Deep Learning for Natural Language Processing in Bilingual Language

Pooja G[1], Gudarada Vandana[2], Usha AH[3], Muskaan[4], Prof.Pathanjali C[5]
[1,2,3,4] *Student, Department of information Science and Engineering, Global Academy of Technology, Bangalore*
[5]*Assistant Professor, Department of Information Science and Engineering, Global Academy of Technology, Bangalore*

***Abstract*- Existing and emerging technologies aid in the resolution of real-time problems without the need for manual intervention. Code-Switching allows people to socialize with others, learn new languages. This paper discusses various approaches to code-switching, including recurrent neural networks (RNN), support vector machines (SVM), bidirectional encoder representations from Transformers (BERT), and others. As a result, an appropriate method must be chosen to achieve maximum accuracy with solutions to all existing problems with minimal enhancements.**

***Index Terms*—Code-Switching,Speech recognition, Neural Machine Translation, BERT model.**

I. INTRODUCTION

Language plays a vital role in life as a medium for human communication. Interacting with others in society is one of them. Speakers in multilingual societies frequently transition or mix between two or more languages in their daily lives. Code-switching is the term for this phenomenon. Multilingual speakers switch between their common language in writing and spoken conversations, a practice known as code-switching. The approach used by bilinguals and multilingual people who often mix multiple languages in a speech with a small change of interlocutor or topic is known as code-switching (CS).

The units and places of the switches might range from one-word switches to entire phrases. Such events present issues for spoken language technologies, such as ASR,because the systems must be capable to handle input in a bilingual or multilingual environment. Various studies developed a code-switching automatic speech recognition (ASR) for a variety of language pairings. However, the most typical goal of designing code-switching ASR is to convert code-switching speech utterances to CS-text sentences within an individual.

In this paper, however, the situational context that occurs during dialogues among code-switching and non-code-switching speakers is addressed, as well as providing support to monolinguals who desire to translate code-switching speaker. The number of international visitors and inhabitants to India is steadily rising, for reasons such as tourism and education. Bilingual speakers are on the rise, according to the survey. People's communication styles are influenced by these shifts. As a result, English-Hindi code-switching is becoming more common, and Some people use it on a regular basis. This phenomenon presents a challenge for spoken language technology, such as ASR, which must handle input in a multilingual environment. As a result, It isimportant to formulate a system that recognizes code-switching speech and transforms it into monolingual texts.

II. LITERATURE REVIEW

Janhavi R. Chaudhary et.al[1] presented a bidirectional RNN encoder for the [RNN] Recurrent neural network model. The forward and the backward RNN make up a bidirectional RNN encoder. Tanaka corpus has been used. It is a freely accessibledatabase. They involve extensive matrix multiplication, parallel processing, and a huge number of calculations during the training phase, deep learning (NeuralMachine Translation) applications require high computations. In comparison to the CPU, thegraphics processingunit (GPU) is an excellent alternative for parallel processing and quick computation. They attempted to

get an accurate translation from Japanese to English using this method. To train the algorithm, they used a tiny set of data. They also use this architecture with an efficient processing unit to handle enormous amounts of data.Premjith.B et.al[2] presented that Neural machine translation NMT was created with encoder and decoder LSTM networks and bi-directional RNN in this paper. The translation is done for certain Indian languages, and diverse language corpora are obtained and then used for the modelling translation system. Both automation (BLEU) and manual (adequacy, fluency, and overall ranking) evaluations are performed. The encoder network was designed using the LSTM and bi-RNN algorithms, and each hidden layer was trained with 200 to 500 hidden layer units. As an outcome, the proposed approach of bi-RNN provided a good translation of various Indian languages.The NLP (natural language processing) technique proposed by Manjula S et.al[3] makes it easier to identify languages in text documents. They used the parallel corpus as input values for the language identification process, and different steps in the processing were used, such as data acquisition, data pre-processing, tokenization, feature extraction, and SVM and Bayesian classification. This proposed methodology was tested on the European Parliament Proceedings Parallel Corpus. They also described different feature extraction approaches, as well as a novel hybrid method for recognizing diverse languages, in which the retrieved features are used to categorize the languages into classes. Sandeep Saini and Vineet Shaula et.al[4] proposed the recurrent neural network (RNN) model with LSTM (long short-term memory). For a long time, statistical phrase-based Machine translation methods have struggled with accuracy and the need for big data sets. In this paper, the idea of using a shallow RNN and LSTM based Neural Machine translator to solve the challenge of MT (machine translation). For experiment, they utilized a tiny dataset and a limited number of layers. The results reveal that NMT has a significant number of layers in the encoder-decoder and can deliver substantially better results for a larger dataset. NMT-based MT considerably outperforms current SMT and PBMT technologies.Omkar Dhariya et.al[5] presented a hybrid technique that combines SMT (statistical machine translation), EBMT (example-based machine translation), and RBMT(rule-basemachine translation). Segmentation, translation, POS tagging, and rearrangement were used to implement the proposed method. However, the proposed approach only works with short sentences and different tenses. The hybrid technique improved the machine translation system's quality, and it also outperformed numerous baseline translation methods based on the SMT, EBMT and RBMT approaches individually. Gualberto Guzman et.al[6] test their hypothesis on different Spanish-English corpora. The results show that their model can distinguish between some corpora based on whether or not they have a dominant Machine Learning, but that the corpora extend a variety of mixing types. The model's performance on the Miami and S7 datasets shows that current metrics are inadequate to predict agreement also when corpora have characteristics that indicate they are primarily insertional. The uncertainty in the model predictions leads us to believe that within the known typology of alter national and insertional mixing, there is a continuum of mixing types. Finally, a deeper syntactic examination of the nature of mixing types, as well as the grammatical structure of Miami and S7 datasets, in particular, improves the model's performance.Deepanshu Vijay et.al[7] suggested a mechanism for detecting irony in Hindi-English code-mixed social media text, they postulate a supervised classification technique for detecting irony in text using various characters, words, and structural features. They assessed datasets based on lexical and semantic features of each word, as well as word occurrences, which were used to train the Decision tree. They created a Hindi-English code-mixed corpus from tweets in JSON format, and the corpus has been annotated. To train a supervised model, a machine learning model was presented that involved pre-processing and classifying features using feature vectors. An experiment with two different classifiers, SVM (Support vector machine) and random forest classifier, was carried out, with the Support vector machine outperforming the random forest classifier.Khyathi Raghavi Chandu et.al[8] proposed a language-informed model for downstream NLP applications such as speech recognition and machine translation. Implemented various language models derived from multi-layered LSTM architectures. The primary purpose of this paper is to investigate the role of language

information in modelling of code-switching text. They suggested different models for training code-switched text using RNN of 3 layered stacked encoder decoders with LSTM and discovered that Language Aware Encoding and Decoding with AWD-LSTM provides the minimum perplexity.Sahoko Nakayama et.al[9] devised a system which recognizes code-switch speech and converts it into monolingual text. They worked on a variety of approaches like a cascade of ASR and neural machine translation, a cascade of ASR and deep bidirectional language model, and ASR that directly outputs monolingual text from CS speech, as well as multitask learning. On a Japanese-English CS to English unilingual task, they evaluate four approaches. They built artificial and natural CS corpora for code-switching corpora. When they compared all four methods, they discovered that ASR and BERT performed better.M. K. Vathsala and Holi Ganga et.al[10] focused on two types of models RNN-based Language Models (which can retain long-term dependencies) and LSTM-based Language Models (which can estimate the conditional probability between the input and output sequences). determined that RNN-LSTM is more precise than traditional (SMT) statistical machine translation models and that Bi-Lingual Evaluation Understudy (BLEU) is one of the most effective methods for assessing machine translation quality. However, because a portion of the translation is not present in the referenced data, this parameter does not influence the accuracy of the translated data. The research aimed to analyze social network data in order to transcode and translate it into English using the models mentioned above and used TensorFlow to express the LSTM. A Python script is used to pull data from various fields and a regular expression is used to clean it. The transliteration and translation of social media data allow for content in the languages that the majority of the people understands.Yinhan Liu et.al[11] proposed RoBERTa, the best model that achieves state-of-art results on GLUE, Squad, and RACE. The pretraining procedure BERT model has been modified and enhanced and is now referred to as RoBERTa (Robustly optimized BERT approach). A model of masked language and (NSP) Next sentence prediction were used to train this RoBERTa. They discovered that RoBERTa outperforms BERT significantly.Guillaume Lample and Alexis Conneau et.al[12] proposed two methods for learning cross-lingual language models (XLMs): Causal Language Modelling (CLM) and Masked Language Modelling (MLM), which are unsupervised training objective that required only monolingual corpora and supervise learning that supports parallel data with new cross-lingual language model objectives. They achieved a new state of the art in Bi-Lingual Evaluation Understudy (BLEU) on Romanian-English, which equates to a gain of greater than 4 BLEU points. They also show how cross-lingual language models can be utilized to enhance uncertainty of a Nepali-language model while also generating unsupervised cross-lingual word embeddings. The Translating Language Modelling (TLM) aim, which supports parallel data and enhances pre-training of multilingual language models, contributes to their work. They discovered that TLM naturally expanded the MLM approaches by employing a series of parallel sentence instead of consecutive words and that using TLM in conjunction with MLM provided a significant gain.Kartikay Khandelwal, Alexis Connemp et.al[13] The article demonstrates that simulating multilingual language models in the scale improves performance on a variety of tasks. They use more than two terabytes of routinely filtered data to create a masked language model based on a transformer per hundred languages. This model, known as XLMR, is suspected of outperforming a number of cross-lingual molds, including + 14.6 % accuracy on XNLI, + 13% of F1 average points on MLQA, and + 2.4s on the NER. XLMR is designed exclusively for Low Resource languages, enhancing XNLI accuracy by 15.7% for Swiahili and 11.4 % for Urdu on XLM models. They also have a detailed experimental analysis needed to achieve these benefits, including compromises between positive transfer, dilution, and high resource language performance. And low on the scale. Finally, they show how XLMR can model multilingual without sacrificing per-language performance by using robust monolingual models on the Common Language Understanding Assessment (GLUE) and benchmarks Natural Multilingual Inference (XNLI).A transformer based pre-trained language model was proposed by Wietse de Vries et.al[14]. BERT improved the performance of many NLP tasks, and they developed, created and analyzed a monolingual Dutch BERT model known as BERTje

using same parameters. BERTje is based on a vast and diverse dataset of 2.4 billion tokens, as compared to the BERT model (Transformers Bidirectional Multilingual Coding Representation), which contains Dutch but is based on text from Wikipedia, newspaper. Because BERT's core model is trained in English, its success in NLP tasks is confined to the English language. Theytrained language specific models with the same BERT architecture for other languages, or they used an existing multilingual BERT model. On Dutch NLP tasks, multilingual BERT model outperformed the single language model, indicating that the single language model should be preferred.BERT is a novel language representation paradigm by Jacob Devlin et.[15]. Many NLP tasks, such as sentence level tasks and token sentence tasks, have been found to benefit from language model pre-training. They used a feature-based technique, such as (Elmo) Embeddings from Language Models, and a fine-tuning method, like the Generative Pre-trained Transformer, to apply pre-trained language representation to downstream tasks. The significant breakthrough is that these discoveries can now be used in deep bidirectional architectures, allowing a single pretrained model to solve wide range of NLP applications.

III.LITERATURE SUMMARY

The summary of the literature is depicted in Table1:

Table 1:Summary of literature Survey

Sl. No	Author	Methodology / Algorithms used	Contribution
1	Janhavi R. Chaudhary and Prof. Ankit C. Patel	(DNN) deep neural network and (RNN) Recurrent neural networks	They attempted to get an accurate translation from Japanese to English using DNN method.
2	B. Premjith	(NMT)Neural machine translation, recurrent neural networks (RNN), (LSTM) long short-term memory, gated recurrent unit (GRU), or bidirectional RNN	The proposed approach of bi-RNN provided a good translation of various Indian languages.
3	Manjula S and Dr. Shivamurthaiah M	Data acquisition, Data pre-processing, feature extraction and classification	Method to Classification of data based on languages from a text document
4	Sandeep Saini	Neural Machine Translation (NMT), Recurrent Neural Networks (RNN)	capability to handle complex sentences with a BLEU score of 38.20
5	Omkar Dhariya	(SMT)statistical machine translation,(EBMT)example-based MT and (RBMT) rule-based MT and(HMT)Hybrid approach for machine translation.	The hybrid technique improves the quality of a machine translation system by increasing fluency, accuracy, and grammatical precision.
6	Barbara E. Bullock	Matrix Language Frame model (MLF)	With an accuracy of 69.3%, forecast the concurrence of multiple machine learning approaches.
7	Deepanshu Vijay	Pre-processing, classification features, (SVM)Support Vector Machines with radial basis function kernel and (RFC) Random Forest Classifier	SVM performs better than Random Forest classifier and gives a highest F1 score of 0.77 when all features are used that achieved an accuracy of 95.76%
8	Khyathi Raghavi Chandu	AWD-LSTM Model, State-of-Art language models	On test sets, Language Aware Encoding and Decoding with the AWD-LSTM has the lowest perplexity score of 19.52.
9	Sahoko Nakayama	Casecade ASR and (NMT) neural machine translation, a cascade of ASR and BERT	ASR +BERT may outperform ASR+NMT, hence BERT is the more powerful model with better performance.
10	M. K. Vathsala and Ganga Holi	RNN-based Language Model and (LSTM Model)long short-term memory	Learning rate of 0.001
11	Chia-Hsuan Chang and Yinhan Liu	(RoBERTa)robustlyoptimizedBERTpretrainingapproachand BERT (Bidirectional Encoder Representations from Transformers)	RoBERTa accuracy of 83.2% when compared with BERT
12	Guillaume Lample	XML Cross-lingual Language model pretraining	accuracy from 83.2% to 85%

	and Alexis Conneau		
13	Alexis Conneau	XLM approach,Masked Language Model(MLM), BERT	Average of 90.2 to 92.8
14	Wietse de Vries	Bidirectional Encoder Representations from Transformers (BERTje)	BERTje beats UDPipe 2.0's accuracy sc ore of 95.98 %.
15	Jacob Devlin	(BERT) Bidirectional Encoder Representations from Transformers, (BiLSTM)Bidirectional long short-term memory	Test F1 score of 92.8

IV. METHODOLOGY

Machine Translation with Deep Learning that is, Neural Machine Translation is a new approach to machine translation that has recently been proposed. The term "machine translation" refers to a translation from one language to another that does not require human intervention. It is also known as automated translation. To modify the dynamics of Representation learning, today's advanced deep neural networks combine algorithms, vast data, and the computational power of GPUs. Many big data challenges, such as computer vision, speech recognition, and (NLP) natural language processing is solved with deep learning. Natural Language Processing is the most advanced version of currently established ASR technology. This variation of ASR gets the closest to permitting true conversation between people and machine intelligence.

NMT is a recently developed approach for automatic translation that makes use of Deep Neural Networks. In the translation of a number of language pairs, NMT has already shown promising results. In contrast to SMT, which requires independently trained subcomponents for translation, NMT trains with a single big neural network. The encoder consumes the input sentences to form a vector representation, which the decoder uses to generate the target language words.

RNN (Recurrent neural network), LSTM (long short-term memory), GRU (gated recurrent unit) or bidirectional RNN are the most common options for RNN for both encoder and decoder networks. Recurrent neural networks (RNNs) are employed in applications like language modelling because the input can flow in either direction. For this purpose, long short-term memory is very useful. NMT is a new technique for MT that employs a specific Neural Network framework known as Encoder-Decoder Architecture and Bidirectional Encoder Representations. BERT (Bidirectional Encoder Representations from Transformers) is a machine learning method based on transformers that is used for NLP (natural language processing) pre-training.

V. APPLICATION

In a bilingual language, code-switching can motivate people to convey meaning accurately, which should be understandable to the listener. This helps people express solidarity with a social group, define themselves, engage in social interactions, and discuss a specific topic.

Children will not learn academic subject material if they do not grasp the language of instruction, therefore the bilingualism model, which explains how concepts learned in one language can be transferred to another, helps children understand academic subject material thoroughly. When traveling or relocating to a new country, the process can be intimidating, frightening, and exciting all at the same time. In any case, the native language spoken in a certain nation may differ from your home tongue, making communication difficult, and sometimes people will be unable to experience the delights of travel owing to weak language skills. This is the motivation that allows people to quickly learn and speak a foreign language.

Language barrier will be eliminated hence helps in all common application. It also assists in industries such as agriculture, tourism, banking, and movies. Understanding a multilingual language allows a person to engage with a variety of individuals and gain a better understanding of another culture's intricacies. People may have more opportunities to meet friends, discover new hobbies, and have a better understanding of code-switch languages.

VI. CONCLUSION

Thispapersummarizeson the research done in the areas of speech-to-text conversion and machine translation models. As described in the articles, this research will aid in the analysis of models that will produce the required output, such as automatic voice

recognition (ASR), neural machine translations (NMT) such as BERT, and RNN over CNN. According to the review, the NMT model performs well when compared to other models. Machine translation is a socially relevant project that will facilitate communication while also providing educational, cultural, and other benefits. This work can be used to research machine translation strategies that aid in the development of effective translators for communication and educational purposes.

REFERENCE

[1] Janhavi R Chaudhary, Ankit C Patel. "Bilingual machine translation using RNN based deep Learning". Int J Sci Res Sci Eng. Technol 4 (4), 1480-1484, 2018

[2] Premjith, Bhavukam et al. "Neural Machine Translation System for English to Indian Language Translation Using MTIL Parallel Corpus." Journal of Intelligent Systems 28 (2019): 387 – 398.

[3] Mallaiah, Shivamurthaiah & Shamarao, Manjula. "Identification of Languages from The Text Document Using Natural Language Processing System".(2021)(TURCOMAT). 12. 2465-2472.

[4] S. Saini and V. Sahula, "Neural Machine Translation for English to Hindi,"Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), 2018.

[5] O. Dhariya, S. Malviya and U. S. Tiwary, "A hybrid approach for Hindi-English machine translation," International Conference on Information Networking (ICOIN), *2017, pp. 389-394.*

[6] Bullock, Barbara & Guzmán, Wally & Serigos, Jacqueline & Sharath, Vivek & Toribio, Almeida. "Predicting the presence of a Matrix Language in code-switching". 68-75. 10.18653/ v1/W18-3208.

[7] Deepanshu Vijay, Aditya Bohra, Vinay Singh, Syed S. Akhtar, and Manish Shrivastava."A Dataset for Detecting Irony in Hindi-English Code-Mixed social media Text" (2018)

[8] Chandu, Khyathi & Manzini, Thomas & Singh, Sumeet & Black, Alan. "Language Informed Modeling of Code-Switched Text". 92-97. 10.18653/v1/W18-3211.

[9] S. Nakayama, T. Kano, A. Tjandra, S. Sakti and S. Nakamura, "Recognition and translation of code-switching speech utterances," 2019.

[10] Vathsala, M.K., Holi, G. "RNN based machine translation and transliteration for Twitter data" . Int J Speech Technol 23, 499–504 (2020).

[11] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov.V. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". ArXiv, abs/1907.11692.

[12] Guillaume Lample, Alexis Conneau. "Cross-lingual Language Model Pretraining"(2019)

[13] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov." Unsupervised Cross-lingual Representation Learning at Scale"(2020)

[14] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, Malvina Nissim. "Berea: A Dutch BERT Model"(2019)

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"(2019)

[16] Dalmia, Siddharth & Liu, Yuzong & Ronanki, Srikanth & Kirchhoff, Katrin. Transformer-Transducers for Code-Switched Speech Recognition.(2021).58595863.10.1109/ICASSP39728.2021.9413562.

[17] X. Zhou, E. Yilmaz, Y. Long, Y. Li, and H. Li, "Multi-Encoder-Decoder Transformer for Code-Switching Speech Recognition," in Proc. Interspeech, 2020.

[18] E. van der Westhuizen, T. Niesler, "Automatic Speech Recognition of Code-Switched Speech"(2016) 121–127.

[19] Emre Yilmaz; Henk van den Heuvel; David van Leeuwen "IEEE spoken language technology workshop"(2016).

[20] Heike Adel, Ngoc Thang Vu, Franziska Kraus, TimSchlippe, Haizhou Li, and Tanja Schultz. "Recurrent neural network language modeling forcode-switching".(ICASSP), 2013 IEEE 8411–8415.