Project Report

Rishu Kumar

Abstract

Significant progress in natural language processing (NLP) and text summarization has led to the development of Transformer-based architectures, driving diverse summarization techniques. In this project, we explored abstractive summarization for the creative tasks to be able to use in the subsequent generation of plays in the THEaiTRE project. As the baseline for the project, we participated in the CreativeSumm shared task at COLING'22. We used a few-shot learning approach for summarization and self-qualitative evaluation to select the model for our submission. To study and overcome the difficulties in understanding implicit references in dialogues and make efforts to meaningfully handle token limits, we also contribute a novel corpus for dialogue summarization based on the Summ-Screen corpus. We also extend the dataset to include plays from Shakespeare to introduce a mixture of soap operas and plays. Through experimentation, we assess the potential of large language models for dialogue summarization and compare the performance in different model architectures. From our evaluation and comparison, we find that 1) BART-based models outperform other models, 2) few-shot fine-tuning is better than increasing the context window, 3) style transfer of a text is an issue when relying on few-shot training, and 4) the current evaluation metrics are not suitable to assess the quality of generated output. We also provide some preliminary examples of why generating summaries from closed-source models such as ChatGPT may not be a good choice.

1 Introduction

In recent years, the landscape of machine learning has seen remarkable development, particularly within the areas of NLP and text summarization. Summarization, an important NLP task, entails distilling extensive text while retaining its core essence. This task finds applications in diverse domains such as information retrieval, text generation, and content summarization for news articles, documents, and social media posts. The ascent of powerful models, notably the Transformer-based architectures (Devlin et al. 2019), has sparked an upsurge in academic research and the spread of diverse summarization techniques in various aspects. The current paradigm frequently involves fine-tuning large pre-trained language models (LMs) (Radford et al. 2019; Devlin et al. 2019) for summarization downstream tasks.

This progression in pre-trained large LMs has revolutionized NLP and extended the horizons of text generation tasks. Extractive and abstractive summarization are two predominant approaches within the summarization sub-field. Extractive summarization involves selecting and combining important sentences or phrases from the original text to create a summary, while abstractive summarization involves generating new text that conveys the essence of the original content. Notably, abstractive summarization emerges as a big challenge, aspiring to maintain the intended meaning of documents in concise but informative summaries (Rush et al. 2015).

Within this context, the focus of our project centres on leveraging the prowess of expansive LMs to achieve abstractive summarization, with a specific emphasis on dialogues extracted from SummScreen (Chen et al. 2022b) and the literary works of William Shakespeare[1]. These dialogues pose unique challenges for the summarization task due to their dynamic and interactive nature. These challenges arise due to the unique ways in which characters communicate, often involving implicit references and concise stage directions. Critical plot developments and essential information are frequently hidden within short stage directions rather than explicitly expressed in the dialogues themselves. Summarizing such dialogues requires a nuanced understanding of context and the art of rephrasing. For instance, transforming first-person pronouns into third-person forms is essential to maintain grammatical coherence. The complexities are further heightened when dealing with the Shakespearean language, which needs to be translated into contemporary English to facilitate the summarization process.

Furthermore, dialogue summarization is another big challenge that has attracted much attention. It attempts to provide a summary in the target language from a given document as turn-taking utterances, and the output is supposed to capture the person and the corresponding important points. In our work, especially with Shakespeare, we also encountered the difference in English from that era to modern English. An interesting addition to this phenomenon was that while the original texts were in Shakespearean English, the summaries were in modern English. There are additional challenges that are unique to conversational summarization. Firstly, in terms of data, pairing naturally occurring source language documents with their corresponding summaries in similar tone/style is not common, unlike in machine translation. This scarcity hinders the collection of extensive human-annotated datasets (Ladhak et al. 2020; Perez-Beltrachini & Lapata 2021). Secondly, from a model perspective, dialogue summarization requires proficiency in both summarization and generation skills (Cao et al. 2020). Consequently, conducting dialogue summarization

[1] https://en.wikipedia.org/wiki/William_Shakespeare

directly from models not trained on dialogues makes it hard to generate accurate summaries.

As a result, motivated by the aim to assess the potential of large LMs in achieving excellence in few-shot dialogue summarization, we dive deeply into this task by building our own corpus, which is extracted from SummScren and Shakespeare's plays. More precisely, this study will shed light on the following research questions:

1. How is the performance of LMs in the context of complex dialogue summarization? Will the challenging dialogue features mentioned above affect results negatively to a large extent?
2. How does increasing the context length of the summarizer model help in dialogue summarization?

In order to answer these questions, we manually collect English dialogues extracted from the plays of William Shakespeare and the SummScreen dataset and the corresponding summaries to build our own custom corpus. On the basis of this corpus, we propose two strategic approaches to creating different datasets on which the Bidirectional and Auto-Regressive Transformers (BART) (Lewis et al. 2020) and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019) is trained. For the baseline, we utilize an existing abstractive summarization dataset designed exclusively for dialogues, known as SAMSum (Gliwa et al. 2019). We train the BART and BERT models on this dataset for dialogue summarization. Considering the linguistic distinctions between the SAMSum dataset and our custom corpus derived from SummScreen Shakespeare's plays, we employ a combining strategy. This approach involves fine-tuning the models using a joint dataset that includes both the SAMSum data and our custom dataset.

Through this project, our contributions are mainly two-folded:

- We create a corpus based on the SummScreen corpus and extend it to Shakespeare's plays to be used in Dialogue Summarization in creative tasks
- By focusing on dialogues extracted from Shakespearean plays, we explore a distinctive and challenging facet of summarization, thereby contributing to the advancement of dialogue summarization methodologies using a few-shot approach
- From the comparison of different architectural choices for models, we note that BART still performs better than other models and increasing the context window may not translate to better summaries.

In this report, our journey begins with a literature review in Section 2 for the summarization task in NLP. After this, we describe in detail the process of constructing a dataset tailored to the creative dialogue summarization task in Section 3. We shed light on data collection, pre-processing, and annotation steps. Building upon this foundation, Section 4 delves into our experimental methodology as used in our submission to the CreativeSumm shared task. Through evaluation metrics and case analysis, Section 5 includes the report on the performance of the LMs with different settings in generating dialogue summaries and the discussion of the hidden reasons leading to these results. Lastly, we summarize and analyze our project's limitations and future work and make final conclusions based on the experiment results in Section 6 and 7.

2 Related Work

With the advances in machine learning, there have been proportional changes in summarization. While in recent years, the focus has been shifted completely towards incorporating state-of-the-art language models for this, early works such as Luhn (1958) talk about more statistical features such as word frequency and phrase frequency. In the following decade, Edmundson (1969) built upon this and proposed to incorporate it as follows:

- Cue method: Relevance of phrase is based on the frequency of certain cue words.
- Title method: Using the words appearing in the Title and heading to calculate the weight of a sentence.
- Location Method: In this method, the authors hypothesize that the sentence appearing at the beginning of a paragraph or document is more relevant.

A study by Sanderson (1998) discusses the summarization system from an information retrieval point of view and discusses the human evaluation of the models. In this paper, they also talk about user-directed, i.e., personalized summaries based on the keywords in the query provided by the user. With the shift to neural-network-based models, the domain of summarization started to diverge more in terms of abstractive and extractive summarization. With a black box approach to summarization, the approaches between abstractive and extractive summarization grew wider.

While the domain of summarization saw progress through these years, it was primarily focused on summarizing documents and articles. After Transformer (Vaswani et al. 2017) based Language Modelling approach came into practice, all downstream tasks, including summarization, switched to using these models. An important upside and caveat of these models was the context length; BERT (Devlin et al. 2019) had a context length of 512 tokens, while BART (Lewis et al. 2020) had a context length of 1,024 tokens. The increased context lengths provided the capability to have more information present in the model for better information retention. However, this alone was insufficient to cover an entire document as it may span over 20,000 tokens or more. Beltagy et al. (2020) proposed Longformer, a BART-based model that increased the context length to a maximum of 16,384 tokens. The proposed approach linearly concatenated the attention in the BART models for this architecture. The proposed approach provided better results in creating summaries that were consistent throughout the generation process.

Even with the models' increased capacity, conversational summarization in the creative domain remained largely unexplored. This, in part, can be attributed to the lack of available datasets, especially in a monolingual setting. SummScreen (Chen et al. 2022b) was the first curated dataset of transcripts from different TV Shows and corresponding summaries as generated by viewers on the internet. The average length for a transcript in this dataset is around 9,000 words, which is far beyond the token limit of readily available large LMs for fine-tuning, except for Longer-based models. To address this issue, Summ^N (Zhang et al. 2022) introduced a greedy approach to split the data into chunks, create coarse summaries for the dataset, and then apply a multi-stage training approach to generate final summaries.

Recently, CreativeSumm (Agarwal et al. 2022) introduced creative summarization as a novel shared task. The first iteration of the shared task saw different approaches for handling summarization, ranging from using Longformers to zero-shot approaches. Chatterjee et al. (2022) proposed using multiple summarizers on different splits of the data along with a multi-layer neural network to handle the token limits of existing models while Kashyap (2022) used an encoder-decoder architecture with Longformers for handling the complete text all at once. Upadhyay et al. (2022) uses several pre-and-post processing methods along with essentially a zero-shot training approach with BART, and we (Kumar & Rosa 2022) proposed using a manually curated subset of the released dataset as few-shot training examples. The results show that using Longformers surprisingly provides the lowest-scoring models while fine-tuning BART-based models with different strategies are ranked in the top three models.

3 Dataset

For this project, we create our own dataset of samples from the SummScreen dataset and Shakespeare's plays and their corresponding summaries. In this section, we describe how we collect the plays and summaries. Additionally, we provide an overview and analysis of the dataset consisting of play chunks and summaries.

3.1 Collection of Shakespeare's plays

We experiment with plays of William Shakespeare in their original English version as well as their modern English translation [2]. The DraCor project (Fischer et al. 2019) provides the full text of the plays in the original setting. In particular, we use the *Shakespeare Drama Corpus*[3]. Among other things, DraCor provides meta-data, information about network relations, and the full text as a structured XML file with tags for speakers, stage directions, and spoken texts for each play. We are mainly interested in the plain text format of the plays, so we extract relevant text passages from the XML file. We take additional preprocessing steps in order to structure the play texts more consistently, thus making them easier to handle for the deep learning model. This includes:

1. Removing blank lines;
2. Tokenization, e.g., splitting punctuation from words;
3. Remove line breaks from a speaker's part, i.e., each turn of a speaker is one line.

Collection of Summaries

For each play, we manually collect summaries. We look for up to three summaries for every play via a simple online search. We consult various sources, including Wikipedia and sites for students studying Shakespeare plays. The summaries vary in length and detail in describing the play's content. We manually remove sections of the summaries that are not concerned with the actual content of the play, such as the time and setting the play was written, as well as analyses that go beyond the content of the play. Direct quotes are also deleted from the play. Each summary file includes the texts after pre-processing and their source, such as the website. The Appendix A contains all sources used for collected summaries.

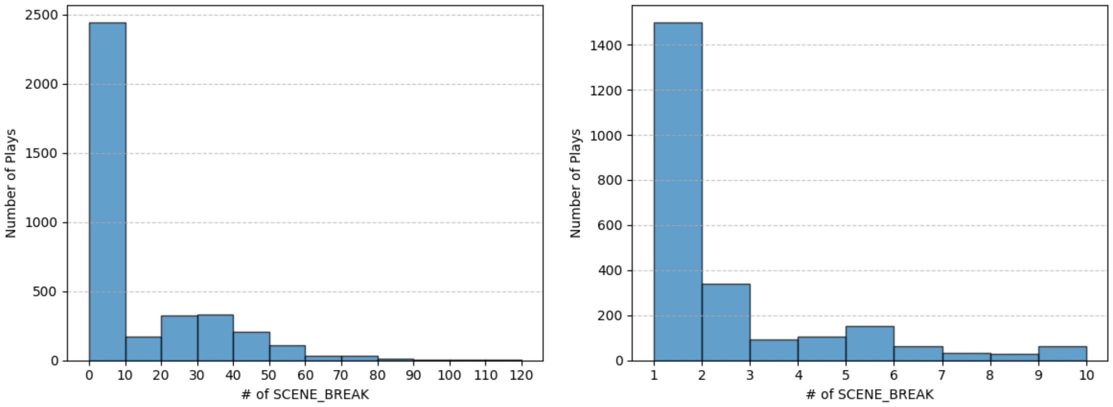
3.2 Collection of SummScreen plays

We explore the dataset to find linguistic cues in the data, which would help in splitting the dataset. Upon review, we select SCENE_CHANGE as our parameter for the same. Figure 1 provides a brief overview of the statistics for our dataset. We take similar preprocessing approaches in this dataset, such as

1. Removing @@ from the text;
2. Tokenization, e.g., splitting punctuation from words;

[2] The project is called NoSweatShakespeare

[3] https://dracor.org/shake



(a) For entire dataset

(b) For dataset where SCENE_CHANGE < 10 times

Figure 1: Statistics about SCENE_CHANGE in SummScreen dataset

3. Remove line breaks from a speaker's part, i.e., each turn of a speaker is one line.
4. (Optionally) Convert all non-dialogue lines to **Narrator: [...]**

Collection of Summaries

While the dataset provides summaries for each episode, we found the plays for older TV shows to be quantitatively and qualitatively worse than the summaries in newer TV shows. Based on the genre of the show, the summaries also varied in length and details of the plot. While curating our dataset, we additionally looked up the data through Internet searches to help in mapping.

3.3 Mapping Process

As there is a length limit for the input in Transformer-based deep learning models, we manually map parts of the plays to parts of the summaries (we refer to this as *chunk*). Although the chunks are created automatically, we post-process the chunks to ensure that each chunk is at least somewhat coherent. This manual creation of the training and validation dataset is done as follows: A chunk of a play is read and compared to the available collected summaries. Parts of the summaries are considered relevant for this particular chunk and mapped. Importantly, not all sections in the chunk have an equivalent in the mapped summary. This is desired, as not all sections in a play are considered relevant or important in a summary. Therefore, a model should also learn what information to drop when generating summaries. In case a chunk has no corresponding section in any of the summaries, we try to write a short description of the scene ourselves. This is more problematic for the older TV Shows as, for some of the lesser-known shows, not many quality summaries are available online.

In total, we map chunks to summaries for sixteen plays for Shakespeare for training our models. For Shakespeare's plays, we manually selected four plays for mapping. The names of the plays are as follows:

1. Macbeth
2. The Winter's Tale
3. The Comedy of Errors
4. The Merry Wives of Windsor

For the test set, we do not perform any mapping as we aim to evaluate the automatic summarization of a complete play, meaning that at inference time, a play is split automatically into chunks, fed to the model, and its output is concatenated. Finally, the complete generated summary is evaluated against corresponding references.

3.4 Dataset analysis

In this section, we describe our qualitative analysis of our dataset. For the mapped data, we chose not to automatically map the text to summary using a greedy approach as described in $Summ^N$ (Zhang et al. 2022). We find that due to the shorter length of summaries in older TV shows, we are deliberately introducing errors in the first part of the training regime, where we generate coarse summaries. Furthermore, introducing this error in the first part hints that it's only the second summarization step, which combines output from this erroneous data, is the most important one. Given the erroneous mapping and yet better performing summaries in some automated evaluation metrics, it may point towards data contamination. While it is a promising research direction, it was not aligned with the goal of our project. Similar variations can be observed with the Shakespeare dataset: The shortest summary for a chunk is only six words long, and the longest is 137 words. However, on average, the summaries for a play chunk are much shorter, with a mean of only 20 tokens. All in all, this speaks to the fact that sections of a play are deemed less or more important and have, therefore, been summarized with varying degrees of detail.

It is interesting to see how summary and chunk differ from one another. For Shakespeare, the overlap of vocabulary items is somewhat higher. There is about a 5% overlap with the respective play chunk's vocabulary if stop-words are included. If they are ignored, this percentage drops to about 2%. Still, it is quite low. However, the overlap of vocabulary with the actual dialogue in the SAMSum dataset is much higher, with about 20%, not including stop-words, and the summaries are much shorter on average, with only about 20 tokens. Therefore, this suggests that the summarization of plays is challenging for our custom dataset, especially compared to the SAMSum corpus.

It also becomes clear that extractive summarization is not an option for complex dialogues like plays, especially for text with antiquated language; the style and vocabulary differ greatly. As a result, copying phrases or sentences from the source material would not be a successful strategy for a model. This is exemplified by the excerpt of a chunk and its corresponding summary in Table 1. Strikingly, there are sections in the summaries that are only implicitly present in the play. For instance, the fact that the character Lady Macbeth is reading a letter from her husband is not clearly mentioned. The same applies to content that is mentioned in previous play chunks: In this example, Lady Macbeth mentions that something was promised to her husband. The corresponding summary condenses this as the witches' prophecies. The fact that it was witches who promised something to Macbeth was mentioned at the beginning of the play and is now not accessible anymore. Of course, this is due to the fact that plays have to be split up. Similarly, other facts are quite obscured by the figurative language. Additionally, we note that not all the summaries have equal amount of details available. While it is possible to create a mapping from only one summary, just using one summary leads to more chunks having no summaries at all, while having multiple summaries reduces this issue. Overall, it is evident that a lot of reasoning and comprehension is necessary in order to condense the dialogue of Shakespeare's plays into a summary.

4 Experiments

The following section will mainly introduce the Transformer-based LMs, the evaluation metrics we employed, and the detailed experimental designs in the cross-lingual and multi-lingual settings.

4.1 Models

BART Lewis et al. (2020) introduced BART which combines Bi-directional and Auto-Regressive Transformers. Lewis et al. (2020) showed that it achieves state-of-the-art performance in a number of tasks, including summarization and abstractive dialogue. During pre-training, documents are corrupted and subsequently encoded with a bi-directional model. Based on these encodings, the auto-regressive decoder then reconstructs the original document. Corrupting techniques include token masking, token deletion, and sentence permutation.

Play Chunk	Mapped Summary
Messenger : The king comes here to-night . LADY MACBETH : Thou 'rt mad to say it : Is [...] LADY MACBETH : Give him tending ; He brings great news . Exit Messenger The raven himself is hoarse That croaks the fatal entrance of Duncan MACBETH : My dearest love , Duncan comes here to-night . LADY MACBETH : And when goes hence ? MACBETH : To-morrow , as he purposes . LADY MACBETH : O , never Shall sun that morrow see ! Your face , my thane , is as a book where men May read strange matters . To beguile the time , Look like the time ; bear welcome in your eye , Your hand , your tongue : look like the innocent flower , But be the serpent under 't . He that 's coming Must be provided for : and you shall put This night 's great business into my dispatch ; Which shall to all our nights and days to come Give solely sovereign sway and masterdom . MACBETH : We will speak further . LADY MACBETH : Only look up clear ; To alter favour ever is to fear : Leave all the rest to me . Hautboys and torches . Enter DUNCAN , MALCOLM , DONALBAIN , BANQUO , LENNOX , MACDUFF , ROSS , ANGUS , and Attendants	Lady Macbeth receives news from her husband [...] of the prophecy and his new title and she vows to help him become king by any means she can. Macbeth's return is followed almost at once by Duncan's arrival.

Table 1: Excerpt of a chunk mapped to a summary from the play Macbeth. Corresponding sections are marked in the same color.

BERT2BERT We also use BERT2BERT (Chen et al. 2022a) for our experiments. It is an encoder-decoder-based model on BERT architecture. The main aim of this model was to introduce a better mapping of text and the corresponding summaries as it leverages two Language models for the same. For our experiments, we use bert2bert-uncased, which ignores the casing of the input.

LED To handle the issues of small context windows in document summarization, Beltagy et al. (2020) introduced LED, which is also a seq2seq model based on BART architecture. This model's local attention mechanism is repeated multiple times to create a global attention mechanism. This method also enables these models to handle longer input sequences, with the longest being 16384.

4.2 Metrics

We mainly use ROUGE to do an automated evaluation of our models. ROUGE (Lin 2004) is a metric that evaluates the number of overlapping units between the generated summary and the references. In particular, we look at overlapping unigrams (ROUGE-1) and overlapping bigrams (ROUGE-2) as well as scores based on the longest common subsequence (ROUGE-L). However, the organizers of the CreativeSumm shared task evaluated on different matrices, and we also include the result in here.

4.3 Hyperparameters

We train our models for a maximum of three epochs in all experiments and use AdamW as an optimization method with an initial learning rate of 5e-05. We also ensure that we use the vanilla hyperparameters released by different studies to provide a fair comparison.

4.4 Experimental Designs

As a baseline model, we use the **BART-LARGE** model as released on huggingface. This model has been trained on extreme summarization of news articles and generates a single sentence as the summary. We fine-tune this model with the only available conversational corpus, i.e. SAMSum, in conjunction with adding our custom dataset to the SAMSum corpus. We also tried just using our dataset (Shakespeare) for training, but as expected, the results were not promising. We will discuss more in the manual evaluation setion.

Model	Rouge1	Rouge2	RougeL	RougeSum
bert2bert	27.8385	9.2011	20.6134	20.6184
BART_large	47.9166	24.4649	38.1257	38.1219
LED_4096	48.0707	24.2645	38.1655	38.1986

Table 2: Evaluation Score on SAMSum test dataset after training with SAMSum data

LED As the LED models were created to handle more context for better summarization, we experimented with different input token limits for our study. Notably, the 1024 token limit gives the same result as the BART model, as it essentially is the same model. We also like to note that due to the extremely demanding memory requirements for the 16K token limit version of LED, we did not run the experiments ourselves for the same. However, a submission in CreativeSumm used the model, and we argue that it can be used for comparison as the model was trained on the complete training set released by organizers.

In total, this resuled in 184 experiment combinations. In these, we also included fine-tuning only on our custom dataset seperately (SummScreen, and Shakespeare's Plays) with different model choices. Considering the novelty of the field and lack of experiments in the domain, we believe that an exploratory approach to be the best approach for our project.

Our extensive experimental design along with only limited number of data points, encouraged us to look beyond the automated evalution for our study. We found that while a 0.20+ point difference in results could help distinguish the quality of the output, smaller differences did not provide any meaningful information.

5 Results & Discussion

In this section, we will report the results from different experiment designs and discuss them from the perspective of dataset content and different architectural choices via quantitative and qualitative analysis.

5.1 Automatic Evaluation

SAMSum Table 2 comprises the results of our evaluations based on different training regimes on the SAMSum dataset. The lack of a conversational dataset forces us to use the SAMSum as the only reliable way to automatically evaluate our models.

We find that LED-based models are slightly better performing than BART-based models in this experiment. It is also very intuitive, as few of the examples in SAMSum corpus exceed the 1024 tokens limit for BART-LARGE, but that's still well within

Model	Rouge1	Rouge2	RougeL	RougeSum
bert2bert	10.5549	2.0721	6.5507	6.5445
BART_large	24.5986	5.4374	12.7077	12.6977
LED_4096	24.7298	3.6579	13.8439	13.8652
BART_large_Dracor	35.6145	8.6837	14.9063	14.8977

Table 3: Evaluation Scores of models finetuned on SAMSum dataset and tested on FD/Dracor test dataset

token limits of 4096 for the LED-4096 model. We also see that bert2bert is the worst-performing model due to the lower context window size of 512. Almost all of the examples in the SAMSum corpus exceed this window, and we only include this in future comparisons for reference, as we have split our custom dataset to fit the 1024 or more token limit of BART-based models.

Few-shot Experiments Table 3 provides a brief overview of our ***essentially zero-shot*** training experiments where we trained our models on SAMSum dataset and evaluated it on the SummScreen/Shakespeare's play dataset. The last entry in this table is the evaluation score of training our model on the SAMSum dataset and evaluating it on the Dracor corpus.

CreativeSumm official Results Table 4 provides an overview of the results officially released by the shared task organizers. Our submission to this shared task consisted of fine-tuning a BART-LARGE model on SAMSum and our custom dataset from the SummScreen corpus. On multiple metrics, we have scored second place in the shared task, irrespective of using only a small sample of the provided dataset.

5.2 Qualitative Evaluation

While automatic evaluation metrics can provide first insights into the performance of the fine-tuned models, these evaluation methods fall short with respect to more thorough evaluation regarding semantic and pragmatic aspects of the generated summaries. Furthermore, automatic metrics only provide insights into how similar the generated summaries are to certain references. Clearly, there are more than three possible summaries for a play. As a result, it is not enough to only consider automatic metrics; we therefore also do a qualitative analysis of the generated summaries.

Following Fabbri et al. (2021), we evaluate the generated summaries according to four dimensions: coherence, consistency, fluency, and relevance. According to Fabbri et al. (2021), coherence concerns the structure of a summary and the quality of

	ROUGE-1	ROUGE-2	ROUGE-L	Length	Density	Coverage	Novel 1-grams	Novel 2-grams
LED_1024	0.1428	0.0154	0.1236	330	1.1440	0.7148	0.3060	0.7801
LED_4096	0.1694	0.0209	0.1501	188	1.4378	0.7343	0.2803	0.7314
LED_16384	0.1514	0.0170	0.1334	192	1.5474	0.7108	0.2904	0.7285
inotum_summscreen-fd.jsonl	0.2860	0.0624	0.2529	86	1.0321	0.6664	0.3715	0.8251
team_ufal_fd.json	0.2469	0.0408	0.2300	289	2.0821	0.7127	0.2484	0.6498
AMRTVSumm_summscreen-fd.jsonl	0.2307	0.0303	0.2106	256	0.8789	0.6137	0.4924	0.8569

Table 4: Result snippet from Creative Summarization shared task at COLING'2022, our submission include training on SAMSum + our custom SummScreen dataset.

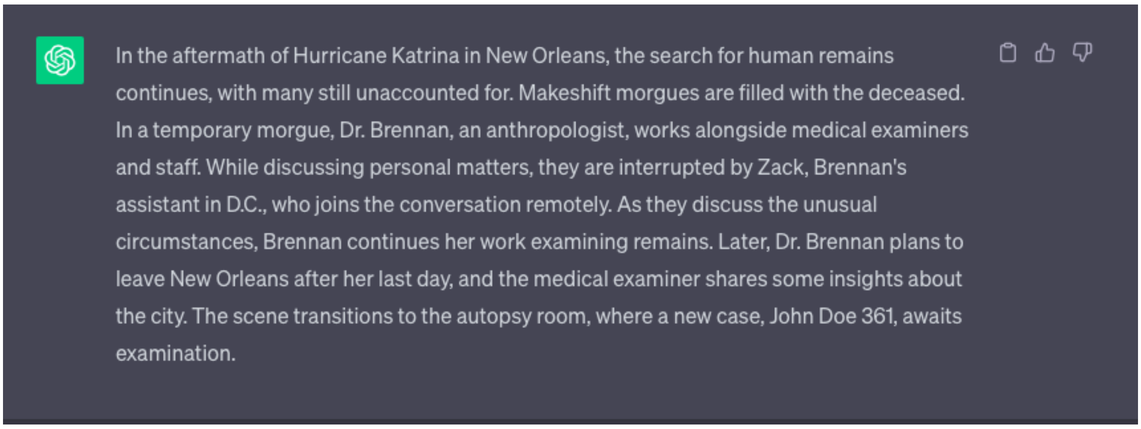
all sentences taken together. Consistency is about the factual content of the source material aligning with the generated summary. Finally, fluency is about the quality of individual sentences with respect to formatting and grammaticality, whereas relevance is concerned with the most important information being extracted (Fabbri et al. 2021).

For the qualitative evaluation, we used Shakespeare's play as it was more in line with the goal of this project, i.e. using summarization for the generation of Theatre plays.

A preliminary look at our results states that fine-tuning on SAMSum corpus provides a better result than finetuning solely on our custom dataset, which also aligns with intuition as the number of data points in these two approaches is not comparable. It often leads to blank summaries in the approach where we fine-tune the models only on the custom dataset. In the play *A Midsummer Night Dream*, we find that most of the chunks do not have any output at all, and the ones that have the output are garbage words. For example, chunk_3 only has "confirmssssssssssssss" as the output, but this word does not occur in our training or evaluation dataset. We hypothesize that the tokenizer is not getting enough instances to update itself due to the inclusion of several non-modern English spelling and words in these data points. We notice a similar occurrence where the last character is repeated multiple times in some other chunks.

The output generated by training on the joint dataset provides better results. However, we note that it also suffers from poor fluency, consistency, coherence, and relevance. For example, in the same play, chunk_3 is summarized as "The play begins with scenes from the play, where Demetrius reveals his true nature and plays out the play-play, including scenes with scenes from Hermia's past. The play continues with Polixenes, who plays Thisby, the lion, and Flute, who plays Robin, the tailor, and Tom Snout, the tinker.", whereas the model only fine-tuned on the SAMSum produces "There is an improper plot line between Demetrius and Thisby in HELENA's play 'The most comedy, and most cruel death of Pyramus and Thisby'. Quince tells FLUTE and BOTTOM that they have to play Thisby's mother, Francis Flute, the bellows-mender, and the lion." [4] In this instance, we see that the first model has several inconsistencies in the output, yet it still follows a so-called *natural* flow of information. In the second output, we see different types of inconsistencies. We

[4] The outputs are copied verbatim, so typesetting issues are from the outputs.



In the aftermath of Hurricane Katrina in New Orleans, the search for human remains continues, with many still unaccounted for. Makeshift morgues are filled with the deceased. In a temporary morgue, Dr. Brennan, an anthropologist, works alongside medical examiners and staff. While discussing personal matters, they are interrupted by Zack, Brennan's assistant in D.C., who joins the conversation remotely. As they discuss the unusual circumstances, Brennan continues her work examining remains. Later, Dr. Brennan plans to leave New Orleans after her last day, and the medical examiner shares some insights about the city. The scene transitions to the autopsy room, where a new case, John Doe 361, awaits examination.

Figure 2: ChatGPT's output of the first chunk for Bones Season 1 Episode 19

assume that the inclusion of our custom dataset helps better in the style transfer of the summaries compared to the summaries generated by the model only trained on the SAMSum dataset. This also highlights the issue in the evaluation, as only by looking at the average number of scores is it difficult to determine the true relevance of summaries generated by our models.

One of the most important aspects of summarization is consistency. Does the content of the generated summary accurately represent the content of the play? In most cases, the answer is clearly no or very little. In case of *Romio & Juliet*, we see the inconsistency of characters and their corresponding role in the play. We also find instances of correct characters from the input, but they are associated with wrong actions from the play. We also find instances of Character Names appearing from other plays that were not introduced to the models in training. It is also noteworthy that when we introduce ***Narrator*** for non-dialogue lines, the output contains Narrator as a character, even though none of the training summaries had Narrator as a character in the summary.

Another interesting pattern is that while the plays contained several words in old English, the summaries do not contain any words from old English but their equivalent word in modern English. This can be explained by the tokenizer using subwords rather than words, and this makes the summaries used for fine-tuning written in modern-day English.

All-in-all, a qualitative review of the results indicates that massively large language model-based summarization is not a reliable source for script generation. While it is impossible to pinpoint the exact reason for every bad inconsistency, we hypothesize that the model must have seen the text during pre-training and is hallucinating during inference.

ChatGPT While using ChatGPT was not in our purview when we started our work, we still did some qualitative evaluation on inference with ChatGPT on the SummScreen dataset to judge its performance. We find ChatGPT to be very random in generating the output and often with prompts such as ***Summarize This:*** *followed by the first chunk of the script* makes it generate the entire summary. While in the continued "conversation" it is possible to modify this behaviour, new session leads to similar issues. The website explicitly mentions factual inconsistencies as one of the downsides of ChatGPT, and we experienced the same in our experiment. Figure 2 provides a screenshot for the same. In the play, it's nowhere mentioned that Dr. Brennan is an anthropologist, but ChatGPT's summary includes it. While this inclusion may not affect the automated evaluation scores, it changed the entire character of the protagonist and would lead to factual inconsistencies throughout the script if used as the base for generating continued text.

6 Limitations

To begin with, the utilization of mixed training data, such as Shakespearean English and modern English, can potentially lead to diminished performance compared to real, curated dialogue data. The few-shot nature of the experiment can introduce potential noise, which is undesirable.

Another prominent constraint in this project is the token limitations inherent to Transformer-based models. Dividing the plays into chunks and processing them sequentially within the model restricts the availability of comprehensive context and information. This becomes particularly problematic when dealing with intricate dialogue texts like those found in Shakespeare's plays. Retaining information from earlier scenes that become relevant later in the text requires storing knowledge across multiple chunks for accurate integration.

Additionally, since both BART and BERT are trained on a version of the Common-Crawl corpus, it is likely that this corpus already contains websites such as Wikipedia, where we collect summaries for our custom corpus. The same goes for the actual plays, which are freely available online. This might mean that the BART and BERT models have already seen the texts. This would be an instance of contamination. Consequently, the results presented here could potentially be overly optimistic as our test set is at least partially present during pre-training. This issue is hardly avoidable since Shakespeare is one of the most well-known authors in the world, and SummScreen consists of well-known and highly-rated TV shows.

7 Conclusion & Future Work

In this work, we experiment with the summarization of creative text, with the additional challenge of having non-standard language and creative settings. We also explore the domain of few-shot summarization in different architectural settings to help alleviate the issue of the lack of conversational corpus. We demonstrate that by including as few as 100 data points, we can generate better results than increasing the context window length of the model.

The automated evaluation metrics used in the domain are another area where we would like to draw the community's attention. Numbers alone for summarization based on n-gram matching do not provide us with an overview of the quality of the result. While the inclusion of BERTScore and other large LM-based evaluation metrics has helped mitigate some of the problems, we do not understand how these scores correlate to the accuracy and robustness of the output. Future work should concentrate more on this problem.

We also note that except for the integration of models for the generation of Theatre play scripts, we fulfil all other criteria of our proposed research plan. We strongly believe that the current lack of a dataset will continue the issue until we have more datasets. To start the mitigation, we release our albeit small but manually curated dataset for Shakespeare's plays and SummScreen. Creating a complete dataset was beyond the project's scope, but we hope it will help motivate researchers in the domain to generate more data.

A Summary sources

Play	Summary Sources
The Winter's Tale	https://nosweatshakespeare.com/play-summary-2/winters-tale/ https://www.playshakespeare.com/the-winters-tale/synopsis https://www.bardweb.net/plays/winterstale.html
The Comedy of Errors	https://nosweatshakespeare.com/play-summary-2/comedy-errors/ https://www.playshakespeare.com/comedy-of-errors/synopsis https://www.supersummary.com/comedy-of-errors/summary
Macbeth (EN)	https://www.playshakespeare.com/macbeth/synopsis https://nosweatshakespeare.com/play-summary-2/macbeth/ https://www.litcharts.com/lit/macbeth/summary
The Merry Wives of Windsor	https://www.sparknotes.com/shakespeare/merrywives/summary/ https://nosweatshakespeare.com/play-summary-2/merry-wives-windsor/ https://www.bard.org/study-guides/synopsis-the-merry-wives-of-windsor/
A Midsummer Nights Dream	https://nosweatshakespeare.com/play-summary-2/midsummer-nights-dream/ https://www.playshakespeare.com/midsummer-nights-dream/synopsis https://www.supersummary.com/a-midsummer-night-s-dream/summary/
As You Like It	https://nosweatshakespeare.com/play-summary-2/as-you-like-it/ https://www.playshakespeare.com/as-you-like-it/synopsis https://www.supersummary.com/as-you-like-it/summary/
Henry VIII	https://www.rsc.org.uk/henry-viii https://www.sparknotes.com/shakespeare/henryviii/summary/ https://nosweatshakespeare.com/play-summary-2/henry-viii/
Romeo and Juliet	https://nosweatshakespeare.com/play-summary-2/romeo-juliet/ https://www.playshakespeare.com/romeo-and-juliet/synopsis https://www.supersummary.com/romeo-and-juliet/summary/
The Tempest	https://www.litcharts.com/lit/the-tempest/summary https://www.sparknotes.com/shakespeare/tempest/summary/ https://www.cliffsnotes.com/literature/t/the-tempest/play-summary
The Two Noble Kinsmen	https://www.bardweb.net/plays/kinsmen.html https://nosweatshakespeare.com/play-summary-2/two-noble-kinsmen/ https://www.playshakespeare.com/two-noble-kinsmen/synopsis
Timon of Athens	https://nosweatshakespeare.com/play-summary-2/timon-athens/ https://www.playshakespeare.com/timon-of-athens/synopsis https://www.bardweb.net/plays/timon.html

Table 5: Sources for all German and English collected summaries.

References

Agarwal, Divyansh, Alexander R. Fabbri, Simeng Han, Wojciech Kryscinski, Faisal Ladhak, Bryan Li, Kathleen McKeown, Dragomir Radev, Tianyi Zhang & Sam Wiseman. 2022. CREATIVESUMM: Shared task on automatic summarization for creative writing. In *Proceedings of the workshop on automatic summarization for creative writing*, 67–73. Gyeongju, Republic of Korea: Association for Computational Linguistics. https://aclanthology.org/2022.creativesumm-1.10.

Beltagy, Iz, Matthew E. Peters & Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150* .

Cao, Yue, Hui Liu & Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 6220–6231. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.554. https://aclanthology.org/2020.acl-main.554.

Chatterjee, Niladri, Aadyant Khatri & Raksha Agarwal. 2022. Summarization of long input texts using multi-layer neural network. In *Proceedings of the workshop on automatic summarization for creative writing*, 13–18. Gyeongju, Republic of Korea: Association for Computational Linguistics. https://aclanthology.org/2022.creativesumm-1.2.

Chen, Cheng, Yichun Yin, Lifeng Shang, Xin Jiang, Yujia Qin, Fengyu Wang, Zhi Wang, Xiao Chen, Zhiyuan Liu & Qun Liu. 2022a. bert2BERT: Towards reusable pretrained language models. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2134–2148. Dublin, Ireland: Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.151. https://aclanthology.org/2022.acl-long.151.

Chen, Mingda, Zewei Chu, Sam Wiseman & Kevin Gimpel. 2022b. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 8602–8615. Dublin, Ireland: Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.589. https://aclanthology.org/2022.acl-long.589.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

Edmundson, Harold P. 1969. New methods in automatic extracting. *Journal of the ACM (JACM)* 16(2). 264–285.

Fabbri, Alexander R., Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher & Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics* 9. 391–409. doi:10.1162/tacl_a_00373. https://aclanthology.org/2021.tacl-1.24.

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling & Peer Trilcke. 2019. Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In *Proceedings of DH2019: "Complexities", Utrecht, July 9–12, 2019*, Utrecht University. doi:10.5281/zenodo.4284002.

Gliwa, Bogdan, Iwona Mochol, Maciej Biesek & Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd workshop on new frontiers in summarization*, 70–79. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-5409. https://aclanthology.org/D19-5409.

Kashyap, Prerna. 2022. COLING 2022 shared task: LED finteuning and recursive summary generation for automatic summarization of chapters from novels. In *Proceedings of the workshop on automatic summarization for creative writing*, 19–23. Gyeongju, Republic of Korea: Association for Computational Linguistics. https://aclanthology.org/2022.creativesumm-1.3.

Kumar, Rishu & Rudolf Rosa. 2022. TEAM UFAL @ CreativeSumm 2022: BART and SamSum based few-shot approach for creative summarization. In *Proceedings of the workshop on automatic summarization for creative writing*, 24–28. Gyeongju, Republic of Korea: Association for Computational Linguistics. https://aclanthology.org/2022.creativesumm-1.4.

Ladhak, Faisal, Esin Durmus, Claire Cardie & Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the association for computational linguistics: Emnlp 2020*, 4034–4048. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.360. https://aclanthology.org/2020.findings-emnlp.360.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov & 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*, 7871–7880. Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. https://aclanthology.org/2020.acl-main.703.

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries .

Luhn, Hans Peter. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development* 2(2). 159–165.

Perez-Beltrachini, Laura & Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, 9408–9423. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. doi:10.18653/v1/2021.emnlp-main.742. https://aclanthology.org/2021.emnlp-main.742.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8). 9.

Rush, Alexander M., Sumit Chopra & Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 379–389. Lisbon, Portugal: Association for Computational Linguistics. doi:10.18653/v1/D15-1044. https://aclanthology.org/D15-1044.

Sanderson, Mark. 1998. Accurate user directed summarization from existing tools. In *Proceedings of the seventh international conference on information and knowledge management*, 45–51.

Upadhyay, Aditya, Nidhir Bhavsar, Aakash Bhatnagar, Muskaan Singh & Petr Motlicek. 2022. Automatic summarization for creative writing: BART based pipeline method for generating summary of movie scripts. In *Proceedings of the workshop on automatic summarization for creative writing*, 44–50. Gyeongju, Republic of Korea: Association for Computational Linguistics. https://aclanthology.org/2022.creativesumm-1.7.

Vaswani, Ashish, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. In *Nips*, https://api.semanticscholar.org/CorpusID:13756489.

Zhang, Yusen, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev & Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 1592–1604. Dublin, Ireland: Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.112. https://aclanthology.org/2022.acl-long.112.