RESEARCH ARTICLE OPEN ACCESS

Kanglish to English Machine Translation using Pretrained Multilingual Transformers

SANDHYA S*

*(MSc Data Science, NMKRV College for Women, Bengaluru-560011
Email: sandhyashanmugam459@gmail.com)

--************************----------------------------------

Abstract:

Kannada is one of the Dravidian languages which is commonly spoken in Karnataka state of India. Kanglish is a code-mixed language which comprises of both Kannada and English languages. Over the course of time with advancement in technology and social media development, it is observed that many people who has a native language different from English tend to mix up their mother tongue with English in day-to-day life. Code-mixed languages play a vital role as most of the text data that are generated on online platforms are posted by people who use such code-mixed languages. Also, such mixed languages are used to browse information on search engines. However, machine translation for Kanglish to English is not explored widely. With the recent advancement in pretrained language transformers, our aim is to achieve the Machine Translation for Kanglish to English using pretrained language models and acquire the best accuracy.

Keywords — **Code-mixed language, Machine translation, Kanglish.**

--************************----------------------------------

I. INTRODUCTION

We say a person is bilingual if he/she can understand and communicate in 2 different languages. A person can develop to be a bilingual naturally when he/she learns a different language apart from their mother tongue. Many tend to mix-up their native language with English in day-to-day life without even noticing it. In a sense it has been implemented as a part of our interaction with other individuals with the growing sense of community and technology.

Since the number of users are also increased on social platform, the text data that are being generated in the form of tweets, comments, or even user questions on web are in the form of bilingual. To fully understand the combination of 2 different languages is bit challenging as the semantic/grammar rules of each language differs.

And it is to be noted that the machine translation of code-mixed languages lacks exploration for Indian languages. And since India has diverse languages, it is an added part of the challenge. But there are pre-trained multilingual transformers that are available for the Indian language in which we can fine-tune to perform our desired monolingual machine translation of Kanglish.

The term Code-Mixing may sound new, but this has been in practice among us even without our own knowledge. In the present world where people live in various parts of the society apart from their native, it is essential to pick up a new language for easy communication. And since the human brain can understand more than one language, it is quite common to mix-up these languages in daily life. And with the recent development in the technology, where people connect to others all over the world, the text data that gets generated online is tremendous. And these data need not be monolingual all the times. It can be of bilingual which might be challenging for processing.

Many applications of Data Science rely upon the text data that is being easily available online which are in turn generated by real humans. Various applications like Sentiment analysis, Prediction of success of a movie or an election and many more such operations are done by analysing these texts. In such requirements, it is essential to preprocess the data in initial stages before the development of the models.

Natural Language Processing (NLP) have grown in recent times which helps in understanding and processing of human language. Though in initial stages it was widely used for foreign languages like English, German etc., lately it has been implemented to Indian language with the advancement in the generation of diverse data.

Machine Translation also been very useful and is being implemented widely in recent days as it is supporting different languages including Indian languages. Machine Translation helps with translating one language script to the other language. Also, it helps in transliterating a language in other language. Though machine translation is helping in understanding the language and performing the user required operations by understanding the rules and semantics of each individual language, it is quite challenging to deal with the Code-Mixed language, where the script formed with the combination of more than one language.

One such Code-Mixed language is the Kanglish (Kannada + English). Though Dealing with Code-Mixed language have been implemented and are in use for foreign languages, it is not explored for Indian languages. The key insight of this project is to build a model which can translate this code-mixed language Kanglish to a monolingual language (English in our case), utilizing the different seq2seq pre-trained multilingual transformer-based models and achieve the best accuracy.

II. Related Work

In this section, we will discuss about the previous works that were done in the field of Code-Mixed Machine Translation by several researchers and data scientists. We can understand how the work has been progressing in this area since the beginning and the research gaps that needs to be filled.

Code-Mixing arises when a person make use of 2 or more languages in the context of the same communication. It is quite common in the present society with the development in social media platforms and multilingual society.

Many workshops (Diab et al.,2014 [7], 2016 [6][8]; Aguilar et al., 2018b [10]) have been conducted in notable conferences to advance in the code-mixed machine translation field. But these workshops were diversely dealing with related tasks like language identification (Solorio et al., 2014 [7][9]; Molina et al., 2016 [8]), NER (Aguilar et al., 2018a [10]; Rao and Devi, 2016 [11]), PoS tagging (Jamatia et al., 2018 [13]), sentiment analysis, and question answering (Chandu et al., 2018 [12]).

Though these workshops were stepping stones to advance and explore code-mixed translation, the inadequacy of datasets were of a major hindrance.

Ganesh Jawahar, El Moatez Billah Nagoudi, Muhammad Abdul-Mageed and Laks V.S. Lakshmanan published a paper on "Exploring Text-to-Text Transformers for English to Hinglish Machine Translation with Synthetic Code-Mixing". In this paper they used the Transformer-based encoder-decoder models like mT5 and mBART on models to translate monolingual English text into Hinglish code-mixed language. Added to which they also proposed a method for generating code-mixed texts from bilingual representations. And the result showed that their mT5 model gives better performance with 12.67 BLEU score.

Satyam Dutta, Himanshi Agarwal and Pradeep Kumar Roy from Indian Institute of Information Technology, Surat, Gujarat, India published a paper on "Sentiment Analysis on Multilingual Code-Mixed Kannada Language" which presents a model that helps in sentiment analysis of Dravidian Code-Mixed Kannada comments. Using the BERT model, it achieved a F1-score of 0.66 on validation dataset, and F1-score of 0.619 on test dataset.

Hinglish to English Machine Translation using Multilingual Transformers by Vibhav Agarwal, Pooja Rao S B, Dinesh Babu Jayagopi from International Institute of Information Technology Bangalore, India and University of Lausanne, Switzerland proposed machine translation of code-mixed Hinglish to English language using the pretrained transformers like mBART and mT5 on the PHINC dataset. Also, they report a significant jump from 15.3 to 29.5 in BLEU scores which was a 92.8% improvement over the same dataset [19].

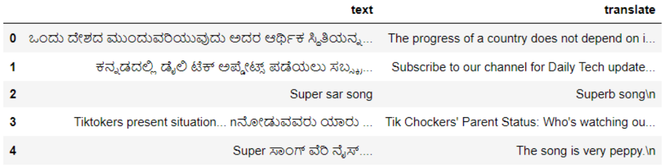
III. Proposed Work

In this section, we explain about the mBART and mT5 based machine translation model and the dataset that will be used to proceed the experiment.

A. Dataset

For the dataset we create a csv file with the Kanglish sentences along with their English translations. And the Kanglish sentences will be of the kind that are fully Kannada, Kannada mixed with English and Kannada transliterated in English wordings. The dataset is created in this way so that the model will perform well to deal with the different styles of input. The data is collected from internet source such as public corpora or social media comments etc.

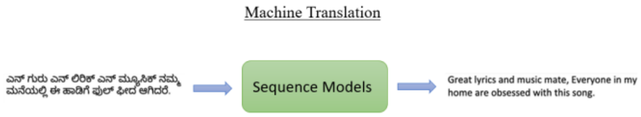
TABLE I
A sample from the collection of data



	text	translate
0	ಒಂದು ದೇಶದ ಮುಂದುವರಿಯುವುದು ಅದರ ಆರ್ಥಿಕ ಸ್ಥಿತಿಯನ್ನ...	The progress of a country does not depend on i...
1	ಕನ್ನಡದಲ್ಲಿ ಡೈಲಿ ಟೆಕ್ ಅಪ್ಡೇಟ್ಸ್ ಪಡೆಯಲು ಸಬ್ಸ್ಕ್ರ...	Subscribe to our channel for Daily Tech update...
2	Super sar song	Superb song\n
3	Tiktokers present situation... nನೋಡುವವರು ಯಾರು ...	Tik Chockers' Parent Status: Who's watching ou...
4	Super ಸಾಂಗ್ ವೆರಿ ನೈಸ್....	The song is very peppy.\n

B. Machine Translation Model

Sequence to Sequence (often abbreviated to seq2seq) models is a special class of Recurrent Neural Network architectures that we typically use (but not restricted) to solve complex Language problems like Machine Translation, Question Answering, creating Chatbots, Text Summarization, etc.


Machine Translation
ಎನ್ ಗುರು ಎನ್ ಲಿರಿಕ್ ಎನ್ ಮ್ಯೂಸಿಕ್ ನಮ್ಮ ಮನೆಯಲ್ಲಿ ಈ ಹಾಡಿಗೆ ಫುಲ್ ಫೀದ ಆಗಿದರೆ.
Sequence Models
Great lyrics and music mate, Everyone in my home are obsessed with this song.

mBART is a seq2seq multilingual bidirectional auto-encoder pretrained model which is an improved model of BART (Lewis et al., 2020 [4][5]) but on large-scale monolingual corpora of 35 languages. It is an architecture that consists of 12 encoder and decoder layers each with 16 attention heads and model dimensions being 1024 resulting in roughly 680 million parameters.

mT5 is a T5 model's multilingual variant pretrained on 101 languages. It has quite similar architecture with 2 encoder and decoder layers each, model dimensions being 1024 and 12 attention heads resulting in approximately 770 million parameters [19].

We propose our model such that the transformers and their associated weights are from HuggingFace's Transformer (Wolf et al., 2020 [1]) package. We planned to use mbart-cc-25 model weight for mBART and mT5-base model weight for mT5. We also plan to finetune the mBART and mT5 models on our created dataset and acquire good performance by altering the hyperparameters and the number of epochs.

Finally, to check the measure the performance accuracy of the model we tend to use the BLEU score for each of the model respectively and compare the results.

IV. CONCLUSIONS

Our key insight of the project is to use the large multilingual transformers and demonstrate their performance on translating Kanglish code-mixed language into English language.

There are deep learning models by which we can achieve this project. But fine tuning the already existing i.e., pretrained models for our requirements will be helpful in translating the code-mixed languages.

As a part of our future work, we would like to improve our model to translate English to code-mixed language translation which in turn will help to improve the Code-mixed to English translation. Also, as an extension to implement code-mixed to English translation to the diverse Kannada slangs which variates as per different places of Karnataka.

ACKNOWLEDGMENT

Apart from the efforts of myself, the success of the project depends largely on the encouragement and guidelines of many others, I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project.

I would like to express a deep sense of gratitude to Mrs.Deepalakshmi Coordinator of Data Science Department, NMKRV College for Women, Bengaluru for guiding me throughout the project.

I would like to extend my heartiest thanks to my parents, friends and well-wishers for being with me and extending encouragement throughout the project.

Finally, I am greatly indebted to the various writers, jurists and all others from whose writings and work I have taken help to complete this project.

REFERENCES

[1] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

[2] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward ¨ Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 8024–8035.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4- 9, 2017, Long Beach, CA, USA, pages 5998–6008.

[4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

[5] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8:726–742.

[6] Mona Diab, Pascale Fung, Mahmoud Ghoneim, Julia Hirschberg, and Thamar Solorio, editors. 2016. Proceedings of the Second Workshop on Computational Approaches to Code Switching. Association for Computational Linguistics, Austin, Texas.

[7] Mona Diab, Julia Hirschberg, Pascale Fung, and Thamar Solorio, editors. 2014. Proceedings of the First Workshop on Computational Approaches to Code Switching. Association for Computational Linguistics, Doha, Qatar.

[8] Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In Proceedings of the Second Workshop on Computational Approaches to Code Switching, pages 40–49, Austin, Texas. Association for Computational Linguistics.

[9] Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In Proceedings of the First Workshop on Computational Approaches to Code Switching, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

[10] Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Thamar Solorio, Mona Diab, and Julia Hirschberg, editors. 2018. Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-switching. Association for Computational Linguistics, Melbourne, Australia.

[11] Pattabhi R. K. Rao and S. Devi. 2016. Cmee-il: Code mix entity extraction in indian languages from social media text @ fire 2016 - an overview. In FIRE.

[12] Khyathi Chandu, Ekaterina Loginova, Vishal Gupta, Josef van Genabith, Gunter Neumann, Manoj Chin- ¨ nakotla, Eric Nyberg, and Alan W. Black. 2018. Code-mixed question answering challenge: Crowd-sourcing data and techniques. In Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, pages 29–38, Melbourne, Australia. Association for Computational Linguistics.

[13] Anupam Jamatia, Bjorn Gamb ¨ ack, and Amitava Das. ¨ 2018. Collecting and Annotating Indian social media Code-Mixed Corpora. In Computational Linguistics and Intelligent Text Processing, pages 406– 417, Cham. Springer International Publishing.

[14] Somnath banerjee, Kunal Chakma, Sudip Kumar Naskar, Amitava Das, Paolo Rosso, Sivaji Bandyopadhyay, and Monojit Choudhury. 2016. Overview of the Mixed Script Information Retrieval (MSIR). In Proceedings of FIRE 2016. FIRE.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[16] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettle- ´ moyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440– 8451, Online. Association for Computational Linguistics.

[17] Alexis Conneau and Guillaume Lample. 2019. Cross lingual language model pretraining. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 7057–7067.

[18] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

[19] Vibhav Agarwal, Pooja Rao S B, Dinesh Babu Jayagopi, International Institute of Information Technology

Bangalore, India; and University of Lausanne, Switzerland. "Hinglish to English Machine Translation using Multilingual Transformers", Proceedings of the Student Research Workshop associated with RANLP-2021, pages 16-21, held online, Sep 1-3, 2021