Abstract

Dzongkha, national language of Bhutan, is under represented in the field of NLP due to the language being written in complex script without explicit word boundaries, and lack of annoted data. This paper describes an attempt to develop transformer-based approach for bidirectional Dzongkha-English Neural Machine Translation (NMT) with linguistic adaptation to address the unique syntactic and orthographic characteristics of the language. Leveraging a curated parallel corpus, subword tokenization, and pretrained embeddings, the proposed model demonstrates superior performance over recurrent architectures in both translation quality and processing efficiency. Designed with real-time communication in mind, the system enables accurate and immediate translation between Dzongkha and English, which not only contributes to the advancement of NMT for low-resource languages but also supports broader efforts in preserving and promoting Dzongkha in the digital era.

Keywords: Neural Machine Translation, Low-Resource Languages, Transformer, Dzongkha, Linguistic Adaptation

1. Introduction

The evolution of Machine Translation (MT) has seen significant progress over the past decades, transitioning from manually engineered Rule-Based Machine Translation (RBMT) systems to probabilistic models in Statistical Machine Translation (SMT), and more recently to end-to-end Neural Machine Translation (NMT) frameworks. While NMT has achieved remarkable success for high-resource language pairs, it remains underdeveloped for low-resource language, like Dzongkha.

Dzongkha, the national language of Bhutan, presents unique linguistic and computational challenges due to its agglutinative morphology, lack of explicit word boundaries, and scarcity of parallel corpora (Jamtsho, 2019). The language's Tibetan script and syntactic structure—characterized by Subject-Object-Verb (SOV) word order—further complicate alignment with languages like English that follow Subject-Verb-Object (SVO) patterns (Dhungyel & Grundspenkis, 2017). These limitations make Dzongkha an ideal testbed for exploring novel, linguistically adapted NMT approaches designed for low-resource environments.

While transformer-based architectures (Vaswani et al., 2017) have significantly improved translation quality in high-resource settings, their performance in low-resource languages remains limited without sufficient linguistic adaptation. To overcome these limitations, recent studies advocate for methods such as: parameter-efficient fine-tuning (Pfeiffer et al., 2021),

Sub word modelling (Neubig, 2017), and self-supervised pretraining strategies (Liu et al., 2021) to enhance translation in underrepresented languages.our work research builds on these advances while introducing:

- Developed a real-time, transformer-based Dzongkha-English bidirectional translation system, optimized for Dzongkha’s SOV structure and linguistic properties.
- Incorporated sub word-level modeling, lexical adaptation, and rule-based preprocessing to handle data scarcity and morphosyntactic divergence.
- addressed limitations of prior models and contributed scalable, linguistically informed tools for modernizing and preserving underrepresented languages.