Neural Text Generation for Urdu Language

Rida Zainab
Department of Electrical Engineering
Stevens Institute of Technology
Hoboken, NJ 07030
rzainab@stevens.edu

Abstract

Text Generation using artificial neural networks is a classical deep learning problem. Models which can learn the dynamic temporal behaviour of a sequential data are popular choice for text generation. There has been considerable progress in this area in the past, currently the focus being adversarial text generation using generative adversarial networks. However, most of the research in this area has been on English, and to a comparable extent, on Chinese language text generation. This presents a huge gap in the field of research in natural language processing. In this project, we present results of using current deep learning text generation techniques on Urdu language.

1 Introduction

The availability of astronomical amounts of data and commensurate progress in the computational capacity and capability of digital procesing units has made it possible to imagine talking, seeing, thinking and feeling artificial machines in not so distant future. While there has been considerable success in the area of image processing and its subsidiary fields, research in natural language processing (NLP) has not yet achieved what can be called a gold standard when it comes to text recognition and prediction. One primary reason for the inherent difficulty when dealing with natural language is the subjectivity of the data. Text and speech are sensitive to the location, time, era, dialect and other such social factors, which is not a significant problem when it comes to images. The picture of a cat is a picture of a cat everywhere in the world. But the script of Punjabi language changes if the source of the data crosses the border between India and Pakistan. The speech, however, remains the same. With text, what might be a negative comment to one annotator might be a clever witty statement to the other. This problem is especially pronounced when the text data is from a resource starved language such as Urdu.

Lack of available datasets, not to mention the poor quality of the data, linguistic and typographical errors in the text documents due to the uniqueness of the script and no information of document hierarchy intensify the problem of appying natural language processing methods to such languages.

This paper discusses the methods and results of neural text generation using an Urdu language dataset recently released. While there are many techniques for generating artificial text such as Markov Models, Recurrent Neural Networks (RNNs), Long Short Term Memory (LSTMs), Gated Recurrent Units (GRUs), Transformer, and more recently, Variational Auto Encoders (VAEs) and Generative Adversarial Networks, we were not able to successfully employ all of them due t lack of data. The best results were found using LSTMs and GRUs which will be discussed later in the paper.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

2 Data

There are a few resources for Urdu language dataset such as OpenSubtitles 2016 and 2018 corpora, XLNI, LDC2010T21, LDC2010T23 LDC corpora and Makhzan. The dataset we used was obtained from Makhzan which has been recently released. This was a particularly good dataset as it was annotated where English text occured in the document or if at any point the text had a different literature element such as poetry.

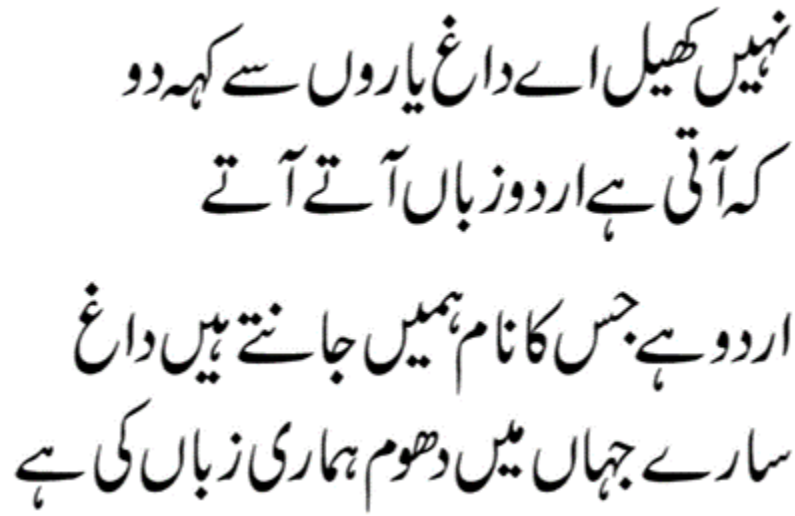


Figure 1: A verse in Urdu language

```
<document>
        <meta>
                <title>میرا جی کے تراجم</title>
                <author>
                        <name>ناصر عباس نیر</name>
                        <gender>Male</gender>
                </author>
                <publication>
                        <name>Bunyad, Volume 4</name>
                        <year>2013</year>
                        <city>Lahore</city>
                        <link>https://gcll.lums.edu.pk/sites/default/files/12_nasir_abbas_nayyar_bunyad_2013.pdf</link>
                        <copyright-holder>Gurmani Centre for Languages and Literature, Lahore University of Management Sciences
                </publication>
                <num-words>6627</num-words>
                <contains-non-urdu-languages>Yes</contains-non-urdu-languages>
        </meta>
        <body>
                <section>
                        <p>ایسی ہی کہانی سناتی ہے۔ اس کہانی کا ایک اہم واقعہ میرا جی کے تراجم ہیں جو اس میں ایک نیا موڑ لاتے ہیں۔</p>
                        <p>نظوم کے چار ایڈیشن چھپ چکے تھے۔ تراجم کی اس روایت کو رسالہ دلگداز اور مخزن نے خاص طور پر آگے بڑھایا۔</p>
                        <p>پر انگریزی ادب کے اثر کی نوعیت کو سمجھنے کے لیے ان دونوں صاحبان کے خیالات پر ایک نظر ڈالنا ضروری ہے۔</p>
                        <p>ین انگریزی اور اردو میں ایک ایسی مغائرت موجود ہے جسے محض ترجمہ یعنی حقیقی لفظی ترجمہ نہیں پاٹ سکتا۔</p>
                        <p>حامل ہوتے ہیں جب کہ مشرقی مصنف 'شخصی، مخصوص، مجسم اور ڈرامائی' انداز رکھتے ہیں۔ لہٰذا لفظی ترجمہ نہیں</p>
                        <p>ادب کینن نہیں تھا۔ انگریزی ادب کا وہ حصہ جو نئے اردو ادب کے زمروں میں موزوں بیٹھتا تھا، وہی کینن تھا۔</p>
                        <p>ی سے زیادہ اور اکثر متناقض شناختیں رکھتے تھے اور اس سے پیدا ہونے والی کش مکش سے شاعری کشید کرتے تھے۔</p>
                        <p>تی، دسودر گیت، عمر خیام نیز کوریائی، چینی، جاپانی گیتوں کے ترجمے کیے اور ان پر تفصیلی نوٹ تحریر کیے۔</p>
                        <p>جھ نہیں تھا، مگر وہ اپنے عہد کی بری طرح سیاست زدہ فضا میں اپنی معنویت باور کرانے کی جرأت سے لیس تھا۔</p>
                        <p>یہاں آگے بڑھنے سے پہلے ہم دیسی نقطۂ نظر سے متعلق دو ایک باتیں کہنے کی ضرورت ہے۔ ہم دیسی یا <annota
                        <p>زمانی شعریت ، ادبیت پر اصرار کا رویہ تھا۔ دیکھیے میرا جی کس قسم کی نظمیں بذریعہ ترجمہ سامنے لاتے ہیں:</p>
                        <blockquote>
                                <verse>
                                        مست عشرت کا کوئی مول نہیں
                                        میرے قریں
```

Figure 2: Text dataset file view

3 Method

3.1 Pre-Processing

The dataset was split into a fixed sequence of words and the followong character as the target element. A total of around 7600 such samples were generated and there were 59 unique characters. The data was vectorized for further application.

```
In [46]: sentence[1], next_character[1]
Out[46]: ('نگلہ دیش کی عدالتِ عالیہ نے طلاق کے ایک ', 'م')

In [47]: sentence[2], next_character[2]
Out[47]: ('گلہ دیش کی عدالتِ عالیہ نے طلاق کے ایک م', 'ق')

In [49]: sentence[3], next_character[3]
Out[49]: ('لہ دیش کی عدالتِ عالیہ نے طلاق کے ایک مق', 'د')
```

Figure 3: Pre Processing

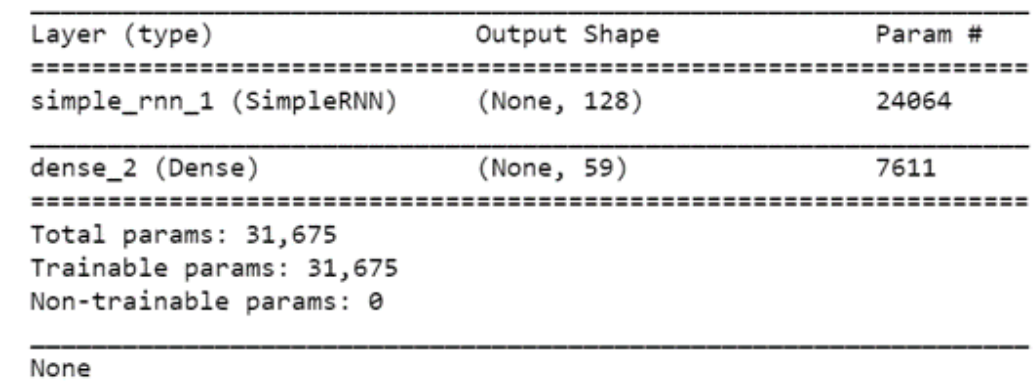
3.2 Model

Single RNN, LSTM and GRU models were implemented. Best results were found using LSTM model, GRU being closely tied with LSTM and RNN did not give good results.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_2 (LSTM)                (None, 128)               96256
_________________________________________________________________
dense_4 (Dense)              (None, 59)                7611
=================================================================
Total params: 103,867
Trainable params: 103,867
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 4: LSTM Model

Results While there are a few methods for evaluating the quality of the generated text, we were doubtful of using them as they are more suited towards resource rich languages. For instance, the bleu score which is a popular metric for evaluating generated text works by translating the machine text and comparing it with human translation and then evaluating it. This is a problem with languages which are do not have plenty of resources as the translation algorithms do not work well. Hence, evaluation was done manually using human judgement.

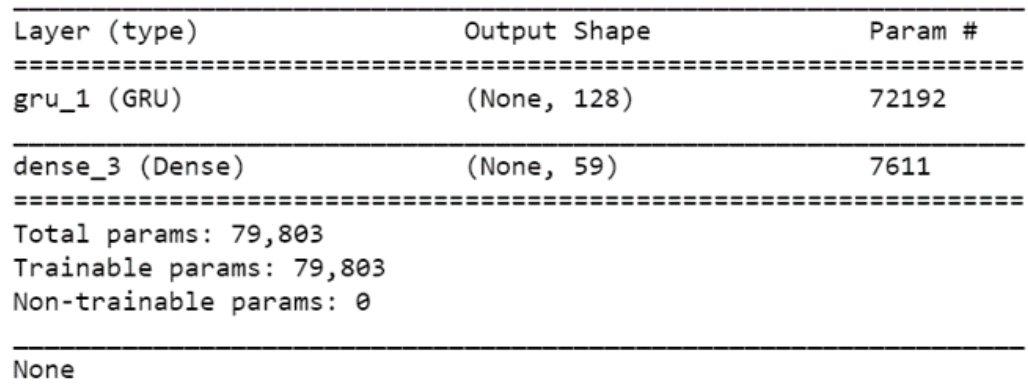


Layer (type)	Output Shape	Param #
simple_rnn_1 (SimpleRNN)	(None, 128)	24064
dense_2 (Dense)	(None, 59)	7611

Total params: 31,675
Trainable params: 31,675
Non-trainable params: 0

None

Figure 5: RNN Model



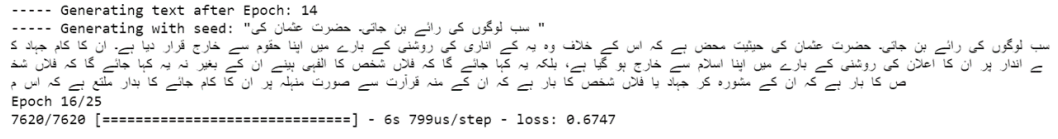
Layer (type)	Output Shape	Param #
gru_1 (GRU)	(None, 128)	72192
dense_3 (Dense)	(None, 59)	7611

Total params: 79,803
Trainable params: 79,803
Non-trainable params: 0

None

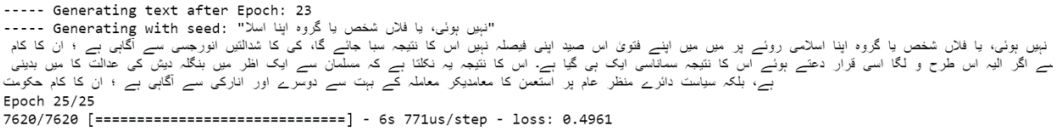
Figure 6: GRU Model

After training for more than 20 epochs, LSTMs and GRUs exhibited much better performance than RNN. The lack of performance of RNN could be due to the problem of vanishing gradient.



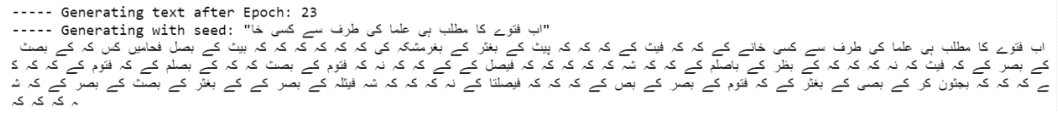
----- Generating text after Epoch: 14
----- Generating with seed: "سب لوگوں کی رائے بن جاتی۔ حضرت عثمان کی "
سب لوگوں کی رائے بن جاتی۔ حضرت عثمان کی حیثیت محض ہے کہ اس کے خلاف وہ یہ کہے انارکی کی روشنی کے بارے میں اپنا حقوم سے خارج قرار دیا ہے۔ ان کا کام جہاد ک
ے اندار پر ان کا اعلان کی روشنی کے بارے میں اپنا اسلام سے خارج ہو گیا ہے، بلکہ یہ کہا جائے گا کہ فلاں شخص کا الفبی بیٹے ان کے بغیر نہ یہ کہا جائے گا کہ فلاں شخ
ص کا بار ہے کہ ان کے مشورہ کر جہاد یا فلاں شخص کا بار ہے کہ ان کے منہ قرأت سے صورت منبلہ پر ان کا کام جائے کا بدار ملتع ہے کہ اس م
Epoch 16/25
7620/7620 [==============================] - 6s 799us/step - loss: 0.6747

Figure 7: Text generated using LSTM



----- Generating text after Epoch: 23
----- Generating with seed: "نہیں ہوئی، یا فلاں شخص یا گروہ اپنا اسلا"
نہیں ہوئی، یا فلاں شخص یا گروہ اپنا اسلامی روئے پر میں میں اپنے فتوی اس صید اپنی فیصلہ نہیں اس کا نتیجہ سبا جائے گا، کی کا شدالتین انورجسی سے آگاہی ہے ؛ ان کا کام
ے اگر الیہ اس طرح و لگا اسی قرار دعتے ہوئے اس کا نتیجہ سماناسی ایک ہی گیا ہے۔ اس کا نتیجہ یہ نکلتا ہے کہ مسلمان سے ایک اظر میں بنگلہ دیش کی عدالت کا میں بدینی
ہے، بلکہ سیاست دائرے منظر عام پر استعمن کا معامدیکر معاملہ کے بہت سے دوسرے اور انارکی سے آگاہی ہے ؛ ان کا کام حکومت
Epoch 25/25
7620/7620 [==============================] - 6s 771us/step - loss: 0.4961

Figure 8: Text generated using GRU



----- Generating text after Epoch: 23
----- Generating with seed: "اب فتوے کا مطلب ہی علما کی طرف سے کسی خا"
اب فتوے کا مطلب ہی علما کی طرف سے کسی خانے کے کہ کہ کہ پیٹ کے بغیر کے بغرمشکہ کی کہ کہ کہ کہ بیٹ کے بصل فحامیں کس کہ کے بصت
کے بصر کے کہ فیٹ کہ نہ کہ کہ کہ کے بظر کے باصلم کے کہ کہ شہ کہ کہ کہ کہ فیصل کے کے کہ کہ نہ کہ فتوم کے بصت کہ کہ کے بصلم کے کہ فتوم کے کہ کہ ک
ے کہ کہ کہ بجنون کر کے بصی کے بغتر کے کہ فتوم کے بصر کے بص کے کہ کہ کہ فیصلتا کے نہ کہ کہ کہ شہ فیتلہ کے بصر کے کے بغتر کے بصت کے بصر کے کہ شہ
ہ کہ کہ کہ

Figure 9: Text generated using RNN

4 Conclusion

Text Generation for Urdu language is, as of now, an unexplored area in the field of research in natural language processing. Only one python library exists for Urdu language text pre processing, and one other library for Urdu language word embedding. There exists only a handful of labelled Urdu language datasets for supervised learning problems. We hope this work will benefit research groups working to apply deep learning techniques in Urdu language.