A MINI-PROJECT REPORT

ON

"JAPANESE TO ENGLISH TRANSLATOR"

BY

Nitish Talekar
Sarvesh Wanode
Sarthak Vage

Under the guidance of

Internal Guide
Prof. Suresh Mestry

MANJARA CHARITABLE TRUST
RAJIV GANDHI INSTITUTE OF TECHNOLOGY

Juhu-Versova Link Road Versova, Andheri(W), Mumbai-53

Department of Computer Engineering

University of Mumbai

April- 2020

MANJARA CHARITABLE TRUST

RAJIV GANDHI INSTITUTE OF TECHNOLOGY

Juhu-Versova Link Road Versova, Andheri(W), Mumbai-53

C E R T I F I C A T E

Department of Computer Engineering

This is to certify that

1. Nitish Talekar
2. Sarvesh Wanode
3. Sarthak Vage

Have satisfactorily completed this project entitled

"JAPANESE TO ENGLISH TRANSLATOR"

Towards the fulfilment of the

FOURTH YEAR OF BACHELOR OF ENGINEERING
IN
(COMPUTER ENGINEERING)

as laid by University of Mumbai.

Guide
Prof. Suresh Mestry

H.O.D.
Dr.Satish Y. Ket

Principal
Dr. Sanjay Bokade

Project Report Approval for B. E.

This project report entitled **"*Japanese to English Translator*"** by ***Nitish Talekar , Sarvesh Wanode , Sarthak Vage*** is approved for the degree of ***Fourth-Year Bachelor of Computer Engineering***.

Examiners:

1----------

2.----------

Date:

Place

DECLARATION

We wish to state that the work embodied in this project titled “Japanese to English Translator” forms our own contribution to the work carried out under the guidance of “Prof. Suresh Mestry” at the Rajiv Gandhi Institute of Technology.

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Students Signatures)

Nitish Talekar (B- 862) _____

Sarvesh Wanode (B-869) _____

Sarthak Vage (B-867) _____

ABSTRACT

Demand for translation has increased due to drastic boost in globalization. Translation services are required across the globe to provide smooth interaction between different communities and organizations. Human based translation is precise and most accurate but at the same time it requires time, expenses and efforts from individual parties. Hence Machine based translators are a need of today to ensure quick and easy translation. The Japanese language is ranked amongst the most complex languages. Thus Japanese to English machine translation is a complex task.

This project endeavours to solve the complexity in Japanese to English Machine translation by using various Natural Language processing techniques in combination with neural networks to obtain precise and accurate translations. The project uses Long Short-Term Memory Network in a sequential model to provide translated sentences.The project also provides a user friendly UI to interact with the user and provide easy translations.

The application of this project is to provide Japanese translations of sentences in a swift and accurate manner to bridge the gap in communication amongst Japanese and English parties.

Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHM

CHAPTER 1

Introduction

1.1 Introduction Description

In todays rapidly changing and interactive world, exchange of information has become prominent. A consequence of this is the need to translate the knowledge base of different organizations to obtain their knowledge. Hence Machine based translation is an essential requirement. Texts in Japanese language are required to be translated in quick and accurate manner for proper, inexpensive information exchange. Hence this project develops a machine-based Japanese to English translation system that is trained on a proper dataset with sufficient accuracy.

The existing techniques use a various set of models and different kinds of datasets to procure the best translation. Japanese to English translation involves complex understanding of the languages and is an area of interest for Machine Translations worldwide.

This project provides a web-based platform to translate Japanese sentences to English sentences with the help of Artificial Neural Networks and several Natural language pre-processing algorithms. The existence of this automated system will help in understanding the Japanese language in a quick, inexpensive and easy way.

1.2 Organization of report

Describe every chapter (what every chapter contain)

- Ch.1 Introduction: An introduction to motivation for the project, proposed plan and system and a brief description of existing system.
- Ch.2 Literature Review: A survey of existing systems and current methods of dealing with Japanese to English translation queries.
- Ch.3 Proposed System: The detailed description of proposed system using Natural Language processing techniques.
- Ch.4 Results & Simulation: The resulting output system and analysis of results obtained during development.
- Ch.5 Conclusion: The overall summary of the project complete with results and analysis

CHAPTER 2

Literature Review

2.1 Survey existing system

2.1.1 Neural Network Assisted System

Machine translation system attempts to draw together three different forms of knowledge representation in an integrated expert system. Procedural techniques are used for the initial pre-processing stage, where Japanese sentences are divided into words, tagged with a part-of- speech identifier, and expressed in a unique form. Declarative techniques are then used to perform the actual translation of sentences from Japanese to English. [1]

2.1.2 Super Function Based System

A Method called Super-Function Based Machine Translation (SFBMT). SFBMT uses Super-Function (SF) to translate without syntactic and semantic analysis as most conventional MT systems do. SF rcpresents correspondence with a source langunge sentence structure and a target language sentciicc structure directon-. According to this feature. the system realizes very fast translation. [2]

2.2 Limitation of existing systems or Research gap

In the future, we need to narrow the translation candidate of SF using a concept dictionan: Besides, we need to improve the quality of SF by adjusting the sentence end of a Japanese sentence.

Kanji is not processed directly first it is converted to hizragana so computational power is wasted over that process of conversion.

2.3 Problem Statement and Objectives

A high demand for translation services due to globalization of industries and cultures has favoured in the need for machine translation because of its quick and inexpensive nature. Japanese language translation to English is an essential block of translating services around the globe that has a high degree of complexity. Hence the propped system is to design a model to resolve the complexity in this translation and provide quick accurate results in a user-friendly system with accessible interface.

2.3.1 Objectives

- To understand and break down the complexity of Japanese language
- To compare different methods and models of translation and select the best path for accurate translation.
- To design a user-friendly interface for easy translation of Japanese sentences.

2.4 Scope

The next stage of this project could be to build a model that take highly complex Japanese sentences into consideration. Models could be built with higher accuracy and less response time. This project could be further developed to provide audio translations using text-speech conversions.

CHAPTER 3

Proposed System

3.1 Framework and Algorithms

3.1.1 Keras

Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano. It was developed with a focus on enabling fast experimentation. Being able to go from idea to result with the least possible delay is key to doing good research.

Keras proper does not do its own low-level operations, such as tensor products and convolutions; it relies on a back-end engine for that. Even though Keras supports multiple back-end engines, its primary (and default) back end is TensorFlow, and its primary supporter is Google. The Keras API comes packaged in TensorFlow as tf.keras, which will become the primary TensorFlow API as of TensorFlow 2.0.

3.1.2 LSTM

The Encoder-Decoder LSTM is a recurrent neural network designed to address sequence-to-sequence problems, sometimes called seq2seq.This architecture is comprised of two models: one for reading the input sequence and encoding it into a fixed-length vector, and a second for decoding the fixed-length vector and outputting the predicted sequence. The use of the models in concert gives the architecture its name of Encoder-Decoder LSTM designed specifically for seq2seq problems.

The Encoder-Decoder LSTM was developed for natural language processing problems where it demonstrated state-of-the-art performance, specifically in the area of text translation called statistical machine translation.

3.1.3 Flask

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. The Flask framework is used in this project along with HTML and CSS to create a user-friendly UI that makes the use of the classification model easy and simple. Flask creates a backend where the model is loaded, and the input is processed. The output of the translator is then displayed.

3.2 Details of hardware and software

3.2.1 Hardware requirements

- Processor - Intel Core i5+
- RAM - 8GB+
- ROM - 1GB+
- GPU - 2GB+

3.2.2 Software requirements

- Windows, MacOS or Linux
- Python –
  - Keras
  - Tinysegmentor
  - Kuromoji
  - Pickle
  - Flask

3.3 Design Details

3.3.1 System Flow/ System Architecture

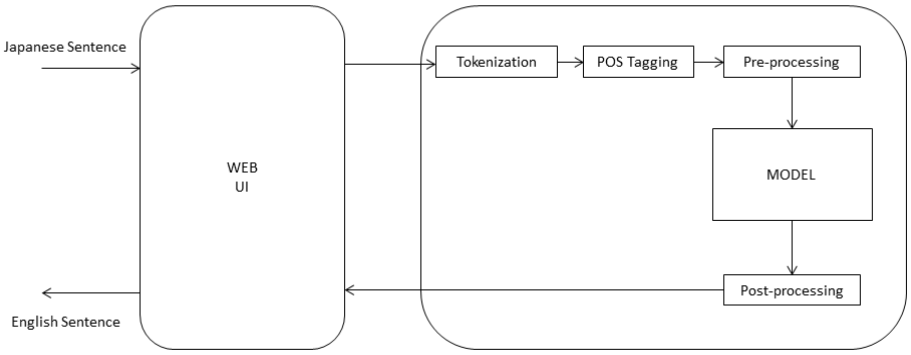

Fig. 3.1 System Architecture

Working of system:

- Input is taken into the system via web interface in Japanese language format.
- Input is sent to server where it is tokenized with help of tokenization techniques for Japanese words and sentences.
- Tokenized sentences are tagged with Parts of speech tagging for each Japanese word.
- A combined vector of sentence along with its tag is produced by the preprocessing module
- The produced vector is delivered to a pretrained model that produces a sequence of English words as output to given vector.
- The output is processed again to form a proper sentence in English and sent to the web UI
- The output is displayed as translated English version of Japanese input sentence.

3.3.2 Detailed Design

3.3.2.1 Use Case

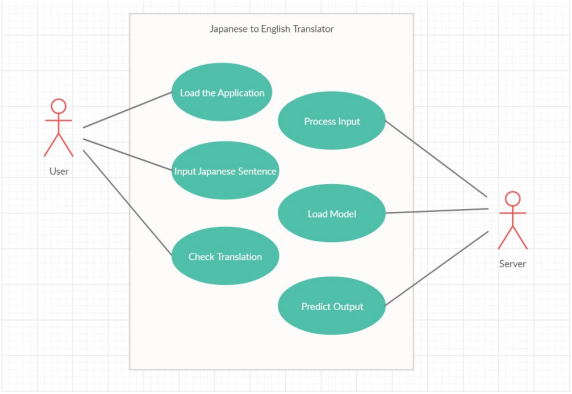

Fig. 3.2 Use Case Diagram

Entities:

- User: The user can access the web interface of the application to present input in Japanese language and receive its accurate translation as output from the system

- Server: The server loads the pretrained model for translation, receives input from web UI, processes given input and predicts the accurate output. This output is displayed on the Web Interface.

3.3.2.2 Sequence Diagram

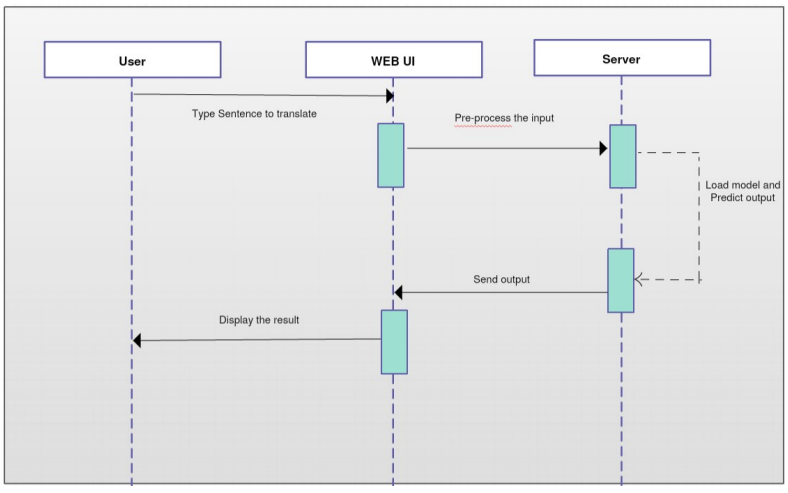

Fig. 3.3 Sequence Diagram

- Input is sent from user to web UI which forwards it to server.
- Server processes the input to produce output and sent it to UI
- UI displayes output that is in turn accessed by the user

3.4 Methodology

3.4.1 Data

Dataset used for the model contains 15000 sentences in Japanese language and their corresponding accurate translations.

Japanese	English
なんてこった！	Oh no!
大丈夫ですよ。	I'm OK.
どうぞお話し下さい。	Go ahead.
誰が来ましたか？	Who came?
もう少し待ってください。	Give me a minute.

Table 3.1 Dataset

The dataset includes sentences ranging from simple to complex in the Japanese language. It is a combined dataset of Kanji and Hiragana sentences of the Japanese vocabulary.

English Word	Japanese (Kanji)	Japanese (Hiragana)
Jump!	飛び越えろ！	とびこえろ！

Table 3.2 Kanji/Hiragana

3.4.2 Pre-processing

Preprocessing is done on the dataset to provide a uniform vector to the model. Each instance in the dataset is pre-processed in 3 steps:

- Tokenization: Japanese sentence is tokenized using a Japanese tokenizer (tinysegmentor) on hence split in different words or segments.
- POS Tagging: With help of Japanese segmentor (Kuromoji) Each segment of tokenized sentence is tagged with a unique Part of speech in relevance to the sentence.
- Vector: A combined sequence is created by joining the segmented word with its POS tag and reconstructing the sentence as a string.

Vector creation for: どうぞお話し下さい。 (Go Ahead)

Segment	POS Tag	Vector
どうぞ	副	どうぞ。副 お話。名 し。動 下さい。動
お話	名	
名し	動	
下さい	動	

Table 3.3 Vector Creation

Hence a unique sequence is created for each sentence in the dataset. The vector and its English translation as label is fed to the neural network for training purposes.

3.4.3 Vocabulary

Generated vectors are taken as sentences and tokenized into words to cumulate a Japanese English vocabulary. The vocabulary is generated by calculating unique Japanese and English words and searching for maximum no of words in a sentence in both vocabularies.

This helps in understanding the depth and impact of the data. The generated vocabulary is used to generate a legitimate word in the language instead of random words.

```
English Vocabulary Size: 3580
English Max Length: 7
Japanese Vocabulary Size: 5625
Japanese Max Length: 19
```

3.4.4 Training Model

The model used to train the dataset is a sequential model that uses LSTM layers. The model is trained with both Japanese and English vocabularies. The model uses ***"adam"*** optimizer and "***categorical_crossentropy"*** loss function.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 19, 256)           1440000
_________________________________________________________________
lstm_1 (LSTM)                (None, 256)               525312
_________________________________________________________________
repeat_vector_1 (RepeatVecto (None, 19, 256)           0
_________________________________________________________________
lstm_2 (LSTM)                (None, 19, 256)           525312
_________________________________________________________________
time_distributed_1 (TimeDist (None, 19, 3580)          920060
=================================================================
Total params: 3,410,684
Trainable params: 3,410,684
Non-trainable params: 0
```

Fig. 3.4 Model

The model is trained over 30 epochs. The model uses 13500 random data as training and remaining 1500 data as validation data for better accuracy.

The model is trained with pre-processed vectors as input and their corresponding English translations at labels of the data. The model gives a validation accuracy of 85.8% during training.

CHAPTER 4

Results and Simulation

4.1 Results

Trained model was tested on the dataset of the dataset after completion of 30 epochs. A Bilingual Evaluation Understudy (BLEU) Score for generated for the data of the model.

A training and testing BLEU score were generated for the model. This score determines the success rate of the models based of a standard number of factors.

TRAINING:

```
japanese sentence = その。連 工場。名 は。助 玩具。名 を。助 製造。名 し。動 て。助 いる。動
target translation = that factory makes toys
predicted translation = the factory makes toys

japanese sentence = 彼。名 は。助 ぐっすり。副 眠っ。動 て。助 い。動 た。助
target translation = he was fast asleep
predicted translation = he was asleep

japanese sentence = 砂糖。名 が。助 ない。形 よ。助
target translation = theres no sugar
predicted translation = we have sugar sugar
```

BLEU SCORE:
BLEU-1: 0.724702
BLEU-2: 0.619604
BLEU-3: 0.560233
BLEU-4: 0.425662

TESTING:

```
japanese sentence = 覚え。動 て。助 い。動 ない。助 ん。名 だ。助
target translation = i dont remember
predicted translation = i dont remember

japanese sentence = トム。名 は。助 卵。名 を。助 1。名 つも。動 買わ。動 なかっ。助 た。助
target translation = tom didnt buy any eggs
predicted translation = tom needs to to visa

japanese sentence = 彼。名 なら。助 できる。動
target translation = he can do it
predicted translation = he can do it
```

BLEU SCORE:
BLEU-1: 0.433501
BLEU-2: 0.299449
BLEU-3: 0.248337
BLEU-4: 0.146549

4.2 Simulation

Web Based UI is developed with a neural network model in the backend.

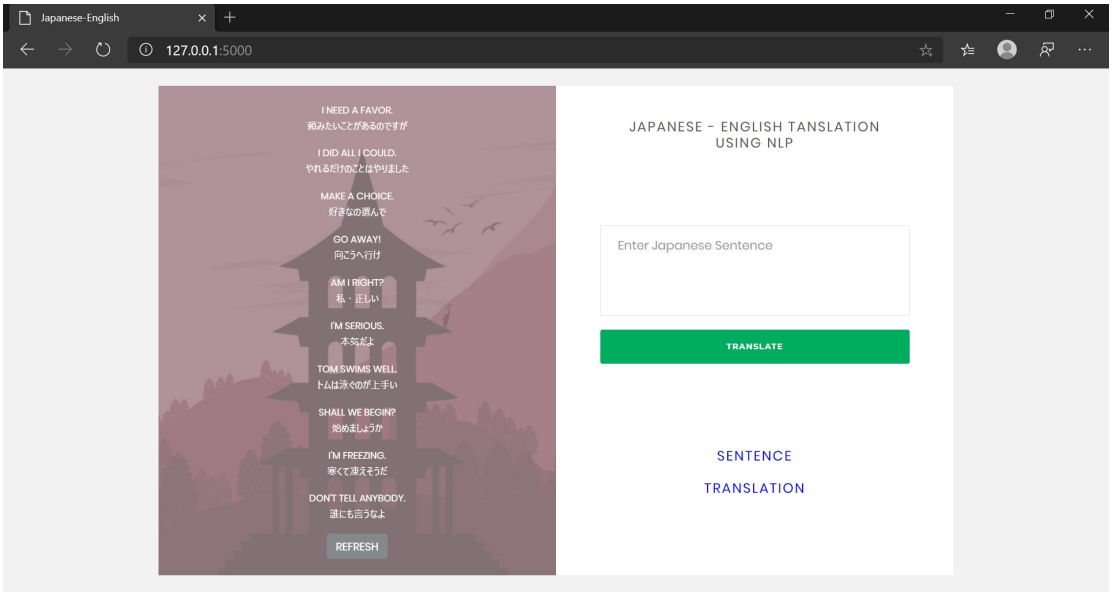

Fig 4.1 Home Page

Input box provided for input of Japanese text and submit button for start of translation. A list of words and translations is provided for checking and trial of the system with a Japanese themed background.

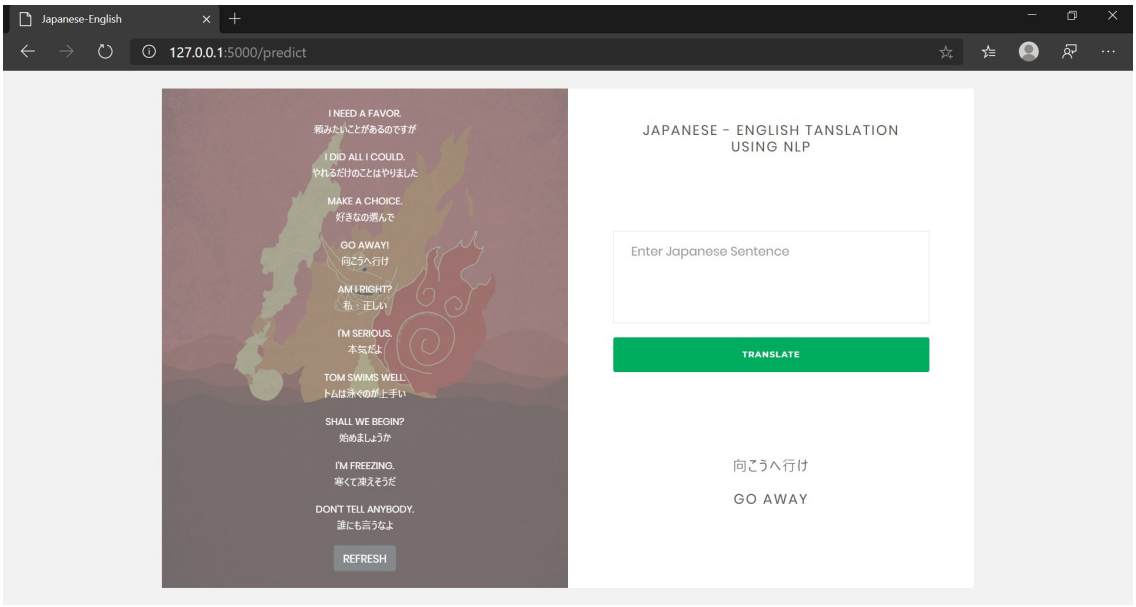

Fig 4.2 Successful Translation

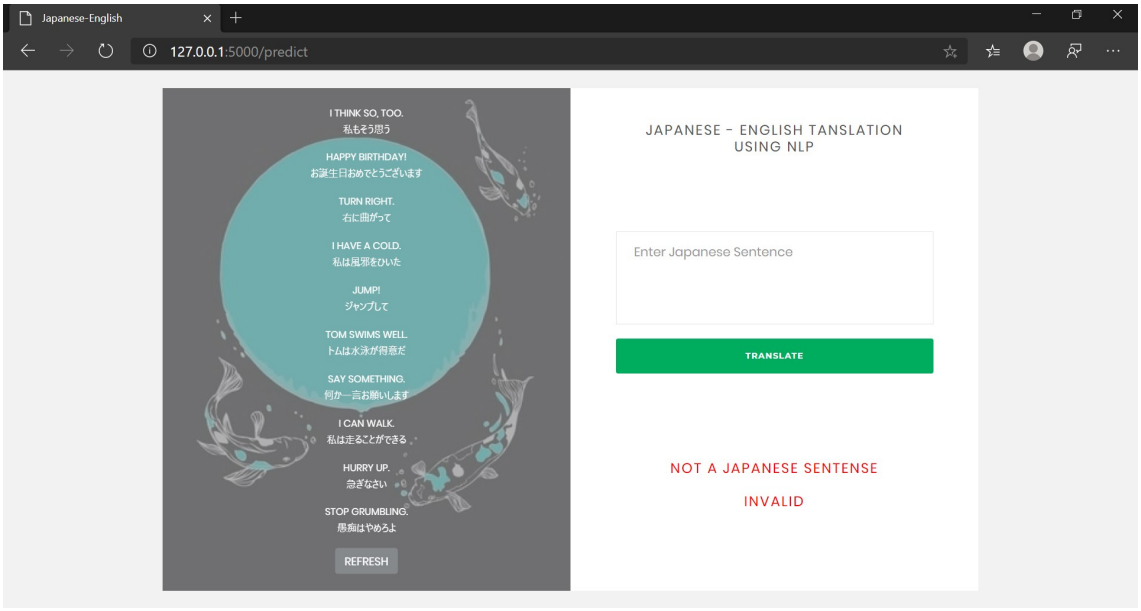

Fig 4.3 Incorrect input provided

CHAPTER 5

CONCLUSION

This Project is a Japanese to English translation system that is developed on an artificial neural network with the help of 15000 Japanese-English labelled data. The project uses segmentation, tokenization and POS tagging to equip the vocabulary of the languages for higher accuracy the Network. The trained model provides a validation accuracy of 85.8% and a Bilingual Evaluation score of 0.72 out of 1.

A web application is deployed with a complete system consisting of server-side components containing a trained model and display of predicted output translation. The web application also provides a testing data list for users to understand the accuracy of the model. Error handling and inappropriate input handling is also done in this project.

This project is useful for better understanding and exchange of knowledge between English and Japanese communities with the help of accurate and quick translation services.

REFERENCES

[1] Todd Law, Hidenori Itoh , Hirohisa Seki, *"A Neural Network- Assisted Japanese- Englis h Machine Translation System", Proceedings* of 1993 International Joint Conference on Neural Networks.

[2] Manabu Sasayama, Fuji Ren, Shingo Kuroiwa, *"SUPER-FUNCTION BASED JAPANESE ENGLISH MACHINE TRANSLATION"*, International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003.

[3] Tohru Shimizu, Yutaka Ashikari, Eiichiro Sumita, ZHANG Jinsong, Satoshi Nakamura, *"NICT/ATR Chinese-Japanese-English Speech-to-Speech Translation System"*, TSINGHUA SCIENCE AND TECHNOLOGY ISSN 1007-0214 18/22 pp540-544 Volume 13, Number 4, August 2008.

[4] https://machinelearningmastery.com/develop-neural-machine-translation-system-keras/

[5] https://machinelearningmastery.com/calculate-bleu-score-for-text-python/