GupShup: Summarizing Open-Domain Code-Switched Conversations *

Laiba Mehnaz
NYU, USA

Debanjan Mahata
Moody's Analytics, USA

Uma Sushmitha Gunturi
Virginia Tech, USA

Amardeep Kumar
Walmart, India

Rakesh Gosangi
Bloomberg, USA

Riya Jain
IIIT-Delhi, India

Gauri Gupta
IIIT-Delhi, India

Isabelle Lee
Univ of Washington, USA

Anish Acharya
Univ of Texas at Austin, USA

Rajiv Ratn Shah
IIIT-Delhi, India

Abstract

Code-switching is the communication phenomenon where the speakers switch between different languages during a conversation. With the widespread adoption of conversational agents and chat platforms, code-switching has become an integral part of written conversations in many multi-lingual communities worldwide. Therefore, it is essential to develop techniques for understanding and summarizing these conversations. Towards this objective, we introduce the task of abstractive summarization of Hindi-English (Hi-En) code-switched conversations. We also develop the first code-switched conversation summarization dataset - *GupShup*, which contains over 6,800 Hi-En conversations and their corresponding human-annotated summaries in English (En) and Hi-En. We present a detailed account of the entire data collection and annotation process. We analyze the dataset using various code-switching statistics. We train state-of-the-art abstractive summarization models and report their performances using both automated metrics and human evaluation. Our results show that multi-lingual mBART and multi-view seq2seq models obtain the best performances on this new dataset. We also conduct an extensive qualitative analysis to provide insight into the models and some of their shortcomings.

1 Introduction

Conversation summarization is the process of generating a condensed version of a given conversation while preserving the most salient aspects. With the extensive use of various chat applications such as messaging apps and virtual assistants (Klopfenstein et al., 2017), there has been a growing interest in the abstractive summarization of written conversations (Mehdad et al., 2014; Goo and Chen, 2018; Zhao et al., 2019). Automatic conversation summarization has potential applications in various fields such as healthcare (Song et al., 2020), call centers (Alam et al., 2016), education (Joshi and Rosé, 2007), and many other areas (Feng et al., 2021).

One of the biggest challenges in conversation summarization has been the lack of large datasets with human-annotated summaries. Most researchers evaluate their summarization techniques on transcriptions of AMI (Carletta et al., 2005), or ICSI meeting corpus (Janin et al., 2003) using the meeting topics as summaries. These corpora are very useful for various speech-related research problems, but they do not represent written conversations in chat applications. Recently, (Gliwa et al., 2019) published the SAMSum corpus, which contains over 16,000 written English conversations and their corresponding manually annotated summaries. Though these conversations were not extracted from actual chat applications, they were created by linguists to replicate natural conversations. To our knowledge, this is the largest summarization dataset for written conversations.

Leon: kya tujeh abhi tak naukari nahi mili?
Arthur: nahi bro, abhi bhi unemployed :D
Leon: hahaha, LIVING LIFE
Arthur: mujeh yeh bahot acha lagta hai, dopahar ko jagata hoon, sports dekhta hoon - ek aadmi ko aur kya chahiye?
Leon: a paycheck? ;)
Arthur: mean mat bano ...
Leon: but seriously, mere dosth ke company mein ek junior project manager offer hai, tujeh interest hai?
Arthur: sure thing, tere pass details hai?
Leon: <file_photo>
English Summary: *Arthur is still unemployed. Leon sends him a job offer for junior project manager position. Arthur is interested.*

Table 1: Example of a code-switched Hi-En conversation and the corresponding En summary. ■: En words, ■: transliterated Hi words, ■: language-agnostic words such as named entities and punctuation marks

The SAMSum corpus is monolingual (English); therefore any model trained on this dataset may

Emails ids of the corresponding authors are, lm4428@nyu.edu, Debanjan.Mahata@moodys.com, rajivratn@iiitd.ac.in. Authors 1,2,3, and 4 have equal contributions. Debanjan Mahata participated in this work as an Adjunct Faculty at IIIT-Delhi.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6177–6192
November 7–11, 2021.

not adapt effectively to multi-lingual, especially code-switched conversations where speakers alternate between different languages within the scope of a conversation or even an utterance (Gumperz, 1977; Muysken et al., 2000; Myers-Scotton, 1997). Code-switching is commonly observed during interactions between peers who are fluent in multiple languages. For example, in the Indian subcontinent, it is common for people to alternate between English and other regional languages like Hindi over the course of a single conversation. Code-switching is an integral part of both written and spoken conversations for various multi-lingual communities across the world (Auer, 2013). Developing models that can accurately process code-switched text is essential to the proliferation of NLP technologies to these communities and contributes towards the diversity and inclusivity of language resources (Joshi et al., 2020). However, building such models would require high-quality human-curated datasets.

This paper introduces the task of abstractive summarization of open-domain code-switched written conversations. Namely, given a multi-party conversation in Hi-En, the objective is to generate a summary in English, as shown in Table 1. A multi-party code-switched conversation C is a sequence of n utterances $\{u_1, u_2, ..., u_n\}$, where the i^{th} utterance u_i is written by p_j one of the k participants. The utterance u_i is a sequence of m tokens $\{x_1, x_2, ..., x_m\}$, where the tokens could either be in English or transliterated from Hindi. The goal of the code-switched conversation summarization task is to generate an English summary that captures the most salient aspects of the conversation.

These English summaries could serve as input to other downstream NLP models, often trained on English data, to perform various tasks such as intent classification, question answering, and item recommendation. To facilitate this task, we present a new corpus named *GupShup*[1], which contains over 6,800 Hi-En[2] code-switched conversations and corresponding human-annotated summaries in En and Hi-En. We build this dataset by manually translating a subset of conversations and summaries from the SAMSum corpus (Gliwa et al., 2019) from En to Hi-En. This effort has not only developed the first code-switched conversation summarization corpus but also a parallel corpus of En and Hi-En conversations with 76,330 utterances. Following are some of the main contributions of this work:

- We present the first open-domain code-switched conversation summarization dataset - *GupShup*[3] with over 6,800 Hi-En conversations and their corresponding annotated summaries in En and Hi-En.
- We characterize the complexity of this dataset through various code-switching statistics.
- Through rigorous experimental work, we benchmark the performances of various state-of-the-art abstractive summarization models.
- We perform a thorough human evaluation and qualitative analysis to provide insight into the strengths and shortcomings of different language models when dealing with code-switched data.

2 Background

In the linguistics community, code-switching typically refers to the change of language or grammatical systems from one utterance to another within the same conversation (Gumperz, 1982). On the other hand, code-mixing refers to the use of linguistic units such as phrases, words, or morphemes of one language in the utterance of another language (Myers-Scotton, 1997; Myers-Scotton et al., 2002). In other words, code-switching is an inter-utterance, and code-mixing is an intra-utterance phenomenon. However, in this paper, we use the term code-switching to refer to both these concepts.

Code-switching has started to gain some traction from computational linguists over the last few years (Barman et al., 2014b; Bali et al., 2014), where they developed datasets for many interesting problems such as language identification (Das and Gambäck, 2014), part of speech tagging (Barman et al., 2014a), question answering (Chandu et al., 2015), and named entity recognition (Singh et al., 2018). Additionally, researchers have also started to develop objective metrics that characterize the complexity of code-switching in a given corpus (Gambäck and Das, 2016; Guzmán et al., 2017). For a thorough review of datasets and other developments in this space, we recommend the review paper from (Sitaram et al., 2019).

Most of the code-switched datasets typically contain individual posts or comments from social media applications like Twitter and Facebook anno-

[1]The word 'gupshup' means conversations for fun in Hindi and Urdu languages. The proposed dataset in this paper is named as 'GupShup' for reflecting the nature of the content in the conversations and has no connection with the messaging platform gupshup.io.

[2]Throughout the paper we denote code-switched Hindi English conversations by Hi-En. This is also popularly known as Hinglish.

[3]https://github.com/midas-research/gupshup

Dataset	Conversational	Task
(Das and Gambäck, 2014)	✗	Language identification and POS tagging
(Barman et al., 2014a)	✗	POS tagging
(Chandu et al., 2015)	✗	Question Answering
(Jamatia et al., 2015)	✗	Language identification
(Jamatia et al., 2016)	✗	Language identification
(Banerjee et al., 2016)	✗	Question Answering
(Chakma and Das, 2016)	✗	Information Retrieval
(Patro et al., 2017)	✗	Language identification
(Bohra et al., 2018)	✗	Hate-speech Text Classification
(Gupta et al., 2018)	✗	Question Answering
(Banerjee et al., 2018)	✓	Close domain conversation system
(Chandu et al., 2018)	✗	Question Answering
(Patra et al., 2018)	✗	Sentiment analysis
(Singh et al., 2018)	✗	Named Entity Recognition.
(Bhat et al., 2018)	✗	Dependency parsing
(Bawa et al., 2018)	✓	Accommodation quantification
(Khanuja et al., 2020)	✓	Natural Language Inference
GupShup (This work)	✓	**Open domain conversation summarization**

Table 2: Comparison of GupShup with existing datasets on Hindi-English code-switched language tasks.

tated for different NLP tasks. There are very few code-switched datasets that contain complete conversations: back and forth utterances from multiple participants. Notable examples are: (1) Bangor Miami corpus (Margaret et al., 2014), which contains audio recordings and transcripts of informal Spanish-English (Sp-En) multi-party conversations, (2) COMMONAMIGOS corpus (Ahn et al., 2020), which contains 587 Sp-En conversations between human users and a dialogue system, (3) DSTC2 corpus (Banerjee et al., 2018), which contains Hi-En translations of the restaurant reservation dataset. As shown in Table 2, there are only three datasets with Hindi-English code-switched conversations, but none of them contain summaries.

A large majority of research in abstractive summarization focuses on news articles (Hermann et al., 2015; Grusky et al., 2018a; Narayan et al., 2018a) and scientific papers (Cohan et al., 2018), because of the availability of large benchmark datasets. The task of summarizing open-domain multi-party conversations has not been investigated until recently with the introduction of SAMSum (Gliwa et al., 2019). (Chen and Yang, 2020) obtained state-of-the-art results on this corpus with their multi-view sequence-to-sequence model, which extracts different *views* of the conversation, encodes them in an LSTM-based model, and then uses an attention-based mechanism in the decoding phase to generate the summary.

Conversation summarization is a challenging research problem because conversations are filled with various complex linguistic phenomenon such as informality, verbosity, interruptions, backchanneling, reconfirmations, hesitations, and many implicit connotations (Sacks et al., 1978). Therefore, it is difficult for current summarization approaches to identify the most relevant and salient aspects of a conversation (Chen and Yang, 2020). The code-switched nature of our data poses additional challenges. We believe the task's challenging nature would encourage the development of better multi-lingual language models and facilitate a new direction of research in the area of summarization, leading to innovations in modeling architectures, especially for code-switched text.

3 Data Collection

The goal of our annotation process is to build a Hi-En code-switched conversation summarization dataset. We first considered the option of creating summaries for an existing code-switched conversational dataset like DSTC2 (Banerjee et al., 2018). Though this dataset is substantially large with over 50,000 utterances, it is not open-domain and only contains human-computer conversations focused on restaurant reservations, thus lacking linguistic diversity. We, therefore, chose the option of manually translating the SAMSum corpus (Gliwa et al., 2019) from En to Hi-En.

Annotation Process - We hired eight annotators who are fluent in both Hindi and English. We first explained the concept of code-switching and provided them with a few reference examples annotated by the authors. Based on our interactions with the annotators, we observed that Hi-En code-switching was an integral part of their vernacular. They also frequently used code-switching on social media and chat applications.

We first provided each annotator with a random sample of ten conversations. We instructed them to first go through the conversation and the corre-

Karen: Hey guys! Is anyone in the office. I forgot my key... :/ **John**: I'll be there in 1 hour. **Patrick**: Oh no! I'm sorry, can't help you. I'm out of office today. **Mary**: Are you by the entrance? I should be there soon. **Karen**: Thanks Mary, yes, I'm here. **Mary**: I think I see you. 2 minutes I'm there. **Karen**: Thanks a lot!	**Karen**: *Hey guys! Kya koi office mein hai. Main apni chaabi bhool gayi... :/* **John**: *Main waha pe hounga in 1 hour.* **Patrick**: *Oh no! I'm sorry, main help nahi kar sakta. Maine office se bahar hu aaj.* **Mary**: *Kya tume entrance pe ho? I should be there soon.* **Karen**: *Thanks Mary, yes, main yaha hu.* **Mary**: *I think main tumhe dekh sakti hu. 2 minutes, main waha hu* **Karen**: *Thanks a lot!*
Karen forgot the key to the office. Mary will be there soon to let her in.	*Karen apni chaabi bhool gayi office ki. Mary waha pe hogi thodi der mein usko andar aane ke liye.*

Table 3: Sample conversation and summary from SAMSum (Gliwa et al., 2019) (first column) and the corresponding Hi-En code-switched conversation and summary in GupShup (second column).

sponding summary in English and then translate the content to Hi-En, assuming it was an interaction between themselves and their friends. However, they could not introduce or remove any utterances during translation but rather do an utterance by utterance translation. We asked the annotators to transcribe the resulting conversations only in Romanized text: transliterate the Hindi words. They used the same process for translating the summaries.

After the annotators completed translating the initial random samples, we provided feedback in terms of format and organization of the data. Once they were comfortable with the process, we assigned them random batches of conversations, and they worked independently based on their schedules. The contributions of each annotator were mainly driven by their availability. Due to time and resource constraints, we only had one translation for a given source conversation. The entire annotation process lasted for around three months, at the end of which we translated 6,831 conversations containing 76,330 utterances. To our knowledge, this is also the largest parallel corpus for Hi-En and En languages containing 109,346 sentences, out of which 48,578 are code-switched.

Table 3 shows a sample conversation and summary in English and the corresponding code-switched translations. As demonstrated in this example, the annotators preserved the use of punctuation and emojis in the original text. This sample also demonstrates different types of code-switching patterns. For example, in the first utterance, an English term *office* is inserted in the middle of a transliterated Hindi sentence. The same applies to the phrase *1 hour* in the second utterance. In the fourth utterance, the speaker *Mary* switches from Hindi to English in the middle of an utterance.

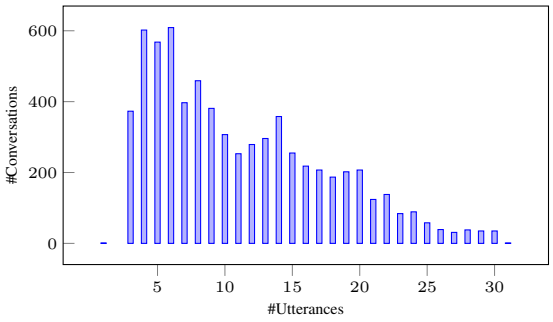

Figure 1: Distribution of number of utterances per conversation.

In the last utterance, the speaker *Karen* switched entirely to English though they initiated the conversation in Hi-En.

Hindi and many other Indian languages exhibit inflection of verbs based on the gender of the speaker (Kumar et al., 2019). For example, the sentence *I read*, when said by someone of masculine gender in Hindi, would be: *main padhtaa hoon*, but when said by someone of the feminine gender would be: *main padhtee hoon*. The verb *padh* was inflected by the gender of the speaker. During the annotation process, we did not provide any explicit instructions about this inflection, but the annotators used the speaker's names or conversational context to derive the gender and used that information for translation. For example, the term *hounga* in the second utterance of the conversation in Table 1 is a reflection of John's perceived masculine gender. Likewise, the term *sakti* in the sixth utterance reflects Mary's perceived gender.

4 Corpus Analysis

Figure 1 shows the distribution of the number of utterances per conversation. Gupshup has 76,330

utterances, where the shortest conversation has one utterance and the longest has 31 utterances. The average length of a conversation is 11.17 utterances. The English version of the corpus had 28.6 words per utterance and 19,252 unique words, whereas GupShup has 31.1 words per utterance and 25,865 unique words. A large portion of the data (73.6% of conversations) had only two participants, 18.7% of the conversations had 3 participants, 5.2% had 4 participants, and the remaining conversations had more than 4 participants.

Language Tagging - Since GupShup is a code-switched corpus, it is essential to analyze and quantify the complexity of code-switching in the corpus, which requires us first to identify the language associated with each token. Since this is a parallel corpus, we used information from the source utterances to determine non-English tokens in the translated utterance. More specifically, for a given pair of utterances, we first removed all the punctuation marks, emojis, and numerical tokens. We then identified named entities in the English utterance, and if these entities also appeared in the translated utterance, the corresponding tokens were excluded from language tagging. Of the remaining tokens, if any of them appeared in the English utterance, we tagged them as English.

We repeated the same process on the remaining tokens but with their lemmatized versions because we observed that sometimes the annotators used the English word from a source utterance in a different part of speech. For e.g. the phrase *"a few years after we graduated"* was translated to *"graduation ke kuch saal baad"*. Here graduation can be accurately tagged as English using lemmatization. Lastly, the remaining tokens in the Hi-En utterance were tagged as Hindi.

We observed that this process captured most English tokens in the translated utterances, except when the annotators introduced a new English token. For example, one annotator translated the phrase *"I have been thinking for a while"* to *"Main kaafi time se soch rahi hu"*, where they introduced the token *time*. However, we observed that this was a rare phenomenon and did not sway the analysis significantly.

Code-switching statistics - Per the language tagging approach described above, 18.15% (13,760) of the utterances were entirely in transliterated Hindi, 23.49% (17,804) were entirely in English. The majority of the utterances, 58.86% (43,407), were code-switched: had a combination of Hindi and English tokens. Table 4 has more detailed code-switching statistics of the corpus.

Vocabulary size	26,865
English vocabulary	11,616
Hindi vocabulary	12,016
Total utterances	76,330
Unique utterances	75,791
Code-switched utterances	43,407
Hindi utterances	13,760
English utterances	17,804
Avg. length of utterance	10.07
Avg. # of code-switched utterances per conversation	6.35
# of code-switched utterances with Hindi matrix	45,644
# of code-switched utterances with English matrix	2,934
# of Hindi insertions into English Matrix	2,810
# of English insertions in Hindi Matrix	38,539

Table 4: Code-switching statistics of the GupShup dataset.

We determine the matrix language (Myers-Scotton et al., 2002)[4] of a sentence using the heuristics proposed in (Dhar et al., 2018). Namely, we define a sentence as Hindi if (a) the majority of tokens are Hindi, (b) we detect the use of any Romanized Hindi verbs, or (c) we detect the use of Romanized Hindi bi-grams. Per this definition on the 43,407 code-switched utterances, the matrix language of 45,644 sentences was Hindi, and 2,934 sentences were in English.

Based on the matrix language and token-level language tags, we further analyzed the code-switching complexity using the metrics C_{avg} (Gambäck and Das, 2016), C_c (Banerjee et al., 2018), and I-index (Guzman et al., 2016). These metrics quantify complexity in terms of the number of foreign language tokens and switch points. On our corpus, we estimated that $C_c = 63.25$, $C_{avg} = 13.57$, and I-index was 0.14. For reference, (Gambäck and Das, 2016) applied these metrics to the different code-switching datasets from (Solorio et al., 2014) and observed that English-Nepalese corpus had the highest levels of switching with C_c value of 49.06 and C_{avg} value of 7.98. These metrics show that GupShup has high levels of code-switching complexity.

5 Empirical Benchmarks

In our experimental work, we employ the following six abstractive summarization models: GPT-2, BART, PEGASUS, T5, multitask T5, mBART,

[4]Matrix language represents the underlying language choice, therefore, driving the grammatical structure of a sentence

and multi-view seq2seq model (Chen and Yang, 2020). Most of these transformer models were pre-trained on English corpora, except for mBART, which was trained on multilingual data. We expect our results to serve as empirical benchmarks for future researchers. We used 5,831/500/500 conversations as our train, dev, and test splits, respectively. We trained all the models for three epochs, with evaluation on the dev set after each epoch. Due to constraints on computing resources, we used a smaller mBART model with a 12-layer encoder and a 6-layer decoder. For the T5 model[5], we tried both fine-tuning and multitask learning approach (T5 MTL); for the latter approach, we trained the model for both summarization and translation. We ran this entire process for three experimental setups: (1) En summaries from Hi-En conversations, (2) En summaries from En conversations, and (3) Hi-En summaries from Hi-En conversations. All models were trained using Huggingface's transformer library on Google colab GPU enabled platform. The model performances are reported in terms of the following automatic evaluation metrics: ROUGE (R1, R2, RL) (Lin, 2004), BLEURT (Sellam et al., 2020), BERT-score (Zhang et al., 2020), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005).

Results - The results are summarized in Table 5. In generating English summaries from Hi-En conversations, the mBART model obtained the best R1 and R2 scores; this is likely because it is the only model to have been trained on multiple languages and therefore better understands code-switching points in the conversation. The multi-view model obtained the best RL, BLEURT, and BLEU scores. Though this model was pre-trained only on English data, it explicitly extracts conversational structure to help with the summary generation.

The T5 MTL model outperformed the T5 model across all metrics and also achieved the best METEOR score. We applied the Wilcoxon-signed rank test for the ROUGE-1 scores of the two T5 models and obtained a p-value of 7.1e-16 at a confidence level of 5%, indicating a statistically significant difference. This observation suggests that summarization task would benefit from translation task due to the multi-lingual and parallel nature of the dataset. Therefore, we expect the other models could also benefit from a multi-task learning setup.

Model	R1	R2	RL	BLEURT	BERTScore	BLEU	METEOR
Hi-En conversations ⟶ En summaries							
mBART	**43.14**	**16.83**	33.87	-0.46	0.90	9.96	26.74
Multi-view	41.21	16.16	**39.80**	**-0.43**	0.90	**11.45**	28.75
PEGASUS	41.60	15.72	32.91	-0.44	0.90	8.91	25.55
T5 MTL	40.84	15.50	30.38	-0.47	0.90	11.05	**30.8**
T5	37.52	12.60	27.55	-0.56	0.89	8.12	26.92
BART	39.75	14.09	31.52	-0.52	0.90	6.92	23.73
GPT2	13.52	2.59	10.5	-1.03	0.84	2.21	12.05
En conversations ⟶ En summaries							
mBART	49.69	24.36	40.40	-0.35	0.91	14.37	32.44
Multi-view	**50.65**	25.04	40.13	-0.30	0.92	**18.34**	**39.04**
PEGASUS	50.53	**25.77**	**41.94**	**-0.28**	0.92	17.47	36.64
T5	45.62	21.65	35.25	-0.46	0.91	14.95	38.24
BART	46.47	21.79	37.74	-0.39	0.91	14.37	32.44
GPT2	15.78	5.42	13.58	-0.98	0.84	2.78	14.75
Hi-En conversations ⟶ Hi-En summaries							
mBART	19.98	2.89	16.7	-0.88	0.83	1.60	10.42
Multi-view	21.92	4.55	18.16	-0.83	0.85	2.27	9.94
PEGASUS	35.69	11.01	28.78	-0.72	0.86	**6.16**	20.91
T5	31.85	7.90	24.13	-0.72	0.86	5.51	20.6
BART	**36.28**	**11.45**	**28.92**	**-0.70**	**0.87**	5.96	**21.82**

Table 5: Performance of all the models across various metrics for three experimental setups.

In the second experimental setup, where English summaries were generated from English conversations, we observed a significant improvement for all the models and across all the metrics. This observation is expected as most of these models have been pre-trained on English data. The multi-view model obtained the best R1, BLEU, and METEOR scores, whereas the PEGASUS model obtained the best R2, RL, and BLEU scores. Once again mBART obtained better scores than BART, but the difference is less pronounced than for Hi-En conversations. It is unclear if the superior performance of mBART is due to multilinguality or because it was pre-trained on more English data. In the future, we would like to explore the use of Devanagari script for the portions of the conversations that are in Hindi as opposed to the Romanized script since models like mBART have been exposed to Devanagari script.

In both these experimental setups, we observed that GPT-2 model obtained the lowest scores, but further analysis helped us realize that the summaries generated by this model were not semantically meaningful and often contained repeated words. This could be attributed to the model's architecture which contains only a decoder and can-

[5]We also trained mT5. However, even after training for 30 epochs with early stopping we barely got a ROUGE-1 score of 31.06, ROUGE-2 score of 9.41 and ROUGE-L score of 25.33, which were lesser than what we obtained using T5. On manually analyzing the results, we observed that the summaries produced by mT5 were not grammatically sound, there were lot of repetitions, the entities mentioned in the summaries were wrong, and relevant points from the conversations were missing in the summaries. This is very surprising. Since we decided to keep only the top performing models in the final results table and could not understand the reason behind such poor performance of mT5 we did not pursue it further.

not, therefore, create a meaningful representation of the source conversation (Rothe et al., 2020). Additionally, we also observed minimal variance in BERTScore values across the models, suggesting it may not be the most effective metric for measuring the quality of summaries.

In the third experimental setup, we generated Hi-En summaries from Hi-En conversations. Here we observed a steep drop in performance of all the models when compared to generating English summaries. However, this drop was less prominent for BART, which obtained the best scores across most metrics. Another interesting observation is that, despite being multilingual, mBART generated poor Hi-En summaries. Overall, from this experiment, we observe that most language models may not be capable of generating Hi-En summaries.

Human Evaluation - We conducted a human evaluation of the English summaries generated by all the models for a random selection of 100 Hi-En conversations. Following (Fabbri et al., 2020), we had three annotators rate the summaries for the following four metrics[6]: consistency, coherence, fluency, and relevance on a Likert scale of 1 to 5. The results, average Likert score normalized to the range 0 to 1, are summarized in Table 6.

The multi-view model obtained the best scores for consistency and relevance, whereas the PEGASUS model obtained the best scores for coherence and fluency. The mBART model also obtained comparable scores. These observations are reasonably consistent with the numbers in Table 5 except for PEGASUS, which may not have obtained the highest scores in automatic metrics but generated more coherent and fluent summaries.

Model	Coherence	Consistency	Fluency	Relevance
mBART	0.81	0.63	0.85	**0.65**
Multi-view	0.83	**0.65**	0.86	**0.65**
PEGASUS	**0.84**	0.58	**0.87**	0.60
T5 MTL	0.70	0.55	0.77	0.55
T5	0.67	0.532	0.77	0.54
BART	0.73	0.56	0.74	0.55
GPT2	0.37	0.36	0.52	0.33

Table 6: Human evaluation of En summaries generated from Hi-En conversations.

Given the variety of automatic metrics for measuring the quality of summaries, we also wanted to understand how they correlate with human judgments. Table 7 summarizes the Spearman correlation between all the automatic metrics and human evaluations. All the ROUGE-based metrics and BLEURT were strongly correlated with human evaluations, but METEOR had a very low correlation. Perhaps, as a next step, we could consider fine-tuning BLEURT on this dataset to serve as a measure of quality for summaries from code-switching conversations.

Metric	R1	R2	RL	BLEURT	BERTScore	BLEU	METEOR
Coherence	0.86	0.82	0.89	0.93	0.80	0.57	0.18
Consistency	0.86	0.93	1.00	0.93	0.80	0.71	0.32
Fluency	0.86	0.82	0.79	0.93	0.67	0.75	0.46
Relevance	0.89	0.96	0.96	0.96	0.80	0.86	0.50
Overall	0.89	**0.96**	**0.96**	**0.96**	0.80	0.86	0.50

Table 7: Correlation between human judgements and automatic metrics for En summaries generated from Hi-En conversations.

6 Qualitative Analysis

Error Analysis - To get a better insight into the generated summaries, we chose a random sample of 100 conversations both English and Hi-En code-switched, and analyzed English summaries generated by all the models. We classified each summary into one or more of the following four error categories: (a) missing information (MI): the generated summary is missing a salient aspect mentioned in the reference summary, (b) erroneous reference (ER): one of the speakers is incorrectly associated with an action or location, (c) incorrect inference (II): summary contains a statement that cannot be inferred or reasoned from the conversation, and (d) wrong pronouns (WP): summary misrepresents a speaker's gender. These classes were not predetermined, but grew out of the error analysis. Table 8 shows an example of a summary generated from Hi-En conversation that has different types of errors. The results are summarized in Table 9.

As with automatic metrics (Table 5), the error analysis also shows that the models summarize English conversations better. We observe that all models miss important information when summarizing; however, this was more pronounced with Hi-En conversations suggesting code-switching adds more complexity to the summarization process. Interestingly, the T5 model has the fewest MI errors when summarizing English conversations. The multi-view model had the fewest ER errors, likely because it explicitly tracks speakers and therefore

[6]More details on the definitions of the metrics can be found in (Fabbri et al., 2020).

Randy: Natalie, we're dead, look what happened!?
Natalie: What's wrong?
Randy: <file_video>
Natalie: No way, how did Sally and Molly get to the parent's bedroom?!!
Randy: I've no idea, someone must have left the door open
Natalie: It looks really nasty! Why are all the sheets so dirty?
Randy: I think they both went to play in the mud and then somehow ended up here
Natalie: Jesus Christ, mum is going to be really mad! Start cleaning it
Randy: I know, but it's hopeless, parents might here any minute
Natalie: Just do it, I'll be back as soon as I can!

Reference summary: Sally and Molly possibly played in the mud and got to parent's bedroom, where they made a mess. Randy will start cleaning and Natalie will join him as soon as she can, because parents might be here any minute.
Summary from Hi-En conversation: Natalie's parents are dead. Randy thinks Sally and Molly are in their parents bedroom. Natalie thinks they're dead.
Summary from En conversation: Sally and Molly got into the parent's bedroom. The sheets are dirty. Randy thinks they went to play in the mud and then ended up here.

Table 8: Example of a generated summary with multiple errors.

Error Type	mBART	Multiview	Pegasus	T5	T5 MTL	BART	Avg
Hi-En conversations ⟶ En summaries							
MI	77	75	81	69	63	87	75.33
ER	41	21	37	77	49	48	45.50
II	82	69	78	87	75	76	77.83
WP	2	0	0	1	1	1	0.83
En conversations ⟶ En summaries							
MI	72	62	67	43	n/a	79	64.60
ER	49	31	11	43	n/a	48	36.40
II	58	27	42	42	n/a	70	47.80
WP	6	2	0	2	n/a	1	2.20

Table 9: Distribution of error types in summaries generated by all models.

less likely to associate them with incorrect actions. The II errors have increased significantly for all the models when summarizing Hi-En conversations, further demonstrating the inability of the models to understand code-switched conversations. We noticed that even though the models select salient aspects of the conversations, the summaries have incorrect inferences because some of the critical tokens in the conversation were code-switched.

Summary Compression - We measured the compression ratio of the summaries generated by all models for both En and Hi-En conversations as show in figure 2. Per (Grusky et al., 2018b), compression ratio is defined as the ratio of the number of tokens in the conversations to that of the summaries. We observe that all the models seem to generate shorter summaries for Hi-En conversations. This is understandable considering that most of these models are pre-trained on English data; therefore, when exposed to Hi-En conversations, the English content becomes less frequent and fragmented, resulting in the models under-generating or even stopping early. Compared to reference summaries, BART, mBART, and PEGASUS seem to generate more compressed summaries, whereas T5 and GPT-2 generate longer summaries. This could also explain why T5 has fewer MI errors.

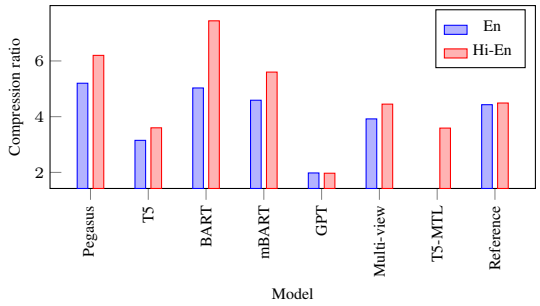

Figure 2: Compression ratios of summaries generated by all the models.

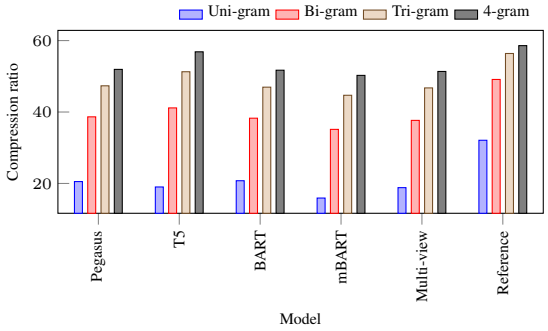

Figure 3: Proportion of the new n-grams generated by each model wrt source conversations.

Quantifying Abstractiveness - During our error analysis, we also observed that the summaries generated from English conversations were often more extractive than abstractive, where the models were selecting the most salient phrases in the source conversations. Following (Narayan et al., 2018b), we quantified the *abstractiveness* of the summaries by calculating the portion of new n-gram generated by the models for English. Namely, we report the ratio of n-grams that were part of the summaries but not the source conversations to the total number of n-grams in the summaries. When compared to reference summaries, all the models seem to be more extractive in nature. Of the five models here, T5 generated the most novel n-grams despite not introducing many new uni-grams, which is consistent with the low compression ratio of T5.

Model Diversity - Since we have employed different language models in our work, we wanted to understand how similar are the summaries gener-

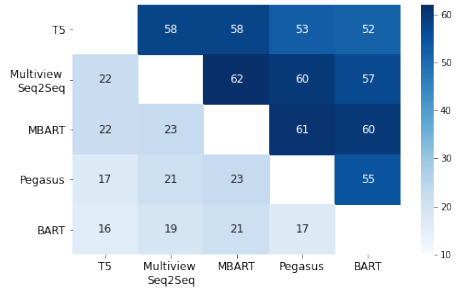

Figure 4: ROUGE-1 (upper triangular) and ROUGE-4 (lower triangular) scores for summaries generated by each pair of models from English conversations.

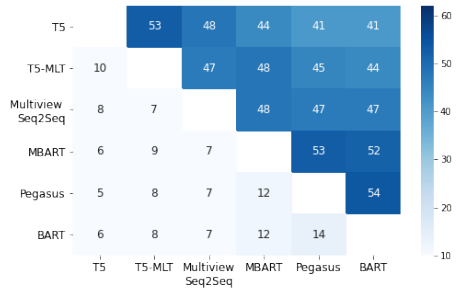

Figure 5: ROUGE-1 (upper triangular) and ROUGE-4 (lower triangular) scores for summaries generated by each pair of models from Hindi-English conversations.

ated by these models. To this end, we measured the ROUGE-1 and ROUGE-4 scores between the summaries from each pair of models. The results are summarized in Figure 5. We observe that the ROUGE-1 scores are significantly higher than ROUGE-4 scores for both En and Hi-En conversations suggesting that the models seem to be choosing a similar set of tokens for summarization. However, the word order is very different, thus suggesting that the eventual summaries are different. Table 10 shows an example of one such conversation and its summaries. Also, the ROUGE-1 and ROUGE-4 scores between the model are higher than w.r.t the reference summaries. These findings are similar to those reported in (Kryscinski et al., 2019). Lastly, as with reference summaries, the pairwise ROUGE scores are higher for En than Hi-En conversations.

7 Conclusions and Future Work

In this work, we presented the first code-switched conversation summarization dataset - *GupShup*, which has over 6,800 multi-party open-domain conversations and their corresponding summaries in En and Hi-En. We quantified the performance of various state-of-the-art neural models using both automatic metrics and human evaluation. mBART and multi-view models obtained the best scores on automatic metrics, whereas PEGASUS and mBART obtained best scores on human judgement. The T5 model generated relatively longer summaries and was therefore effective in capturing all the salient aspects of the conversation. ROGUE-based metrics and BLEURT were highly correlated with the human judgements but METEOR, BERTScore, and BLEU proved relatively ineffective for this task. We also observed that multi-task learning setup showed promise and we would like to explore this further.

We conducted an extensive qualitative analysis of the summaries, which provided very interesting insights into different language models. In the future, we would like to explore new techniques for overcoming the various challenges that we observed in this work and would like to dive deeper in understanding the failures of different language models when applied to code-switched text.

Clark: hi, Tumhaari profile kaafi interesting hai **Louise**: thank you :) **Clark**: yaha jaada John Cage ke fans nazar nhi aate **Louise**: mujhe uske music ke fans khi bhi nazar nhi aate **Clark**: sahi baat **Clark**: yaha pr Saturday ko ek gig hai **Louise**: tribute to John Cage **Louise**: ha mujhe pta hai! **Clark**: do you wanna go? **Louise**: mene phle hi ticket leli hai so sure :) **Clark**: great! :)
Reference Summary: *Clark and Louise are John Cage's fans. They will meet at the tribute concert on Saturday. Louise already has her ticket.* **BART**: Clark and Louise will go to John Cage's tribute gig on Saturday. **MBART**: Clark and Louise are going to a tribute to John Cage gig on Saturday. **Multiview**: Louise is going to John Cage's gig on Saturday. Clark is going too. **Pegasus**: Clark and Louise are going to a gig on Saturday to pay tribute to John Cage. **T5**: Clark and Louise are going to John Cage's tribute gig on Saturday on Saturday. They have a ticket for the gig, so they're going to buy it. **T5_MTL**: Clark and Louise want to go to a tribute to John Cage gig on Saturday. Louise has a ticket for the gig.

Table 10: Example of a conversation where the models generated summaries have high pairwise ROUGE-1 scores but low ROUGE-4 scores.

Acknowledgements

Rajiv Ratn Shah is partly supported by the Infosys Center for AI and the Center of Design and New Media at IIIT-Delhi.

References

Emily Ahn, Cecilia Jimenez, Yulia Tsvetkov, and Alan Black. 2020. What code-switching strategies are effective in dialogue systems? *Proceedings of the Society for Computation in Linguistics*, 3(1):308–318.

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2016. Can we detect speakers' empathy?: A real-life case study. In *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000059–000064. IEEE.

Peter Auer. 2013. *Code-switching in conversation: Language, interaction and identity*. Routledge.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "I am borrowing ya mixing ?" an analysis of English-Hindi code mixing in Facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126, Doha, Qatar. Association for Computational Linguistics.

S. Banerjee, S. Naskar, P. Rosso, and Sivaji Bandyopadhyay. 2016. The first cross-script code-mixed question answering corpus. In *MultiLingMine@ECIR*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Suman Banerjee, Nikita Moghe, Siddhartha Arora, and Mitesh M. Khapra. 2018. A dataset for building code-mixed goal oriented conversation systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3766–3780, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014a. Code mixing: A challenge for language identification in the language of social media. In *CodeSwitch@EMNLP*.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014b. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23, Doha, Qatar. Association for Computational Linguistics.

Anshul Bawa, Monojit Choudhury, and Kalika Bali. 2018. Accommodation of conversational code-choice. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 82–91, Melbourne, Australia. Association for Computational Linguistics.

Irshad Ahmad Bhat, R. Bhat, Manish Shrivastava, and D. Sharma. 2018. Universal dependency parsing for hindi-english code-switching. In *NAACL-HLT*.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

K. Chakma and Amitava Das. 2016. Cmir: A corpus for evaluation of code mixed information retrieval of hindi-english tweets. *Computación y Sistemas*, 20:425–434.

Khyathi Raghavi Chandu, Manoj Kumar Chinnakotla, and Manish Shrivastava. 2015. "answer ka type kya he?": Learning to classify questions in code-mixed language. In *WWW '15 Companion*.

Khyathi Raghavi Chandu, E. Loginova, Vishal Gupta, J. Genabith, G. Neumann, Manoj Kumar Chinnakotla, Eric Nyberg, and A. Black. 2018. Code-mixed question answering challenge: Crowd-sourcing data and techniques. In *CodeSwitch@ACL*.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed Indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387, Goa, India. NLP Association of India.

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.

A. R. Fabbri, Wojciech Kryscinski, B. McCann, R. Socher, and D. Radev. 2020. Summeval: Re-evaluating summarization evaluation. *ArXiv preprint*, abs/2007.12626.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *ArXiv preprint*, abs/2107.03175.

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).

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742. IEEE.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018a. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018b. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

John J Gumperz. 1977. The sociolinguistic significance of conversational code-switching. *RELC journal*, 8(2):1–34.

John J Gumperz. 1982. *Discourse strategies*, volume 1. Cambridge University Press.

Vishal Gupta, Manoj Chinnakotla, and Manish Shrivastava. 2018. Transliteration better than translation? answering code-mixed questions over a knowledge base. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 39–50, Melbourne, Australia. Association for Computational Linguistics.

Gualberto A Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. In *INTERSPEECH*, pages 67–71.

Gualberto A. Guzman, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2016. Simple tools for exploring variation in code-switching for linguists. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 12–20, Austin, Texas. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2015. Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 239–248, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2016. Collecting and annotating indian social media code-mixed corpora. In *CICLing*.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Mahesh Joshi and Carolyn Penstein Rosé. 2007. Using transactivity in conversation for summarization of educational dialogue. In *Workshop on Speech and Language Technology in Education*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020. 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.

Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. 2017. The rise of bots: A survey of conversational interfaces, patterns, and paradigms. In *Proceedings of the 2017 conference on designing interactive systems*, pages 555–565.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Yaman Kumar, Debanjan Mahata, Sagar Aggarwal, Anmol Chugh, Rajat Maheshwari, and Rajiv Ratn Shah. 2019. Bhaav-a text corpus for emotion analysis from hindi stories. *ArXiv preprint*, abs/1910.04073.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Deuchar Margaret, Peredur Davies, Jon Russell Herring, M Carmen Parafita Couto, and Diana Carter. 2014. Building bilingual corpora. *Advances in the study of bilingualism. Bristol: Multilingual Matters*, pages 93–110.

Yashar Mehdad, Giuseppe Carenini, and Raymond T. Ng. 2014. Abstractive summarization of spoken and written conversations based on phrasal queries. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1220–1230, Baltimore, Maryland. Association for Computational Linguistics.

Pieter Muysken, Pieter Cornelis Muysken, et al. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge University Press.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Carol Myers-Scotton et al. 2002. *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford University Press on Demand.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task @icon-2017. *ArXiv*, abs/1803.06745.

Jasabanta Patro, Bidisha Samanta, Saurabh Singh, Abhipsa Basu, Prithwish Mukherjee, Monojit Choudhury, and Animesh Mukherjee. 2017. All that is English may be Hindi: Enhancing language identification through automatic ranking of the likeliness of word borrowing in social media. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2264–2274, Copenhagen, Denmark. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Vinay Singh, Deepanshu Vijay, S. Akhtar, and Manish Shrivastava. 2018. Named entity recognition for hindi-english code-mixed social media text. In *NEWS@ACL*.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *ArXiv preprint*, abs/1904.00784.

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.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhou Zhao, Haojie Pan, Changjie Fan, Yan Liu, Linlin Li, and Min Yang. 2019. Abstractive meeting summarization via hierarchical adaptive segmental network learning. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3455–3461. ACM.

A Appendix

A.1 Examples of model generated summaries

Table 11, 12, 13, 14, 15 and 16 show the top 3 generated summaries by models trained on summarizing Hi-En code-switched conversations to En summaries.

Peter: Kya mai tumhari car borrow kar sakta hu? **Hugh**: sure **Hugh**: but tumhari car ke saath kya hua? **Peter**: pata nahi **Peter**: aur time nahi hai check karne ka **Peter**: mai wese bhi late hu! **Hugh**: ok, ok, aake lele
Pegasus generated summary: Peter will borrow Hugh's car. Gold summary: Peter will borrow Hugh's car.
Monica: Kaha ho tum? **Monica**: Mai tumhe dekh ni pa rhi. **Lexie**: Bs yahi hu. **Monica**: OK, waiting.
Pegasus generated summary: Monica is waiting for Lexie. Gold summary: Monica is waiting for Lexie.
Agatha: meri book read karna khatam kar diya? **Johnny**: abhi nahi **Agatha**: I see
Pegasus generated summary: Johnny hasn't read Agatha's book yet.
Gold summary: Johnny hasn't finished reading Agatha's book yet.

Table 11: Top 3 summaries generated by Pegasus from Hi-En code-switched conversations, in terms of ROUGE-1 scores.

A.2 Correlation between model performances in terms of ROUGE-1 and dataset attributes.

Fig.7 shows the correlation between the several conversation-summary pair attributes and the performance of the models for generating English summaries from Hi-En conversations. We notice that the only positive correlation that exists in the figure is for of English tokens in the conversation to be summarized.This makes sense because268more English content in the code-switched conversations

Scott: Hume kaha milna chahiye? **John**: at Oculus? **Scott**: ok! 7.30 baje **John**: yup!
mBART generated summary: Scott and John will meet at Oculus at 7.30. Gold summary: Scott and John will meet at 7.30 at Oculus.
Ralph: Tum abhi bhi yahi kahi ho? **Mary**: Bathroom! **Ralph**: Oh! TMI! **Mary**: ek sec mein aa raha huu... **Ralph**: jitna time chaiye utna lo...
mBART generated summary: Mary is in the bathroom. Gold summary: Mary is in the bathroom.
Peter: Kya mai tumhari car borrow kar sakta hu? **Hugh**: sure **Hugh**: but tumhari car ke saath kya hua? **Peter**: pata nahi **Peter**: aur time nahi hai check karne ka **Peter**: mai wese bhi late hu! **Hugh**: ok, ok, aake lele
mBART generated summary: Peter will borrow Hugh's car. Gold summary: Peter will borrow Hugh's car.

Table 12: Top 3 summaries generated by mBART from Hi-En code-switched conversations, in terms of ROUGE-1 scores

Scott: Hume kaha milna chahiye? **John**: at Oculus? **Scott**: ok! 7.30 baje **John**: yup!
Multi-view Seq2Seq generated summary: John and Scott will meet at Oculus at 7.30. Gold summary: Scott and John will meet at 7.30 at Oculus.
Jude: Tumhara wallet kaha hai mujhe kahi nahi mil raha **Faith**: maine apne saath le liya tha **Jude**: kyu? I need your credit card to pay the bills
Multi-view Seq2Seq generated summary: Jude needs Faith's credit card to pay the bills. Gold summary: Jude needs Faith's credit card to pay the bills.
Emir: Etna ki financial statement bhej sakti ho? **Britta**: Sure, konsa saal? **Emir**: 2017 **Britta**: Ok **Emir**: English mei please
Multi-view Seq2Seq generated summary: Emir will send Britta Etna's financial statement in English. Gold summary: Britta will send Emir Etna's 2017 financial statement in English.

Table 13: Top 3 summaries generated by Multi-view Seq2Seq from Hi-En code-switched conversations, in terms of ROUGE-1 scores

will help the models in making sense of the conversations, as they are all trained on only English corpus except for mBART. Its worth noting that mBART, which is our only multilingual model has a negligible positive correlation of 0.07, suggesting it's probably using other signals as well, to make sense of the code-switched conversations benefiting from its multilingual pretraining.

Fig.6 shows the correlation between the several conversation-summary pair attributes and the performance of the models for generating English summaries from En conversations. We see that PEGASUS, BART, mBART, and Multi-view Seq2Seq are all negatively correlated to *avg. utterance length of the conversation, number of turns in a conversation, number of turns/number of participants of a conversation, length of the conversation, and the compression ratio of the gold summary with respect to the conversations to be summarized.* T5 largely remains uncorrelated with any of the conversation-summary attributes.

A.3 Sample for the Error Analysis performed in section 6

Table 17 shows the difference in the errors in summaries generated from the same conversation, one from the respective En conversation and one from Hi-En code-switched conversation, as mentioned in section 6.

A.4 Hyperparameters for models trained for summarization of Hi-En code-switched conversations to En summaries.

Table 18 reports the hyperparameters of the models trained for summarization of Hi-En conversations to English summaries.

Jamal: < $file\_photo$ > **Terry**: Taj Mahal! **Maria**: Yes, we visited it today with Jamal **Ken**: yeh bahut sundat mosque hai! **Maria**: it's not a mosque! **Ken**: what? **Maria**: it's a mausoleum **Ken**: mujhe hamesha lagta tha it's a mosque **Jamal**: bahut logo ko lagta hai **Maria**: it is a mausoleum that an emperor commissioned for his favourite wife **Maria**: shayad uska naam Mumtaz Mahal that **Jamal**: correct! :D what a good pupil! **Maria**: haha, because it's such a romantic story **Maria**: 20000 logo ne Taj Mahal banaya, it's so monumental **Ken**: iss naam kaa kya matlab hai? **Maria**: Taj is a short version of Mumtaz **Maria**: and Mumtaz Mahal means "Crown of the Palace" **Ken**: wow **Maria**: Jamal was an amazing guide today **Ken**: I wish I was there with you
T5 generated summary: Maria and Jamal visited the Taj Mahal today. The mausoleum was commissioned by an emperor for his favourite wife, Mumtaz. Gold summary: Maria and Jamal visited Taj Mahal today. It's a mausoleum that an emperor commissioned for his wife Mumtaz Mahal.
Lidia: Maine bahi tumhe office jaate hue dekha. What the hell are you wearing? Is it shells? **Brooke**: Bas ek purana brooch hai jo maine apni dadi se inheritkiya tha. Acha hai naa? **Lidia**: You crazy woman! Bekar hai. Mujhe nahi pata log iske baare mein kya kahenge. **Brooke**: Mujhe farak nahi padta log kya kahenge. For me it's truly wonderful! **Lidia**: Wanna go for a cup of coffee? **Brooke**: Sure. kitchen mein milte hai in 5 minutes, okay? **Lidia**: Okay
T5 generated summary: Lidia and Brooke are going to the office in 5 minutes. Brooke and Lidia are going for a cup of coffee. Lidia will be in the kitchen in 5 min. Gold summary: Brooke wears a brooch in the office that she inherited from her grandmother. Lidia and Brooke will meet in the kitchen in 5 minutes to go for a cup of coffee.
Robert: Hi Matt **Matt**: Hi! **Robert**: How are you doing? **Matt**: good, thanks, aur tum? **Robert**: mai bhi thk hu. **Robert**: kal dubara tumse milkr bhut acha lga, although by chance **Matt**: yes, bhut mjaa aya tha, I didn't expect to meet you on the subway **Robert**: me neither **Matt**: Maybe we should meet again, and not on the subway? **Robert**: I'd really like to **Matt**: tum kab free ho? **Robert**: har Mondays and Tuesdays **Matt**: So maybe Monday evening? **Robert**: great **Matt**: hum sath me dinner kar sakte hai **Robert**: theek hai esa hi krenge **Matt**: kya mai tumhe pick krne auu after you're done? **Robert**: haan yeh theek rhga ! **Matt**: ok,toh I'll be there at 7PM **Robert**: good!
T5 generated summary: Robert will meet Matt on the subway on Mondays and Tuesdays. Matt will pick him up at 7PM. Robert will be there at 7pm.
Gold summary: Matt and Robert will meet on Monday at 7PM. Matt will pick Robert up. They will have a dinner together.

Table 14: Top 3 summaries generated by T5 from Hi-En code-switched conversations, in terms of ROUGE-1 scores.

Jamal: < $file\_photo$ > **Terry**: Taj Mahal! **Maria**: Yes, we visited it today with Jamal **Ken**: yeh bahut sundat mosque hai! **Maria**: it's not a mosque! **Ken**: what? **Maria**: it's a mausoleum **Ken**: mujhe hamesha lagta tha it's a mosque **Jamal**: bahut logo ko lagta hai **Maria**: it is a mausoleum that an emperor commissioned for his favourite wife **Maria**: shayad uska naam Mumtaz Mahal that **Jamal**: correct! :D what a good pupil! **Maria**: haha, because it's such a romantic story **Maria**: 20000 logo ne Taj Mahal banaya, it's so monumental **Ken**: iss naam kaa kya matlab hai? **Maria**: Taj is a short version of Mumtaz **Maria**: and Mumtaz Mahal means "Crown of the Palace" **Ken**: wow **Maria**: Jamal was an amazing guide today **Ken**: I wish I was there with you
T5-MTL generated summary: Jamal visited the Taj Mahal today. It's a mausoleum that an emperor commissioned for his favourite wife, Mumtaz Mahal. Gold summary: Maria and Jamal visited Taj Mahal today. It's a mausoleum that an emperor commissioned for his wife Mumtaz Mahal.
Lidia: Maine bahi tumhe office jaate hue dekha. What the hell are you wearing? Is it shells? **Brooke**: Bas ek purana brooch hai jo maine apni dadi se inheritkiya tha. Acha hai naa? **Lidia**: You crazy woman! Bekar hai. Mujhe nahi pata log iske baare mein kya kahenge. **Brooke**: Mujhe farak nahi padta log kya kahenge. For me it's truly wonderful! **Lidia**: Wanna go for a cup of coffee? **Brooke**: Sure. kitchen mein milte hai in 5 minutes, okay? **Lidia**: Okay
T5-MTL generated summary: Brooke has inherited a brooch from her father. Brooke will meet Lidia in the kitchen in 5 minutes. Lidia will go for a cup of coffee. Gold summary: Brooke wears a brooch in the office that she inherited from her grandmother. Lidia and Brooke will meet in the kitchen in 5 minutes to go for a cup of coffee.
Justin: hey max, kya tum abhi free ho? **Max**: mai huu **Max**: Kya tumhe kuch chahiye? **Justin**: mai homeless logo ke liye volunteer kar raha huu at the soup kitchen aur hume help karne ke liye log chahiye **Max**: count me in, Mai vaha 40 minutes mein pahuch jaaunnga
T5-MTL generated summary: Justin is volunteering at the soup kitchen to help the homeless. Max will be there in 40 minutes and he will be home in about an hour. Gold summary: Justin is volunteering at the soup kitchen for the homeless. Max will be there to help out in around 40 minutes.

Table 15: Top 3 summaries generated by T5 trained in multi-task setting from Hi-En code-switched conversations, in terms of ROUGE-1 scores

Ralph: Tum abhi bhi yahi kahi ho? **Mary**: Bathroom! **Ralph**: Oh! TMI! **Mary**: ek sec mein aa raha huu... **Ralph**: jitna time chaiye utna lo...
BART generated summary: Mary is in the bathroom. Gold summary: Mary is in the bathroom.
Peter: Kya mai tumhari car borrow kar sakta hu? **Hugh**: sure **Hugh**: but tumhari car ke saath kya hua? **Peter**: pata nahi **Peter**: aur time nahi hai check karne ka **Peter**: mai wese bhi late hu! **Hugh**: ok, ok, aake lele
BART generated summary: Peter will borrow Hugh's car. Gold summary: Peter will borrow Hugh's car.
Paula: Mai iss week apni thesis submit kr raha hu **Marcela**: Great! **Laura**: Congrats!!
BART generated summary: Paula submitted her thesis this week. Gold summary: Paula is submitting her thesis this week.

Table 16: Top 3 summaries generated by BART from Hi-En code-switched conversations, in terms of ROUGE-1 scores

Randy: Natalie, we're dead, look what happened!? **Natalie**: What's wrong? **Randy**: <file_video> **Natalie**: No way, how did Sally and Molly get to the parent's bedroom?!! **Randy**: I've no idea, someone must have left the door open **Natalie**: It looks really nasty! Why are all the sheets so dirty? **Randy**: I think they both went to play in the mud and then somehow ended up here **Natalie**: Jesus Christ, mum is going to be really mad! Start cleaning it **Randy**: I know, but it's hopeless, parents might here any minute **Natalie**:Just do it, I'll be back as soon as I can!	**Randy**: Natalie, we're dead, ye dekh kya hua!? **Natalie**: Kya hua? **Randy**: <file_video> **Natalie**: No way, Sally aur Molly parents bedroom par kese pohoche? **Randy**: Mujhe nahi pata, kisi ne door khula chhod diya hoga. **Natalie**: Ye toh bohot ganda lag raha hai. Ye sheets kyu gandi hai? **Randy**: I think wo dono keechad mei khelke yaha aa gaye somehow. **Natalie**: Jesus Christ, mumma bohot gussa hongi. Jaldi saaf karo. **Randy**: Pata hai, but it's hopeless. Parents yaha aate hi honge wese bhi. **Natalie**:Bas karle, mai jaldi aati hu!
Summary A: *Sally and Molly got into the parent's bedroom. The sheets are dirty. Randy thinks they went to play in the mud and then ended up here.* **Reference summary**: *Sally and Molly possibly played in the mud and got to parent's bedroom, where they made a mess. Randy will start cleaning and Natalie will join him as soon as she can, because parents might be here any minute.*	**Summary B**: *Natalie's parents are dead. Randy thinks Sally and Molly are in their parents bedroom. Natalie thinks they're dead.* **Reference summary**: *Sally and Molly possibly played in the mud and got to parent's bedroom, where they made a mess. Randy will start cleaning and Natalie will join him as soon as she can, because parents might be here any minute.*

Table 17: An example of conversation summarization, where the same model summarizes the given example. On the left, the model generates an English summary from the respective English conversation, and on the right, the model generates an English summary from the respective Hindi-English code-switched conversation. Both are erroneous summaries but differ in the error classes that each belong to. Summary A which is generated from the respective English conversation contains sentences that are irrelevant and misses out on relevant information, exhibiting the MI (*relevant information is missing*) error class as mentioned in section 6. Irrelevant lines in the summary are shown in (■). On the other hand, summary B which is generated from the respective Hi-En code-switched conversation also misses relevant information (MI) but also shows a case of *instrinsic hallucination*. Phrases exhibiting intrinsic hallucination in the summary are shown in (■).

Model	Learning Rate	Optimizer	No. Epochs	Batch Size	No. Beams
GPT2	3e-4	Adam	3	1	4
BART	3e-5	Adam	3	1	4
PEGASUS	5e-4	Adam	3	1	4
T5 MLT	3e-5	Adam	3	1	8
T5	3e-5	Adam	3	1	1
mBART	3e-5	Adam	10	1	1
Multiview	3e-4	Adam	3	32	4

Table 18: Training details of summarization models.

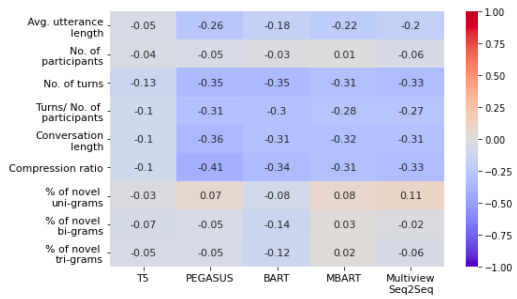

Figure 6: shows the correlation between the models' ROUGE-2 scores and the attributes of conversation-summary pairs that could affect the performance. This heatmap represents the correlation for English summaries generated from English conversations.

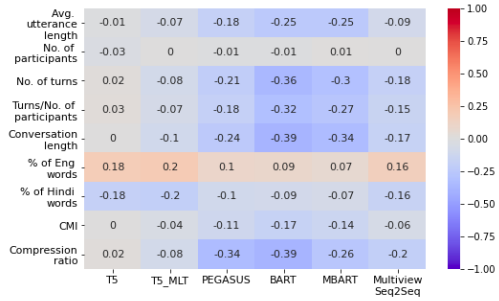

Figure 7: Shows the correlation between the models' ROUGE-2 scores and the attributes of conversation-summary pairs that could affect the performance. This heatmap represents the correlation for English summaries generated from Hindi-English conversations.