IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 10, October 2021

DOI: 10.17148/IARJSET.2021.81032

An Effective Approach to Perform Language Translation

Rohith Kalvakota[1], Vatsal Vaghasia[2], Suyash Gadiwan[3], Prof. Sonia Relan[4]

Student, Information Technology, Svkm's-Nmims, India[1]

Student, Information Technology, Svkm's-Nmims, India[2]

Student, Information Technology, Svkm's-Nmims, India[3]

Assistant Professor, Department of Information Technology, Svkm's-Nmims, India[4]

Abstract: As we know that English is the foremost Language spoken in the world but not everyone can speak English. Translation is a solution to this problem and is a bridge between different cultures from all over the world. Language can become a major barrier in our day-to-day life, when communicating with your team members while working on a new project or when visiting another country.

Keywords: Translation, Translators, NLP, Language.

I. INTRODUCTION

Translation basically refers to converting a source language into a target language for the purposes of communication. A translator is said to be a system or model which performs translation. Some basic things to take into consideration while performing translation are to include source-language keywords, grammar and syntax into the target-language rendering. Translation is also used in the field of biology where it refers to the process of translating the sequence of messengerRNA (mRNA) molecules to a sequence of amino acids during protein formation. Translators have helped us in many ways, mainly to decipher the sacred texts, which have helped us to learn many different things about life.

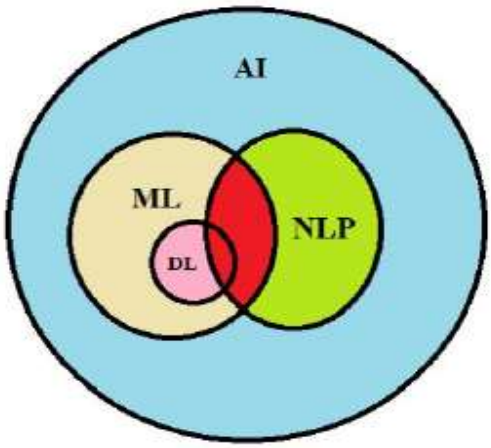

Fig. 1 Domains of AI

For the purpose of translation, NLP (Natural Language Processing) and NLU (Natural Language Understanding) is used. They are both domains of Artificial Intelligence and build add up to sub domains of AI like Machine Learning and Deep Learning. The algorithms which work on text or phrases basically lie in the fields of NLP and NLU. Activities like Translation, Semantic Analysis, Chatbots, Recommendation systems and many more.

II. MODELS USED

A. Symbolic NLP (1950s - Early 1990s)

Symbolic NLP was an idea proposed by John Searle's Chinese room experiment. The chinese room experiment basically states that when given a group of rules, the computer applies Natural Language Understanding to the data it is confronted with.

1) 1950s: An experiment was conducted where translation was fully automated and Russian sentences were converted into English. The authors of the experiment suggested that translation was going to be a problem

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 10, October 2021
DOI: 10.17148/IARJSET.2021.81032

which would be solved in the upcoming 3-5 years. The real progress was very slow due to which funding for these projects were also reduced drastically.

2) 1960s: Some successful translators were also created like SHRDLU, which worked with restricted vocabulary and ELIZA. Without almost no information about human thought or emotion, ELIZA provided a very human-like interaction.
3) 1970s: In this period of time, some programmers began to write "conceptual ontologies", which helps to structure real world information into language that the computer knows and can understand easily.
4) 1980s: In this period of time, there were significant developments in the field of Symbolic NLP methods. This period included research on rule-based decoding. Semantics, reference, etc. and other areas of Natural Language Understanding.

B. Statistical NLP (1990s - 2010s)

Before the 1980s, most NLP systems were based on complex sets of hand-written rules. It was not until the late 1980s that the introduction of machine learning algorithms for language processing brought a revolution in natural language processing. This was mainly due to the steady increase in computational power.

1) 1990s: Much of the early success in Statistical NLP can be credited to IBM Research for their work in machine translation.
2) 2000s: In this period of time, due to the internet the amount of data increased which led to using supervised and unsupervised algorithms for translation. Supervised learning algorithms are much easier to perform than unsupervised algorithms.

C. Neural NLP (Present)

In the 2010s, representation learning and deep neural network-style came into use and are currently being used in fashion. They have been proven to show and achieve state-of-the-art results which results in many language tasks. The use of this particular thing is increasing due to demand in healthcare and medicine, where NLP is being used to analyze notes and text in health records that otherwise would be very hard to get a hand on.

III. NATURAL LANGUAGE PROCESSING

Natural Language Processing is an AI based solution which allows/helps computers to understand, make note of and modify/manipulate human language. [3] It also helps to make use of techniques like audio to text conversion, which can make sure that the systems can understand human language by speech recognition. It can help in implementing voice control of languages on various platforms. NLP is used for various purposes like:

1) Semantic Based Search
2) Product Recommendations
3) Digital Assistants
4) Conversational AI
5) Computational Phenotyping
6) Dictation
7) Credit Scoring
8) Fraud Detection

The challenges of Natural Language Processing are speech recognition, Natural Language Understanding and Natural Language Generation. There are also many methods in Natural Language Processing like Symbolic Natural Language Processing (1950 to Early 1990s), Statistical Natural Language Processing (1990s to 2010s) and nowadays Neural Natural Language Processing is being used. [2] The NLP pipeline consists of seven steps which are:

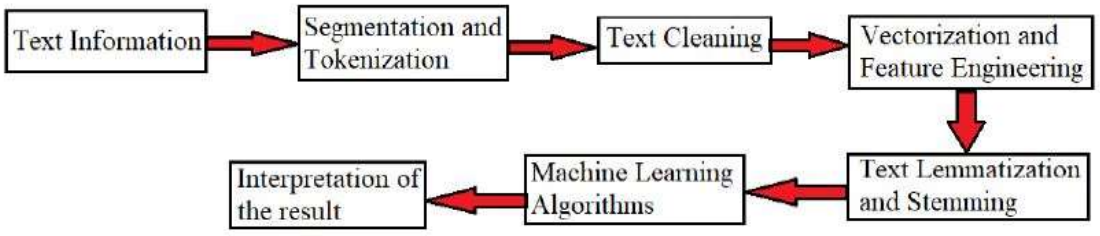

Fig. 2 Pipeline of NLP

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 10, October 2021
DOI: 10.17148/IARJSET.2021.81032

1) Enter the text or sound to be converted to text.
2) Segmentation of text into components using segmentation or tokenization.
3) Text Cleaning.
4) Text is vectorized and feature engineering is performed.
5) Lemmatization and stemming which means to reduce inflections for words.
6) Using machine learning or deep learning methods to train and create models.
7) Using the models on new data and interpreting the results.

IV. SYSTEM ARCHITECTURE

The source language is the one which will be taken as input through the interfaces provided by the application or website. Then the website or application will use the algorithms. The algorithms will use the data-sets which are fed to them at the stage of building the model. After that the algorithms will perform the mapping and translation will be done after which the accuracy will be checked then the algorithm with highest accuracy will be considered for that particular execution. The output then will be displayed as text which can then be played as audio if the user requires it.

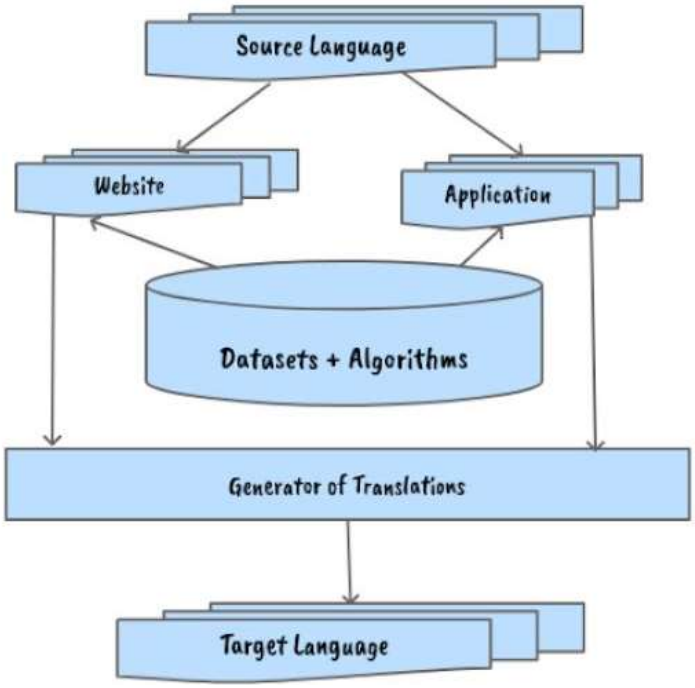

Fig. 3 System Architecture

V. ISSUES SEEN IN EARLIER TRANSLATORS

1) **Number Of Words as Input:** The translation system should be made such that the number of words inputted shouldn't be limited like many translators which have a fixed input size.
2) **Words with multiple meanings:** There are some words that have multiple meanings which can lead to improper translation being done. Example: Scale of fish and scale for weighing.
3) **Compound Words:** Some words have combined words with two different meanings which can lead to the output being translated wrongly. Example: Mass-produced, here mass and produced , although used for the same context as a single word can be taken as two different words which can lead to translation of the two words being done separately which can lead to improper translation.
4) **Missing Terms:** There are some words which might have some meaning in one language but don't exist in another.

VI. TRANSFORMERS

The transformer models were basically made to solve seq2seq tasks alongside handling dependencies of long range with ease. There are two different parts in the architecture of Transformers, one is encoder and the second one is decoder.

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 10, October 2021
DOI: 10.17148/IARJSET.2021.81032

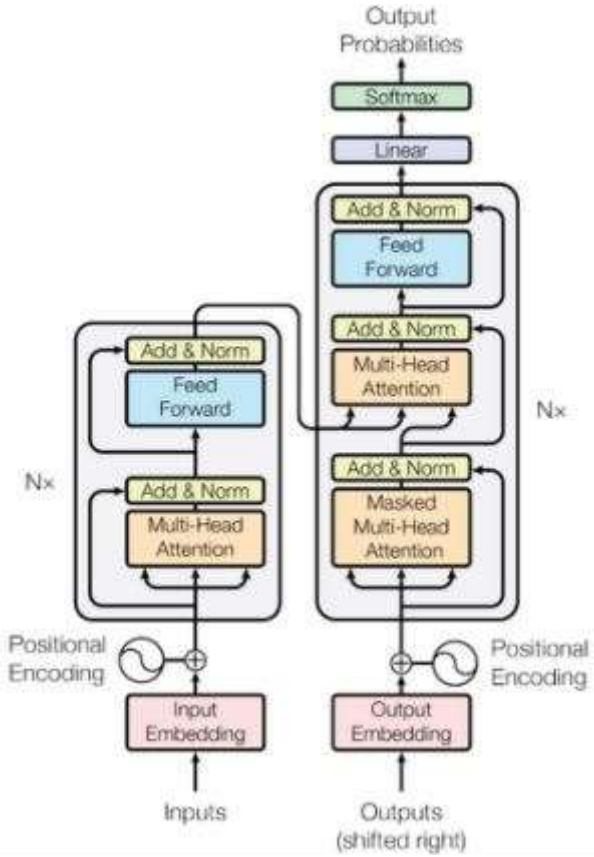

Fig. 4 Architecture of Transformers

The encoder and decoder are multiple encoders and decoders which are stacked on each other to form a transformer. Both of them have a set of equal units where each unit is called a hyperparameter. Self attention is also used by transformers. Self attention is also known as intra-attention, which is a type of attention mechanism relating different variations of a sequence to form multiple sequences.

Vectors used by transformers:

1) Query Vector
2) Key Vector
3) Value Vector

VII. BERT

The algorithm is classified into two steps: Pre-training and fine-tuning. While pre-training, the model is trained on data which is unlabeled over many pre-trained models. As for fine tuning, the BERT model is first initialized with the parameters which are pre-trained, and all these parameters are fine- tuned with the help of labeled data from the downstream tasks. These downstream tasks have separate models for fine-tuning, even though they are initialized with the same pre-trained parameters. [5]

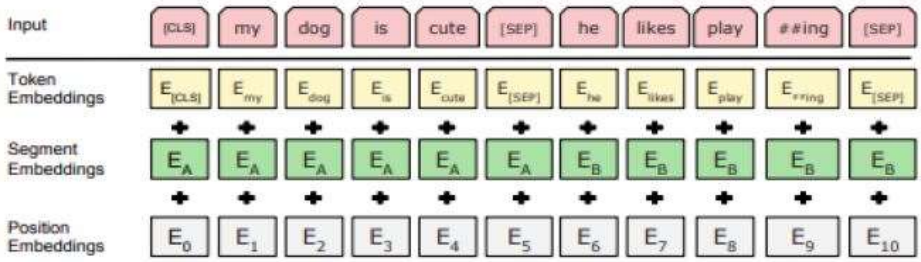

Fig. 5 Architecture of BERT

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 10, October 2021
DOI: 10.17148/IARJSET.2021.81032

It is a multi-layer bidirectional Transformer formed by combining a series of encoders from a Transformer. BERT name roots from this definition, named as Bidirectional Encoder Representation of Transformer.

VIII. RNN

RNN specifically focuses on taking sequences of text as inputs or returning sequences of text as outputs. The network is called recurrent because the hidden layers of the network have a loop in which the output and cell state from each time step becomes inputs to the next time step. [1] Precise Output is produced from previous time steps so that they can be applied to network operations at the current time step.

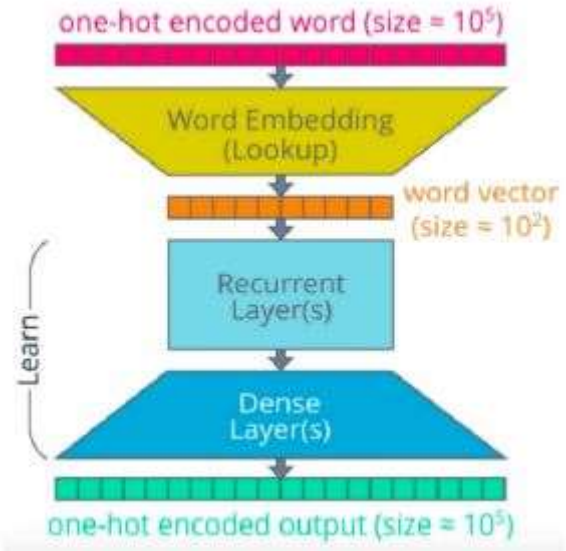

Fig. 6 Architecture of RNN

Layers in RNN:

1) **Inputs:** For every time step, input sequence is fed with one word into the model. Each word is encoded as a unique integer or one-hot encoded vector that maps to the input dataset vocabulary.
2) **Embedding Layers:** This layer is used in the conversion of word to vector. The vector's size depends on the complexity of the vocabulary.
3) **Encoder (Recurrent Layer):** Input is taken from word vectors which were part of previous time steps and is applied to the current word vector.
4) **Decoder (Dense Layer):** Decode the input which is encoded into a precise translation sequence.
5) **Outputs:** Output is generated as a sequence of integers or one-hot encoded vectors and then can be mapped to the desired language dataset.

IX. LONG SHORT-TERM MEMORY NETWORKS

Long Short-Term Memory Networks or LSTMs are basically an extended version of RNNs. They have capabilities to perform long term dependencies. They work on a lot of various problems due to which they are being used nowadays. [4] LSTMs have been designed explicitly to remember things for the long term so it is practically their default behavior and not something that they require or struggle to learn. [7] It is basically a chain of repeating modules of neural networks.

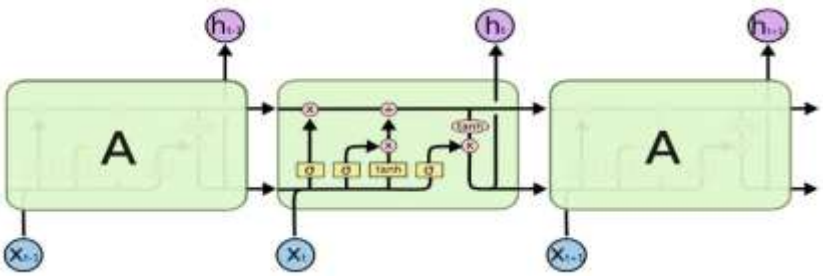

Fig. 7 Architecture of LSTM

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 10, October 2021

DOI: 10.17148/IARJSET.2021.81032

Working of LSTMs:

1) We need to decide the information we are going to throw away from the cell state.
2) Next, we need to decide what information we are going to store in the cell state.
3) Then the final step will be to output the data we want to which will depend on earlier neurons and will be filtered. [8]

X. DATA DICTIONARY

A typical dataset will contain mostly two parameters, source language and target language. These parameters play an important role in the language translation system. Then, each of these datasets will be sent forward to train the algorithm.

Sr. No.	Data Name	Data Type
1.	Source Language	string
2.	Target Language	string

Fig. 8 Data Representation

XI. IMPLEMENTATION

The system is basically used for translation purposes so the user needs to input text in the form of text or speech alongside choosing the input language and then can translate to their desired language. After translation is done the user can then view the output in the form of text or speech.

1) **Input will be taken by the System:** A GUI will be provided to the user where a textarea or textbox will be present. The user will be allowed to select the language in which input will be done. The user can give input to the system in the form of text or speech.
2) **Translation is done by the System:** User will give input to the system in the form of text/speech. For text, it can be directly used but in the case of speech it will rst be detected then converted into text which can then be used for the models. Using the algorithms, translation of text will be done where after translation all accuracies of the models will be checked and the model with the highest accuracy will be considered for translation. The model considered for translation will then translate the input text to text in desired language.
3) **Conversion of Translated text to Speech:** After the translation is done, it will be displayed in a textbox. If the user requires they can then convert it into the form of speech.

XII. FUTURE SCOPE

Translators and the concept of language translation has come a long way since it was created. In the future, automated translation will be a concept/ technology where a lot of development would be done. [6] Automated translation will help to perform multilingual translation in an efficient way which will help in performing SEO (Search Engine Optimization) and help access new platforms. Translators will also start to use more and more AI technologies to get better results and reach a broader audience. Increase in the number of cloud environments will also be a factor in the development of Translation.

XIII. CONCLUSION

Translation and translators are used by a lot of people when they travel around the world, need to communicate with people of other cultures, etc. The translators were first built as Symbolic models then soon came statistical models and the latest models used now are Neural models. [9] The increase in NLP models will only help to make translation more efficient so that people can use it effectively. NLP is a concept which can be performed using open source languages which will help people modify and make contributions to the concept of translation. With the concept of 5G, which guarantees faster network connection it will only help people to use these translation systems for their own good.

XIV. REFERENCES

[1]. L. Yao and Y. Guan,"An improved lstm structure for natural language processing," in 2018 IEEE International Conference of Safety Produce Informatization (IICSPI), 2018, pp. 565 - 569.
[2]. S. Gogineni, G. Suryanarayana, and S. K. Surendran, "An effective neural machine translation for english to hindi language," in 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 209 - 214.
[3]. G. Tiwari, A. Sharma, A. Sahotra, and R. Kapoor, "English-hindi neural machine translation-lstm seq2seq and convs2s," in 2020 International Conference on Communication and Signal Processing (ICCSP), 2020, pp. 871 - 875.
[4]. S. P. Singh, H. Darbari, A. Kumar, S. Jain, and A. Lohan, "Overview of neural machine translation for english-hindi," vol. 1, pp. 1 - 4, 2019.

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 10, October 2021

DOI: 10.17148/IARJSET.2021.81032

[5]. Jacob Devlin and Ming-Wei Chang and Kenton Lee and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." CoRR, abs/1810.04805.

[6]. Y. Fan, F. Tian, Y. Xia, T. Qin, X. Li and T. Liu, "Searching Better Architectures for Neural Machine Translation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1574-1585, 2020, doi: 10.1109/TASLP.2020.2995270.

[7]. L. Yao and Y. Guan, "An Improved LSTM Structure for Natural Language Processing," 2018 IEEE International Conference of Safety Produce Informatization (IICSPI), 2018, pp. 565-569, doi: 10.1109/IICSPI.2018.8690387.

[8]. G. Tiwari, A. Sharma, A. Sahotra and R. Kapoor, "English-Hindi Neural Machine Translation-LSTM Seq2Seq and ConvS2S," 2020 International Conference on Communication and Signal Processing (ICCSP), 2020, pp. 871-875, doi: 10.1109/ICCSP48568.2020.9182117.

[9]. S. Satpathy, S. P. Mishra and A. K. Nayak, "Analysis of Learning Approaches for Machine Translation Systems," 2019 International Conference on Applied Machine Learning (ICAML), 2019, pp. 160-164, doi: 10.1109/ICAML48257.2019.00038.

