ENGLISH TO URDU NEURAL MACHINE TRANSLATION

Muhammad Aqeel

Saint Louis University

ABSTRACT

In the modern world, machine translation is a necessary piece of technology to break down communication boundaries. It has advanced significantly in recent years and is now widely used in both the commercial and nonprofit sectors. Due to a lack of resources, the technology for the English and Urdu language pair is still in its infancy. Neural networks are currently dominating the field of translation techniques given their ability to provide a single, large neural network with an attention mechanism, sequence-to-sequence translation, and long-short term model. This paper discusses two approaches of machine translation from the source language to target language using neural networks. The techniques discussed in the paper include character-level sequence-to-sequence modeling and word-level sequence-to-sequence neural machine translation with attention. BLEU score has been used as an evaluation criterion for both techniques. A BLEU score of 0.5 has been achieved on the character-level seq2seq model and state of the art result of 0.69 has been achieved on NMT with attention model.

Index Terms— machine translation, machine learning, Seq2seq model, LSTM, Encoders-Decoders, Attention Layers, BLEU Score

1. INTRODUCTION

The Indo-Aryan language family includes the free-order language Urdu. Arabic, Persian, Turkish, and Hindi all had an impact on the development of Urdu. Even the name "Urdu" is Turkish in origin. Urdu is the official language of Jammu and Kashmir and the national language of Pakistan. Other South Asian nations like Bangladesh and Afghanistan are home to many speakers of it. Nearly 70 million people are native speakers of Urdu. English has emerged as an international language and has become standard of communication. In NLP applications, Urdu is an under resourced language and there is a strong need to develop translation systems from English to Urdu which can help such a massive number of native speakers in many ways.

Artificial intelligence is used in machine translation to translate text from one language to another without the need for a human translator. The whole meaning of the text in the original language is communicated in the target language using modern machine translation, which goes beyond mere word-to-word translation. It examines every aspect of the text and finds the relationships between the words. It has been difficult to use machine translation (MT) for languages with limited resources When translating between a morphologically rich and a morphologically low language, the difficulty is more multidimensional. English is a language with a simple morphology, but Urdu has a rich inflectional and derivational morphology. In this study, we will provide one such pair, English to Urdu translation and will analyze various methods for machine translation from English to Urdu.

One method discussed in this paper is the character level, encoder-decoder model. The method uses two recurrent neural networks: an encoder, which encodes the input sequence, and a decoder, which converts the encoded input sequence into the target sequence. The second technique discussed in the paper is the encoder-decoder model with attention. At each time step, the decoder's output is combined with the encoder's output to predict the next word.

2. RELATED WORK

Possibly one of the earliest translations from English to Urdu was made by [1] , who did it by applying transformation rules to the parse tree of an English sentence. In order to address the issue of long-distance word reordering between the two languages, [2] provide a simple English-to-Urdu RBMT system that makes use of phrase-based models. They used the Quran and Bible corpora, Treebank (Marcus et al., 1993), Emille (Baker et al., 2002), and Treebank corpora. They report an improvement in BLEU scores thanks to the suggested reordering strategy.

The English to Hindi translation algorithm for the English and Urdu language pair was tested due to the few similarities between Urdu and Hindi. The Hindi language served as an interlingua with a Hindi-Urdu mapping table to produce final output in the system, which was based on a pseudo interlingua rule-based method. Due to verb forms, gender mismatch, and phonetic discrepancies between Hindi and Urdu, a low BLEU score was detected. [3]

A basic Phrase-based system for English to Hindi Translation was presented by [4] using a relatively modest amount of data. They employed human evaluation metrics as their evaluation criteria. Compared to the currently available automatic evaluation metrics, these evaluations were more expensive.

[5] paper addresses the Hierarchical Phrase-based (HPB) models which are used in development of different Statistical Machine Translation (SMT) Systems for many modern languages. This paper considers English as Source and Urdu as target language for experiments. For this study, Hierarchical phrase-based Baseline SMT system is used for English to Urdu translation. At the end automatic evaluation of system is performed by using BLEU and NIST as evaluation metrics. The developed system was able to get 0.13 Bleu score.

[6] paper present results for four major domains including Biomedical, Religious, Technological and General using Statistical and Neural Machine Translation. They performed series of experiments in attempts to optimize the performance of each system and also to study the impact of data sources on the systems. Finally, they established a comparison of the data sources and the effect of language model size on statistical machine translation performance. The best BLEU score they got from religious datasets (Quran+Bible+QBJ) is 0.19

3. METHODOLOGY

In this project, UMC005 (Jawaid and Zeman, 2011) dataset and the dataset from tatoeba.org has been used. UMC005 dataset provides 6414 sentence pairs from Bible and 7957 sentence pairs from the Quran corpus. The dataset from tatoeba.org provides 1143 parallel sentences in English and Urdu.

Two different techniques of neural machine translation have been utilized to achieve translation in Urdu language. One of the techniques is Character-level recurrent sequence-to-sequence model. The second method used in this paper is Word-level Neural machine translation with attention. BLEU score has been used as an evaluation criterion for both techniques.

3.1. Preprocessing

During the preparation of the dataset for the translation task, a few preprocessing steps have been performed such as merging all the datasets and performing clean-up tasks which will help the model learn important features. All the datasets were merged and for consistency, all sentences of source language were converted into lowercase letters. Quotes from the source as well as the target language were removed as part of preprocessing step.

The target sentences were also padded with start and end characters to enable the decoder to learn the start and end of sentences. Sets for English as well as Urdu language have been maintained to store unique characters from both languages.

The Max encoder length was calculated as the max length of the input sentence from the source language and the max decoder length was calculated as the max length of the target sentence from the target language.

Next, we created three different NumPy arrays to store input and output, and intermediate data. These three arrays are for encoder input data, decoder input data, and decoder output data.

Using the Scikit-learn train-test split function, we split the dataset into training and validation data by a 90/10 ratio. (90% training data and 10% validation data)

4. EXPERIMENTAL SETUP OF MODELS

As discussed earlier, two different machine translation techniques have been implemented and we will go over details on both techniques in this section. Character-level sequence-to-sequence modeling was implemented as a baseline to achieve translation and to compare it with a more advanced technique such as Word-level neural machine translation.

4.1. Char-Level Seq2Seq Model

Char-level seq2seq model was implemented in following steps. [7]

I. We start with input sequences from a domain (English sentences) and corresponding target sequences from another domain (Urdu sentences).

ii. An encoder LSTM turns input sequences to 2 state vectors (we keep the last LSTM state and discard the outputs).

iii. A decoder LSTM is trained to turn the target sequences into the same sequence but offset by one timestep in the future, a training process called ”teacher forcing” in this context. It uses as initial state the state vectors from the encoder. Effectively, the decoder learns to generate targets[t+1...] given targets[...t], conditioned on the input sequence.

iv. In inference mode, when we want to decode unknown input sequences, we encode the input sequence into state vectors. Start with a target sequence of size 1 (just the start-of-sequence character). Feed the state vectors and 1-char target sequence to the decoder to produce predictions for the next character. Sample the next character using these predictions (we simply use argmax). Append the sampled character to the target sequence. Repeat until we generate the end-of-sequence character, or we hit the character limit.

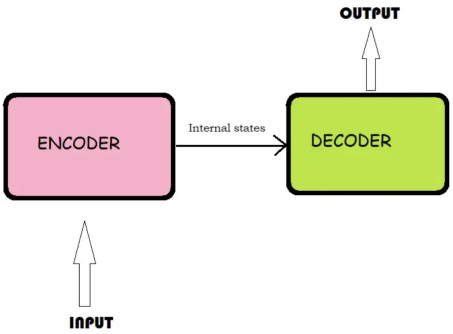

Fig. 1. Char-Level Encoder-Decoder Architecture

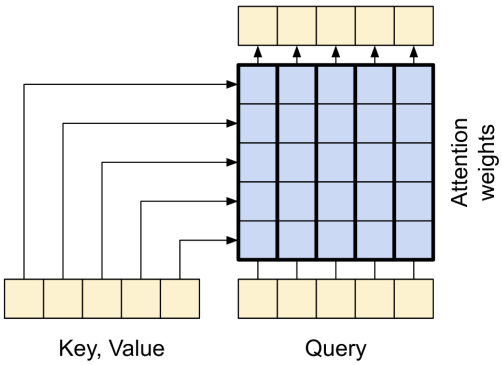

Fig. 2. NMT Attention Layer Architecture[8]

4.2. Word-Level Neural Machine Translation with Attention

An attention mechanism has lately been used to improve neural machine translation (NMT) by selectively focusing on parts of the source sentence during translation. Below we will examine a simple and effective class of attention mechanism that always attends to all source words. [8]

The attention model implementation follows the following steps

i. The goal of the encoder is to process the context sequence into a sequence of vectors that are useful for the decoder as it attempts to predict the next output for each timestep.

ii. The attention layer lets the decoder access the information extracted by the encoder. It computes a vector from the entire context sequence and adds that to the decoder's output.

iii. The decoder's job is to generate predictions for the next token at each location in the target sequence.

When training, the model predicts the next word at each location. So, it's important that the information only flows in one direction through the model. The decoder uses a unidirectional (not bidirectional) RNN to process the target sequence. When running inference with this model it produces one word at a time, and those are fed back into the model.

5. RESULTS

In our experiment for both models, we used a setup of encoders and decoders. In the character level model, we used a simple encoder and decoder setup while in the word level model we used an encoder and decoder with an attention layer that will pay attention to all source words at every step.

The table below show the output of results from both techniques. BLEU score has been used to evaluate the performance of both techniques. BLEU (Bilingual Evaluation Understudy) is a metric for automatically evaluating machine-translated text. The BLEU score is a number between zero and one that measures the similarity of the machine-translated text to a set of high-quality reference translations.

Models	Source Language	Target Language	BLEU Score
Char-Level Seq2Seq	English	Urdu	0.50
NMT with Attention	English	Urdu	0.69

Fig. 3. BLEU Score for Experimental Models

For the attention model, we can examine the training and validation accuracy of the model during training against epochs. The accuracy we are considering here is actually cross-entropy/token.

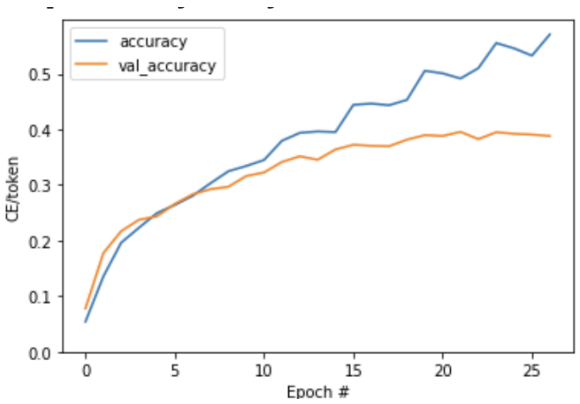

Fig. 4. Accuracy Vs Epoch Attention Model

A few instances of translated sentences from both models have been provided below in tables.

6. FUTURE WORK AND CONCLUSION

In this paper, Urdu, an under-resourced language has been chosen to carry out machine translation analysis using various different techniques. Analysis with two major techniques

and whosoever liveth and believeth in me shall never die . believest thou this ?	جاکر ان کے لئے اس کے ساتھ اس کے ساتھ ایسا ہؤا ہے اور اس کے ساتھ اس کے ساتھ ایسا ہؤا اور اس کے ساتھ
and whosoever liveth and believeth in me shall never die . believest thou this ?	جیکا کیا تم اس کے ساتھ ایسا ہؤا ہے اور اس کے ساتھ اس کے ساتھ ایسا ہؤا اور اس کے ساتھ اس کے ساتھ ایک

Fig. 5. Char-Level Seq2Seq Decoded Translations

and whosoever liveth and believeth in me shall never die . believest thou this ?	اور جو کوئی مجھ پر ایمان لائے اور مجھ پر ایمان لاتا ہے وہ ابد تک کبھی کبھی نہ مریگا . کیا تو یہ ہے ؟ .
and whosoever liveth and believeth in me shall never die . believest thou this ?	اور جو کوئی مجھ پر ایمان لائے اور مجھ پر ایمان لاتا ہے وہ ابد تک کبھی کبھی نہ مریگا . کیا تو یہ ہے ؟ .

Fig. 6. NMT with Attention Decoded Translations

has been carried out which includes, character level sequence to sequence model, and word level with attention. Due to the insufficiency of the data, the results are not very much impressive for character level but quite improved results have been achieved on the word level model with attention. The technique which stood out among all others is word level sequence-to-sequence model with attention as it was able to achieve the highest BLEU score of 0.69

As future work on this paper could include training the model on a much larger dataset as the best-performing models are usually trained on larger datasets. We can use dropout and other forms of regularization techniques to mitigate overfitting. Perform Hyper-parameter tunings such as changing learning rate, batch size, and dropout rate. We can also try using multi-layered LSTMs. All these techniques can be applied as an improvement to existing models. More advanced techniques like sequence-to-sequence machine translation with transformers can also be applied to see if are able to get any improvement over the BLEU score.

7. REFERENCES

[1] Tafseer A. and S. Alvi, "English to urdu translation system.," in *manuscript*. University of Karachi, 2019.

[2] Naila Ata and Bushra Jawaid A. K., "Rule based english to urdu machine translation.," in *Proceedings of Conference on Language and Technology (CLT'07).*, 2007.

[3] R. Mahesh and K. Sinha, "Developing english-urdu machine translation via hindi," in *Third Workshop on Computational Approaches to Arabic-Script-based Languages*, 2009.

[4] Nakul Sharma, Parteek Bhatia, and Varinderpal Singh, "English to hindi statistical translation," in *International Journal in Computer Networks and Security (IJCNS)*, 2011.

[5] Nadeem Jadoon Khan, Muhammad Waqas Anwar, Usama Ijaz Bajwa, and Nadir Durrani, "English to urdu hierarchical phrase-based statistical machine translation," 2013.

[6] Sadaf Abdul-Rauf, Syeda Abida, Noor e Hira, Syeda Zahra, Dania Parvez, Javeria Bashir, and Qurat ul-ain Majid, "On the exploration of english to urdu machine translation," in *SLTU*, 2020.

[7] François Chollet, "Character-level recurrent sequence-to-sequence model," 2020.

[8] Pham H. Manning C. D. Luong, M., "Effective approaches to attention-based neural machine translation," 2015.