Challenges in Urdu Machine Translation

Anonymous ACL submission

Abstract

Machine translation systems have witnessed significant advancements in various tasks, raising questions about their performance for low-resource languages, particularly those based on Indo-Aryan scripts like Urdu. This study delves into the challenges faced by machine translation systems when dealing with Urdu, a low-resource Indo-Aryan language. We conduct a comprehensive evaluation of three language models: `GPT-3.5`, a large language model; `opus-mt-en-ur`, a publicly available bilingual translation model; and `IndicTrans2`, a specialized translation model for Indian languages, particularly low-resource ones. Our results reveal that `IndicTrans2` outperforms the other models, signifying its potential in handling low-resource language translation. Additionally, this study sheds light on the specific challenges encountered by models in Urdu translation, offering valuable insights for future improvements in the field of machine translation for low-resource Indo-Aryan languages.

1 Introduction

Urdu is spoken by over 100 million people worldwide (Haider, 2018). It is predominantly spoken in Pakistan, where it serves 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, where it has a sizable population of speakers.

Neural Machine Translation (NMT) has exhibited remarkable performance on benchmark datasets, particularly following the introduction of transformer architectures (Vaswani et al., 2017) tailored for machine translation tasks. Among these advancements, large language models like `GPT-3.5` have demonstrated promising potential for machine translation. Primarily trained on the English corpus, with supplementary segments from the Latin corpus, `GPT-3.5` showcases significant capabilities in handling translation tasks. 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).

In this work, empirically evaluate three 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, and `IndicTrans2` — a multilingual translation model designed for low-resource Indian languages. `IndicTrans2` demonstrates the highest `SacreBLEU` on five diverse machine translation datasets, followed by `GPT-3.5` and `opus-mt-en-ur`. To identify the challenges in Urdu machine translation, we examine the translation capability of the three 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 and context-aware translations.

The transformer architecture has improved how well machines can translate languages. Large language models, such as `GPT-3.5`, have emerged as potent candidates for machine translation

	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
IndicTrans2	**30.76**	**27.41**	**21.73**	**20.41**	**16.50**

Table 1: The SacreBLEU score of three models on five datasets for Urdu machine translation

tasks. Numerous studies have been conducted to assess the effectiveness of ChatGPT in the domain of NMT. Hendy et al. (2023) demonstrate that ChatGPT, GPT-3.5 (text-davinci-003), and text-davinci-002 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 specific low-resource languages, such as Urdu.

3 Methodology and Experiments

We conduct empirical evaluation for Urdu machine translation on three types of language models: Large Language Models (LLMs), bilingual models, and 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

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

Bilingual. For our bilingual experiments, we utilize the opus-mt-en-ur model (Tiedemann, 2020), which has been meticulously trained from scratch to cater to the Urdu language. To facilitate the deployment of this model, we make use of the HuggingFace platform[1]. This enables us to efficiently conduct our experiments and assess the performance of the bilingual model in the context of our research.

Multilingual. We use IndicTrans2 as a multilingual translation model (Gala et al., 2023), a specialized model designed to cater to Indian languages, including Urdu, characterized as a low-resource language. 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. For our study, we exclusively focus 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

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

Issue	Source	System	Reference
NER	A **piano** is expensive.	ایک نہایت **قیمتی** ہے	**پیانو** کافی مہنگا ہے۔
Mistranslation	That will be **funny**.	یہ سُن کر **حیران** رہ جائے گا	وہ بہت **مزاحیہ** ہو گا۔
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؟

Table 2: Translation problems identified for `opus-mt-en-ur`

Issue	Source	System	Reference
Word-Repetition	An **inquiry** was established to investigate.	**تحقیق** کرنے کے لئے ایک **تحقیق** کا اندراج کیا گیا تھا	**تفتیش**کیلئے ایک انکوائری تشکیل دی گئی تھی
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 غیرتیار فرد کو کرنا پڑے گا
Word Order Error	A hostel collapsed in Mecca, the holy city of Islam at **about 10 o'clock** this morning local time	ایک ہاسٹل مکہ المکرمہ میں آج صبح کریب **10 بجے** مقامی وقت پر گر گیا۔	آج صبح علاقائی وقت کے مطابق **10 بجے** اسلام کے مقدس شہر مکہ میں ایک ہوسٹل گر گیا۔

Table 3: Translation problems identified for ChatGPT

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 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.

3.4 Results

We present the `SacreBLEU` scores in Table 1 to assess the translation efficacy of the designated models. Our observations indicate that the `GPT-3.5` model exhibits notably superior performance compared to the bilingual counterpart, particularly evident in relatively straightforward assessments such as the (Tiedemann, 2020) `tatoteba-test.eng-urd` test set. In this context, the bilingual model achieves a `SacreBLEU` score of 12.06, whereas the `GPT-3.5` model excels with a `SacreBLEU` score of 21.68. However, when scrutinizing more challenging evaluations, as exemplified by the TED Talk test set (Zweigenbaum et al., 2018), the performance of `GPT-3.5` only marginally surpasses the bilingual model, with scores of 12.29 and 11.84, respectively. These outcomes underscore that neither `GPT-3.5` nor the bilingual model demonstrates adeptness as proficient Urdu translators.

In stark contrast, the multilingual translation model, `IndicTrans2`, emerges as the frontrunner, surpassing both `GPT-3.5` and the bilingual model in translation proficiency. This is evident in its `SacreBLEU` scores of 30.76 for the `tatoteba-test.eng-urd` test set and 16.50 for the TED Talk set. Notably, when focusing exclusively on the Flores test set, which stands as a diverse benchmark assessment, the results are compelling. The `opus-mt-en-ur` model yields a score of 7.09, `GPT-3.5` records 16.67, and `IndicTrans2` significantly outperforms with a score of 27.41. Together, these results highlight that `IndicTrans2` performs better in translating Urdu compared to the other models we considered. A plausible hypothesis for the superior performance of `IndicTrans2` stems from the specialized training methodology tailored specifically for Indian languages. Conversely, `GPT-3.5`'s predominant training on Latin corpora might contribute to its comparatively diminished performance in this context.

3.5 Challenges

Our research has unveiled various challenges associated with translation models. Some of these challenges are universal across all models, while

Issue	Source	System	Reference
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
Word-Repetition	After the fire, the fortress was **preserved and protected**, remaining to be one of Bhutan's most sensational attractions.	آگ لگنے کے بعد، قلعے کو **محفوظ اور محفوظ** کر لیا گیا، جو بھوٹان کے سب سے سنسنی خیز پرکشش مقامات میں سے ایک رہا	قلعے کو آتش زنی کے بعد **محفوظ** کیا گیا، وہ بھوٹان کا ایک سب سے زیادہ سنسنی خیز مقام رہ گیا۔
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

Table 4: Translation problems identified for `IndicTrans2`

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 entities. This issue is observable in the first row of Table 2. Interestingly, we did not notice this issue in the `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 exemplified in the second row of Table 2. 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`.

4. Machine translation systems have long been noted for their tendency to produce overly literal translations (Dankers et al., 2022). Results show that `GPT-3.5` was less literal in the case of high-resource languages (Raunak et al., 2023). We observed literal translations for all selected models in our experiments, and an example for `GPT-3.5` can be seen in the second row of Table 3.

5. Transliteration errors can arise from ambiguous transliterations or inconsistent segmentations between the source and target text (Sennrich et al., 2015). We observe that `IndicTrans2` faces this challenge (see the third row of Table 4).

6. 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 `IndicTrans2` model still faces this challenge for Urdu translation (first row of Table 4).

4 Conclusion and Future Work

Our investigation encompassed the assessment of these models in the realm of elementary Urdu translation, a member of the Indo-Aryan language family. Moving forward, our focus could extend to the evaluation of these models across additional low-resource languages, integral components of the broader Indo-Aryan linguistic spectrum.

5 Limitations

Our evaluation of Urdu machine translation can be extended to additional, domain-specific datasets to uncover specific issues and to better understand the Urdu translation capabilities of large language models. We report only the `SacreBLEU` score in our study. CHRF++ (Popović, 2017) scores can be useful for evaluating translation quality, especially when dealing with languages that have complex word structures and word order.

References

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*.

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*.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan Awadallah. 2023. Do gpts produce less literal translations? *arXiv preprint arXiv:2305.16806*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. 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.

6 Hyperparameters

The table 6 lists the hyperparameters we used in our experiments.

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 `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

7 Resources

we conduct our experiments on the cloud and used Tesla's k80 GPU for running the inference of the models.