Nepali Lemmatization with Multilingual Transformers: Intrinsic and Extrinsic Evaluation in a Low-Resource Setting

Sunil Regmi[1], **Sundeep Dawadi**[1], **Bal Krishna Bal**[2]
[1]Department of Artificial Intelligence, Kathmandu University, Dhulikhel, Kavre, Nepal
[2]Information and Language Processing Research Lab,
Department of Computer Science & Engineering, Kathmandu University, Dhulikhel, Kavre, Nepal
sunilregmi@ku.edu.np[1], sundeepdwd@gmail.com[1], bal@ku.edu.np[2]

Abstract

The Nepali language has a rich and complex morphology. Existing lemmatization research focuses on traditional rule-based or TRIE-based approaches. These methods often fail when encountering out-of-vocabulary or misspelled words. This paper investigates neural lemmatization for the under-resourced Nepali language using multilingual transformer models. We formulate lemmatization as a text-to-text generation problem and evaluate its impacts on downstream tasks by finetuning mBART-large-50, mT5-base, and mT5-small. The models were trained on a combination of publicly available and human-annotated word-lemma pair (8,000 instances) dataset. The performance is evaluated using Character Error Rate (CER), accuracy, character-level Bilingual Evaluation Understudy (BLEU), and morphological coverage. The mT5-base model achieved the highest overall performance. The model achieved 96.1% accuracy and a 1.1% CER using a learning rate of 5×10^{-4}. However, it showed slightly weaker performance in handling complex morphological variations. The mBART-large-50 model followed closely with 96.0% accuracy and 0.970 morphological coverage. To assess the efficacy of these models, we applied lemmatization to downstream tasks. In Hindi-Nepali cross-lingual alignment, performance improved significantly from 12.86% to 41.61% using mBART model. In information retrieval, the Mean Average Precision (MAP)@1 using binary index increased from 0.71 to 0.90 using mBART model. These results demonstrate that multilingual transformers effectively learn morphological transformations for low-resource languages through text-to-text generation.

Keywords: Nepali lemmatization, mT5, mBART, morphological processing

1. Introduction

In Natural Language Processing (NLP), lemmatization transforms inflected words into their base or root forms which reduces data sparsity and improves model performance across tasks such as sentiment analysis, text classification, and information retrieval(Bhat and Rai, 2012; Dave and Balani, 2015; Ghimire et al., 2024). Similarly, studies on word embeddings (Koirala and Niraula, 2021) have emphasized the need for robust lemmatization algorithms to reduce morphological complexity and improve semantic representation. This is particularly important for morphologically rich languages like Nepali, where a single lemma possesses numerous inflected variants (Prasain, 2011a).

As an Indo-Aryan language written in the Devanagari script, Nepali presents unique challenges for lemmatization. These include complex inflectional patterns, adpositional suffixation, orthographic variation, and scarce annotated resources (Yadav and Shakya, 2020). Table 1 provides representative examples of Nepali inflections and their canonical lemmas. For instance, consider the verb root गर् (gar) 'to do': it yields inflected forms such as गर्छ (*garchha*) 'does', गरेको (*gareko*) 'did/done', गर्दै (*gardai*) 'doing', and गरिरहेको (*gariraheko*) 'has been doing'. All of these map to the canonical lemma गर्नु (*garnu*) 'to do'. In an ideal lemmatization, this specific canonical form of verb is preferred as this is widely accepted in published dictionaries for Nepali. The गर् (gar) 'to do' variant is popular in every rule based method as this form is easy to work with suffix stripping method. Similarly, nouns take case-marking postpositions (e.g., घरमा (*gharamā*) 'in the house' $\rightarrow$ घर (*ghara*) 'house'), creating additional surface variation.

Inflected Form	Lemma	Gloss
गर्छ *(garchha)*	गर्नु *(garnu)*	'does'
गरेको *(gareko)*	गर्नु *(garnu)*	'did/done'
गर्दै *(gardai)*	गर्नु *(garnu)*	'doing'
खान्छु *(khānchhu)*	खानु *(khānu)*	'I eat'
घरमा *(gharamā)*	घर *(ghara)*	'in the house'

Table 1: Examples of Nepali inflections and their canonical lemmas.

These constraints highlight the need for scalable, data-driven methods capable of learning morphological transformations directly from linguistic examples, motivating the transition toward neural and transformer-based architectures.

Despite this complexity, Nepali remains a low-resource language with scarce annotated datasets for morphological analysis. To address this, we look toward multilingual encoder–decoder architectures such as multilingual Text-to-Text Trans-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3462–3469
11–16 May 2026.

fer Transformer (mT5) (Xue et al., 2021) and multilingual Bidirectional and Auto-Regressive Transformer (mBART) (Liu et al., 2020a). These models utilize large-scale cross-lingual pretraining. This process transfers linguistic knowledge from high-resource languages to low-resource ones. Such architectures have already improved performance in machine translation and text normalization for Indo-Aryan languages.

By leveraging pretrained representations, these models reduce the need for massive manual annotation. They provide increased robustness against linguistic noise. Furthermore, they generalize effectively across diverse inflectional patterns. This allows the models to mitigate vocabulary sparsity and handle previously unseen word forms. Their ability to learn shared linguistic structures makes them ideal candidates for Nepali lemmatization.

This paper provides empirical insights into the suitability of these transformers for Nepali and makes the following contributions:

1. A systematic evaluation of multilingual encoder-decoder models for Nepali lemmatization.
2. A comparative analysis of `mT5` and `mBART` for morphological normalization in a low-resource Nepali language.
3. An extrinsic evaluation via cross-lingual alignment and information retrieval improvements.

2. Related Work

Early research on Nepali lemmatization relied on rule-based and dictionary-driven approaches. Foundational systems such as (Bal and Shrestha, 2004; Bal et al., 2004; Bal and Shrestha, 2007) used handcrafted morphological analyzers which were built upon spelling and grammar checking tools. These systems applied manually designed linguistic rules to analyze and normalize inflected word forms. Subsequent studies introduced TRIE-based or hybrid affix-removal approaches (Bhat and Rai, 2012; Sitaula, 2013; Paul et al., 2014; Koirala and Shakya, 2020). These methods combined prefix–suffix stripping rules with dictionary lookups and achieved up to 90% accuracy on curated datasets.

However, these systems faced significant limitations. They lacked scalability and struggled with out-of-vocabulary (OOV) words. They also failed to process informal and misspelled text. Their performance further degrades when applied to raw text corpora due to inconsistent suffix usage is common in Nepali.

Modern NLP relies on tagged datasets like Universal Part-of-Speech (UPOS), eXtended Part-of-Speech (XPOS), and lemma-tagged corpora. These resources are widely available for high resource languages like Hindi and English. Unfortunately, they remain unavailable for Nepali (Nivre et al., 2020). For example, IndoWordNet (Bhingardive and Bhattacharyya, 2016) has only 6,000 synsets for Nepali compared to 40,000 for Hindi. This resource scarcity limits the development of word sense disambiguation tools. Some lexical resources exist within tools such as the Hunspell dictionary, which contains approximately 37,000 normalized word forms primarily developed for spelling correction (Bal et al., 2004). However, these are not well studied and optimized for complex morphological tasks. Research on transformers and neural models has highlighted the need for sentence or phrase level morphological studies as morphological processes are not confined to just words (Acikgoz et al., 2022).

Recently, several studies following multilingual transformer models have shown promising results for low-resource languages. The Cross-lingual Language Model-RoBERTa (XLM-R) (Conneau et al., 2019), `Trankit` (Nguyen et al., 2021), and other Bidirectional Encoder Representations from Transformers (BERT) based models can generalize effectively to various Devanagari-script languages and downstream tasks (Acharya et al., 2025; Subedi et al., 2024). Despite progress in neural lemmatization for high-resource languages, the systematic exploration of encoder-decoder based models such as mT5 and mBART for Nepali remains limited.

Existing approaches for Nepali lemmatization generally fall into two broad paradigms. The first includes rule-based and affix-stripping stemmers, which rely on predefined linguistic rules and suffix lists. While computationally efficient, these methods often fail to capture the full range of linguistic variability present in naturally occurring text. The second paradigm consists of neural morphological analyzers, which attempt to model morphological transformations using supervised learning frameworks. However, these approaches typically depend on rich linguistic annotations that Nepali lacks.

We address these limitations by treating lemmatization as a pure sequence-to-sequence generation task. This allows encoder-decoder models to learn the mapping between inflected forms and their lemmas. We investigate multilingual transformers without requiring explicit morphological tag supervision or POS conditioning, considering the low-resource constraints of Nepali. This approach is well suited for low-resource languages such as Nepali. We perform a comprehensive evaluation of this method using multiple performance metrics. Furthermore, we assess the practical impact of our models on downstream NLP

tasks, including information retrieval and cross-lingual alignment.

3. Methodology

3.1. Data Collection

We curated a dataset from human-annotated and publicly available sources to support the core lemmatization experiment. The primary dataset consists of 5,000 gold-standard word–lemma pairs obtained from the publicly available Nepali Lemmatization dataset (dpakpdl, 2025). To increase lexical diversity and expose the model to a broader range of morphological variations, we added 3,000 additional samples based on common Nepali morphological transformation rules and suffix variations observed in inflectional patterns (Prasain, 2011b). These additional pairs were manually verified for linguistic correctness by a native Nepali speaker with linguistic background. These 8,000 unique word-lemma pairs were then split into 80% and 20% for training and evaluation. We ensured strict separation between splits, with no overlap between the train and test split to prevent data leakage.

For the error analysis, we prepared an isolated error evaluation set of 100 words from each of the 7 parts of speech, totaling 700 words, from a dictionary where the input word itself acts as the gold-standard lemma (Nepal Academy, 2022). Their frequency data was obtained from the fineweb2 web corpus (Penedo et al., 2025).

3.2. Data Preparation

To frame lemmatization as a sequence-to-sequence generation task, each input word in the corpus was reformulated as an instruction-style prompt (e.g., *"lemmatize: <word>"*) to explicitly condition the model for the lemmatization task. We inserted the SentencePiece boundary marker ("_") between tokens before tokenization to preserve morphological boundaries and ensure consistent subword segmentation. All Nepali text was normalized to Unicode Normalization Form C (NFC) to maintain canonical representation for the Devanagari script.

Tokenization was performed using SentencePiece for both models (mT5 and mBART), which provides language-independent subword segmentation and supports shared multilingual vocabularies (Kudo and Richardson, 2018; Liu et al., 2020b). Each input–output pair was truncated or padded to a maximum sequence length of **32** tokens to match the subword-level nature of the tokenizer. Although the input typically contains only one word, multilingual tokenizers often split Devanagari words into multiple subword tokens. A maximum sequence length of 32 ensures sufficient capacity while avoiding unnecessary padding. Padding token IDs were replaced with `-100` to exclude them from loss computation. The final preprocessed dataset was built with the `datasets` library and stored as model-ready input-target pairs for supervised fine-tuning.

3.3. Model Implementation and Fine-Tuning

The lemmatization models were implemented and fine-tuned using the *Hugging Face Transformers* library, leveraging encoder–decoder architectures designed for text-to-text generation. Specifically, two multilingual sequence-to-sequence models were employed: `mT5-base` (∼580M parameters), `mT5-small` (∼300M parameters) (Xue et al., 2021) and `mBART-large-50` (∼610M parameters) (Liu et al., 2020b). Although both architectures are Transformer-based, they differ in pretraining objectives and multilingual coverage. Each model was fine-tuned on the lemmatization dataset described above. The input was formatted in a prompt-based structure (e.g., *lemmatize:* " खान्छु " → " खानु ") to enable instruction-based conditioning during training. To ensure consistency across models, the same data preprocessing pipeline was applied as discussed in the above section.

This comparative fine-tuning approach allowed the assessment of `mT5` and `mBART` in handling morphological normalization tasks under a unified text-to-text framework, highlighting differences in their multilingual pretraining strategies and transferability to low-resource Nepali text.

3.4. Evaluation

In this paper, we have used intrinsic and extrinsic evaluation methods. The intrinsic methods include: accuracy, which estimates the proportion of exact matches between predicted and gold-standard lemmas; Character Error Rate (CER) provides fine-grained measures of orthographic differences; and character-level BLEU assesses subword correctness in partial matches. Morphological coverage measures whether the predicted lemma preserves the correct morphological ending. Because Nepali lemmas typically end with characteristic suffixes, matching the final three characters provides a simple heuristic for evaluating morphological correctness.

For extrinsic evaluation, we evaluated our models firstly on the error analysis set. This set was used to evaluate how the model behaves with dictionary base forms. Secondly, we evaluate our

model on two downstream tasks as described in the results and discussions section below.

4. Experimentation

All experiments were conducted in a GPU-enabled environment with access to NVIDIA Tesla T4 GPUs (16 GB each) and ∼30 GB system RAM. The fine-tuning experiments evaluated the multilingual encoder–decoder models `mT5` and `mBART` across a hyperparameter grid.

We searched learning rates in $\{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}\}$ and batch sizes $\{2, 4, 8, 16\}$, training each configuration for 5–20 epochs. Optimization used AdamW (Loshchilov and Hutter, 2019) with weight decay $= 0.01$ and gradient clipping to stabilize training. The best checkpoint was selected by the lowest evaluation loss. For this task, we treat `CER` as the primary metric because it directly measures character-level edit distance between predicted and gold lemmas; accuracy, BLEU char, and morphological coverage are presented as complementary metrics.

5. Results and Discussions

The evaluation results in Table 2 show that `mT5-base` achieves the best overall performance, with the lowest Character Error Rate (CER) of **1.1%** and the highest accuracy of **96.1%**. Despite a slightly lower BLEU (Char) score (**0.980**) compared to `mBART-large-50`, its combination of low CER and marginally high accuracy indicates stronger consistency in surface-level prediction. A learning rate of 5×10^{-4} with batch size **16** yielded optimal convergence, benefiting from mT5's lighter regularization.

The `mBART-large-50` model follows closely, attaining a CER of **1.6%**, accuracy of **96.0%**, and the highest BLEU (Char) score (**0.986**). Its superior morphological coverage (**0.970**) reflects stronger generalization across inflectional forms, confirming mBART's denoising pretraining advantage for morphologically rich languages. The model needed a lower learning rate (1×10^{-5}) for stable optimization.

The smaller `mT5-small` variant performs competitively with a CER of **1.7%** and accuracy of **95.2%**, demonstrating efficiency despite reduced capacity. Its slightly lower morphological coverage (**0.955**) indicates limited performance in handling rare suffixes.

Overall, `mT5-base` excels in precision and has low error, while `mBART-large-50` offers greater morphological robustness and fluency. Both experiments confirm the effectiveness of multilingual encoder–decoder pretraining for Nepali lemmatization. The following sections present extrinsic evaluation results using our fine-tuned mT5-base, mT5-small, and mBART-large-50 models.

Metric	mBART-large-50	mT5-small	mT5-base
CER	0.016	0.017	**0.011**
Accuracy	0.960	0.952	**0.961**
BLEU (Char)	**0.986**	0.983	0.980
Morph. Coverage	**0.970**	0.955	0.964

Table 2: Comprehensive evaluation of multilingual transformer models for Nepali lemmatization.

While direct comparison is limited by differing evaluation datasets and preprocessing pipelines, the consistent superiority of neural methods across all metrics suggests that data-driven approaches better handle the morphological variability and spelling inconsistencies characteristic of real-world Nepali text.

5.1. Error Analysis

The model when evaluated on error analysis dataset showed a failure rate of 23.8% (167 out of 700 words). Crucially, cross-referencing these errors revealed that 106 of the 167 failed words (63.4%) had a corpus frequency of zero, representing Out-Of-Vocabulary (OOV) scenarios.

Without sentential context to resolve Part-of-Speech (POS) ambiguity, the neural architecture defaults to a strong **Corpus Frequency Bias** as shown in Table 3 below. Rather than performing a conservative identity mapping for rare or unseen lemmas, the model aggressively overwrites them with orthographically similar, high-frequency stems from its training distribution.

This frequency-driven fallback behavior highlights a structural vulnerability in subword-level neural lemmatization for low-resource languages using a multilingual transformer: the model learns to aggressively strip complex affixes or morphologically alter valid, low-frequency noun and verb roots (e.g., stripping the zero-frequency noun *kaṭagranthi* into the broken stem *kaṭagra*, or mapping the rare postposition *khātira* to the common verb *khānu*) where it attempts to satisfy its decoder's probabilistic language modeling prior.

5.2. Downstream Task Evaluation

5.2.1. Task Overview

We now turn to extrinsic evaluation to assess the impact of lemmatization in the following downstream tasks:

- Hindi–Nepali cross-lingual word alignment
- Nepali information retrieval (headline-to-content matching)

Dictionary Lemma	Predicted	POS Shift	Error Type	Transliteration
उसिन्निनु (0)	उनी (1.2M)	Verb → Pronoun	Frequency Bias	*usinninu* → *unī*
कुरेत (15)	कुरा (1.7M)	Adj. → Noun	POS Ambiguity	*kureta* → *kurā*
कटग्रन्थि (0)	कटग्र (0)	Noun → Noun	Overstemming	*kaṭagranthi* → *kaṭagra*
खातिर (64k)	खानु (260k)	Postpos. → Verb	Context. Omission	*khātira* → *khānu*

Table 3: Representative lemmatization failures when processing rare lemmas. Numbers in brackets represent corpus frequency.

These tasks were used exclusively for extrinsic evaluation and did not influence lemmatizer training.

5.2.2. Hindi–Nepali Cross-Lingual Alignment

VecMap is a tool that maps words from two different languages into a shared mathematical space to find translations (Artetxe et al., 2018). We used Hindi and Nepali word embeddings from official Fasttext implementation (Mikolov et al., 2018).

5.2.2.1 Data: A supervised Hindi–Nepali word alignment dataset comprising 14,500 bilingual pairs was constructed from a dictionary (Shivaramakrishna, 1977) to facilitate alignment learning for the VecMap model. For evaluation 3,000 validation word pairs were derived from a Nepali–Hindi parallel corpus generated synthetically from Nepali sentences (for Indian Language Technology), 2025). Exact translation matches were excluded as they are implicitly used during semi-supervised training. This data of Hindi-Nepali word pair consists of inflected word form as used in sentences. The Nepali counterpart obtained from corpus is considered to be the desired golden translation against which candidates from VecMap model will be compared.

5.2.2.2 Evaluation Settings: Three comparison conditions were evaluated. Firstly, the raw Nepali counterpart from parallel corpus was compared with the candidates from VecMap model. Secondly, only the Nepali counterpart from corpus was lemmatized and compared to candidates. Thirdly both the Nepali word and the Nepali candidates were lemmatized. Performance was measured using Top-1 accuracy, Top-5 accuracy, and MAP@5.

5.2.2.3 Example: Given the Hindi word गांव *gaon* 'Village', the VecMap model gives a list of Nepali counterparts (गाउँमा, गाउँ, गाउँका, गाउँस्थित, गाविस, गाउँको,) (available in the raw fasttext word vector) via cosine similarity. The raw prediction shall not match the gold translation गाउँलाई *gaunlai* 'For village' as per parallel corpus, but after lemmatization, the prediction is reduced to (गाउँ, गाउँ, गाउँ, गाउँस्थित, गाविस, गाउँ....) and the gold standard to गाउँ *gaun* 'Village', yielding a correct match.

The results in Table 4 demonstrate that lemmatization is critical for improving low-resource cross-lingual word alignment. In the baseline condition, matching raw Nepali words to raw VecMap candidates yields a low Accuracy@1 of 12.86% due to morphological sparsity. Lemmatizing only the gold target degrades performance further by forcing an unnatural comparison between canonical roots and highly inflected candidate vectors. Conversely, lemmatizing both the gold Nepali words and the candidate outputs have effectively reduced this inflectional variance. This approach increases Accuracy@1 to 41.61% and more than doubles the MAP@5 score (from 0.2031 to 0.4883) using mBART. Consistent with our intrinsic evaluations, mBART-large-50 marginally outperforms mT5-base which shows that robust lemmatization successfully maps dispersed morphological variants into singular, alignable semantic representations.

5.2.3. Nepali Information Retrieval

5.2.3.1 Data: 600 randomly selected Nepali news headlines and their corresponding article snippets were compiled from open-domain sources (Disisbig, 2025). The headline-to-content matching task is highly morphology-sensitive because headlines often use abbreviated or inflected forms that need to be matched against the differently-inflected forms in the main text body.

5.2.3.2 Methods: Two information retrieval (IR) baselines were evaluated:

- Term Frequency–Inverse Document Frequency (TF-IDF) vectorization
- Binary term indexing

For preprocessing, we first applied a standard stopword removal and tokenized the text using a whitespace-based approach. We then compared the IR performance on the original (unlemmatized) text and the lemmatized text to measure the effectiveness of our lemmatization as a stemming alternative.

Evaluation Condition	Acc@1 (%)	Acc@5 (%)	MAP@5
Gold Nepali word vs. Candidates	12.86	29.60	0.2031
L. Gold Nepali vs. Candidates (mT5-base)	10.57	17.91	0.1338
L. Gold Nepali vs. L. Candidates (mT5-base)	40.80	59.08	0.4809
L. Gold Nepali vs. Candidates (mBART-50)	10.38	17.81	0.1318
L. Gold Nepali vs. L. Candidates (mBART-50)	**41.61**	**59.99**	**0.4883**

Table 4: Cross-lingual translation accuracy under various evaluations (L. = Lemmatized). Symmetric lemmatization (L. vs. L.) provides the most significant gain in alignment.

5.2.3.3 Evaluation Metrics: Performance was evaluated using Accuracy@1, Accuracy@5, and MAP@5.

IR Method	Preprocessing	Acc@1	Acc@5	MAP@5
TF-IDF	Lemmatized(mT5-base)	0.5345	0.7884	0.6491
	Lemmatized(mBART)	0.5533	0.7821	0.6594
	Unlemmatized	0.4984	0.7821	0.6243
Binary Index	Lemmatized(mT5-base)	0.8245	0.9467	0.8779
	Lemmatized(mBART)	0.8574	0.9702	0.9064
	Unlemmatized	0.6317	0.8166	0.7159

Table 5: IR performance on headline-to-content matching task.

The results in Table 5 demonstrate that lemmatization consistently improves retrieval performance across both methods. The Binary Index method benefits most substantially, with mBART-based lemmatization improving Accuracy@1 from 0.6317 to 0.8574 (+35.7%). This is expected, as Binary Index relies on exact lexical matches, making it highly sensitive to morphological variation that lemmatization resolves. TF-IDF achieves more modest improvements, as its weighted term frequency partially compensates for inflectional differences. Across both methods, mBART-based lemmatization slightly outperforms mT5-based lemmatization, consistent with mBART's higher morphological coverage observed in the intrinsic evaluation.

6. Conclusion

The findings of this study establish a strong case for multilingual encoder–decoder architectures and their effectiveness in Nepali lemmatization. By framing lemmatization as a text-to-text generation task, we demonstrate that multilingual transformers can learn morphological transformations effectively, even with limited training data. Among the models that were evaluated, `mBART-large-50` achieves better performance in terms of morphological coverage, reaffirming the capacity of denoising autoencoder-based architectures like mBART to capture cross-lingual morphological patterns, particularly between typologically similar languages such as Hindi and Nepali.

The practical implications of these accurate lemmatization models are substantial for downstream applications in Nepali. They can serve as robust preprocessing modules for Machine Translation (MT), improve post-processing for Automatic Speech Recognition (ASR), aid in OCR correction, and significantly enhance indexing for Nepali search engines.

7. Limitations

While the presented results are promising, there are several limitations to acknowledge. The dataset lacks full morphological diversity and should be considered insufficient but it was valuable for our experiment. Moreover, the current study focuses only on word-level lemmatization, excluding contextual dependencies that influence meaning and cross part-of-speech morphological transformations. Future work should expand the dataset through expert linguistic annotation, incorporate sentence-level context for disambiguation, and explore hybrid architectures that combine neural models with rule-based morphological analyzers. Considering POS and polysemy awareness, proper introduction of rare dictionary words for self lemma representation in training, named and other entities, spelling errors, and its proper standard evaluation data are a must for a truly reliable and more deterministic lemmatization system for any language. This is often realized in a UPOS-style tagging scheme.

8. Ethics Statement

This work focuses on developing morphological tools for the Nepali language and does not introduce novel ethical risks. The datasets utilized for training and evaluation consist of publicly available open-domain texts, such as news headlines and snippets, which were responsibly sourced. We manually reviewed a subset of this data to ensure the absence of personally identifiable information (PII) or explicit harmful content. While pre-trained language models like mT5 and mBART may reflect societal biases present in their massive multilingual pre-training corpora, the task of lemmatization primarily involves syntactic transformations

rather than generating free-form semantic text, mitigating the risk of generating toxic outputs.

9. Data and Code Availability

To promote reproducibility and support future research in low-resource language processing, all artifacts associated with this study are openly accessible. The code for generating the synthetic morphology datasets, training scripts, and evaluation data can be found in our official repository at https://github.com/sunilRegmi-ai/Nepali-Lemmatizer. Furthermore, the fine-tuned models and corresponding tokenizers are hosted on the Hugging Face model hub (https://huggingface.co/sunilregmi/nepali-lemmatizerV1-mT5-base), allowing immediate access for the research community. All curated datasets adhere strictly to the licenses of their original sources.

10. Acknowledgments

This research was supported by Information and Language Processing Research Lab, Department of Computer Science & Engineering, Kathmandu University.

11. Bibliographical References

Darwin Acharya, Sundeep Dawadi, Shivram Saud, and Sunil Regmi. 2025. Paramananda@nlu of devanagari script languages 2025: Detection of language, hate speech and targets using fasttext and bert. In *Proceedings of the First Workshop on Challenges in Natural Language Understanding of Devanagari Script Languages*.

Emre Acikgoz, Tilek Chubakov, Müge Kural, Gözde İşgüder, and Deniz Yuret. 2022. Transformers on multilingual clause-level morphology. pages 100–105.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

B. K. Bal and P. Shrestha. 2004. The nepali spellchecker using hunspell framework. In *PAN Localization Working Papers 2004–2007*. PAN Localization Group.

Bal Krishna Bal, Basanta Karki, and Prajol Shrsetha. 2004. Nepali spellchecker 1.1 and the thesaurus, research and development.

Bal Krishna Bal and Prajol Shrestha. 2007. Architectural and system design of the nepali grammar checker.

Sunil M. Bhat and Rajendra Rai. 2012. Building morphological analyzer for nepali. *Journal of Modern Languages*, 22(1):45–58.

Sudha Bhingardive and Pushpak Bhattacharyya. 2016. Word sense disambiguation using indowordnet. In *The WordNet in Indian Languages*, pages 243–260. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 8440–8451. Association for Computational Linguistics.

R. Dave and P. Balani. 2015. Survey paper of different lemmatization approaches. In *Proceedings of the International Conference on Advances in Technology and Engineering (ICATEST 2015)*, volume 8, pages 366–370, India. IJRAT.

Dadhi Ram Ghimire, Sanjeev Panday, and Aman Shakya. 2024. Information extraction from a large knowledge graph in the nepali language. *National College of Computer Studies Research Journal*, 3(1):33–49.

Pravesh Koirala and Nobal B Niraula. 2021. Npvec1: Word embeddings for nepali-construction and evaluation. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 174–184.

Pravesh Koirala and Aman Shakya. 2020. A nepali rule-based stemmer and its performance on different nlp applications. *arXiv preprint arXiv:2002.09901*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of EMNLP: System Demonstrations*, pages 66–71.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation.

Yinhan Liu, Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2020b. Multilingual denoising pre-training for neural machine translation. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 162–176, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*, pages 1674–1685. Association for Computational Linguistics.

R. Paul, S. Sah, and S. Joshi. 2014. An affix removal stemmer for nepali language. *Proceedings of the National Conference on Information Technology*.

Balaram Prasain. 2011a. *A Computational Analysis of Nepali Morphology: A Model for Natural Language Processing*. Ph.D. thesis, Central Department of Linguistics, Tribhuvan University, Kathmandu, Nepal.

Puskar Prasain. 2011b. Nepali lemmatizer: A rule based approach. In *Proceedings of the 9th International Conference on Natural Language Processing (ICON)*, Kharagpur, India. NLP Association of India.

Chiranjibi Sitaula. 2013. A hybrid algorithm for stemming of nepali text. *Intelligent Information Management*, 5(4).

B. Subedi, S. Regmi, B. K. Bal, and P. Acharya. 2024. Exploring the potential of large language models (llms) for low-resource languages: A study on named-entity recognition (ner) and part-of-speech (pos) tagging for nepali language. In *Proceedings of the 2024 Joint International Conference on Computational Intelligence and Language Technology (CILT)*. Cited by 9.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Ankur Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 483–498, Online. Association for Computational Linguistics.

S. Yadav and S. Shakya. 2020. A neural morphological analyzer for nepali language. In *Proceedings of LREC*.

12. Language Resource References

Disisbig. 2025. *Nepali News Dataset*. Kaggle Datasets. PID https://www.kaggle.com/datasets/disisbig/nepali-news-dataset. Accessed 1 October 2025.

dpakpdl. 2025. *Gold Data for Manually Annotated Corpus in the NepaliLemmatizer Project*. GitHub. PID https://github.com/dpakpdl/NepaliLemmatizer. Accessed 1 October 2025.

CFILT (Computational Foundations for Indian Language Technology). 2025. *RoundTripOCR-Nepali*. Hugging Face Datasets. PID https://huggingface.co/datasets/cfilt/RoundTripOCR-nepali. Post-OCR error correction dataset for Nepali language using RoundTripOCR technique, accessed 1 October 2025.

Nepal Academy. 2022. *Pragya Nepali Brihat Shabdakosh*. Nepal Pragya Pratisthan, 10th (Revised).

Nivre, Joakim and de Marneffe, Marie-Catherine and Ginter, Filip and Hajić, Jan and Manning, Christopher D. and Pyysalo, Sampo and Schuster, Sebastian and Tyers, Francis and Zeman, Daniel. 2020. *Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection*. European Language Resources Association (ELRA). PID https://universaldependencies.org.

Guilherme Penedo and Hynek Kydlíček and Vinko Sabolčec and Bettina Messmer and Negar Foroutan and Amir Hossein Kargaran and Colin Raffel and Martin Jaggi and Leandro Von Werra and Thomas Wolf. 2025. *FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language*.

C. V. Shivaramakrishna. 1977. *Trilingual Dictionary: Hindi–English–Nepali*. Jai Gyan (via Digital Library of India). PID https://archive.org/details/dli.language.1977. Digitised version, accessed 1 October 2025.