchmarks. However, these systems still face numerous challenges when translating low-resource languages such as Urdu. In this work, we highlight the specific issues faced by machine translation systems when translating Urdu language. We first conduct a comprehensive evaluation of English to Urdu Machine Translation with four diverse models: `GPT-3.5` (a large language model), `opus-mt-en-ur` (a bilingual translation model), `NLLB` (a model trained for translating 200 languages) and `IndicTrans2` (a specialized model for translating low-resource Indic languages). The results demonstrate that `IndicTrans2` significantly outperforms other models in Urdu Machine Translation. To understand the differences in the performance of these models, we analyze the Urdu word distribution in different training datasets and compare the training methodologies. Finally, we uncover the specific translation issues and provide suggestions for improvements in Urdu machine translation systems.

1 Introduction

Neural Machine Translation (NMT) has shown remarkable performance on benchmark datasets, particularly following the introduction of transformer architectures (Vaswani et al., 2017). Among these advancements, large language models like `GPT-3.5` and 4 have demonstrated promising potential for machine translation, particularly for resource-rich languages including English, French, and German. However, these models face numerous challenges in translating low-resource languages (e.g., Urdu) due to limited training compared to their high-resource counterparts (Hendy et al., 2023).

Urdu is spoken by over 100 million people worldwide (Haider, 2018). It is predominantly spoken in Pakistan, serving as the national language (Metcalf, 2003) and holds significant cultural importance. Urdu is also spoken in various regions of India, particularly in states like Uttar Pradesh, Bihar, and Telangana, with a sizable population of speakers. However, due to the scarcity of available linguistic resources for Urdu, it is considered a low-resource language (Daud et al., 2017).

In this work, we empirically evaluate four language models for Urdu machine translation: `GPT-3.5` – a large language model, `opus-mt-en-ur` — a bilingual model specifically trained for Urdu translation, `NLLB` — a multilingual translation model designed to cover 200 languages, incorporating a mix of both high-resource and largely low-resource languages and `IndicTrans2` — a multilingual translation model designed for low-resource Indian languages. `IndicTrans2` demonstrates the highest `SacreBLEU` and `Chrf` on five diverse machine translation datasets, followed by `NLLB` , `GPT-3.5` and `opus-mt-en-ur`. To identify the challenges in Urdu machine translation, we examine the translation capability of the four different models qualitatively and highlight the key areas where the bilingual, multilingual, and large language models struggle to perform.

2 Background

Machine translation is a crucial aspect of NLP, automating text translation between languages. It has evolved from rule-based to data-driven and neural approaches. Traditional rule-based systems faced challenges with language complexities, while statistical methods improved but still struggled with syntax and semantics (Okpor, 2014). Neural machine translation (NMT) has significantly improved the performance, employing deep learning models like sequence-to-sequence architectures (Sutskever et al., 2014) for more fluent

Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024), pages 44–49
August 15, 2024 ©2024 Association for Computational Linguistics

and context-aware translations.

The transformer architecture has improved the overall quality of machine translation. Therefore, large language models, such as GPT-3.5, have emerged as potent candidates for machine translation tasks. Numerous studies have been conducted to assess the effectiveness of these modals for neural machine translation. Hendy et al. (2023) demonstrate that GPT-3.5 can generate remarkably fluent and competitive translation outputs, particularly in the zero-shot setting, especially for high-resource language translations. Prior research has demonstrated the remarkable performance of Large Language Models (LLMs) in high-resource bilingual translation tasks, such as English-German translation (Vilar et al., 2022; Zhang et al., 2022). Jiao et al. (2023) observed that GPT-4 performs competitively with commercial translation products for high-resource European languages but demonstrates a notable drop in performance for low-resource and distant languages. Stap and Araabi (2023) show that GPT-4 is unsuitable for extremely low-resource languages. However, there is currently a lack of cross-evaluation of different types of language models for Urdu machine translation.

3 Methodology and Experiments

We conduct empirical evaluation for Urdu machine translation on four types of language models: Large Language Model (LLM), bilingual model, and two multilingual models using five diverse datasets. Through this investigation, we aim to gain insights into the translation capabilities of these language models for the Urdu language.

3.1 Models

GPT-3.5. Large Language Models (LLMs), like GPT-3.5, have demonstrated strong and consistent performance across a range of NLP tasks. We investigate the performance of GPT-3.5 in translating the English source language into Urdu. Leveraging the API for the model GPT-3.5-turbo-0125, we use a specific translation prompt: "Please translate the sentence into Urdu." Additionally, we add the contextual information, "You are a machine translation system", to facilitate the translation process.

Bilingual. For the bilingual experiments, we utilize the opus-mt-en-ur model (Tiedemann, 2020), which has been specifically trained for En → Ur machine translation. To facilitate this model's deployment, we use the HuggingFace platform[1]. This enables us to conduct our experiments efficiently and standardize the evaluation process.

Multilingual. We use two multilingual translation models, NLLB and IndicTrans2.

NLLB (Costa-jussà et al., 2022) is a multilingual translation model that supports 200 languages, incorporating a combination of high-resource and low-resource languages. Within the inference process, we specify the source language as English and the target language as Urdu, each identified by their respective language codes eng-Latn and urd-Arab.

IndicTrans2 (Gala et al., 2023) is a specialized model designed to cater to 22 Indic languages, including Urdu. During the inference process, we explicitly specify the source language as English and the target language as Urdu, denoted by the language codes eng-Latn and urd-Arab, respectively.

3.2 Datasets

We evaluate the performance of the selected models on five publicly available test data sets. We utilize the tatoteba-test.eng-urd (Tiedemann, 2020) test set, which is a component of the Tatoeba Translation Challenge. This challenge encompasses numerous test sets created for over 500 languages. Our study exclusively focuses on the publicly available Urdu test set. Secondly, we utilize the Flores 101 dataset (Goyal et al., 2022), which provides a valuable resource for evaluating models on low-resource languages, encompassing 101 such languages. For our study, we concentrate on the Urdu subset of Flores 101 to gauge our model's effectiveness in handling low-resource scenarios. Additionally, we evaluate our models using the Mann Ki Baat (Siripragada et al., 2020) test dataset, which exclusively contains Urdu language content extracted from speeches delivered by the Indian Prime Minister in various Indian languages. Our focus centers on the Urdu subset of Mann Ki Baat. Moreover, we incorporate the UMC005 dataset (Jawaid and Zeman, 2011), a parallel corpus comprising English-Urdu alignments sourced from multiple texts, including the Quran, Bible, Penn Treebank, and EMille corpus. Given the publicly available test sets for the Quran and Bible, we merge these subsets to

[1] https://huggingface.co/Helsinki-NLP/opus-mt-en-ur

	tatoteba-test.eng-urd	**Flores101**	**MKB**	**UMC 005**	**Ted Talk**
opus-mt-en-ur	12.06	7.09	6.62	14.51	11.84
GPT-3.5	21.68	16.67	12.79	11.87	12.29
NLLB	25.04	21.37	18.52	20.68	**19.55**
IndicTrans2	**30.76**	**27.41**	**21.73**	**20.41**	16.50

Table 1: The SacreBLEU scores of four models on five datasets for Urdu machine translation

	tatoteba-test.eng-urd	**Flores101**	**MKB**	**UMC 005**	**Ted Talk**
opus-mt-en-ur	0.39	0.28	0.28	0.35	0.34
GPT-3.5	0.48	0.44	0.40	0.37	0.40
NLLB	0.50	0.48	0.45	**0.48**	0.43
IndicTrans2	**0.53**	**0.53**	**0.49**	0.45	**0.44**

Table 2: The CHRF++ scores of four models on five datasets for Urdu machine translation

Model	**Train Sentence Pairs**	**Test Sentence Pairs**	**Languages**	**Params**
opus-mt-en-ur	1M	1663	1	76.42M
GPT-3.5	NA	NA	NA	NA
NLLB-1.3B	18B	1012	200	1.3B
IndicTrans2	**230.5M**	**2036**	**22**	**1B**

Table 3: A comparison of the training & test splits, number of languages, and the number of parameters of different models.

conduct comprehensive evaluations. Lastly, our models undergo assessment using the TED Talk test dataset (Zweigenbaum et al., 2018). Before evaluation, we preprocess the test data by removing pairs containing symbols in their translations, ensuring a standardized and reliable evaluation process.

3.3 Metrics

We use SacreBLEU (Post, 2018) metric to evaluate the translation performance, which has built-in support for scoring detokenized output using standardized tokenization methods, ensuring a fair and unbiased evaluation of models' translation performance. Additionally, we use CHRF++ (Popović, 2017) scores for assessing translation quality, which is particularly useful when dealing with languages featuring complex sentence structures.

3.4 Results

We present SacreBLEU scores in Table 1 to assess the translation efficacy of the designated models. We observe that GPT-3.5 model exhibits notably superior performance compared to the bilingual model but lags behind the multilingual translation models. NLLB emerges as the runner-up, surpassing both GPT-3.5 and bilingual translation proficiency. IndicTrans2 outperforms all other models on four out of five datasets. However, when scrutinizing more challenging evaluations, as exemplified by the TED Talk test set (Zweigenbaum et al., 2018), the performance of IndicTrans2 surpasses that of the bilingual model and the large language model with scores of 16.50, 12.29, and 11.84. Nevertheless, the NLLB model slightly performs better with a score of 19.55. Additionally, we present Chrf++ scores in the table 2, and our observations indicate that IndicTrans2 outperforms all other models.

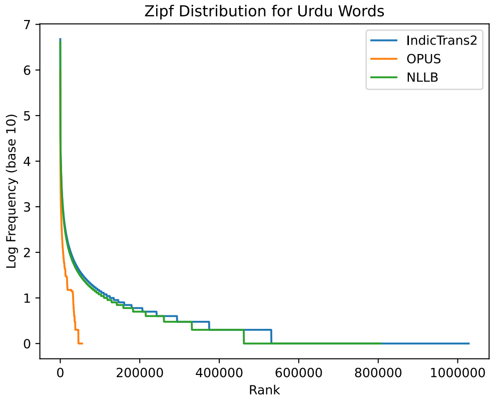

Figure 1: Comparison of Zipf distribution of the training data used in NLLB, IndicTrans2, and OPUS.

To understand why IndicTrans2 performs better than other models, we compare the Zipf distribution of Urdu words present in the training data of NLLB, OPUS and IndicTrans2. Figure 1 shows a significant difference in the Zipf distribution of OPUS compared to other datasets, with significantly fewer types. In contrast, the Zipf distribution of IndicTrans2 and NLLB is more similar, especially for higher-frequency words.

Issue	Source	Actual Translation	Correct Translation	Issue Detail
NER	A **piano** is expensive.	ایک نہایت قیمتی ہے	**پیانو** کافی مہنگا ہے۔	پیانو (piano) is missing in the translated text.
Mistranslation	That will be **funny**.	یہ سُن کر **حیران** رہ جائے گا	وہ بہت **مزاحیہ** ہو گا۔	مزاحیہ (funny) is replaced by حیران (surprised) in the translated text.
Word-Repetition	Is this your **first time** in Japan?	کیا یہ جاپان میں **پہلی بار** آپ کی **پہلی بار** ہے؟	کیا تم **پہلی دفعہ** جاپان آئی ہو؟	پہلی بار (first time) is mentioned twice in the translated text.
Literal translation	Cold **weather is perhaps** the only real danger the unprepared will face.	سرد **موسم شاید تیار نہیں** ہونے والوں کے لئے واقعی خطرہ ہوگا	ٹھنڈا **موسم شاید وہ واحد** حقیقی خطرہ ہے جس کا سامنا غیرتیّار فرد کو کرنا پڑے گا	Incorrect phrase in the translated text: موسم شاید تیار نہیں (cold weather is unprepared).
Word-Omission	The protest started around **11:00** local time (UTC+1) on Whitehall opposite the police-guarded entrance to Downing Street, the Prime Minister's official residence	احتجاج کا آغاز مقامی وقت کے مطابق یو ٹی سی وائٹ ہال پر وزیر اعظم کی سرکاری رہائش گاہ ڈاؤننگ اسٹریٹ کے پولیس کے حفاظتی دروازے کے سامنے ہوا	وزیر اعظم کی سرکاری رہائش گاہ کے داخلی راستے کے سامنے پولیس کی حفاظت والے ڈاوننگ اسٹریٹ کے وائٹ ہال پر مقامی وقت کے مطابق تقریبا **11:00** بجے یہ احتجاج شروع ہوا	11:00 missing in the translated text.
Transliteration	These **scarps** were, found all over the moon and appear to be minimally weathered, indicating the geologic events that created them were fairly recent	یہ **سکارپس** پورے چاند پر پائے گئے تھے اور کم سے کم آب و ہوا کے دکھائی دیتے ہیں، جس سے یہ ظاہر ہوتا ہے کہ ان کو پیدا کرنے والے ارضیاتی واقعات کافی حالیہ تھے	چاند کی سطح پر جا بجا پائی جانے والی **کھائیوں** سے معلوم ہوتا ہے کہ وہ کم موسم دیدہ ہیں۔ ان سے ظاہر ہو تا ہے کہ جن جیولوجک حادثات سے ان کی تخلیق ہوئی وہ بہت حالیہ زمانہ کے	سکارپس instead of کھائیوں in the translated text.

Table 4: Different Urdu translation issues present in Neural Machine Translation models.

In the tail of the distribution, we notice a higher frequency for words present in the IndicTrans2 dataset compared to NLLB, which corresponds to a higher BLEU score as well for IndicTrans2 model. This suggests that for training Urdu NMT models, datasets with optimal Urdu word distribution should be prioritized, as observed in the superior performance of IndicTrans2 that shows a Zipf distribution with a better long tail compared to other datasets.

We now outline the training process of IndicTrans2 to understand the reasons behind its superior performance. The training comprises two phases: auxiliary training and downstream training. The auxiliary phase involves back translation to augment large amounts of monolingual corpora (Sennrich et al., 2015a). Subsequent fine-tuning is done on high-quality, human-generated seed data, including BPCC-H-Wiki and the NLLB seed (Costa-jussà et al., 2022). In the second phase, they train on the augmented parallel corpora which combines original data with the back-translated data. Tagged back translation is used (Caswell et al., 2019) for providing additional supervision to the model such that it distinguishes between different data sources during training. This training process combined with high-quality data sources allows IndicTrans2 to perform better than other models on En → Ur machine translation.

3.5 Challenges

Our research has unveiled various challenges associated with Urdu machine translation. Some of these challenges are universal across all models, while certain issues are present only in specific models. We enumerate these challenges below.

1. The opus-mt-en-ur model encounters a challenge in the domain of Named Entity Recognition (NER), specifically, its ability to produce accurate translations for certain entities. This issue is observable in the first row of Table 4. This issue was not widely present in the translations done through GPT-3.5 or IndicTrans2 models.

2. When the translation diverges from an accurate representation of the source, it is termed 'Mistranslation' (Freitag et al., 2021). The opus-mt-en-ur model consistently grappled with this issue across all datasets, as demonstrated in the second row of Table 4. In contrast, GPT-3.5 and IndicTrans2 exhibited notably superior proficiency in addressing this challenge.

3. The issue of repetition, which has been noted in almost all text generation models, significantly undermines their overall generation performance (Fu et al., 2021). The

word repetition problem was observed in all three models, namely `opus-mt-en-ur`, `GPT-3.5`, and `IndicTrans2` (third row of Table 4).

4. Machine translation systems have long been noted for their tendency to produce overly literal translations (Dankers et al., 2022). We observe a few instances of literal translations for all selected models in our experiments. An example of literal translation with `GPT-3.5` can be seen in the fourth row of Table 4.

5. NMT systems exhibit a tendency to exclude vital words from the source text, thereby significantly diminishing the overall adequacy of machine translation (Yang et al., 2019). The results indicate that the models still face this challenge for Urdu translation. An example from the text translated by `IndicTrans2` is given in the fifth row of Table 4.

6. Transliteration errors can arise from ambiguous transliterations or inconsistent segmentations between the source and target text (Sennrich et al., 2015b). We observe this issue in different models and an example is given in the last row of Table 4.

4 Limitations and Conclusion

In this work, we investigate the Urdu translation capabilities of four diverse models and uncover the specific challenges. We find that `IndicTrans2` outperforms other models for English to Urdu translation, demonstrating superior performance on `SacreBLEU` and `CHRF++` scores, primarily due to its specialized training process and superior Urdu word distribution in its dataset. We uncover specific Urdu translation issues including named entity recognition, mistranslation, word repetition, literal translations, word omissions, and transliteration errors. Addressing these challenges requires focused efforts on constructing high-quality Urdu training datasets, refining model training methods, and incorporating more robust evaluation metrics. For future work, our evaluation of Urdu machine translation can be extended to additional domain-specific datasets and other low-resource Indic languages to uncover additional issues.

References

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. *arXiv preprint arXiv:1906.06442*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Verna Dankers, Christopher G Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. *arXiv preprint arXiv:2205.15301*.

Ali Daud, Wahab Khan, and Dunren Che. 2017. Urdu language processing: a survey. *Artificial Intelligence Review*, 47:279–311.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12848–12856.

Jay Gala, Pranjal A Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Samar Haider. 2018. Urdu word embeddings. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Bushra Jawaid and Daniel Zeman. 2011. Word-order issues in english-to-urdu statistical machine translation. *Prague Bull. Math. Linguistics*, 95:87–106.

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Barbara D Metcalf. 2003. Urdu in india in the 21st century: A historian's perspective. *Social Scientist*, pages 29–37.

Margaret Dumebi Okpor. 2014. Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues (IJCSI)*, 11(5):159.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Shashank Siripragada, Jerin Philip, Vinay P Namboodiri, and CV Jawahar. 2020. A multilingual parallel corpora collection effort for indian languages. *arXiv preprint arXiv:2007.07691*.

David Stap and Ali Araabi. 2023. Chatgpt is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Jörg Tiedemann. 2020. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. *arXiv preprint arXiv:2010.06354*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. arxiv e-prints, page. *arXiv preprint arXiv:2211.09102*.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th workshop on building and using comparable corpora*, pages 39–42.

5 Hyperparameters

The hyperparameters used in our experiments are listed below.

Hyperparameters for GPT-3.5	
Batch Size	500
Tokens	1024
Temperature	0
Language Pair	eng-urd

Hyperparameters for `IndicTrans2`	
Batch size	100
Pad token id	1
Scale embedding	True
Model type	IndicTrans
Language pair	eng-urd

Hyperparameters for `NLLB`	
Batch size	100
Pad token id	1
Scale embedding	True
Model type	m2m_100
Language pair	eng-urd

Hyperparameters for `opus-mt-en-ur`	
Batch size	100
pad token id	1
Scale embedding	True
Number of beams	4
Model type	Marian
Language pair	eng-urd

6 Resources

We conduct all our experiments on a privately hosted server on the cloud and use a Tesla K80 GPU to run the inference.