KLE Society's
KLE Technological University

A Mini Project Report

On

LANGUAGE TRANSLATION USING MACHINE LEARNING ALGORITHMS

submitted in partial fulfillment of the requirement for the degree of

Bachelor of Engineering in

Computer Science and Engineering

submitted by:

JAYASHREE A V	**01FE16BCS082**
SHRIYA HIREHOLI	**01FE16BCS079**
JYOTHI S HOSAMANI	**01FE16BCS083**
NAYONIKA KADABI	**01FE16BCS118**

Under the guidance of
MRS. MANJULA K.PAWAR

.

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

HUBLI – 580 031 (India).

Academic year 2018-19

KLE Society's
KLE Technological University

2018 - 2019

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that Mini Project entitled **LANGUAGE TRANSLATION USING DEEP LEARNING ALGORITHMS** is a bonafied work carried out by the student team Ms. Jayashree A V – 01FE16BCS082, Ms. Shriya Hireholi – 01FE16BCS079, Ms. Jyothi S Hosamani - 01FE16BCS083, Ms. Nayonika Kabadi - 01FE16BCS118, in partial fulfillment of completion of Fifth semester B. E. in Computer science and Engineering during the year 2018 – 2019. The project report has been approved as it satisfies the academic requirement with respect to the project work prescribed for the above said program.

Guide **Head of SoCSE**

Asst. Prof. Mrs Manjula Pawar **Dr. Meena S. M**

External Viva:

Name of the Examiners	**Signature with date**
1.	
2.	

ABSTRACT

In large societies like India there is a huge demand to convert one human language into other. Lots of work has been done in this area. Many transfer based MTS (Machine Translation System) have been developed for English to other languages, as MANTRA CDAC Pune, MATRA CDAC Pune, SHAKTI IISc Bangalore and IIIT Hyderabad. Our focus is to design an app that translates the text from English to Hindi or vice versa using machine learning algorithms. This system takes an input text check its structure through parsing. Machine Learning algorithms are used to generate the text in target language. We get correct translation for simple assertive sentences and almost correct for complex and compound sentences.

ACKNOWLEDGEMENTS

Every project big or small is successful largely due to the effort of a number of wonderful people who have always given their valuable advice or lent a helping hand. We sincerely appreciate the inspiration; support and guidance of all those people who have been instrumental in making this project a success.

We the team 4 of 5th sem B division, KLE Tech University are extremely grateful to the faculty of the college for the confidence bestowed in us and entrusting our project entitled "Language Translation using Deep Language Algorithms".

We are obliged to have such energy provoking mentor Asst. Prof. Mrs. Manjula Pawar who made our work easier by enhancing our enthusiasm and passion for learning new and interesting things and guided us thoroughly.

We feel deeply honored in expressing our sincere thanks Asst. Prof. Mr. Mahesh Patil and Asst. Prof. Mr. Prakash Hegde for providing valuable insights leading to the completion of design process of our project.

We would also like to thank all the faculty members of our college for their critical advice and guidance without which this project would not have been possible.

1. INTRODUCTION

1.1 OVERVIEW OF THE PROJECT

Translation is the communication of the meaning of a source-language text by means of an equivalent target-language text. Translators have helped shape the very languages into which they have translated. The significance of translation in our daily life is extensively multidimensional. Not only does translation pave the way forward for global interaction, but allows nations to forge interactive relationships when it comes to making advancements in technology, politics, etc. Hence we intend to provide a solution to develop an app that does translation in native language (from English to Hindi).

1.2 MOTIVATION

English is the third most widely-spoken language in terms of native speakers, of which it has at least 330 million. But if you count the people who speak it as a second language, it's the most popular language in the world. So, here are 5 factors that motivate us to take this project.

- Not everyone speaks English
- People prefer their native language
- Connects the global economy
- Translation spreads ideas and information
- Emerging markets mean emerging languages

1.3 OBJECTIVES

- To read the output in the desired target language
- To translate from one language to other

1.4 PROBLEM STATEMENT

To develop an app that translates text input from English to Hindi language and vice versa using domain specific model.

2. PROPOSED SYSTEM

2.1 DESCRIPTION OF SYSTEM WITH BLOCK DIAGRAM

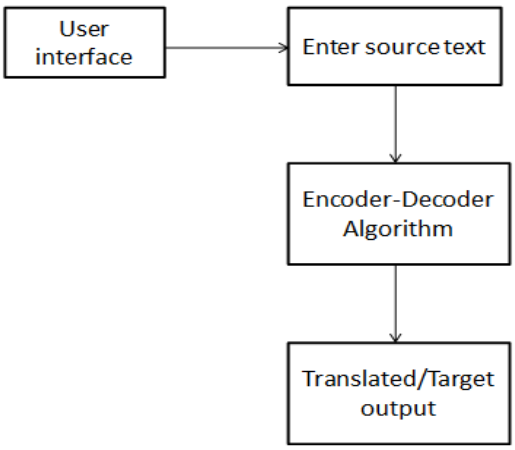

Figure: 1 Block Diagram

Translation is the communication of the meaning of a source-language text by means of an equivalent target-language text. A translator always risks inadvertently introducing source-language words, grammar, or syntax into the target-language rendering.
Computer-assisted translation (CAT), also called "computer-aided translation", "machine -aided human translation" (MAHT) and "interactive translation," is a form of translation wherein a human translator creates a target text with the assistance of a computer program. The machine supports a human translator.

Machine translation (MT) is a process whereby a computer program analyzes a source text and, in principle, produces a target text without human intervention. In reality, however, machine translation typically does involve human intervention, in the form of pre-editing and post-editing. With proper terminology work, with preparation of the source text for machine translation (pre-editing), and with reworking of the machine translation by a human translator (post-editing), commercial machine-translation tools can produce useful results, especially if the machine-translation system is integrated with a translation-memory or globalization-management system.

2.2 DESCRIPTION OF PROPOSED SYSTEM

Target users are:

- Common people who don't know English or Hindi
- Students who are interested in higher education or jobs in foreign countries
- Employees for better communication
- People who want to learn different languages (here English or Hindi)
- Business people not to get cheated because of miscommunication

2.3 ADVANTAGES OF PROPOSED SYSTEM

- Curiosity to learn languages
- No miscommunication
- Deliver information in multiple languages
- Increase human translation productivity
- Create and manage enterprise language as a corporate asset
- Integrate with enterprise application
- Understanding cultural identities and differences
- Accurate transfer of news

2.4 SCOPE

- The system has 2 languages Hindi and English for translation
- Length of input text should be maintained
- No spelling mistakes
- Proper sentence or word

3. SOFTWARE REQUIREMENT SPECIFICATION

3.1 OVERVIEW OF SRS

This section describes the nature of a project, software or application. It provides a detailed overview of our software product, its parameters and goals. This includes the purpose, scope, functional and non-functional requirements, software and hardware requirements of the project.

3.2 REQUIREMENT SPECIFICATIONS

3.2.1 FUNCTIONAL REQUIREMENTS

System: should be able to

- recognise the word
- provide Hindi and English language
- translate into particular language
- search the word in database
- maintain vocabulary

User: shall be able to

- enter text
- select language
- read the output

3.2.2 USE CASE DIAGRAMS

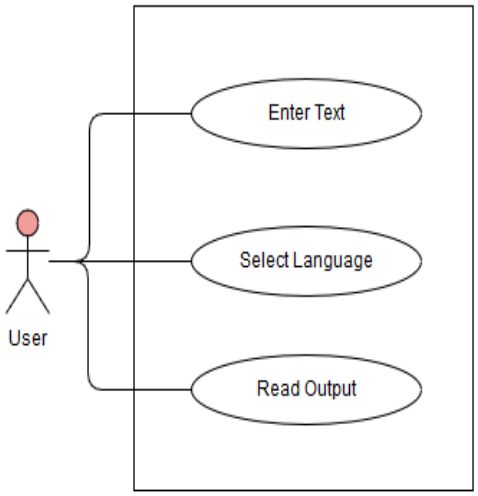

Figure: 2 Use Case Diagram

3.2.3 USE CASE DESCRIPTIONS USING SCENARIOS

Use case: Enter Text
Actors: End- Users
Pre-condition:

- Application should give a text field to enter the text.
- User should be able to type the text.
- Entered text must be either English or Hindi

Post condition:

- Text should be grammatically correct.

Main Success scenario:
Step 1: Enter the text, either in English or Hindi.
Step 2: Entered text must be grammatically correct.
Exception Scenario:

- Vocabulary exception
- Syntax exception
- Spelling exception

Use case: Select Language

Actors: End- Users

Pre-condition:

- Application should provide dropdown field to select the language.

Post condition:

- Converts the entered text into selected language.

Main Success scenario:

Step 1: Select the language specified in the dropdown filed.

Exception Scenario:

- When language is not selected, exception is raised**.**

Use case: Read output

Actors: End- Users

Pre-condition:

- Application should give a text field to display the text.
- User should be able to read the text.
- Displayed text must be either English or Hindi

Post condition:

- Should display the output.

Main Success scenario:

Step 1: Displayed output should be understandable by the user and should be in particular format.

Exception Scenario:

- Vocabulary.

3.2.4 NONFUNCTIONAL REQUIREMENTS

- Response time of 0.25-0.30 seconds
- Accuracy of 80%
- Resource behaviour 80%
- Install ability 90%

3.3 SOFTWARE REQUIREMENT SPECIFICATIONS

Software:

- anaconda3 v5.2.0
- numpy
- python v2.7 and above
- tensorflow v1.10 and above
- jupyter notebook v4.40

3.4 GUI OF PROPOSED SYSTEM

Figure: 3 GUI

3.5 ACCEPTANCE TEST PLAN

User enters the text

Test case id	Input Description	Expected output	Actual output
1.	Verify the morphological analysis of the source text with punctuation (eg : Hello!!)	Punctuation are removed (eg : Hello)	
2.	Verify the morphological analysis of the source text without punctuation (eg : Hello)	Goes to the verification of syntactical analysis	
3.	Verify the syntactical analysis of the source text with correct input (eg : How are you?)	Goes to the verification of semantic analysis	
4.	Verify the syntactical analysis of the source text without correct input (eg : How you are?)	Word sequence will be rejected	
5.	Verify the semantic analysis of the source text with correct input (eg : come)	Mapping is done (eg : आइए)	
6.	Verify the semantic analysis of the source text without correct input (eg : cum)	Word will be rejected	

Select language

Test case id	Input Description	Expected output	Actual output
1.	If the selected language is English	The target language must be displayed in English	
2.	If the selected language is Hindi	The target language must be displayed in Hindi	
3.	If the selected language is other than these two languages	Pop up message “Choose specified language”	
4.	If no language is selected	Error message is displayed “Choose the language ”	
5.	If more than one language is selected	Error message is displayed “Choose only one language”	

Read the output

Test case id	Input Description	Expected output	Actual output
1.	If the output has grammatical mistakes	Reframe the sentence	
2.	If the output is not readable	Reframe the sentence	
3.	If the output is in a proper format	Display the output	
4.	If the output has no grammatical mistakes	Display the output	
5.	If the output target language is not selected	Popup message “Select the target language”	
6.	If the target language selected by user is same as source language	Popup message “Select proper target language”	

System:

Search the word in database

Test case id	Input Description	Expected output	Actual output
1.	Word is found in database	Translate word	
2.	Word is not found in database	Pop up error message “Word not found”	
3.	System gives wrong result for search	Search again	
4.	System provides more than one result for search	Choose appropriate one	
5.	System provides no result for search	Popup message “Word not found”	
6.	System provides proper result for search	Translate	

Maintain vocabulary

Test case id	Input Description	Expected output	Actual output
1.	If the input text does not maintain vocabulary	Popup message “Maintain vocabulary”	
2.	If the input text maintains vocabulary	Goes for splitting for sentence	
3.	If the output text does not maintain vocabulary	Reframe the sentence	
4.	If the output text maintains vocabulary	Display the output	

Recognize the word

Test case id	Input Description	Expected output	Actual output
1.	If the system recognizes the word	Find the meaning and translate	
2.	If system does not recognize the word	Pop up error message “The word not found”	
3.	If the system recognizes the word in a wrong way	Pop up error message “Enter the correct word ”	
4.	If the system recognizes the word in an improper format	Pop up error message “Enter the correct format ”	

4. SYSTEM DESIGN

4.1 ARCHITECTURE OF THE SYSTEM

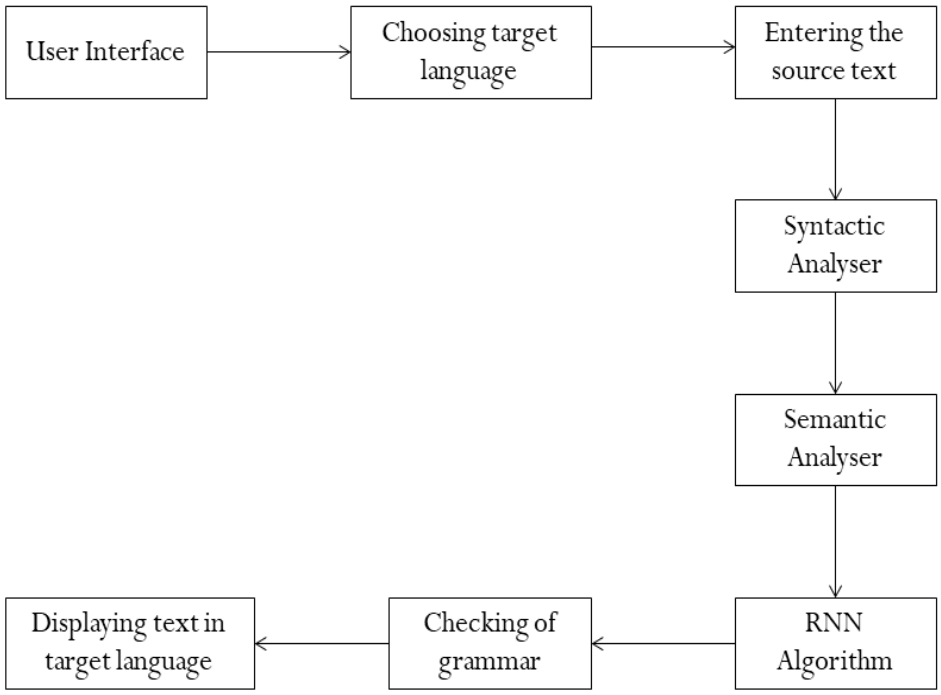

Figure: 4

4.2 LEVEL 0 DFD

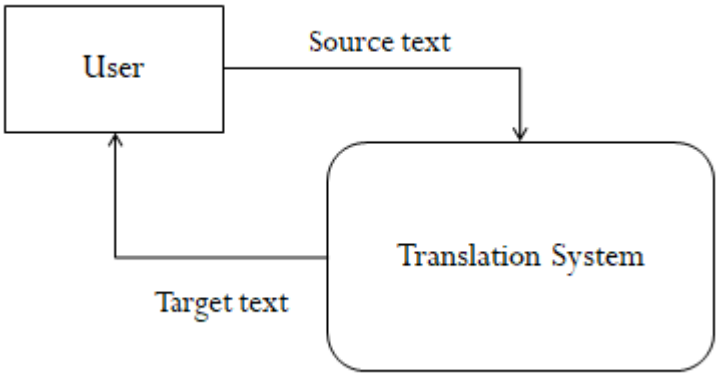

Figure: 5

4.3 DETAILED DFD FOR THE PROPOSED SYSTEM

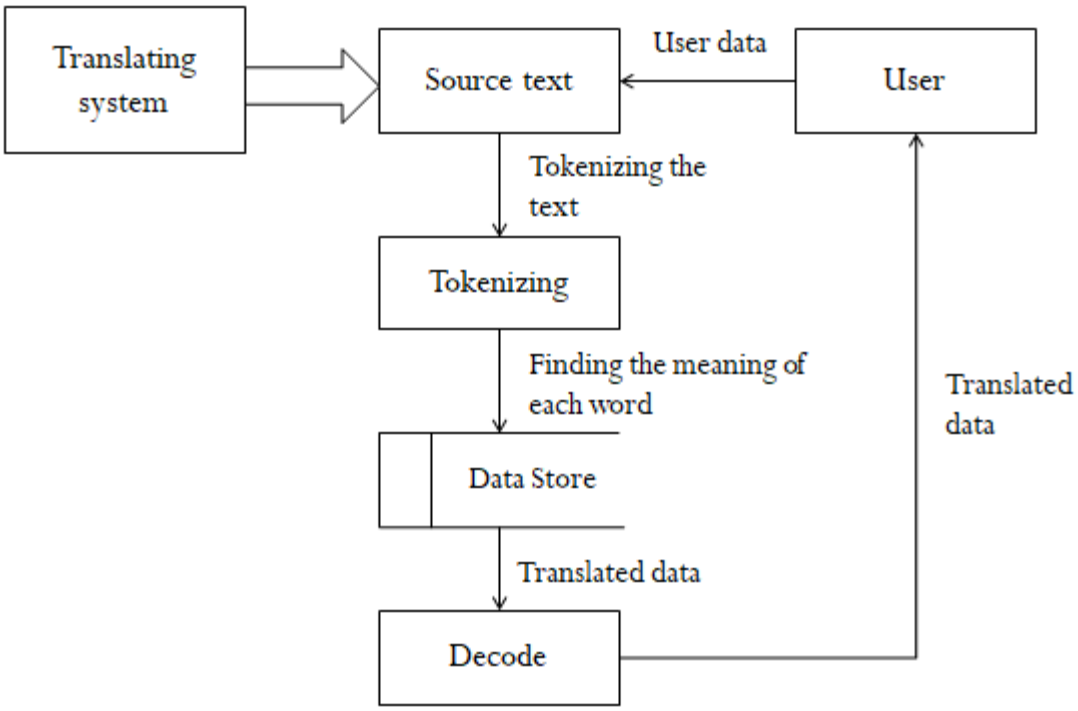

Figure: 6

4.4 ACTIVIY DIAGRAM

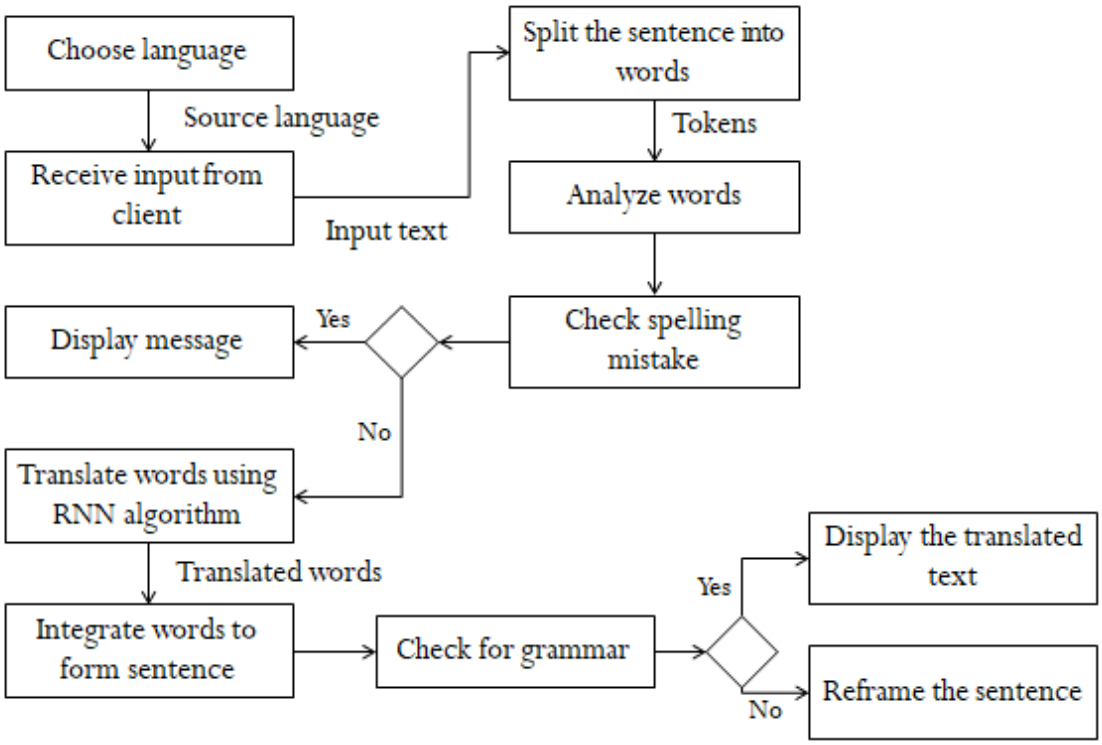

Figure: 7

5. IMPLEMENTATION

5.1 PROPOSED METHODOLOGY

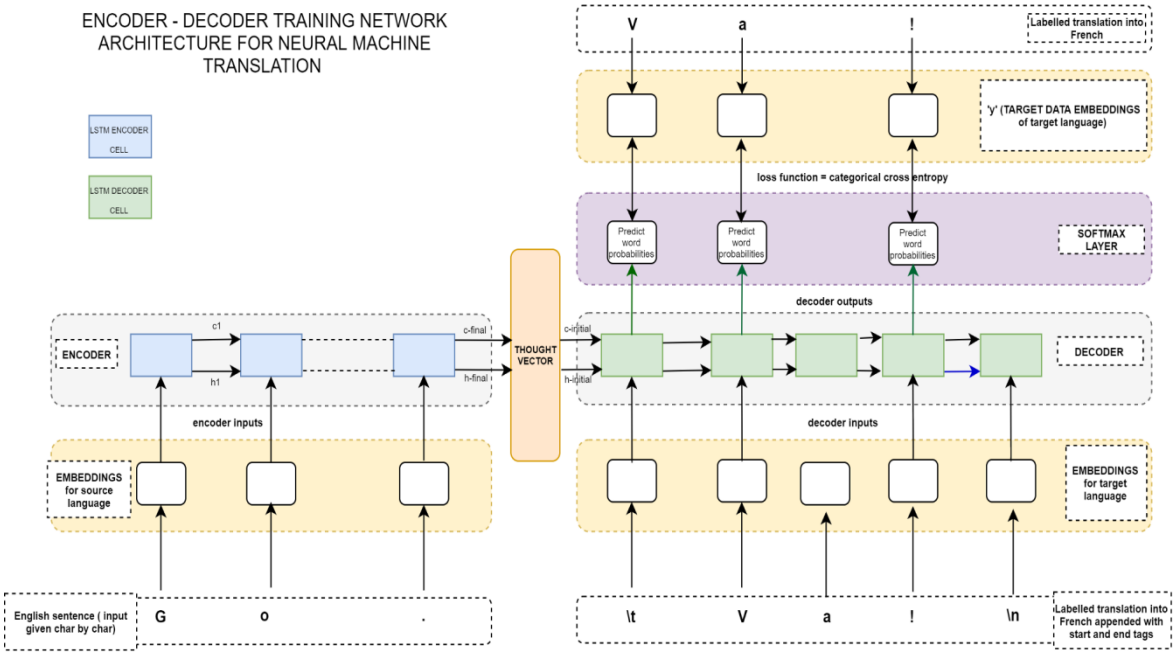

Figure: 8

This is the detailed network architecture used for training the seq2seq Encoder-Decoder network (RNN network).

The first step is **Training the network—**

1. Create one-hot character embedding for English and Hindi sentences. These will be the inputs to the encoder and the decoder. The Hindi one-hot character embeds will also be used as target data for loss function.
2. Feed character by character embeds into the encoder till the end of the English sentence sequence.
3. Obtain the final encoder states (hidden and cell states) and feed them into the decoder as its initial state.
4. Decoder will have 3 inputs at every time step—2 decoder states and the Hindi character embed fed to it character by character.
5. At every step of the decoder, the output of the decoder is sent to softmax layer that is compared with the target data.

The next step is **Testing (Inference mode) —**

Below is the architecture used for inference models —The inference model will leverage all the network parameters learnt during training but we define them separately because the inputs and outputs during inference are different from what they were during training the network. Here we feed the new English sentence (one hot character embedded) vector as input sequence to the encoder model and obtain the final encoding states.

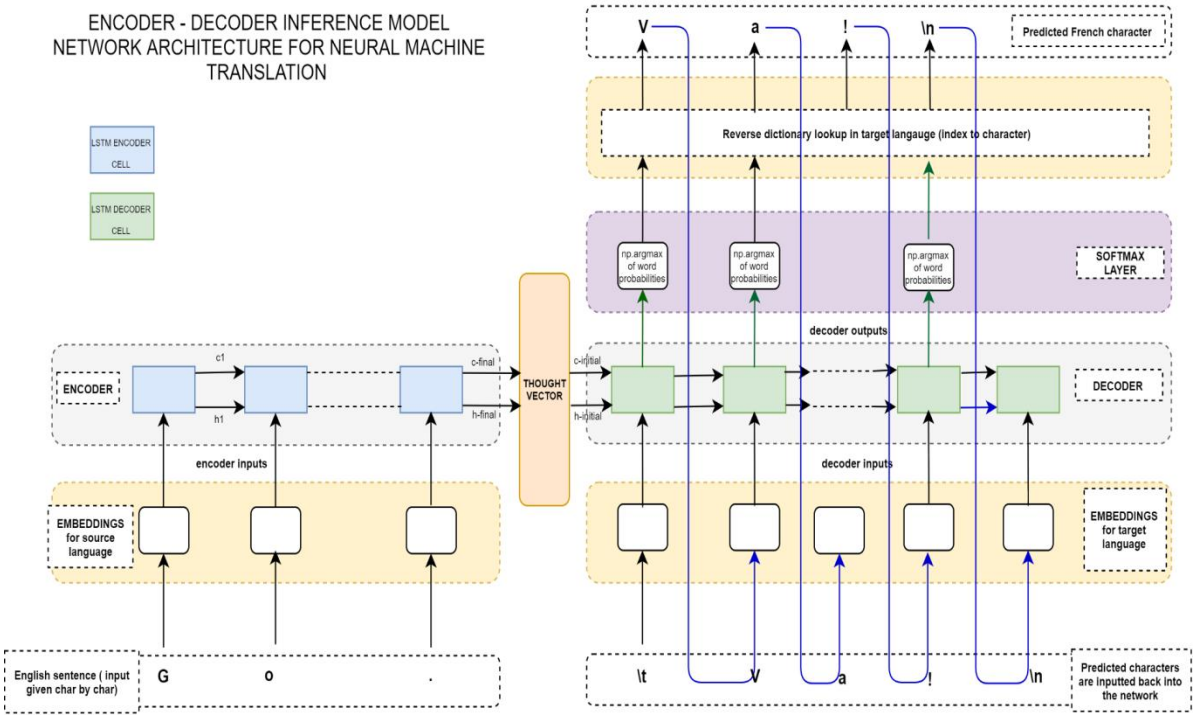

Figure: 9

Contrast this figure with previous figure on the decoder side. The major changes can be seen are as below —

- At the first time step, the decoder has 3 inputs—the start tag '\t' and the two encoder states. We input the first character as '\t' (its one hot embed vector) into the first time step of the decoder.
- Then the decoder outputs the first predicted character (assume it is 'V').
- Observe how the blue lines are getting connected back into the decoder input for the next time step. So this predicted character 'V' will be fed as an input to the decoder at the next timestep.
- Also note that we only obtain the one hot embed vector of the predicted character using the np.argmax function on the output of the softmax layer at each timestep. So we do a reverse dictionary lookup on the index to obtain the actual character 'V'.
- From next time step on wards the decoder still has 3 inputs but different from the first time step. They being—one hot encode of previous predicted character, previous decoder cell state and the previous decoder hidden state

6. TESTING

6.1 TEST CASES

Test case id	Input Description	Expected output	Actual output
1.	Index out of range (>1000 and <0)	No output	No output
2.	Proper index value	Correct output	Correct output
3.	No index value	No output	No output
4.	Input sentence in dataset is incorrect	System error	No output

7. RESULTS AND DISCUSSIONS

```
-
Input sentence: how are you
Decoded sentence: START_ तू कैसा है _END
-
Input sentence:  मुझे बताओ क्या हुआ। _END
Decoded sentence: tell me what happened
```

Figure: 10

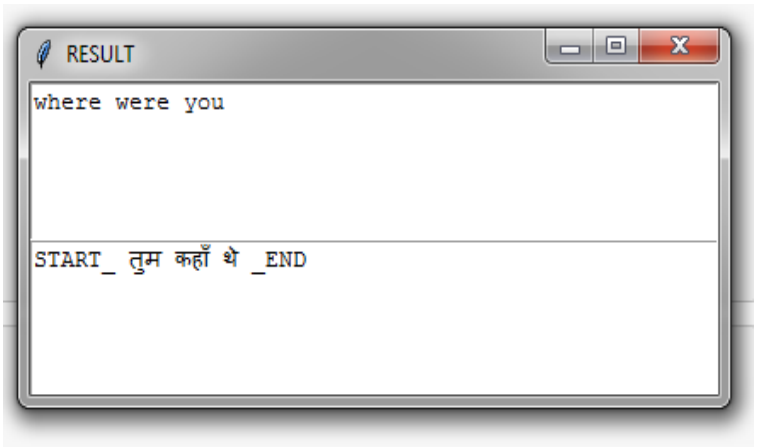

Figure: 11

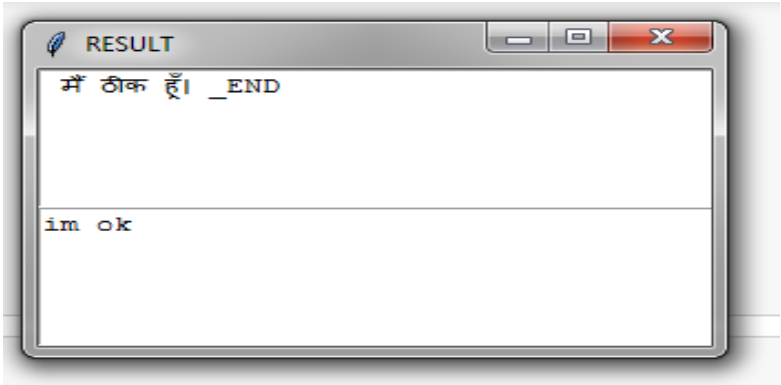

Figure: 12

8. REFERENCES/BIBLIOGRAPHY

- http://colah.github.io/posts/2015-08-Understanding-LSTMs/
- https://krazytech.com/projects/sample-software-requirements-specificationsrs-report-airline-database
- https://www.coursera.org/lecture/language-processing/encoder-decoder-architecture-bGV7m
- http://anie.me/rnn-encoder-decoder/
- https://www.softwaretestingclass.com/software-requirement-specification-srs/