NAJD-MT: High-Fidelity Saudi Najdi–English Training Data for Bidirectional Neural Machine Translation

Nour Qandos[1*], **Samar Ahmed**[1], **Omer Nacar**[2], **Ahmad Alrabghi**[1]
Rahaf Al Hallay[1], Aya Hamod[1], Shaden Alsuhaim[1]
[1]NAMAA Community, Riyadh, Saudi Arabia
[2]Tuwaiq Academy, Riyadh, Saudi Arabia
[*]nooramerq0@gmail.com

Abstract

Dialectal Arabic remains significantly underrepresented in parallel resources for direct machine translation with English, particularly for regional varieties such as Saudi Najdi Arabic. In this work, we introduce **NAJD-MT**, a systematically constructed Saudi Najdi→English parallel corpus designed for training bidirectional neural machine translation models. Starting from the Saudi Arabic Dialectal Annotated (SADA) dataset, we generate English translations using GPT-4.1 and subsequently apply cross-lingual embedding-based cosine similarity filtering to improve semantic alignment and reduce translation noise. We analyze the impact of varying semantic similarity thresholds on corpus size and downstream translation performance. Using the constructed datasets, we train and evaluate multiple Transformer-based models, including NLLB-200, OPUS-MT, mBART, and AraT5v2, in both Najdi→English and English→Najdi directions. Experimental results demonstrate that stricter semantic filtering (cosine ≥ 0.7) consistently improves translation quality despite reducing dataset size, highlighting that data purity plays a critical role in dialectal machine translation training. Our findings provide a reproducible framework for constructing high-fidelity dialect English parallel corpora and emphasize the importance of semantic alignment filtering in low-resource dialectal settings.

Keywords: Dialectal Arabic, Machine translation, Semantic Similarity Filtering, Cross-Lingual Embeddings, Multilingual Transformers

1. Introduction

Recent advances in multilingual neural machine translation (NMT) have significantly improved translation quality across high-resource language pairs (Tang et al., 2020; Costa-Jussà et al., 2022). However, dialectal Arabic remains substantially underrepresented in parallel corpora, particularly for direct translation between regional dialects and English. While Modern Standard Arabic (MSA) benefits from large-scale datasets and extensive pretraining coverage, regional varieties such as Najdi Arabic continue to face severe data scarcity and limited standardized resources.

Dialectal Arabic differs from MSA at multiple linguistic levels, including phonology, morphology, syntax, and lexicon (Habash, 2010). These divergences are not merely surface-level variations; they often involve distinct morphosyntactic constructions and culturally embedded expressions. As a result, models trained predominantly on MSA frequently struggle when applied to dialectal input, leading to semantic drift and reduced robustness in conversational translation scenarios.

Previous research has explored pivot-based approaches for dialectal machine translation, typically translating dialectal Arabic to MSA before translating to English (Salloum and Habash, 2013). While pivoting can mitigate data scarcity, it may introduce cumulative translation errors and obscure dialect-specific nuances. Direct dialect-to-English translation has been shown to be preferable when sufficient parallel data is available, yet constructing such corpora remains costly and labor-intensive.

More recently, large language models (LLMs) have been employed for synthetic data generation and corpus augmentation in low-resource settings (Wang et al., 2023). However, automatically generated translations may introduce semantic inconsistencies or hallucinations, particularly when applied to dialectal input. Ensuring high-quality alignment therefore requires additional filtering mechanisms.

Our main contributions are:

- We propose a scalable pipeline for constructing semantically aligned Najdi→English parallel data using large language model translation and embedding-based filtering.
- We analyze the impact of semantic similarity thresholds on corpus size and downstream NMT performance.
- We provide empirical evidence that stricter semantic filtering improves bidirectional dialectal translation quality.

2. Related Work

Neural machine translation for Arabic has achieved strong performance for Modern Standard Arabic (MSA), largely due to the availability of large parallel corpora and multilingual pretraining (Tang et al.,

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 88–93
May 11, 2026.

2020; Costa-Jussà et al., 2022). However, dialectal Arabic remains under-resourced and presents additional challenges stemming from lexical variation, morphosyntactic divergence, and orthographic inconsistency (Habash, 2010; Zaidan and Callison-Burch, 2014).

Early approaches to dialect-to-English translation relied on pivoting through MSA (Salloum and Habash, 2013), where dialectal input is first normalized or translated into MSA before translating to English. Although pivoting reduces data requirements, it may introduce cascading errors and weaken dialect-specific semantic preservation. More recent studies emphasize the importance of direct dialect-to-English translation when parallel resources are available, as this approach better preserves colloquial and culturally embedded expressions (Bouamor et al., 2018).

Despite these advances, Saudi dialects particularly Najdi remain sparsely represented in publicly available parallel datasets. The SADA corpus (Alharbi et al., 2024) provides dialectal speech transcripts but does not include aligned English translations, leaving a gap for direct dialect–English MT training.

Synthetic data augmentation has become a common strategy in low-resource machine translation. Back-translation (Sennrich et al., 2016) and self-training approaches have demonstrated that synthetic parallel data can significantly improve NMT performance. More recently, large language models (LLMs) have been used to generate synthetic instruction-following and translation data (Wang et al., 2023), offering scalable solutions for data construction.

However, automatically generated translations may introduce semantic drift, hallucination, or stylistic inconsistencies, particularly in dialectal contexts where linguistic variability is high. Consequently, quality control mechanisms are necessary to ensure that synthetic parallel data maintains semantic fidelity.

Cross-lingual sentence embeddings enable semantic comparison across languages by mapping text into a shared representation space (Reimers and Gurevych, 2019; Gao et al., 2021). Such representations have been widely used for bitext mining and parallel sentence extraction (Schwenk et al., 2021; Artetxe and Schwenk, 2019). Cosine similarity thresholds are commonly employed to filter noisy sentence pairs and improve corpus quality.

In contrast to large-scale web-mined bitext extraction, our work applies embedding-based cosine similarity filtering to synthetic dialect-generated translations, evaluating how semantic filtering thresholds impact downstream NMT training performance.

While prior research has explored dialectal MT, pivot-based strategies, synthetic data augmentation, and embedding-based bitext mining, limited work has examined systematic pipelines for constructing semantically aligned Saudi dialect–English training corpora. Our contribution lies in integrating LLM-based translation with cross-lingual semantic filtering for Najdi Arabic and empirically analyzing how alignment thresholds influence bidirectional NMT training outcomes.

3. Methodology

This section describes the construction of the NAJD-MT corpus and the training setup used to evaluate its impact on bidirectional Najdi→English neural machine translation (NMT). The overall pipeline is illustrated in Figure 1. The process consists of three stages: (1) source preparation and cleaning, (2) parallel data construction with semantic alignment filtering, and (3) bidirectional NMT training and evaluation.

3.1. Source Dataset and Preprocessing

We construct our corpus from the Saudi Arabic Dialectal Annotated (SADA) dataset (Alharbi et al., 2024), a speech-based resource containing transcribed Saudi dialect conversations. Since SADA includes both Najdi and Hejazi varieties, we restrict our experiments to the Najdi subset to reduce dialectal variability and promote homogeneity during model training.

Because SADA is derived from speech-to-text (STT) transcription, it contains noise artifacts, segmentation errors, and occasional transcription markers. We apply the following preprocessing steps: Removal of single-character or non-linguistic fragments, Deduplication of identical utterances, and Removal of STT artifacts such as the token "غير واضح". After preprocessing, the cleaned Najdi corpus serves as the source-side data for parallel corpus construction.

3.2. LLM-Based Translation

To construct the English side of the parallel corpus, we translate the cleaned Najdi sentences using GPT-4.1. Prior to full translation, we conducted a pilot comparison across multiple large language models (GPT-4.1, LLaMA-4-Maverick, Gemini-2.5-Pro, Claude-Sonnet-4, and Command-R-Plus) on 20 representative samples. Models were assessed across four criteria: hallucination rate, named entity preservation, semantic faithfulness, and dialectal term capture. valuation was conducted via blind human assessment by a bilingual Najdi Arabic–English expert and automated LLM-as-a-judge scoring using chatGPT. GPT-4.1

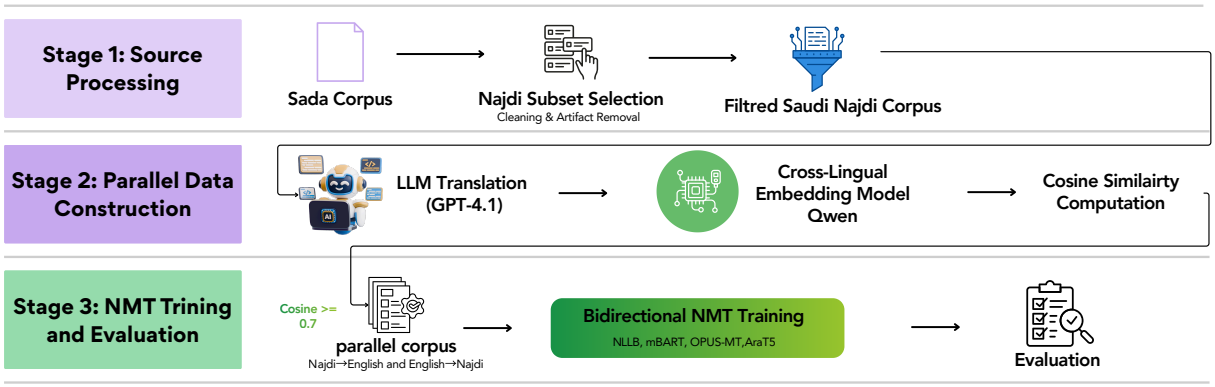

Figure 1: Methodology Pipeline

Stage	Remaining	Removed
Original	69,978	0
Single-character removal	69,945	33
Duplicate removal	69,663	282
Artifact removal	68,158	1,505
Embedding filter ($\tau \geq 0.6$)	54,280	15,698
Embedding filter ($\tau \geq 0.7$)	30,421	39,557

Table 1: Dataset size after each filtering stage.

demonstrated stronger semantic preservation and fewer hallucinations during manual inspection and was therefore selected for full-dataset translation.

The output of this stage consists of synthetic Najdi→English sentence pairs.

3.3. Semantic Similarity Filtering

Although large language models provide high-quality translations, automatic generation may introduce semantic drift, contextual expansion, or partial meaning loss. To ensure alignment fidelity, we apply cross-lingual embedding-based filtering using **Qwen3-Embedding-8B** (Zhang et al., 2025), a multilingual embedding model producing 4096-dimensional representations trained on large-scale multilingual corpora. Table 1 summarizes the corpus size after each preprocessing and filtering stage.

Let $x_{ar} \in R^d$ and $x_{en} \in R^d$ denote the cross-lingual sentence embeddings of the Arabic source sentence and its English translation, respectively. The cosine similarity between the two embeddings is computed as:

$$sim(x_{ar}, x_{en}) = \frac{x_{ar}^{\top} x_{en}}{\|x_{ar}\|_2 \, \|x_{en}\|_2} \quad (1)$$

where $\| \cdot \|_2$ denotes the ℓ_2 norm. A sentence pair (x_{ar}, x_{en}) is retained in the filtered dataset if:

$$sim(x_{ar}, x_{en}) \geq \tau \quad (2)$$

where $\tau \in \{0.6, 0.7\}$ is the semantic similarity threshold.

3.4. Human Validation

To validate the reliability of embedding-based filtering, we conduct targeted manual inspection. First, all sentence pairs with cosine similarity ≤ 0.3 (158 samples) are reviewed. The majority exhibit severe semantic drift, confirming that low similarity correlates with misalignment.

Second, we randomly sample 170 sentence pairs with cosine similarity ≥ 0.7. None contain critical mistranslations, and only minor deviations are observed in a small subset.

Validation is performed by two bilingual evaluators with native Najdi Arabic proficiency and advanced English competency, who independently assessed translation correctness and flagged dialectal inaccuracies, with disagreements resolved through discussion.

3.5. Dataset Splits

For each filtered dataset ($\tau = 0.6$ and $\tau = 0.7$), we apply an 80/20 split into training and validation sets. Splitting is performed after filtering to prevent leakage of low-quality pairs into evaluation data. No sentence overlap exists across splits.

3.6. Neural Machine Translation Models

To evaluate the effectiveness of the constructed corpus, we fine-tune multiple Transformer-based models:

- mBART (Tang et al., 2020)

- NLLB-200-distilled (Costa-jussà et al., 2022)
- OPUS-MT (Tiedemann and Thottingal, 2020)
- AraT5v2 (Elmadany et al., 2023)

We train models in both Najdi→English and English→Najdi directions.

3.7. Training Setup

All models are fine-tuned using the HuggingFace Transformers framework. Training is conducted on NVIDIA RTX 5090 GPUs with the following hyperparameters: Learning rate: 2×10^{-5}, Linear decay scheduler, Batch size: 64, Maximum epochs: 30, Early stopping based on validation BLEU. Evaluation is performed using BLEU to ensure standardized and reproducible scoring.

4. Experiments and Results

We evaluate bidirectional Najdi→English translation using BLEU (Papineni et al., 2002) as the primary automatic metric. Additionally, we report BERTScore (Zhang et al., 2020) for the top-performing model in each translation direction to provide complementary semantic-level evaluation. All scores are reported on the validation set.

4.1. Najdi → English

Table 2 presents BLEU scores for translation from Najdi Arabic to English under two semantic filtering thresholds ($\tau = 0.6$ and $\tau = 0.7$).

Across all models, training on the more strictly filtered dataset ($\tau = 0.7$) consistently yields higher BLEU scores despite the reduction in corpus size. NLLB-200 achieves the highest BLEU score (30.60) at $\tau = 0.7$, followed closely by OPUS-MT (29.10) and mBART (28.40). BERTScore evaluation of NLLB-200 at $\tau = 0.7$ further supports this result (P: 0.9406, R: 0.9370, F1: 0.9387), indicating strong semantic adequacy beyond surface n-gram overlap. AraT5v2 shows slightly lower performance but remains competitive. These results suggest that stricter semantic filtering improves translation quality in the Najdi→English direction.

4.2. English → Najdi

Table 3 reports results for English→Najdi translation. Overall BLEU scores are substantially lower than in the reverse direction, reflecting the greater difficulty of generating dialectal output. mBART achieves the highest BLEU score (15.60) at $\tau = 0.7$, followed closely by OPUS-MT (15.10). Similar to the previous direction, models trained on the $\tau = 0.7$ dataset outperform those trained on $\tau = 0.6$. BERTScore evaluation of mBART at $\tau = 0.7$ (P: 0.9122, R: 0.9101, F1: 0.9111) further reveals that despite low BLEU scores, the model captures substantial semantic meaning even when exact dialectal surface forms are not reproduced. The relatively small performance differences across models suggest that dialect generation remains challenging and may be constrained by limited dialectal representation in multilingual pretraining.

Model	$\tau = 0.6$	$\tau = 0.7$
AraT5v2	25.53	26.80
mBART	26.0	28.40
OPUS-MT	25.83	29.10
NLLB-200	27.60	**30.60**

Table 2: Najdi→English BLEU scores.

Model	$\tau = 0.6$	$\tau = 0.7$
AraT5v2	12.40	13.50
mBART	14.43	**15.60**
OPUS-MT	13.70	15.10

Table 3: English→Najdi BLEU scores.

5. Discussion

Our experiments highlight two central findings. First, stricter semantic filtering consistently improves translation performance across all evaluated models and both translation directions. Despite reducing the corpus size from 54,280 to 30,421 sentence pairs, increasing the cosine similarity threshold from $\tau = 0.6$ to $\tau = 0.7$ yields consistent BLEU improvements. This suggests that, in dialectal settings, semantic alignment quality may outweigh data volume. Noisy or weakly aligned sentence pairs appear to introduce greater harm during training than the benefit gained from additional quantity.

Second, a clear directional asymmetry is observed. Najdi→ The English translation achieves substantially higher BLEU scores than the reverse direction. This asymmetry likely reflects the broader representation of English in multilingual pretraining corpora, as well as the relative difficulty of generating dialectal output. While multilingual NMT models are generally exposed to English during pretraining, dialectal Arabic particularly Najdi remains sparsely represented. Consequently, generating dialect-specific morphology, vocabulary, and colloquial constructions presents a greater challenge than translating dialectal input into standardized English.

Across models, NLLB-200 demonstrates the strongest performance in the Najdi→English direction, suggesting that large-scale multilingual pre-

training combined with distillation remains effective even for underrepresented dialectal input. However, performance differences between models are relatively modest, particularly in the English→Najdi direction, indicating that architectural variation alone is insufficient to overcome dialectal data scarcity.

Overall, the results reinforce the importance of corpus construction methodology in low-resource dialectal machine translation. The consistent gains obtained through embedding-based filtering demonstrate that semantic similarity can serve as an effective quality-control mechanism for synthetic parallel data.

6. Error Analysis

To systematically investigate the model's limitations in English→Najdi direction, an error analysis was conducted using a representative sample of outputs generated by the best-performing model in this translation direction. Overall, the model demonstrates strong performance in translating short, everyday, and relatively simple sentences. However, several recurring issues were observed, particularly related to gender agreement, where masculine and feminine forms are occasionally confused (e.g., translating possessive forms incorrectly). The model also shows clear limitations when handling longer and more complex sentences. This may be attributed to the loss of semantic nuances and contextual information during the translation process between English and Najdi. Notably, the model's performance improves when the source sentence is well-structured and unambiguous.

Following the same methodology, an error analysis was performed on samples generated by the best-performing model in the Najdi→English direction. One prominent issue is the repetition of certain words, especially named entities, where the model tends to repeat frequently recognized tokens (e.g., proper names such as "Fatimah") at the expense of preserving the overall context. Despite this, the model demonstrates relatively better handling of longer texts, producing translations that are moderately coherent. This improvement may be due to the relative clarity and structural richness of the source Najdi Arabic text compared to the generated English output.

7. Conclusion

In this work, we introduced Najdi-MT, a semantically aligned Najdi→English parallel corpus constructed through large language model translation and cross-lingual embedding-based filtering. We demonstrated that applying stricter semantic similarity thresholds improves downstream bidirectional neural machine translation performance, even at the cost of reduced dataset size. Our experiments show that data quality plays a critical role in dialectal machine translation, particularly in low-resource settings. While multilingual models such as NLLB-200 and mBART perform competitively, improvements remain constrained by limited dialectal representation during pretraining. The observed directional asymmetry further highlights the challenges of generating dialectal Arabic compared to translating it into English. These findings suggest that careful corpus construction and alignment filtering are essential for advancing dialect-aware machine translation systems. Future work will extend Najdi-MT to additional Saudi dialects, incorporate broader domain coverage.

8. References

Sadeen Alharbi, Areeb Alowisheq, Zoltán Tüske, Kareem Darwish, Abdullah Alrajeh, Abdulmajeed Alrowithi, Aljawharah Bin Tamran, Asma Ibrahim, Raghad Aloraini, Raneem Alnajim, et al. 2024. Sada: Saudi audio dataset for arabic. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10286–10290. IEEE.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3197–3203.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

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*.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, 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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil

Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Octopus: A multitask model and toolkit for Arabic natural language generation. In *Proceedings of ArabicNLP 2023*, pages 232–243, Singapore (Hybrid). Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6894–6910.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

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. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Wael Salloum and Nizar Habash. 2013. Dialectal arabic to english machine translation: Pivoting through modern standard arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–358.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 86–96.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt — building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508.

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.