Rationale-based Learning using Self-Supervised Narrative Events for Text Summarisation of Interactive Digital Narratives

Anonymous submission

Abstract

This paper explores using rationale-based learning with supervised attention to focus the training of text summarisation models on words and sentences surrounding choice points for Interactive Digital Narratives (IDNs). IDNs allow players to interact with the story via choice points, making choices central to these narratives. Exploiting such knowledge about narrative structure during model training can help ensure key narrative information appears in generated summaries of narrative-based text and thus improve the quality of these summaries. We experiment with using word-level and sentence-level rationales indicating the proximity of words and sentences to self-supervised choice points. Our results indicate that rationale-based learning can improve the ability of attention-based text summarisation models to create higher quality summaries that encode key narrative information better for different playthroughs of the same interactive narrative. These results suggest a promising new direction for narrative-based text summarisation models.

Keywords: Interactive Digital Narratives, Summarization, Rationale based learning

1. Introduction

Interactive Digital Narratives (IDNs), such as choose-your-own-adventure games and story-rich video games, are narratives that support player interaction. IDNs are becoming increasingly more prevalent with the growing popularity of narratives in mediums such as video games and interactive mixed-reality experiences. However, while there are some studies on how external information about narrative structures can be introduced into narrative summarisation(Papalampidi et al., 2020), there is not much research investigating what prior information about *interactive* narrative structure can be introduced for interactive narrative summarisation and how this can be done. This is what we address in this paper.

In IDNs, while interaction can occur in many ways, making choices that affect the course of the story is a popular interaction pattern, with the plot and gameplay being closely entwined with the choices made by the player. In such IDNs, the context in which choices are presented, the player choices and their consequences heavily influence which parts of the narrative are salient enough to be included in the summary. Therefore, understanding the significance of narrative events is often enhanced by considering them in the context of player choices. For example, the player may have chosen to kill a Non-Player Character (NPC) who appeared to be the evil, but later in the story, they may find out that they were innocent. Finding out about the NPC's innocence becomes more significant in the context of the choice the player had to make earlier in the game. In this paper, we investigate leveraging this knowledge regarding the importance of choices to enhance IDN summarisation.

To incorporate this knowledge into the training process, we explore for the first time, choice-focussed rationale-based learning for extractive summarisation of IDN. Our approach is motivated by the text classification model of (Kanchinadam et al., 2020), which used word-level rationale-based learning with supervised attention to help focus model training on areas of the text that human annotators considered important. Inspired by this approach, we explore sentence-level and word-level rationale-based learning for extractive summarization of IDN narratives, using proximity to choice points as a self-supervised proxy for human rationales. This paper is focussed on IDNs and choice points but the proposed approach can also be extended to traditional narrative-based text to incorporate knowledge about narrative structure like the importance of emotion using emotion detection techniques to automatically generate rationales.

The novelty of our approach is in the formulation of the data and training objectives for this unique domain (IDN). While the outlined approach can be extended to various types of attention-based architectures, applying supervised attention to model architectures with multi-head attention can involve additional layers of complexity. Therefore, in this paper, we first investigate the efficacy of this approach on variants of the classic SummaRuNNer model equipped with simple attention layers. Our results show that choice-focussed rationale-based learning delivers a significant improvement in ROUGE scores when compared against gold-standard human-authored abstractive reference summaries, encouraging further research in this direction. To summarise, the contributions of this paper are as follows:

1. A novel method using word and sentence level rationales applied to an existing RNN-based model (SummaRunner) for Interactive Digital Narratives (IDN) summarisation, addressing a

domain that remains relatively underexplored.

2. Empirical results showing that using choice points for self-training rationales outperforms similar models trained traditionally.

3. Manual Qualitative and Fault analyses providing deeper insights into model limitations to guide future researchers in this area.

4. To the best of our knowledge, this is the first self-trained rationale-based method for narrative summarization.

We review related work in section 2 before outlining, in detail, our rationale-based training approach and the models we train in section 3. Section 4 reports results from our automatic and manual evaluation and analysis of variability of generated summaries across different playthroughs of the same interactive narrative, which we discuss and conclude in section 5 and section 6.

2. Related Work

Previous studies on extractive summarisation have focussed on various techniques including RNN-based models (Nallapati et al., 2017), language model-based methods (Liu, 2019) and graph-based methods (Antognini and Faltings, 2019). However, these methods are most commonly trained and tested on datasets like news (Hermann et al., 2015) and academic articles (Gupta et al., 2021). While some approaches for summarisation of traditional narratives have been explored, like using GCNs for screenplay summarisation (Lapata, 2021) and taking turning point information into account (Papalampidi et al., 2020), summarisation of interactive narratives has not been explored in much depth. IDN-Sum (Revi et al., 2022) is a dataset introduced for studying interactive narrative extractive summarisation and is used for the experiments in this paper. Interactive narratives are unique from other domains where summarisation has been explored in that they often have complex structures arising from the ability of players to interact with the story.

Rationale-based learning, or explanation-based learning, is an approach that uses rationales to guide the training of machine learning models (Gao et al., 2022a). This has been applied in a variety of NLP tasks including Text Classification (Arous et al., 2021; Choi et al., 2020), Natural Language Inference (Camburu et al., 2018; Stacey et al., 2022) and Sentiment Analysis (Zhong et al., 2019). Both local explanations (Gao et al., 2022b) and global explanations have been applied to guide training (Liu and Avci, 2019) in this way. Rationales are incorporated into training through various means including supervised attention (Kanchinadam et al., 2020), which is the approach we have used in this paper. However, in this paper, we investigate the effectiveness of choices as rationales in the novel context of summarising IDNs. We also experiment with different kinds of explanations applied at both word and sentence levels.

3. Method

3.1. Choice Focussed Rationales

We will introduce information regarding the importance of choices in IDN summarisation into the training process through rationales that indicate the proximity of words and sentences to choice points. In IDN-Sum dataset(Revi et al., 2022) used in this paper, choice points are marked using a choice tag, "CHOICE :". Using this tag, sentence and word rationales were embedded as tensors in the following way:

$$rs_i = \begin{cases} 1 & \text{if } CT \in [s_{i-ws}, s_{i+ws}] \\ 0 & \text{otherwise} \end{cases}$$

$$rw_i = \begin{cases} \text{tfidf}(w_i) & \text{if } w_i \in CW \\ 0 & \text{otherwise} \end{cases}$$

where CW is the set of all words that fall inside a window of size ws around the choice tag given by,

$$CW = \{w_i \, \epsilon \, W \mid CT \; in \; (w_{i-ws} : w_{i+ws})\}$$

CT stands for the choice tag, rs_i and rw_i stand for the rationale for sentence/ word at index i, ws stands for window size, s_i and w_i stands for sentence/ word at index i and notations $s_i : s_j$ and $w_i : w_j$ represents concatenation of sentences/ words at indexes from i to j.

Then, following the method used in previous work in supervised attention (Kanchinadam et al., 2020), to use rationales in training, training loss was calculated in the following way:
For sentence attention model:

$$L = \alpha * L_l + (1 - \alpha) * L_s$$

For word attention model :

$$L = \alpha * L_l + (1 - \alpha) * L_w$$

For attention model with sentence and word level attention :

$$L = \alpha * L_l + \alpha_1 * L_s + \alpha_2 * L_w$$

where: α + α_1 + α_2 = 1,
L = Total Loss,
L_l = Cross-entropy loss calculated for the output of the model against the target labels,

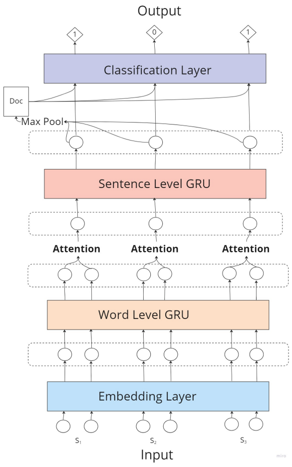

Figure 1: Summarunner modified to use attention instead of max pooling at word level (wordonlyAttnRNN).

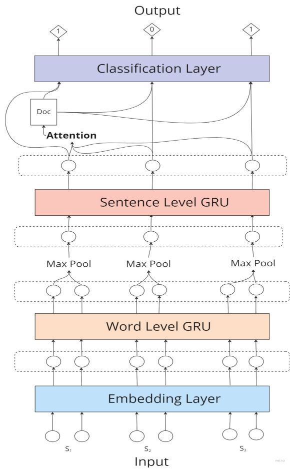

Figure 2: Summarunner modified to use attention instead of max pooling at sentence level (sentonlyAttnRNN).

L_s = Cross-entropy loss calculated for sentence attention scores against sentence rationales and L_w = Cross-entropy loss calculated for word attention scores against word rationales.

This essentially tells the model to pay more attention to sentences and words surrounding the choice points when generating internal representations and deciding whether to include the given sentence in the extractive summary or not.

3.2. Base Models

While our training approach could theoretically be applied to any model with an attention layer, introducing supervised attention to recent Pretrained Language Models (PLMs) and other transformer based models with multi-head attention introduces additional layers of complexity when applying supervised attention (eg. how many and which attention heads do we align with the rationales). Another significant limitation of many PLMs is their fixed context length, making them unsuitable for direct application to datasets like IDNSum with an average document length of 22,900 tokens. Therefore, in this paper, we first test our approach on a simple attention layer, saving other attention types for future research.

In our experiments, we utilize models based on SummaRunner, an RNN-based model for extractive summarisation with simple attention layers added to it. We chose SummaRunner as the base model because of its superior performance on the IDNSum dataset, outperforming even PLM based models like Longformer(Beltagy et al., 2020) on this dataset(Revi et al., 2022) and its renowned and consistent performance as a standard for extractive summarisation, allowing us to contextualize the efficacy of our proposed approach within a widely recognized model. The model referred to as RNN, in this paper, represents the original architecture used in Summarunner, modified to truncate documents at 3000 sentences instead of 100.

In Summarunner, word representations are combined into sentence representations and sentence representations are combined into document representations using max pool. Attention layers are added to this model so that rationales can be incorporated through supervised attention. In order to test the effectiveness of rationale-based learning at both the word and sentence level, max pool is replaced with attention layers at different levels in the following three ways, inspired by Hierarchical Attention Networks (HAN) (Yang et al., 2016) to produce three types of attention models: The first attention model is the Word level AttnRNN model(**wordonlyAttnRNN**), which only uses attention at the word level to combine the outputs of the word level GRU into sentence representations. This model architecture is illustrated in Figure 1. The second modified architecture is the Sentence level AttnRNN model (**sentonlyAttnRNN**), where attention is used only to pool the outputs of the sentence level GRU into document representations. This is illustrated in Figure 2. The third modified architecture is **AttnRNN**, modelled after Hierarchical Attention Networks (Yang et al., 2016), which uses attention at both the word and sentence level and is illustrated in Figure 3. In this paper, versions of these models trained with rationales is indicated by the suffix "+ rationale".**sentonlyAttnRNN + rationale** represents sentonlyAttnRNN trained with sentence rationale labels. **wordonlyAttnRNN + rationale** represents wordonlyAttnRNN trained with word rationale labels. **AttnRNN + rationale** rep-

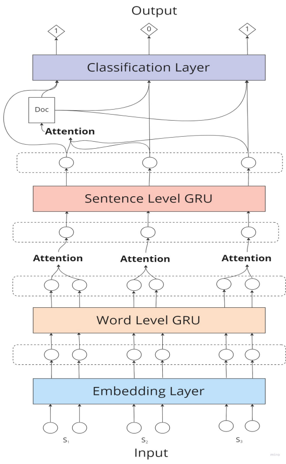

Figure 3: Summarunner modified to use attention instead of max pooling at both word and sentence level(AttnRNN).

resents AttnRNN trained with both rationales. All these models have approx 1.7M parameters.

3.3. Experiment Set Up

3.3.1. Dataset

The IDN-Sum dataset (Revi et al., 2022) was used to train each model. The dataset contains 10000 documents consisting of 1250 simulated playthroughs per episode of two interactive narrative games: *Before the Storm* developed by Deck Nine and released in 2017 and *Wolf Among Us* developed by TellTale Games and released in 2013. The dataset also contains the fan-written abstractive summaries for each episode and automatically generated extractive summaries for each playthrough. The extractive summary is represented through sentence-wise binary annotation indicating whether the sentence is included in the summary or not. The models were trained using the default split of this dataset (playthroughs of 3 episodes from Wolf Among Us in the training set, the remaining 2 episodes of Wolf Among Us in the validation set and the 3 episodes from Before the Storm in the test set.)

3.3.2. Models

We use the implementations provided on Github [1] as the starting point for the modifications described in section 3.2. These modified versions are made available on Gihutb[2]. Default settings were used except for the following parameters - since IDN documents are larger, the models were trained using batches of 1 document at a time to fit GPU memory. The parameter "report every" was reduced to 30 to monitor the training process more closely since IDN-Sum has many repeated sentences between data points making models more prone to overfitting when training on this dataset. The parameters window size (ws) and the coefficients (alpha) were tuned manually using the validation set within the bounds 0.99 - 0.25 for alpha and values [2,4,8,16] for was for sentence rationales and values [20,40,80,160] for word rationales. The best model, according to validation f1 scores, for which results are reported, was trained with parameters - ws = 2, alpha =0.95 for sentonly AttnRNN + rationale, ws=20, alpha = 0.5 for wordonly AttnRNN + rationale and ws=8,80 and alpha = 0.5, alpha1 = 0.25, alpha2 = 0.25 for AttnRNN + rationale.

In addition to the original SummaRuNNer model, we also show the performance using the more recent Longformer(Beltagy et al., 2020) (PLM for long documents with approx 149M parameters) and a zero shot LLM-based approach using Google's flan-t5-base model(Chung et al., 2022a) (instruction tuned LLM with 250m parameters) for comparison. For Longformer, we use the implementation from TransformerSum[3]. This implementation had a 4096 context window for pretrained extractive summarisation models at the time this experiment was run and documents were truncated at this length. We finetune the model using the same training, validation, test split and default parameters. For flan-t5, we get the pretrained model from Huggingface[4]. Summaries are generated in a zero-shot setting, 25 sentences at a time, to fit the context window and strung together at the end to get the final summary. The prompt and hyperparamters were manually tuned. The prompt used was: *"Create an extractive summary for the document. The summary should contain up to 3 sentences from the original text that best capture the essence of the document. \n Document: {25 sentence document} \n Extractive Summary:"* Refer Appendix for the full list of hyperparameters, hardware details and training times for all models.

3.3.3. Evaluation

We evaluated the performance of our models using ROUGE-1(R1), ROUGE-2(R2), and ROUGE-L(RL). The performance of these models with and without attention and trained with and without rationales for the attention models were also compared. ROUGE scores were calculated against the human-authored abstractive summaries. ROUGE scores

[1] The RNN model and Hierarchical Attention Network model from https://github.com/hpzhao/SummaRuNNer are used in this paper as RNN and AttnRNN, respectively.

[2] github link will be provided if paper accepted

[3] https://github.com/HHousen/TransformerSum

[4] https://huggingface.co/google/flan-t5-base

against the branch-wise extractive summaries and ROUGE scores calculated with and without the stop word filter will be shown in the Appendix .

Some studies rely solely on ROUGE for comparing summarisation approaches(Zhong et al., 2020; Yuan et al., 2023; Cui et al., 2020). The ROUGE metric and automatic evaluation for summarisation face many challenges and several studies supplement the ROUGE based evaluation with manual human evaluation. However, the novelty of the domain and length of source documents and summaries for the IDN-Sum dataset makes large-scale human evaluation challenging and resource intensive. Therefore, following the approach used in recent work(Tang et al., 2022), we provide examples of the model-generated summaries and reference summaries for human evaluation in the Appendix and perform a qualitative analysis to compare and illustrate intuitive aspects of quality that the ROUGE-based evaluation is unable to capture.

IDN-Sum dataset is characterised by a high overlap of text between data points caused as a result of generating different playthroughs through the same game. IDN summaries are hence most useful when these differences are captured. We analyse the variation between summaries generated by the model for different playthroughs through the same episode by calculating the average overlap of sentences between each pair of model summaries of the same episode in the test set to understand how varied the generated summaries are.

In addition to the comparison of approaches, we also perform a manual fault analysis to understand the limitations of our approach and encourage further research. The fault analysis was performed on 10 summaries generated by the best model (SentAttn + rationale) from each of the three episodes in the test set. These summaries were sampled randomly from the set of summaries that had a ROUGE score below the mean for that episode. This was done to get a deeper insight into the type of errors made by the model. In the first pass, the main error classes in the model-generated summaries were identified. Then, in the second pass, each sentence in model generated summary was coded against the error classes.

4. Results

4.1. Automatic Evaluation

Table 1 shows the ROUGE scores calculated against the human-authored abstractive summary. The corresponding validation scores will be shown in the Appendix. A breakdown of these scores by episode is also provided in Appendix.The evaluation script is available on GitHub [5].ROUGE score was calculated with Porter stemmer on and the stop filter turned off. Additional analysis showing ROUGE scores with the stop word filter turned on and ROUGE scores calculated against the automatically aligned extractive summary will also be provided in the Appendix. The versions of the attention models trained with different types of rationales are compared with those trained without rationales and the RNN model which does not incorporate attention or rationales.

The rationale-based models outperform the RNN model and the corresponding attention models trained without rationales. This indicates that choice focussed rationale-based learning can improve the performance of summarization models for IDN. The model that incorporated rationales at the sentence level (sentonly AttnRNN + rationale) shows the most improvement when measured against human-annotated abstractive summaries. R1 and R2 scores show an increase of approximately 14% and 12% respectively compared to the sentonly AttnRNN model and by 7% and 5% respectively compared to the RNN model.

We also show the performance of more recent approaches (using Longformer and flan-t5 models) for comparison. The relatively lower performance of Longformer is mainly because the documents had to be truncated to fit the context window. Despite the instruction to generate extractive summaries, the flan-t5 model tended to paraphrase the sentences from the original text and produced many hallucinations leading to lower scores. These results are reported to contextualise the performance of our method rather than claim state of the art. Fine tuning flan-t5 and optimising the prompt could result in better performance. Similarly, alternate strategies for handling long documents in case of both models could improve their performance. Additionally, by employing our rationale based learning on them, we could potentially get even better performance. This will be explored in future work.

4.2. Human Evaluation

The best and worst scoring summaries from Episode 1 of Before the Storm from the base RNN model (RNN) and Sentence Attention model trained with and without rationales (sentonlyAttnRNN and sentonlyAttnRNN+rationale) were reviewed manually to get an understanding of subjective aspects of quality that automatic metrics are unable to capture. All the output summaries will be made available on GitHub if the paper is accepted and some examples will be shown in Appendix for reference.

Summaries produced by the RNN model appear to contain more sentences from the beginning scenes of the games, with a lot of redundant information in the earlier scenes and missing information in the middle and later scenes. The attention

[5] github link will be provided if paper accepted

Model	R1(abs)	95% CI	R2(abs)	95% CI	RL(abs)	95% CI
SummaRuNNer (RNN)	0.47757	0.47689 - 0.47825	0.12379	0.12323 - 0.124358	0.46460	0.46403 - 0.4651
sentonly AttnRNN	0.44569	0.44464 - 0.44671	0.11624	0.11550 - 0.11697	0.43477	0.43382 - 0.43572
sentonly AttnRNN + rationale	**0.50852**	0.50767 - 0.50936	**0.13036**	0.12977 - 0.13095	**0.49223**	0.49150 - 0.49299
wordonly AttnRNN	0.46508	0.46446 - 0.46568	0.12082	0.12012 - 0.12155	0.45205	0.45152 - 0.45258
wordonly AttnRNN + rationale	0.48124	0.48032 - 0.48209	0.12386	0.12331 - 0.12439	0.46764	0.46681 - 0.46839
AttnRNN	0.44044	0.43983 - 0.44107	0.11081	0.11018 - 0.11142	0.42832	0.42782 - 0.42884
AttnRNN + rationale	0.48637	0.48542 - 0.48725	0.13337	0.13265 - 0.13407	0.47231	0.47147 - 0.47309
Longformer	0.30881	0.30754 - 0.31007	0.06692	0.06641 - 0.06748	0.30237	0.30117 - 0.30354
Google flan-t5-base	0.46577	0.46519 - 0.46637	0.11833	0.11800 - 0.11866	0.41051	0.40997 - 0.41112

Table 1: Mean ROUGE-1 (R1), ROUGE-2 (R2) and ROUGE-L (RL) scores and confidence interval (CI) of generated summaries of IDNSum playthroughs calculated against gold standard human written abstractive summaries(abs).

based models cover all the scenes in a more balanced way. When comparing the attention-based models trained with and without rationales, it is not immediately obvious if the improvement in the scores comes from including more information that is related to choices and their consequences. While summaries from the rationale-based model appear to be clearer and more relevant overall, both summaries contain sentences that are related to choice points. Further research is required to understand what aspects of summarisation improve when making use of rationales and why.

The summaries from the best model with the best ROUGE score from each episode were also analysed qualitatively. To someone reading the summary without any other context,it only provides a vague, fragmented view of the plot. However, to someone already familiar with the story, the summary can serve as a recap of the plot to some extent. This is because most of the plot elements are not directly conveyed but can be inferred. However, there is some variability in how easy it is to do so from the extracts. It was also noted that even with increased attention to choice points, many important choices and related events were missed.

4.3. Variability Analysis

Table 2 shows the average amount of overlap in the summaries produced by the SummaRuNNer variants with and without rationales. This is calculated by taking the average number of overlapping sentences between each pair of summaries produced by the model of playthroughs from the same episode. Models incorporating sentence-level rationales show lower overlap indicating that they are able to produce summaries that better capture the differences between playthroughs. For example, sentonly Attn + rationale model shows 6% less overlap compared to RNN and 16% less overlap compared to sentonlyAttn model.

Model	Avg overlap
RNN	47.85
sentonly Attn	53.48
sentonly Attn + rationale	44.76
wordonly Attn	50.84
wordonly Attn + rationale	49.66
AttnRNN	49.21
AttnRNN + rationale	45.88

Table 2: Average number of overlapping sentences for every pair of summaries from each episode for each model (out of a total of 81 sentences).

4.4. Fault Analysis

Through manual inspection of the model summaries from the best model, sentonly Attention model trained with rationales, four error classes were observed. The error classes are described below and the frequency of occurrence of the error classes is shown in table 3.

1. **Irrelevant Information** *(Common)*: Sentence cannot be matched to any part of the refer-

ence summary. This includes sentences like *"two firefighters show up as well, and one of them speaks to the officer"* which is from a section of the text not covered by the reference summary. The information contained in such extracts is not contained in the reference summmary and is hence considered irrelevant. This also includes sentences like *"frank and his friend are hanging out next to his rv at the old mill."* which is roughly from the portion of the script covered by a sentence in reference summary: *"the episode ends showing each character's reaction to the wildfire seen in the sky."*, but since the extract itself does not talk about their reaction to the fire, it is considered irrelevant.

2. **Incomplete Information** *(Common)*: Given the model summary, reference summary and the script, the sentence can be matched, but the model summary alone is insufficient to convey the relevant information. It needs additional extracts to be useful. This is different from the previous error case in that some relevant information is contained within this sentence, however, the summary lacks enough context for it to convey the necessary information. This mainly happens due to unclear references to pronouns, need for additional information or inference. An example is, *"chloe: (thinking) let 's get these to david so he can drive away."* which can be matched to a sentence in the reference summary : *"chloe'll have to pick the keys from her stepfather, david madsen, and take them to him since he'll be taking her to school today."*. However, the information is not clearly conveyed by that extract alone. This also includes cases where the reference summary contains a brief mention of a high level event and the model summary captures some detail of the event without conveying the big picture. For example, the reference summary contains the information, *"Cloe can talk to hayden jones , dana ward , and travis keaton"*, and the model summary contains the extract *"budding dramaturge , may your propitious appearance counteract the tragedy of stephanie gingrich 's sudden recusal ."* which is from the conversation between Chloe and Travis Keaton and can be matched as such, but, the fact that a conversation between Chloe and Mr Keaton is happening is not explicitly captured by the extract.

3. **Redundant Information** *(Common)*: Information covered by this sentence is better captured by other sentences already present in the summary. For example, the information conveyed by the extract, *"then she falls on her back and continues crying on the ground."* is better conveyed by *"chloe approaches the car and starts hitting its hood with her fists and crying ."* where the associated sentence in the reference summary is *"she then has a meltdown upon seeing her late father 's car."*.

4. **Unclear /Short Sentences** *(Rare)* : Sentence is too short and generic to be useful. This includes sentences like *"figures."* and *"yeah."* that appear in the summary without any surrounding context. Note that such sentences were coded as such only when the relevant context was not provided by in the surrounding sentences in the summary.

Error Type	Ep 1	Ep 2	Ep 3	Avg
Redundant	16.5	13.9	22.3	17.57
Incomplete	18.9	17.4	13.9	16.73
Irrelevant	15.2	17.4	21.5	18.03
Unclear	0.1	0.5	0.1	0.23

Table 3: Fault Analysis: Error types in model summaries and the average number of sentences exhibiting these errors out of a total 81 sentences per summary.

Analysis was done at the sentence level. Ten summaries from each episode were sampled randomly from the summaries that had a ROUGE score below the mean for that episode. Each of the sentences in the sampled summaries was coded against the above error classes. In cases where when there is more than one extract that indirectly or incompletely conveys the same information, the least indirect or incomplete sentence is coded as "Incomplete" and the others are coded as "Redundant". For example, the reference summary for episode 1 says that Chloe has the option of playing a role playing game. The introductory sentence of the game *"you are both famous heroes in the kingdom of avernon , a once peaceful land , now laid to waste by the bloodthirsty raiders of the black well"* conveys this better than an extract from the middle, *"to your left , the raiders ' training ground ."*. Therefore, even though both are indirect, the former is coded as "incomplete" and the latter is coded as "redundant" since it conveys no new information that was not better captured by other sentences in the summary. The results showing prevelance of these errors in the summaries generated by the best model (sentonly attention + rationale) in terms of average number of sentences coded with the error for each of the episodes is shown in table 3. Redundant sentences, sentences having incomplete information and irrelevant sentences are more prevalent than unclear sentences, but these three errors are similarly prevalent.

5. Discussion

The results of the experiments show that incorporating rationales in the form of annotations indicating proximity of sentences to choice points improves the performance of attention-based models for extractive summarization of IDN by up to 14% while producing more varied summaries across playthroughs. This suggests that automatically generated choice point annotations can act as effective rationales for IDN since choices are central to the narrative structure of IDN.

Rationale-based learning provides a way to incorporate knowledge and assumptions about narrative structure into training. The work presented in this paper has demonstrated this successfully in the case of choice-based rationales in interactive narratives. This encourages future work that experiments with using rationale-based learning for the summarisation of other types of narratives with rationales indicating aspects that are central to those types of narratives. For example, for traditional narratives including novels and movie scripts, elements like emotion and plot are considered to be central. Approaches used in previous work for tasks like emotion detection in narratives (Kim et al., 2017), turning point identification (Papalampidi et al., 2019) and other heuristics inspired by narrative structure may be used to generate such rationales automatically.

Choices and plot are often heavily entwined in IDNs. This work demonstrates a way to control the relative emphasis placed on choices while generating summaries by setting different values for alpha and window sizes. By focusing on parts of the text that vary most across playthroughs, this could potentially lead to a better understanding of how to generate summaries with more variability. Further analysis exploring the relationship between setting different values for these parameters and the resulting document representations for each playthrough is another future direction that could be explored.

Some limitations of this work are that the fault analysis was only done by one annotator. This creates some subjectivity in the relative prevalence of the error classes. Currently, there are very few resources available for interactive narrative summarisation, so another limitation is that we had only one type of IDN to use in the study. The effectiveness of this approach on other types of IDN is yet to be determined. In this paper, we have used a simple attention mechanism as provided by SummaRuNNer's GitHub repository[6]. While this approach can be applied to other model architectures and other types of attention, testing them empirically is outside the scope of this paper. We also do not empirically prove our results can be transferred to non-interactive narrative text sumamrisation, even though we hypothesise this based on our experience in this domain. The results reported are for single runs with specified hyperparameters. While we have used default values for most hyperparameters, it is worth noting that IDN-Sum has many differences from datasets like CNN-DM on which model hyperparameters were tuned by their original creators. Note that the smaller size and repeated sentences across documents in IDN-Sum, can potentially make the model more prone to overfitting and hence more sensitive to hyperparameters and non-determinism. However, due to time and resource constraints, hyperparameter tuning was performed only on the newly introduced hyperparameters - window_size and alpha.

[6]https://github.com/hpzhao/SummaRuNNer

6. Conclusion

Choices are central to interactive narratives and in this paper, we have explored choice focussed self-supervised rationale-based learning at the word and sentence level to improve IDN extractive text summarisation. We believe that our experience developing better summaries for IDNs could transfer to non-interactive narrative summarisation models as well.

Evaluation using ROUGE metrics shows that models trained using these rationales perform up to 14% better than those trained without. An analysis of variability of the produced summaries also indicates that summaries produced by models placing special emphasis on the choices are up to 16% more varied across playthroughs. Manual fault analysis and qualitative analysis were performed which highlighted that the main types of errors present are redundant information, incomplete information and irrelevant information. These analyses also indicate that summaries may be useful in giving a recap of events to readers already familiar with the narrative. However, coverage of choices and differences across playthroughs still appears low.

These results suggest a promising new direction for narrative-based text summarization models. Future work will include evaluation of this approach on more datasets and model architectures with different attention mechanisms, and performing task-based evaluations with IDN authors to assess the utility of these summaries as authoring feedback.

7. Acknowledgements

Will be filled in if the paper is accepted.

8. Bibliographical References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

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.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge & Information Systems*, 62(8).

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Prithviraj Ammanabrolu, Wesley Cheung, Dan Tu, William Broniec, and Mark Riedl. 2020. Bringing stories alive: Generating interactive fiction worlds. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 16(1):3–9.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

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.

Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. 2021. Marta: Leveraging human rationales for explainable text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):5868–5876.

Ruth Aylett. 2000. Emergent narrative, social immersion and "storification". In *Proceedings of the 1st international workshop on narrative and interactive learning environments*, pages 35–44.

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*.

Sebastian Hurup Bevensee, Kasper Alexander Dahlsgaard Boisen, Mikael Peter Olsen, Henrik Schoenau-Fog, and Luis Emilio Bruni. 2012a. Aporia – exploring continuation desire in a game focused on environmental storytelling. In *Interactive Storytelling*, pages 42–47, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sebastian Hurup Bevensee, Kasper Alexander Dahlsgaard Boisen, Mikael Peter Olsen, Henrik Schoenau-Fog, and Luis Emilio Bruni. 2012b. Project aporia – an exploration of narrative understanding of environmental storytelling in an open world scenario. In *Interactive Storytelling*, pages 96–101, Berlin, Heidelberg. Springer Berlin Heidelberg.

Michal Bída, Martin Černý, and Cyril Brom. 2013. Towards automatic story clustering for interactive narrative authoring. In *Proceedings of the 6th International Conference on Interactive Storytelling - Volume 8230*, ICIDS 2013, page 95–106, Berlin, Heidelberg. Springer-Verlag.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Claire Bonial, Tommaso Caselli, Snigdha Chaturvedi, Elizabeth Clark, Ruihong Huang, Mohit Iyyer, Alejandro Jaimes, Heng Ji, Lara J. Martin, Ben Miller, Teruko Mitamura, Nanyun Peng, and Joel Tetreault, editors. 2020. *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*. Association for Computational Linguistics, Online.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Jeanne H. Brockmyer, Christine M. Fox, Kathleen A. Curtiss, Evan McBroom, Kimberly M. Burkhart, and Jacquelyn N. Pidruzny. 2009. The development of the game engagement questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, 45(4):624–634.

Amy Bruckman. 1990. The combinatorics of storytelling: Mystery train interactive.

Luis Emilio Bruni and Sarune Baceviciute. 2013. Narrative intelligibility and closure in interactive systems. In *Interactive Storytelling*, pages 13–24, Cham. Springer International Publishing.

Luis Emilio Bruni, Sarune Baceviciute, and Mohammed Arief. 2014. Narrative cognition in interactive systems: Suspense-surprise and the p300 erp component. In *Interactive Storytelling*, pages 164–175, Cham. Springer International Publishing.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Rick Busselle and Helena Bilandzic. 2009. Measuring narrative engagement. *Media psychology*, 12(4):321–347.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Elin Carstensdottir, Nathan Partlan, Steven Sutherland, Tyler Duke, Erika Ferris, Robin M. Richter, Maria Valladares, and Magy Seif El-Nasr. 2020. Progression maps: Conceptualizing narrative structure for interaction design support. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Ashley Castleberry and Amanda Nolen. 2018. Thematic analysis of qualitative research data: Is it as easy as it sounds? *Currents in Pharmacy Teaching and Learning*, 10:807–815.

Marc Cavazza and R. Michael Young. 2017. *Introduction to Interactive Storytelling*, pages 377–392. Springer Singapore, Singapore.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Snigdha Chaturvedi, Dan Goldwasser, and Hal Daumé. 2016. <i>ask, and shall you receive?</i> understanding desire fulfillment in natural language text. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2697–2703. AAAI Press.

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story comprehension for predicting what happens next. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1603–1614, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Seungtaek Choi, Haeju Park, Jinyoung Yeo, and Seung-won Hwang. 2020. Less is more: Attention supervision with counterfactuals for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6695–6704, Online. Association for Computational Linguistics.

Katheryn R Christy and Jesse Fox. 2016. Transportability and presence as predictors of avatar identification within narrative video games. *Cyberpsychology, Behavior, and Social Networking*, 19(4):283–287.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022a. Scaling instruction-finetuned language models.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022b. Scaling instruction-finetuned language models.

Davide Colla, Enrico Mensa, and Daniele P. Radicioni. 2020. Novel metrics for computing semantic similarity with sense embeddings. *Knowledge-Based Systems*, 206:106346.

Peng Cui, Le Hu, and Yuanchao Liu. 2020. Enhancing extractive text summarization with topic-aware graph neural networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5360–5371, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Peter David. Zork. Video Game.

Edirlei Soares de Lima, Bruno Feijó, Simone Barbosa, Antonio L. Furtado, Angelo Ciarlini, and Cesar Pozzer. 2011. Draw your own story: Paper and pencil interactive storytelling. In *Entertainment Computing – ICEC 2011*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

Edirlei Soares de Lima, Bruno Feijó, and Antonio L Furtado. 2018. Video-based interactive storytelling using real-time video compositing techniques. *Multimedia Tools and Applications*, 77(2):2333–2357.

Marie-Catherine De Marneffe, Anna N Rafferty, and Christopher D Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Andrea Di Pastena, Dennis Jansen, Brian de Lint, and Amanda Moss. 2018. “the link out”. In *Interactive Storytelling*, pages 206–216, Cham. Springer International Publishing.

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. 2021a. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

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

Dontnod Entertainment. Life is strange. Video Game. PC, PS4, PS3, Xbox 360, Xbox One.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Guy Feigenblat, Haggai Roitman, Odellia Boni, and David Konopnicki. 2017. Unsupervised query-focused multi-document summarization using the cross entropy method. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’17, page 961–964, New York, NY, USA. Association for Computing Machinery.

Rui Figueiredo and Ana Paiva. 2011. “i’m sure i made the right choice!” - towards an architecture to influence player’s behaviors in interactive stories. In *Interactive Storytelling*, pages 152–157, Berlin, Heidelberg. Springer Berlin Heidelberg.

Lucie Flekova and Iryna Gurevych. 2015. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816, Lisbon, Portugal. Association for Computational Linguistics.

S. Fortunato. 2010. Community detection in graphs. *Phys. Rep.-Rev. Sec. Phys. Lett.*, 486:75–174.

Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. 2022a. Going beyond xai: A systematic survey for explanation-guided learning. *arXiv preprint arXiv:2212.03954*.

Yuyang Gao, Tong Steven Sun, Liang Zhao, and Sungsoo Ray Hong. 2022b. Aligning eyes between humans and deep neural network through interactive attention alignment. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

Jacob Garbe. 2020. *Increasing Authorial Leverage in Generative Narrative Systems*. University of California, Santa Cruz.

Barney G Glaser, Anselm L Strauss, and Elizabeth Strutzel. 1968. The discovery of grounded theory; strategies for qualitative research. *Nursing research*, 17(4):364.

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.

Daniel Green. 2018. Novella: An authoring tool for interactive storytelling in games. In *Interactive Storytelling*, pages 556–559, Cham. Springer International Publishing.

Melanie C Green and Timothy C Brock. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology*, 79(5):701.

Melanie C Green and Keenan M Jenkins. 2014. Interactive narratives: Processes and outcomes in user-directed stories. *Journal of Communication*, 64(3):479–500.

Arne Grinder-Hansen and Henrik Schoenau-Fog. 2013. The elements of a narrative environment. In *Interactive Storytelling*, pages 186–191, Cham. Springer International Publishing.

Xinyu Guan, Qinke Peng, Xintong Li, and Zhibo Zhu. 2019. Social emotion prediction with attention-based hierarchical neural network. In *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 1, pages 1001–1005.

Vivek Gupta, Prerna Bharti, Pegah Nokhiz, and Harish Karnick. 2021. SumPubMed: Summarization dataset of PubMed scientific articles. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 292–303, Online. Association for Computational Linguistics.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mohamed Hamroun and Mohamed Salah Gouider. 2020. A survey on intention analysis: successful approaches and open challenges. *Journal of Intelligent Information Systems*, 55(3):423–443.

Hákon Jarl Hannesson, Thorbjørn Reimann-Andersen, Paolo Burelli, and Luis Emilio Bruni. 2015. Connecting the dots: Quantifying the narrative experience in interactive media. In *Interactive Storytelling*, pages 189–201, Cham. Springer International Publishing.

Benjamin Hättasch, Nadja Geisler, Christian M. Meyer, and Carsten Binnig. 2020. Summarization beyond news: The automatically acquired fandom corpora. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6700–6708, Marseille, France. European Language Resources Association.

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.

Zhichao Hu and Marilyn Walker. 2017. Inferring narrative causality between event pairs in films.

In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 342–351, Saarbrücken, Germany. Association for Computational Linguistics.

E. Hullermeier and M. Rifqi. 2009. A fuzzy variant of the rand index for comparing clustering structures. In *in Proc. IFSA/EUSFLAT Conf.*, pages 1294–1298.

Matthew L. Jockers and David Mimno. 2013. Significant themes in 19th-century literature. *Poetics*, 41(6):750–769. Topic Models and the Cultural Sciences.

Michael Joyce. Afternoon, a story. online.

Teja Kanchinadam, Keith Westpfahl, Qian You, and Glenn Fung. 2020. Rationale-based human-in-the-loop via supervised attention.

Anna Kazantseva and Stan Szpakowicz. 2010. Summarizing short stories. *Computational Linguistics*, 36(1):71–109.

Pooja Kherwa and Poonam Bansal. 2020. Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, 7(24).

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26, Vancouver, Canada. Association for Computational Linguistics.

Sofia Kitromili, James Jordan, and David E. Millard. 2020. What authors think about hypertext authoring. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, page 9–16, New York, NY, USA. Association for Computing Machinery.

Erica Kleinman, Karina Caro, and Jichen Zhu. 2020. From immersion to metagaming: Understanding rewind mechanics in interactive storytelling. *Entertainment Computing*, 33:100322.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Lobke Kolhoff and Frank Nack. 2019. How relevant is your choice? In *Interactive Storytelling*, pages 73–85, Cham. Springer International Publishing.

Vonnegut Kurt. Shapes of stories.

Vincent Labatut and Xavier Bost. 2019a. Extraction and analysis of fictional character networks: A survey. *ACM Comput. Surv.*, 52(5).

Vincent Labatut and Xavier Bost. 2019b. Extraction and analysis of fictional character networks: A survey. *ACM Comput. Surv.*, 52(5).

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.

Boyang Li, Beth Cardier, Tong Wang, and Florian Metze. 2018. Annotating high-level structures of short stories and personal anecdotes. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3290–3296, Miyazaki, Japan. European Language Resources Association (ELRA).

Zhongyang Li, Tongfei Chen, and Benjamin Van Durme. 2019. Learning to rank for plausible plausibility. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4818–4823, Florence, Italy. Association for Computational Linguistics.

Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. *arXiv preprint arXiv:1906.08286*.

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.

Inderjeet Mani. 2012. Computational modeling of narrative. *Synthesis Lectures on Human Language Technologies*, 5(3):1–142.

Chris Martens and Owais Iqbal. 2019. Villanelle: An authoring tool for autonomous characters in interactive fiction. In *Interactive Storytelling*, pages 290–303, Cham. Springer International Publishing.

Marcel Marti, Jodok Vieli, Wojciech Witoń, Rushit Sanghrajka, Daniel Inversini, Diana Wotruba, Isabel Simo, Sasha Schriber, Mubbasir Kapadia, and Markus Gross. 2018. Cardinal: Computer assisted authoring of movie scripts. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, page 509–519, New York, NY, USA. Association for Computing Machinery.

Joshua McCoy, Mike Treanor, Ben Samuel, Aaron A Reed, Michael Mateas, and Noah Wardrip-Fruin. 2014. Social story worlds with comme il faut. *IEEE Transactions on Computational intelligence and AI in Games*, 6(2):97–112.

Mohsen Mesgar and Michael Strube. 2018. A neural local coherence model for text quality assessment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4328–4339, Brussels, Belgium. 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.

Matthew K. Miller and Regan L. Mandryk. 2016. Differentiating in-game frustration from at-game frustration using touch pressure. In *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces*, ISS '16, page 225–234, New York, NY, USA. Association for Computing Machinery.

Alex Mitchell and Kevin McGee. 2011. Supporting rereadability through narrative play. In *Interactive Storytelling*, pages 67–78, Berlin, Heidelberg. Springer Berlin Heidelberg.

Andreea Molnar, David Farrell, and Patty Kostova. 2012. Who poisoned hugh?-the star framework: integrating learning objectives with storytelling. In *International Conference on Interactive Digital Storytelling*, pages 60–71. Springer.

Christopher Moser and Xiaowen Fang. 2014. Narrative control and player experience in role playing games: Decision points and branching narrative feedback. In *Human-Computer Interaction. Applications and Services*, pages 622–633, Cham. Springer International Publishing.

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.

Rajdeep Mukherjee, Hari Chandana Peruri, Uppada Vishnu, Pawan Goyal, Sourangshu Bhattacharya, and Niloy Ganguly. 2020. Read what you need: Controllable aspect-based opinion summarization of tourist reviews. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1825–1828, New York, NY, USA. Association for Computing Machinery.

Janet Horowitz Murray. 2017. *Hamlet on the holodeck: The future of narrative in cyberspace*. MIT press.

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*.

Elena Novak. 2015. A critical review of digital storyline-enhanced learning. *Educational Technology Research and Development*, 63(3):431–453.

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.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie plot analysis via turning point identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1707–1717, Hong Kong, China. Association for Computational Linguistics.

Nathan Partlan, Elin Carstensdottir, Sam Snodgrass, Erica Kleinman, Gillian Smith, Casper Harteveld, and Magy Seif El-Nasr. 2018. Exploratory automated analysis of structural features of interactive narrative. In *Fourteenth Artificial Intelligence and Interactive Digital Entertainment Conference*, pages 88–94.

Bronwin Patrickson. 2011a. Multi-user interactive drama: A micro user drama in process. In *Interactive Storytelling*, pages 199–206, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bronwin Patrickson. 2011b. Multi-user interactive drama: The macro view - three structural layers. In *Interactive Storytelling*, pages 317–321, Berlin, Heidelberg. Springer Berlin Heidelberg.

David Pizzi and Marc Cavazza. 2008. From debugging to authoring: Adapting productivity tools to narrative content description. In *Interactive Storytelling*, pages 285–296, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Elahe Rahimtoroghi, Thomas Corcoran, Reid Swanson, Marilyn A Walker, Kenji Sagae, and Andrew Gordon. 2014. Minimal narrative annotation schemes and their applications. In *Seventh Intelligent Narrative Technologies Workshop*, pages 31–37.

Elahe Rahimtoroghi, Jiaqi Wu, Ruimin Wang, Pranav Anand, and Marilyn Walker. 2017. Modelling protagonist goals and desires in first-person narrative. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–369, Saarbrücken, Germany. 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.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):1–12.

Ashwathy T. Revi, Stuart E. Middleton, and David E. Millard. 2022. IDN-sum: A new dataset for interactive digital narrative extractive text summarisation. In *Proceedings of The Workshop on Automatic Summarization for Creative Writing*, pages 1–12, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ashwathy T. Revi, David E. Millard, and Stuart E. Middleton. 2020. A systematic analysis of user experience dimensions for interactive digital narratives. In *Interactive Storytelling*, pages 58–74, Cham. Springer International Publishing.

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.

David L Roberts and Charles L Isbell. 2007. Desiderata for managers of interactive experiences: A survey of recent advances in drama management. In *Proceedings of the First Workshop on Agent-Based Systems for Human Learning and Entertainment (ABSHLE 07)*.

Melissa Roemmele and Andrew Gordon. 2018. An encoder-decoder approach to predicting causal relations in stories. In *Proceedings of the First Workshop on Storytelling*, pages 50–59, New Orleans, Louisiana. Association for Computational Linguistics.

Christian Roth. 2019. The 'angstfabriek' experience: Factoring fear into transformative interactive narrative design. In *Interactive Storytelling*, pages 101–114, Cham. Springer International Publishing.

Christian Roth, Christoph Klimmt, Ivar E. Vermeulen, and Peter Vorderer. 2011. The experience of interactive storytelling: Comparing "fahrenheit" with "façade". In *Entertainment Computing – ICEC 2011*, pages 13–21, Berlin, Heidelberg. Springer Berlin Heidelberg.

Christian Roth and Hartmut Koenitz. 2016. Evaluating the user experience of interactive digital narrative. In *Proceedings of the 1st International Workshop on Multimedia Alternate Realities*, AltMM '16, page 31–36, New York, NY, USA. Association for Computing Machinery.

Ben Samuel, Michael Mateas, and Noah Wardrip-Fruin. 2016. The design of writing buddy: a mixed-initiative approach towards computational story collaboration. In *International Conference on Interactive Digital Storytelling*, pages 388–396. 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.

Rushit Sanghrajka, Daniel Hidalgo, Patrick Chen, and Mubbasir Kapadia. 2021. Lisa: Lexically intelligent story assistant. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 13(1):221–227.

Sebastian Sauer, Kerstin Osswald, Xavier Wielemans, and Matthias Stifter. 2006. U-create: Creative authoring tools for edutainment applications. In *Technologies for Interactive Digital Storytelling* *and Entertainment*, pages 163–168, Berlin, Heidelberg. Springer Berlin Heidelberg.

Henrik Schoenau-Fog. 2011. Hooked! – evaluating engagement as continuation desire in interactive narratives. In *Interactive Storytelling*, pages 219–230, Berlin, Heidelberg. Springer Berlin Heidelberg.

Henrik Schoenau-Fog, Luis Emilio Bruni, Faysal Fuad Khalil, and Jawid Faizi. 2010. First person victim: Developing a 3d interactive dramatic experience. In *Interactive Storytelling*, pages 240–243, Berlin, Heidelberg. Springer Berlin Heidelberg.

Henrik Schoenau-Fog, Luis Emilio Bruni, Faysal Fuad Khalil, and Jawid Faizi. 2013. Authoring for engagement in plot-based interactive dramatic experiences for learning. In *Transactions on Edutainment X*, pages 1–19, Berlin, Heidelberg. Springer Berlin Heidelberg.

Magy Seif El-Nasr, David Milam, and Tony Maygoli. 2013. Experiencing interactive narrative: A qualitative analysis of façade. *Entertainment Computing*, 4(1):39–52.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Yotam Shibolet, Noam Knoller, and Hartmut Koenitz. 2018. A framework for classifying and describing authoring tools for interactive digital narrative. In *Interactive Storytelling*, pages 523–533, Cham. Springer International Publishing.

Victor Socas-Guerra and Carina S. González-González. 2012. User attention in nonlinear narratives: A case of study. In *Communicability, Computer Graphics and Innovative Design for Interactive Systems*, pages 104–111, Berlin, Heidelberg. Springer Berlin Heidelberg.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 950–961, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Swapna Somasundaran, Xianyang Chen, and Michael Flor. 2020. Emotion arcs of student narratives. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 97–107, Online. Association for Computational Linguistics.

Ulrike Spierling and Nicolas Szilas. 2009. Authoring issues beyond tools. In *Interactive Storytelling*, pages 50–61, Berlin, Heidelberg. Springer Berlin Heidelberg.

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Supervising model attention with human explanations for robust natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11349–11357.

Ingibergur Stefnisson and David Thue. 2018. Mimisbrunnur: Ai-assisted authoring for interactive storytelling. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 14(1):236–242.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv e-prints*, pages arXiv–1904.

Adam Summerville, Sam Snodgrass, Matthew Guzdial, Christoffer Holmgård, Amy K Hoover, Aaron Isaksen, Andy Nealen, and Julian Togelius. 2018. Procedural content generation via machine learning (pcgml). *IEEE Transactions on Games*, 10(3):257–270.

Tomi "bgt" Suovuo, Natasha Skult, Tapani N. Joelsson, Petter Skult, Werner Ravyse, and Jouni Smed. 2020. *The Game Experience Model (GEM)*, pages 183–205. Springer International Publishing, Cham.

Neil Suttie, Sandy Louchart, Ruth Aylett, and Theodore Lim. 2013. Theoretical considerations towards authoring emergent narrative. In *Interactive Storytelling*, pages 205–216, Cham. Springer International Publishing.

Ivo Swartjes and Mariët Theune. 2009. Iterative authoring using story generation feedback: Debugging or co-creation? In *Interactive Storytelling*, pages 62–73, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ayesha Ayub Syed, Ford Lumban Gaol, and Tokuro Matsuo. 2021. A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access*, 9:13248–13265.

Nicolas Szilas and Ioana Ilea. 2014. Objective metrics for interactive narrative. In *Interactive Storytelling*, pages 91–102, Cham. Springer International Publishing.

Peggy Tang, Kun Hu, Rui Yan, Lei Zhang, Junbin Gao, and Zhiyong Wang. 2022. OTExtSum: Extractive Text Summarisation with Optimal Transport. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1128–1141, Seattle, United States. Association for Computational Linguistics.

Bryan Temprado-Battad, José-Luis Sierra, and Antonio Sarasa-Cabezuelo. 2019. An online authoring tool for interactive fiction. In *2019 23rd International Conference Information Visualisation (IV)*, pages 339–344.

Mariët Theune, Jeroen Linssen, and Thijs Alofs. 2013. Acting, playing, or talking about the story: An annotation scheme for communication during interactive digital storytelling. In *Interactive Storytelling*, pages 132–143, Cham. Springer International Publishing.

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.

Milo N.R. Utsch, Gisele L. Pappa, Luiz Chaimowicz, and Raquel O. Prates. 2020. A new non-deterministic drama manager for adaptive interactive storytelling. *Entertainment Computing*, 34:100364.

Josep Valls-Vargas, Santiago Ontanón, and Jichen Zhu. 2014. Toward automatic character identification in unannotated narrative text. In *Seventh intelligent narrative technologies workshop*, pages 38–44.

Josep Valls-Vargas, Jichen Zhu, and Santiago Ontañón. 2017. From computational narrative analysis to generation: A preliminary review. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*, FDG '17, pages 1–4, New York, NY, USA. Association for Computing Machinery.

Josep Valls-Vargas, Jichen Zhu, and Santiago Ontañón. 2021. Toward automatic role identification in unannotated folk tales. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 10(1):188–194.

Renske van Enschot, Iris Boogaard, Hartmut Koenitz, and Christian Roth. 2019. The potential of interactive digital narratives. agency and multiple perspectives in last hijack interactive. In *Interactive Storytelling*, pages 158–169, Cham. Springer International Publishing.

K Vani and Alessandro Antonucci. 2019. Novel2graph: Visual summaries of narrative text enhanced by machine learning. *Text2Story@ ECIR*, pages 29–37.

Maria Vayanou, Yannis Ioannidis, George Loumos, and Antonis Kargas. 2019. How to play storytelling games with masterpieces: from art galleries to hybrid board games. *Journal of Computers in Education*, 6(1):79–116.

Ivar E. Vermeulen, Christian Roth, Peter Vorderer, and Christoph Klimmt. 2010. Measuring user responses to interactive stories: Towards a standardized assessment tool. In *Interactive Storytelling*, pages 38–43, Berlin, Heidelberg. Springer Berlin Heidelberg.

Claudia Volpetti, K. Vani, and Alessandro Antonucci. 2020. Temporal word embeddings for narrative understanding. In *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*, ICMLC 2020, page 68–72, New York, NY, USA. Association for Computing Machinery.

Mirjam Vosmeer and Ben Schouten. 2014. Interactive cinema: Engagement and interaction. In *Interactive Storytelling*, pages 140–147, Cham. Springer International Publishing.

Tyrone Vriesede and Frank Nack. 2011. Storystream: Unrestricted mobile exploration of city neighbourhoods enriched by the oral presentation of user-generated stories. In *Interactive Storytelling*, pages 231–242, Berlin, Heidelberg. Springer Berlin Heidelberg.

David Wilmot and Frank Keller. 2020. Modelling suspense in short stories as uncertainty reduction over neural representation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1763–1788, Online. Association for Computational Linguistics.

Bob G Witmer and Michael J Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence*, 7(3):225–240.

Zongda Wu, Li Lei, Guiling Li, Hui Huang, Chengren Zheng, Enhong Chen, and Guandong Xu. 2017. A topic modeling based approach to novel document automatic summarization. *Expert Systems with Applications*, 84:12–23.

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.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Ruifeng Yuan, Shichao Sun, Zili Wang, Ziqiang Cao, and Wenjie Li. 2023. Separating context and pattern: Learning disentangled sentence representations for low-resource extractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7575–7586, Toronto, Canada. Association for Computational Linguistics.

I Schankler Z Quinn, P Lindsey. Depression quest. online.

Nelson Zagalo, Sandy Louchart, and Maria T. Soto-Sanfiel. 2010. Users and evaluation of interactive storytelling. In *Interactive Storytelling*, pages 287–288, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Ruiqi Zhong, Steven Shao, and Kathleen McKeown. 2019. Fine-grained sentiment analysis with faithful attention. *arXiv preprint arXiv:1908.06870*.

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.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2020. A joint sentence scoring and selection framework for neural extractive document summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:671–681.

Suyang Zhu, Shoushan Li, and Guodong Zhou. 2019. Adversarial attention modeling for multi-dimensional emotion regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 471–480, Florence, Italy. Association for Computational Linguistics.

9. Language Resource References