Multilingual Chatbot for Low Resource Language

Harshithavalli , Monisha Raju , Sri Padmavathi

1 Abstract

This project introduces a multilingual chatbot aimed for supporting low-resource languages, specifically focusing on Urdu. We utilized cross-lingual transfer learning from Hindi to fine-tune a multilingual transformer model (mT5) and compared its performance with a custom Seq2Seq model. The results indicate that the mT5 model, when adapted through a series of fine-tuning stages, generates more fluent and context-aware responses in Urdu. Our findings give the effectiveness of leveraging high-resource languages to improve chatbot performance in low-resource environments.

2 Introduction

Developing intelligent and effective chatbots that can support a wide variety of languages has been an enduring goal of Natural Language Processing (NLP). While a wide range of progress has been made in the field of pre-trained models, such as mT5 and GPT, for high-resource languages such as English and Hindi. The success these pre-trained models have achieved with high-resource languages wasn't extended to low-resource languages. Pre-trained models often underperform for low-resource languages.

Urdu, a popular low-resource language extensively spoken in South Asian Countries, is often underrepresented in the current Natural Language Processing techniques. Hindi, other popular South Asian language, estimated to have about 609 million active speakers, is a high-resource language largely due to the presence of data online. Hindi and Urdu, both languages both originate from Hindustani, sharing similarities in syntax, structure, and vocabulary. Considering the close linguistic proximity of Hindi and Urdu,cross-lingual techniques might be useful to transfer knowledge from Hindi to Urdu (Wu Dredze, 2019; Durrani et al., 2014). This project dwells more on whether these transfer learning strategies can be applied successfully to create Urdu responses that are apt.

To explore and investigate the cross-lingual transfer techniques, we designed two modelling experiments. The first experiment uses a pre-trained multilingual transformer, mT5, further refining that into two stages. First Stage includes fine-tuning on Hindi to help the model understand the general conversational structure, and then fine-tuning on Urdu data. The setup for the second experiment includes training a custom Seq2Seq model with an attention mechanism, with initial training on Hindi, and then training using the Urdu dataset.

With the above dual experiments, we can analyze how the cross-lingual knowledge transfer technique affects the output of dialogues of the chatbots and how effectively each approach mentioned above facilitates Urdu response in a low-resource setting. Also, the evaluation emphasizes on qualitative understanding of generated responses and quantitative indicators such as the BLEU score.

3 Problem Statement

Despite the pretrained model being trained on a large multilingual corpus, it underperforms for low-resource languages like Urdu when used in a zero-shot scenario. Although the pretrained models are trained on a large corpus, they often fail to utilize and capitalize on the linguistic similarity between languages. This problem persists and emphasizes the need for fine-tuning to tackle the shortcomings, such as repetitive and irrelevant responses.

To fill in these gaps, our project aims to use the idea of knowledge transfer techniques. We use the hypothesis that cross-lingual transfer learning with fine-tuning can improve Urdu response qual-

ity when directed by structural and semantic overlap between Hindi , a high-resource language, and Urdu, a low-resource language.

4 Research Questions

- What transfer learning strategies can be effectively leveraged to enhance chatbot training for low-resource languages?
- How does fine-tuning a language-specific model in a large resource-pretrained multilingual model impact low-resource language task performance compared to full-model fine-tuning?
- Whether leveraging linguistic similarities within the Indic language family enhance chatbot performance in a low-resource setting?

5 Related Work

Developing a low-resource language chatbot has become crucial for Natural Language Processing Taks. There has been a vast success rate in the evolution of multilingual chatbots for high-resource languages, but these lack good performance for low-resource languages. Certain pertained models, such as BERT, mBART, and mT5, underperform in a low-resource setting (Hedderich et al., 2020; Xia et al., 2021).

Multilingual transformer-based pertained model has a significant zero-shot capability. However, on performing the tasks, it has become evident that despite being trained on a large multilingual training corpus, these models remain naive to generalization. This emphasizes the importance of fine-tuning in a low-resource setting, although prior multilingual training lays the foundation for generalization for models that are cross-lingual (Hedderich et al., 2020; Xia et al., 2021).

Over time, with enough training data fed, the sequence-to-sequence models with an attention mechanism have demonstrated to be strong baseline models for text generation tasks(Bahdanau et al., 2015). For these models to perform well, they highly depend on fine-tuning and supervised training. Another observation that was noticed without fine-tuning is that the models often tend to be repetitive and irrelevant.

To overcome all the above contrasts, the hypothesis of cross-lingual transfer came into place. Hindi and Urdu, two linguistically similar languages, can use transfer of structural and semantic concepts(Pires et al., 2019; Thangaraj et al., 2024; Nag et al., 2023). With additional fine-tuning, these models have promising results on tasks such as classification, Named Entity Recognition, and Machine Translation (Xia et al., 2021). Previous work on Bantu languages, presented in the paper (Thangaraj et al., 2024), leverages linguistic similarity and cross-lingual learning, which further helped us and laid a strong base to test out our hypothesis on Hindi and Urdu, a high-resource and low-resource language, respectively.

Evaluation of chatbot responses is a challenge. Popular automated metrics such as BLUE, ROGUE can be used, but they fail to capture semantic importance. To evaluate fluency, accuracy, coherence, harmony, and context alignment of the outputs, it is necessary to enforce both qualitative and quantitative methods for evaluation (Hedderich et al., 2020).

6 Dataset Description

IndicDialogue is a comprehensive collection of subtitles from ten Indic languages designed to support NLP research in low-resource Indian languages. With 7,750 raw subtitle files containing a total of 6,853,518 dialogues, it is one of the largest available conversational datasets for Indic languages. This dataset offers rich, natural dialogue structures extracted from media content, making it highly valuable for training AI-driven chatbots. It includes:

- **Languages Covered:** Ten Indic languages
- **Raw Subtitle Files:** 7,750
- **Total Dialogues:** 6,853,518
- **Total Words:** 42,188,569
- **Total Characters**: 218,062,531

6.1 Hindi Data(High-Resource Language

We chose 100,000 Hindi dialogue pairs (as input text and target text) as a starting point. Due to a wide range of online material, the Hindi dataset offers a strong content of data with semantic context, vocabulary richness, and general conversational patterns. Each sample contains an input sentence and its corresponding response.

This dataset was used in Phase 1 of both experimental approaches in mT-5 fine-tuning and Custom Sequence2Sequence, making sure the model grasps fluent and contextually coherent responses before transferring capability to Urdu.

6.2 Urdu Dataset (Low-resource Language)

The Urdu Dataset was used in Phase 2. We selected around 1000-1500 Urdu samples to refine the model trained on Hindi models. Certain constraints we have noticed in the Urdu dataset include:

- limited vocabulary usage
- Noisy sample with spelling and formatting errors
- Frequent repetition of samples

With these constraints in place, this subset was crucial for assessing the model's ability to transfer knowledge from Hindi to Urdu, adapting to the linguistic similarities.

6.3 Data Preprocessing

- Input-output pairs are created after loading and parsing a JSON file.
- **Text Cleaning:**Eliminate unnecessary whitespace, HTML elements, and special symbols. Correct frequent spelling mistakes and irregular punctuation. For uniformity, convert all text to Unicode NFC format.
- **Normalization:**Lowercase the text as required, and remove any unnecessary punctuation or signs. Standardize characters and ensure that all entries have the same encoding (UTF-8).
- **Language check to retain samples:**If any samples don't match the desired language, filter them out. This is necessary to ensure language consistency.
- **Noise Reduction:**Eliminate low-information or repetitive statements that hinder model learning.
- **Tokenization:**Use tokenization for sentences or subwords (e.g., BPE using HuggingFace tokenizer). Usage of a pretrained tokenizer that matches your underlying model (such as Falcon or mBART). Manage attention masks, truncation, and padding appropriately for training.

7 Methodology

We used and compared two different approaches to examine the success rate of transfer learning approaches in creating multilingual responses in a low-resource setting. The approaches are namely (1) Fine-tuning a pretrained multilingual transformer model and (2) training a custom-made sequence-to-sequence model with attention. Both approaches leverage cross-lingual transfer techniques with the usage of a Hindi dataset in Phase 1 and a low-resource language in Phase 2

7.1 Approach 1: Fine-Tuning the mT5 Transformer

7.1.1 Phase 1: High-Resource Fine-Tuning on Hindi Dialogue

The first step was to fine-tune the pretrained mT5-small model (Google's multilingual T5) on 100,000 Hindi conversation pairs in order to adapt it to Hindi language. This process includes:

- Training the model in fundamental conversational skills, such as taking turns and responding to questions.
- Improve its understanding of Hindi-specific structure and semantics in a high-resource setting.

Implementation Details

- **Base Model:** google/mt5-small (a lightweight but efficient multilingual variant).
- **Tokenizer:** SentencePiece (pretrained, subword-based).
- **Sequence Length:** Capped at 64 tokens for balance between context and efficiency.
- **Training:** 2 epochs (sufficient for initial convergence).
- **Batch Size:** 4 (limited by GPU memory constraints).

Outcome: The model achieved coherent Hindi response generation, demonstrating transferable capabilities to Urdu

7.1.2 Phase 2 – Urdu Adaptation

As part of the second phase , the model which was pretrained on Hindi is future fine-tuned on a small 1,000 Urdu samples dataset .

- **Tokenier:** mT5 multilingual SentencePiece tokenizer.
- Avoid starting from scratch while training the model to Urdu. It is observed that even with a small domain specific fine tuning with urdu dataset . BLEU score improved when compared over zero-shot Urdu performance.

Outcome: This method shows that using multilingual pretraining along with specific fine-tuning can improve response generation in languages that are less commonly used.

7.2 Approach 2: Custom Seq2Seq Model with LSTM and Attention

We built a custom model that uses an encoder-decoder structure with LSTM and Bahdanau attention. This model is a lighter option compared to larger transformer models.

Model Architecture:

- Encoder-Decoder: Stacked LSTM layers
- Embedding Dimension: 256
- Latent Dimension (LSTM Hidden State): 256
- Attention: Bahdanau-style context vector fusion
- Decoding Strategy: Greedy decoding

7.2.1 Phase 1 – Training on Hindi dataset

The Seq2Seq model was initially trained on 100,000 Hindi dialogue pairs.

- Tokenizer: Custom Keras Tokenizer trained on Hindi vocabulary
- Max Sequence Length: 30 tokens
- Epochs: 10
- Batch Size: 32

This allowed the model to learn the mapping between input and response sequences in Hindi.

7.2.2 Phase 2 – Fine-Tuning on Urdu

After training, the model weights for Hindi were reused to fine-tune the system on Urdu data

- Tokenizer: A separate Keras tokenizer was trained for Urdu.
- Challenge: There were issues with vocabulary mismatch and index alignment between the Hindi and Urdu tokenizers.
- Outcome: Although the model generated some coherent Urdu phrases, the quality of the output was affected by repetition and token degeneration due to a scarcity of data.

8 Experiments and Results

Our aim was to find out how well we can transfer knowledge from Hindi, which is a high-resource language, to Urdu, a low-resource language, to improve how chatbots generate responses.

8.0.1 Task Definition

To evaluate the performance of the chatbot models, we designed a structured response generation task. We created a curated set of 10 Urdu input prompts to serve as the evaluation test set. These prompts were inspired and simulated by real-world dialogue queries. For each input prompt, the trained model was tasked with generate a response, which was stored in a list of generated responses. Additionally, we manually constructed a corresponding list of expected responses (target texts) to reference points to compare against.

8.0.2 Evaluation Framework

To evaluate how well each approach works, we used two main methods: automated metrics and human analysis.

- **BLEU(Bilingual Evaluation Understudy) score:** This measures how many n-grams match between the generated response and the human response. It provides a standard way to compare how similar the language is.
- **Qualitative Review:**An analysis of responses that is human-readable, assessing semantic fluency, coherence, and appropriateness. This review focuses on identifying strengths and limitations that are not captured by BLEU scores

While BLEU is a useful tool for evaluation, it doesn't always reflect how good a conversation is, especially in languages like Urdu that have complex grammar. That's why we used both scores and human evaluation to get a clearer picture.

8.0.3 Experimental Setup:

We evaluate the chatbot with a specific set of Urdu prompts taken from real-life situations. Each model was asked to respond to these prompts.

```
Prompt: کیا آپ میری مدد کر سکتے ہیں؟
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 60ms/step
Response: نہیں آپ
----------------------------------------
Prompt: آپ کا نام کیا ہے؟
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 62ms/step
Response: نہیں کا
----------------------------------------
```

8.0.4 Results from Approach 1: Fine-Tuning mT5

Phase	Description	BLEU Score	Observations
Zero-shot	mT5 without any fine-tuning	0.0082	Model output was repetitive
Fine-tuned on Hindi	Tested directly on Urdu	0.0105	Slight improvement, but failed to generalize
Phase 2: Fine-tuned on Urdu	Hindi + Urdu fine-tuning	0.0255	Best BLEU score, more coherent responses

After the first phase of training in Hindi, the model struggled to understand Urdu inputs. However, once we trained it further on Urdu, it began to respond with phrases like " " and " " more fluently and appropriately. This shows how effective cross-lingual transfer learning can be in improving multilingual models.

8.0.5 Result from Approach 2: Seq2Seq LSTM with Attention

Phase	Description	BLEU Score	Observations
Trained on Hindi	Tested directly on Urdu	~0.00	Model could not generalize to Urdu
Fine-tuned on Urdu	Reused weights, trained on 1K Urdu samples	0.0017	Some phrases generated but mostly repetitive

The LSTM model faced challenges when working with Urdu, even with weights that had been trained on Hindi. The differences in vocabulary and the lack of a shared token system between the Hindi and Urdu tokenizers complicated the model's effectiveness. While the attention mechanism provided some focus on specific context, the model's limited resources ultimately resulted in subpar performance.

8.0.6 Overall Summary of Models from Both Approaches

Model	Transfer Type	BLEU Score	Output Quality
mT5	Transformer + Fine-tune (Hindi → Urdu)	0.0255	More fluent, generalizes better
Seq2Seq LSTM + Attention	RNN-based, retrained (Hindi → Urdu)	0.0017	Outputs are repetitive, quality low

Overall, the mT5 model performed much better than the custom LSTM model in both measurements and quality. Its training in multiple languages and use of common word parts helped connect Hindi and Urdu effectively.

8.0.7 Insights and Implications

- Training a model on Hindi before adapting it for Urdu is a good approach for languages that lack sufficient resources.
- Multilingual transformers improve performance by leveraging shared characteristics between languages
- Without enough Urdu data, even well-designed LSTM models tend to struggle with generating diverse and meaningful responses.
- Additionally, many contextually appropriate answers received low BLEU scores, indicating a need for enhanced evaluation methods in dialogue systems.

8.1 Research Question Refections and Discussion

Our results show that transfer learning is very important for improving chatbot performance in situations where there aren't many resources. When we fine-tuned the mT5 model by first training it on Hindi and then on Urdu, its BLEU score increased significantly, going from 0.0082 (in a zero-shot setting) to 0.0255. This means that even with only a few Urdu samples, the model was able to use what it learned from Hindi to give better responses. On the other hand, models that weren't fine-tuned on Urdu had a hard time understanding and responding correctly. This shows that simply knowing multiple languages isn't enough when there's no training on a specific language.

Aspect	mT5	Seq2Seq (LSTM + Attention)
Transferability	Strong due to multilingual pretraining	Weak — training from scratch
Response Quality	Acceptable, short but meaningful	Poor, mostly repetitive
BLEU Score	0.0255	0.0017
Resource Usage	High (requires GPU and more memory)	Lightweight and easy to run

mT5 consistently produces outstanding results in all areas, particularly in transferability and response quality.

8.1.1 Discussion on Research Questions

- **What transfer learning strategies can be effectively leveraged to enhance chatbot training for low-resource languages like Urdu?**
  Our research shows that fine-tuning in two steps works really well. First, we trained the model using a language that has a lot of resources, Hindi. Then, we did training on a Urdu language. This method helped the mT5 model improve its BLEU score by more than three times and produce responses that make sense in context.

- **How does fine-tuning a language-specific model within a pretrained multilingual architecture affect Urdu dialogue generation performance compared to full-model training from scratch?**
  The pretrained mT5 model performed better than the Seq2Seq model in both numbers (BLEU: 0.0255 compared to 0.0017) and quality of results. This shows that making small adjustments to a multilingual model helps it work better and use data more efficiently for tasks with less available information.

- **Can leveraging linguistic similarities within the Indic language family improve cross-lingual transfer and response generation in low-resource chatbot settings?**
  Yes, using the similarities between Hindi and Urdu has been helpful. It made it easier to share knowledge and allowed the model to give good responses in Urdu, even though it had only a small amount of Urdu data to learn from.

9 Conclusion

This project focused on developing a multilingual chatbot for low-resource language settings, specifically targeting Urdu. We implemented and evaluated two approaches: (1) fine-tuning a pretrained multilingual transformer model (mT5) and (2) training a custom Seq2Seq model with attention from scratch. The results indicated that fine-tuning the mT5 model first on a high-resource language, Hindi, and then on Urdu facilitated effective cross-lingual transfer. The mT5 model produced responses that were more fluent and contextually appropriate, resulting in a higher BLEU score compared to the Seq2Seq model. In contrast, the Seq2Seq model struggled with generalization and produced repetitive outputs due to its limitations in Urdu data and lack of pretraining

Team Members And Contributions

- **Harshithavalli** - Focused on implementing and fine-tuning the mT5 multilingual model. Handled the two-stage training process—initial fine-tuning on Hindi and subsequent adaptation to Urdu, and evaluated model performance both quantitatively and qualitatively.

- **Monisha Raju** - Contributed to dataset preparation, including cleaning, normalization, and language filtering. Helped in evaluation setup, test prompt design.

- **Sri Padmavathi** - Designed and implemented the custom Seq2Seq model using LSTM with Bahdanau Attention. Managed model training on Hindi and Urdu datasets, tokenizer creation, and inference generation, and evaluated model performance both quantitatively and qualitatively

References

- Arnob, Noor Mairukh Khan, et al. 2024. IndicDialogue: A dataset of subtitles in 10 Indic languages for Indic language modeling. Data in Brief, 55:110690.

- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations (ICLR).

- Devine, Peter. 2024. Tagengo: A multilingual chat dataset. arXiv preprint arXiv:2405.12612.

- Durrani, Nadir, Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2014. Joint projection of multiple annotations for cross-lingual parsing. Transactions of the Association for Computational Linguistics, 2:207–220.

- Hedderich, M. A., D. I. Adelani, X. Zhu, P. Nabende, D. Klakow, and N. D. M. M'hamdi. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. arXiv preprint arXiv:2010.03179.

- Nag, Abhik, Bapi Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. 2023. Transfer learning for low-resource multilingual relation classification. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(2): Article 50.

- Pires, Telmo, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? arXiv preprint arXiv:1906.01502.

- Thangaraj, Hariharan, Alexis Chenat, J.S. Walia, and Vukosi Marivate. 2024. Cross-lingual transfer of multilingual models on low-resource African languages. arXiv preprint arXiv:2409.10965.

- Wu, Shijie, and Mark Dredze. 2019. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 833–844.

- Xia, Mengjie, Guanghui Zheng, Sopan Mukherjee, Milad Shokouhi, Graham Neubig, and Ahmed Hassan Awadallah. 2021. MetaXL: Meta representation transformation for low-resource cross-lingual learning. arXiv preprint arXiv:2104.07908.

- Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).