Towards Code-Mixed Hinglish Dialogue Generation

Vibhav Agarwal, Pooja Rao S B*, Dinesh Babu Jayagopi
International Institute of Information Technology Bangalore, India
* University of Lausanne, Switzerland
{vibhav.agarwal,jdinesh}@iiitb.ac.in, pooja.rao@unil.ch

Abstract

Code-mixed language plays a crucial role in communication in multilingual societies. Though the recent growth of web users has greatly boosted the use of such mixed languages, the current generation of dialog systems is primarily monolingual. This increase in usage of code-mixed language has prompted dialog systems in a similar language. We present our work in Code-Mixed Dialog Generation, an unexplored task in code-mixed languages, generating utterances in code-mixed language rather than a single language that is more often just English. We present a new synthetic corpus in code-mix for dialogs, CM-DailyDialog, by converting an existing English-only dialog corpus to a mixed Hindi-English corpus. We then propose a baseline approach where we show the effectiveness of using mBART like multilingual sequence-to-sequence transformers for code-mixed dialog generation. Our best performing dialog models can conduct coherent conversations in Hindi-English mixed language as evaluated by human and automatic metrics setting new benchmarks for the Code-Mixed Dialog Generation task.

1 Introduction

Due to the popularity of different social media and messaging platforms over the last decade, there has been a significant increase in internet users, mainly from multilingual societies. Multilingual speakers regularly combine languages in what is commonly called code-mixing or code-switching while communicating with other multilingual speakers. This has resulted in a substantial influx of mixed language data in the form of comments, conversations, and other forms of communication. Traditional natural language processing tasks like tokenization and tagging, semantic processing, machine translation, and text generation face new and interesting challenges due to this language mixing.

Dialog Systems have been of great interest amongst the natural language processing community for widespread applications. These systems are broadly categorized into three categories: task-oriented dialog system (Wen et al., 2017; Williams and Zweig, 2016), open-ended conversational systems (Shang et al., 2015; Xing et al., 2017), and interactive question answering system. Traditional Dialog Systems have mostly relied on a rule or template-based approach (Williams et al., 2013).

The success of deep neural networks with a considerable amount of training data has led towards end-to-end trained sequence-to-sequence (seq2seq) (Sutskever et al., 2014) models that enhance the generality and diversity of the text generated. Recent advances in attention-based mechanisms (Bahdanau et al., 2015) and Transformers (Vaswani et al., 2017) have shown significant performance improvement and shifted the communities' approach and interest in training larger models. However, all of these prior works use monolingual data and specifically English.

The increasing use of code-mixed languages and ubiquitous nature of multilingual speakers call for catering to the needs of such users in a multilingual fashion with a dialog system, the need for a Code-Mixed Dialog Generational System. A recent study by Bawa et al. (2020) shows that in a real-life setting, people prefer chatbots that engage in code-mixed language.

Code-mixing in informal contexts like newsgroups, tweets, comments and blogs has made it difficult to define a uniform structure to the language. However, linguists have formulated various hypotheses (Belazi et al., 1994; Pfaff, 1979; Poplack, 1981) and constraints (Sankoff and Poplack, 1981; Sciullo et al., 1986; Joshi, 1982) that can define a general rule for code-mixing.

With the rise of large pretrained language models like (Devlin et al., 2019; Radford et al., 2019; Conneau et al., 2020), there's been a lot of improve-

Proceedings of the Third Workshop on Natural Language Processing for Conversational AI, pages 271–280
November 10, 2021.

ment in machine translation and multilingual models. Prior works (Khanuja et al., 2020b; Gupta et al., 2020) show the effectiveness of large multilingual pretrained language models like mBERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019) on code-mixed data.

Our work attempts to utilize these large seq2seq pre-trained multilingual transformer-based models for the code-mixed dialog generation task. Specifically, our contributions are as follows.

- We propose a synthetic code-mixed dialog dataset, CM-DailyDialog. This is the first benchmark dataset in Hindi-English mixed language for dialog generation generated from translating DailyDialog (Li et al., 2017) to code-mix.

- We set new benchmarks for the code-mixed dialog generation task on the CM-DailyDialog. We finetune the mBART model for the dialog generation in two ways to generate coherent dialog, as seen from both automatic and human evaluation metrics.

- To create CM-DailyDialog, we train a machine translation model. We use monolingual English data as our input instead of a parallel English and Hindi corpus. This differs from earlier work on code-mixed Machine Translation (Gupta et al., 2020; Garg et al., 2018) where they process a parallel corpus for training, making our approach less resource-intensive.

2 Related Work

Code-mixing refers to the interleaving of words belonging to different languages. Code-mixing is most common in multilingual cultures, and it is becoming more common as the number of people using social media and messaging services on the internet grows (Ansaldo et al., 2008). This has lead to a rising research interest in recent years, and several tasks have been conducted as part of code-switching workshops (Diab et al., 2014, 2016). There has been a lot of advancements in solving different code-mixed tasks like language identification (Solorio et al., 2014; Molina et al., 2016), named-entity recognition (Rao and Devi, 2016; Aguilar et al., 2018), question answering (Chandu et al., 2018), part-of-speech tagging (Jamatia et al., 2018), and information retrieval (banerjee et al., 2016). Very recently, a code-mixed version of the GLUE benchmark was proposed by Khanuja et al. (2020b) which introduced a common benchmark for all the tasks.

2.1 Code-Mixed Machine Translation

Machine Translation on Code-Mixed language is a relatively less explored area. There are only a few studies on English to Hinglish Code-Mixed language despite a large population of Code-Mixed speakers in South Asian countries. Dhar et al. (2018) collects a dataset of 6,096 Hinglish-English bitexts and proposes a pipeline where they identify the languages involved in the code-mixed sentence, compute the matrix language and then translate the resulting sentence into a target language. Srivastava and Singh (2020) collects a diversified large parallel corpus called PHINC, consisting of 13,738 Hinglish-English bitexts. Gupta et al. (2020) propose a code-mixed text generator built upon the encoder-decoder framework. They propose using features from a pre-trained cross-lingual transformer based model XLM (Conneau and Lample, 2019) along with Pointer-Generator (See et al., 2017) model as its decoder for the code-mixed text generation.

2.2 Conversational Dialog Systems

This work focuses on open-ended conversational dialog systems, which are interchangeably also called dialog systems here. Conversational systems engage in more open-ended conversations with no specific objective or task to solve in contrast to a task-oriented dialog system. In the open domain, conversational dialog systems fall again into two categories: retrieval-based and generative dialog systems. We take the generative model approach for a code-mixed dialog model rather than retrieve them from a fixed set as it is more dynamic and interactive.

Traditional dialog systems have mostly relied on hand-crafted rules or templates (Williams et al., 2013). Recently, a more data-driven approach is used to enhance the generality and diversity of the text generated in other domains. Researches used RNN (Rumelhart et al., 1986) and LSTM (Hochreiter and Schmidhuber, 1997) based encoder-decoder architectures for the dialog systems. Since the popularity of attention-based mechanisms (Bahdanau et al., 2015), these also have been widely adapted to boost performance. Recent works like DialoGPT (Zhang et al., 2020), Blender-

Bot (Smith et al., 2020) and Meena (Adiwardana et al., 2020) are just a few examples of the open domain conversational agents.

Although much work has been done in dialog systems, mostly all of it is for English conversations. This is because, in most of the work, the dataset used is monolingual. The only recent work involving a multilingual conversational system is by Chen et al. (2019) which performs dialog generation on English and Chinese data. It uses a shared memory mechanism with a seq2seq encoder-decoder like architecture and is trained using multi-task learning (Caruana, 1997).

3 System Overview

This section describes the benchmark dataset for code-mixed dialog generation: CM-DailyDialog, the English to Hinglish translation model used to generate this dataset, and our mBART based dialog generation model.

3.1 CM-DailyDialog dataset

There is no standardized dataset available for multilingual dialog generation; therefore, we choose to generate a synthetic dataset to train our model for code-mixed dialog. We use a standardized and popular English dialog dataset called DailyDialog (Li et al., 2017) and translate the utterances and conversations from English to Code-Mixed using our mBART model (**mBART-en_cm**) defined in Section 3.1.1. This results in the CM-DailyDialog dataset consisting of 11,118 conversations in the training set and 1,000 conversations in both test and validation sets.

We also use the Code-Mixed NLI conversation dataset from the GLUECoS benchmark (Khanuja et al., 2020a). This dataset contains roughly 1,800 training and roughly 500 test conversations extracted from movies.

We first process the multi-turn conversations from the CM-DailyDialog dataset into triplets of utterances using a sliding window approach. The first two utterances in that triplet are served as contextual inputs to the model, while the third utterance is served as the ground truth on which the loss is calculated. This processing of conversations into triplets of utterances increased the size of our dialog dataset from 13,118 to 76,745 data points. Similarly, we process the Code-Mixed NLI dataset into similar triplets, expanding our working dataset from 1,800 to 2,128 unique dialog triplets in the train set and by roughly 500 triplets in the test set. We choose to process these multi-turn conversations into splits of 3 and not say 5 or any other number because of increased computational costs for the mBART model to process such long conversations.

We describe our English to Hinglish translation model mBART-en_cm and the dataset used for its training in the following section.

3.1.1 Machine Translation model

We use an mBART model finetuned on English to Code-Mixed data described in Section 3.1.2 as our machine translation model to convert the English DailyDialog dataset to Code-Mixed form. We denote this model as **mBART-en_cm**. mBART is a multilingual seq2seq denoising bidirectional auto-encoder pre-trained using the same objective as BART (Lewis et al., 2020) but on large-scale monolingual corpora of 25 languages. It is based on the transformer (Vaswani et al., 2017) architecture and consists of 12 encoder and decoder layers, each with 16 attention heads and model dimensions being 1024 resulting in roughly 680 million parameters.

3.1.2 Dataset for Translation

We use the following datasets to finetune and test our *mBART-en_cm* model for English to Hinglish translation task:

- **CMU Hinglish** is an extended Code-Mixed form of the Document Grounded Conversation dataset by Zhou et al. (2018). It consists of roughly 10,000 English and Hinglish Code-Mixed sentences.
- **Reverse PHINC** is the reverse version of the PHINC (Srivastava and Singh, 2020) dataset but we switch the source and target pairs for our task. It contains roughly 13,000 Hinglish and parallel English translations.
- **LinCE** (Aguilar et al., 2020) Benchmark provides an English to Hinglish Code-Mixed dataset as part of their Code-Mixing benchmark for Machine Translation and GLUE. The dataset consists of roughly 10,000 English and Code-Mixed pairs.

We use these datasets individually and in conjunction to see any improvement with increased data.

Dataset	BLEU	Perplexity	CMI_{test}
mBART-dialog			
CM-DailyDialog	4.11	17.4	27.5
CM-DailyDialog + Code-Mixed NLI	1.54	13.6	22.6
mBART-dialog$^+$			
CM-DailyDialog	**8.11**	20.54	**28.9**
CM-DailyDialog + Code-Mixed NLI	5.74	12.93	26.8

Table 1: mBART performance on Dialog Generation

3.2 Code-Mixed Dialog Model

We use the end-to-end multilingual training of the mBART. In literature, there is minimal work utilizing the BART architecture for dialog generation. De Bruyn et al. (2020) is one such work. It utilizes BART for knowledge grounding and knowledge retrieval in dialogs. Our approach attempts to leverage multilingualism by using the pre-trained BART for multilingual dialog generation and presents a few baselines for future work. We compare two strategies for finetuning mBART model for dialog generation:

- **mBART-dialog:** We finetune the mBART model on a Code-Mixed dialog dataset. In our case, we use triplet utterances to train our model.
- **mBART-dialog$^+$:** We finetune the mBART model in a dual curriculum learning method where we first finetune the mBART on an English to Code-Mixed translation task and then on a Code-Mixed dialog dataset.

4 Experimental Setup

In this section, we describe our experimental setup for both our mBART-en_cm model and the dialog model described in Section 3.1.1 & 3.2 respectively.

Our proposed approach is written in Pytorch (Paszke et al., 2019), and the mBART model weights and architecture used are from the HuggingFace's Transformer (Wolf et al., 2020) package. We use only the *mbart-cc-25* weights in all our modeling. All the mBART based models were trained using the AdamW optimizer with weight decay. We used all the default hyperparameters except the number of training epochs. We finetuned all our mBART models for five epochs only. As discussed in Section 2, there is extremely limited prior work and literature on multilingual dialogs. Therefore, there is no baseline for us to compare our model to and we report our numbers as it is.

Datasets	BLEU	CMI_{test}
LinCE Benchmark	11	28.3
CMU Hinglish	**11.53**	**32.5**
CMU Hinglish + LinCE Benchmark + Reverse PHINC	11.25	31.9

Table 2: BLEU score on test set for English to Hinglish Translation using mBART-en_cm model

As discussed in Section 3.1, we process our datasets from English to Code-Mixed and then into triplets. We split our processed CM-DailyDialog dataset into 8:1:1 splits for training, validation, and test set and use the additional Code-Mixed NLI dataset in conjunction with the CM-DailyDialog dataset to see any performance improvement with the increased data. We evaluate both our mBART-dialog and mBART-dialog$^+$ models on BLEU and perplexity metric and report our scores in the Table 1. To gauge the language mixing performance of our models, we also use the Code-Mixing Index (CMI) (Gambäck and Das, 2016). We report sacrebleu as our BLEU metric using the HuggingFace's Dataset package.

We also show the performance of our *mBART-en_cm* model on different datasets for the monolingual English to Hinglish translation task using metrics like sacrebleu(reported as BLEU) and Code-Mixing Index in Table 2.

5 Results

Table 1 shows that the model trained using the dual curriculum learning method (mBART-dialog$^+$) performs better both on the BLEU as well as the CMI

English Dialogs to Code-Mixed Translation		
S1: Good afternoon. This is Michelle Li speaking, calling on behalf of IBA. Is Mr Meng available at all?	→	**S1:** Accha afternoon. Ye Michelle Li speaking hai, IBA ka on behalf calling. Kya Mr Meng available hai?
S2: This is Mr Meng speaking, Michelle.		**S2:** Ye hai Mr Meng speaking, Michelle.
S1: Oh, hello! Sorry about that. I'm just calling to say that we've received your new Corporate Credit Card from HQ.		**S1:** Oh, hello! Sorry iske bare mein. Main bas kah raha hoon ki hamen apna new Corporate Credit Card mil gaya hai HQ.

Table 3: Translating DailyDailog dataset from English to Code-Mixed. Blue tokens refer to the Hindi tokens in the Roman script. S1 and S2 refer to Speaker 1 and 2 respectively.

Code-Mixed Dialogs
S1: actually, fruits aur veggies tumhe ache lagte hain
S2: haan, muje patha hein, lekin chicken ke baare mein kya?
S1(generated): Mujhe lagta hai I'm going to make a slice of it.
S1: Mike! Tumhare se sunke accha laga. Aap kaise hain?
S2: everything is fine , aur tum kaise ho?
S1(generated): Main thik hoon. Tumhare sath baat krke accha laga.

Table 4: Examples of the response generated by mBART-dialog^{+} on the CM-DailyDialog. S1 and S2 refer to Speaker 1 and 2 respectively.

metric. This boost in performance might be due to the model understanding code-mixed language after the first finetune and, as a result, adapting better over the code-mixed dialogs in the second finetune. We show some of the examples of our dialog model in Table 4. We also observe that simply increasing the data does not necessarily increase the model performance and leads to a significant drop in this case. This drop might be due to the inconsistent Hindi vocabulary in the romanized form in different datasets. The same Devanagari token can be represented in various Roman scripts in different datasets. This can cause the model not to have a fixed code-mixed vocabulary, causing this confusion and, hence, a drop in model performance.

Table 2 shows our mBART-en_cm model performance on different datasets. As observed previously, increasing the data leads to a drop in performance, which may be due to different datasets' vocabulary discrepancies. We use the best performing model, i.e. trained on the CMU Hinglish dataset and use that to generate our CM-DailyDialog dataset as described in Section 3.1. Table 3 shows some of the translation examples from English DailyDialog to CM-DailyDialog.

Table 5 shows some of the statistics of the CM-DailyDialog dataset. The CMI scores for all the splits for our generated dataset are close to that of the real world code-mixed datasets like Dhar et al. (2018). This strengthens our intent to utilize this synthetic code-mixed dialog dataset for our dialog generation model.

5.1 Error Analysis of mBART-en_cm Translations

Considering BLEU and CMI ratings do not give insight into translation errors, we use error analysis to further assess the quality of our CM-DailyDialog dataset. We assess the quality of our mBART-en_cm model's translations on the test set by grouping the different errors generated by the model into three error categories and a no error category. We randomly sample 50 sentences from our test set and bucket them into categories. We follow the error analysis categories from Gautam et al. (2021). We employ three human raters that classify the sampled translations into the error buckets. Graduate students (non-native English speakers) familiar with the usage of code-mixed language, specifically Hinglish, in everyday life are the human annotators involved in this research. We report our numbers as a mode of three rater evalu-

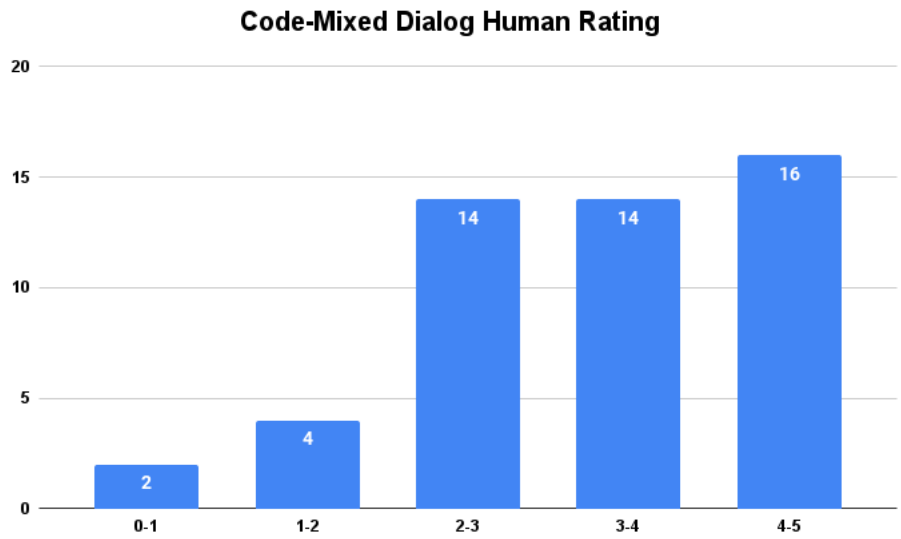

Figure 1: Bucketed average rating of 3 raters over the Code-Mixed dialogs quality from the range of 1-5

	Train set	Dev set	Test set
# of English tokens	295,565	54,150	32,676
# of Hindi tokens	546,535	86,231	67,505
# of Other tokens	77,519	13,493	10,721
Code-Mixing Index (CMI)	35	38.5	32.6

Table 5: Statistics of the generated CM-DailyDialog dataset

ations to account for the subjectivity among the raters. Mistranslated/Partially translation category indicates if the translation has low or no semantic resemblance with the source sentence. Morphological/Syntactical errors indicate if the translation has the same semantic meaning as the source sentence but has minor grammatical or syntax errors. NER mistranslations refer to the situation where the model translates the named entities in the generated output.

Table 6 shows the results of the error analysis over the described errors categories for the 50 test translations. We observe that the model makes 12 syntactical errors and 13 partial/mistranslations out of the 50 samples. After a more nuanced analysis of these numbers, we find that most of the syntax errors were 1-2 token errors or misalignment of those tokens. We also found that out of the 13 partial/mistranslations, only 15% of the translations were complete mistranslations. Most of the sentences in this error category were partial translations where the model failed to translate and code-mix simultaneously.

Error Category	Freq.
Mistranslated/Partially Translated	13
Morphology/Syntax Issues	12
NER mistranslation	1
No error	24

Table 6: Error Analysis on 50 randomly sampled test translated sentences from our best performing mBART-en_cm model on CMU Hinglish Dataset

5.2 Human Evaluation of Code-Mixed Dialog

To further strengthen the assessment of the generated code-mixed dialog, we perform a human evaluation of our best performing dialog model (mBART-dialog^{+} trained on CM-DailyDialog). We employ three human raters who rate the generated follow-up dialog given prior contextual dialogs. These contextual dialogs refer to the first two utterances in the triplets that we processed in Section 3.1. As previously stated, the raters here are Graduate students familiar with the usage of Hinglish. The

raters were instructed to rate the quality of the dialog on a scale of 1-5, with 1 being the lowest. The quality was assessed in terms of both the coherence in the dialog and the code-mixing. We do this analysis on 50 randomly sampled dialog generations from the test set. The results of the human ratings can be seen in Figure 1 as the mean of all three raters.

As it can be seen from Figure 1, 60% of the dialog utterances achieve a score greater than 3. 88% of the dialog utterance are scored above 2. These numbers indicate that our machine-generated code-mixed dialog followups are of good quality both in terms of coherence as well as code-mixing.

6 Conclusion

We introduce a new benchmark dataset for code-mixed dialog generation, CM-DailyDialog, a code-mixed version of the DailyDialog. Our work proposes using multilingual Transformers (mBART) and demonstrates how they help in code-mixed dialog generation. We also introduce a new monolingual English to Code-Mixed machine translation model using mBART. With our comprehensive experiments, we show the effectiveness of our approach in terms of machine translation and dialog generation and set new benchmarks in the Code-Mixed dialog generation task. The manual error analysis illustrates the quality of the new dataset, although it is synthetically generated. In terms of both automatic and human evaluation metrics, we show that the dialog generated from our model is of high quality.

As part of the future work, we would like to improve our machine translation model to improve our CM-DailyDialog data that further boosts our dialog generation. Another huge scope of improvement is in the vocabulary discrepancy in different datasets, and we wish to resolve this to further boost our modeling and performance.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv:2001.09977 [cs, stat]*. ArXiv: 2001.09977.

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.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Ana Inés Ansaldo, Karine Marcotte, Lilian Scherer, and Gaelle Raboyeau. 2008. Language therapy and bilingual aphasia: Clinical implications of psycholinguistic and neuroimaging research. *Journal of Neurolinguistics*, 21(6):539–557.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Hedi M. Belazi, Edward J. Rubin, and Almeida Jacqueline Toribio. 1994. Code switching and x-bar theory: The functional head constraint. *Linguistic Inquiry*, 25(2):221–237.

Rich Caruana. 1997. Multitask Learning. *Machine Learning*, 28(1):41–75.

Khyathi Chandu, Ekaterina Loginova, Vishal Gupta, Josef van Genabith, Günter Neumann, Manoj Chinnakotla, Eric Nyberg, and Alan W. Black. 2018. Code-mixed question answering challenge: Crowdsourcing 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.

Chen Chen, Lisong Qiu, Zhenxin Fu, Dongyan Zhao, Junfei Liu, and Rui Yan. 2019. Multilingual Dialogue Generation with Shared-Private Memory. *arXiv:1910.02365 [cs]*. ArXiv: 1910.02365.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, 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.

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.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2020. Bart for knowledge grounded conversations. In *Converse@ KDD*.

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.

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.

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.

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.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855, Portorož, Slovenia. European Language Resources Association (ELRA).

Saurabh Garg, Tanmay Parekh, and Preethi Jyothi. 2018. Code-switched language models using dual RNNs and same-source pretraining. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3078–3083, Brussels, Belgium. Association for Computational Linguistics.

Devansh Gautam, Prashant Kodali, Kshitij Gupta, Anmol Goel, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. CoMeT: Towards code-mixed translation using parallel monolingual sentences. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 47–55, Online. Association for Computational Linguistics.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2267–2280, Online. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Anupam Jamatia, Björn Gambäck, 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.

Aravind K. Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020a. A new dataset for natural language inference from code-mixed conversations. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 9–16, Marseille, France. European Language Resources Association.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020b. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training 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.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, 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.

Carol W. Pfaff. 1979. Constraints on language mixing: Intrasentential code-switching and borrowing in spanish/english. *Language*, 55(2):291–318.

Shana Poplack. 1981. *Syntactic structure and social function of code-switching*, pages 169–184.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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*.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

David Sankoff and Shana Poplack. 1981. A formal grammar for code-switching. *Papers in Linguistics - International Journal of Human Communication*, 14:3–46.

Anne-Marie Di Sciullo, Pieter Muysken, and Rajendra Singh. 1986. Government and code-mixing. *Journal of Linguistics*, 22(1):1–24.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

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.

Vivek Srivastava and Mayank Singh. 2020. PHINC: A parallel Hinglish social media code-mixed corpus for machine translation. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 41–49, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

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.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, Metz, France. Association for Computational Linguistics.

Jason D. Williams and Geoffrey Zweig. 2016. End-to-end LSTM-based dialog control optimized with supervised and reinforcement learning. *arXiv:1606.01269 [cs]*. ArXiv: 1606.01269.

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.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3351–3357. AAAI Press.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.