IDN-Sum: A New Dataset for Interactive Digital Narrative Extractive Text Summarisation

Ashwathy T Revi and **Stuart E. Middleton** and **David E. Millard**
University of Southampton
University Rd, Highfield, Southampton SO17 1BJ

Abstract

Summarizing Interactive Digital Narratives (IDN) presents some unique challenges to existing text summarization models especially around capturing interactive elements in addition to important plot points. In this paper we describe the first IDN dataset (IDN-Sum) designed specifically for training and testing IDN text summarization algorithms. Our dataset is generated using random playthroughs of 8 IDN episodes, taken from 2 different IDN games, and consists of 10,000 documents. Playthrough documents are annotated through automatic alignment with fan-sourced summaries using a commonly used alignment algorithm. We also report and discuss results from experiments applying common baseline extractive text summarization algorithms to this dataset. Qualitative analysis of the results reveal shortcomings in common annotation approaches and evaluation methods when applied to narrative and interactive narrative datasets. The dataset is released as open source for future researchers to train and test their own approaches for IDN text.

1 Introduction

Automatic summarization has often been been studied for domains such as news and scientific reports. While there is some work on narratives like movies and books, there is limited work surrounding automatic summarization of interactive and game narratives. Extrapolating IDN performance from news article summarization results is non trivial due to longer texts and the existence of elements like characters and plot. IDN also differs from movies and books due to the presence of interactivity and game elements that make summarisation of IDN different to that of general text and/or linear narratives. Unlike novel/movie summarization, IDN has the concept of choices, structure and multiple plot lines which also affect the relative importance of sentences. Additionally, IDN text formats vary significantly and can look like novels, movie scripts, gameplay logs, or a mixture of all three.

The IDN-Sum dataset is generated from fan made transcripts of two narrative games, both sourced from Fandom[1] - *Before the Storm* published by Square Enix and *Wolf Among Us* published by TellTale Games. Different simulated playthroughs through the game are generated by implementing a ReaderBot like the one described in (Millard et al., 2018), assuming a different combination of choices for each playthrough. While these two sources account for only one type of IDN (narratives in the form of a Gauntlet, see section 3.1), it takes a step towards increasing resources available for research in this area. An analysis of dataset characteristics and performance of some baseline summarisation methods on this dataset is presented. Novel contributions of this paper are (a) a new text summarization dataset for IDN (IDN-Sum), with abstractive summaries for overall IDN and aligned extractive summaries for multiple IDN playthroughs, and (b) baseline evaluation of standard benchmarks on IDN-Sum and qualitative analysis of the predictions made by them.

2 Related Work

Most text summarization work is targeted at news, academic papers and reviews. The most commonly used summarisation dataset is the CNN/DailyMail dataset which is a collection of news articles and human written summaries (Hermann et al., 2015; Nallapati et al., 2016). Summarisation datasets for narratives include datasets with novel chapters and corresponding human written summaries from online guides, (Chaudhury et al., 2019) (Ladhak et al., 2020), extractive summaries that read like telegraphs(Malireddy et al., 2018), stories and summaries from Wattpad(Zhang et al., 2019a), transcripts and summaries of movies(Gorinski and Lapata, 2015), transcripts of TV shows (Papalampidi

[1] www.fandom.com

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1–12
July 5–10, 2020.
https://doi.org/10.26615/978-954-452-056-4_001

et al., 2020; Chen et al., 2021) and subtitles (Aparício et al., 2016). Papers on game summarisation are few and usually involve game logs from online games like *DOTA* (Barot et al., 2021; Cheong et al., 2008) or commentary from sports(Sandesh and Srinivasa, 2017). However, IDN text is typically more similar to movie scripts or novels than game logs. The critical role dataset (Rameshkumar and Bailey, 2020) is a dataset of transcripts and summaries from critial role episodes. This is a transcript of several voice actors playing a Table top role playing game and hence captures only one playthrough of a narrative. To the best of our knowledge, IDN-Sum is the first dataset for IDN that captures multiple playthroughs of an IDN.

Unsupervised methods for automatic extractive summarisation use several methods to determine the importance of sentences including statistical methods using features like sentence position and TF-IDF, concept based methods that use external databases like WordNet, topic based methods to infer important topics, graph based methods that build intermediate graphs computed through metrics like semantic similarity, semantic methods using techniques like semantic role annotation, optimization methods that involve optimising for constraints (like maximising coverage or minimising redundancy) and fuzzy logic based methods (El-Kassas et al., 2021). Supervised methods include different RNNs and Tranformers, using pretrained models such as Bert for summarisation (Mridha et al., 2021; Liu, 2019). Variations of BertSum(Liu, 2019), SummaRuNNer(Nallapati et al., 2017), MatchSum(Zhong et al., 2020), Discobert(Xu et al., 2020), HiBert(Zhang et al., 2019b), Banditsum(Dong et al., 2018) and neusum(Zhou et al., 2018) are among the most commonly used baselines for extractive summarisation in the past three years. However, most of these were designed for short documents (CNN/DM). Longformer(Beltagy et al., 2020) is an adpatation of BertSum for longer documents. There are also summarisation approaches that are specific to the narrative domain (Gorinski and Lapata, 2015; Tran et al., 2017; Papalampidi et al., 2020).

3 IDN Dataset

3.1 Methodology for Dataset Creation

The IDN-Sum dataset consists of several simulated playthroughs through two narrative games - *Before the Storm* and *Wolf Among Us*. Both of these are narrative games in which the choices made by the player change how they experience the story. Playthroughs are simulated by assuming a different combination of choices each time. The script that generates these playthroughs is referred to as ReaderBot in this paper, following terminology used in (Millard et al., 2018). Both of these have what are referred to as a Gauntlet structure (Rezk and Haahr, 2020) which means the story changes based on player choices but then eventually all paths converge back onto a common storyline making a gauntlet shape. While this is not the only type of IDN, they were chosen based on availability of resources and smallest variation in domain from existing work.

Fan made transcripts and summaries are scraped from Fandom. The transcripts on Fandom contains the script of the game and tabs showing how the dialogue changes based on different options the player might chose throughout the game. This html page is parsed and different playthroughs are then generated by a ReaderBot(Millard et al., 2018) by choosing different combinations of options for each scene. Fandom much like Wikipedia, is a major community site with more than 31 million registered users[2]. Through the authors' own inspection, the summaries were found to be of good quality. The limitations of the ReaderBot, details of implementation and the game mechanics that are supported are described on the Github page[3].

There is only one human authored abstractive summary per episode. We take this overall abstractive plot summary from Fandom and produce extractive summaries for each playthrough using the TransformerSum[4] library. This library follows the method used in (Nallapati et al., 2017) to convert abstractive summaries to extractive summaries by greedily selecting extracts that maximise the ROUGE score with the abstractive summary until the sentence limit is hit or ROUGE score cannot be improved. Summaries were generated with target lengths of 3 (similar to CNN/DM) but also longer target lengths of 9 and 27, since for narrative datasets the source text and reference summaries are much longer. For IDN and CRD3, we also generate target length of 81 since the reference summaries for the these datasets are considerably larger than 27. The human authored abstractive summary

[2] stats taken from https://community.fandom.com/wiki/Special:Statistics
[3] https://github.com/AshwathyTR/IDN-Sum
[4] https://github.com/HHousen/TransformerSum

for each episode is also provided along with the dataset so that annotations can be generated using any alignment algorithm.

3.2 Dataset Characteristics and Comparison

Property	CNN DM	Novel	CRD3	SB	IDN
#docs	280K	4366	159	850	10K
#sents	10M	630K	524K	2M	26K
doc length	40	278	2400	2797	2290
ref length	3.8	24	141	34	72
tokens/sent	21	24	18	11	10
vocab size	681K	115K	53K	202K	10K

Table 1: Dataset Metrics: number of instances in dataset (#docs), number of unique sentences (#sents), average number of sentences in source text (doc length) and human authored reference summary (ref length), average number of tokens per sentence (tokens/sent) and number of words in vocabulary (vocab size) for each dataset

Table 1 compares **IDN-Sum (IDN)** with several other narrative datasets. The Novel Chapter dataset from (Ladhak et al., 2020) is included since it contains narrative elements like plot but is not as structurally different from the CNN/DM as the screenplay datasets. **Scriptbase (SB)**(Gorinski and Lapata, 2015) was chosen for comparison because the IDN text that is generated by the ReaderBot is very similar to screenplays. **Critical Role Dataset (CRD3)**(Rameshkumar and Bailey, 2020) was chosen since this is an example of a kind of interactive narrative, even though it does not show alternate storylines that are possible through the story world. The metrics for CNN/DM (Hermann et al., 2015; Nallapati et al., 2016) dataset is also shown for comparison since this is a widely used dataset by the NLP community for text summarisation. IDN, SB and CRD3 datasets are structured like screenplays so they were preprocessed into a format that captures the structure for consistency. The tag ':SC:' was used to separate scenes, '[EX]' was used to denote beginnings and ends of extracts and 'S0:' was used to denote non dialogue sentences (narration).

As can be observed from the table, CNN/DM has a lot more datapoints than the narrative datasets. The narrative datasets are much longer (refer length of source column). ScriptBase and IDN tend to have shorter sentences than the other datasets. The extractive summaries were generated using the alignment technique described in the last section

Dataset	no filter	stop filter
CNN/DM_3	0.56	0.56
Novel_3	0.31	0.19
Novel_9	0.44	0.29
Novel_27	0.50	0.35
CRD3_3	0.19	0.18
CRD3_9	0.34	0.31
CRD3_27	0.49	0.44
CRD3_81	0.62	0.55
SB_3	0.17	0.09
SB_9	0.3	0.18
SB_27	0.45	0.31
IDN_3	0.08	0.06
IDN_9	0.18	0.14
IDN_27	0.36	0.31
IDN_81	0.56	0.49

Table 2: ROUGE1 F1 scores of automatically aligned extractive summaries (oracle) against human authored abstractive summaries with and without stop words. Target lens 9, 27 and 81 for CNN/DM and 81 for Novel and SB was not generated since these target lengths are much greater than the average length of human written abstractive reference summaries

for target lengths 3, 9, 27 and 81 depending on the average length of the reference summaries (9, 27 and 81 was not run for CNN/DM and 81 was not run for Novel and SB datasets). The ROUGE1 F1 scores of the generated summary against the human written summary are shown in table 2. IDN has lower unique sentences and vocab size because unlike other datasets, the IDN dataset has a lot of overlap in text between datapoints since it contains hundreds of playthroughs of each episode. Since it follows the gauntlet structure, both in *Before the Storm* and *Wolf Among Us* a major portion of the story is present in all branches. This is illustrated in figures 1 and 2. Fig 1 shows the amount of token overlap between one data point in the IDN dataset with all the other data points. A similar graph showing variation in the aligned extractive summaries is also shown. As can be seen in the figure, a set of other data points have high overlap. These are other playthroughs of the same episode where only some parts of the text are different. For comparison, a similar graph is shown from ScriptBase which contains screenplays that are entirely unrelated to each other in fig 2. In this case, all data points have only a small overlap. Examples of the data are shown in Appendix A.

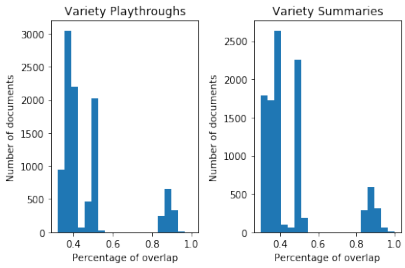

Figure 1: Variety in IDN Dataset

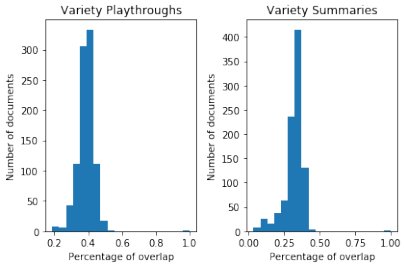

Figure 2: Variety in Scriptbase Dataset

4 Baseline Experiments

4.1 Methods

Baseline models used in this paper represent a good coverage of standard methods used for extractive text summarisation today. The baselines were chosen so that they include two simple baselines, Random-N and LEAD-N, a commonly used unsupervised method, TextRank (Mihalcea and Tarau, 2004) and two neural network based methods (transformer based approaches BertSum(Liu, 2019), Longformer(Beltagy et al., 2020) and an RNN based sequence model, SummaRuNNer(Nallapati et al., 2017)). Out of the popular baselines mentioned in section 2, SummaRuNNer was chosen because it was the most easily extendable to longer documents. BertSum was included since this was the most popular baseline and variation of it for longer documents, Longformer was included so that a more recent model is also included as a baseline. Narrative summarisation models mentioned in section 2 work at a scene level and hence return huge summaries for complete narratives/IDN's, so these methods are not included.

Random-N selects a random N sentences as the summary and Lead-N selects the first N sentences of the source text as its summary where N for each dataset is set to summary lengths 3,9,27 and 81. TextRank is similar to Google's PageRank(Page et al., 1999) algorithm where each sentence is considered in place of web pages. A sentence similarity graph is computed and used to calculate importance of sentences which are then ranked accordingly. For supervised methods, training data for extractive summarisation is generated by automatically aligning abstractive summaries with the original text by greedily maximising ROUGE scores as in (Nallapati et al., 2017). Both in case of BertSum and SummaRuNNer extractive summarisation is framed as a sequence classification task where text is first split into segments (sentences, in this case) and then each sentence is sequentially classified as either belonging to the summary or not. SummaRuNNer uses a GRU-RNN based architecture for this. We report results on two variations of Summarunner - one with default document truncation at default 100 sentences (SR) and one with document truncation changed to 3000 sentences (SRL) for narrative datasets that are long. BertSum takes a transformer based pretrained Bert model and fine-tunes it for summarisation tasks. However, it is only able to handle 512 tokens as input. Since, all of the narrative datasets are much bigger than this, we report results on LongFormer for these as well. Longformer modifies this approach for longer documents using windowed attention. While there is still a limitation on the number of tokens it can take as input, it improves on BertSum by allowing longer input sequences. Since more recent models like MatchSum and DiscoBert uses an underlying Bert model, they suffer from this limitation as well and hence, were not included as baselines.

4.2 Experiment Setup

We use gensim[5] library's implementation of the TextRank algorithm. For BertSum, we use TransformerSum library's[6] implementation of BertSum and LongFormer. At the time of running experiments, this implementation of LongFormer supported upto 4096 tokens as input. SummaRuNNer uses implementation from hpzao [7]. First 3 episodes of Wolf among us was used as training set , last 2 episodes of Wolf Among Us was used as validation and Before the storm was used as test set. Using a different game for the test ensures that there is no data leakage into the test set. Both models were trained with default parameters (except for max_epochs in TransformerSum's BertSum implementation which was set to 10 epochs rather the the default 100). Summarunner was originally

[5]https://pypi.org/project/gensim/
[6]https://github.com/HHousen/TransformerSum
[7]https://github.com/hpzhao/SummaRuNNer

Dataset_ Length	RAND-N	LEAD-N	TextRank (TR)	BertSum (BS)	SummaRuNNer (SR)	LongFormer (LF)	SummaRuNNer Long(SRL)
CnnDm_3	0.29	0.4	0.35	0.4	0.35	N/A	N/A
Novel_3	0.15	0.18	**0.26**	0.17	**0.26**	0.16	**0.26**
Novel_9	0.28	0.29	0.34	0.28	0.33	0.3	**0.35**
Novel_27	0.33	0.31	0.31	0.33	0.35	0.32	**0.36**
CRD3_3	0.02	0.03	0.08	0.03	0.03	0.02	**0.17**
CRD3_9	0.07	0.09	0.16	0.06	0.07	0.06	**0.31**
CRD3_27	0.17	0.18	0.27	0.14	0.18	0.27	**0.4**
CRD3_81	0.3	0.31	0.35	0.15	0.27	0.36	**0.47**
SB_3	0.05	0.05	0.12	0.05	0.1	0.07	**0.14**
SB_9	0.12	0.13	0.22	0.1	0.18	0.16	**0.27**
SB_27	0.24	0.23	0.32	0.21	0.27	0.27	**0.36**
IDN_3	0.02	0.04	0.04	0.008	0.04	0.04	**0.06**
IDN_9	0.06	0.09	0.1	0.05	0.11	0.08	**0.13**
IDN_27	0.17	0.17	0.24	0.12	0.2	0.2	**0.29**
IDN_81	0.35	0.32	0.4	0.16	0.27	0.31	**0.42**

Table 3: ROUGE1 F1 scores against human authored abstractive summary. SummaRuNNer (long) performs best overall. Note that Longformer (LF) and Summarunner (long) were not run for CNN/DM since these are meant for long documents and CNN/DM documents are short.

truncates documents at 100 sentences. We report performance of this model for this default case (SR) and a variation where it accepts longer documents with truncation at 3000 sentences(SRL) for narrative datasets since they are longer. In the long version, batch size had to be reduced to 1 to fit GPU memory. Each summarisation method was run with target length 3 for each dataset. Narrative datasets were also run with target lengths 9 and 27 since they have longer source documents and reference summaries. IDN and CRD3 were also run with target length 81 since reference summaries are much larger than 27 for these datasets.

4.3 Evaluation

The trained models were used to make predictions on the test set and ROUGE scores for all models were evaluated using the evaluation script from SummaRuNNer for consistency. The option setting the limit to the first x bytes was removed. This script uses the pyROUGE library[8]. ROUGE1 F1 score is calculated against the human authored abstractive summary with porter stemming (as commonly done in papers such as (Agarwal et al., 2018)) for all models and datasets and is compared in Table 3.ROUGE2 F1 scores are shown in the Appendix C. Scores against aligned extractive reference summaries can be found in Appendix B. The best and worst summaries (according to ROUGE) from the best model were also analysed qualitatively. The qualitative investigations help assess aspects of quality that are not captured by the ROUGE scores.

5 Results

Table 3 shows the performance of the baseline models. SummaRunner scales for longer documents and the long version (SRL) outperforms the other models in all cases. Another observation is that even though the narrative datasets are considerably smaller than CNN/DM, the use of pretrained language models does not seem to be helping. While Longformer improves on performance of BertSum in many cases, it does not significantly outperform the truncated version of SummaRunner. In many cases, truncated version of SummaRunner even performs better in terms of ROUGE scores in spite of only having access to the first 100 sentences of the text, whereas Longformer has access to significantly more (4096 tokens is between 200 and 400 sentences). Average sentence lengths for each of the datasets can be seen in Table 1. A manual inspection of sample summaries was performed and the results of this analysis are discussed below.

[8]https://pypi.org/project/pyROUGE/

5.1 Quality of aligned extractive summaries

The ROUGE1 F1 scores of the automatically aligned extractive summary overlap to human authored summary is shown in table 2. The ROUGE1 F1 for the narrative datasets at higher target lengths (27, 81) are comparable to that of CNN/DM at target length 3, which reflects the need for longer summaries to capture important information for longer narratives. Manual inspection of the original text and reference summaries also suggest that if all information in the human authored abstractive summary is considered equally important, it is hard to find sentence level extracts from the original text that cover all the information in case of smaller target lengths, especially for SB, CRD3 and IDN.

ROUGE F1 degrades from Novel to CRD3 to SB to IDN, especially for lower target lengths. To understand this further, the best and the worst summaries for each of the datasets were examined manually. This revealed that since words aren't weighted, many irrelevant sentences are picked up due to matching on common words (like character names) and stop words. ROUGE1 F1 scores for each of these datasets computed with the remove stopwords argument is also shown in Table 2 under 'stop filter'. The ROUGE scores of the narrative datasets degrade significantly compared to CNN/DM which stays approximately the same. This indicates the necessity of using weighted versions of ROUGE for alignment of narrative datasets, supporting findings from (Ladhak et al., 2020). It also shows CRD3 and Novel having higher scores when compared to SB and IDN. This can be traced to the presence of a few quotes from the original text in the human authored abstractive summaries for some instances in the Novel and CRD3 datasets. Since there is limited paraphrasing in these sentences, they get picked up and get higher ROUGE scores, but since there are only a few of these kinds of sentences, these datasets only have this advantage at lower target lengths.

It was also observed that summaries for SB had many sentences that are too short or are not coherent without context. Due to the presence of narration-like sentences in the Novel and IDN datasets, the overall readability of the summary was better at lower target lengths. However, in the case of IDN, much of the important information was also embedded in dialogue and was missed in the same way at higher target lengths.

Sample	%relevant (manual)	%coverage (manual)	ROUGE1 F1
IDN(b)	0.67	0.45	0.48
IDN(w)	0.40	0.30	0.36
Novel(b)	0.77	0.76	0.67
Novel(w)	0.07	0.01	0.05
Cnn (b)	1.0	1.0	1.0
Cnn (w)	0.0	0.0	0.02

Table 4: Analysis of best and worst ROUGE1 scoring generated summaries by SRL model. '% relevant' shows percentage of sentences in generated summary that match the ground truth abstractive summary (manual judgement used if there is a good sentence match or not). '% coverage' shows percentage of sentences in ground truth abstractive summary that match sentences in the generated summary.

5.2 Quality of Summaries from Best Model

Automatic metrics to evaluate summarisation is known to have many limitations (Fabbri et al., 2021). To get a better understanding of the quality of the summaries a manual inspection of the best and worst summaries from the best performing model for a non narrative (CNN/DM), narrative non interactive (Novel), and interactive narrative (IDN) was performed. The best performing models used were BS at length 3 for CNN/DM, SRL at length 27 for Novel, and SRL at length 81 for IDN. For each of the sentences in the model generated extractive summary, if it could be matched to any part of the abstractive summary it was marked as relevant. The number of relevant extracts divided by the total number of extracts is denoted as %relevant in table 4. For each sentence in the abstractive reference summary, if any part of the sentence could be matched to any of the extracted sentences it was marked as covered. The number of covered sentences divided by total number of sentences in the reference summary is denoted as %coverage in table 4. The corresponding ROUGE1 F1 score is also shown in the table for comparison.

The ROUGE metrics seems to capture relevance and coverage of sentences to some extent. The difference between best and worst summaries is less pronounced in case of IDN. This is because of shared text between datapoints and smaller differences between datapoints as discussed in section 3.2. However, the manual inspection of summaries revealed issues that were not reflected in the ROUGE scores. A sentence in the reference summary was marked covered if any of the sentences

Reference sentence from human written summary:
the dream abruptly ends with a truck crashing through william 's car .

Extracts:

[ex] s0 : chloe hears a horn three times and approaches william in panic .
a truck crashes into the left side of the car , hitting william , and then everything goes black .

Figure 3: Example of good quality extract

Reference sentence from human written summary:
upon a brief dialogue , in which rachel reveals the man they had seen at the park was her dad , and that he was cheating on her mother with that woman .

Extracts:

[ex] chloe : the ones who were making out ? [ex]
so when i saw he got a text from an unknown number ... asking him to meet ...

Figure 4: Example of low quality extract

in the model summary could be seen to be related to it. However, in most cases these sentences in the extractive summary do not convey all of the information that the corresponding parts of the abstractive reference summary do, even though both sets of sentences can be seen to be related. Additionally, the inspection suggests that even though many relevant extracts get picked up, the quality of selected extracts varies in terms of readability. To demonstrate the range of the quality of the selected extracts, Fig 3 shows an example of a high quality snippet of model summary and fig 4 shows and example of a low quality one. In the first example the information contained in the human written sentence is captured by the retrieved extracts. In case of the second example however, while it can be inferred that they are related, the information contained in the abstractive summary is not fully conveyed by the extracts and has poor readability. This issue is especially obvious in IDN where, due to its screenplay like structure, information captured by a single sentence in the abstractive summary is spread across several extracts. In CNN/DM on the other hand, information is presented in a concise way and sentences are dense with information.

6 Discussion

The main contribution of this piece of work is the generated IDN-Sum dataset. This is the first dataset for IDN that shows different branches that are possible through an interactive story. IDN is different from other forms of narrative text due to the presence of choice points that affect how the story unfolds. This dataset captures many different paths through such narratives. It is hence unique compared to other summarisation datasets because the high amount of overlapping text between data points. The dataset was created as a resource that enables us to investigate summarisation approaches for interactive and game narratives. It may also be used to study how summarisation models respond to small changes in text and target summary.

Capturing important differences between different playthroughs is a significant aspect of IDN summarisation. IDN is essentially a collection of linked literary documents. Summarization of multiple linked literary documents has not been studied previously, although multi-document summarization and plot (literary) summarization have been addressed separately. Unlike domains like news where multi document summarization(Antognini and Faltings, 2019) has been studied, IDN documents have a narrative structure and elements (plot, protagonist, emotions, etc) which influence the relative importance of sentences. The nature of differences between documents is different from domains like academic papers where comparative summarization has been studied(He et al., 2016). The differences are not solely topical and the links and link texts influences what is different between groups of documents. Therefore, this would also be a useful resource to study new NLP problems like comparative plot summarisation.

The dataset has 1250 playthroughs per episode and 8 episodes overall, but the code and JSONs for the ReaderBot will also be made available on GitHub[9]. This can be used to generate more playthroughs of the game, although they will need to be modified to adapt to different games. There are many types of IDN, both in terms of types of text and narrative design. While it is a limitation of this dataset that only one type of IDN is included, it takes a step towards making resources available for exploration of some aspects of IDN summarisation.

We also report and analyse performance of some standard baseline approaches quantitatively and qualitatively. In spite of a smaller number of data points, much longer input documents and difference in domain from CNN/DM, SummaRunner seems to scale for these longer documents and work well across domains, when considering ROUGE scores. However, manual inspection reveals several drawbacks of the ROUGE metric in terms of accurately reflecting summary quality. This is in line with findings from similar experiments performed on SummScreen in (Chen et al., 2021) where new entity centric evaluation metrics are pro-

[9]https://github.com/AshwathyTR/IDN-Sum

posed. Finding a good evaluation metric to assess summary quality is a known challenge, even in case of the CNN/DM dataset(Fabbri et al., 2021). For this reason, evaluation strategies usually include a human evaluation step in addition to automated metrics like ROUGE. However, in the case of narrative datasets, due to the large source length and relatively large reference summaries, human evaluation is resource intensive when compared to datasets like CNN/DM and more subjective since it needs to account for subjective aspects like coverage of plot points. Attempts to decrease subjectivity include strategies like judging the ability of the evaluator to answer questions about major plot points from the summary (Lapata, 2021). However, interactive narrative summarisation needs to account for interactive elements in addition to plot elements and important differences between playthroughs. Future work will augment this dataset with a similar list of plot points and interactive elements like decision points that can be used for evaluation.

The human written summaries against which scores are calculated summarise the entire IDN and represent variations between playthroughs through sentences like : *"If Chloe goes along with Rachel, she will be suspended. If Chloe takes the blame for Rachel, she will be expelled."* This means that in a playthrough where Chloe chose to take blame, there will be keywords relating explusion and in other branches, those relating suspension, but neither branch will have both. Hence, even if the model works perfectly, it cannot get a perfect ROUGE score since some of the keywords in the abstractive summary will not be present in that playthrough. Paraphrasing also causes some keywords to not be present in the original text. While these are drawbacks of the automatic evaluation, these scores give insight into relative performance of models and can be put into context by considering the score of the oracle as the upper bound and Random-N as the lower bound. These issues are mitigated by also providing ROUGE F1 scores against the oracle extractive reference summaries in Appendix B.

The qualitative analysis of the Oracle summaries also reveals some characteristics of narrative datasets that makes it worse if only keyword overlap is considered. News articles are structured differently to narrative text and are more likely to have summary sentences in the original text that capture the important information. Important information in narrative datasets are spread across several sentences. Presence of short sentences and sentences in utterances being broken up to include narration-like sentences in between screenplay-like text produces extracts that have high keyword overlap but are not useful or coherent. While scene-level summaries might be too large, selecting multi-sentence extracts instead of single sentence extracts might alleviate this issue to some extent. Additionally, sentences with many character names or short sentences with character names get high ROUGE scores even if they do not contain any relevant information because the reference summary contains them. A version of ROUGE that gives lower weights to words that are common in the document like the weighted ROUGE from(Ladhak et al., 2020) might do better in this regard. This study indicates that several aspects of the summarisation approaches that are commonly used for CNN/DM need to be re-examined and potentially redesigned for narrative and interactive narrative datasets, including: 1) The size and nature of extracts 2) automatic methods for conversion of abstractive summary to extractive summary 3) evaluation metrics and methodology. Hopefully, this dataset can help aid future research in these directions.

7 Conclusion

In this paper, we present the first summarisation dataset for interactive narratives. This was done by collecting fan made transcripts and abstractive summaries from Fandom and generating simulated playthroughs by assuming different combinations of choices. Annotation for extractive summarisation were created automatically from the abstractive summaries through greedy selection of extracts that maximised the ROUGE score with the abstractive summary. Even though narrative datasets have less data and longer text, SummaRunner with document truncation set to 3000 appears to scale when considering ROUGE scores. However, a qualitative analysis of generated summaries revealed several short comings in the ROUGE metric and oracle summaries suggesting that even though ROUGE scores for narrative datasets are comparable to CNN/DM, the summaries are not on the same level qualitatively. We hope that this dataset can be used for future research into better annotation methods, evaluation strategies. and summarisation approaches for interactive digital narratives.

References

Sanchit Agarwal, Nikhil Kumar Singh, and Priyanka Meel. 2018. Single-document summarization using sentence embeddings and k-means clustering. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 162–165.

Diego Antognini and Boi Faltings. 2019. Learning to create sentence semantic relation graphs for multi-document summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 32–41, Hong Kong, China. Association for Computational Linguistics.

Marta Aparício, Paulo Figueiredo, Francisco Raposo, David Martins de Matos, Ricardo Ribeiro, and Luís Marujo. 2016. Summarization of films and documentaries based on subtitles and scripts. *Pattern Recognition Letters*, 73:7–12.

Camille Barot, Michael Branon, Rogelio Cardona-Rivera, Markus Eger, Michelle Glatz, Nancy Green, James Mattice, Colin Potts, Justus Robertson, Makiko Shukonobe, Laura Tateosian, Brandon Thorne, and R. Young. 2021. Bardic: Generating multimedia narratives for game logs. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 13(2):154–161.

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

Atef Chaudhury, Makarand Tapaswi, Seung Wook Kim, and Sanja Fidler. 2019. The shmoop corpus: A dataset of stories with loosely aligned summaries. *arXiv preprint arXiv:1912.13082*.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2021. Summscreen: A dataset for abstractive screenplay summarization. *arXiv preprint arXiv:2104.07091*.

Yun-Gyung Cheong, Arnav Jhala, Byung-Chull Bae, and Robert Michael Young. 2008. Automatically generating summary visualizations from game logs. In *AIIDE*, pages 167–172.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748, Brussels, Belgium. Association for Computational Linguistics.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076, Denver, Colorado. Association for Computational Linguistics.

Lei He, Wei Li, and Hai Zhuge. 2016. Exploring differential topic models for comparative summarization of scientific papers. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1028–1038, Osaka, Japan. The COLING 2016 Organizing Committee.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown. 2020. Exploring content selection in summarization of novel chapters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5043–5054, Online. Association for Computational Linguistics.

Pinelopi Papalampidi Frank Keller Mirella Lapata. 2021. Movie summarization via sparse graph construction.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Chanakya Malireddy, Srivenkata NM Somisetty, and Manish Shrivastava. 2018. Gold corpus for telegraphic summarization. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 71–77.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

David Millard, Charlie West-Taylor, Yvonne Howard, and Heather Packer. 2018. The ideal readerbot:: Machine readers and narrative analytics. In *NHT?18, July 2018, Baltimore, USA*. ACM.

MF Mridha, Aklima Akter Lima, Kamruddin Nur, Sujoy Chandra Das, Mahmud Hasan, and Muhammad Mohsin Kabir. 2021. A survey of automatic text summarization: Progress, process and challenges. *IEEE Access*, 9:156043–156070.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-first AAAI conference on artificial intelligence*.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. 2020. Screenplay summarization using latent narrative structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1920–1933, Online. Association for Computational Linguistics.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134, Online. Association for Computational Linguistics.

Anna Marie Rezk and Mads Haahr. 2020. The case for invisibility: understanding and improving agency in black mirror's bandersnatch and other interactive digital narrative works. In *International Conference on Interactive Digital Storytelling*, pages 178–189. Springer.

BJ Sandesh and Gowri Srinivasa. 2017. A framework for the automated generation of paradigm-adaptive summaries of games. *International Journal of Computer Applications in Technology*, 55(4):276–288.

Quang Dieu Tran, Dosam Hwang, O Lee, Jai E Jung, et al. 2017. Exploiting character networks for movie summarization. *Multimedia Tools and Applications*, 76(8):10357–10369.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Weiwei Zhang, Jackie Chi Kit Cheung, and Joel Oren. 2019a. Generating character descriptions for automatic summarization of fiction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7476–7483.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019b. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.

A Appendix A

Examples of the data are shown in this appendix. Appendix A.1 shows some lines from the beginning of a sample source document to be summarised. The complete document is not shown here due to its large size, but can be downloaded from the github repository. The corresponding lines from the human authored abstractive summary and aligned extractive summary is shown in appendix A.2 and appendix A.3 respectively. The complete summaries can be seen in the github page.

A.1 Example lines from preprocessed source text

S0 : ' [EX] :SC: S0 : Principal Wells, Rachel Amber, Joyce Price enter the office. [EX] PRINCIPAL WELLS : Ms. Price. How good of you to join us. [EX] JOYCE : I'm so sorry we're late. My—my shift ran late at the diner and then...just, sorry. [EX] PRINCIPAL WELLS : Let us proceed. One of you here is new to the Blackwell disciplinary process... And the other is all too familiar with it. Blackwell's code of conduct is built upon a foundation of mutual respect meant to foster an environment conducive to education and enrichment. When that respect is violated, actions are taken. When that respect is repeatedly disregarded, a more consequential response is required. [EX] CHLOE : (thinking) Okay, reality check time. Yesterday did actually happen. I ditched school with Rachel Amber. And then Rachel really did start that fire. And that was after we actually agreed to run away from here...right? [EX] PRINCIPAL WELLS : Are you paying attention to me, Chloe? [EX] CHLOE : Um...what? [EX] PRINCIPAL WELLS : Ms. Price, the last time we met, an agreement was brokered. Do you recall what that was? [EX] S0 : CHOICE: Don't screw up? [EX] CHLOE : Uh, don't get in trouble again? [EX] PRINCIPAL WELLS : Trouble is merely the byproduct, Ms. Price. What's at issue is your attitude. [EX] PRINCIPAL WELLS : We agreed that you would rededicate yourself to

becoming an exemplary Blackwell citizen. [EX] CHLOE : We did? [EX] PRINCIPAL WELLS : In the event that you were unable or unwilling to do so, we also agreed that it would become pertinent to reassess your future status at the academy. Despite all this, you engaged in the following actions yesterday: Insubordinate language... [EX] S0 : CHOICE: (Trespassed on stage) [EX] PRINCIPAL WELLS : Disregarding posted signs about trespassing on the stage. [EX] PRINCIPAL WELLS : Shall I continue? [EX] S0 : CHOICE: (Didn't sabotage Victoria's homework) [EX] PRINCIPAL WELLS : Witnesses saying you were involved in bullying Nathan Prescott. [EX] S0 : CHOICE: (Didn't help Nathan) [EX] CHLOE : If "involved" means not sticking out my neck for Blackwell's richest asschild. I didn't realize that was a crime. [EX] PRINCIPAL WELLS : Your lack of awareness does not absolve you of anything, Ms. Price. [EX] S0 : CHOICE: (Was nice to Joyce) [EX] JOYCE : Say what you will about my daughter, but she is not a bully. [EX]

A.2 Example of human authored abstractive summary

Episode 2: Brave New World begins with Rachel Amber and Chloe Price in Principal Wells' office. Both Rachel and Chloe are questioned about their absence the day before. The conversation varies depending on how Chloe treated Joyce, if she sabotaged Victoria's homework, if she went onstage and smoked weed, whether she helped Nathan or not, and if she won or lost the backtalk against Drew (if she helped Nathan).

A.3 Example lines from automatically aligned extractive summary

I ditched school with Rachel Amber . [EX] S0 : CHOICE : (Did n't sabotage Victoria 's homework) [EX] PRINCIPAL WELLS : [EX] S0 : CHOICE : (Was nice to Joyce) [EX] PRINCIPAL WELLS : Mr. North 's situation requires ... sensitivity .

B ROUGE1 Scores against automatically aligned extractive summaries

Table 5 shows ROUGE1 scores computed against automatically aligned extractive summaries.

C ROUGE2 F1 Scores against human authored abstractive summaries

Table 6 shows ROUGE2 scores computed against human authored abstarctive summaries.

Dataset+Target Length	RN	LN	TR	BS	SR	LF	SRL
CnnDm3	0.34	0.5	0.45	0.51	**0.59**	N/A	N/A
Novel3	0.24	0.28	0.36	0.27	**0.38**	0.26	**0.38**
Novel9	0.38	0.38	0.42	0.38	0.42	0.41	**0.43**
Novel27	0.42	0.4	0.38	0.42	0.44	0.42	**0.47**
CRD3_3	0.11	0.14	0.23	0.1	0.11	0.19	**0.68**
CRD3_9	0.17	0.21	0.33	0.16	0.18	0.16	**0.74**
CRD3_27	0.31	0.31	0.42	0.26	0.32	0.45	**0.65**
CRD3_81	0.45	0.43	0.47	0.24	0.4	0.49	**0.61**
SB3	0.17	0.14	0.27	0.15	0.23	0.19	**0.36**
SB9	0.26	0.25	0.35	0.22	0.31	0.31	**0.44**
SB27	0.39	0.35	0.4	0.33	0.39	0.4	**0.49**
IDN3	0.15	0.23	0.21	0.07	0.34	0.22	**0.37**
IDN9	0.26	0.34	0.3	0.25	0.36	0.29	**0.45**
IDN27	0.39	0.41	0.4	0.34	0.44	0.4	**0.50**
IDN81	0.54	0.49	0.55	0.3	0.45	0.48	**0.62**

Table 5: ROUGE1 F1 scores against automatically aligned extractive summary

Dataset+Target Length	RN	LN	TR	BS	SR	LF	SRL
CnnDm3	0.084	0.174	0.143	0.177	**0.154**	N/A	N/A
Novel3	0.018	0.032	0.039	0.025	0.041	0.025	**0.042**
Novel9	0.039	0.05	**0.059**	0.046	0.056	0.053	**0.059**
Novel27	0.06	0.062	0.067	0.06	0.067	0.058	**0.074**
CRD3_3	0.005	0.005	0.018	0.004	0.004	0.007	**0.142**
CRD3_9	0.012	0.016	0.037	0.01	0.013	0.012	**0.244**
CRD3_27	0.031	0.03	0.067	0.024	0.038	0.119	**0.265**
CRD3_81	0.065	0.074	0.087	0.026	0.055	0.135	**0.255**
SB3	0.005	0.006	0.017	0.005	0.012	0.008	**0.021**
SB9	0.013	0.016	0.034	0.013	0.024	0.021	**0.041**
SB27	0.028	0.03	0.051	0.027	0.038	0.034	**0.061**
IDN3	0.004	0.011	0.009	0.002	0.011	0.009	**0.016**
IDN9	0.11	0.025	0.023	0.011	0.026	0.019	**0.03**
IDN27	0.03	0.038	0.05	0.03	0.047	0.04	**0.059**
IDN81	0.06	0.06	0.087	0.036	0.052	0.067	**0.096**

Table 6: ROUGE2 F1 scores against human authored abstractive summary