Interpretable Semantic Denoising via Neuron-level Functional Separation for Low-resource Multilingual Translation

Jiarui Zhang*†(✉), Mingzhe Lu*†, Yifan Deng*†
*Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
†School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
Email: {zhangjiarui,lumingzhe,dengyifan}@iie.ac.cn

***Abstract*—Low-resource multilingual neural machine translation is essential for extending language technologies to underserved communities, yet progress remains limited by scarce parallel data and fragile cross-lingual generalization. Existing solutions primarily rely on data augmentation or architectural modifications, which often fail to disentangle semantics from syntax, leaving noisy representations that degrade translation quality. We propose Neuron-Distinction-Based Semantic Denoising (NDS), a neuron-level framework that generates semantically equivalent data via entropy-guided controlled sampling, compares parallel model activations to identify syntactic-noise neurons, and prunes them before fine-tuning. Our findings reveal a functional dissociation between semantic and syntactic neurons, providing a novel perspective on how large-scale neural networks encode cross-lingual invariants. Experimental results demonstrate that NDS bridges the performance gap between low- and high-resource translation while remaining highly parameter-efficient.**

***Index Terms*—Interpretable and Explainable AI, Transformer Networks, Representation & Reasoning, Neural Network Applications**

I. Introduction

Multilingual neural machine translation (MNMT) models [1]–[3] enable cross-lingual transfer through shared representations, offering parameter efficiency, better transfer across related languages, and practical scalability.

Despite these benefits, MNMT still struggles in low-resource settings. Existing augmentation and architectural strategies often fail to disentangle semantics from syntax, leading to noisy cross-lingual representations and degraded translation quality. Recent low-resource MNMT efforts reveal persistent limitations: data-centric methods (e.g., back-translation [1], cross-lingual transfer [2]) incur semantic drift from synthetic noise; pre-trained models (mBART [3]) conflate semantics and syntax, propagating noise; and architectural approaches (adapters [4]) mitigate interference but operate at coarse granularity. Critically, no method targets neuron-level noise propagation or enables interpretable pruning without architectural overhead—a gap that severely compromises low-resource performance where noise amplification disproportionately dominates behavior.

To address these issues, we introduce **Neuron-Distinction-Based Semantic Denoising (NDS)**. As shown in Fig. 1, NDS (1) expands data via entropy-guided sampling, (2) pre-trains two identical models on original vs. augmented data, (3) identifies and prunes syntax-sensitive neurons by activation comparison, and (4) fine-tunes the pruned model. This neuron-level intervention improves low-resource robustness without extra parameters; probing on UD and PAWS-X (Sec. VI) links functional decoupling to translation gains.

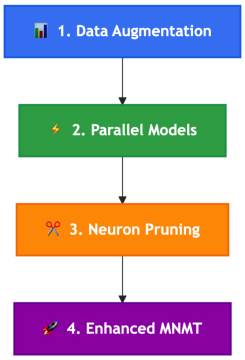

Fig. 1. The brief flowchart of NDS

Our contributions are summarized as follows:

- We propose NDS, a neuron-level semantic denoising framework that improves low-resource MNMT by decoupling semantic and syntactic noise.
- We design an entropy-guided controlled sampling mechanism that generates high-quality, semantically equivalent data and enables precise identification of syntactic-noise neurons via parallel-model comparisons.
- Extensive experiments show that NDS improves translation quality without adding parameters and remains competitive with large language models.

II. Related Work

Prior pruning work mainly targets model compression using generic importance metrics such as magnitude or gradient criteria [5], while representation probing focuses on post-hoc analysis of what models encode. In contrast, NDS is an online, neuron-level semantic purification mechanism: it actively constructs semantically equivalent inputs and uses dynamic

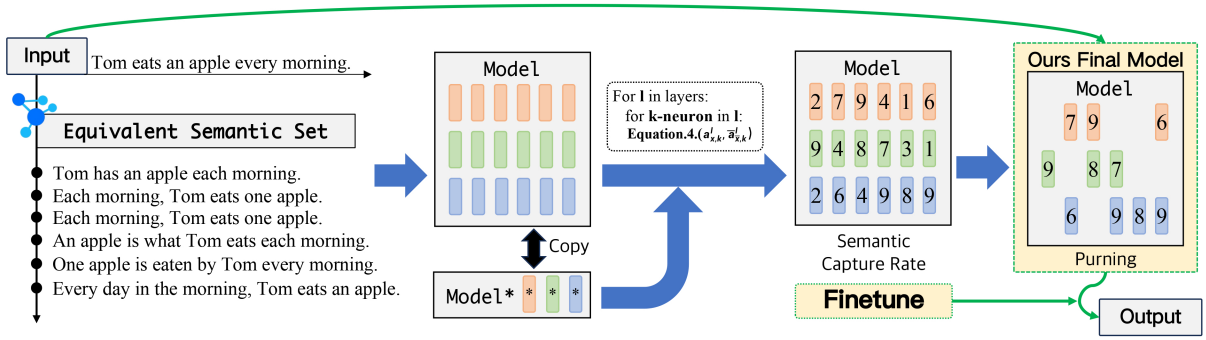

Fig. 2. The whole training process of the proposed method (NDS).

contrast to isolate syntax-sensitive neurons for pruning. Table I summarizes the key differences.

Low-resource MNMT and data augmentation. Low-resource MNMT typically relies on transfer learning and data-centric strategies such as back-translation or synthetic parallel data to compensate for limited supervision [1], [2]. Pre-trained multilingual sequence-to-sequence models (e.g., mBART) further improve generalization by leveraging large-scale pretraining [3], yet they can still entangle semantics and syntax, which makes them sensitive to noisy synthetic inputs [6]. Our approach complements these lines by explicitly controlling semantic equivalence during data expansion and using neuron-level signals to suppress syntactic noise.

Parameter-efficient adaptation and pruning. Adapter-based methods and related lightweight tuning strategies aim to reduce task-specific parameters while maintaining performance [4]. Pruning methods, in contrast, often focus on compression and speed rather than semantic disentanglement [5]. NDS shares the efficiency motivation but departs by targeting functional roles of individual neurons using paired, semantically equivalent inputs, which provides a mechanism-level explanation for performance gains rather than solely reducing capacity.

Interpretability and probing. Probing studies analyze what linguistic properties are encoded in representations, typically via offline classifiers or diagnostic tasks [7], [8]. Our method is closer in spirit to interpretability work, but it integrates probing signals into the training loop to actively refine the model. This bridges analysis and optimization by turning representational diagnostics into actionable pruning decisions.

TABLE I
COMPARISON WITH RELATED PARADIGMS.

Feature	Pruning	Probing	NDS (Ours)
Core goal	Compression	Offline analysis	Online denoising
Granularity	Weights/layers	Vectors/statistics	Single neurons
Feedback	Static importance	Passive observation	Dynamic contrast sets

III. APPROACH

The overview of our **Neuron-Distinction-Based Semantic Denoising (NDS)** framework is illustrated in Fig. 2.

Algorithm 1 Neuron-Distinction-Based Semantic Denoising (NDS)

Require: Original corpus $\mathcal{D}_{ori} = \{(\mathbf{x}_i, \mathbf{y}_i)\}$, reverse model θ, thresholds ε (entropy) and τ (pruning)
Ensure: Pruned and fine-tuned model M^*
1: Generate semantic equivalents $\mathcal{D}_{syn}$ via entropy-guided controlled sampling with θ
2: Initialize two identical models M_1, M_2
3: Pre-train M_1 on $\mathcal{D}_{ori}$ and M_2 on $\mathcal{D}_{syn}$
4: Compute neuron-wise semantic capture rate and importance $I(k)$ across paired inputs
5: Prune neurons $\mathcal{P} = \{k \mid I(k) < \tau\}$ in the target model
6: Fine-tune the pruned model on $\mathcal{D}_{ori}$ to obtain M^*

A. Theoretical Distinction of NDS

Unlike traditional neuron-pruning methods that primarily target model compression or generic importance metrics (e.g., magnitude- or gradient-based pruning) [5], NDS is designed for **semantic purification**.

NDS is distinguished by a **dynamic contrastive probing mechanism**:

- **Decoupling vs. Compression:** Conventional pruning targets redundant neurons to reduce size. NDS targets *interference neurons* that propagate syntactic noise, which hinders language-agnostic generalization.
- **Contextual Stability as a Proxy for Semantics:** NDS operationalizes *semantic invariance* at the neuron level by using entropy-guided variations as a stress test for stability across syntactic realizations.

Together, these steps form a closed loop: controlled perturbation, contrastive identification, and targeted pruning of noise pathways.

B. Semantic Data Expansion

Machine translation is inherently one-to-many: for a target sentence $\mathbf{y}$, multiple source sentences in $\Omega(\mathbf{y})$ can be semantically equivalent [9]. Training on a single pair $(\mathbf{x}, \mathbf{y})$ underexplores this space and weakens generalization under syntactic variation [6].

We therefore use entropy-guided controlled sampling to balance semantic fidelity and syntactic diversity: when confidence

is high (low entropy), decode with $\arg\max$; when confidence is low (high entropy), sample to introduce variation [10].

Given a target sentence $\mathbf{y} = \{y_1, \ldots, y_T\}$ and a pre-trained reverse translation model θ, we generate a source-language sentence $\overline{\mathbf{x}} = \{\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_T\}$ token-by-token. At decoding step t, the sampling strategy is governed by the distribution steepness $\mathcal{H}$:

$$\begin{cases} \overline{\mathbf{x}}_t \sim P\left(\cdot \mid \overline{\mathbf{x}}_{<t}, \mathbf{y}; \theta\right), & \text{if } \mathcal{H} \geq \varepsilon \\ \overline{\mathbf{x}}_t = \arg\max\left(P\left(\cdot \mid \overline{\mathbf{x}}_{<t}, \mathbf{y}; \theta\right)\right), & \text{else} \end{cases} \tag{1}$$

where ε is the steepness threshold, and $\mathcal{H}$ is computed as the conditional entropy of the distribution [11]:

$$\mathcal{H} = -\sum_j P\left(x^j \mid \overline{\mathbf{x}}_{<t}, \mathbf{y}; \theta\right) \log P\left(x^j \mid \overline{\mathbf{x}}_{<t}, \mathbf{y}; \theta\right) \tag{2}$$

where $P\left(x^j \mid \overline{\mathbf{x}}_{<t}, \mathbf{y}; \theta\right)$ represents the conditional probability of the j-th word in the vocabulary appearing after the sequence $\overline{\mathbf{x}}_{<t}$ given the target sentence $\mathbf{y}$.

Expected property of entropy-guided sampling. Let $p_t(v) = P\left(v \mid \overline{\mathbf{x}}_{<t}, \mathbf{y}; \theta\right)$ and $X_t \sim p_t$. The expected token surprisal equals the entropy:

$$\mathbb{E}_{X_t \sim p_t}\left[-\log p_t(X_t)\right] = -\sum_v p_t(v) \log p_t(v) = \mathcal{H}. \tag{3}$$

Thus, larger $\mathcal{H}$ encourages exploration, while smaller $\mathcal{H}$ favors deterministic decoding. The threshold ε provides a principled switch between fidelity and diversity.

C. Neuron Similarity Analysis and Pruning

We identify noise neurons via semantic capture rate, i.e., activation consistency across semantically equivalent inputs. Using the original corpus $\mathcal{D}_{ori}$ and synthesized set $\mathcal{D}_{syn}$, we pre-train two identical translation models.

For neuron k in layer l, we define its **Semantic Capture Rate (SCR)** as the consistency of activations across semantically equivalent inputs from the paired set $\mathcal{D}_{pair} = \{(\mathbf{x}, \overline{\mathbf{x}})\}$:

$$\mathrm{SCR}(k) = \frac{1}{|\mathcal{D}_{pair}|} \sum_{(\mathbf{x}, \overline{\mathbf{x}}) \in \mathcal{D}_{pair}} \mathrm{CosSim}\left(\mathbf{a}_{\mathbf{x},k}^{(l)}, \overline{\mathbf{a}}_{\overline{\mathbf{x}},k}^{(l)}\right) \tag{4}$$

where $\mathbf{a}_{\mathbf{x},k}^{(l)}$ and $\overline{\mathbf{a}}_{\overline{\mathbf{x}},k}^{(l)}$ are the activations of neuron k at layer l in the original and augmented models for inputs $\mathbf{x}$ and $\overline{\mathbf{x}}$, respectively. SCR measures semantic stability: high values indicate robust semantic encoding, while low values indicate syntactic sensitivity. **Rationale.** Cosine similarity is bounded in $[-1, 1]$, so $\mathrm{SCR}(k)$ is a normalized stability score. Consider a simple decomposition for semantically equivalent inputs:

$$\mathbf{a}_{\mathbf{x},k}^{(l)} = \mathbf{s}_k + \mathbf{n}_k, \quad \overline{\mathbf{a}}_{\overline{\mathbf{x}},k}^{(l)} = \mathbf{s}_k + \overline{\mathbf{n}}_k, \tag{5}$$

where $\mathbf{s}_k$ is the shared semantic component and $\mathbf{n}_k, \overline{\mathbf{n}}_k$ are zero-mean, isotropic noise with variance σ^2. Then, using a first-order approximation of cosine similarity,

$$\mathbb{E}\left[\mathrm{CosSim}\left(\mathbf{s}_k + \mathbf{n}_k, \mathbf{s}_k + \overline{\mathbf{n}}_k\right)\right] \approx \frac{\|\mathbf{s}_k\|^2}{\|\mathbf{s}_k\|^2 + \sigma^2}, \tag{6}$$

which is monotonic in the signal-to-noise ratio. Hence, neurons that consistently encode semantics across equivalent inputs yield higher SCR, while syntax-sensitive neurons with large noise components yield lower SCR, justifying its use as a pruning criterion. Based on semantic capture rates, we compute neuron importance [12], [13] as:

$$I(k) = \mathrm{SCR}(k) \cdot \left(1 - \mathrm{Var}_{(\mathbf{x}, \overline{\mathbf{x}})}\left[\|\mathbf{a}_{\mathbf{x},k}^{(l)}\|\right]\right) \tag{7}$$

where Var measures activation variance across equivalent inputs, penalizing neurons with unstable magnitudes. Finally, we prune neurons failing the semantic capture threshold τ:

$$\mathcal{P} = \{k \mid I(k) < \tau\} \tag{8}$$

where $\mathcal{P}$ is the set of neurons to prune. This targeted pruning removes syntax-sensitive noise neurons and improves generalization; in practice we prune 30% neurons.

Mechanism hypotheses and evidence. (H1) Entropy-guided sampling improves equivalent-set quality (A1 drop, Table VI). (H2) Dual-model SCR better isolates syntax-sensitive neurons (A2 drop). (H3) SCR+variance pruning removes syntactic noise rather than semantic signal (PG/SRG split, Table VIII). (H4) A 30% ratio best balances denoising and retention (Table VII).

D. Fine-tuning the Pruned Model

After pruning, we fine-tune on $\mathcal{D}_{ori}$ with standard translation objectives (e.g., cross-entropy), adapting cleaner representations to the target task [14].

IV. Experiments

A. Datasets

We evaluate on WMT-10 [15], [16] and PMIndia [17]. WMT-10 contains 32.5M sentence pairs across 10 language pairs (En $\leftrightarrow$ Fr/Cs/De/Fi/Lv/Et/Ro/Hi/Tr/Gu), spanning mixed resource levels and typologies. PMIndia is a low-resource benchmark with 26–53K pairs per direction across nine Indian languages (bn, gu, hi, ml, mr, kn, or, te, pa). Together they test NDS under both resource abundance and scarcity. An overview of the datasets is in Table II. We report three settings: many-to-English ($X \rightarrow$ En), English-to-many (En $\rightarrow X$), and many-to-many ($X \rightarrow X$).

TABLE II
Overview of the datasets.

Dataset	X $\leftrightarrow$ en	language families	# sent. per direction
PMIndia	{bn, gu, hi, ml, mr, kn, or, te, pa}	Indo-European, Dravidian	26-53K
WMT-10	{fr, cs, de, fi, lv, et, ro, hi, tr, gu}	Indo-European, Uralic, Altaic	0.08-10M

TABLE III
SUMMARY OF TRAINING HYPERPARAMETERS.

Configuration	Value
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98$)
Learning Rate	4×10^{-4} (inverse square-root schedule)
Warm-up Steps	4,000
Label Smoothing	0.1
Dropout Rate	0.1
Batch Size	64k tokens (update freq = 2)
Pruning Ratio	30%
Random Seed	42 (fixed)

B. Model Settings and Preprocessing

All experiments are implemented in **FAIRSEQ** [18]. We use SentencePiece [19] with a shared 250K vocabulary, set the entropy threshold to $\varepsilon = 2.5$ (validation-tuned), and truncate all inputs/outputs to 256 tokens for consistent comparison across NMT and LLM baselines.

C. Implementation Details and Reproducibility

To ensure reproducibility, we provide a comprehensive list of hyperparameters and environmental settings in Table III.

1) Computational Resources and Efficiency: **Hardware:** Experiments are conducted on a cluster of 8 $\times$ NVIDIA A100 (80GB) GPUs to accommodate the dual-model pre-training stage. **Checkpointing:** We validate every epoch and average the top-5 checkpoints by validation loss.

D. Baselines and Evaluation Metrics

Our method is compared against several representative baselines: 1) BiNMT [20]: bilingual NMT models trained separately for each language pair; 2) MNMT [21]: standard multilingual NMT model trained on all language pairs; 3) mBART [3]: a 12-layer pre-trained multilingual sequence-to-sequence model fine-tuned on the translation task; 4) DeltaLM [22]: a pre-trained multilingual translation model based on incremental pre-training; 5) mGPT [23]: a large language model with 1.3B parameters, evaluated using in-context learning with 5 examples; 6) Qwen3-1.7b and Qwen3-4b [24], [25]: large language models with 1.7B and 4B parameters, respectively, evaluated using in-context learning with five examples. We use BLEU scores as our primary evaluation metric.

E. Fairness in LLM Comparison

To ensure a rigorous and fair comparison with Large Language Models (LLMs), we adopt the 5-shot In-Context Learning (ICL) paradigm for mGPT [23] and Qwen3 [24], [25]. **Prompt template:** We use a consistent bilingual instruction format to guide translation: *"Translate the following [Source Language] text into [Target Language]: [Source Sentence]. Translation: "*.

Exemplar selection: The five demonstrations are randomly sampled from each language pair's validation set. We report average BLEU over three random exemplar sets.

Input constraints: For all LLMs, maximum generation length is 256 and decoding is greedy, reducing variance from prompt sampling and decoding randomness.

F. Results and Analysis

We evaluate **NDS** on WMT-10 and PMIndia. Table IV and Table V show consistent gains with strong parameter efficiency.

Strong low-resource performance: On PMIndia, NDS reaches 33.0 BLEU (+2.7 over MNMT). On WMT-10, it outperforms non-LLM baselines and stays competitive with larger LLMs, especially on low-resource subsets.

Parameter efficiency: NDS uses far fewer parameters (242M) than mGPT (1.3B), Qwen3-1.7b (1.7B), and Qwen3-4b (4B), while matching or surpassing them on several evaluated settings.

Robust cross-lingual generalization: NDS remains strong across typologically diverse pairs and resource levels, indicating improved transfer.

V. Ablation Study and Sensitivity Analysis

To quantify each component and key hyperparameters, we run ablations on PMIndia and WMT-10.

A. Component-wise Ablation

We define the following variants:

- **A0 (Full NDS):** The proposed framework with all components.
- **A1 (w/o Entropy Sampling, Greedy):** Replacing entropy-guided sampling with simple greedy decoding for data expansion.
- **A2 (Single Model SCR):** Computing SCR from paired inputs $(\mathbf{x}, \overline{\mathbf{x}})$ using a single model, rather than comparing two parallel models.
- **A3 (Augmentation Only):** Using expanded data for training without any neuron pruning.
- **A4 (Pruning Only):** Applying SCR-based pruning on the original corpus without data expansion.
- **A5 (w/o Variance Term):** Pruning neurons based solely on $\mathrm{SCR}(k)$, removing the stability penalty $(1 - \mathrm{Var}[\cdot])$ in (7).

Fairness of A2/A4. All variants use the same backbone, training budget, and optimizer schedule. For A2, we still use the paired inputs $(\mathbf{x}, \overline{\mathbf{x}})$ but compute SCR from two forward passes of the *same* model (with the same training configuration as A0), removing only the dual-model comparison while keeping the data and training cost unchanged. For A4, we do not train on $\mathcal{D}_{syn}$; synthetic pairs are used only to estimate SCR, and the final fine-tuning is performed on $\mathcal{D}_{ori}$ alone.

Analysis of Results: Table VI shows that full NDS (A0) outperforms all variants under bidirectional averaging. The A3/A4 gaps indicate that augmentation and pruning are complementary; the larger drop in A2 validates the importance of dual-model comparison; and A5 confirms that the variance penalty improves pruning reliability.

TABLE IV

BLEU SCORES OF TRANSLATION TASKS (LEFT: ENGLISH→10 LANGUAGES; RIGHT: 10 LANGUAGES→ENGLISH). AVG$_{high}$ DENOTES AVERAGE OF HIGH-RESOURCE LANGUAGES (FR-ET, 6 LANGUAGES IN TOTAL), AVG$_{low}$ DENOTES AVERAGE OF LOW-RESOURCE LANGUAGES (RO-GU, 4 LANGUAGES IN TOTAL). '*': MGPT DOES NOT SUPPORT GUJARATI(GU).

En→X test set						X→En test set			
System		#Params	Avg$_{high}$	Avg$_{low}$	Avg$_{all}$	System	Avg$_{high}$	Avg$_{low}$	Avg$_{all}$
En→1	BiNMT	242M	24.3	14.1	20.2	BiNMT	28.5	17.6	23.0
En→X	MNMT	242M	24.4	16.7	21.3	MNMT	31.7	23.3	26.3
	mBART	611M	24.7	17.8	22.0	mBART	32.4	27.0	28.2
	DeltaLM	360M	25.6	19.8	23.3	DeltaLM	30.7	28.7	29.9
	NDS	**242M**	**27.0**	**22.6**	**25.2**	**NDS**	**34.8**	**30.7**	**33.2**
X→X	MNMT	242M	24.5	17.5	21.5	MNMT	30.8	23.8	26.0
	mBART	1.2B	24.5	18.1	21.9	**NDS**	**34.4**	**30.9**	**33.0**
	NDS	**242M**	**26.7**	**22.8**	**25.1**				
LLMs	mGPT*	1.3B	25.9	20.1	24.0	mGPT*	33.9	28.0	31.9
	Qwen3-1.7b	1.7B	27.2	21.9	25.1	Qwen3-1.7b	35.1	29.1	32.7
	Qwen3-4b	4B	28.2	22.1	25.8	Qwen3-4b	35.9	30.0	33.5

TABLE V

BLEU SCORES ON PMINDIA DATASET (ALL LOW-RESOURCE LANGUAGES)

Model	**Test Set**		**Avg. (All Directions)**
	En→X	**X→En**	
BiNMT	9.7	13.7	11.7
MNMT	29.3	31.3	30.3
NDS	**32.1**	**33.9**	**33.0**

TABLE VI

ABLATION RESULTS (BLEU) ON PMINDIA AND WMT-10 (AVERAGE OVER $X \rightarrow En$ AND $En \rightarrow X$).

Variant	**PMIndia**	**WMT-10**
A0 (Full NDS)	**33.0**	**29.2**
A1 (w/o Entropy Sampling)	31.8 (-1.2)	27.5 (-1.7)
A2 (Single Model SCR)	30.9 (-2.1)	26.2 (-3.0)
A3 (Augmentation Only)	31.2 (-1.8)	26.5 (-2.7)
A4 (Pruning Only)	30.4 (-2.6)	25.8 (-3.4)
A5 (w/o Variance Term)	32.1 (-0.9)	27.9 (-1.3)

B. Sensitivity to Pruning Ratio (A6)

We vary pruning ratio r from 10% to 40% (step 10%). Table VII shows a bell-shaped trend.

TABLE VII

IMPACT OF PRUNING RATIO r ON BLEU (PMINDIA: BIDIRECTIONAL AVERAGE; WMT-10: $X \rightarrow En$ AVERAGE).

Ratio r	10%	20%	**30%**	40%
PMIndia	31.2	32.4	**33.0**	32.2
WMT-10	30.5	32.1	**33.2**	31.8

Discussion: At 10%, denoising is insufficient; at 40%, useful semantic neurons are over-pruned. The peak at 30% supports our choice.

VI. INTERPRETABILITY: PROBING NEURONAL FUNCTIONS

A core premise of NDS is that SCR-based pruning can effectively decouple semantic information from syntactic noise. To rigorously validate this claim and address concerns regarding functional specialization, we conduct probing experiments following Conneau et al. [7].

A. Experimental Setup

We divide the neurons in the final layer of the encoder into two distinct functional subsets based on their Semantic Capture Rate (SCR) from (4):

- **Semantic-Rich Group (SRG):** The top 30% of neurons with the highest SCR, which we hypothesize encode stable language-agnostic meanings.
- **Pruned Group (PG):** The 30% of neurons with the lowest SCR, which we hypothesize are sensitive to superficial syntactic variations.

For each group, we train a linear probe (logistic regression) on activations to predict linguistic properties, so performance reflects representation quality rather than probe capacity.

B. Probing Tasks and Datasets

We evaluate two task types across languages:

- **Syntactic Probing (Bigram Shift):** On UD treebanks, the probe predicts whether adjacent words were swapped, measuring sensitivity to local word order.
- **Semantic Probing (Paraphrase Identification):** On PAWS-X [26], the probe predicts semantic equivalence under high lexical overlap, requiring robustness to syntactic traps.

C. Results and Discussion

The results, summarized in Table VIII, provide strong empirical support for our "semantic denoising" hypothesis.

TABLE VIII
PROBING ACCURACY (%) OF DIFFERENT NEURON SUBSETS ON SYNTACTIC (BSHIFT) AND SEMANTIC (PAWS-X) TASKS.

Neuron Subset	Syntactic	Semantic
All Neurons (Baseline)	71.2	72.1
Pruned Group (PG)	**81.4**	45.2
Semantic-Rich Group (SRG)	52.4	**88.7**

The **Pruned Group (PG)** is much stronger on Bigram Shift (+29% over SRG), indicating syntax sensitivity, while the **Semantic-Rich Group (SRG)** is strongest on PAWS-X. This dissociation supports that NDS removes syntax-sensitive noise pathways rather than generic low-frequency features.

VII. CONCLUSION

We propose Neuron-Distinction-Based Semantic Denoising (NDS), a neuron-level framework for low-resource MNMT. By combining entropy-guided semantic expansion, dual-model SCR-based pruning, and fine-tuning, NDS decouples semantic and syntactic information without adding parameters. On PMIndia, NDS reaches 33.0 BLEU (+2.7 vs. MNMT); on WMT-10 $X \rightarrow En$, it reaches 33.2 BLEU (+6.9 vs. MNMT), while remaining competitive with much larger LLMs. These results support neuron-level functional separation as an effective path for robust low-resource translation and suggest broader applicability to generation tasks requiring semantic stability.

VIII. ACKNOWLEDGMENTS

We extend our sincere gratitude to the anonymous reviewers for their invaluable and insightful feedback. This research was supported by the National Natural Science Foundation of China (Grant No.U21B2009).

REFERENCES

[1] B. Sennrich, B. Haddow, and A. Birch, "Back-translation: A simple and effective method for low-resource machine translation," in *COLING*, 2016.

[2] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[3] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[4] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[5] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[6] M. Ott, M. Auli, D. Grangier, and M. Ranzato, "Analyzing uncertainty in neural machine translation," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3956–3965.

[7] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, "What you can cram into a single vector: Probing sentence embeddings for linguistic properties," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2126–2136.

[8] J. Hewitt and C. D. Manning, "A structural probe for finding syntax in word representations," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4129–4138.

[9] P. Koehn, *Neural machine translation*. Cambridge University Press, 2020.

[10] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[11] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[12] M. Zhang, M. Tu, F. Zhang, and S. Liu, "A cross search method for data augmentation in neural machine translation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 071–11 075.

[13] J. Zhang, H. Huang, Y. Hu, P. Guo, and Y. Xie, "Importance-based neuron selective distillation for interference mitigation in multilingual neural machine translation," in *International Conference on Knowledge Science, Engineering and Management*. Springer, 2023, pp. 140–150.

[14] X. Wei, H. Yu, Y. Hu, R. Weng, L. Xing, and W. Luo, "Uncertainty-aware semantic augmentation for neural machine translation," *arXiv preprint arXiv:2010.04411*, 2020.

[15] Y. Wang, C. Zhai, and H. H. Awadalla, "Multi-task learning for multilingual neural machine translation," *arXiv preprint arXiv:2010.02523*, 2020.

[16] S. Ma, J. Yang, H. Huang, Z. Chi, L. Dong, D. Zhang, H. H. Awadalla, A. Muzio, A. Eriguchi, S. Singhal *et al.*, "Xlm-t: Scaling up multilingual machine translation with pretrained cross-lingual transformer encoders," *arXiv preprint arXiv:2012.15547*, 2020.

[17] B. Haddow and F. Kirefu, "Pmindia–a collection of parallel corpora of languages of india," *arXiv preprint arXiv:2001.09907*, 2020.

[18] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," *arXiv preprint arXiv:1904.01038*, 2019.

[19] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.

[22] S. Ma, L. Dong, S. Huang, D. Zhang, A. Muzio, S. Singhal, H. H. Awadalla, X. Song, and F. Wei, "Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders," *arXiv preprint arXiv:2106.13736*, 2021.

[23] O. Shliazhko, A. Fenogenova, M. Tikhonova, A. Kozlova, V. Mikhailov, and T. Shavrina, "mgpt: Few-shot learners go multilingual," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 58–79, 2024.

[24] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[25] Q. Team, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, vol. 2, 2024.

[26] Y. Yang, Y. Zhang, C. Tar, and J. Baldridge, "Paws-x: A cross-lingual adversarial dataset for paraphrase identification," 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)*, 2019, pp. 3687–3692.