Machine Translator using ML

Vansh Raj Soam*

Abstract

This paper presents a neural machine translation model for the purpose of translating English to Hindi using a sequence-to-sequence approach with Long Short-Term Memory (LSTM) networks. Traditional methods of translation are unable to work for languages like Hindi and English because of structural differences between them, whereas neural networks offer better adaptability toward complex patterns. Our model is an encoder-decoder model using LSTM units, where the encoder takes the English sentences and converts them into context vectors, and the decoder uses those vectors to generate Hindi translations. We have trained the model on a dataset of English-Hindi pairs using tokenization, padding, and vocabulary indexing to manage text. Preliminary results show that the model can translate basic sentence structures but faces difficulties in translating complex syntax. These findings support neural machine translation's potential for English-Hindi translation, with opportunities for further optimization through dataset expansion and model tuning.

Introduction

Background and Motivation:

MT is the way to communicate between languages, but translation between such languages as English and Hindi remains a challenging task due to structural and grammatical variations. The structure of sentences is different between English (SVO) and Hindi (SOV), with a difference in morphological and lexical use, hence less effective rule-based and statistical approaches to these languages. With advancements in deep learning, NMT has been promising with seq2seq models, especially those using LSTMs. Such networks can capture complex patterns that are present in language just by learning directly from the data, which makes these models a good fit for tasks like English-Hindi translation. This research involves an LSTM-based seq2seq architecture specifically designed for the unique challenges of English to Hindi translation so that fluency and accuracy in the translation can be enhanced.

Objective:

This paper proposes a neural machine translation model in the sequence-to-sequence architecture using LSTM networks for translation between the English sentences and Hindi. The translations of these languages are wrong when mapped between each other because of the variations in the structures and grammatical usage of two independent languages. The challenges here are going to be addressed with the development of an encoder-decoder model that can represent the encoding of English sentences as context vectors and the decoding coherently as Hindi translations. Specific objectives include data preprocessing-for example, tokenization and padding, training the model on a dataset of English-Hindi sentences, and evaluation of model translation accuracy in relation to different sentences. This work finally aims to improve the quality of machine translation and further investigate improvements using increased datasets and better model tuning.

Literature Review

Much has been achieved by MT since the rule-based systems that involve human-defined linguistic rules for translation of words and grammar structures, though earlier methods were effective for simple pairs of language but lacked flexibility and adaptability towards differing language structures, especially like Hindi, which differs in syntax as well as morphology from that of English.

Statistical Machine Translation replaces the rule-based models in the 1990s. It employed high-parallel corpora where the probability of words or phrases could be estimated such that the fluency could increase with adaptability; however, SMT has huge demands for data such that accuracy in translation decreases if there is little consideration in terms of representing a meaningful translation in complex sentences.

The advent of Neural Machine Translation (NMT) introduced new possibilities through sequence-to-sequence (seq2seq) architectures. Sutskever et al. (2014) first introduced RNNs in NMT through an encoder-decoder architecture to map sentences from one language to another. The success of this approach demonstrated massive gains in translation accuracy compared to the traditional models based on predefined rules that can be learned directly from the data.

Bahdanau et al. (2015) extended NMT with an attention mechanism, enabling models to focus on the most relevant parts of input sentences and achieve better accuracy in long and complex translations. Although all of these advances hold true, there is still room for improvement due to structural dissimilarities between the SVO-ordered English language and SOV-ordered Hindi language. Recent experiments demonstrate the ability of LSTM networks in handling such intricacies and complexities as dependencies can now be captured across longer sequences. This work builds upon that and uses a seq2seq model-based LSTM model that enhances the quality of the translation from the English-Hindi language.

Related work:

Machine translation has been a journey from rule-based systems to statistical and recent neural approaches. Rule-based methods, based on prior linguistic rules, were pretty inflexible in all senses, especially about the differences in languages spoken, such as from English to Hindi. SMT methods came as an improvement over this rule-based system by implementing probabilistic models to help align words and phrases according to larger parallel corpora, where they suffered limitations in syntax complexity and long-distance dependency.

The emergence of NMT, especially sequence-to-sequence (seq2seq) architectures, presented a significant improvement. Sutskever et al. (2014) pioneered this approach by employing the Recurrent Neural Networks (RNNs) for encoding input sentences into context vectors that are then decoded as the target language. LSTMs and GRUs have proven to be effective in this context due to their ability to capture long dependencies. Seq2seq models, especially those with attention mechanisms, as proposed in 2015 by Bahdanau et al., work very well for general purposes of translation-especially in languages with diverse structures.

Despite these developments, translation between English and Hindi is still difficult because of structural differences and idiomatic expressions. This research is a continuation of these foundations with an LSTM-based seq2seq model specifically tuned to address the linguistic differences between the two languages, thus improving the coherence and accuracy of translation.

Methadology

This study's goal is to develop a translation model that translates English sentences into Hindi using a machine learning approach. The key parts of this methodology are data preparation, building the model, training it, and testing its ability to translate.

Data Preparation:

Our dataset would comprise two languages: the English one and the corresponding Hindi translated sentence. In this way, a translation of each one of the English sentences with which the model is trained into understanding that. Preprocess the data of our algorithm by making sentences tokenized based on word and padded each one to match the length.

Model structure:

It has a two-part structure known as an "encoder-decoder" that aids in its ability to read and translate sentences.

Encoder: A. The input is reduced to an English sentence. That gets read into this "context vector," like the distilled core of its meaning.

Decoder : This summary form the decoder uses in order to generate Hindi word for word translation. Both use a particular type of neural network layer: Long Short-Term Memory, or LSTM. It is well-known that this type functions well with word order when performing translation.

Model Training:

It then structures this to train thousands of English-Hindi bilingual pairs with their sentences. Every training fine-tunes a model that is always better at prediction. It measures how close its predictions are to the actual translations with a measure known as "cross-entropy loss". In addition to this, for thousands of rounds-called epochs-it is further trained to be more precise.

Translation Process:

We use this encoder and decoder part of the model for translating some new sentence. We begin by passing the sentence in English through the encoder, while such information is passed on to the decoder to start coming up with the complete sentence in Hindi one word after the other.

Evaluation:

We also judge the fluency of the output translation by comparison to a reference translation and examine the BLEU score of the model output regarding human translations.