Bidirectional English-Nepali Machine Translation System for the Legal Domain

Shabdapurush Poudel[1], **Bal Krishna Bal**[1], **Praveen Acharya**[2]
[1]Department of Computer Science and Engineering, Kathmandu University, Nepal
[2]School of Computing, Dublin City University, Ireland
{poudelshabda@gmail.com, bal@ku.edu.com, acharyaprvn@gmail.com}

Abstract

Nepali, a low-resource language belonging to the Indo-Aryan language family and spoken in Nepal, India, Sikkim, and Burma has comparatively very little digital content and resources, more particularly in the legal domain. However, the need to translate legal documents is ever-increasing in the context of growing volumes of legal cases and a large population seeking to go abroad for higher education or employment. This underscores the need for developing an English-Nepali Machine Translation for the legal domain. We attempt to address this problem by utilizing a Neural Machine Translation (NMT) System with an encoder-decoder architecture, specifically designed for legal Nepali-English translation. Leveraging a custom-built legal corpus of 125,000 parallel sentences, our system achieves encouraging BLEU scores of 7.98 in (Nepali → English) and 6.63 (English → Nepali) direction.

Keywords: English-Nepali, Low-resource, Legal Domain MT, Machine Translation, Neural Machine Translation

1. Introduction

Machine Translation (MT) Systems are performing better lately with advanced methods and techniques coming along the way in Deep Learning and Natural Language Processing. Correspondingly, the reliability of MT systems and the trust of the general public towards them have also increased.

Large Language Models (LLMs) are offering a helping hand to Machine Translation (MT) systems for languages that don't have a lot of digital resources (low-resource languages) (Moslem et al., 2023). They act as a kind of "platform" that can be fine-tuned utilizing different aspects of a specific language. This flexibility largely facilitates for creating entirely new and more robust MT systems for these languages.

The transition from a Statistical Machine Translation System (SMT) to Neural Machine Translation (NMT) has been reasonably smooth for high-resource languages but this has not been the case for low-resource languages. The primary reason behind this is that the NMT models are more data-hungry. To make things worse, the challenges of developing a suitable dataset for domain-specific work are manifold.

Nepali, which is the official language of Nepal and spoken in parts of India and Burma is a low-resource language (Bal, 2004) with considerably fewer resources and has limited research in the field despite the growing interest (Duwal and Bal, 2019); (Chaudhary et al., 2020); (Acharya and Bal, 2018). This scarcity of resources extends to domain-specific MT applications, particularly within the legal domain, where the lack of specialized translation tools presents a significant challenge.

In this research work, we have:

- Developed the first transformer-based bidirectional Machine Translation (MT) system (Vaswani et al., 2017) for English-Nepali and vice-versa in the legal domain, specifically focusing on legal terminology and nuances.
- Created a parallel corpus consisting of 125k sentences in the Nepali legal domain, a pioneering effort in this field.

2. Related Works

Machine Translation (MT) systems for Nepali have primarily focused on general domains, leaving a notable gap in addressing the specific requirements of legal translation. This lack of domain-specific tools impedes efficient and accurate legal communication in Nepali. However, insights from studies conducted in other languages offer valuable perspectives and methodologies for addressing this gap.

(Defauw et al., 2019) explored the use of Recurrent Neural Network (RNN)-based MT for legal content in Irish, highlighting challenges and dataset requirements for optimal results. Their study emphasizes the importance of domain-specific considerations in legal translation tasks.

Additionally, discussions on resource sharing for under-resourced European languages by (Bago et al., 2022) provide an understanding of potential works and challenges in the legal domain. This study stresses on the collaborative efforts needed to overcome resource limitations in addressing legal translation needs.

(Martínez-Domínguez et al., 2020) developed a customized Neural Machine Translation (NMT) system named "LexMachina," explicitly tailored for legal contexts in French. Their work showcases the effectiveness of specialized NMT systems in achieving high translation accuracy in legal domains.

Similarly, (Briva-Iglesias et al., 2024) analyzed various state-of-the-art models in Large Language Models (LLM) and NMT for legal translations across multiple language pairs. Their study offers valuable insights into the effectiveness of different technology approaches in legal translation tasks.

A common theme among these studies is the utilization of domain-specific corpora tailored explicitly for legal translation tasks. These specialized datasets play a crucial role in enhancing translation accuracy and addressing the unique linguistic nuances present in legal documents.

Despite advancements in related language pairs, such as Nepali-English translation, previous studies primarily focused on general domains, utilizing Transformer models. Works by (Duwal and Bal, 2019) and (Garcia et al., 2020) achieved promising results, setting the foundation for further experimentation with NMT models in the Nepali legal domain.

Moreover, (Nemkul and Shakya, 2021) explored alternative translation methods beyond state-of-the-art NMT approaches using RNN with LSTM(Long Short-Term Memory) providing a valuable understanding of potential avenues for experimentation in Nepali legal translation.

Overall, while the lack of domain-specific works in Nepali legal translation presents challenges, insights from existing studies offer valuable guidance and methodologies for addressing this gap. Our study aims to build upon this foundation and contribute to developing specialized translation tools tailored for the Nepali legal domain.

3. Methodology

3.1. Data Collection

Our research faced an initial challenge concerning the lack of a suitable parallel dataset for the legal domain in Nepali. Previous works exploring Nepali Machine Translation (MT) relied primarily on general corpora for various language pairs. While we initially considered adopting a general corpus for our project, we quickly dropped the idea keeping into consideration the following reasons:

- Legal translations predominantly use a passive voice and tone.

- Legal language possesses unique characteristics distinct from general discourse. Employing a general corpus could introduce noise and bias, hindering the translation accuracy for legal terminology and nuances.

- Utilizing a general corpus would require extensive filtering and data cleaning to extract domain-specific content, leading to inefficiency and potential loss of valuable domain-specific data.

Therefore, we undertook the extensive task of creating a new, domain-specific dataset tailored to our project. This involved:

- Manual translations by legal professionals: We commissioned experts to translate legal documents, including constitutional acts, court cases, and general legal proceedings, ensuring linguistic accuracy and domain expertise. Confidentiality agreements ensured sensitive information was redacted.

- Website scraping: To expand the dataset, we utilized custom legal keywords to filter and collect relevant legal documents from the Supreme Court website and news websites focusing on legal topics[1]. However, this raw data required significant cleaning to remove noise and errors.

3.2. Dataset

Through the efforts mentioned in the previous section, we built a final dataset of approximately 125,000 parallel sentences (Table 1). The curated dataset included a balanced mix of general and complex sentence structures while excluding shorter sentences for overall quality in the legal domain. The sentences consisted of legal terminologies which helped in the better training of the model. Shorter sentences were removed during filtering, to improve the general quality of the training data thereby matching with the general trend of legal texts (long and complex sentences).

Corpus Source	Corpus Size
Manually translated data	60K
Legal website scraped data	25K
News site scraped data	40k

Table 1: Data source and corpus size. The data mentioned are cleaned from noise and filtered.

3.3. Data Preprocessing

For this work, we collected data from multiple sources which were raw and considerably noisy. The noises were texts from non-Unicode encoding, XML, and HTML tags in the text and issues

[1] Documents: www.supremecourt.gov.np

with improper date and time conversion. Each scraped data was stored as an individual file and also cleaned for any noise individually.

Further preprocessing was done thus creating a final larger dataset following the steps below:

- Normalization and tokenization: We used IndicNLP[2] library (Kunchukuttan, 2020) to both normalize and tokenize the Nepali language, and then used Sacremoses[3] library for English language.

- Vocabulary Building: Translation cannot always include all the words in a model. Byte-Pair-Encoding (BPE)[4] (Sennrich et al., 2016) is also used in this work to learn the legal vocabulary for both source and target language. Earlier works on Nepali MT employed a small vocabulary size of 5k. Hence, for this work, we have used a vocabulary size of around 10000. Sentencepiece[5] library (Kudo and Richardson, 2018) was used to learn BPE for the source language.

3.4. Choosing the Right Model

Initially, we explored Recurrent Neural Networks (RNNs) as proposed by (Defauw et al., 2019). However, the results obtained revealed several weaknesses of RNNs for the English-Nepali pair. The training was slow and resource-intensive owing to the following reasons:

- Lack of parallelization and recursion: Processing took longer than expected.

- High memory usage: Dealing with large text segments strained resources.

- Limited long-range dependency handling: Capturing distant relationships within sentences was challenging.

Seeking significant improvements, we shifted our focus to Transformer-based Neural Machine Translation (NMT)(Vaswani et al., 2017).

The Transformer model, renowned for its fast training, inherent parallelization, and ability to handle long-range dependencies, offered a promising solution. Equipped with six encoder-decoder layers, the NMT architecture effectively addressed the challenges encountered in previous models, leading to demonstrably improved performance for both English-to-Nepali and Nepali-to-English translations.

Table 2: Tuning Parameters for models used in experimentation.

Parameters	RNN Model	NMT Model
Batch Size	32	96
Learning Rate	3e-3	5e-4
Epochs	100	150
Optimizer	Adam	Adam
Beam Size	5	6
Dropout rate	0.5	0.5

4. Experiments

For our experiments, we utilized a server equipped with an NVIDIA RTX 3090 GPU, 96 GB RAM, and 2TB RAID storage. Opting for the more promising Neural Machine Translation (NMT) approach, we employed the Fairseq[6] toolkit(Ott et al., 2019) for training our models.

To tackle data sparsity, a common challenge in NMT, we employed preselected and custom legal domain-specific word lists of varying sizes (10k and 20k words). This helped in creating training data with relevant terminology, enhancing the model's ability to translate legal text accurately.

Further details regarding the experimental parameter setup specific to the models are presented in a separate table (Table 2). This information allows for in-depth analysis and potential adjustments in the future.

5. Results and Discussion

Since this work is the first of its kind on the MT System in the Nepali legal domain, we do not have a baseline model to compare our work with. Nevertheless, we have considered the BLEU scores of other Nepali MT systems in the general domain alongside for tentative analysis purposes. We used the BLEU[7] (Papineni et al., 2002) for evaluation and the results are presented in Table 3.

Our research explored multiple MT models for the legal domain in Nepali. We started by exploring Recurrent Neural Networks (RNNs) with LSTM architecture. While the initial RNN model achieved scores of **6.19** and **5.89** for Nepali-English and English-Nepali translation, respectively, the translated documents lacked proper readability and fluency.

Subsequently, we transitioned to using a Transformer-based Neural Machine Translation (NMT) model. During our efforts in building a bidirectional translation model, we achieved scores of **7.98** and **6.63** for Nepali-English and English-Nepali translations, respectively.

[2] https://github.com/anoopkunchukuttan/indic_nlp_library

[3] https://github.com/alvations/sacremoses

[4] A data compression technique.

[5] https://github.com/google/sentencepiece

[6] https://github.com/facebookresearch/fairseq

[7] https://github.com/mozilla/sacreBLEU

Additionally, when we compared our model's performance on general domain data, we attained scores of **13.76** and **9.47** for Nepali-English and English-Nepali translations, respectively. These results surpassed the performance of previous studies (Duwal and Bal, 2019); (Guzmán et al., 2019), demonstrating the effectiveness of our approach in improving translation quality.

The model's better performance in the general domain compared to previous work could be due to sources for the data collection. We gathered data from news sites like OnlineKhabar[8] in both English and Nepali. Initially, we created a legal terminology dictionary to guide our data extraction. However, the extracted articles were primarily intended for a general audience, potentially resulting in a mismatch with the actual legal language. Additionally, documents from the Supreme Court websites, aimed at a general audience, were included. This mix of general and legal domain content may have influenced the model's performance, providing better results in the general domain as well.

Our findings underscore the challenges inherent in translation tasks, particularly between Nepali and English, and highlight the ongoing efforts required to enhance accuracy and fluency in specific domains. The adoption of an NMT-based architecture resulted in an improved score compared to previous works, indicating progress in the right direction, particularly for low-resource languages like Nepali. The modest increase in score from previous experiments signifies a positive advancement, considering the scarcity of available datasets and the inherent challenges in constructing a comprehensive legal domain corpus for Nepali. These challenges include difficulties in achieving proper alignment and the limited availability of publicly accessible data sources for training purposes. While the Transformer model shows promise, further efforts are needed to improve accuracy and domain-specific fluency

6. Conclusion and Future work

We present a Neural Machine Translation (NMT) based approach utilizing a Transformer model for an English-Nepali machine translation system in the legal domain. To the best of our knowledge, this is the first research work carried out in the English-Nepali legal domain which also achieves results on par with the general-domain English-Nepali machine translation systems. The results of this experiment set a baseline for future domain-specific research in low-resource legal MT.

While MT technology is rapidly evolving, many improvements are required in the legal domain. Building on our work, future efforts could focus on:

	Nepali →English		English→ Nepali	
Model	Legal	General	Legal	General
(Guzmán et al., 2019)	-	7.6	-	4.3
(Duwal and Bal, 2019)	-	12.17	-	7.49
NMT Model	**7.98**	**13.67**	**6.63**	**9.47**
RNN Model	6.19	-	5.89	-

Table 3: BLEU score comparison between models by (Guzmán et al., 2019), (Duwal and Bal, 2019) and our work.

- Enhanced Out-of-Vocabulary (OOV) handling: Implementing better methods to address out-of-vocabulary words.
- Improved fluency: Refining techniques to generate smoother and more natural translations.
- Date and time conversion: Integrating a tool for seamless conversion between English Gregorian and Nepali Bikram Sambat calendars.
- Exploring the usefulness and appropriateness of the SMT(Statistical Machine Translation) model especially because the word order for English and Nepali is different (S-V-O, S-O-V) and the previous study by (Acharya and Bal, 2018) has reported some promising results for the English-Nepali pair using this approach.

Furthermore, we aim to explore newer translation architectures to enhance the translation process. By conducting thorough comparisons of results obtained from these architectures on the same dataset, we can gain deeper insights into their effectiveness. Additionally, to facilitate better testing and validation, we plan to deploy the model as software and distribute it to legal professionals for their input and understanding of the output. Leveraging feedback from these professionals, we intend to refine the architecture further to ensure more robust and accurate translations.

7. Limitations

The research work is the first one in the Nepali legal domain, hence has several limitations which are:
Challenges with Legal Terminologies:
The model struggles to accurately translate intricate legal terms.
Complexity of Legal Nuances:
Legal language varies according to contexts and

[8] https://www.onlinekhabar.com

nuances making it difficult to capture the intended meanings in the translation.

Adaptation to Legal Variability:
Legal terminology and conventions vary across jurisdictions, requiring additional model adaptation for accurate translation across diverse legal contexts.

In addition, due to confidentiality constraints and restrictions associated with legal documents from Nepal, we are unable to make our dataset publicly available. We also acknowledge this as a limitation in terms of reproducibility and replicability of this research work.

8. Ethics Statement

In accomplishing this research work we had to deal with proprietary legal data, which we acquired through the signing of the NDA agreement, that restricts the sharing of the data openly. Other than that there are not any issues that affect individuals or groups, hence the research ethics have been properly followed in due course of the research.

9. Acknowledgements

We would like to express our sincere gratitude to **law firms (NDA[9])** of Nepal along with students of **AI, Kathmandu Univeristy[10]** for providing us with data and helping us with cleaning and creating parallel datasets. We would also like to thank the reviewers for their feedback and comments.

10. References

Praveen Acharya and Bal Krishna Bal. 2018. A Comparative Study of SMT and NMT: Case Study of English-Nepali Language Pair. In *Proc. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 90–93.

Petra Bago, Sheila Castilho, Edoardo Celeste, Jane Dunne, Federico Gaspari, Niels Gíslason, Andre Kåsen, Filip Klubička, Gauti Kristmannsson, Helen McHugh, et al. 2022. Sharing high-quality language resources in the legal domain to develop neural machine translation for under-resourced european languages. *Revista de Llengua i Dret (Journal of Language and Law)*, 78:9–34.

Bal Krishna Bal. 2004. Structure of nepali grammar.

Vicent Briva-Iglesias, Joao Lucas Cavalheiro Camargo, and Gokhan Dogru. 2024. Large language models" ad referendum": How good are they at machine translation in the legal domain? *arXiv preprint arXiv:2402.07681*.

Binaya Kumar Chaudhary, Bal Krishna Bal, and Rasil Baidar. 2020. Efforts towards developing a tamang nepali machine translation system. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 281–286.

Arne Defauw, Sara Szoc, Tom Vanallemeersch, Anna Bardadym, Joris Brabers, Frederic Everaert, Kim Scholte, Koen Van Winckel, and Joachim Van den Bogaert. 2019. Developing a neural machine translation system for irish. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 32–38.

Sharad Duwal and Bal Krishna Bal. 2019. Efforts in the development of an augmented english-nepali parallel corpus. Technical report, Technical report, Kathmandu University.

Xavier Garcia, Aditya Siddhant, Orhan Firat, and Ankur P Parikh. 2020. Harnessing multilinguality in unsupervised machine translation for rare languages. *arXiv preprint arXiv:2009.11201*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

[9] NDA: Non-Disclosure Agreement signed with various law firms regarding data used.

[10] B.Tech in AI Program, Kathmandu University

Rubén Martínez-Domínguez, Matīss Rikters, Artūrs Vasiļevskis, Mārcis Pinnis, and Paula Reichenberg. 2020. Customized neural machine translation systems for the swiss legal domain. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 217–223.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models.

Kriti Nemkul and Subarna Shakya. 2021. Low resource english to nepali sentence translation using rnn—long short-term memory with attention. In *Proceedings of International Conference on Sustainable Expert Systems: ICSES 2020*, pages 649–657. Springer.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.