Code Switched Machine Translation

Manas Bundele
University of Texas, Dallas

Randeep Ahlawat
University of Texas, Dallas

Siddhartha Sahai
University of Texas, Dallas

Abstract

Code switching is defined as the practice of switching between two or more languages or varieties of languages in conversation. The proliferation of code-switched languages, in the context of informal communication, has risen tremendously over the recent years, resulting in an increasing need for machine translation systems catered towards code switched languages. However, due to the low resource nature of code-switched language translations, existing machine translation systems fail to perform adequately. In this paper we propose a novel machine translation system which leverages a pre-trained language model to generate translations for low resource code-switched languages. In particular, we devise a transformer NMT model, fine tuned using fastText embeddings over a corpus of code-switched Hindi-English to English sentence pairs. Our model outperforms other recent approaches aimed at Hindi-English translation

1 Introduction

The massive adoption of social media platforms and communication applications like Facebook, Email, Whatsapp etc. as means of communication has led to an enormous increase in informal discourse. Consequently, the use of hybrid code-switched languages has increased in such contexts.

Code-switched languages are especially prevalent in multilingual societies (Choudhury et al., 2007). Hindi-English or *Hinglish* is ubiquitous on Social media platforms in India. Hinglish sentences follow the syntax of Hindi but borrow some vocabulary from English. Below is an example of an Hinglish sentence along with its translation:

> Hi-En : Main kal *movie* dekhne jaa rahi thi *and* raaste me *I met Sam*.
>
> English : I was going to a movie yesterday and I met Sam on the way.

According to [1]InternetWorldStats.com, there are over 1 Billion English speakers on the internet out of 4 billion users. Amongst these unaccounted 3 billion users, many users belonging to dual-lingual or multilingual countries utilize code-switched languages. While a lot of focus has been placed on translation of one major language to another language, code-switched translation is an area which has not received much attention in spite of its prevalence.

The problems associated with code-switched language translation lies in the ephemeral nature of Social Media content in which it is used. Social media content is generated and updated at a fast pace with a short lifespan and immense volumes of it are produced, rules out the possibility of translations by human subjects. Consequently, there is a dearth of gold parallel data. Neural machine translation systems are the de facto standard for translation over Statistical models due to ability to train an end to end system without the need to deal with word alignments, translation rules and complicated decoding algorithms((Bahdanau et al., 2014); (Cho et al., 2014); (Sutskever et al., 2014)). Neural machine translation system, due to their data hungry nature, perform poorly on low resource task such as code-switched translation(Zoph and Le, 2016).

2 Related Work

There has been an increasing interest in supporting code-mixing for language models as the amount of text online grows. (Vyas et al., 2014)deal with POS tagging of English-Hindi code-mixed data that they extracted from social media. (Sharma et al., 2016) have addressed shallow parsing of English-Hindi code-mixed data. (Raghavi et al., 2015) explored the problem of classification of code-mixed questions. WebShodh, developed by (Chandu et al.,

[1] https://internetworldstats.com

2017), is an online web based question answering system based on their work. (Dhar et al., 2018) have developed an augmentation approach for code-mixed data sets that can be plugged into existing NMT systems. Their system classifies languages into matrix and target languages using heuristics such as number of words, syntactic structure and presence of keywords. They evaluate their results on BLEU scores with 3 NMT systems - Moses (Koehn et al., 2007), Googles Neural Machine Translation System (NMTS) (Wu et al., 2016), and Bing Translator. Research in code-mixed language models has historically been constrained by the lack of quality data sets. Several efforts address this problem - the IIT Bombay parallel corpus by (Kunchukuttan et al., 2017) aims to address this problem and was used in the Workshop on Asian Language Translation Shared Task in 2016 and 2017. (Dhar et al., 2018) also created a code-mixed data set with the help of 4 human translators.

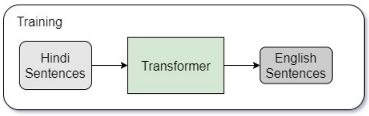
3 Details

3.1 Data set

We utilize 2 data sets for the purposes of training and fine-tuning the NMT model. One large data set is used for training the NMT as well as creating word embeddings (Kunchukuttan et al., 2017). This consists of 800,000 translation pairs of monolingual Hindi-English sentences. The Devanagari Hindi sentences are translated to Roman Hindi using the indic translation library. This corpus is referred to as large corpus and is only used for training, not testing. A smaller data set of 6000 code-mixed sentences is used to fine-tune and test the NMT model (Dhar et al., 2018). This model contains crawled Hinglish-English sentence pairs from social media pages.

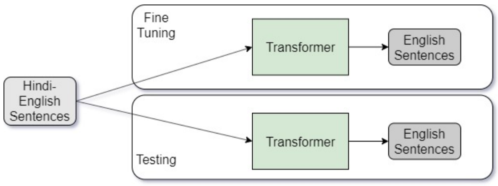
3.2 Base Model

We implement a custom Neural Machine Translation system that uses Transformer architecture(Vaswani et al., 2017). The neural architecture includes 6 encoder-decoder layers with positional encoding. The RNN size is 512, the word vector size is 512, with 8 heads. This NMT was trained on the large corpus ((Kunchukuttan et al., 2017)). The resulting model is used to translate the small corpus test set (Dhar et al., 2018). The predicted sentences are evaluated on the basis of BLEU score.



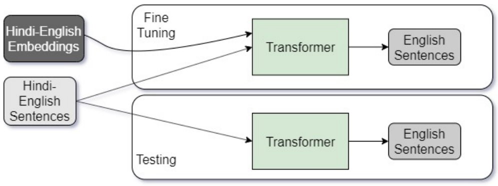
3.3 Fine Tuning

We re-use the transformer NMT architecture from the previous section, and add an additional step of fine-tuning the model on 5400 sentences of the small corpus (Chu et al., 2017). The resulting model is again used to translate the remaining small corpus as test set. Thus transfer learning is used on a low-resource data set (Hinglish - English sentences) utilizing the model trained on a larger different data set.



3.4 Model Augmentation with Bilingual Embeddings

We create a text corpus containing concatenated Hindi and English sentences. This corpus was used to initialize a skip-gram word embedding using FastText (Bojanowski et al., 2016). These embeddings are used to initialize the pre-trained NMT model while fine-tuning on the small corpus as above. The embedding created has dimensions of 512 with maximum and minimum length of character n-grams set to 5 and 2 respectively. The embeddings are used to better link the grammar of Hindi sentences to the words in the English sentences of the parallel corpus.



3.5 Experimental Settings

All models are trained on google colab notebooks with common CPU, GPU, RAM and disk space. There is 1 Tesla K80 GPU with 2496 CUDA cores and 12 GB GDDR5 VRAM. The CPU include 1 single core hyper-threaded Xeon Processor with a clock speed of 2.3Ghz and 45MB cache. The RAM available is approximately 12.5 GB and disk space around 320 GB.

Model	**BLEU**	**WER**
Augmentation + NMTS	0.07	0.83
Base Transformer	0.02	1.22
Fine Tuned Transformer	0.1226	0.699
Fine Tuned Transformer + Embeddings	0.1254	0.697

Table 1: BLEU and WER scores over 610 Hindi-English to English sentence pairs

All data is lower cased and limited to 175 characters. We utilize byte pair encoding to handle out of vocabulary words. We train byte pair encoding on the large data set for source and target sentences and apply the trained encoding to the small data set sentence pairs.

We train the Transformer NMT from scratch using 800,000 sentence pairs of the large corpus. It is trained for 54000 epochs, with learning rate = 2, Adam optimizer with beta2 = 0.998, using glorote param initialization, label smoothing set to 0.1, dropout set to 0.1, and batch size as 4096.

We fine-tune the pre-trained NMT using 5400 sentence pairs of the small corpus. This fine tuning is done for 5000 epochs, with learning rate = 2, Adam optimizer with beta2 = 0.998, using glorote param initialization, label smoothing set to 0.1, dropout set to 0.1, and batch size as 4096.

The word embeddings are created using a single text file containing 630478 transliterated Hindi sentences followed by 630478 English sentences. This embedding is converted to torch embeddings using OpenNMT tools, and passed into the NMT model when starting the fine-tuning phase. It is fine-tuned for 5000 epochs.

4 Results

We have chosen BLEU scores as a metric for quantifying the performance of our translation model. 1 presents the BLEU scores for the Hindi-English to English parallel corpus.

5 Conclusion and Future Work

We present a transformer based translation model for code-switched sentences. The transformer was trained on monolingual to monolingual (Hindi to English) parallel data. We fine tuned the trained model on code-switched to monolingual (Hindi-English to English) parallel data and further added bilingual(Hindi and English) embeddings to the model. Improvements were observed after fine tuning the monolingually trained transformer. Further improvements were discerned by augmenting the fine tuned transformer with word embeddings. From the results, we infer that improvements in performance by fine tuning and addition of bilingual embeddings can be attributed to the resemblance of the structure of the code-switched data to the source language. In the future, we plan to integrate pre-trained models, namely ELMo ((Peters et al., 2018)) and BERT ((Devlin et al., 2018)), with our framework to provide contextualized word embeddings, with the aim of improving performance even more. We plan to devise data augmentation approaches to artificially generate code-switched translation pairs.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information.

Khyathi Raghavi Chandu, Manoj Chinnakotla, Alan W. Black, and Manish Shrivastava. 2017. Webshodh: A code mixed factoid question answering system for web. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 104–111, Cham. Springer International Publishing.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. *International Journal of Document Analysis and Recognition (IJDAR)*, 10(3):157–174.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Mrinal Dhar, Vaibhav Kumar, and Manish Shrivastava. 2018. Enabling code-mixed translation: Parallel corpus creation and MT augmentation approach. In

Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, pages 131–140, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The iit bombay english-hindi parallel corpus.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Khyathi Chandu Raghavi, Manoj Kumar Chinnakotla, and Manish Shrivastava. 2015. "answer ka type kya he?": Learning to classify questions in code-mixed language. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 853–858, New York, NY, USA. ACM.

Arnav Sharma, Sakshi Gupta, Raveesh Motlani, Piyush Bansal, Manish Shrivastava, Radhika Mamidi, and Dipti M. Sharma. 2016. Shallow parsing pipeline - Hindi-English code-mixed social media text. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1340–1345, San Diego, California. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. POS tagging of English-Hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979, Doha, Qatar. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Barret Zoph and Quoc V. Le. 2016. Neural architecture search with reinforcement learning.