Universiteit Leiden

Master Computer Science

Video Game Dialogue Summarization using Natural Language Processing

Name:	Leon Li
Student ID:	s2448475
Date:	12/07/2024
Specialisation:	Data Science
1st supervisor:	Mike Preuss
2nd supervisor:	Giulio Barbero

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

With the increased use of Natural Language Processing many models have become available for a variety of language tasks. One such task includes autamatically summarizing text or dialogue. This paper looks at using these text summarization models to automate the creation of summaries for video games and sees the feasibility of using summarization models in real-world applications. The research focus lies in finding the difficulty in summarizing video game dialogues as well as looking at the actual performance of the generated summarizations. Experiments were conducted using multiple models and different preprocessing steps to find the best method to summarize video game dialogue. The results showed that summarization models fine-tuned on book summarization performed the best. Furthermore, the results showed that adding preprocessing increased the performance of the generated summaries in most cases. The generated summaries did not manage to capture all the context of the dialogue. However, we deemed that the amount of context captured in the summaries would still be enough for a user to have a high grasp of the story.

Contents

1 Introduction

Keeping track of events that have happened in video game stories can be difficult. This can be even more difficult if a user plays multiple games at the same time or if they only play a game occasionally. This can cause a user to forget what has happened in the story that they are currently playing.

Making a summary of prior events can help users recall what has happened so far in the story. If within a video game there is a quick way to review a summary of major story events that have happened up until the point that a user has played, a user will be able to better recall where they are in the story in case they forgot it. Adding a summary, therefore, can be beneficial for a user if they want a refresher of the story.

However, manual text summarization can be expensive in both monetary and time aspects since companies need to hire people to write the summaries and invest time to create them. This is especially true for dialogue-heavy games such as visual novels or role-playing games. With the increase in Natural Language Processing (NLP) models, many language tasks can be automated. One of these tasks is automatic text summarization. Using automatic summarization models to summarize the text, the cost of creating summaries can be reduced. However, automatic summarizers can have the drawback of not summarizing all the important points of the dialogue as well as factual inconsistency, also known as hallucination [1] [16].

Automatic text summarizers already exist. An example is being able to summarize news articles very well [14]. The amount of text for video game dialogue is, on the other hand, often larger in quantity compared to news articles, as some video games have scripts that include thousands of lines of dialogue. Furthermore, dialogues within video games can be more complex as there is the possibility of having dialogue trees or conditional choices, increasing the complexity of creating these summaries.

The complex nature of video game dialogue makes it an interesting point of research for dialogue summarization. This is especially the case compared to the summarization of news articles, which often focus on shorter dialogue. This paper is researching the feasibility of automatic dialogue summarization for video game dialogue to see if this can possibly be a viable method to be implemented in real-time for video games. This is accomplished by looking at the complexity of the summarization and by looking at the quality of the generated summaries made by NLP models.

This paper explores important factors in text summarization in video game dialogue. This is accomplished by first comparing different models found on Huggingface and selecting the best possible candidates for further experimentation. Secondly, we evaluate the impact of how input can influence the quality of the generated summaries, such as how the input text is formatted or how removing names from dialogue can have an impact. Finally, we will look at how much text should be inputted for a summarization. This compares short and long passages and determines if this has an influence on summary performance.

To determine the performance of the text summarization, the generated summaries are compared to humanly made reference summary. The performance of the generated summary will be compared with the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric [15].

The research questions this paper will focus on are the following: *What are important parameters/features for general video game dialogue summarization using NLP?* and *What are challenges with text summarization for video game dialogue?*

1.1 Paper outline

In Section 2, the history of text summarization is given as well as information on different types of Natural Language Models. In Section 3, information about the data used for the experimentation is given. In Section 4, information about the performance metric is included. Furthermore, this section also includes an initial test for selecting models and gives information about the preprocessing steps planned for experimentation. In Section 5, the results of the experiments are shown. In Section 6, the results found are discussed. In Section 7, an overall conclusion is given. In Section 8, potential future work is discussed.

2 Related Work

Text summarization is the task of taking some input text, which is often a long passage, and trying to condense the content of the passage into a shorter passage. Automatic text summarization uses computation to shorten a passage of text. With the advent of Natural Language Processing (NLP), the process of automatic text summarization has become easier, more reliable, and more widespread [10].

There are two main methods for text summarization that are commonly used. The first method to summarize text is extractive text summarization [12][18]. This method attempts to identify significant sentences within a passage and then adds them to the summary. This summarization method will contain exact sentences from the original text. This can be an accurate way to summarize text. However, the summarization might not be written in a natural way. The main benefit of extractive text summarization is that it can be a good way to summarize text automatically without losing the original text.

The second method to summarize text is Abstractive Text Summarization [17][19]. This method attempts to identify important sections, interpret the context, and intelligently generate a summary. This method generates a text summarization that might be more natural. However, the downside is that some information might be lost in the generated summarization. Furthermore, there is a chance of hallucination [1][16]. This leads the generated summarization to include facts that are not present in the original text, which causes the summarization to be incorrect. This research looks at models that generate abstractive summaries as the summaries generated are more naturally written.

Transformers models have been at the forefront of NLP models since the release of Attention is All You Need [26]. This, together with the advent of the Bidirectional Encoder Representations from Transformers (BERT) [4], caused an increase in NLP models. Bert and its related off-shoots are encoder-type models that are good for classification such as sentiment analysis and sequence labeling such as named entity recognition. There are also decoder models such as the Generative Pre-Training Transformer (GPT) and family [22], which are mainly used for text generation. However, when decoder models become large enough, they can also be used for text summarization. Finally, encoder-decoder models such as Text-to-Text Transfer Transformer (T5) [23] and PEGASUS [27] can be used for text summarization and translation.

There has already been prior work related to dialogue summarization. This includes A Survey on Dialogue Summarization: Recent Advances and New Frontiers [9], which looked at the summarization of real-world dialogue such as email or customer service dialogue. Furthermore, there exist datasets with dialogue and a reference summary that can be used to help fine-tune a model, such as DialogSum [3] or Booksum [13]. The DialogSum dataset

focuses on real-world conversation, while the Booksum dataset focuses on summaries for books. Video game dialogue does not necessarily follow either dataset completely, but looking at models that are already fine-tuned on these datasets can give insight into the summarization of video game dialogue.

3 Data

The video game dialogue dataset used during the experiment was retrieved from the The Video Game Dialogue Corpus by Rennick and Roberts dataset [25]. The dataset includes over 6 million words of dialogue from mainly games in the role-playing game genre. This genre of video game is often long and dialogue-heavy. This makes it a good genre to try and summarize the dialogue for, as the large amount of dialogue can cause people to struggle with remembering the entire story. For this reason, the dataset was chosen to try and summarize the dialogue. The dataset includes multiple metadata tags. The ones important for the research are the following tags:

- **game**: Full name of a game.
- **series**: Name of the game series (e.g., "Final Fantasy").
- **character groups**: Mapping from group names to a list of character names who are members of that group

The dataset includes multiple different franchises of video games. Within a video game franchise, there is a possibility to include multiple games from within the same franchise. The dialogue for a video game is stored in JSON format and can be found within the previously mentioned franchise and game folders. The keys within the data are a field named *text*, and the values of the keys are each a line of dialogue. These values are a chronologically ordered list of dialogue occurrences for a given game. Each value has an extra selection of keys within. These keys are one of the following:

- **actions**: dialogue not said by any character
- **character_name**: dialogue said by a character (key is the character_name of the dialogue)
- **location**: the location where the story currently takes place
- **choice**: when there is a possibility for branching choices.
- **status**: includes additional context as character joining party..

3.1 Data Retrieval

The dataset can be retrieved by following the instructions on the GitHub page provided by Rennick and Roberts [24]. The repository includes a script that automatically scrapes websites to retrieve the data and format it in the above-mentioned JSON form. The games are split into franchise folders, and within a franchise folder, games within the franchise can be found, which includes the JSON data. It is to note that some video game dialogue were unable to be retrieved, as the source for the video game data has been removed or the format has been altered on the website where the data was scraped.

3.2 Game selection

As the dataset includes many different games, a choice had to be made on which games the experiment should be conducted on. For the experiment, games from the Final Fantasy franchise had been chosen for further research. There are multiple reasons this game franchise was chosen to conduct the experiments on. The first reason is that the Final Fantasy franchise included many games. This makes it possible to verify the results within one game and see if the performance is similar to other games within the same franchise. The second reason the Final Fantasy franchise was chosen is due to the fact that all games include similar themes. This includes things such as the magic system or animals in the world, such as Chocobo's. This also includes the fact that the games within the franchise are fantasy-based. This makes the comparison between the games more fair, as the summaries created are for games that are very similar. The third reason is that Final Fantasy includes lots of dialogue, which makes it an interesting point to try and summarize all the dialogue as not every single game has much dialogue to summarize, such as the Legend of Zelda.

3.3 Branching choices

As mentioned before, it is possible for some games to include branching choices within the dialogue. Final Fantasy games also include some branching choices. For the experiment, only the first choice is considered for the summary. This means that if there is a branching path with five choices, only the first choice will be chosen to be summarized. The reason for this choice is that the Final Fantasy games are very linear, and the choices within the game have little influence on the overarching story of the game. By only choosing the first possible choice, the experiments can be simplified. If these summaries want to be applied in a real-world playthrough of a game, the summary can be made in real-time as a user makes a choice or the choice can be saved to generate a summary at a later point.

4 Method

This section includes information about the performance metrics used, models selected for experimentation, and any preprocessing applied on the data. In Figure 1, the outline of how to summarization are made is laid out. This figure shows how we plan to generate the summaries. The first decision is to decide how much dialogue is used to summarize each part. Next, a preprocess to apply for the summarization is selected. Finally, a summarization model is used to summarize the dialogue, resulting in a generated summary.

4.1 Performance metrics

The performance metric used to determine the performance of the generated summaries is Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [15]. ROUGE is a metric that is often used to determine the performance of summaries generated by models and gives an indication of the accuracy of a generated summary. ROUGE works by looking at overlapping N-grams between a reference summary and a generated summary. The N-grams can be of any arbitrary length (N), such as an unigram (1) or a bigram (2) up to

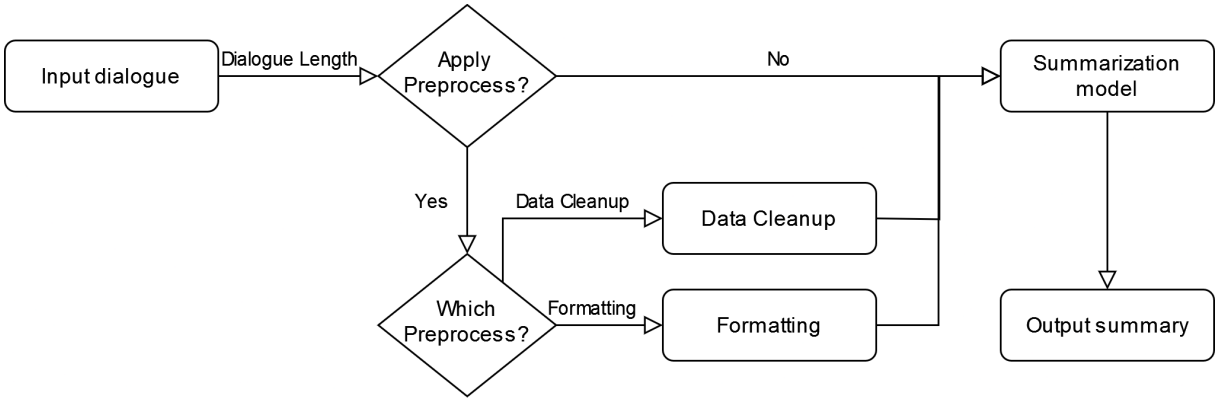

Figure 1: Summarization Process Flowchart

an N-gram of length N. The more overlapping N-grams there are between the reference summary and the generated summary, the better the generated summary is.

This paper will focus on the performance of ROUGE-1 and ROUGE-2. ROUGE-1, as mentioned above, is about having overlapping unigrams between a reference summary and a generated summary. ROUGE-1 is an important metric to see if the generated summary includes the general context of the reference summary. When there is a high ROUGE-1 score, there is a high overlap in context between a reference summary and a generated summary. ROUGE-2, as mentioned above, focuses on overlapping bigrams. A high ROUGE-2 score indicates more fluency for the generated summary and is preferable when the sentence structure of the generated summary is desired to be similar to a reference summary. For this research, a high ROUGE-1 is preferred, as the goal is to summarize as much context as possible without losing information. A high ROUGE-2 score would be an extra benefit. There is also ROUGE-L, which matches the longest sequence of matching words between a reference and generated summary with gaps allowed between the matching words. A high ROUGE-L shows that a generated summary has a similar sentence structure as a reference summary.

4.1.1 Example ROUGE-2 calculation

The following shows an example of calculating the ROUGE-2 value. ROUGE-2 looks at the overlap between word pairs between a reference summary and the generated summary. This means that the bigrams for both the reference and the generated summary have to be found first. Firstly, we start by getting a proposed reference summary and a generated summary:

- **Reference summary**: The cat is on the mat
- **Generated summary**: The cat and the dog

Now the bigrams for both the summaries can be retrieved:

- **Reference bigrams**: the cat, cat is, is on, on the, the man
- **Generated bigrams**: the cat, cat and, and the, the dog

Finally the overlapping bigrams are compared to each other according to equation 1.

$$ROUGE - N = \frac{number\_overlaping\_N\_grams}{total\_N\_gram\_reference} \tag{1}$$

In this case there exist only one overlap in bigrams being "the cat" and with a total number of 5 bigrams in the reference summary gives the following ROUGE score:

$$ROUGE - 2 = \frac{number\_overlaping\_2\_grams}{total\_2\_gram\_reference} = \frac{1}{5} = 0.20$$

4.2 Model selection

For the experiments on text summarization, a model needed to be selected. Many summarization models are publicly available at Huggingface and are easy to setup and use, as well as being fine-tuned for specific tasks by the community. From the available models, models were selected that were fine-tuned for either booksum, a collection of datasets for long-form narrative summarization, or dialogsum, a dataset focusing on casual dialogue conversation summarization. The models have been fine-tuned to summarize large passages of text. These fine-tuned models can be used for video game dialogue, even though the dialogue is not entirely similar. The dialogue within video games is similar in a sense of narrative as well as quantity. The selected models are all abstractive summarization models. The reason for this choice is that the summaries generated will be more natural. The selected models are the following:

- pszemraj/led-large-book-summary [20]
- Falconsai/text_summarization [5]
- chanifrusydi/t5-dialogue-summarization [2]
- pszemraj/long-t5-tglobal-base-16384-book-summary [21]
- gauravkoradiya/T5-Finetuned-Summarization-DialogueDataset [11]

To determine the best model, a small experiment is run to test the performance of the summarization models on video game dialogue. The dialogue used for the experiment is the first sequence from the video game Final Fantasy VII, consisting of 83 lines of dialogue from different characters. As mentioned in Section 3, the dataset includes multiple games from the Final Fantasy franchise. This allows the experiments to be run on multiple different games that are not only similar in the same genre but also have similar themes. Furthermore, early Final Fantasy games were very linear in story due to the limitations of the technology of consoles at that time. This allows the summarization to deal with fewer branching paths compared to more recent games. The data itself is sent into the models without preprocessing. The models will be determined based on the metrics ROUGE-1, ROUGE-2, and ROUGE-L [15]. A reference summary is used to compare the generated summary with [8]. The results of the experiment are visible in Table 1.

The model pszemraj/led-large-book-summary [20] stands out when looking at the ROUGE-1 score. ROUGE-1 focuses on finding overlaps between unigrams with a reference summary and a generated summary and indicates a high degree of context being captured. This means that roughly 28% of the words from the generated summary are also in the

Model Name	ROUGE-1	ROUGE-2	ROUGE-L
led-large-book-summary [20]	**0.2809**	0.0499	**0.1606**
text_summarization [5]	0.1955	0.0190	0.1121
t5-dialogue-summarization [2]	0.1584	0.0317	0.0945
long-t5-tglobal-base-16384-book-summary [21]	0.2353	0.0446	0.1266
T5-Finetuned-Summarization-DialogueDataset [11]	0.2363	**0.1468**	0.1468

Table 1: Performance of models on first dialogue sequence of Final Fantasy VII

reference summary. The models pszemraj/long-t5-tglobal-base-16384-book-summary [21] and gauravkoradiya/T5-Finetuned-Summarization-DialogueDataset [11] also perform well. However, they are both unable to find as many unigrams as pszemraj/led-large-book-summary [20] and thus capture less context. The models Falconsai/text_summarization [5] and chanifrusydi/t5-dialogue-summarization [2] perform the worst out of the models for ROUGE-1.

The model gauravkoradiya/T5-Finetuned-Summarization-DialogueDataset [11] performs the best when looking at ROUGE-2. ROUGE-2 looks at the amount of overlap between bigrams in the reference summary and the generated summary. All other models struggle to find many bigrams present in the reference summary. With ROUGE-2, word ordering is more important. This can indicate more fluency in the generated summary compared to ROUGE-1.

Finally, when looking at ROUGE-L pszemraj/led-large-book-summary [20] performs the best. ROUGE-L looks at the longest common sub-sequence between a reference summary and the generated summary and indicates similar sentence structure with a reference summary.

For further experimentation, the model pszemraj/led-large-book-summary [20] will be used, since performance in both ROUGE-1 and ROUGE-L was the best. The model gauravkoradiya/T5-Finetuned-Summarization-DialogueDataset [11] will also be used for testing, as this model was the best model when looking at the ROUGE-2.

4.3 Data Preprocessing

As mentioned in section 3, the data retrieved is in JSON format. As the data only includes one key *text*. The dialogue can be extracted from this key by looping through the *text* key. The data can then be preprocessed per line of dialogue. This first step of preprocessing includes the removal of unnecessary characters. These characters are **{** and **}** and are used to separate the lines of dialogue. However, for the summary, these characters are unnecessary as they provide no additional context to the dialogue and are removed. Any extra blank spaces are also removed during this step, as these similarly do not provide extra information about the dialogue. Next, the **:** is removed. The semicolon is used to separate an action, such as a character name, from the dialogue. This step is done to allow for formatting changes within the dialogue.

For the experiments, two more preprocessing steps are planned. The first, as mentioned above, is that dialogue is split from a character. This allows the dialogue to be summarized without specifically mentioning by which character it has been said. This preprocessing step therefore removes the names to see if the performance changes with the removal of some context. The second preprocessing step includes names explicitly to summarize the

dialogue in a similar format as with book writing. Since, one of the selected models was fine-tuned for book summarization. Seeing if formatting the dialogue in a similar way might show interesting results. An example of the dialogue for different preprocessing steps are the following:

- **Raw Data**: ["Barret": "There IS a way! Look! What's that look like?"]
- **Names removed**: "There IS a way! Look! What's that look like?"
- **Names explicitly**: "There IS a way! Look! What's that look like?", says Barret.

Finally, we also conduct a test by varying the amount of dialogue inputted in the summarization model. Calling the model more often can generate more detailed results. However, it is unclear if the performance of the summary will change based on this. This is not necessarily a preprocessing step, but mainly looks at the performance changes when more detailed or concise summaries are tried to be made.

5 Results

This section includes the results found using the text summarization models on different Final Fantasy data. Firstly, the two best performing models found in Section 4, are compared using different preprocesses and their influence on the performance of the summarization for dialogue in Final Fantasy VII. Secondly, the best model is then used on different Final Fantasy games to validate the findings.

5.1 Final Fantasy VII

The best-performing models found in 4.2 are tested using different preprocesses. The first preprocess includes no preprocessing and inputs the raw data into the summarization model. The second preprocess cleans the data. However, this preprocess removes the names of characters saying the dialogue. The final preprocess adds names explicitly after the cleanup process. The experiments are run with varying amounts of input tokens. ROUGE is used as the performance metric, and a reference summary is used [8]. The results for the experiments with 4096 input tokens are visible in Table 2,3 and the results for the experiments with 2048 input tokens are visible in Table 4,5.

5.1.1 4096 tokens

Firstly, the model performance using 4096 tokens are compared. As with the initial testing with the models in Section 4.2 the led-large-book-summary model performed better for ROUGE-1 compared to the T5-Finetuned-Summarization-DialogueDataset model. This shows that the led-large-book-summary is able to retrieve more unigrams from the reference summary. The higher ROUGE-1 score also means that the led-large-book-summary was able to capture more context of the reference summary compared to the T5-Finetuned-Summarization-DialogueDataset model. The T5-Finetuned-Summarization-DialogueDataset also had high ROUGE-1 scores, especially when looking at ACT 4. However, the model is still outperformed by led-large-book-summary.

For ROUGE-2, both models performed worse. This means that the models were unable to retrieve many bigrams compared to the reference summary. This shows that for longer

pszemraj/led-large-book-summary

4096		Act 1			Act 2	
Preprocess	ROUGE1	ROUGE2	ROUGEL	ROUGE1	ROUGE2	ROUGEL
None	0.216	0.018	0.126	0.307	0.040	0.163
No_names	0.303	0.015	0.193	0.243	0.023	0.171
Names Explicit	0.250	0.022	0.130	0.327	0.040	0.136
		Act 3			Act 4	
Preprocess	ROUGE1	ROUGE2	ROUGEL	ROUGE1	ROUGE2	ROUGEL
None	0.270	0.034	0.162	0.343	0.073	0.195
No_names	0.302	0.037	0.188	0.256	0.054	0.146
Names Explicit	0.364	0.053	0.184	0.258	0.047	0.142

Table 2: ROUGE scores for different acts in Final Fantasy VII using pszemraj/led-large-book-summary and with 4096 input tokens

gauravkoradiya/T5-Finetuned-Summarization-DialogueDataset

4096		Act 1			Act 2	
Preprocess	ROUGE1	ROUGE2	ROUGEL	ROUGE1	ROUGE2	ROUGEL
None	0.143	0.027	0.133	0.182	0.014	0.127
No_names	0.217	0.020	0.152	0.120	0.025	0.091
Names Explicit	0.128	0.000	0.104	0.133	0.052	0.101
		Act 3			Act 4	
Preprocess	ROUGE1	ROUGE2	ROUGEL	ROUGE1	ROUGE2	ROUGEL
None	0.144	0.038	0.112	0.291	0.074	0.182
No_names	0.143	0.062	0.107	0.230	0.036	0.142
Names Explicit	0.195	0.034	0.133	0.133	0.018	0.109

Table 3: ROUGE scores for different acts in Final Fantasy VII using gauravkoradiya/T5-Finetuned-Summarization-DialogueDataset and with 4096 input tokens

summaries, the actual fluency is different when comparing the generated summary to the reference summary. For the initial testing, led-large-book-summary already had a low ROUGE-2 score. However, it is notable that T5-Finetuned-Summarization-DialogueDataset initially was able to capture some fluency compared to a reference summary with less dialogue being summarized. However, be unable to when a large quantity of dialogue is summarized.

Finally, for ROUGE-L the models both perform similarly. The led-large-book-summary does get a better score compared to T5-Finetuned-Summarization-DialogueDataset in more cases. However, the actual scores remain close in most cases for both the models.

When looking at specific preprocesses, no clear best performing preprocess is found. Depending on the act, any preprocess might result in a higher ROUGE score. This indicates that the performance of the summary can be influenced by the type of dialogue within each act. And therefore, the generated summaries can also vary when changing games. However, it is important to note that in most acts, one of the two tested preproccess had the highest results compared to using no preprocessing. This shows signs that, even though adding preproccessing is not always beneficial, it did generate the best performing summaries in most cases.

pszemraj/led-large-book-summary

2048		Act 1			Act 2	
Preprocess	ROUGE1	ROUGE2	ROUGEL	ROUGE1	ROUGE2	ROUGEL
None	0.252	0.013	0.172	0.329	0.037	0.167
No_names	0.254	0.036	0.134	0.234	0.030	0.130
Names Explicit	0.213	0.017	0.164	0.324	0.040	0.173
		Act 3			Act 4	
Preprocess	ROUGE1	ROUGE2	ROUGEL	ROUGE1	ROUGE2	ROUGEL
None	0.203	0.030	0.133	0.198	0.034	0.135
No_names	0.239	0.032	0.153	0.325	0.060	0.165
Names Explicit	0.311	0.041	0.159	0.386	0.109	0.281

Table 4: ROUGE scores for different acts in Final Fantasy VII using pszemraj/led-large-book-summary and with 2048 input tokens

gauravkoradiya/T5-Finetuned-Summarization-DialogueDataset

2048		Act 1			Act 2	
Preprocess	ROUGE1	ROUGE2	ROUGEL	ROUGE1	ROUGE2	ROUGEL
None	0.152	0.022	0.130	0.182	0.035	0.127
No_names	0.211	0.043	0.168	0.139	0.054	0.133
Names Explicit	0.115	0.022	0.115	0.171	0.019	0.115
		Act 3			Act 4	
Preprocess	ROUGE1	ROUGE2	ROUGEL	ROUGE1	ROUGE2	ROUGEL
None	0.071	0.000	0.071	0.170	0.040	0.170
No_names	0.098	0.047	0.091	0.250	0.062	0.167
Names Explicit	0.210	0.036	0.177	0.321	0.074	0.214

Table 5: ROUGE scores for different acts in Final Fantasy VII using gauravkoradiya/T5-Finetuned-Summarization-DialogueDataset and with 2048 input tokens

5.1.2 2048 tokens

Next, the experiment is run with 2048 tokens for the summarization models. Since fewer tokens are used for the summary, more detailed summaries can be made. This experiment again shows that led-large-book-summary model outperforms T5-Finetuned-Summarization-DialogueDataset model. The ROUGE scores specifically for ROUGE-1 are again higher for the led-large-book-summary model for each preprocess. As with the experiment using 4096 tokens, both models perform worse for ROUGE-2 and have similar performance for ROUGE-L.

When comparing the different ROUGE scores for the different preprocesses to 4096, it becomes visible that the amount of dialogue has some influence on the performance of the summarization. However, this influence is small, as the actual ROUGE scores do not differ much and the experiments for both 4096 tokens and 2048 tokens performed similarly. Some cases show that 4096 tokens perform better, while other cases show 2048 tokens performing better. This is a good indication that it is possible to generate shorter and more concise summaries by increasing the amount of tokens inputted into the model without losing much performance or to generate longer more detailed summaries by decreasing the amount of input tokens.

5.2 Final Fantasy II/III

Finally, we look at the best performing model pszemraj/led-large-book-summary from the previous experiments and the performance of the summaries from the same franchise as Final Fantasy VII. The games tested are Final Fantasy II and Final Fantasy III. As these games are from the same franchise as Final Fantasy VII, they are similar in themes and are used to validate the results from the best-performing model. The summaries generated use 4096 tokens for the summary, as the performance was similar compared to 2048 tokens. Since the performance was similar, it was decided that the shorter summary would be preferred for the generated summary. Reference summaries for both Final Fantasy II and Final Fantasy III used to calculate the ROUGE scores are retrieved from FANDOM [6][7]. The results of the generated summaries for Final Fantasy II and Final Fantasy III are visible in Table 6.

pszemraj/led-large-book-summary

4096		FFII			FFIII	
Preprocess	ROUGE1	ROUGE2	ROUGEL	ROUGE1	ROUGE2	ROUGEL
None	0.247	0.021	0.145	0.312	0.095	0.201
No_names	0.308	0.029	0.162	0.185	0.023	0.159
Names Explicit	0.229	0.029	0.138	0.269	0.038	0.167

Table 6: ROUGE scores for Final Fantasy II & Final Fantasy III using pszemraj/led-large-book-summary and with 4096 input tokens

The results for both Final Fantasy II and Final Fantasy III show the same behavior as Final Fantasy VII. The ROUGE-1 scores stay in a similar range around $[0.200 - 0.300]$ as with the prior experiment. This again shows that the model was able to capture a similar amount of context as before and indicates that the model might be able to be used for more video games of a similar genre. The model also performs worse when looking at ROUGE-2 for these games. This again indicates that the generated summaries have a different fluency compared to the reference summary. The ROUGE-L also stayed in a similar range as the prior experiment.

This experiment also does not show a clear best preprocess. All of the preprocessing performed similarly for both Final Fantasy II and Final Fantasy III. The only preprocess that performed worse was removing names as the preprocess for Final Fantasy III. However, it is worth noting that the same preprocess performed the best for Final Fantasy II.

6 Discussion

This section includes a discussion of the results found in Section 5. The discussion includes the performance of the two tested models on Final Fantasy VII, the performance metrics ROUGE-1 and ROUGE-2, the performance of the best-performing model on Final Fantasy II and III, and lastly, a discussion about the difficulty of summarizing video game dialogue.

6.1 Model Performance on Final Fantasy VII

The main observation from the results showed that between the two models tested, the led-large-book-summary [20] performed better in general compared to T5-Finetuned-

Summarization-DialogueDataset [11] in generating summaries when looking at ROUGE-1 scores. This observation was already present when selecting the models. However, the test used to select the models summarized a very limited amount of dialogue. Compared to the complete experiment, where all the dialogue was summarized. Running the experiments with all the dialogue showed that led-large-book-summary performed better for summarizing longer passages, such as video game dialogues. The likely reason for the discrepancy is due to the way both the led-large-book summary and T5-Finetuned-Summarization-DialogueDataset models were fine-tuned. The T5-Finetuned-Summarization-DialogueDataset was fine-tuned using dialogsum [3], while led-large-book-summary was fine-tuned on booksum [13]. Both datasets include examples of dialogue and how it should be summarized. However, summaries found in dialogsum focus on real-world dialogue. This includes topics such as buying a house, planning a vacation, and movie discussions. On the other hand, summaries for booksum include examples of long passages from novels and books. The discrepancy in performance between the two models can probably be attributed to this fact. Final Fantasy dialogue mainly consists of the state of affairs within the video game world. The booksum dataset aligns more with this type of data compared to the dialogsum dataset, which focuses on casual talk.

6.2 ROUGE-1

The performance of the best-performing model, pszemraj/led-large-book-summary, showed ROUGE-1 scores around the range of $[0.20 - 0.35]$. ROUGE-1 is important to show that the generated summary contains similar context as a reference summary. The calculated ROUGE-1 scores show that the generated summary was able to find the context of the reference summary. This shows promise, especially since the generated summary and the reference summary are both long. However, the summaries generated by the model did not manage to find the majority of all the context in the reference summary. The amount of context summarized should still be enough information for a user to have a grasp of the story.

A possible way to improve the ROUGE-1 score in the future is to use an extractive summarizer instead. The models used are of an abstractive variant, meaning the model tries to understand the context of the dialogue and make a more humanly readable summary. This leads to more fluently coherent summaries. However, this can also cause the model to hallucinate, causing it to think about things that are present but not important. If a hallucination is summarized, the ROUGE-1 score will decrease since the hallucination would not be present in the reference summary. An extractive model does not have this issue as only the actual sentences in the dialogue are summarized word for word, giving it a higher possibility of capturing more of the context. It is important to note that using an extractive summarizer may not necessarily lead to a better ROUGE-1 score, as the summaries created this way will be less fluent compared to a reference summary. Furthermore, with the downside of being less fluent, the summaries generated will likely not be of a quality that a user might want to read it.

6.3 ROUGE-2

Notable results during the experiment showed that both models did not perform well when looking at ROUGE-2. ROUGE-2 focuses on the overlap between bigrams of words between

a reference summary and a generated summary. Having a higher ROUGE-2 is preferred to show that the generated summary has similar fluency to a reference summary. However, the summaries generated by the tested models have low ROUGE-2 scores, indicating low fluency compared to the reference summary. There are multiple reasons for the low ROUGE-2 score.

The first possible reason for the low ROUGE-2 scores is that both the reference summary and the generated summaries are long, with both containing multiple hundreds of words, up to multiple thousands of words. Because of the length of the summaries, many bigrams are possible for both the reference and the generated summary. This increases the chance that that will me a missmatch between bigrams between a reference summary and a generated summary.

The second possible reason for the low ROUGE-2 scores is due to the method of actual text summarization. The models only receive a certain amount of tokens before they produce the summary, and the model will only make a summary for those tokens. However, this method leads to possible fluency errors, as something might be mentioned in the prior summarized section but may have been relevant for the upcoming section. The summarization model will be unable to fluently combine the two sections together since the summaries are made independently of each other.

The final possible reason for the low ROUGE-2 scores is that there is a mismatch between the contexts of the reference summary and the generated summary. This means that the generated summary was unable to find all the relevant plot points in the reference summary. This might be true for these models. However, the ROUGE-1 scores showed that a portion of the context was present. This makes it likely that the context mismatch is not the main problem. However, it might still have an influence. Since the ROUGE-1 scores did not retrieve all of the context.

6.4 Performance on Final Fantasy II and Final Fantasy III

When looking at the results specifically for Final Fantasy II and Final Fantasy III, a difference in the best-performing preprocessing is noticeable. For Final Fantasy II, the best-performing preprocessing was removing the name. On the other hand, for Final Fantasy III, this was the worst-performing preprocess. Final Fantasy III had the best performance when no actual preprocessing was applied for the summary. The most likely reason why Final Fantasy III has the worst performance when removing the names is most likely due to the type of dialogue within the game. First of all, Final Fantasy III has comparatively few lines when compared to Final Fantasy II, which has around 40% more lines of dialogue. Secondly, most of the dialogue is said by actual characters in Final Fantasy III, with around 5% of the dialogue not being said by any character. Compared to Final Fantasy II, where around 18% of the dialogue is not said by any character. Since in Final Fantasy III there is already very little context due to the fewer amount of dialogue as well as most of the dialogue being spoken by characters, when the character names are removed from the dialogue, even less context is available for the model to help understand the actual story and therefore produce a worse summary.

6.5 Difficulty with dialogue summarization

The experiments showed some promising results for video game dialogue summarization. However, there were some difficulties left when summarizing these types of dialogue.

The first difficulty with summarizing dialogue is that the experiments showed that the input method of the data in the models has an influence on the performance of the generated summaries. The results showed that including some preprocessing gave the best results in the generated summaries in most cases and is suspected to be the general best method for summarizing the dialogue. It is to note that it is not always clear which preprocess of removing names or adding names explicitly is the best. However, there were few cases where no preprocessing at all showed similar performance.

The second difficulty with summarizing dialogue is the way the data is structured. In the case of Final Fantasy, the dialogue is linear, which allows for easy summarization. However, there are also games such as Hades that require certain conditions to be met before new dialogue will be used. This is an extra layer of complexity that has to be accounted for in certain games when trying to summarize the dialogue. Finally, there is a particularity to the type of story-telling in video games. Since video games often have some visuals to accompany the dialogue, less context needs to be mentioned within the dialogue. This can make it more difficult for a dialogue summarizer.

7 Conclusions

This paper looked at automatic dialogue summarization for video game dialogue using NLP models and whether they might be viable to be used in real-world applications. To determine the effectiveness of the summaries generated by the summarization models, experiments were conducted with various preprocessing and NLP models to see the impact of these changes on the generated summaries.

Firstly, to answer the research question *What are important parameters/features for general video game dialogue summarization using NLP?*. The results showed that including some preprocessing can be beneficial for generating better summaries. It was true that in most cases adding some preprocessing generated better summaries. However, the results also showed that, in some cases, adding no preprocessing and using raw data generated the best summaries.

This leads into the second research question *What are challenges with text summarization for video game dialogue?*. There are many ways, as well as many models, available to try and summarize video game dialogue. There are a multitude of ways to format the dialogue, which can change the performance of the summaries. Also, with many different models available, finding the one best suited for the situation proved to be difficult. A best summarization model was found for the experiments tested. However, there might be different models available that prove to be better for generating summaries for video game dialogues that have not been looked at. This makes it difficult to confidently conclude that the method found was the best method to try and summarize video game dialogue. However, the results did show that the models were able to generate summaries that capture the same context as a reference summary.

The reason for a summary is to have a more condensed version of a story available. Even though the generated summaries did not include all the context of a reference summary, they managed to summarize major parts of the story. Even with the difficulty in creating

summaries, this research shows that automatic summarization can possibly be used for summaries in video games, as the context of the dialogue was captured. The generated summaries might also be used as a starting point for a human-made summary. This can remove a large part of the labor involved in making summaries from scratch and therefore reduce costs. And including automated summaries might prove to be a useful addition in future video games with large quantities of dialogue.

8 Future Work

8.1 Additional models

There exist many different types of NLP models at looking at more different models might be interesting in the future. This does not have to be exclusively be one of the models available on Huggingface. The models can also include GPT-type models With the multitude of models available on Huggingface, there is a high likelihood that there are some models that might perform better for this task. Since only a handful of models have been looked at during this research.

Another possibility with different models is to look at fine-tuning an encoder-decoder model such as T5 [23]. The models selected for the experiment were fine-tuned for longer dialogues, such as books, which managed to generate summaries for the tested games. With extra fine-tuning on the tested models or on another model, the performance might be able to be increased.

8.2 ChatGPT

As mentioned above, looking at different models might be interesting. With the increased popularity of ChatGPT, we conducted a small experiment to see how well the ChatGPT model performs in summarizing the Final Fantasy data and if it might be possible to use it in the future to summarize video game dialogue. The ChatGPT model used was the free-to-use ChatGPT-3.5. For the experiment, raw JSON data for Final Fantasy VII Act 1 is used. The following prompt was used in ChatGPT to create the summaries:

- *I have to following lines of dialogue can you summarize this for me it's in JSON format. Assume for choices that the first choice is chosen. Write the summary in a human-readable format.*

Due to the token limit within ChatGPT, the act was split into 300 lines of dialogue per prompt. This was large enough to not be limited by the token limit, but not too small that the model had to be called many times, making that experimental test quicker. The results for ChatGPT performance in summarizing Act 1 of Final Fantasy VII are visible in Table 7.

Model Name	ROUGE1	ROUGE2	ROUGEL
ChatGPT-3.5	0.277	0.038	0.176

Table 7: ROUGE scores for ChatGPT-3.5 on Final Fantasy VII act 1 and 300 lines of dialogue per prompt

This experiment shows that ChatGPT can generate similar performance to the tested models. However, a major issue with ChatGPT was that, even given the prompts, the way the ChatGPT model tried to summarize the dialogue was inconsistent. Sometimes it would write it like a small excerpt, and sometimes it would write a summary using bullet points. This is not necessarily a problem, as a summary is created with the general outline of a story. However, it shows the downside of ChatGPT being inconsistent in the way it writes its summary over multiple sections, even when a prompt is given to write in a specific way. It is to note that this limitation might be a limit of ChatGPT-3.5, which is not the newest and most advanced option available for ChatGPT. Newer versions of ChatGPT have been released, such as ChatGPT-4 and ChatGPT-4o, which are larger models. These models might perform better and generate more consistent summaries. Furthermore, the experiment only included raw data and did not look at the other tested preprocesses. These preprocesses showed to give better performance in more cases compared to no preprocessing and could also be looked at in the future using ChatGPT. Finally, there might still be more restrictions that could be given in the prompt to try and format the text in a more consistent method.

8.3 Additional preprocessing steps

For the experiments, three different preprocesses were tested. The tested processes were no processing, the removal of names, and adding names explicitly with varying amounts of total dialogue. Testing different preprocesses might show different results. During the testing, it was found that, in most cases, adding some preproces was beneficial. This raises the question of what influence other preproccesing might have. For example, instead of summarizing only a fixed amount of dialogue each time, it might prove beneficial to try and dynamically summarize each time a location changes. This method might show better results as context for location is better preserved for each section within a game, instead of having a chance that some sections might miss context since the dialogue got cut off to fit the character limit. However, this will require more manual labor, as within a video game, it might not always be clear when a new section starts, as sometimes the section change is not specifically mentioned or is only visible within the game itself. Another preprocess that could be possible is to add the context explicitly to the dialogue before it is sent to the summarizer. Extra context that could be added includes, for example, who the characters are, what their motives are, or what the location of the current section is. This could alleviate the previously mentioned problem of missing context. However, this again will be costly to do since getting the context in itself might not be straightforward in all cases.

References

[1] Cao, M., Dong, Y., and Cheung, J. C. K. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. *arXiv preprint arXiv:2109.09784* (2021).

[2] chanifrusydi. t5-dialogue-summarization, 2023.

[3] Chen, Y., Liu, Y., Chen, L., and Zhang, Y. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762* (2021).

[4] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Falconsai. text_summarization, 2023.

[6] Fandom. Final fantasy ii synopsis, 2024.

[7] Fandom. Final fantasy iii synopsis, 2024.

[8] Fandom. Final fantasy vii synopsis, 2024.

[9] Feng, X., Feng, X., and Qin, B. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175* (2021).

[10] Gambhir, M., and Gupta, V. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review 47* (2017), 1–66.

[11] gauravkoradiya. T5-finetuned-summarization-dialoguedataset, 2023.

[12] Gupta, V., and Lehal, G. S. A survey of text summarization extractive techniques. *Journal of emerging technologies in web intelligence 2*, 3 (2010), 258–268.

[13] Kryściński, W., Rajani, N., Agarwal, D., Xiong, C., and Radev, D. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209* (2021).

[14] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[15] Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (2004), pp. 74–81.

[16] Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661* (2020).

[17] Moratanch, N., and Chitrakala, S. A survey on abstractive text summarization. In *2016 International Conference on Circuit, power and computing technologies (ICCPCT)* (2016), IEEE, pp. 1–7.

[18] Moratanch, N., and Chitrakala, S. A survey on extractive text summarization. In *2017 international conference on computer, communication and signal processing (ICCCSP)* (2017), IEEE, pp. 1–6.

[19] Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023* (2016).

[20] Peter Szemraj. led-large-book-summary (revision 38be53c), 2022.

[21] Peter Szemraj. long-t5-tglobal-base-16384-book-summary (revision 4b12bce), 2022.

[22] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training.

[23] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research 21*, 1 (2020), 5485–5551.

[24] Rennick, S., and Roberts. The video game dialogue corpus. `https://github.com/seannyD/VideoGameDialogueCorpusPublic`, 15-11-2023.

[25] Rennick, S., and Roberts, S. G. The video game dialogue corpus. *Corpora* (2023).

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

[27] Zhang, J., Zhao, Y., Saleh, M., and Liu, P. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning* (2020), PMLR, pp. 11328–11339.