FINAL DEFENSE - FINAL YEAR PROJECT

Automated Dubbing System

Submitted by:
Amna Abid
Zainab Binte Iftikhar

Advisor: Ma'm Rafia Mumtaz

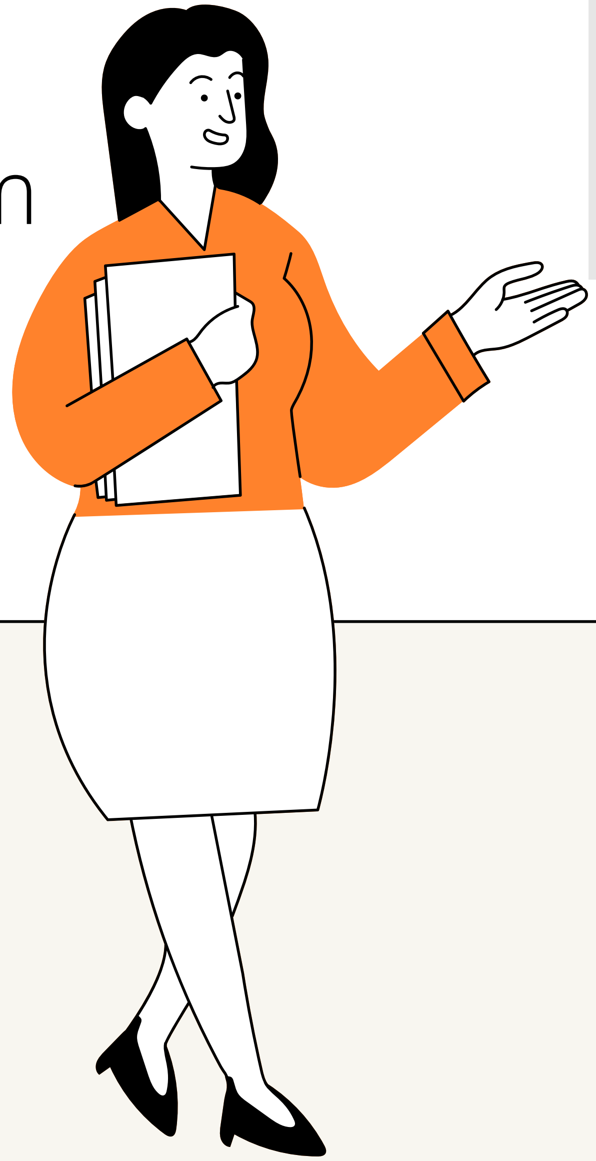
Co-Advisor: Sir Ali Tahir

Content

- Introduction
- Problem Statement
- Methodology
- Work Division
- Results & Conclusions
- References

WHAT IS A DUBBING SYSTEM?

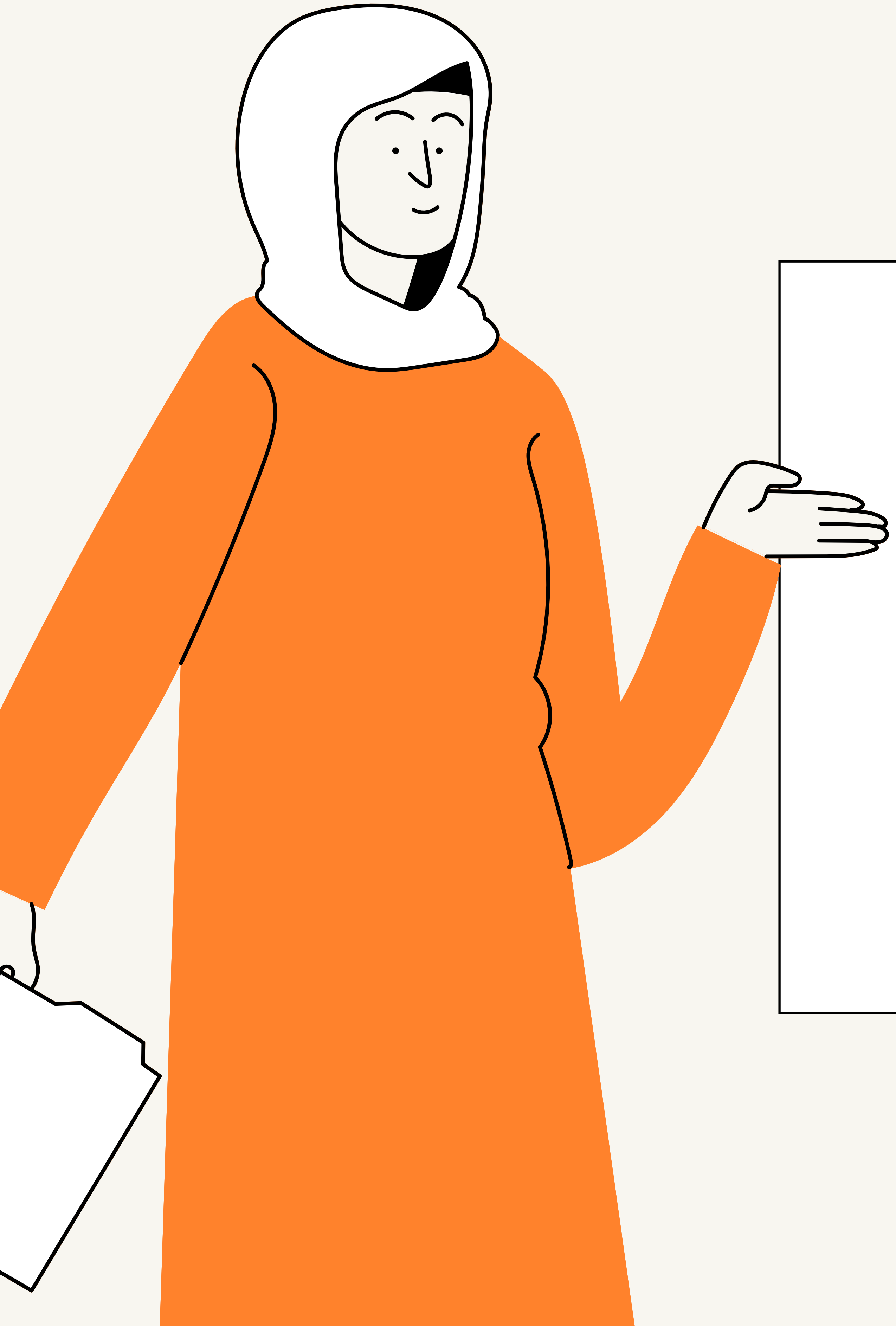
- Post Production process used in film making and video production
- Additional or supplementary recordings are lip-synced with original production sound to create the finished soundtrack.
- This is also termed as revoicing in the film industry.



Introduction

Problems with existing Dubbing System

- Manual Dubbing is currently used in film industry
- This process is
  -Tedious
  -Costly
  -Time Taking
  -Resource Bounding



Solution:

An Automated Dubbing System

- Automate the dubbing process instead of manual effort.
- Reduced manual input
- Less time consuming
- Lenient on financial overheads
- Can be extended across multiple linguistic barriers simultaneously.

What is An Automated Dubbing System?

Researchers at Amazon define ADS as:

“Automatic dubbing involves transcribing speech to text and translating that text into another language before generating speech from the translated text”

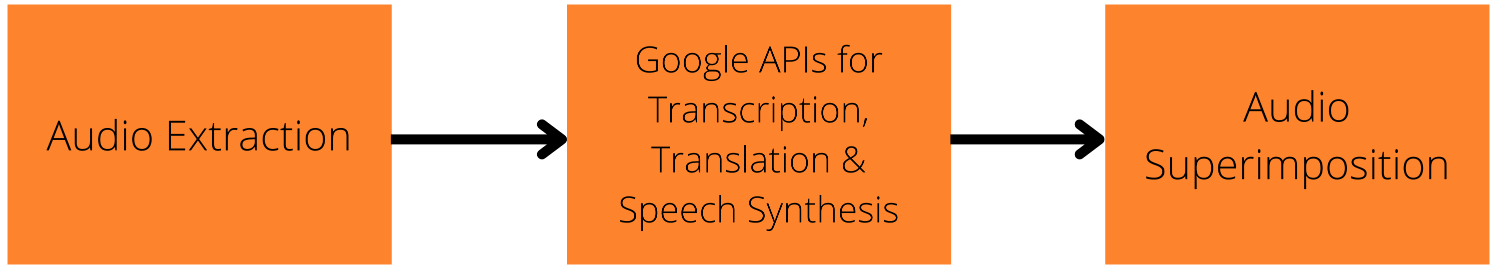
We define ADS as:

“An Automated System that takes in an input video file in a source language i-e English and uses machine translation and speech synthesis to produce a dubbed output file i-e Urdu in the required target language.”

Our Approach is more on “**content-based**” criterion than “**fluency based**” criterion

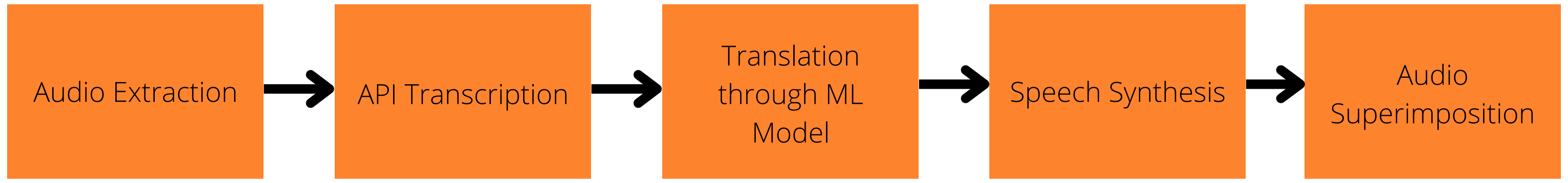
Methodology

ETHROUGH EXISTING APIs



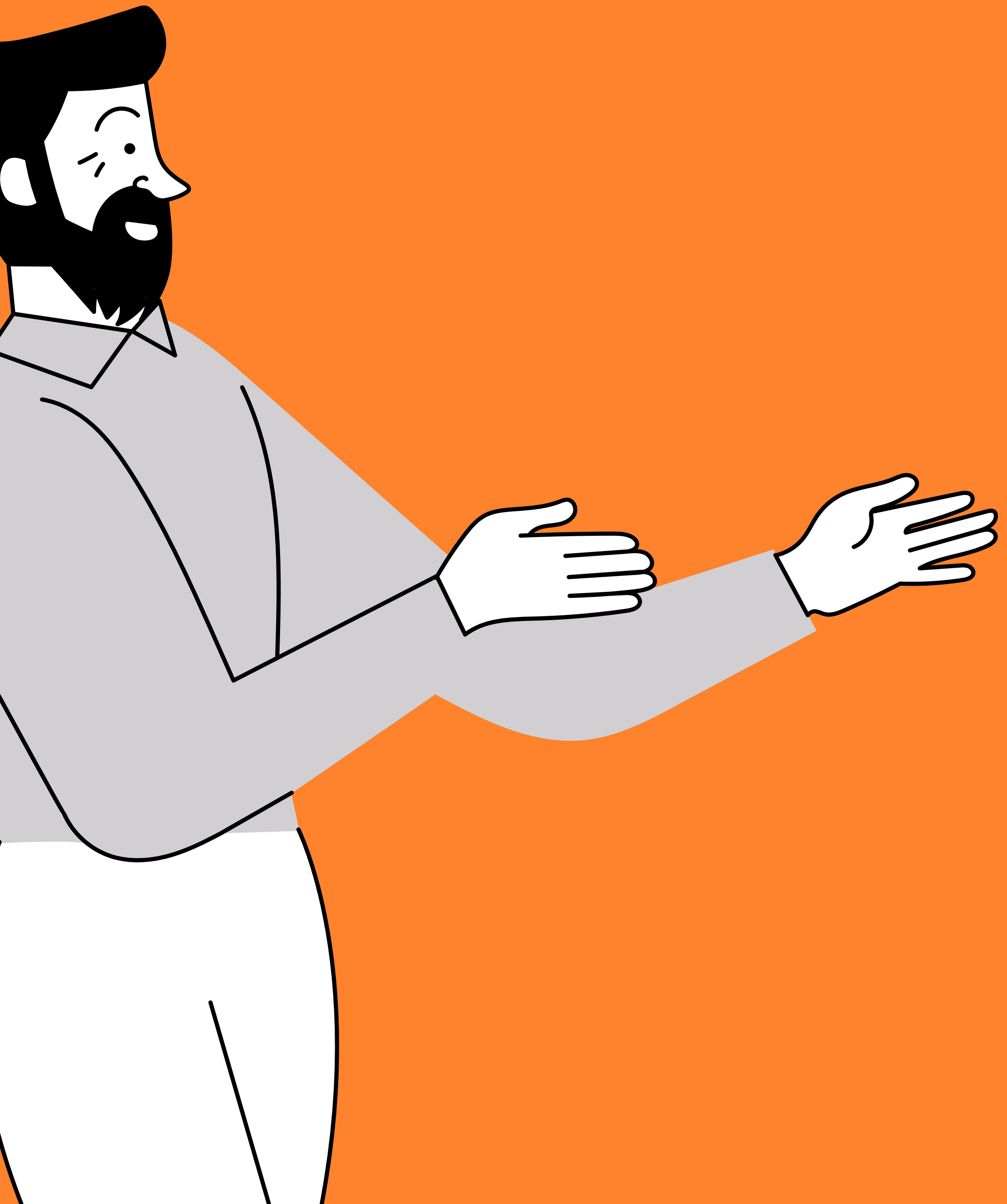
Audio Extraction → Google APIs for Transcription, Translation & Speech Synthesis → Audio Superimposition

THROUGH MACHINE LEARNING



Audio Extraction → API Transcription → Translation through ML Model → Speech Synthesis → Audio Superimposition

Applications



Can be used to dub educational English videos into Urdu for better understanding

Is helpful for automatic voiceovers in poems, shows, movies, and other videos

Automatic Dialogue Replacement (ADR) for actors

TOOLS USED FOR API MODEL

Tools including libraries and APIs used for the API end-to-end system

Google Speech-to-Text

Google Translate

MoviePy

Google Text-to-Speech

Google Storage

API Model Implementation

Phase 1: Audio Extraction

- Achieved through python library called MoviePy
- User Uploads a video
- MoviePy extracts audio from the video
- The audio is saved as an mp3 file

Phase 2: Transcription

- Achieved through Google Speech-to-Text API
- Uses Asynchronous Transcription Method
- Converts Audio to Text

Phase 3: Translation

- Achieved through Google Translate API
- Translates English text extracted from user video into Urdu language

Implementation

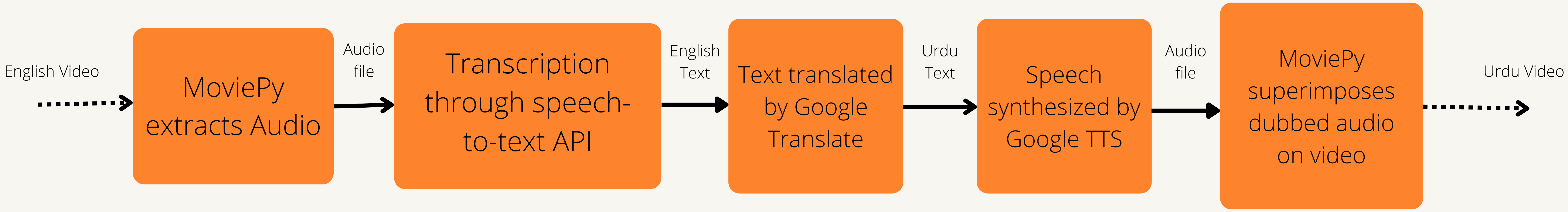
Phase 4: Text-to-Speech (TTS)

- Uses Python library *gTTs* built upon Google Text-to-Speech
- Converts the Urdu translation into audio
- Synthesizes Speech

Phase 5: Adding Dubbed Audio

- Achieved through MoviePy library
- Converts the Urdu translation into audio
- Adds the dubbed audio file to the original video

The Flow



English Video

MoviePy extracts Audio

Audio file

Transcription through speech-to-text API

English Text

Text translated by Google Translate

Urdu Text

Speech synthesized by Google TTS

Audio file

MoviePy superimposes dubbed audio on video

Urdu Video

API Model Results

Original Video

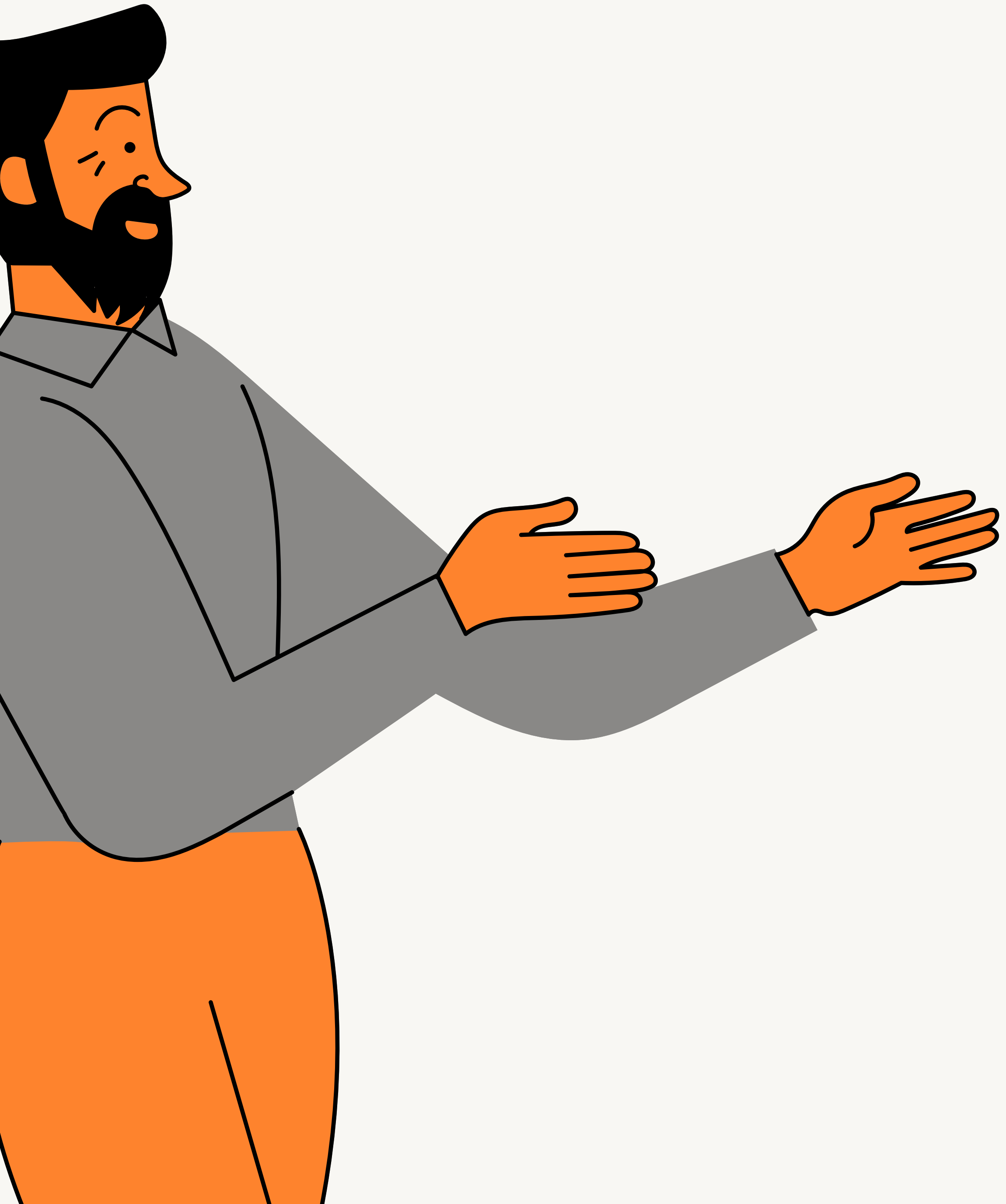
API System Dubbed Video

NEURAL NETWORKS

RNN or Recurrent Neural Network is commonly used for speech recognition and natural language processing.

It recognizes data’s sequential characteristics and uses patterns to predict the next likely scenario.

RNNS Explored

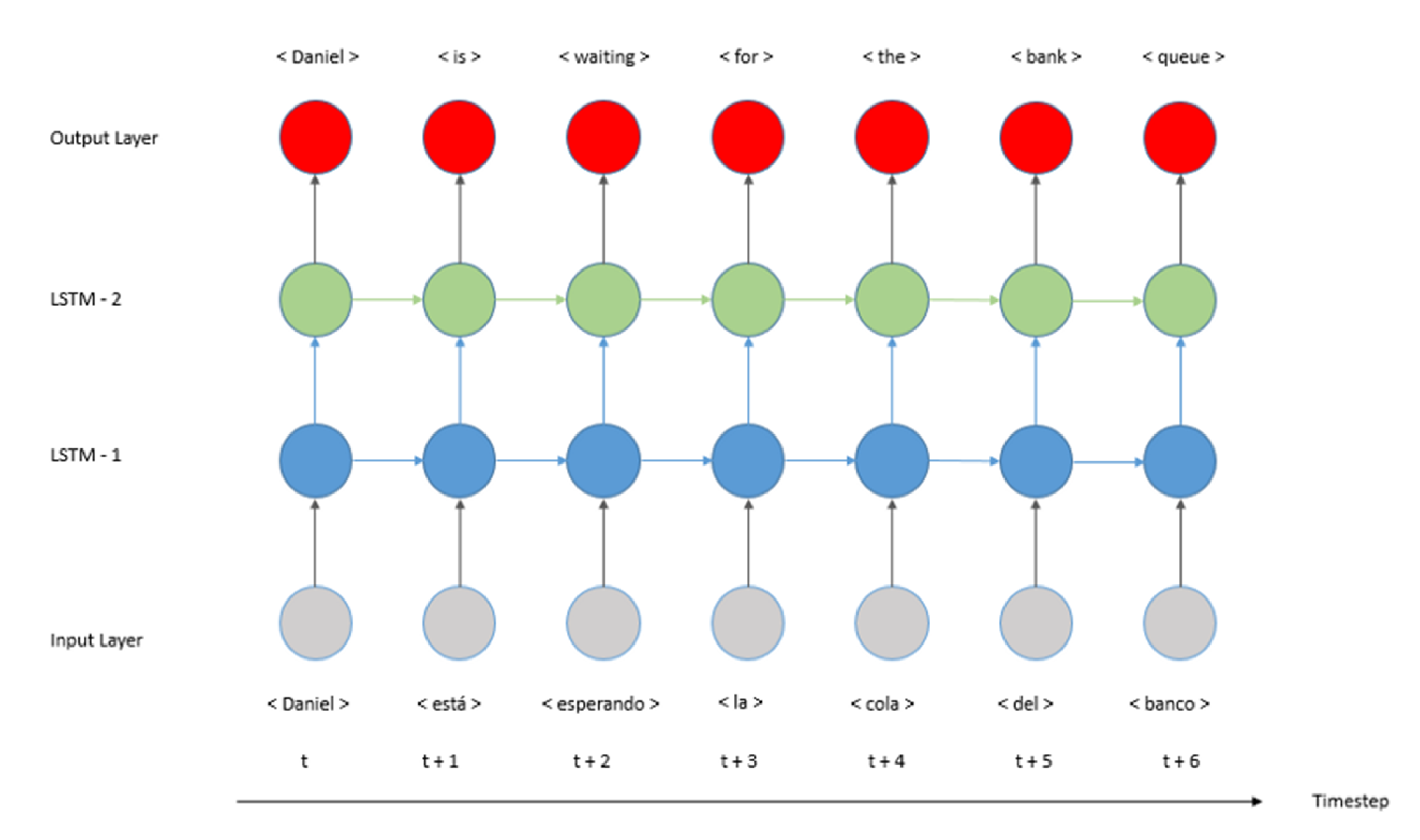


- Many to Many RNNs

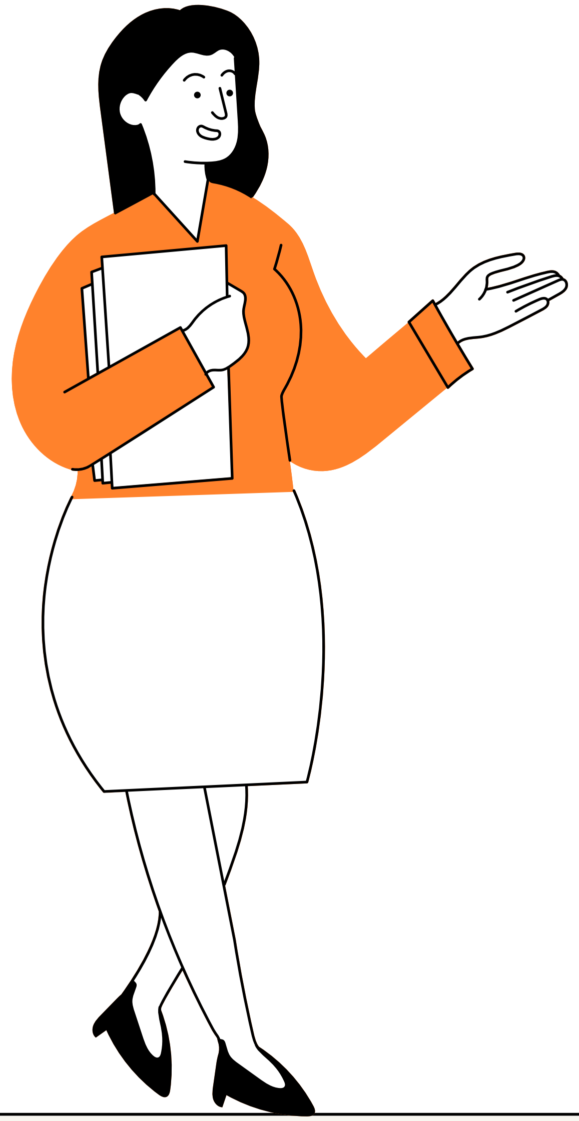
- Encoder Decoder RNNs

- Encoder Decoder RNN with LSTM

Issues with Many to Many RNN



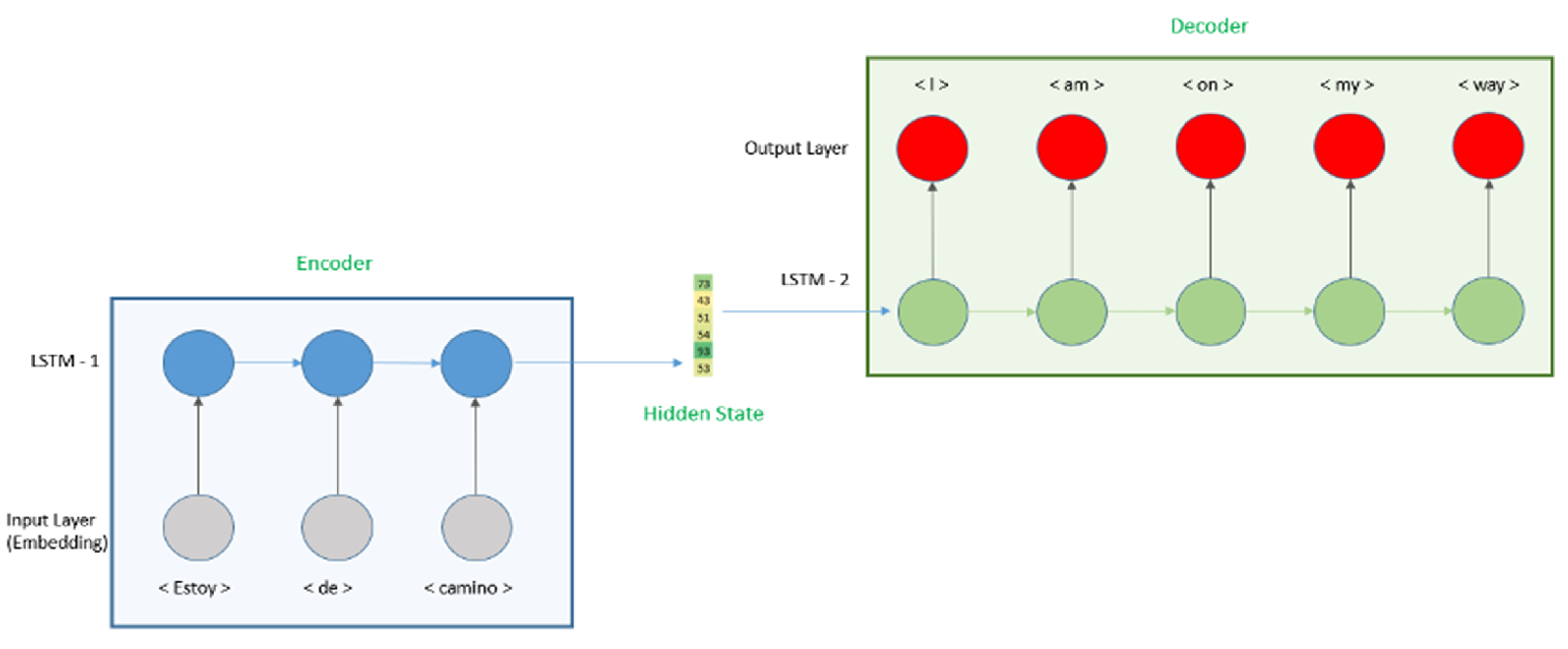
Encoder Decoder RNN



The **Encoder-Decoder LSTM** is a recurrent neural network designed to address sequence-to-sequence problems, sometimes called seq2seq. It consists of 3 main parts:

1. Encoder
2. Hidden States
3. Decoder

Architecture of Our RNN

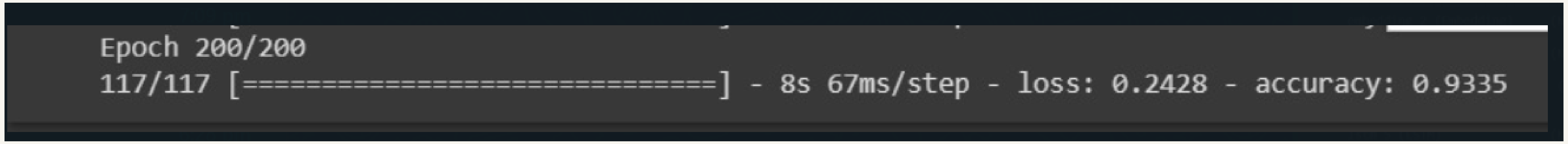


Dataset Analysis

73%	~2000 sentences
95%	~2000 sentences with Preprocessing
88%	~15000 sentences

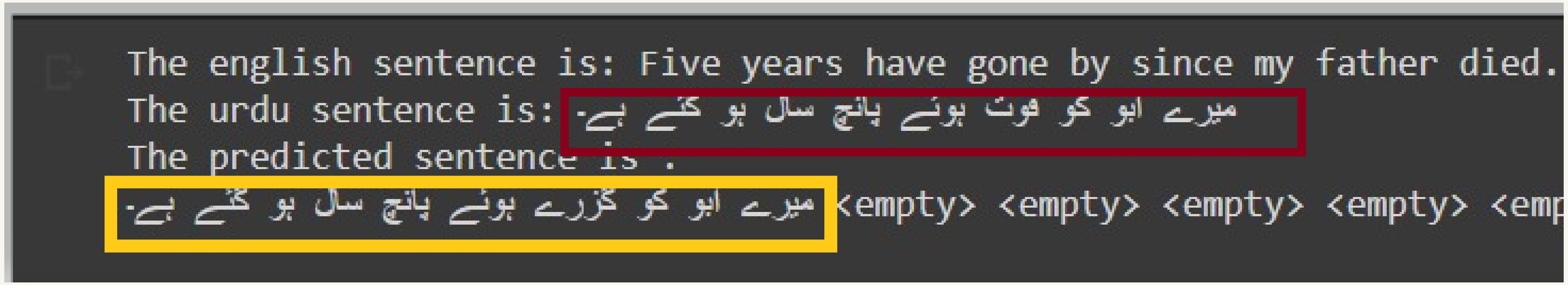
Results

- Accuracy was initially 63%
- This was increased to a 95% accuracy
- This was done in by experimenting with batch sizes and epochs as following results show:



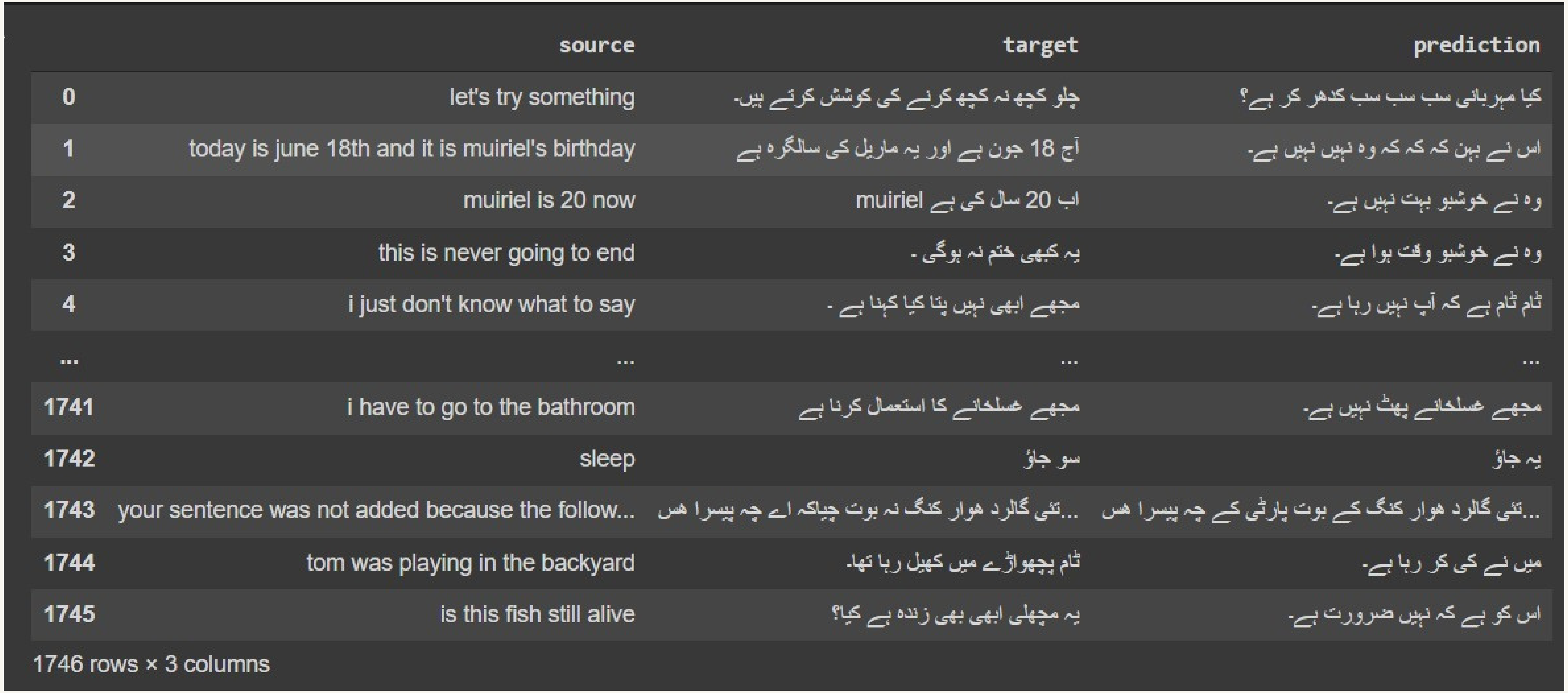
```
Epoch 200/200
117/117 [==============================] - 8s 67ms/step - loss: 0.2428 - accuracy: 0.9335
```

Accuracy



```
The english sentence is: Five years have gone by since my father died.
The urdu sentence is: میرے ابو کو فوت ہوئے پانچ سال ہو گئے ہے۔
The predicted sentence is :
میرے ابو کو گزرے ہوئے پانچ سال ہو گئے ہے۔ <empty> <empty> <empty> <empty> <emp
```

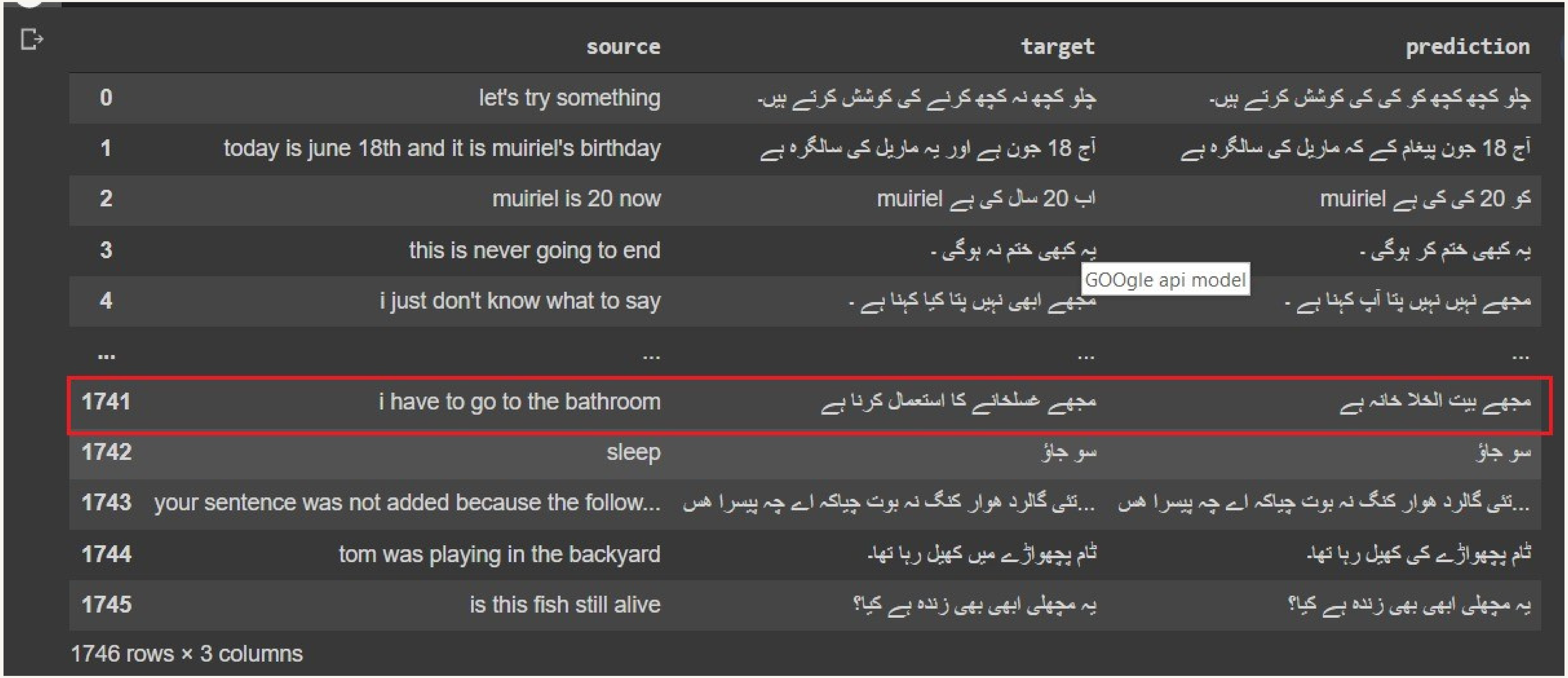
End Predictions



	source	target	prediction
0	let's try something	چلو کچھ نہ کچھ کرنے کی کوشش کرتے ہیں۔	کیا مہربانی سب سب کدھر کر ہے؟
1	today is june 18th and it is muiriel's birthday	آج 18 جون ہے اور یہ ماریل کی سالگرہ ہے	اس نے بہن کہ کہ کہ وہ نہیں نہیں ہے۔
2	muiriel is 20 now	اب 20 سال کی ہے muiriel	وہ نے خوشبو بہت نہیں ہے۔
3	this is never going to end	یہ کبھی ختم نہ ہوگی ۔	وہ نے خوشبو وقت ہوا ہے۔
4	i just don't know what to say	مجھے ابھی نہیں پتا کیا کہنا ہے ۔	ٹام ٹام ہے کہ آپ نہیں رہا ہے۔
...	...	...	...
1741	i have to go to the bathroom	مجھے غسلخانے کا استعمال کرنا ہے	مجھے غسلخانے پھٹ نہیں ہے۔
1742	sleep	سو جاؤ	یہ جاؤ
1743	your sentence was not added because the follow...	...تئی گالرد ھوار کنگ نہ بوت چیاکہ اے چہ پیسرا ھس	...تئی گالرد ھوار کنگ کے بوت پارٹی کے چہ پیسرا ھس
1744	tom was playing in the backyard	ٹام پچھواڑے میں کھیل رہا تھا۔	میں نے کی کر رہا ہے۔
1745	is this fish still alive	یہ مچھلی ابھی بھی زندہ ہے کیا؟	اس کو ہے کہ نہیں ضرورت ہے۔

1746 rows × 3 columns

After 200 epochs at 89% Accuracy



	source	target	prediction
0	let's try something	چلو کچھ نہ کچھ کرنے کی کوشش کرتے ہیں۔	چلو کچھ کچھ کو کی کی کوشش کرتے ہیں۔
1	today is june 18th and it is muiriel's birthday	آج 18 جون ہے اور یہ ماریل کی سالگرہ ہے	آج 18 جون پیغام کے کہ ماریل کی سالگرہ ہے
2	muiriel is 20 now	اب 20 سال کی ہے muiriel	کو 20 کی کی ہے muiriel
3	this is never going to end	یہ کبھی ختم نہ ہوگی ۔	یہ کبھی ختم کر ہوگی ۔
4	i just don't know what to say	مجھے ابھی نہیں پتا کیا کہنا ہے ۔	مجھے نہیں نہیں پتا آپ کہنا ہے ۔
...	...	...	...
1741	i have to go to the bathroom	مجھے غسلخانے کا استعمال کرنا ہے	مجھے بیت الخلا خانہ ہے
1742	sleep	سو جاؤ	سو جاؤ
1743	your sentence was not added because the follow...	...تئی گالرد هوار کنگ نہ بوت چیاکہ اے چہ پیسرا هس	...تئی گالرد هوار کنگ نہ بوت چیاکہ اے چہ پیسرا هس
1744	tom was playing in the backyard	ٹام پچھواڑے میں کھیل رہا تھا۔	ٹام پچھواڑے کی کھیل رہا تھا۔
1745	is this fish still alive	یہ مچھلی ابھی بھی زندہ ہے کیا؟	یہ مچھلی ابھی بھی زندہ ہے کیا؟

1746 rows × 3 columns

GOOgle api model

After 400 epochs at 93% Accuracy

RNN Model Results

Original Video

RNN System Dubbed Video

Work Division

Zainab Binte Iftikhar

- Understanding existing systems
- Google API Model
- Assorting a framework of APIs for text extraction, translation and speech synthesis
- Front End Development
- Data Creation for RNN Model
- Data Preprocessing for ML models

Amna Abid

- Understanding existing research work and models
- Designing an architecture of the system
- Implementing a Statistical Model (HMM)
- Designing a neural network for machine translation
- Inferring their results for machine translation

Conclusion & Future Work

The current effort has produced an end to end usable dubbing system with around 90-93% accuracy in translation of text from English to Urdu. This work can further be extended to improve upon:

- Speech Synthesis
- Prosodic Alignment
- Dataset

THANK YOU!

References

- Federico, M., Enyedi, R., Barra-Chicote, R., Giri, R., Isik, U., Krishnaswamy, A. and Sawaf, S., 2020. *FROM SPEECH-TO-SPEECH TRANSLATION TO AUTOMATIC DUBBING*. [online] Available at: <https://arxiv.org/pdf/2001.06785.pdf> [Accessed 7 December 2021].
- Oktem, A., Farrus, M. and Bonafonte, A., 2019. [online] Available at: <https://arxiv.org/pdf/1908.07226.pdf> [Accessed 7 December 2021].
- Yang, Y., Shillingford, B., Assael, Y. and Freitas, N., 2020. .
- Rauf, S., Abida, S. and Bashir, J., 2021. [online] Available at: <https://aclanthology.org/2020.sltu-1.40.pdf> [Accessed 7 December 2021].
- Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark and R. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," 29 March 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1703.10135. [Accessed 10 May 2022].
- Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," 8 June 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2006.04558. [Accessed 10 May 2022].