

MTG-LIRA at WMT 2026: Data-Centric Arabic-Hindi Machine Translation through Corpus Augmentation, Synthetic Parallel Data Generation and Transformer Models.

Vedarth Hede
C-DAC, Pune

Neha Gupta
C-DAC, Pune

Harish Bapat
C-DAC, Pune

Harsh Ekbote
C-DAC, Pune

Abstract

This paper describes the MTG-LIRA submission to the WMT 2026 Arabic–Asian Machine Translation Shared Task (Laskar et al., 2026) for Arabic–Hindi machine translation. Neural Machine Translation (NMT) for the Arabic–Hindi language pair remains challenging due to the limited availability of high-quality parallel corpora and the linguistic differences between the two languages. To address these challenges, we propose a data-centric translation pipeline that combines extensive corpus preprocessing, external corpus augmentation, synthetic parallel data generation, and Transformer-based Neural Machine Translation. Our system is developed using the official datasets provided by the organizers of the WMT 2026 Arabic–Asian MT Shared Task and additional publicly available parallel corpora to improve translation quality for this low-resource language pair. The details of the shared task and the official datasets are described in the WMT 2026 Arabic–Asian MT Shared Task overview paper.

Our training corpus is constructed by integrating the official WMT 2026 shared task data with multiple publicly available parallel corpora collected from OPUS. In addition, we generate synthetic Arabic–Hindi parallel data by translating the English side of the BPCC (Gala et al., 2023) English–Hindi corpus using Meta’s NLLB model (Costa-jussà et al., 2022), thereby expanding the available Arabic training data. All datasets undergo a comprehensive preprocessing pipeline consisting of Unicode normalization, language-specific text normalization, script validation, sentence length filtering, duplicate removal, and alignment verification to ensure corpus quality. We further investigate the impact of two widely used subword segmentation techniques, Byte Pair Encoding (BPE) (Sennrich et al., 2016) and SentencePiece, on translation performance.

The translation models are implemented using the OpenNMT-py framework (Klein et al., 2017) with a Transformer encoder–decoder architecture (Vaswani et al., 2017) and trained from scratch on the curated parallel corpus. Experimental results demonstrate the effectiveness of large-scale corpus augmentation and rigorous data preprocessing for improving Arabic–Hindi translation quality. Our work highlights the importance of high-quality data preparation and subword tokenization in developing robust Neural Machine Translation systems for under-resourced language pairs.

1 Introduction

Arabic is one of the world’s most widely spoken languages and serves as a primary language for communication, education, religion, business, and digital content across the Middle East and North Africa. Hindi, one of the most widely spoken languages in India, represents a large and rapidly growing population of digital users. Accurate Arabic–Hindi machine translation has practical applications in international trade, tourism, education, cultural exchange, media, and cross-border communication. However, compared to widely studied language pairs such as English–French or English–German, Arabic–Hindi remains relatively under-resourced, with limited publicly available high-quality parallel corpora spanning diverse domains (Ahmed et al., 2019).

The WMT 2026 Arabic–Asian Shared Task (Laskar et al., 2026) provides a common benchmark for advancing machine translation research across Arabic and several Asian languages. Beyond the official training data, publicly available parallel corpora often differ considerably in domain, sentence quality,

formatting, and linguistic consistency. Consequently, simply combining multiple datasets does not necessarily improve translation performance. Careful corpus preparation, language-specific normalization, quality filtering, and duplicate removal are essential for constructing reliable training data suitable for Neural Machine Translation.

In this paper, we examine the specific dataset challenges associated with the low-resource Arabic-Hindi language pair, analyze their impact on model performance, experiment with tokenization and present benchmark results for these languages

Rather than proposing a new model architecture, this work focuses on the systematic construction and evaluation of training resources for Arabic-Hindi machine translation. Our contribution lies in combining heterogeneous OPUS resources with synthetic data generated through pivot-based augmentation, applying a unified preprocessing pipeline, and empirically analyzing the effect of these choices across both translation directions.

2 Data

The performance of Neural Machine Translation systems is largely determined by the quality, diversity, and coverage of the parallel training corpus. To construct a comprehensive Arabic-Hindi training dataset, we combined the official WMT 2026 shared task data with multiple publicly available parallel corpora collected from the OPUS repository (Tiedemann, 2012). Additionally, we generated synthetic Arabic parallel data using Meta's multilingual NLLB model to further enrich the training corpus. Before model training, all datasets were subjected to a unified preprocessing and quality assurance pipeline to remove noise, normalize language-specific text, and preserve sentence alignment.

2.1 Official WMT 2026 Shared Task Data

The official data given as part of WMT 2026 Arabic-Asian Shared Task (Laskar et al., 2026) was used as primary dataset and/or high quality dataset for fine tuning a base model. This dataset served as the primary benchmark for system development, hyperparameter tuning, and evaluation. Although the official corpus provides high-quality sentence alignments, its relatively limited size motivated the use of additional external resources to improve lexical coverage and domain diversity. The corpus statistics are summarized in Table 1.

Dataset	Sentence Pairs
Train Set	20,100
Dev Set	1,000
DevTest Set	500

Table 1: Official WMT 2026 Arabic-Hindi Dataset Statistics

2.2 External Parallel Corpora

To complement the official WMT dataset, we incorporated several publicly available Arabic-Hindi parallel corpora obtained from the OPUS repository (Tiedemann, 2012). These datasets cover a wide range of domains, including religious texts, software localization, conference talks, humanitarian communication, and general multilingual resources. Since the quality and characteristics of these corpora vary considerably, each dataset was individually preprocessed and filtered before being merged with the official training data. The corpus statistics are summarized in Table 2.

Corpus	Sentence Pairs
Tanzil	12,472
Bible (UEDIN)	31,102
TED 2020	46,369
TICO-19	46,065

KDE4	122,975
Neulabs	221,375
NLLB (OPUS Mined)	3,571,635
Total	4,051,993

Table 2: External Parallel Corpora Used

The external corpus is dominated by the NLLB (OPUS-mined) component, which contributes approximately 3.57M of the 4.05M external sentence pairs. We retained this corpus to maximize parallel-data coverage in the low-resource setting, while keeping the same external-data composition across all experiments. However, appropriate up-weighting of other components was done while training the models to ensure NLLB(OPUS-mined) does not completely overshadow other components.

2.3 Synthetic Parallel Data Generation

To further increase the amount of training data, we generated additional synthetic parallel corpora using Meta's multilingual NLLB-200-instruct-3.3B model (NLLB Team et al., 2022). We translated the English side of the BPCC English–Hindi corpus into Arabic, thereby creating approximately 7 million synthetic English–Arabic sentence pairs. This synthetic corpus was incorporated into our overall training pipeline (pivot-based augmentation) after undergoing the same preprocessing and quality assurance procedures as the naturally occurring parallel datasets. The use of synthetic data substantially increased the scale and diversity of the available training resources while reducing dependence on limited manually curated parallel corpora.

2.4 Data Preprocessing

The Arabic–Hindi parallel corpora collected from multiple sources exhibited considerable variations in formatting, NFKC encoding, sentence quality, and alignment. Therefore, all

datasets were processed through a unified preprocessing pipeline before being used for model training. The generated synthetic pairs underwent the same preprocessing pipeline as the parallel data, including removal of empty or malformed sentences, duplicate pairs, and basic source–target consistency checks.

Language-specific Normalization: Arabic and Hindi texts were independently normalized to remove Unicode inconsistencies and standardize punctuation, whitespace, and language-specific characters. This step reduced vocabulary fragmentation caused by multiple representations of identical characters.

Quality Filtering: Sentence pairs containing excessive punctuation, invalid characters, empty lines, abnormal length ratios, duplicated punctuation, or extremely short or long sentences were removed. Script-based validation was also performed to ensure that the source sentences predominantly contained Arabic script and the target sentences predominantly contained Devanagari script.

Duplicate Removal and Corpus Validation: Exact duplicate sentence pairs were removed from every corpus before merging. The deduplicated corpus was further validated to ensure sentence alignment and eliminate redundant parallel instances that could bias model training.

Corpus Merging: After preprocessing and validation, the cleaned OPUS corpora and the synthetic parallel corpus were merged to construct the final training corpus used for model development.

3 Methodology

3.1 Transformer Architecture

The proposed Arabic–Hindi NMT systems were implemented using the OpenNMT-py framework (Klein et al., 2017) based on the Transformer encoder–decoder architecture (Vaswani et al.,

2017). The Transformer architecture employs multi-head self-attention mechanisms to model long-range dependencies while enabling efficient parallel training. Compared to recurrent neural network-based architectures, Transformers provide improved scalability and translation quality, making them the de facto standard for modern Neural Machine Translation.

Recent WMT shared-task work has explored Transformer-based and transfer-learning approaches for low-resource Indic translation (Wagh et al., 2025). We sought to extend this strategy to the Arabic-Hindi language pair.

Parameter	Value
Framework	OpenNMT-py
Type	Transformer
Encoder layers	6
Decoder layers	6
Hidden size	512
Word Embedding Size	512
Attention Heads	8
Self Attention	Scaled Dot Product
Dropout	0.30
Training Steps	50000
Validation Interval	5000
Early Stopping	5

Table 3: Transformer Model Configuration

3.2 Subword Tokenization

To effectively handle the rich morphology and large vocabulary of Arabic and Hindi, we investigated two widely adopted subword segmentation techniques: SentencePiece (SPM) (Kudo and Richardson, 2018) and Byte Pair Encoding (BPE) (Sennrich et al., 2016). In both

approaches, separate source and target vocabularies of 32,000 subword units were constructed for Arabic and Hindi.

3.3 Training Strategy

Two training strategies were investigated for both Arabic→Hindi and Hindi→Arabic translations.

Training Strategy I: A base Transformer model was first trained on the cleaned external OPUS parallel corpus. The resulting model was subsequently fine-tuned using the official WMT 2026 Arabic–Hindi training corpus, which served as a high-quality in-domain dataset for task adaptation. During fine-tuning on the official WMT 2026 dataset, the encoder parameters were frozen while the decoder was updated.

Training Strategy II: To further increase corpus diversity, the OPUS corpus was augmented with approximately seven million synthetic parallel sentence pairs generated using Meta’s multilingual NLLB model. The technique of pivot-based data augmentation was followed for this purpose. A base Transformer model was trained on the combined OPUS and synthetic corpus before being fine-tuned on the official WMT 2026 dataset.

For both strategies, identical model architectures and optimization settings were employed to ensure a fair comparison.

Parameter	Value
Optimizer	Adam
Learning Rate	0.001
Adam β_1	0.9
Adam β_2	0.998
Label Smoothing	0.1
Gradient Clipping	1.0

Precision	BF16
Beam Size	5
n-best	1
Selection Criteria of checkpoint	Validation Perplexity
Penalty Length	Average

Table 4: Training Configuration

4 Experiments

4.1 Experiment I: Tokenization Comparison

To evaluate the impact of subword segmentation on Arabic–Hindi translation, we trained two independent Transformer-based NMT systems using Byte Pair Encoding (BPE) (Sennrich et al., 2016) and SentencePiece (SPM) (Kudo and Richardson, 2018), respectively. Both systems employed identical architectures, training configurations, datasets, and decoding settings, differing only in the tokenization method. Separate source and target vocabularies of 32,000 subword units were constructed for Arabic and Hindi in both approaches, enabling a fair comparison of the two segmentation techniques.

4.2 Experiment II: Impact of Synthetic Data Augmentation

To assess the effectiveness of synthetic data augmentation, two training strategies were explored. The first strategy trained the base model using only the cleaned OPUS parallel corpus before fine-tuning on the official WMT 2026 dataset, while the second strategy incorporated approximately seven million synthetic sentence pairs generated using Meta’s multilingual NLLB model during base model training, followed by the same fine-tuning procedure. All other model architectures, optimization settings, and training parameters

were kept identical to isolate the effect of synthetic data.

5 Results

5.1 Evaluation Metrics

System performance was evaluated using three widely adopted automatic machine translation metrics: BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), and COMET (Rei et al., 2022). BLEU measures n-gram overlap between the generated translation and the reference, chrF++ evaluates character-level and word-level similarity, and COMET is a neural evaluation metric that estimates translation quality based on semantic adequacy and fluency. Together, these metrics provide a comprehensive assessment of translation performance.

5.2 Experimental Results

Table 5 shows a clear direction-dependent effect of synthetic data. For Hindi→Arabic (Hi→Ar), adding synthetic data improves all metrics for both tokenizers: BLEU increases from 7.2 to 9.5 with BPE and from 7.0 to 9.3 with SPM, while chrF++ and COMET show similar improvements. In contrast, Arabic→Hindi (Hi→Ar) performance decreases substantially, with BLEU dropping from 25.3 to 17.7 for BPE and from 25.8 to 18.1 for SPM. The similar trends across both tokenizers suggest that data composition has a stronger effect than the choice of subword segmentation. Overall, BPE and SPM show small numerical differences across the experiments (maximum BLEU difference: 0.5), without statistical significance testing.

It can be noticed that adding synthetic data is hurting only Arabic to Hindi translation direction. A possible explanation is the different role of synthetic Arabic in the two directions. The synthetic corpus was generated by translating English into Arabic using NLLB. Thus, for Hindi→Arabic, synthetic Arabic serves as the target language and may provide

additional lexical coverage. For Arabic→Hindi, it serves as the source language, where generation errors or unnatural constructions may introduce source-side noise and unreliable alignments. This may explain the observed degradation, although further controlled analysis is required to establish causality.

Direction	tokenizer	dataset	BLEU	chrF++	Comet
Hi→Ar	BPE	OPUS	7.2	42.7	0.8166
Hi→Ar	BPE	OPUS + Syn	9.5	53.6	0.8262
Hi→Ar	SPM	OPUS	7.0	42.5	0.8100
Hi→Ar	SPM	OPUS + Syn	9.3	53.5	0.8167
Ar → Hi	BPE	OPUS	25.3	51.0	0.6969
Ar → Hi	BPE	OPUS + Syn	17.7	43.46	0.6381
Ar → Hi	SPM	OPUS	25.8	51.1	0.7014
Ar → Hi	SPM	OPUS + Syn	18.1	43.6	0.6401

Table 5: Results on Devtest (WMT 2026 official data)

Furthermore, The lower BLEU scores for Hindi→Arabic direction should not be interpreted as directly indicating lower overall translation quality than Arabic→Hindi direction. BLEU measures n-gram overlap and can be sensitive to lexical and word-order differences between translation directions. Notably,

Hindi→Arabic achieves substantially higher COMET scores (0.8100–0.8262) than Arabic→Hindi (0.6381–0.7014), despite its lower BLEU scores. Therefore, we compare systems primarily within each translation direction rather than directly comparing absolute BLEU values across directions.

6 Conclusion

This paper presented Team MTG-LIRA’s submission to the WMT 2026 Low resource Arabic–Asian Shared Task for Arabic–Hindi and Hindi–Arabic machine translation. We developed a data-centric Neural Machine Translation pipeline by combining the official WMT training data with large-scale external parallel corpora from OPUS and additional synthetic parallel data generated using Meta’s multilingual NLLB model. Extensive corpus preprocessing, including language-specific normalization, quality filtering, deduplication, and alignment verification, was performed to construct a high-quality training corpus. We further investigated the impact of two widely used subword tokenization techniques, SentencePiece and Byte Pair Encoding, under two different training strategies. Experimental results demonstrate that while both tokenization methods achieved comparable performance, the effectiveness of synthetic data augmentation varied across translation directions, highlighting the importance of corpus quality and domain compatibility in multilingual Neural Machine Translation. In future work, we plan to investigate higher-quality synthetic data generation, multilingual and transfer learning approaches, and large language model-assisted data augmentation to further improve translation quality for Arabic–Asian language pairs.

Acknowledgments

We would like to thank the organizers of the WMT 2026 Low-Resource Arabic-Asian Language Translation Shared Task for providing the datasets and facilitating this research. Furthermore, we extend our gratitude to the developers of OpenNMT, OPUS and NLLB for

making their models and data publicly accessible.

References

- Ahmed, S. A., Joshi, N., Mathur, I., & Katyayan, P. (2019). Implications of English as a Pivot Language in Arabic-Hindi Machine Translation. *International Journal of Engineering and Advanced Technology*, 8(4), 271–277.
- Bhagyashree Wagh, Harish Bapat, Neha Gupta, and Saurabh Salunkhe. (2025). Transformers: Leveraging OpenNMT and Transfer Learning for Low-Resource Indian Language Translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1227–1232, Suzhou, China. Association for Computational Linguistics.
- Costa-jussà, M. R., et al. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics
- Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).
- Laskar, S. R., Alam, F., Paul, B., Ahmad, I., Lydia, M. S., & Dadure, P. K. (2026). Findings of the WMT 2026 Shared Task on Low-Resource Arabic-Asian Language Translation. *Proceedings of the Eleventh Conference on Machine Translation*.
- Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Papineni, K., et al. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. *ACL*.
- NLLB Team et al. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv:2207.04672*.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 2017. *Attention Is All You Need*. *Advances in Neural Information Processing Systems (NeurIPS 30)*.