UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Sciences
School of Electronics and Computer Science

An Investigation into NLP Techniques for Generating Intelligent Narrative Feedback to Support IDN Authoring

by

Ashwathy T Revi

MSc in AI

ORCiD: 0000-0002-9936-8141

A thesis for the degree of
Doctor of Philosophy

November 2024

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences
School of Electronics and Computer Science

Doctor of Philosophy

An Investigation into NLP Techniques for Generating Intelligent Narrative Feedback to Support IDN Authoring

by Ashwathy T Revi

Authoring Interactive Digital Narratives (IDN) is challenging since past a certain size, it becomes hard to keep track of the user's experience along all the different storylines. Natural Language Processing (NLP) provides us with the opportunity to generate such intelligent feedback that can help authors keep better track of the story space. This is what this PhD addresses. In the first phase a systematic review of IDN literature is performed and list of User experience (UX) dimensions that could form the basis of feedback to authors is compiled. The second phase then maps these onto related areas of NLP research to see where these could be estimated automatically. This reveals 47 dimensions of UX covering 8 categories—23 of these map to 12 areas of NLP research, leading on to 5 specific examples of how they might help IDN authors: plotting emotional arcs, visualising emotion type and intensity, revealing the predictability of events, debugging internal story logic, and branch-wise summarization. One of these NLP areas (Automatic Text Summarisation) is chosen for deeper investigation in Phase 3. A dataset is generated by simulating playthroughs of eight episodes from two narrative games - *Before the Storm* and *Wolf Among Us* using fan-created transcripts online. Annotations for extractive summarisation were created automatically by aligning extracts with fan-made abstractive summaries available online. The dataset is released as open source for future researchers to train and test their approaches for IDN text. On applying common baseline extractive text summarization approaches to this dataset, several shortcomings in standard approaches are revealed when applied to narrative and interactive narrative datasets. The last phase of this work experiments with using rationale-based learning with word-level and sentence-level rationales indicating the proximity of words and sentences to choice points. The 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 suggesting a promising new direction for narrative-based text summarisation models. In this way, this thesis takes a step toward generating authoring feedback to assist IDN authors as well as understanding the complexities and unique challenges posed by the domain.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

7. Parts of this work have been published as:
   Ashwathy T. Revi, David E. Millard, and Stuart E. Middleton. A systematic analysis of user experience dimensions for interactive digital narratives. In Anne-Gwenn Bosser, David E. Millard, and Charlie Hargood, editors, *Interactive Storytelling*, pages 58–74, Cham, 2020. Springer International Publishing. ISBN 978-3-030-62516-0

   Ashwathy T. Revi, Stuart E. Middleton, and David E. Millard. 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, October 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.creativesumm-1.1

   Ashwathy T Revi, Stuart E Middleton, and David E Millard. Rationale-based learning using self-supervised narrative events for text summarisation of interactive digital narratives. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13557–13585, 2024

Signed:..... Date:.....

Acknowledgements

This thesis stands as a testament to the support and guidance I've been fortunate to receive from a group of remarkable people.

Foremost, I express my deepest appreciation to my supervisors, Prof David Millard and Dr Stuart Middleton, whose encouragement, insight, and invaluable feedback have been the backbone of my research journey. A special thanks to Dave, for introducing me to the world of interactive narratives and inspiring me to pursue this research.

My sincere gratitude goes to the Engineering and Physical Sciences Research Council (EPSRC) for their generous funding, which made this research possible. I would also like to acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

I thank Prof Jonathon Hare and Dr Nicholas Gibbins for their feedback on my work. I would also like to thank the anonymous reviewers who provided insightful feedback on intermediate chapters in this thesis that were submitted as papers to various venues.

To my mother, from whom I inherited my love of learning, to my father, who passed on his love of stories, and my grandmom for her unconditional support, I am forever thankful.

For the countless hours of creativity and escape, my thanks go to Joey, Lesia, and Sam for our unforgettable role-playing sessions. Equally, to Appu, Febin, Ammini, and Vava, thank you for being a constant source of support and encouragement. I am also immensely grateful to Alex, Morgan, Denis, Marie, Grace, Dev, and Eirini, for being my found family in the UK for a little while.

Most importantly, I want to thank my best friend and partner, Pav - for everything from giving me feedback on my work to keeping me sane with his love and support. Thank you seems an understatement, but here it is anyway.

To everyone who has been a part of this journey, in ways big and small, I extend my heartfelt thanks.

To Matacha,
whose memory remains a guiding light . . .

Definitions and Abbreviations

NLP	Natural Language Processing
IDN	Interactive Digital Narratives
UX	User Experience
AI	Artificial Intelligence
RNN	Recurrent Neural Networks
LLM	Large Language Models
LSTM	Long Short Term Memory
GRU	Gated Recurrent Units

Chapter 1

Introduction

1.1 Motivation

A large amount of work in AI for creative projects has focused on Generative AI that tries to replicate human creativity. Instead, this research joins a rapidly growing interest in exploring ways in which AI can be used in an assistive or collaborative capacity to augment the creative process. It looks specifically at the area of Interactive Narrative Authoring where AI has the potential to help creators keep track of and manage complex state spaces during the design process.

Interactive Digital Narratives (IDNs) are a medium of storytelling that allows the audience to actively participate in the narrative by interacting with it. Through their interactions, they can often influence the course and the outcome of the narrative. IDNs take various forms including story-rich video games, digital choose-your-own-adventure style games, and hypertext fiction. Unlike traditional linear narratives like novels and movies, IDNs often involve additional elements of complexity like branching plot lines arising from decision-making by the audience. This could result in nonlinear structures of varying levels of complexity (an example of this is illustrated in Figure 1.1). This makes the process of creating IDNs challenging for authors as they have to keep track of multiple storylines and envision how the audience will experience their work along all those different storylines.

Authors often need to conduct iterative playtesting to understand how the audience will experience their work, but this can be time-consuming and resource-intensive. Giving the author automatically generated feedback on the potential experiences possible within their work (referred to as Narrative Analytics in [156] and Intelligent Narrative Feedback in [217]), has been proposed as a way to overcome this issue. Artificial Intelligence (AI) and Natural Language Processing (NLP) open up many opportunities for generating intelligent narrative feedback; for example, sentiment

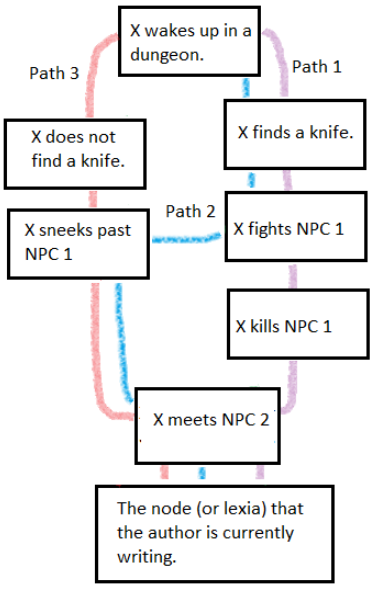

FIGURE 1.1: Lines in different colours show different story paths that lead up to the same piece of text (or lexia). Feedback based on automatic analysis along all the paths can be shown to the author in real time while they're writing that particular lexia to help them keep context. For example, here, the author needs to account for the fact that at this stage, NPC 1 may be dead or alive depending on the path they have taken.

networks depicting the evolution of relationships between characters[132] and emotional arcs [182] can be used to generate different perspectives of the narrative and can potentially serve as useful feedback to the authors. By simulating playthroughs through an interactive narrative and providing insights into the potential reader experiences, AI can assist authors in creating more complex and engaging narratives.

While NLP research has been done on many narrative domains including novels and movie scripts, NLP approaches for interactive narratives are relatively under-explored. Applying general NLP approaches to the IDN domain is non-trivial since IDN has distinct features such as interactivity and non-linearity. Hence, existing approaches need to be tested and adapted for this domain. With interactive media becoming more and more prevalent, this poses an important yet under-explored domain for NLP research.

1.2 Summary

This work investigates NLP techniques for generating feedback to enhance the IDN authoring process. To achieve this, the research will first determine which aspects of

the reader's/player's experience are most important to IDN authors. It then investigates if and how NLP techniques can be applied to simulated playthroughs of interactive narratives to give insight into these aspects of player experience. The practical challenges and effectiveness of one of these approaches (Automatic Text Summarization), are then investigated more deeply. A new dataset to study IDN summarisation is created and standard summarisation approaches are applied to this dataset to determine the extent to which they work on IDN text. Finally, a modification of an existing summarisation approach to better suit the IDN domain is proposed and evaluated. Through this process, this research hopes to pave the way for future research in Human-AI collaborative creation, NLP techniques for interactive narratives and improving the IDN authoring process by utilizing the power of AI-generated feedback.

1.3 Research Questions

The overarching research objective motivated in the previous section is summarised and broken down below:

Research Objective: How can NLP be used to generate intelligent narrative feedback to assist authoring of IDN?

While previous work that proposed intelligent narrative feedback [217] and Narrative Analytics [156] provide some examples of automatically generated feedback, they do not systematically study what types of feedback would be useful to IDN authors and what is feasible to generate using advanced NLP techniques. This leads to the first Research Question:

RQ1: What type of feedback has the potential to be both useful to IDN authors and feasible to generate using NLP techniques?

1. What concrete aspects of the reader/player's experience interest IDN authors? (impact)
2. How can NLP techniques be applied to generate feedback that can give insight into these aspects of player experience? (feasibility)

One form of feedback is then picked for deeper investigation. The analysis described in Chapters 3 and 4 answers RQ1 and suggests that feedback in the form of automatically generated summarisation has the potential to be both impactful and feasible. This leads to RQ2:

RQ2: How can IDN text be summarised automatically?

1. Can standard summarisation approaches be applied to this domain (IDN text)?
2. Can they be adapted to better suit this domain (IDN text)?

In this way, this research approaches the research objective through different levels of increasing depth - first answering it in terms of opportunities revealed at the literature search level, then in terms of the practicalities of applying standard approaches to exploit one of these opportunities and finally by investigating one way to adapt these standard approaches to better suit the domain.

1.4 Research Framework

The research is broken down into four phases, each of which builds on the findings from previous experiments to take the investigation a step deeper. The experiments, their outcomes and how the outcomes of each experiment inform the next are summarised in figure 1.2 and explained below:

The first phase (carried out in 2020) investigates which aspects of user experience (UX) would be useful to IDN authors if estimated automatically. Through a systematic analysis of IDN literature and thematic coding of UX dimensions discussed in them, it was found that there were 47 codes spanning 8 categories that represent aspects of UX that could be useful to IDN authors if made available as automatically generated feedback.

The second phase (carried out in late 2020 and early 2021) examines which of the UX dimensions identified in Phase 1 are feasible to estimate using NLP techniques. Through an exploratory review of NLP literature and mapping NLP problems to UX dimensions, it was found that 24 UX dimensions have some associated NLP research that could be applied to automatically generate feedback that could give insight into them. The keywords used for the exploratory review were informed by the UX dimensions from Phase 1. 5 types of feedback items related to some of these UX dimensions were identified that could potentially be implemented using existing NLP techniques.

Out of the 5 types of feedback identified in Phase 2, feedback in the form of automatically generated extractive summaries was selected for further investigation in Phase 3, since it gives the author insight into several important UX dimensions at once and is well-studied within the NLP community. Summaries can also a more intuitive form of feedback than metrics or graphs for the author. Extractive summaries (summaries consisting of the most important extracts from the original text similar to recaps of previous episodes in TV shows) were chosen over abstractive summaries (summaries where new text that synthesises and condenses the original text is

generated) since the latter is prone to hallucination and harder to trace back to the original text.

The third phase (carried out over 2021 and 2022) investigates how well standard summarization approaches work on IDN data. To do this, a dataset for IDN summarisation was compiled using resources available online and computationally simulating different paths through two popular narrative games. Out of the standard extractive summarisation models tested, an RNN-based model, SummaRuNNer [169], gave the best performance. This exercise revealed four challenges regarding the applicability of standard approaches to narrative and interactive narrative summarisation suggesting directions of research to improve IDN summarisation. One of these is that existing approaches do not place any special emphasis on the regions of text that correspond to interaction (for example points in the IDN where the player can make choices) when summarising, but doing so could help with some of the challenges. This was chosen as the direction for further research since interactions are an important aspect of IDN.

The last phase investigates focusing on choices and decision points when summarizing IDN. Specifically, it examines whether choice-based explanations improve IDN summarization by comparing the performance of modified versions of the classic SummaRuNNer model trained with different choice-based explanations and without. The dataset from phase 3 was used to train these models. Models trained with sentence-level choice-based explanations outperformed all other models showing that annotations indicating the proximity of sentences to choice points are effective explanations for IDN summarisation. The work for this phase was done primarily in 2022, with some experiments using Google Flan T5 conducted in 2023 due to the increasing success and attention devoted to Large Language Models in the NLP research community since December 2022.

1.5 Contribution and Novelty

The four main contributions made by this PhD research are listed below:

1. **The Codebook shown in Tables 3.2, 3.3 and 3.4 showing 47 concrete aspects of UX that are likely to be of interest to IDN authors**

   Through a systematic literature review, Phase 1 brings together and untangles different interpretations of User Experience (UX) in the interactive digital narratives (IDN) literature, resulting in a list of 47 concrete aspects of UX that could be useful to authors if provided as feedback. It also provides insight into the relative interest and usefulness of modelling different dimensions of UX in the IDN community and offers a starting point for generating automated

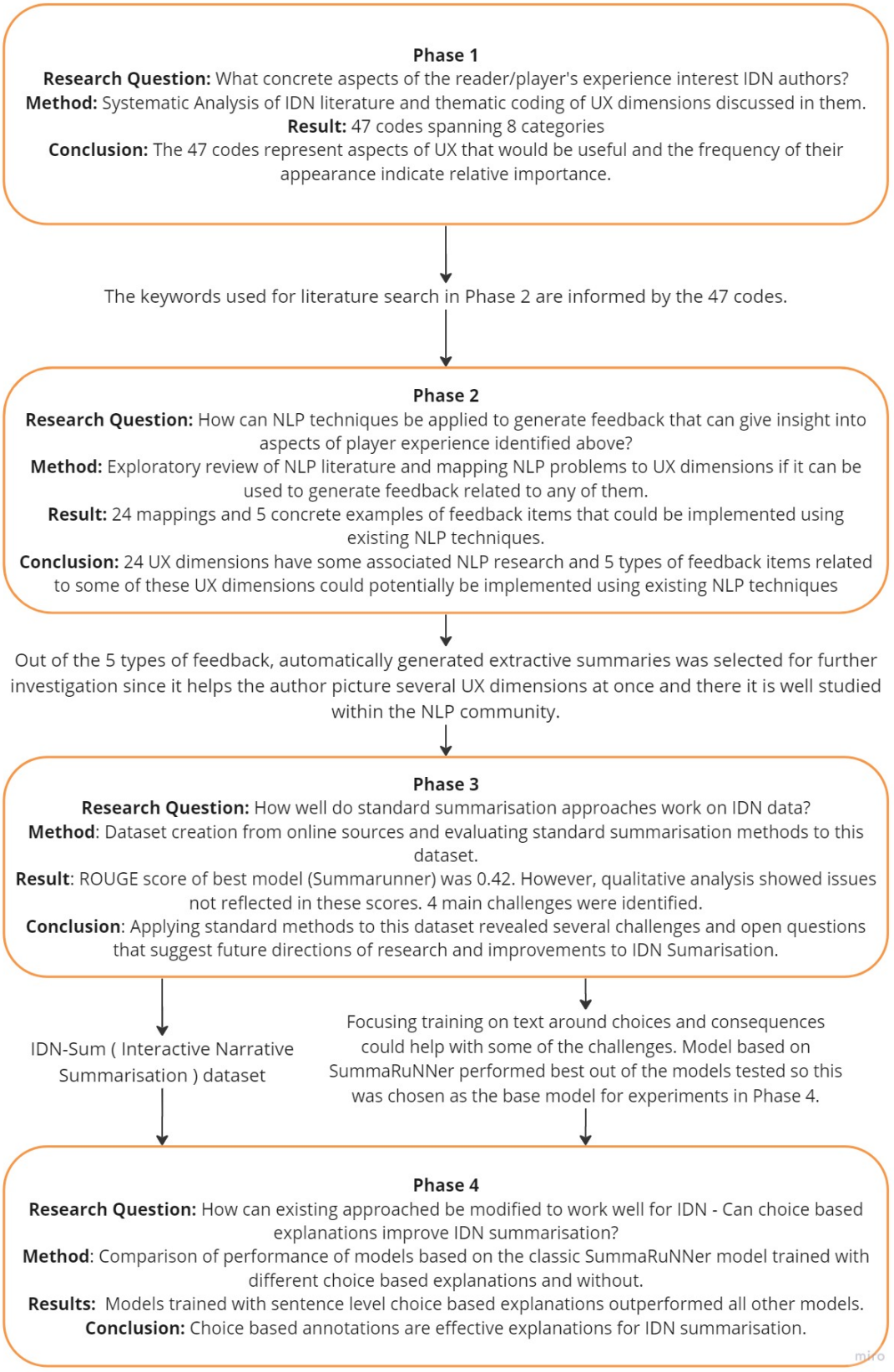

FIGURE 1.2: Research Framework

feedback for IDN authors to assist in authoring interactive narratives. The research also gives a broader and more complete understanding of UX for IDN. This research is published as part of ICIDs proceedings [184].

2. **Tables 4.1 and 4.2 showing how UX Dimensions in IDN map to problems being investigated by the NLP research community**

   Phase 2 bridges the Natural Language Processing (NLP) and IDN research communities by mapping theoretical problems that are of interest to both communities. It shows the untapped potential of applying NLP to generate automatic feedback to assist planning and authoring of IDN and gives concrete examples of possible feedback items that can be generated using NLP methods. It also highlights the value of adapting NLP techniques to fit the IDN domain and this use case. Additionally, it identifies new directions of research for the NLP community in terms of modelling and estimating concepts like Dissonance from the text for which no associated NLP research exists. This research is under review for publication in IEEE Multimedia.

3. **IDN-Sum Dataset - The first dataset for interactive narrative summarisation and evaluation of standard summarisation approaches on this dataset**

   Phase 3 led to the creation of the IDN-Sum dataset, the first dataset for interactive digital narrative (IDN) which captures many different paths through interactive narratives. The dataset has a high amount of overlapping text between data points, making it unique compared to other summarisation datasets. The dataset can be used to investigate summarization approaches for interactive narratives and study new NLP problems such as comparative plot summarization. The dataset includes 10000 playthroughs split equally over 8 episodes of 2 IDN games, made available online. This experiment also applied some standard summarization approaches to linear playthroughs of the IDN and analyzed the performance of these approaches quantitatively and qualitatively. This research was published in the proceedings of Automatic Summarisation of Creative Writing Workshop at COLING2022 [185].

4. **A novel approach for interactive narrative summarisation that leverages rationale-based learning with self-supervised rationales**

   Experiments in Phase 4 is the first to apply rationale-based learning to interactive narrative summarization. The results showed that using rationales in training can improve performance both in terms of ROUGE scores and the variety of sentences in the generated summary across playthroughs. The analysis of error types in the model-generated summaries also provides insights into where the model can improve. This research is published at LREC/COLING 2024[186].

1.6 Thesis Outline

Background outlined in Chapter 2 situates this work in the context of related areas of research - Interactive Digital Narratives, Natural Language Processing and Human-AI Collaboration. Background reading suggests that while automatic feedback has been proposed as a way to assist authoring and there are ways to generate such intelligent feedback using NLP, this has not been investigated in much depth. From existing work, it is unclear what exactly this feedback could be. This motivates Phases 1 and 2. It also shows how very few resources could be found for the automatic summarisation of interactive and game narratives. This motivates dataset creation in Phase 3. And finally, it gives some background on attention mechanisms and rationale-based learning, illustrating how this presents a unique opportunity to focus the training of IDN summarisation models on regions of text surrounding the choice points and consequences, motivating Phase 4.

Chapter 3 describes a systematic literature review of IDN literature that was performed to get a better understanding of what feedback items would be useful to authors. This is done by focusing on the most emphasised concern of authoring - the user's experience of the authored content. The review identifies 47 concrete aspects of user experience that are of interest to IDN authors by looking at how user experience has been talked about and evaluated in IDN literature.

Chapter 4 extends this work by mapping the identified UX dimensions to fields of NLP research and discussing how they might be estimated automatically. The results of this review indicate that automatically generated summaries are a promising form of feedback in terms of usefulness to authors and feasibility of implementation since it gives insight into several UX dimensions and is well studied within the NLP community.

Chapter 5 describes the creation of the IDN-Sum dataset, which is generated from fan-made transcripts of two narrative games and includes abstractive summaries for the overall interactive narratives and automatically generated extractive summaries for multiple interactive narrative playthroughs. The chapter also presents a baseline evaluation of standard summarisation approaches on the IDN-Sum dataset and a qualitative analysis of the summaries generated by these approaches.

Chapter 6 proposes a new method for extractive summarization that incorporates information about the narrative structure of the text by using self-supervised annotations regarding the proximity of words and sentences to choice points. The method is based on rationale-based learning where choice-based rationales are used to guide the learning process of the model. Results comparing models trained with and without explanations suggest that incorporating the choice-based rationale improves the extractive summarization of interactive narratives.

Chapter 7 summarises the key findings, and discusses the impact, limitations, and future directions for this research.

Chapter 2

Background

This chapter summarises the background literature that forms the context and motivation for this work. It will do this work first from an IDN perspective, then from an NLP perspective and finally from a Human-AI Collaboration perspective.

2.1 Interactive Digital Narratives

2.1.1 Definition of Terms

"Interactive digital narrative" is a general term encompassing a wide variety of digital narrative experiences that allow active participation from the audience. A simple example of interactive narratives are books such as the *Choose Your Own Adventure* books where the reader is given choices at different points in the story and they're prompted to turn to a different page depending on the choice they want to take. In digital format, there is a lot more variety and interactive stories can be found in the form of parser-based adventure games like *Zork*, hypertext fiction like *Afternoon, A Story*, and story-rich video games like *Life is Strange* and *Witcher 3*. There have been many attempts to pin down a definition for Interactive Digital Narratives since it encompasses a wide variety of formats (What counts as in "interaction"? What is a "narrative"?)[100]. Without going into the nuances of these terms, in this thesis, the term is used somewhat generally. However, the approaches described in this work are most relevant for narratives that have some level of non-linearity (caused by player interactions or otherwise) that makes the story space more cumbersome to envision during authoring.

In this thesis, *story space* refers to the set of all possible stories through an IDN. This encompasses all possible trajectories that a player can take through the game. This is also referred to as the Protostory[120]. Each trajectory is referred to as a *branch* or a

playthrough. In Koenitz's SPP model[121], the story space corresponds to the system and the playthrough corresponds to the product.

The person or persons involved in the creation of the interactive narrative is referred to as the *author*, *creator*, *designer* or *writer* and the person playing the interactive narrative is referred to as the *audience*, *player*, *reader* or *user*.

This work proposes simulating different playthroughs through an IDN that is being authored, and then using NLP to analyse the playthroughs to automatically generate feedback for the author. This feedback is referred to as *Authoring Feedback* in this thesis. Such feedback has also been referred to in IDN literature using terms like Intelligent Narrative Feedback[217] and Narrative Analytics[156].

2.1.2 Types of Interactive Narratives

Interactive digital narratives encompass a wide variety of digital experiences. They include many different mechanisms for enabling interactivity. In hypertext fiction, a non-linear narrative is represented as web pages containing sections of the story that are connected through hyperlinks[1]. Emergent Narratives consist of narratives that emerge from agent interactions in a simulated world[201]. Some interactive narratives have specialised narrative engines or experience managers that support the narrative mechanics in the game and ensure that a coherent and interesting interactive narrative is presented to the player[229].

IDNs have been categorised in many ways - for example, based on the medium of delivery (text, video, VR, multimodal), the level of interactivity and narrativity [20] or based on the design patterns used in them [157].

The study described in Chapter 3 covers a wide range of interactive narrative types but is limited to interactive narratives that are non-linear and have a significant narrative component. An IDN is considered predominantly narrative if it prioritises narrative goals over other goals, following the framework described in [31]. Narrative goals refer to the IDN's goal of communicating a narrative to the user as opposed to system goals, which refer to any other goal that the system may have. For example, edutainment is an example of IDN applications that prioritize system goals (teaching a skill or subject) over narrative goals.

Chapters 5 and 6 focus on text transcripts of two popular narrative games: *Life is Strange: Before the Storm* by Square Enix and *Wolf Among Us* by TellTale Games. These narrative games have a primarily gauntlet-like narrative structure. This design pattern refers to interactive narratives that have a mostly linear plot line with some variations and deviations[157]. This choice was driven by the availability of resources online. While the general approach described in these chapters can be adapted to work on

interactive narratives with different design patterns, the results of these experiments have only been empirically evaluated on this type of interactive narrative. Therefore, the results and discussion of these chapters should be interpreted as applying to IDNs with similar design patterns with decreasing confidence in generalisability to more different IDNs. While this research does not cover all the different types of IDNs, it takes a reasonable step forward in mapping out, making resources available and starting the investigation of NLP techniques for this domain.

This thesis focuses on interactive narratives represented as text since it investigates NLP techniques. These could apply to purely text-based interactive narratives or other types of interactive narratives that are represented as text (for example, at the planning stage of other types of interactive narrative where a lot of the narrative is represented as text, or game logs or transcripts that capture player trajectories as text).

2.1.3 The Role of Player Choices in IDN

There are four ways in which interactive narratives progress the narrative - through player choices (where a player makes a choice that drives the story forward), through scripted scenes (where the narrative progresses without player input, Discovery (where the player needs to locate some story content) and In-Game systems (like combat or tasks that the player needs to complete to drive the story forward)[39]. Out of these, player choices are the narrative mechanic that often leads to complex non-linear structures that are hard to manage. They are also a commonly used mechanic in interactive narratives, and are a unique affordance of the medium[243].

Choice points refer to points in the narrative where the player makes decisions that influence either the *fabula* (the raw sequence of events) or the *syuzhet* (the way these events are presented to the audience) [137]. The conception of a choice point can vary depending on the level at which the choice is happening. For example, in Michael Joyce's *Afternoon, a story*, the text itself remains constant, but the order and implicit relationships between narrative elements change based on the reader's navigation, affecting the *syuzhet* rather than the *fabula*. In contrast, the examples used in this work mainly contain dialogue choices that directly impact the narrative events, thus altering the *fabula*.

This motivates the experiments in Chapter 6 which experiments with adapting an existing summarisation technique to place special emphasis on text around choice points when generating summaries.

2.1.4 Authoring Interactive Digital Narratives

2.1.4.1 The Author

Murray[164] notes how the *cyberbard* or the IDN author is often not one person, but a group of people. In practice, this may include several roles including writers, narrative designers, game designers, narrative directors, creative directors and graphic designers. Creating IDNs also often involves designing and developing narrative engines which may include AI actors (as NPCs[1], story sifters[126], Drama Managers[229]). NLP-generated feedback discussed in this thesis mostly concerns the roles acting as the architect of the interactive narrative — this could be the narrative director, narrative designer and/or the writer. However, this could also benefit directors or cross-disciplinary teams where an overview of the narrative or certain aspects of it are desired.

2.1.4.2 The Author's Goals

To understand how IDN authors could be supported in their creative process, we must first look at what goals and concerns an IDN author might have during the process of authoring. This could vary from author to author, but some such goals that have been described in IDN literature are listed below:

1. **Ensuring a good experience for the player** or creating a certain effect in them is commonly emphasised as the author's primary goal [164]. The importance of user experience is also reflected in how IDN creators often use user experience evaluation as a measure of the IDN's success [229].
2. **Expression and communication of authorial intent** - Authors are usually intrinsically motivated by wanting to express an idea or vision that is specific and wish to have enough control to mould their creation in some specific way or express a specific authorial intent [217]. Successful communication of the intended message has also been described as factors that determine the success of an IDN[31].
3. **Advancing IDN as a medium** IDN is a medium that offers unique affordances. Exploring and maximising the use of these affordances so that IDN matures as an art form has also been described as an authoring goal [164].

Since user experience has been described as the primary concern for authors across the literature [164], this work explores supporting IDN authors by using AI to help them

[1] https://charisma.ai

better understand how a player might experience their story. Authoring feedback in the form of automatically generated summaries of possible playthroughs (which the second part of this work focuses on), could also help the author keep better track of the story space facilitating better expression and communication. It could also reveal experience through story paths that are afforded by the narrative engine but the author had not expected to be taken by the player. Armed with greater authorial control and understanding of user experience, the hope is that, this also takes a step towards empowering authors to explore and more fully take advantage of the unique affordances of IDN.

2.1.4.3 The Authoring Process

IDN authoring goes beyond the act of writing and extends to designing the overall experience including the rules of the story world and the interaction. For this reason, they are often described as procedural authors [164] and experience designers [117]. *"Bringing an Interactive Narrative into Existence"* involves a number of steps and often, a number of people[117]. Combining insights from several studies on the topic, Kitromili et.al [117] proposes an iterative IDN authoring process model involving four main stages - Ideation, pre-production, production and post-production. This model is summarised below:

1. **The ideation stage** is when the concept of the IDN is developed. This includes ideas on the narrative space, plot lines, and how the player will access it (types of interactions). This stage may also involve any training the author might need to take in order to create the IDN (for example, to familiarise themselves with the authoring tools or software) as well as some initial planning such as sketching possible plot lines and characters.
2. **The pre-production stage** involves the development of early prototypes. This includes the development of the storylines as well as the interactions - how the player will interact with and affect the story. This stage may also include visually or graphically structuring the story using different views provided by an authoring tool including how the narrative changes in response to interaction, mapping out the relationships between characters or events or dividing the experience into chapters or scenes.
3. **The production stage** involves the development of the complete product including all the content and assets that the IDN requires (this may consist of text, images, videos, 3D spaces, and so on depending on the type of IDN) as well as editing and compiling to check that the IDN runs without errors.
4. **The post-production stage** includes additional testing (for example playtesting) and debugging and finally, packaging and publishing the completed IDN.

The IDN author often goes back and forth through these phases, repeating some of these steps many times and in different orders. For example, in the pre-production stage, when trying to plan out the interactions and storylines, the author may discover that the idea is not feasible to implement and go back to the ideation stage. Alternatively, errors or shortcomings can also be discovered in the post-production stage during playtesting which might require going back to the production or pre-production stage to plan out and restructure the intertwining storylines and interactions more carefully. This can be costly, but the complex nature of IDN makes it hard for the author to recognise these issues without playtesting. This is because IDN authoring, which mainly consists of authoring the story space is somewhat dissociated from the possible instantiations of that IDN that the players would experience as a result of playing them. Koenitz's SPP model[121] frames an IDN *system* as the set of all possible stories, and one possible instantiation as a *product*.

Using automatically generated feedback has been proposed as a way to help the author identify and address such issues earlier on and iterate on ideas faster[218, 156, 217] and this is the line of inquiry in this research.

2.1.4.4 The Authoring Problem

Authoring interactive narratives is challenging for a number of reasons. In a branching narrative, the author has to write exponentially more content as the story gets longer. This is one of the main problems faced by IDN authors and is referred to as combinatorial explosion[30] or the **Authoring Wall** [85]. Another challenge is that each additional piece of content may make the IDN more complex and harder to manage. The effort involved in adding a new piece of content is referred to as a high **Complexity Ceiling** by Garbe [85].

The framework proposed by Garbe[85] also breaks down authoring challenges into concerns that pertain to:

- the mechanics of writing - this includes the complexity of the format the author would need to write in and the number of different components that the author would have to manage to produce a single unit of content and,
- the conceptual art of writing (which includes **clarity** or the complexity of state/system dynamics that the author would have to mentally track and **controllabililty** or the ability to test how and when units of content get presented to the player).

Other frameworks[209] categorise authoring issues as

- arising because of story ideas that don't align with the underlying engine's approach
- arising because it is hard to deliberate the user's experiences and,
- related to the authoring process being painful.

Recent advances in generative AI have inspired a large amount of research in automatically generating creative content that can be applied to address the Authoring Wall [82]. This PhD research addresses the problem of a high Complexity Ceiling (due to low clarity and controllability) and the problem of deliberating user experiences by using NLP to generate intelligent feedback giving insight into possible player experiences through the IDN.

IDN authors take different approaches to tackle the authoring problem. Jones[107] lists five categories that such strategies fall under - reducing and reusing content using clever design patterns, decoupling units of narrative to avoid explosion, automatically generating the content, for example, using NLP or simulations and finally, embracing the complexity through employing more resources or better tools.

Strategies used to reduce, reuse and decouple narrative units tend to place constraints on the type of interactive narratives that can be created while also raising the complexity ceiling. For example, one common way to reduce content is to use design patterns like the bottleneck where different narrative paths converge at an important plot point. Like the example shown in Figure 1.1, this means that the player could have arrived at a node through many different paths, making it hard for the author to envision the player's perspective.

Intelligent feedback generated for the author using NLP (for example, a question answering (QA) interface that allows the author to ask questions about the different paths through which the player could have arrived there) can allow the author to maintain context easier. In the case of generated content, intelligent narrative feedback can help the author ensure that the generated content is aligned with authorial intent. By reducing many of the constraints created by high complexity ceilings, the hope is that better authoring tools that provide intelligent narrative feedback will empower IDN authors to both more efficiently implement these strategies to reduce, reuse, decouple and generate content and also better embrace the complexities of IDN.

2.1.4.5 Authoring Tools

Generalized game design tools like Unity[2] and Unreal[3] do not offer many features to manage narrative content. However, [166] notes that visual programming languages

[2]https://unity.com
[3]https://www.unrealengine.com

like Unity's Visual Scripting system and Unreal Engine's Blueprints have similar concerns as interactive narrative authoring systems and may eventually be able to support narrative-oriented features.

There are also many IDN authoring tools ranging from research prototypes to commercial applications[204] that share the goal of helping authors manage the complexity of writing non-linear stories[166]. A review of IDN authoring tools by Shibolet et. al[204] proposes categories and descriptors and classifies over 300 such authoring tools. The most commonly used authoring tools and frameworks include Twine, Bitsy and Ink [64].

Some authoring tools support visualization of the underlying structure[89] to help authors manage the complexity of the interactive narrative. Some of them are also specifically designed with the intention of helping the author have both high visibility and generativity [85]. [166] gives an overview of how popular authoring tools support authors in visualising and structuring the narrative space through their visual aids and graphical interfaces. They refer to this process of structuring the narrative and how they respond to interaction as "mapping" and find four ways in which these authoring tools visualise the narrative space - spatial mapping, scene-driven mapping, nodal mapping and traversal mapping. While these types of visual aids are vital to the process of creating IDNs, they focus on illustrating the low-level structure of how units of content (or lexia) make up the narrative space and how they transition and respond to player interaction. Understanding and interpreting the semantics of the content contained within each lexia - for example, the events, characters, the relationships and dependencies between them as well as imagining how they translate to player experience is still left to the author. As the size and complexity of the IDN increases these visualisations become hard to fully comprehend and reason over in this way[166].

Tools also often have a way of letting the author play through the narrative like a player would[166]. While playing through the narrative helps the author gain some insight into how a player might experience the narrative, they can only test a few different playthroughs manually.

This work investigates the idea of simulating many different playthroughs through the interactive narrative and then using NLP to analyse the playthroughs automatically to get more of those insights about aspects of the player's experience of the narrative that the authors might check for when they manually play through it themselves or playtest the games. A system like this involves many components - an agent that simulates player traversals or playthroughs of the game, NLP techniques to analyse the playthroughs and an authoring tool with an intuitive interface that has these systems integrated to provide feedback to IDN authors. All three components have their own challenges but this research focuses specifically on the second

component - investigating how NLP techniques can be applied and adapted for this domain and use case. We use random traversals of IDN games as playthroughs in this study and leave integrating this into an authoring tool alongside more sophisticated agents that can approximate player interaction[156] for future work. In this way, this work takes a step towards increasing the visibility of the *semantic* aspects of narrative space in a way that is player-centric and scales with complexity.

2.1.4.6 Authoring Feedback

Several papers [218, 156][217] draw attention to the issue of lack of visibility of the story space and propose automatically generated feedback as a way to help with this. They also give some high-level categories of feedback items with examples to illustrate how such feedback can be useful. While these serve as good starting points, feedback discussed in [217] is only in the context of emergent narratives and [156] is more focused on detecting specific problems rather than increasing overall visibility. Other work also discusses automatic structural analysis using graph theory[174] and low-level computational metrics like frequency and diversity of choices[221]. The interest in this thesis is instead in higher-level, more intelligent insights that are generated from analysing the semantic aspects of the content.

Efforts to encode various semantic aspects of the narrative and player experience rather than the structure include Story Intention Graphs [147] that present an annotation schema to record underlying facts about the story and story world (or fabula), Interactive Cinematic Experience (ICE) schemata [165] which was used to align data containing player responses (such as video recordings and heart rate) to the content of the game (such as feature locations and choice points) and Progression Maps [40] which proposes and evaluates a framework for visualising the interaction design. Using NLP techniques offers a variety of ways to automatically analyse potential playthroughs of the interactive narrative and generate insights about narrative and semantic aspects of the story space including emotional arcs[182] and sentiment networks[132], but this has not been investigated to much depth.

The inability to easily visualize the user's experience of written interactive work has been identified as a problem that is faced by IDN authors. Using automatically generated intelligent feedback has been proposed as a way around this, however, there is a gap in existing literature here since UX is a very broad concept and it is not clear what exactly this feedback needs to be. Chapter 3 addresses this through a systematic review of IDN literature to identify what aspects of the user's experience would be useful to IDN authors if provided as automatically generated feedback. Moreover, NLP techniques for generating intelligent, intuitive feedback over freely written IDN text have not been implemented and experimented with. This is what Chapter 4 addresses. By mapping the UX dimensions identified in Chapter 3 to NLP

AI Role	number of papers	References
Generating Content	6	[8, 45, 62, 79, 197, 63]
Game Playing AI	5	[112, 28, 113, 69, 146]
Player Modelling	4	[239, 96, 134, 88]
Co-Creation	3	[142, 136, 125]
Experience Managers	3	[212, 160, 127]
Automated Game Design	2	[205, 61]
AI Characters	2	[14, 108]
Narrative Planning	2	[196, 194]
Feedback AI	1	[78]

TABLE 2.1: AI Roles in IDN at AIIDE 2022

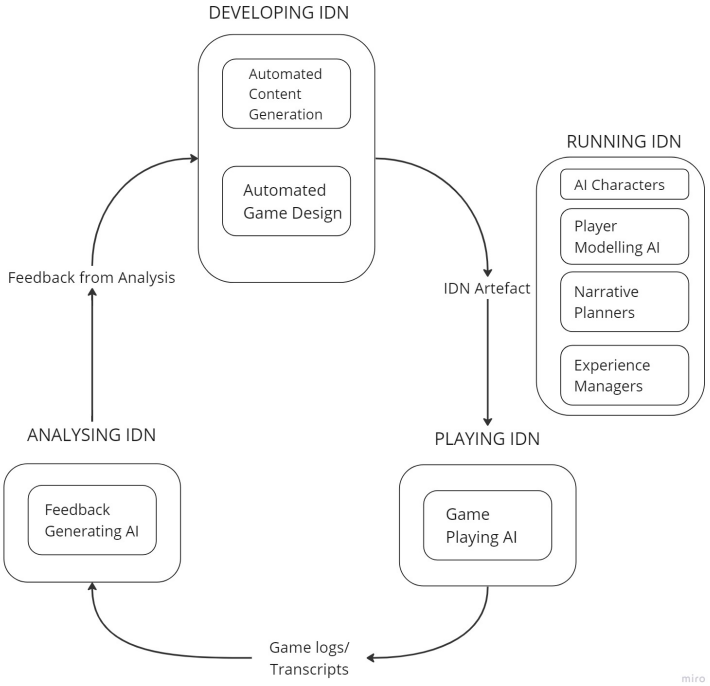

FIGURE 2.1: Different stages of iterative IDN development and how AI can be applied (AI roles) in each stage. The focus of this research is on Feedback Generating AI.

tasks, types of authoring feedback that would be both impactful to IDN authors and feasible to generate using NLP techniques are identified. Feedback in the form of automatically generated summaries of playthroughs was selected for further investigation in Chapters 5 and 6.

2.1.5 AI in Games and IDN

AI has been applied in Games and Interactive Narratives in a variety of roles from creating and running them to playing them. Table 2.1 shows how papers involving AI in different roles were distributed at AIIDE in 2022. A large body of work is focussed

on Automated Content Creation for Games and Interactive Narratives[8]. There has also been some work on Automated Game Design [61] and interactive narrative generation[82]. Some papers also address aspects of human - AI interaction in co-creation[125]. AI has been used in Narrative and Game Engines for AI-driven elements in the story world (eg NPCs) and generating believable simulations as well as Automated Experience Management where AI is used to direct and manipulate game events to ensure an enjoyable and dramatic experience and Player Modelling where AI is used to understand player intention and preferences to tailor the experience for them. Several papers also address Game Playing AI either using games as a test bed for reinforcement learning approaches or for automated playtesting [69]. Most similar to this work, a small body of work applies AI for Automated Analysis to gain insight into the story space [78]. Such work can be applied to generate feedback to support its creation. These approaches with AI in different roles complement each other. Figure 2.1 shows an example how these roles might fit into an IDN development lifecycle. Feedback-generating AI approaches such as those described in this thesis may be applied in such a workflow with either AI agents or human actors assuming the other roles where possible.

2.2 Natural Language Processing

Natural Language Processing involves computational processing of natural language - or language commonly used for human communication (e.g. English, Hindi) as opposed to machine languages (eg - Python, assembly code) or logical expressions (eg - mathematical formulas and proofs). Some over-arching NLP problems include Machine Translation (using computational methods for automatically translating from one language to another), Text Classification (classification of text data into pre-defined classes), Information Extraction (extracting structured information from unstructured text data), Machine Reading Comprehension (getting the model to "understand" language and then perform a variety of tasks like question answering and inference) and Automatic Text Summarisation (or producing a shorter version of the text that retains the most important information). Chapter 4 reviews NLP literature and maps out ways in which different types of NLP problems can be applied to generate authoring feedback that can help authors envision different aspects of player experience. Chapters 5 and 6 then explore one of these types of NLP-generated feedback further (Automatic Text Summarisation (ATS) of IDN). Section 4.5.3 in this chapter gives an overview of ATS research.

Natural Language data covers a variety of domains like news, social media and conversation text, media including narratives and fiction, legal, financial, medical and academic documentation. While the recent large language models seem to show good zero-shot performance across domains and use cases, the generalisability of NLP

approaches across domains and tasks is not obvious since the patterns embedded in the text vary across domains and different language-based tasks have varying levels of complexity. NLP approaches for IDN text have not been explored in much depth previously and this research takes steps in this direction. Section 2.2.1 gives an overview of how NLP has been applied to traditional narratives (which is a closely related domain) as well as how it has been used in the context of assisting in authoring narratives.

AI approaches used for Natural Language Processing have evolved from rule-based approaches rooted in theoretical knowledge derived from linguistics and domain knowledge (eg- using regular expressions and context-free grammar) to more powerful deep learning-based methods. Common deep learning methods used for NLP have evolved from word embeddings and seq2seq models [216] to attention-based models (2015)[15], transformers (2017)[235] to language models(2018)[70] which have been scaled to larger and larger models with increasingly impressive capabilities in the recent years demonstrating near-human performance in many NLP tasks[33]. Section 2.2.3.2 gives a brief overview of NLP approaches for text summarisation that are used in this research and section 2.2.4 gives an overview of Rationale Based Learning - the technique this research adopts to adapt a classic summarisation approach to the IDN domain by training the model to give special attention to the text around the parts of the IDN that involve player interaction.

2.2.1 NLP and Narratives

NLP has been applied in many ways to narrative text including novels and movie scripts. There has been some effort to define formal and computational models of different aspects of narratives including suspense, pacing, causality and characters [151, 231]. There have also been efforts to come up with a standard annotation scheme for narratives[151, 179, 138]. Work on narrative information extraction tries to extract structured information including automatically identifying characters[230], relationships between them[132], their evolution[240] and summarisation and visualization of stories[234]. Narrative understanding techniques attempt to reason and answer questions about the story[26]. Automatic literary analysis includes a variety of studies including plotting emotional arcs of narratives and validating narrative theories[182]. An exploratory search for NLP techniques applied to narrative text brings up work including emotion detection[182], sentiment networks[132], plot summarization[49], tuning point identification[172] and identifying temporal relationships[44]. NLP techniques are applied to different kinds of narratives including Novels[49], short stories[244], personal narratives[208], common sense stories[44] and movie scripts[172].

The many different ways of applying NLP to narratives suggest many different forms of authoring feedback that could be automatically generated. To understand which forms of feedback would be most impactful if generated automatically, Chapter 4 reviews NLP literature and maps NLP tasks and applications to aspects of UX that were identified as useful to IDN authors through the review in Chapter 3. Work exploring NLP on narrative text suggests that such approaches would work similarly on IDN as well but this has not been investigated and IDN has several differences from traditional narratives caused by interactivity. In the context of interactive narratives, NLP has previously been applied for the procedural generation of interactive narratives using language models [214] extracting structured information for automatic world-building for interactive fiction games using an NLP-driven Question Answering (QA) system[6].

2.2.2 NLP-Assisted Authoring

NLP has previously been applied to assist IDN authoring. Mimisbrunner[211] is an IDN authoring tool which uses AI to support authors with automatically generated suggestions and allows them to author in restricted Natural Language. On the other hand, this thesis investigates NLP techniques to better envision the story space.

Cardinal[152] assists the authoring of movie scripts by allowing writers to visualise the script from different perspectives, including the interaction between characters and 2d and 3d previews. LISA[195] is an authoring tool which tries to automatically detect logical flaws in authored content. It also includes a QA system that allows the author to interact with the story world as its being created allowing them to maintain context. However, they use a knowledge-based approach that requires extensive authoring of "rules" that the system will check against. Similar to these approaches, this thesis attempts to investigate NLP approaches to help the author manage the complexity of the authored artefact and better envision the end product, but for IDNs.

Many of these approaches also require that the authoring be done in specific ways (for example the rules in LISA and restricted natural language in Mimisbrunner). This thesis investigates NLP techniques for analysing the story space agnostic of how the IDN was authored. The methods introduced in Chapters 5 and 6, however, do require that the playthroughs for the IDN can be generated for analysis as text in the form of transcripts or game logs and experiments in 6 require that the choice points are marked as such in the transcripts.

2.2.3 Automatic Text Summarisation

After identifying feasible and impactful forms of feedback in Chapters 3 and 4, Chapters 5 and 6 investigate one of these forms of feedback - Automatically Generated Summaries, more deeply. Automatic Summarisation was chosen for further investigation since it mapped to many important UX dimensions and at the same time, it is well-researched within the NLP community. This section gives an overview of ATS, motivates experiments in Chapters 5 and 6 and contextualises the approaches used.

2.2.3.1 Types of Summarisation

Many different classification frameworks have been used to talk about different types of automatic text summarisation[248]. Some relevant categorisations are discussed in this section.

Based on the framing of the problem and nature of the output, automatic text summarisation approaches can be classified in the following ways:

1. **Extractive or Abstractive Summarisation** Extractive summarisation refers to extracting sentences or "extracts" from the original text[74]. Abstractive summarisation refers to generating a summary that succinctly captures important information using paraphrased sentences. Either type of summarisation may be used as authoring feedback[74]. However, in case of abstractive summarisation, the information contained in the summaries may or may not be reliable. In case of extractive summarisation, since extracts from the text are directly used in the summaries, the reliability of information is less in question. We focus on extractive summarisation in this study since it is more transparent and is less likely to give a false sense of security to the authors even if it performs poorly.

2. **Generic or Query-based Summarisation** Generic summarisation typically tries to include salient content in the summary and minimise redundancies. Query-based summarisation also considers relevance with respect to a query when deciding whether to include the information in the final summary[101]. While query-based summaries could also be useful as authoring feedback, it is a less mature field than generic summarisation. Therefore, this PhD focuses on generic summarisation approaches as a first step in Chapter 5 and then explores increasing emphasis on the text surrounding interaction points in Chapter 6. Extending this work to query-focused summarisation for specific queries is left for future work.

3. **Indicative or Informative Summarisation** [248] classifies narrative summaries as indicative or informative. Indicative summaries try to indicate what the content is about without revealing all the important information in its content. In the case of narratives, this could mean summaries without spoilers. Informative summaries try to capture the most important content in the original text. This categorisation is closely related to the intention and context in which the summaries will be used. In the context of authoring feedback, ideally, we want the summaries to act as a recap of the different ways in which a player can traverse the story space, with an emphasis on player choices, consequences and how player experience varies across playthroughs. This is not exactly the same as either the indicative or informative summary and is instead a more nuanced use case. Implications of this are discussed further in Chapter 6.

4. **Domain Dependant and Domain-Independent Summarisation** Domain independent summarisation approaches place no intrinsic restrictions on the type or domain of text it is summarising, whereas, domain dependant approaches specialise the system for a domain by incorporating knowledge about the domain into the summarisation approach[248]. Since IDN is a domain that has not been explored before for summarisation, in Chapter 5, the performance of domain-independent approaches is first evaluated to determine baselines. In Chapter 6, domain-specific information is then introduced through rationale-based learning to improve summarisation.

Based on the nature of the input, they can also be classified as follows:

1. **Short and Long Document Summarisation** The size of a document considered "long" has varied across NLP history [123]. For example, previous work[43] considered news dataset, CNN/DailyMail(CNN/DM) long, but in current literature, these are considered short. As of 2021, long documents are commonly considered to be 2000 tokens or higher [19, 123]. This is because state-of-the-art summarisation models, for example, those based on transformer-based language models, commonly have a token limit of 512 - 1024 tokens[260, 247]. However, the interactive narrative dataset used in this thesis is much longer (approx 23K tokens).

2. **Single and Multidocument Summarisation** Multidocument summarisation refers to summarising multiple documents, for example, multiple accounts of the event in news[9]. Different playthroughs of an interactive narrative can be seen as multiple documents pertaining to the same IDN to be summarised. Additionally, comparative summarisation[102] refers to summarising similarities and differences between documents which would be especially useful in the context of authoring feedback. However, most existing supervised

approaches for multi-document summarisation require datasets with a number of different document sets and summaries. Due to the lack of such resources for IDN Summarisation, this thesis focuses on single-document summarisation as a first step, considering each playthrough as a single document and leaving multidocument summarisation for future work.

2.2.3.2 Standard Approaches for Extractive Summarisation

Methods: Early approaches used for text summarisation included methods like:

- Statistical methods where the importance of sentences was determined based on statistic and linguistic features like frequency and positioning of words and sentences[74].
- Topic-based methods where the main topics in the documents are identified and then sentences are selected for inclusion in the summary based on relevance to those topics[74].
- Optimization-based methods where the summarisation problem is cast as an optimisation problem (for example, through constraints such as reducing redundancy and maximising coverage) [74].
- Graph-based approaches where relationships between different parts of the text are modelled as graphs before applying graph theory and heuristics to select important sentences[74].

In recent years, supervised deep learning approaches are increasingly used for text summarisation. Some common deep learning model architectures applied for extractive summarisation include:

- Recurrent Neural Networks (RNNs): Many of the initial deep learning models used were based on RNNs which are a type of neural network where information from model states in previous time steps loops back into the network, allowing it to model sequential data[202]. Variations of this architecture like Long Short-Term Memory (LSTMs) [86] and Gated Recurrent Units (GRUs)[54] were introduced for longer sequences. These are used to generate internal representations of words, sentences or overall documents. These representations can then be classified as to whether or not they belong in the extractive summary[169, 262]. For example, SummaRuNNer used word-level and sentence-level GRUs along with fully connected layers to capture salience, position, nature of content and novelty of content to classify sentences.

- Transformers and Pre-trained Language Models: Transformers are a type of deep learning model architecture that has been influential in the field of NLP[235]. They are built on the idea of "Attention". Attention mechanisms allow the model to focus on some parts of the sequential input over others, similar to how humans pay attention to some parts of the text more than others[15]. The transformer uses a type of attention called "Self Attention" which allows the model to compare each element of the input sequence to every other element in the sequence, determining how much attention each word should receive when creating a representation for the other. Pre-trained models are models (often built using transformers) which are initially trained on large amounts of data[70]. They can then be fine-tuned on specific tasks like text summarisation. An example of this is BertSum [145] which fine-tunes a pre-trained language model called BERT for text summarisation. Similar to RNN-based approaches, these view extractive text summaristaion as a sequence classification task where internal representations are computed using the transformer based language model instead of the RNNs, allowing them to better capture relationships between words and sentences[145, 19].
- Graph Neural Networks: In graph neural networks, extractive text summarisation is cast as a node classification problem. The graph is constructed based on concepts like semantic similarity[9] and discourse relationships [247]. GraphTP explores the use of GCNs for screenplay summaisation [135].

These models are often used in combination with other approaches including formulating extractive summarisation as text matching[260] and reinforcement learning[50].

SummaRuNNer (an RNN based method) and BertSum[145] and a variation of it modified for longer sequences, LongFormer[19] (which are language model-based approaches) are included in the baseline evaluation in Chapter 5. TextRank, which is a graph-based unsupervised approach, is included also in the baseline evaluations. Future work exploring evaluation of graph-based neural networks for IDN is discussed in Chapter 7. Commonly used approaches for abstractive summarisation include models based on BART[237] and PEGASUS[255]. Recent advancements in abstractive summarization with large language models (LLMs) like GPT-4[33] and Llama-3[73] have demonstrated impressive performance, especially in zero-shot settings, often outperforming fine-tuned models[12]. However, challenges like position bias, where models disproportionately focus on certain sections of input text, persist [53]. Reinforcement learning techniques are being explored to improve summary quality. Emerging approaches include using LLMs as reference models to enhance smaller models and testing model robustness through paraphrasing.

Evaluation Strategies: The standard evaluation metric used to compare model output to reference summaries is the ROUGE score[140] which is based on keyword overlap between the two. This metric has the limitation that it does not account for paraphrasing. Other evaluation metrics like BertScore measure the semantic overlap to some extent as well, however, these could be less intuitive to interpret[256]. Manual evaluation involves using metrics like readability, coherence, conciseness, coverage, clarity, grammaticality and non-redundancy to measure overall quality or performing a task-based evaluation to judge the utility of the summaries to the given use case (eg for information retrieval)[74]. However, both automatic and manual evaluation strategies have several challenges ranging from the effectiveness of the automatic metrics[76] to choosing gold standards to use as a reference since many valid summaries can be created for the same text. In this research, quantitative evaluations using the ROUGE metric are performed along with qualitative analyses to assess more subjective aspects of summary quality. Further discussion regarding the challenges of evaluation of text summarisation, recent research in this area and how they relate to findings in this work is discussed in the Conclusion in Chapter 7.

Oracle Summaries: Most summarisation datasets consist of the original text and an associated human-written abstractive summary. Therefore, extractive summarisation approaches usually include a step where an extractive summary is first created from the abstractive summary so that supervised training strategies for text classification can be applied. This typically involves aligning parts of the original text to sentences in the abstractive summary based on a similarity metric like the ROUGE score. For example, the method introduced by Nallapathi et.al [169] involves greedily selecting extracts from the original text until the ROUGE score between the abstractive summary and the aligned extractive summary can not be improved anymore. The extractive summaries generated this way are then used as labels for training. These summaries are referred to as *automatically aligned extractive summaries* or *oracle summaries* in this thesis. While alternatives to this way of generating better alignments have been explored[133], commonly used extractive summarisation baselines like SummaRuNNer and BertSum[145] use this method to create oracle extractive summaries for training. Therefore, this research uses this greedy alignment approach and leaves the exploration of alternative alignment methods for future work.

2.2.3.3 Narrativity and summarisation

Most text summarization work is targeted at news, academic papers and reviews. However, there is some work on novel and movie summarization. This section gives an overview of work in this space. Details related to relevant and available datasets used in these papers and available from other sources online as of 2020 are summarised in table 2.2.

Name	Content	Size
ScriptBase[87]	Movie scripts and 3 summaries each	1276
CMU Movie Summary Corpus[17]	movie plot summaries and metadata	42,306
Film corpus 2.0 [141]	movie scripts	1068
The Movie Corpus[4]	movie scripts with imdb links	25,000
Novel Chapters[133]	Novel chapters and summaries	8088
The Schmoop Corpus[49]	Novel chapters and summaries	7231
Project Gutenberg[5]	large corpus of ebooks	60000+
CMU Book Dataset[16]	plot summaries	16,559
The TV Corpus[6]	TV episode scripts, IMDb links	75,000
WikiPlots[7]	all types of plots	112,936
Telegraphic summaries[150]	telegraphic summaries of short stories	200
NarrativeQA [119]	summaries, links to full stories	1572

TABLE 2.2: Datasets for Narrative Summarisation, gathered in August 2020

In the field of novel summarisation, [49] introduces a dataset for novel chapter summarization. It includes the results of training some models on it, evaluated on multiple choice abstractive summarization. Work focusing on summarisation approaches for narratives include summarising novels based on topic modelling[246] and extracting information from stories to generate character descriptions that form the introductory section of plot summaries[257]. [133] proposes an alignment method to get extractive summaries from reference abstractive summaries and finds that using a weighted version of the ROUGE metric for alignment gave better results. There has also been some investigation into the idea of extractive summaries that read like telegraphs (using smaller units for extraction instead of full sentences) since in the case of literary text, the relevant information is spread over many sentences[150]. Work on movie summarisation includes exploring summarisation techniques for movie subtitles [10], applying rule based approaches to identify salient scenes for movie summarization[228, 87], and incorporating information as to whether a scene is a turning point(TP) or not into the summarisation model[173]. While there is some existing work on screenplay summarisation[135, 87], these are applied at the scene level in resulting summaries would be still pretty huge to use in the context of assisted authoring.

Redundancy and topic diversity are constraints that are commonly used to generate summaries across domains [246]. In [246], the only narrative-specific assumption used was selecting content according to the ratio 20:60:20 for the beginning, middle and end of the story. The narrative structure is made use of in [173] to identify turning points in the story and let that influence scoring. Narrative-specific features that are assumed to influence saliency according to these papers are, sentence position[173], presence of

lead characters[228], sentiment intensity[87], whether a scene is a turning point and similarity of scene embedding with TP sequence and global screenplay embedding[173]. While [49] and [133] tests out deep learning-based architectures on narrative datasets, only [173] and [135] propose model architectures that were designed specifically for narratives, but both these work at a scene-level rather than at sentence level.

2.2.3.4 Interactivity and summarisation

There has been some work on game log summarisation. In this domain, Bardic[18] generates narrative query-based reports (as text, map and machinima) from game logs for the online game, DOTA. They apply intention recognition and identify a fixed set of narrative tropes from action sequences (eg - Chase flight, failure, etc). There has also been work on extracting plan steps from custom game logs[52] where essential events are extracted based on causal relationships to story goals. [193] tries to generate summaries for sports games from commentary. However, the nature of text in all these logs is significantly different from what you would typically expect from interactive narratives since these logs do not contain much narrative. No papers talking about summarization of IDN work could be found but several summaries and transcripts are available online from sources like Fandom[8], IFDB[9] and Wikipedia[10]. They also have associated APIs which can be used to automatically scrape some data. For example, [101] introduces a way to collect query-specific summary dataset from Fandom. The most closely related dataset that could be found was the Critical Role Dataset[181] which contains transcripts of Critical Role episodes (this is a show where people play a tabletop RPG, Dungeons and Dragons) which could be considered to be a single playthrough of an interactive narrative.

2.2.4 Rationale Based Learning

Explainable AI deals with attempting to understand why AI models make certain decisions or making AI models that are interpretable. Some models are interpretable by design. However, in the case of neural networks, this involves getting the model to produce explanations that show why the model made those decisions. Model agnostic methods like SHAP[148] can be applied to generate explanations for predictions indicating how different features contribute to the model prediction. Example-based approaches use specific examples from the dataset to uncover insights about how the model works. Model-specific approaches involve using aspects of the model like

[8] www.fandom.com
[9] ifdb.tads.org
[10] www.wikipedia.org

attention scores and gradient saliency maps to better understand the model decisions[83].

Recently there has been a growing interest in rationale-based learning, where human annotated explanations are collected and used for data augmentation. These are used for better performance on a predictive task or to train models to produce explanations for their outputs. Model-generated explanations can be evaluated against the collected human annotated explanations[83]. These explanations are also known as *rationales* leading to the term *rationale-based learning*. One way in which rationales are incorporated into training is through Supervised Attention[109]. Attention mechanisms were introduced into the field of deep learning in 2016[15]. Attention mechanisms allow networks to learn weights that determine how much emphasis is placed on different parts of the input sequence when producing the output. There has been some debate on whether attention weights (indicating which parts of the input the model paid the most attention to when producing the output) count as explanations[22]. However, explicitly training the model to focus on parts of text representing rationale using supervised attention (through minimising the loss between model attention and human annotated rationales) has been shown to yield positive results in text classification[109]. Chapter 6 extends this approach to interactive narrative summarisation using automatically generated rule-based annotations indicating the proximity of sentences and words to choice points in place of human explanations.

2.2.5 Discussion

There are many NLP applications that can extract information from narratives that could serve as potential feedback items. This motivates the work described in chapters 3 and 4 to identify the IDN author's requirements and opportunities afforded by NLP. Through these literature reviews, feedback in the form of automatically generated summaries is found to be a type of feedback that has the potential to be both impactful and feasible. However, Automatic Summarisation approaches for IDN have not been studied before and will be investigated as part of this PhD.

The most commonly used datasets to study and benchmark extractive summarisation approach include news datasets like CNN/DM[103]. IDN Summarisation is different from such domains in many ways, so experimentation is required to check if and how the difference in domain affects performance of NLP approaches. There also is not enough IDN data available for supervised learning, but some can be scraped for evaluation from online sources like Fandom. Fandom provides both transcripts and summaries of popular narrative games whose content structure resembles screenplays. This motivates creation of the first dataset for IDN summarisation in Chapter 5.

The differences in IDN compared to other domains include the presence of narrative and interactive elements, longer text, and higher overlap between data points. Some of these differences like the presence of narrative elements and long document summarisation have been studied in the context of traditional narratives like novels and movie scripts. While many insights from these works are relevant to IDN summarisation, there are no established benchmarks for which state-of-the-art can easily be tracked and compared. Automatic Text Summarisation is also a fast-moving field where newer approaches are constantly being proposed and evaluated. As of 2020, MatchSum[260] has been considered state-of-the-art of extractive summarisation [254], but the approach described in the paper involves using BertSum for an intermediate step which has a token limit of 512 tokens. Trying to determine the best combination of approaches for IDN summarisation would therefore require adapting, testing and comparing different combinations of these approaches. Rather than doing this, Chapter 5 experiments with a representative subset of standard or widely adopted summarisation approaches that form the foundation of most of these approaches.

After this, the suitability of a more specialised approach for interactivity through player choices (which is a novel aspect of this domain) is further explored in Chapter 6 by exploring rationale-based learning to increase emphasis on the text surrounding interaction points. The suitability of newer and more specialised approaches for other aspects of summarisation (including alignment techniques for generating oracle extractive summaries, better evaluation metrics and strategies, narrative-specific modifications and newer training strategies and model architectures) are discussed as part of future work in Chapter 7.

2.3 Human-AI Collaboration

The focus and core work of this thesis is primarily rooted in advancing NLP for IDN research. However, the NLP techniques are investigated in this thesis with the primary intention of supporting IDN authoring, so the motivation of this research and future directions has roots in themes from human - AI collaboration. While this thesis does not delve into designing and developing an authoring tool (or the complete Human-AI System), the NLP techniques it investigates are for systems that would involve Human-AI Collaboration and co-creation. This section will contextualize this work from this perspective.

2.3.1 Themes and Perspectives

Several related terms and concepts, sometimes with overlapping meanings, have been used to talk about systems involving both Human and AI input. This section will go over themes and perspectives most relevant to this thesis.

2.3.1.1 Human-Centered AI

[38] reviews papers that use the term Human centered AI and find that the term has been used to describe four types of "human-centeredness" - Humans teaming with AI, using human centered approaches to design and evaluate AI, Explainable or Interpretable AI and Ethical AI. Humans teaming with AI involved humans and AI working together with varying degrees of control for the AI. This is the type of Human-AI interaction envisioned in this research. Using human-centered values in design and evaluation approaches involved using principles from HCI to better incorporate and align AI to human needs. Since this thesis does not directly deal with designing an authoring tool, it does not delve into HCI approaches - this is, however, part of future work for this research. Explainable or Interpretable AI involves using AI techniques that give humans a better understanding of how and why the AI makes certain decisions. Chapter 5 reports performance of NLP approaches with varying degrees of transparency - TextRank, a highly transparent summarisation approach works the second best of out all the approaches tested in terms of ROUGE scores. Ethical AI involves research into concepts such as human-AI value alignment, fairness, bias, transparency and trust. Research in HCAI finds that explainability and transparency of AI approaches used impact trust. However, [38] notes that transparency does not always mean more trust - particularly in case where the system outputs are incorrect or the explanation is not in line with expectations. While these concepts are not addressed directly by this work, the choice of extractive summaries over abstractive summaries as the type of feedback for deeper investigation was motivated by the fact that extractive summarisation is more transparent than abstractive summaries[254]. The difference between these two types of summaries is discussed in section 2.2.3.1.

2.3.1.2 Human-AI Teaming

Human-AI Teaming refers to Human and AI agents working together to accomplish a task. This is a subset of Human Centered AI approaches discussed in the previous section which is most in line with this work. The framework described in [38] classifies approaches along a dimension of being AI-led to human-led. Generative AI or AI simulating humans is more AI-led commonly used in creative AI because low

stakes decisions and subjectivity. Human-in-the-loop approaches involves using human feedback to improve the AI over time. This adds some human influence, but is still, largely AI-led. Human-AI teaming often involves greater human control and tend to be human-led or collaborative. In line with human-AI teaming, this PhD tries to increase authorial control by increasing the visibility of the story space.

2.3.1.3 Augmented Intelligence

Augmented intelligence is another term used in this context that has some overlap with Human AI teaming. [177] describes three perspectives towards the role of Human and Artificial Intelligence - A techno-centric view which holds that AI will eventually be superior to human intelligence and will replace humans, a Human-centered view which holds that AI will always be lacking in some aspects and should only be used as tool, deployed only with human involvement and a collective intelligence perspective which hold that a hybrid intelligence is more powerful than both and intelligence should be studied at the level of Human-AI teams and this can be improved by improving interactions and workflows between them. In this framework, this project takes a view that is human-centered to collective for co-creation of IDN - it aims to give the IDN author more control and oversight - allowing more efficient Human-AI collaboration and communication.

2.3.1.4 Mixed-Initiative Approaches

Mixed initiative approaches[263] in AI involve systems that allow for both human and machine input and decision-making. This can include methods that allow for human oversight and intervention in autonomous decision-making, as well as systems that allow for collaboration between humans and machines to achieve a common goal. The goal of mixed-initiative approaches is to combine the strengths of both humans and machines to improve overall performance and decision-making. This is a similar concept to those described above, but the stress here is on the degree to which the Human and AI "take initiative" in terms of progressing the task. NLP methods discussed in this thesis can be used to generate feedback that takes the form of suggestions for improvement or pointing out potential issues - which could be interpreted as AI taking initiative. Feedback in the form of summaries, which was picked for deeper investigation, does not, however, make these types of suggestions. It simply serves as a recap for the author of the different ways in which the player could have arrived at a lexia - It is fully left to the author to take any initiative on what is to be done with the information.

2.3.1.5 Distribution of Agency

Human - AI collaboration approaches assign varying degrees of agency to human and AI actors[38]. A study of human-ai co-creation in a collaborative story writing setting found that writers desired more control in cases where they prioritised emotional values in implementing their ideas well more than productivity, and distrusted the AI to accomplish challenging subtasks and match their writing strategies [24]. While there has been some discussion on how the degree to which human control comes at the cost of the degree to which processes can be automated and generativity can be leveraged, there have also been some efforts towards decoupling these concepts and achieving high control and generativity [85, 38]. This research is an effort in this direction as well. Better visibility of the story space through authoring feedback can facilitate the use of sophisticated narrative engines and non-linearity, maybe even those that incorporate generative models without compromising on authorial control.

2.3.2 Creative AI and Co-Creation

In this thesis, the term creative AI is used in a broad sense - AI as used in creative processes. In creative industries, AI is increasingly being used in many ways including Content Creation (for example, generating images or animation), Content Enhancement (eg. deblurring or denoising images) and Analysis (eg. analysing sentiment of reviews, recommendation engines, assisting with research in the creative process and intelligent assistants) [7]. This work is in line with the latter application and explores AI tools for assisting in the creative process. There have been many varieties of Creativity Support Tools (CSTs). [58] reviews and categories 111 such tools. AI is used in different roles including Idea Generation, Curation, Execution Assistance, Production, Understanding the current state of the creation and Evaluating it. The majority of papers reviewed address idea generation or execution assistance. This work, on the other hand, investigates NLP methods that can be used for the last two cases (Understanding and Evaluating) for supporting IDN creation. They also differentiate between tools that Implement parts of the creative artefact from those that Influence the creator, which is the type of use case that this work addresses. NLP for authoring feedback has no direct input in the creative artefact, it just allows the author have more visibility and control. They are free to form and take creative decisions based on these insights. Mixed inititative approaches have been used in game development for challenge approximation [104] and automated playtesting [46]. This PhD investigates NLP techniques for similar use cases, but for narrative aspects of an IDN rather than the ludic aspects.

2.3.3 Discussion

It is important to note that this research chooses to focus on adapting NLP techniques for IDN creation rather than building an authoring tool with feedback generated using existing techniques. Designing, developing and maintaining such tools comes with its own challenges [249]. However, to enable the design and testing of such AI-powered tools and the effect they might have on IDN authoring, AI approaches need to be tested and adapted for the IDN domain and so far, NLP approaches have not been tested or validated on IDN data. Therefore, this thesis focuses on investigating the NLP techniques that might be used in such an authoring tool rather than dealing with the challenges of developing the authoring tool itself.

However, decisions taken at several stages in this research are informed by insights from this field since the NLP techniques are investigated with the primary aim of being integrated into a system that would involve Human-AI Collaboration. These decisions include the focus on extractive summarisation for better transparency and posing RQ1 to ground the choice of type of feedback for which NLP techniques are investigated in this thesis on the author's goals and requirements by asking RQ1.1.

2.4 Conclusion

Background outlined in sections 2.1 and 2.2 suggests that automatic intelligent feedback can help in authoring IDN and that such feedback may be generated using NLP, but this has not yet been investigated in much depth. This leads to the high-level research objective - Can NLP be used to help solve the authoring problem in IDN by generating intelligent feedback?

Through literature reviews described in Chapters 3 and 4, 5 concrete forms of feedback that have the potential to be useful to IDN authors as well as feasible to implement using NLP techniques was identified. Out of the types of feedback identified, extractive text summarisation was chosen for further investigation. Text summarisation was chosen since it mapped to many important UX dimensions and at the same time, is well-researched within the NLP community. Extractive summarisation was chosen for investigation over abstractive summarisation since it is more transparent.

No IDN summarisation datasets that can be readily used for this investigation could be found. Therefore, an IDN dataset using fan-made resources available online is created. Chapter 5 describes the methodology used for creating this IDN dataset in more detail and reports the performance of some baseline summarisation approaches (described in section 2.2.3.2) on this dataset.

Section 2.1.3 highlights how choices form a central aspect and unique affordance of the IDN medium. In Chapter 6, enhancing a classic summarisation approach by giving more attention to the text surrounding choice points is experimented with. This is done using Rationale based learning, which is described in Section 2.2.4.

Chapter 3

Systematic Literature Review of UX Dimensions in IDN

3.1 Introduction

As noted in the background, authoring Interactive Digital Narratives (IDN) can be very challenging. Creators often have to compromise on either the interactive complexity or the quality of the IDN artefact created[30, 13]. Most efforts at increasing interactivity, by relying on emergent narratives, for example[13], do so at the expense of authorial control and/or quality. Subsequent efforts, like drama managers[187] try to retain complexity and improve quality by introducing new architectures and more sophisticated technology[229]. While some authoring tools support debugging and visualization of the underlying structure [89], as complexity increases these become hard to fully comprehend.

A mixed initiative approach has been proposed as a way to overcome this issue of dissociative authoring[217] by giving the author feedback on the potential experiences possible within their work, referred to as Narrative Analytics in [156] and Intelligent Narrative Feedback in [217]. Similarly, Artificial Intelligence (AI) and Natural Language Processing (NLP) open up a lot of opportunities for generating intelligent feedback; for example, sentiment networks [131], emotional arcs [182]. By using this feedback to inform authoring, the author could make use of the affordances offered by a complex system while retaining visibility and control, and by extension, quality.

But what exactly is the feedback required by authors? Due to IDN's interdisciplinary and relatively novel nature, collecting these by finding and interviewing a representative set of IDN creators would be challenging. In this chapter an alternative path was taken, presenting a systematic review of IDN literature, focusing on the goals and concerns of authors in order to identify an appropriate set of feedback

items. Many papers talk about authoring goals, including expressing a specific intent [217], maximizing affordances of IDN[164] and creating a certain effect in the user [31] [164]. However, the most emphasised goal is a good User Experience (UX). Importance of UX is also reflected in how IDN creators often use UX evaluation to measure their success [229]. However, user experience (UX) is a very broad concept. To be able to implement NLP approaches that can give insight to UX, it needs to first be broken down into more concrete concepts. This leads to the first research question - *What concrete aspects of UX are of interest to IDN creators?*

Therefore, this chapter focuses on identifying the UX dimensions of IDN, with the idea that this could then form the basis of useful automated feedback to authors. This chapter attempts to answer the research question through a systematic review of how interest has been expressed in UX in the IDN community. This is done to define in a more concrete way, what types of feedback would be useful. The work outlined in this chapter was published as part of ICIDS 2020 proceedings[184]. The chapter is structured as follows: Section 2 discusses related work and background, Section 3 outlines the methodology used for the systematic review, Section 4 presents the results, Section 5 discusses findings and potential applications, and Section 6 outlines future work and conclusions.

3.2 Related Work

Previous work has identified some high level categories of useful feedback items for authors. But these deal either with specific problems, for example structural analysis to identify dead ends or short experiences [156], or are not comprehensive in that they focus on specific aspects such as emotional experience [217]. Some other papers on automatic analysis and feedback are discussed in the previous chapters as well [40, 174, 221], but the interest here is in higher level insights than the ones addressed here. For example, [32] talks about a similar idea of collecting parameters and then figuring out how to map them to corresponding cognitive processes but limits the scope of their discussion to two feedback items - suspense and surprise.

UX is a very broad area. Audience Studies is a whole field devoted to studying and developing theories surrounding audience's reception of media including IDN [91], and there are conceptualizations of UX (like those presented in [176] and [215]) which describe the process of experience or the relationship between design and experience. However, these do not easily extend to evaluation frameworks or feedback. A number of evaluation frameworks of UX have been proposed for IDN that could form this basis. For example, [190] consolidates Murray's high level interpretation of UX (as Immersion, Agency and Transformation [164]) with Roth's framework[238], to get twelve concrete UX dimensions. Whereas [118] uses GEQ[29], NEQ[36], and NTQ[90]

to create a specialised UX questionnaire. These are overlapping, but non-identical frameworks. Concepts like affect, curiosity, suspense and identification from [190] are closely related to the emotional engagement dimension in NEQ but are not quite the same. NEQ includes a narrative understanding dimension which is not talked about in [190]. Roth and Koenitz[190] notes how immersion is defined in different ways and settles on its broader high level definition, whereas in work by Kleinman et al.[118] immersion is simply the "capacity of the game contents to be believable".

There is clearly inconsistency and overlap in how UX is defined and understood by different researchers [31]. It is this that motivates our systematic literature review of papers talking about user experience in IDN.

3.3 Methodology

Our systematic literature review follows the established methodology set out in [116], this is formally five steps: outlining the research question, selecting keywords, selecting appropriate electronic resources, constructing a search method, and defining inclusion and exclusion criteria. The research question being asked in the review is: *What concrete aspects of UX are of interest to IDN creators?*, and the following section outlines our approach to the other steps.

3.3.1 Constructing the Sample

Springer[1] was chosen as the electronic resource because it is a database that has good coverage of IDN specific research (for example, ICIDS proceedings). While other resources like CHI Play contain literature on HCI, they tend to be more focused on games. To get IDN focused literature a lot of filtering would have been required. Similarly a wider search on Scholar resulted in many irrelevant results. Therefore, the search was focused on Springer.

Only papers from the past ten years (2010 - 2020) were included in order to ensure that the UX dimensions identified were relevant to current approaches and technology. This literature review was performed in June 2020 so it covers literature up till this date. Saturation sampling was chosen as the search method since the potential set of matches was too large to exhaustively analyse. The following search phrase was built by listing commonly used keywords for UX and IDN, searching for the intersection and adjusting to reduce number of irrelevant results :
((user OR player) NEAR/1 (evaluation OR experience OR experiences OR study OR studies OR engagement OR satisfaction OR enjoyment)) AND ("adventure game" OR "adventure

[1]Springer Link - https://link.springer.com/

games" OR "hypertext fiction" OR "emergent narrative" OR "emergent narratives" OR (interactive NEAR/2 (media OR cinema OR narrative OR narratives OR drama OR dramas OR fiction OR story OR stories OR storytelling) OR (game OR games) NEAR/1 (narrative OR narratives))
Any paper having the above keywords is likely to talk about some aspect of user experience of IDN in some way. However, for practical reasons, the following inclusion and exclusion criteria were chosen to select papers that are likely to give the most insight into which parameters are of interest:

1. *Does the paper focus sufficiently on narrativity and interactivity?* There are many types of IDN including Interactive Cinema, Mixed Reality, Storytelling Games and Documentaries and these were all included. Papers were excluded if they were discussing linear narratives, or did not put enough focus on narrativity. The framework proposed in [31] distinguishes narrative goals from system goals. Edutainment and games with a weak narrative component are examples of IDN applications that prioritize system goals over narrative goals. Only papers that focus primarily on narrative goals are included. For example, [170] is excluded because while it touches on narrative goals (affect, immersion), the primary focus is on learning.

2. *Is the paper about formalizing, measuring or evaluating user experience or some aspect of user experience of IDN or does it include some evaluation of it?* The kind of papers that are most likely to tell us which aspects of UX are of interest to IDN creators are those that include user experience studies or evaluation frameworks. Such papers also break down user experience into more concrete, measurable parameters. Papers that conduct computational evaluation instead of a user study also give us similar insights. Papers that theoretically formalize UX or discuss it in the context of IDN theory could help concretize UX and make the list more complete.

3. *Is the discussion on user experience in the paper detailed and concrete enough to provide relevant insight?* Some papers that discuss UX theoretically do so at a very high level [215, 175, 42] so including them is not useful for our purpose of concretizing it.

3.3.2 Coding Process

To enable saturation sampling, the results of the search were filtered and coded in batches of 20 papers. Each paper in the batch was compared to the criteria, and if it matched was reviewed, and coded as per the following process:

Batch number	number of selected papers	number of new codes
0 (seed papers)	4	28
1	10	13
2	6	4
3	6	2
4	5	0

TABLE 3.1: Systematic literature review - saturation sampling

1. UX dimensions were interpreted based on how UX was structured or evaluated in each paper. This was sometimes explicit, for example [190], but sometimes it had to be interpreted from how the authors discussed UX, such as [20] where they evaluate UX in terms of felt and actual understanding, perceived interactivity, narrativity and dissonance.

2. Sometimes, the papers include a hierarchical representation of UX dimensions[190] but since the interest here in concrete concepts only leaf nodes (called *low level concepts* in this paper) are added to the codebook.

3. If any overlap between the low level concepts is encountered while merging to codebook, the conflicting low level concepts are deconstructed based on their definitions and separated out.

This process is continued until all the papers in the batch are processed and took between 30-60 minutes per paper. The process was repeated for the next batch, until a batch with no new codes was encountered (saturation point). The number of papers included and new codes added per batch can be seen in Table 3.1.

3.3.3 Subjectivity in the Coding Process

Some subjectivity is intrinsic to the task so it cannot be prevented. In this section, two broad areas where such subjectivity is present are discussed: decisions surrounding inclusion of concepts and and those surrounding code definitions. This section aims to increase transparency by explaining the nature of this subjectivity with some examples of how it was handled.
Inclusion of concepts: In order to scope and contain growth of the codebook, concepts that are specific to a certain kind of narrative layer (eg - video quality), type of IDN (eg - distance between locations) or multiplayer experiences (eg - social relatedness) and concepts collected for contextualization (eg reasons for quitting, suggestions for improvement) are excluded.

Splitting up old codes to avoid overlaps and revising the definitions are not seen as adding new concepts. When a code is split, the code counts in the codebook are revised retrospectively. If enough information is not available to resolve an overlap

between two concepts, the more concrete or well defined concept is kept and the other one is discarded. If the overlap is minimal, both are kept. Concepts that are very similar are merged into a single low level concept and subtle differences are kept track of in the last column of the codebook.

Concepts can be placed on a spectrum based on whether the experiences are more intrinsic to system or user - for example, the degree to which user feels anxious would fall closer to user whereas perceived logical consistency and realism would fall closer to system. The interest here is in subjective user experience. Properties completely intrinsic to either system or player are excluded (eg- details pertaining to interaction design like number of choices and extrinsic goals, motivation to start playing, player skills). Some properties, though subjective, are still so intrinsic or specific to either the user or the system that modelling them as intelligent feedback is unlikely to be either feasible or useful - eg Loss of self consciousness, or the desire to save some particular non player character(NPC). In such cases, if an underlying generalizable concept can be discerned based on why the author was interested in this, then it is this concept that is coded. For example loss of self consciousness may have been collected as an indicator of presence. Desire to save an NPC may be interesting because it indicates the degree of attachment or identification with that character.

Some properties are too dependent on either specific user or system - for example degree to which user relates to a story is too dependent on the specific user and desire to save some particular NPC is too dependent on specific system. In this case the underlying generalizable concept is coded (save npc is coded as degree of believability, intrinsic objective) depending on why the author's were interested in this. If one cannot be interpreted from given information, it is ignored.

Subjectivity of decision regarding concreteness is mitigated to some extent by defining concrete nodes as the leaf nodes according to the structure of UX defined in the paper. However, the structure of UX is specified to different degrees in different papers. For example, it is very clear in [190] and vague in [201]. When the structure of UX is not clear in a paper, its interpretation and consequently the process of identifying the leaf nodes becomes more subjective. This impacts the decision regarding which concepts are concrete enough to be coded as low level. When a concept's concreteness is not clear from a paper, it is decided by considering the context and its description in other papers.

Papers sometimes talk about UX concepts that are not central to the scope of the paper - for example, in the background sections, follow up questions, when describing causal relationships to other concepts or in general discussion. Such mentions are often so brief that interpretation of meaning and concreteness would be too subjective, making merging them into the codebook difficult. So concepts that are not central to the framework or evaluation presented in the paper are excluded.

Code definitions: There is some subjectivity in how the UX concepts as described in each paper were considered to fall under the same code or not since there is some variability in how these terms are used and described in each paper. The code definitions were reviewed by an NLP expert and an IDN expert to reduce bias in the coding process. The code definitions as described in section 3.4 may vary slightly from how they were originally used in the paper. This is because sometimes only a subset of the code is mentioned. For example papers with an interest in just excitement or anxiousness, were counted as interested in in game or at game affect type and/or affect intensity accordingly.

Also, as noted in the previous section, this work codes for low level or concrete concepts that each paper talks about, or in other words, the leaf nodes of their UX breakdown. This is because this work is interested in concepts that are specific enough to attempt modelling and automatic analysis using NLP techniques. When the discussion of UX is too high level (eg. engagement and immersion) without breaking it down further, it is not considered useful since it is too vague and general to be modelled and implemented using NLP techniques.

To increase transparency and account for these factors, the variation caused by these factors is captured in the last column of Table 3.2 which shows the references as well as the sense in which concepts were originally used in those papers before they were split up or absorbed either fully or partly into the corresponding code. For some edge cases, there is some subjectivity around which concepts were coded using the same code and which codes are included in one category. For example Usability could have been placed under a different category or been a category on its own. However, here, it is placed under Agency following Roth et.al 's classification in [190]. Similarly, the concept "effort to change the story" was coded as usability since the way it was used in the paper[66], better usability corresponds to lower effort required to interact with the story and change it, but this isn't strictly the same thing as usability. There is subjectivity involved in this process and a different coder might have grouped things a little differently. However, what this work aims to do here is to give a concise view of the landscape of UX dimensions in IDN, presenting one way it can be untangled, rather than claiming this to be the definitive structure of UX in IDN.

3.3.4 Quantifying Subjectivity

As a result of feedback from reviewers, 16 months after the original coding was performed, I recoded a random sample consisting of 10% of the papers against the codebook. This helps quantify intra coder consistency. There was 88.23% overlap in the resulting codes, which is considered normal [41]. No new codes were added to the code book during recoding. 15 of the codes that were found during recoding overlapped with codes that were found from these papers in the first round of coding.

1 code that appeared during recoding was not included in the first round. This code was "motivation to continue" from [252]. This is due to ambiguity surrounding interpretation of the high level concept of engagement. One code (suspension of disbelief) was coded from [233] in the first round, but not in the second. This is due to a similar ambiguity in interpretation of "believability". It was coded as "perceived realism" and "suspension of disbelief" in the first round but simply as "perceived realism" in the second round based on how it was talked about in the paper. Such subjective decisions are described and clarified further in section 3.3.3.

To determine inter-rater reliability, another PhD student, who has a background in IDN research and authoring, independently recoded a randomly sampled 20% of the papers against the codebook. Agreement ranged from 0.98 in [226] to 0.76 in [124] and average agreement was 0.86. Cohen's Kappa is a metric commonly used to account for chance agreement when calculating inter-rater reliability [59]. The average Cohen's Kappa was 0.72 which is considered substantial [59]. 59% of the disagreements were due to codes being missed by either coder due to subjectivity regarding which concepts are considered central to the scope of the paper and whether the paper focused on a UX dimension enough. Additionally, there was an underlying assumption in the methodology design that the author's overall interpretation of UX would be the same as the breakdown they used while evaluating their systems. However, this is not the case in [32] where the goals of design mention concepts like usability whereas this is not reflected in the final evaluation. This resulted in some disagreements regarding which concepts were central to the scope of the paper. The remaining 41% of disagreements were due to subjectivity in interpretation and breakdown of codes based on the way it was talked about in the paper. For example, in [124], choice frequency was coded as interactivity by the first coder and as variety by the second coder.

3.4 Results

This process yielded 47 codes which can be placed under 8 categories as shown in Table 3.2. Note that the use of each concept in its original paper might vary slightly from the definitions given below. Sometimes only a subset of the code is mentioned. For example papers with an interest in just excitement or anxiousness, were counted as interested in in game or at game affect type and/or affect intensity accordingly. Papers that don't mention the code but a higher level concept, for example, believability, were counted for all sub-codes based on its interpreted meaning. Papers where a code is mentioned very briefly or not as part of the central work were not counted. The last column shows the references as well as the sense in which concepts were originally used in those papers before they were split up or absorbed either fully

or partly into the corresponding code. The following sections describe each category in more detail.

3.4.1 Agency

Six dimensions related to player agency were identified. **Autonomy** or the perceived freedom to do as the user wanted is related to the number and quality of options as well as navigational freedom. High levels of autonomy is usually desirable but may make the IDN more resource intensive. Therefore, authors often to make sure that they strike the right balance, providing the player with enough options that cover enough of what they might want to explore in the game. **Effectance** or perceived meaningfulness and impact of choices is related to being able to recognize when and how the storyworld was causally affected by the player's actions through clear feedback. Having many options that make no difference to the game is rarely engaging. Effectance serves to reinforce player engagement and immersion by validating player actions with tangible narrative consequences. This is also a pre-requisite for **control** which means being able to intentionally bring about specific goals and outcomes. Recognising patterns in how actions bring about certain consequences can lead to players being able to pursue their objectives in the game. Understanding the level of control players have is essential for the author when designing in game tasks and challenges for the player. [80] is interested in the idea of persuasion or degree to which the player was persuaded to take a particular action. Conversely, [229] talks about the degree to which the player felt like he was being manipulated by the system. These concepts were coded as **manipulation**. Understanding this dimension allows IDN creators to balance narrative guidance with player autonomy, ensuring that users do not feel overly directed which can negatively impact their sense of agency. [176] talks about **personalisation** or the extent to which the user feels that they experienced a story unique to their actions. This is related to the extent to which a user feels like they expressed their intention and extent to which they feel like the system understood this expression and has responded to it accordingly. Additionally **usability**, which refers to the user's experience with both the hardware and the software from a HCI perspective is also put under this category.

3.4.2 Cognition

Eight dimensions related to cognition were identified. **Logical consistency** is consistency of events and character behaviour as well as the themes and messages of the narrative. As an IDN author, maintaining logical consistency along all possible traversals of the IDN is important to ensure that the reader does not experience any jarring discrepancies in the story logic, breaking immersion. **Ambiguity** is the level of

Category	Code	num	references
Agency	Autonomy	6	navigational freedom[200, 199], availability of desired choices[124, 99] autonomy[190, 229]
	Effectance	7	effectance[191, 161, 190, 241] unnecessary choices[124] meaningful interaction [31], actions had no effect[99]
	Control	4	flow[236, 161] control[124, 201]
	Manipulation	3	likelihood of successful manipulation[80] autonomy[190]non limitation[229]
	Personalization	1	personalization[176]
	Usability	7	usability[191, 66, 124, 220, 190, 201]effort to change story[66]
Cognition	Narrative Understanding	9	Epiphany [71], observed understanding[233] understanding theme, intent [200, 199, 189, 20] narrative understanding[252, 118] intelligibility[31]
	Game Understanding	9	flow[236, 176]clear feedback,goals[161] expectations[191, 220, 189] understanding how to interact[66, 206]system intelligibility[31]
	Perceived Understanding	4	epiphany[71]closure[31]perceived understanding[233, 20]
	Logical Consistency	11	epiphany[71] believability[233, 220, 190, 191, 189, 201]visual communication[252] surprise, incongruency[32] immersion[118]coherence[229] inconsistencies[229, 201]
	Ambiguity	1	level of abstractness[31]
	Perceived Realism	10	believability[233, 220, 190] character believability[191, 189, 201] intelligent response [161] perceived realism[252]presence, immersion, naturality[118] breaks - sense of strangeness[229]
	Challenge	5	difficulty[236, 161, 176] was demanding[124] flow[190, 236]
	Storification	4	variation in experienced story[200] degree of storification[99] emergent narrative[198] narrative understanding, mental models[118]

TABLE 3.2: Codebook: UX dimensions

Category	Code	num	references
Immersion	Presence	7	sensory, imaginative immersion[236] presence[191, 190, 252, 118] Loss of Self Consciousness[161] emotional, spacial immersion[241]
	Suspension of disbelief	4	believability[233, 220, 190] role identification[189]
	Degree of focus	10	Absorption, attention, focus[206, 252, 124, 118, 161, 236, 191, 118, 176], transformation of time[161], attraction[92], awareness of surroundings[66]
	Object of focus	3	attraction towards[92, 206] reference[226]
	Identification	9	role adoption[191, 220] cognitive/behavioral responses[161]emotional engagement[252]suspense[190] identification[190, 229, 252, 189] perspective[226] like/dislike[242]
	Continuity	7	flow[236, 191, 67, 176], inconsistencies[229, 201], breaks[229], relatedness[242]
	Aesthetics	4	sensory immersion[236] pleasantness[191, 220, 124]
	Safety	1	safety[176]
Affect	in game affect intensity	17	Suspense, tension, anxiety [233, 236, 200, 241, 23, 220, 190], Affect, emotional state [233, 236, 191, 161, 220, 190], Enjoyment[233, 191, 67], Flow[236, 118], emotional engagement/immersion[199, 118, 252, 241], behavioural responses[161], Reception[252], closure[176], Curiosity[190], Pleasure[229], Surprise[201, 190]
	in game affect type	15	suspense[233, 236, 200, 23, 190] affect[233, 236, 191, 161, 198, 190] enjoyment[233, 191] flow[236, 118, 190] reception[252] closure[176] emotional state[92, 220] curiosity[190] emotional engagement[118] pleasure[229] surprise[201]
	at game affect intensity	5	annoyance[124], enjoyment[220, 229, 190] interest,fun[118] flow[118, 190]
	at game affect type	6	annoyance[124]affect-technical[198] enjoyment[220, 229, 190] flow[118, 190] interest, fun[118]

TABLE 3.3: Codebook: UX Dimensions contd

Category	Code	num	references
Drama	Curiosity	10	curiosity[191, 233, 199, 189, 124, 67, 220, 190] [201] temporal immersion[241]
	Closure	2	narrative closure[31, 176]
	Uncertainty	13	Epiphany [71], Suspense [233, 236, 191, 200, 32, 220, 190], imaginative/emotional immersion[236, 241] curiosity[220, 190] believability[191] predictability[124, 32, 201] Surprise[32, 201]
	Expectation	9	Suspense[233, 220, 191, 200, 32] imaginative/ emotional immersion [236, 241] expectation[189, 32] surprise[201, 32]
	Desired outcomes	3	satisfaction with ending[124] dreaded/desirable outcomes[190, 124]
	Novelty	1	novelty[124]
	Variety	2	variation in experienced story[200]variety[229]
	Themes	4	theme[199, 200]images[229] escalating climax[201]
Rewards	Eudaimonic appreciation	4	eudaimonic appreciation [233, 190] meaningfulness, take-away[189] pleasures of reflection[176]
	Sense of reward	5	Auteletic Experience, intrinsic rewards[161] feeling rewarded[176, 124, 201] curiosity[190]
	Learning	1	cognitive responses[161]
	Interest	2	increase of interest in the topic[189] edurability[124]
Motivation	to continue	5	continuation desire[21, 198, 20, 118] engagement[67]
	to replay	5	desire to replay[236, 200, 159, 198, 124]
	to interact	2	desire to explore/get involved[199]motivation to change story[66]
	Objectives	1	objectives[198]
	Activities	1	activities[198]
	Reinforcement	2	catharsis[176]accomplishments[198]
Dissonance	Interactivity	3	frequency choices[124]participation[241] interactivity[20]
	Narrativity	1	perceived narrativity[20]
	Dissonance	4	disruption[124, 241] narrative play[159] separation of interactivity and narrativity[20]

TABLE 3.4: Codebook: UX Dimensions contd

abstraction or clarity of the content. According to [31], narrative is said to be unambiguous when the content predisposes audience towards one and only one interpretation. High ambiguity can encourage interpretation and curiosity whereas low ambiguity facilitates a clear understanding. The author might want to balance this differently in different aspects of the narrative to guide the player's experience while allowing a level of interpretive freedom. **Degree of storification** is the extent to which a self-narrated story and mental models are created internally in the player. This becomes important especially in the case of IDN experiences that are more emergent or less closely authored, requiring active user participation in creating a narrative from the experience. **Narrative understanding** is a measure of how much the user understands the story as intended by the author. Having insight and control over this dimension of UX allows authors to ensure that all information is set up and released appropriately and that the core messages of the narrative is communicated as intended. Similarly, **game understanding** refers to how much the user understands game elements like clarity of goals, rules, boundaries and how to interact with and influence story. This allows the author to ensure that the players know how to effectively navigate and influence the story's progression. **Perceived understanding** is the user's subjective sense of understanding or the degree to which users felt like they understood the narrative rather than their interpretation of it being conjecture. Narratives often include moments when the narrative elements "click" into place. Authors may want to intentionally time moments of realization to make the experience both intellectually stimulating and rewarding. **Challenge** is a measure of how difficult users found the game and if they found that level of difficulty necessary, meaningful and enjoyable. Optimising the level of challenge is essential for effective design of game elements in the IDN, allowing the players to achieve an enjoyable "flow" state. **Perceived realism** is the game's closeness or resonance with reality judged on plausibility of events and character behaviour, perceived intelligence of system and characters and the degree to which the experience does not feel engineered. Implausible storylines can break immersion making the experience less engaging.

3.4.3 Immersion

Eight dimensions of Immersion were identified. **Presence** is related to the degree to which the user feels like they have left the actual world and entered the story, the feeling of being in the mediated space with mediated people. **Suspension of disbelief** refers to the degree to which the player loses awareness of the medium through which the experience is transmitted. **Degree of focus** or absorption refers to the degree to which the user's abilities and attention is focused on the experience. Sometimes there is also interest in the **Object of Focus** - game, narrative or reality frame. **Identification** or connection refers to the perspective adopted by the user as well as affective

disposition towards different story elements. It includes the degree to which users identify with the role and the story as well as the degree of attachment, empathy, and sympathy felt towards different characters. **Continuity** is the degree and duration of ongoing continuous involvement in the storyworld, merging action and awareness and the absence of breaks in the narrative caused by sudden changes in tone or the occurrence of abrupt, unconnected events. **Aesthetic pleasantness**, or the degree to which the user finds the setting and layout appealing, is also included in this category. Though not commonly discussed, [176] also talks about immersion in relation to the user's perceived **Safety** and how past a certain level of immersion, the user is at the risk of feeling unsafe. Monitoring these dimensions could help IDN authors identify parts of the IDN that may break immersion.

3.4.4 Affect

Authors usually want to trigger certain emotions in the player and create a certain pattern of arousal and relief. While affect encompasses a vast range of experiences, it is listed in the codebook as in game and at game affect type and intensity. **Affect intensity** is the intensity of emotional arousal and engagement felt. **Affect type** refers to the type of affect. More than 40 types of affects were listed from all papers together (e.g. Exhilaration, Anger, Frustration, etc.). Listing them out as separate codes does not seem useful but an important distinction to make is between affect felt towards the application or the game itself versus the same emotions aroused by events in the narrative. This is similar to the idea of at-game and in-game frustration described in [158]. This is differentiated as **at-game** or **in-game** affect.

3.4.5 Drama

This category relates to traditional narratology and drama. **Curiosity** is defined here as the degree of interest in the story, progression and actionable possibilities, or simply, a desire to find out more. **Themes** refers to topics, images and tropes that the user identified in the experience. **Novelty** refers to perceived newness and innovation in different elements of the experience. **Variety** refers to number and diversity of choices, experiences and actions. **Closure** is the degree to which users felt like the experienced story was complete and that the relevance of all story elements was revealed [31]. Suspense and surprise were absorbed into other concepts including **uncertainty** or predictability of progression and system responses and **expectation** aroused by a situation or narrative prompt. Suspense also includes the code - **desired outcomes** which refers to the user's dreaded and desirable consequences as well as satisfaction with how the story progressed. These are narrative devices that authors

use and can easily keep track of when writing linear narratives, but this becomes increasingly hard to manage with non-linear plots.

3.4.6 Rewards

Reward systems are essential to retain engagement both for narrative and interactive experiences. Four dimension related to rewards were identified. **Eudaimonic appreciation** is a measure of perceived cognitive and emotional meaningfulness of the experience (in terms of deducing general life lessons, insights into the meaning of life or how much the source challenges perceptions and life stories of the user.) **Sense of accomplishment** is related to the degree to which the player found the experience intrinsically rewarding and considers their investment in it worthwhile. **Learning** is a measure of how much playing game improved skill, knowledge or intelligence and arousal of **interest** stands for the degree to which the experience created an interest in the topic or in IDN. Keeping track of these dimensions can help IDN authors set up and pace their rewards to create an engaging experience.

3.4.7 Motivation

Six dimensions related to motivation were identified - **Objectives** refers to intrinsic objectives that the user developed while playing. **Activities** refers to what types of actions (interface/solve/ sense/ socialize/experience story and characters/ explore/experiment/create/destroy) users planned to or wanted to perform. **Reinforcement** refers to types of rewards that kept them motivated (completion, advancement or achievement). This was included in this category rather than Rewards because they were interested in the reward in the context of continuation desire. The remaining dimensions - intensity of desire **to continue** playing, desire **to interact** and desire **to replay** are self explanatory. In the context of user studies, understanding these dimensions helps authors place narrative milestones, challenges, choices and rewards in a way that is streamlined for player behaviour and measure the overall success of the game. As simulated feedback using game playing AI that mimic human play patterns, these can help gain some insight into these dimensions in the early stages of prototyping and development.

3.4.8 Dissonance

The final category has only three codes. **Interactivity** refers to the user's perception and satisfaction of the degree of participation or interactivity. **Narrativity** refers to user's perception of the game's focus on narrative elements as compared to its game elements. **Dissonance** stands for the degree of perceived dissonance or harmony

experienced between game and narrative elements. A satisfactory level of interactivity and narrativity with low level of dissonance indicating seamless integration of gameplay with story are key objectives in IDN design.

3.5 Discussion

The codebook shown in tables 3.23.3 and 3.4 tells us what aspects of UX IDN creators are interested in, answering RQ 1.1 — *What concrete aspects of the reader/player's experience interest IDN authors?*. While there have been many efforts to formalize and break down UX into simpler dimensions, they have resulted in many different interpretations - each concept being defined slightly differently in different papers and concepts overlapping each other to varying degrees in their many definitions. The intention of this work is not to promote a new belief of what UX should look like, or to claim that this is a definitive list of UX dimensions, rather the work presented here brings together and untangles those interpretations of UX expressed in the IDN literature, showing us ultimately what concrete dimensions of UX can be considered to be of interest to this specific community. The counts associated with each concept also gives us some insight into the relative interest and usefulness of modelling different dimensions of UX, although there will be other factors at play (for example, how commonly they are discussed in other communities, or the availability of instruments with which to measure them). The references column also tells us in what sense the interest was originally expressed. Researchers are often interested only in specific aspects of UX but this table may be used in evaluation to give a broader and more complete understanding of UX for IDN, and to identify dimensions that are considered less frequently. For example, while effectance, autonomy and usability are widely evaluated, concepts like control, manipulation and personalization are given less attention, even though they might provide useful insights about the user's experience of agency. Other commonly used evaluation frameworks like [190] and [29] can be seen as focusing on a subset of the codes listed.

Springer was chosen as the electronic resource for the UX review since it has enough breadth and it contains ICIDs proceedings, so it was considered sufficient to map out the space. It includes enough complexity for it to be sufficient, but it cannot be considered complete since by not including the other resources, this work might have missed out on some relevant work.

The main motivation for our list is so it can be used as a starting point for generating automated feedback for the author to assist authoring. Authors usually want to create a certain pattern of effects for the user. For example, [190] talks about cyclical building up and relieving of curiosity. While it might be possible for the author to visualize, predict and create this effect when writing linear stories, it becomes hard to keep track

of this when the space of potential stories grows. However, if UX dimensions like curiosity, expectation, uncertainty and affective responses could be automatically modelled, then it should be possible to reflect this to the author for all the possible paths through their narrative, allowing them to more efficiently tailor the content and tune its effects on the user along all branches, resulting in better authorial control and user experience.

It is important to note that while the work presented in this chapter gives us enough insight into the research question to guide further research into NLP techniques that might be used to generate feedback related to them, creating an exhaustive list of UX dimensions and authoritatively determining their relative importance would require covering more sources and validating the results with a representative set of IDN authors.

3.6 Conclusion

This chapter examines what the IDN community considers to be the important UX dimensions for its users, readers and players. The goal is to understand what automated feedback might be useful to IDN authors. To gather feedback items that will help the author a systematic review of UX in the IDN literature is performed. This process untangles the many overlapping interpretations of UX by different IDN researchers and yields a list of 47 feedback items covering 8 categories: Agency, Cognition, Immersion, Affect, Drama, Rewards, Motivation, and Dissonance. The next chapter will investigate NLP techniques that will help automatically estimate these. Integrating such feedback into an authoring environment would not only help detect problems but would also allow authors to closely tailor the experience for their users without massive-scale iterative playtesting. It could free them to write more complex narratives without losing sight of how each branch of those narratives impacts the user.

Authoring goals go beyond UX, and these could also be assisted by automation. For example, the desire to express a specific authorial intent calls for feedback at a lower level than UX. This is in part accomplished by visualizations such as those in Novella[89] and progression maps[40], but as complexity grows, more insightful views like sentiment networks, maps, timelines and dramatic arcs are also worth considering. Reviewing commonly applied narrative devices, formalisms, conventions and authoring practices might tell us which would be most useful. Similarly reviewing critical analyses of IDN works and IDN theory might show us what feedback items can help the author maximise the use of IDN affordances or train them in the art of IDN [122]. This might also mean feedback that helps them fluently use the authoring tool (e.g. system feedback [217]).

By focusing on the dimensions of UX specifically of interest to the IDN community this work has shown the range of feedback that automation might usefully provide, answering RQ 1.1. In the next chapter, these UX dimensions are mapped to NLP tasks and applications to understand which of these forms of feedback NLP techniques can feasibly generate and gain insight into RQ 1.2.

Chapter 4

Mapping UX Dimensions to NLP Research

4.1 Introduction

The nonlinear structure of IDNs makes authoring them very challenging since even in the case of simple interactive narrative structures (like choice based branching structures), the authors quickly end up having to keep track of many different story permutations. It becomes impossible to keep track of the internal narrative space above a certain size (in terms of either the length or breadth of the tree) and it becomes harder and harder for the author to picture how a reader might experience the content that they are writing. Usually, the author has to write a lot of content and perform iterative playtesting to get an understanding of the user's experience, but this an expensive process that is difficult to do exhaustively. The creation of a specific experience for the user is commonly emphasised as the author's primary goal [164]. So this inability to picture user experience while writing is a major challenge for authors. Our work is aimed at reducing the burden on human authors, not in terms of the amount of content that needs to be written, but by helping them manage the complexity of the narrative by giving them more oversight and control over the the interactive experience that they are creating.

A mixed-initiative approach has been proposed as a way to overcome this issue by giving the author feedback on the potential experiences possible within their work, referred to as Narrative Analytics in [156] and Intelligent Narrative Feedback in [217] but this has not yet been investigated in much detail. This is referred to as *Authoring Feedback* in this thesis. Artificial Intelligence (AI) and Natural Language Processing (NLP) open up a lot of opportunities for generating intelligent feedback. Some types of feedback can also be provided in real time while the author is writing to help them keep track of all the paths in the interactive narratives that lead up to the portion of

text that they are currently writing. This work was written up as a position paper and is under review at IEEE Multimedia.

The last chapter looked at what aspects of user experience could be useful to IDN authors if estimated automatically and given as feedback. This chapter tackles the next research question - *How can NLP techniques be applied to generate feedback that can give insight into the identified UX dimensions?* An exploratory review of NLP work related to UX dimensions identified in the previous chapter and on narrative text in general is performed and the UX dimensions are mapped to areas of NLP research that can be leveraged to generate feedback related to them. The mappings were made based on the similarity of the description of NLP problems and the definition of the corresponding UX dimensions as well as the applicability of NLP techniques to gain insight into a given UX dimension. For 23 of the identified UX dimensions, related NLP work that might be applied to provide automatic insights was found. This work then takes a pragmatic look at three of these NLP mappings in more detail and discuss challenges, state of the art, and what these might look like if integrated into an authoring tool, leading to five concrete examples of feedback items that could be generated using existing NLP approaches.

4.2 Related Work

Previous works like [156] draw attention to the issue of lack of author's visibility of the story space and user experience and proposes automatically generated feedback as a way to help. Later work on UX Dimensions in IDN[184] builds upon these ideas by investigating what kind of feedback items could be useful to authors and replace iterative playtesting. This chapter looks at how NLP techniques might be applied to do this automatically. Even though some authoring tools use AI to assist authoring, it is used generatively to reduce authoring burden [211] or to ensure consistency[153] it has not been used analytically - to analyse the space of possible player experiences that can be had from the content written and give this as feedback that assists authoring. The authoring tools that do give feedback to assist authoring like [40] do not investigate using NLP techniques to do so. AI and NLP could be used to generate a variety of more sophisticated feedback and that is what is addressed in this paper.

4.3 Methodology

To get a sense of what techniques are available to estimate these UX dimensions automatically, NLP literature related to each of these dimensions was reviewed. Replicating the systematic review methodology for NLP work is not feasible due to

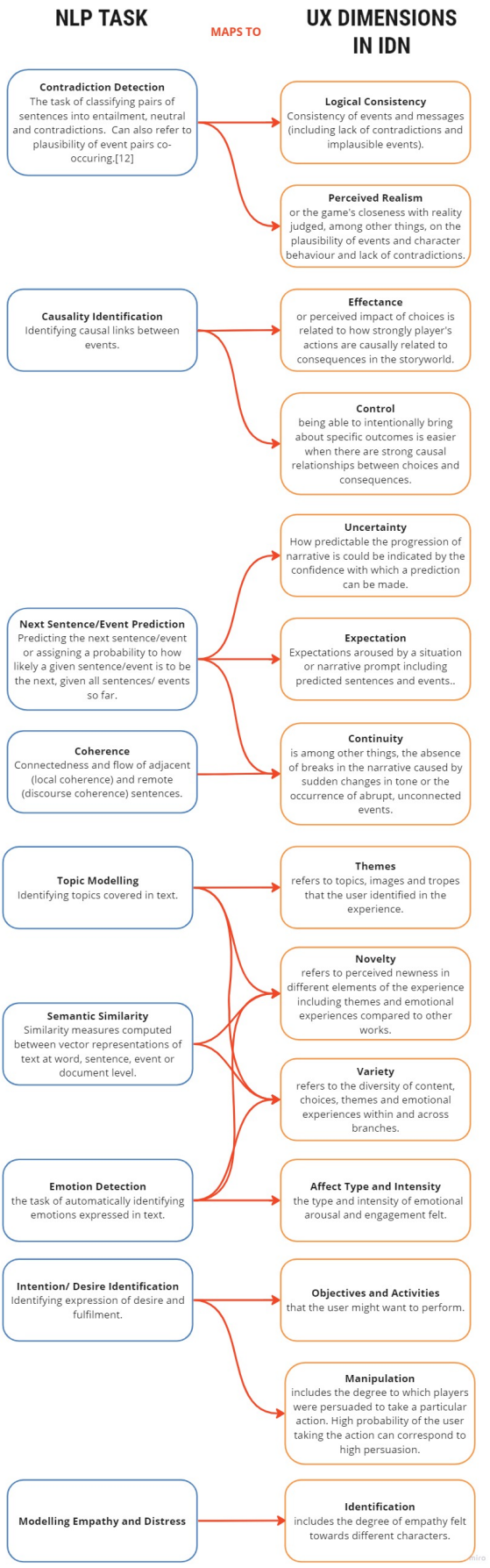

FIGURE 4.1: Mapping NLP Tasks to UX Dimensions based on their definitions

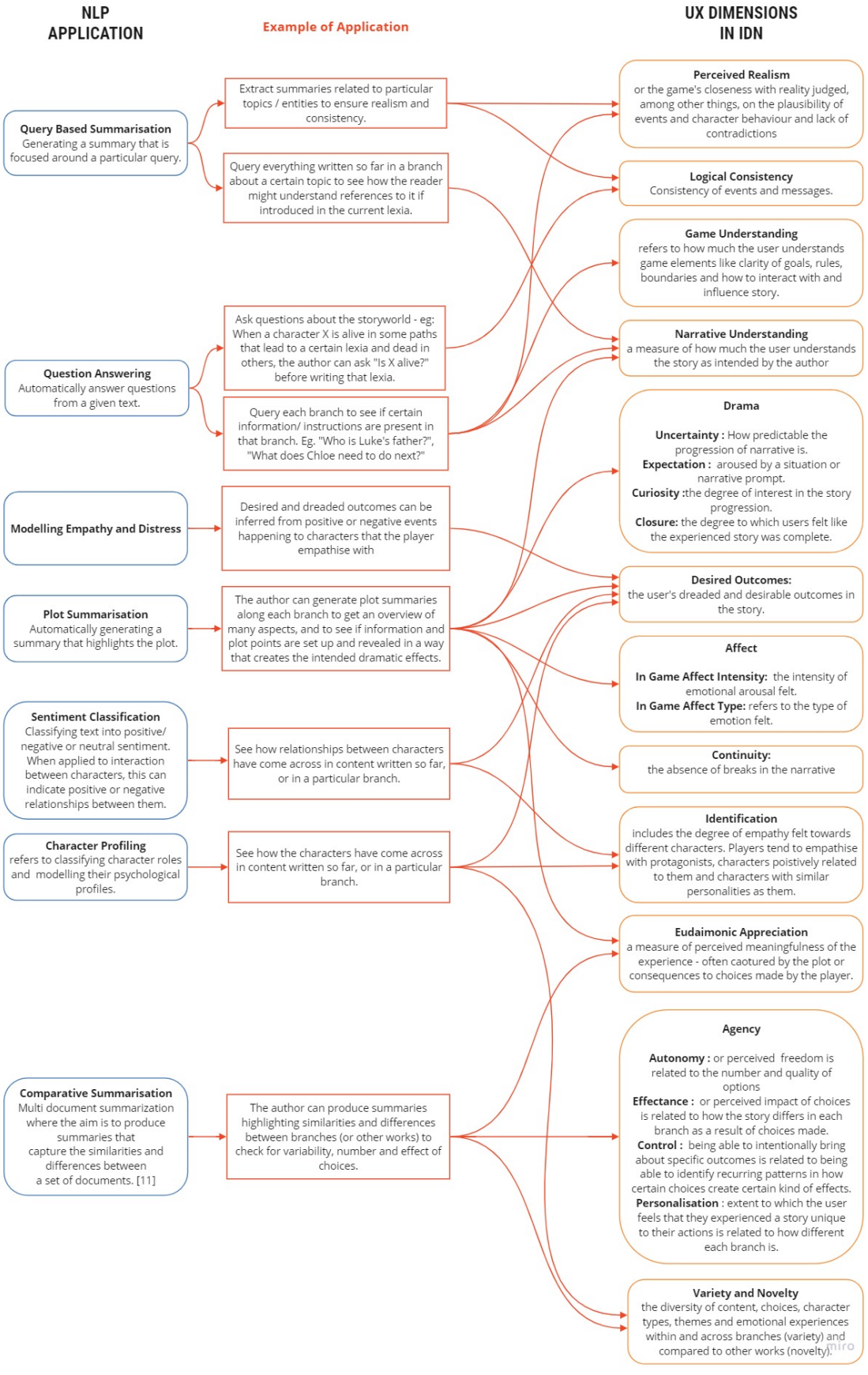

FIGURE 4.2: Mapping NLP Tasks to UX Dimensions based on their applicability

size of the field and because relevant work is spread so widely, it is also not feasible to focus the search to a subset of NLP research. Additionally, unlike UX dimensions for IDN which was talked about differently in different papers, NLP research areas covered in this chapter are well understood and do not need to be discovered or unified through a systematic review. Therefore, for this second phase, an exploratory approach is adopted, basing our search on the UX dimensions identified in the previous section.

The search was performed primarily on ACL[1](Association for Computational Linguistics) since it contains a variety of good quality NLP research. Google Scholar was also used to be more comprehensive. This is a lot wider than the review constrained to Springer in the previous chapter. The main reason for that was that for the last chapter, the objective was to untangle the many different ways in which UX was talked about, but here the research areas don't need to be discovered in the same way. It is more about understanding the relationships between the UX concepts and fields of NLP researcg so its more exploratory and does not need to be constrained in the same way. Each of the UX dimensions was searched for with terms associated with NLP. Search terms were determined by identifying potentially relevant NLP research areas from surveys and literature reviews of natural language processing [227] or through trial and error by trying out terms associated with each UX dimension. For example, corresponding the UX dimension in-game affect, NLP search terms like affect detection and emotion recognition was used and for effectance terms like 'causality' and 'causal relationships' returned more relevant results. Literature on narrative NLP was also searched for. This included keywords like 'narratives', 'fiction', 'literary texts', 'novels','movies', 'screenplays', 'plot', 'story', 'games' and 'interactive narratives'.

Even though the ease with which NLP work could be found for each UX dimension reflects its availability to some extent, note that this type of exploratory review cannot provide a comprehensive, quantitative view of how much NLP research is available for each UX dimension. The primary intention behind this mapping is to illustrate the potential of NLP within IDN by drawing connections between well-known areas of NLP research with UX dimensions that are interesting to IDN authors with some examples that back up these mappings and can act as starting points for researchers from either field.

Mappings were refined and validated through review by an IDN expert, an NLP expert and an experienced IDN author. The exploratory search and mapping were done by the author of this thesis and then reviewed and revised by the experts and the IDN author through iterative validation steps to mitigate bias and control subjectivity.

[1]https://aclanthology.org/

The mappings are created in two ways: based on definition and based on applicability. Some NLP tasks (for example, Emotion Detection) by definition, correspond to the task of estimating a certain UX dimension automatically from text (affect). In some other cases, even though the NLP problem does not directly correspond to a UX dimension, it can be applied indirectly to gain insight into a certain UX dimension. Both these views are useful in the context of generating authoring feedback, and are described in more detail in the following sections.

4.3.1 Mapping Based on Definitions

These are mappings that were created because the paper tackled an NLP problem that directly corresponded to any of the IDN UX dimensions. These mappings were decided based on the description of the NLP problem and definition of the UX dimension. A mapping was made when the descriptions suggested a potential for positive or negative correlation between these concepts. Figure 4.1 shows which UX dimensions map to which NLP tasks and the reasoning behind these mappings by showing the definitions of the NLP tasks and UX dimensions and how they are related.

A mapping between a UX dimension and NLP task indicates that that particular NLP task is closely related to estimating the corresponding UX dimension automatically from the text - and hence methods from these research areas could be applied to this end. To clarify, this doesn't necessarily mean that existing methods in these fields are directly applicable or already good enough to generate the necessary feedback at a usable quality. It will require further experimentation and research to see to what extent they would help in practice and how existing approaches would need to be modified for this domain and use case. The mapping does however indicate that potential and serves as a call for collaboration between these two communities by connecting affordances offered by NLP techniques to requirements from the IDN authors.

4.3.2 Mappings Based on Applicability

These are mappings that were created because the NLP technique could be applied indrectly in some way to give insight into a UX dimension. Figure 4.2 outlines NLP applications which can indirectly give insight into mapped UX dimensions, even though they don't directly try to estimate the UX dimension. For example, plot summaries along different branches help the author better understand how the plot progresses along each branch, providing the author insight into several dimensions related to Drama like *uncertainty*, *expectation* (and by extension, suspense and surprise), *curiosity*, *closure* and *desired outcomes*. Figure 4.2 summarises these mappings

along with an example of how the NLP application might be applied to gain insight into the corresponding UX dimension.

4.4 Results

After the UX dimensions relevant to IDN were identified, the second phase of this research looks for relevant NLP techniques. This second exploratory search revealed 12 NLP research areas that could be applied to 23 of the 47 UX dimensions (either directly, indirectly or partially). No mappings could be found for the remaining 24 UX dimensions. Table 4.2 summarises the mappings shown in Figures 4.1 and 4.2.

Six NLP research areas were identified that are related to the broader research area of Machine Reading Comprehension (MRC). Different types of Automatic Text Summarisation (ATS) approaches were also found that could give insight to many of the UX dimensions at once by providing a compact view of the story space. Of the remaining NLP problems, five are are Text Classification (TC) tasks.

These NLP research areas are shown in Table 4.2 along with references to relevant papers from these fields. The References (General) column gives examples of papers tackling these tasks or recent literature reviews of the field that give an overall landscape of that task. The References (In Domain) column contains examples of papers where the technique was applied to narrative text like literary texts or movie scripts (these are closest in domain to the kind of texts found in IDN and narrative games). Note that the character profiling task has examples only under 'In Domain' since this task is specific to the narrative domain. The mappings are described in more detail in the following sections. The mappings are also shown as a heatmap in table 4.2.

4.4.1 Machine Reading Comprehension

MRC is a broad research area that includes several subtasks of interest to IDN. Contradiction detection is a classification task of classifying pairs of sentences into entailment, neutral and contradictions which is similar to the idea of ensuring *logical consistency* [213] between parts of a narrative. Sometimes, instead of strict contradictions, highly implausible event pairs are also considered contradictions in order to fit real-world use cases better [68]. There are also papers like [139] that directly address the issue of plausibility. This maps to *perceived realism* which also deals with plausibility of events in the story. Next Sentence Prediction is the task of predicting a sentence embedding of the next sentence given all the sentences so far [70]. The task is often framed as assigning probabilities to a set of potential next

sentences such as in suspense. It is often used in Language Modelling and deals deals with flow of textual data which is directly linked to *continuity*. Next event prediction [44] is a similar task where instead of sentences future events are predicted based on past ones. These can both be indicative of *uncertainty* (how uncertain the model is about the next event or sentence) and *expectation* (what is the expected next sentence/event).

Even though there is no NLP work on effectance, there is some work on automatically identifying causal links between events [188]. This could potentially be applied to capturing causal links between choices and their effects - giving an indication of *effectance*. Consistency of causal relationships between events also gives some insight into *control* as it indicates identifiable patterns in actions and consequences that the players can pick up on.

Local coherence [154] represents connectedness of adjacent sentences and discourse coherence represents the connectivity of remote sentences and overall flow structure of the document. This is partially similar to the UX dimension, *continuity*. Work based on intention, desire and belief identification like in [47], [180] might be applied to identify the user's *objectives and* activities that they might want to perform based on their actions so far. Predicting what actions the user is likely to perform also gives some insight into *manipulation* in terms of how likely they are to take the intended actions in the game. Tasks like question answering help the author keep track of the narrative space better, indirectly helping them with many UX dimensions. For example, letting the author interact with the storyworld they are creating through a QA interface like in [195] can help them maintain context and avoid *logical inconsistencies* and ensure *realism*. While it can be challenging for QA systems to perform complex inferences to retrieve answers, the author can also query the system with simple questions to get some indication of if certain information (indicating *Narrative Understanding*) or instructions (indicating *Game Understanding*) are present in a branch.

4.4.2 Automatic Text Summarization

Feedback in the form of automatically generated summaries can potentially help the author visualize many of the UX dimensions at once by providing a more compact view of the narrative space. Summaries help visualize the narrative space, so it helps in visualizing all UX dimensions a little better, but in particular it helps with the dimensions mapped to it in table 4.1. Different types of summarisation may be categorised as single or multidocument summarization based on type of input, generic or query focused summarization based on nature of output summary, general or domain-specific based on summarization domain [74]. This section discusses these mappings and different types of summarization in more detail.

Generic summarisation in the narrative domain usually refers to summarising the overall plot [173, 133]. IDN summaries contain game and interactive elements in addition to plot elements with these elements sometimes being closely entwined with each other. Summaries along different branches of an IDN can help the author visualize plot progression, providing the author insight into dimensions related to Drama like *uncertainty*, *expectation* (and by extention, suspense and surprise), *curiosity*, *closure* and *desired outcomes*. It also helps the author get a sense of the overall *affective dynamics* of the story and assists in ensuring *continuity* and *narrative understanding*. The message of the story is often tied into the overall plot so plot summaries can also give the author an indication of *eudaimonic appreciation*.

Summarising IDN can also be seen as multi document summarization [9] since each lexia or different readings through the story can be thought of as separate linked documents. Comparative summarization [102] is a type of multi document summarization where the aim is to produce summaries that capture the similarities and differences between a set of documents. This is very interesting in the context of IDN as it can be applied to compare different branches and highlight differences. This can help the author visualize variety and many agency related dimensions like *autonomy*, *effectance*, *control* and *personalization*. Since meaningfulness in IDN is often hidden in the choices and differences across branches, it can also help the author picture the *eudaimonic appreciation* dimension as well. It is also a way to produce a compact summary of the entire IDN, allowing the author to keep better track of both the story and structure which indirectly helps the author improve all of the UX dimensions.

Query based summarization can also be applied in the form of feedback to let the author extract summaries related to particular topics and entities in the story to ensure consistency and realism. This would provide more visibilty than a similar QA based interface. It could also be used to query everything that was written so far in a branch about a certain topic. This could help the author see how well the reader might understand references to that particular topic if introduced in the current lexia that the author is writing. This links to *narrative understanding*. There is also some work around controllable summarization[163] that takes user contraints into account when creating summaries which can be applied in a similar manner.

Additionally, controllable summarisation [163] allows taking user defined constraints into account while producing summaries, allowing the IDN author to have more control over the summary produced. Extractive summarization usually involves assigning sentence scores according to some redundancy and diversity constraints. To adapt automatic summarization techniques to a particular use case or domain, the scores are influenced by additional criteria. For example, in comparative summarization in [102] topic diversity across documents influences sentence scores and for Query/Entity focused summarization like in [101], relevance with regard to

query influences the scores. Following this logic adding additional constraints related to any of the UX dimensions to produce specialized summaries could also be investigated. For example, NLI models could be combined with summarization to produce summaries that highlight *uncertainty* or *expectation* of a given event (by assigning higher scores to sentences that make the event likely or unlikely). This would be more informative and intuitive to the author than looking at metrics alone. Similarly emotion detection models could be combined with summarization to assign higher weights to emotion heavy sentences, and models for causality could be combined to produce summaries that show a causal chain of events to illustrate *effectance*. Section 4.4 goes through each of the mappings and explains why they were created and section 4.5 illustrates them with concrete examples .

4.4.3 Text Classification

NLP work around emotion detection, empathy estimation and character profiling are classed under the broader area of Text Classification since these tasks predominantly involve classifying text spans into predefined labels. Emotion detection is the task of automatically identifying emotions expressed in text [5, 2]. There is also some work specifically focused on prediction reader emotion or emotion aroused in the reader by some text [94]. This is directly related to the task of automatically estimating *in game affect type* and *intensity*. It also indirectly allows the author to ensure *variety* in possible emotional experiences.

No NLP work could be found on estimating identification as such, but there is some preliminary work on automatically modelling empathy [35] which is related to *identification*. Papers on character classification like [232] can also be helpful since players tend to empathise with protagonist characters more. Character profiling [81] can indirectly give the author an indication of how the character has come across in content written so far (or in a particular branch). Profiling also helps the author ensure a diverse cast of characters. Character profiles and empathy aroused can also indicate what the player's *desired and dreaded outcomes* might be.

Topic modelling approaches like in [106] may be applied to identifying themes and tropes present in different branches. Topic analysis for generating feedback related to themes has also been proposed in [156]. Visualising the major themes in different branches also allows the author to compare them against each other in terms of topics and get a sense of the *variety* of topics across branches. Similarly, the author can compare major themes in their work to those in existing works to get a sense of *novelty*.

Variety can be expressed using similarity measures computed between vector representations of text at word, sentence, event or document level using similarity metrics like those discussed in [60]. For example, greater distance between

embeddings for different branches indicate greater variety and lower redundancy in the type of content across branches. Similarly, greater distance between embeddings for choices indicate more diverse choices. Similarity measures between the work currently being authored and existing creative works in the field could also give some insight into *novelty*.

Topic modelling and semantic similarity are included under this category even though they are not strictly text classification tasks. Topic models are an automated way to create a set of classes, which can then be used to do text classification. Vector similarity for classifying variety and novelty can be thought of as a task to categorize properties of text.

4.5 Case Studies: Examples of Potential Feedback Items

To explore whether the theoretical mappings identified above do, in fact, correspond to real opportunities that can be exploited, this work took a pragmatic look at what algorithms and data were available in three of the NLP research areas identified via the process above (Emotion Detection, Next Sentence Prediction and Generic Summarisation), culminating in concrete examples of the types of feedback that could be generated using existing technology. Generic Summarization was chosen since it indirectly maps to many UX dimensions. Next Sentence Prediction and Emotion Detection was chosen because they map to UX dimensions that are of high interest to the IDN community and are well researched in the NLP community.

This work explores the different approaches used, the type of data, the status of state of the art in these fields, what integrating these approaches into an authoring tool might look like, and potential challenges to adapting existing methods to our domain and use case by referring to recent literature. This leads to 4 concrete examples of potential feedback items linked to specific techniques, and use cases of how an author might use that feedback to inform authoring.

The existence of these concrete examples validates that there are real opportunities that can be investigated. Additional validation of the mappings will need to be undertaken empirically by integrating them into an authoring tool and running experiments to see how they affect authors, preferably as part of longitudinal studies with expert communities. These examples aim to illustrate how there is ongoing NLP work and existing approaches that attempt to automatically estimate features of text that have the potential to assist the author by providing intelligent insight on some of the UX dimensions that IDN researchers have previously tried to understand through iterative playtesting. These mappings thus indicate a potential for further research and collaboration and set out one possible agenda for future mixed initiative research into IDN authoring.

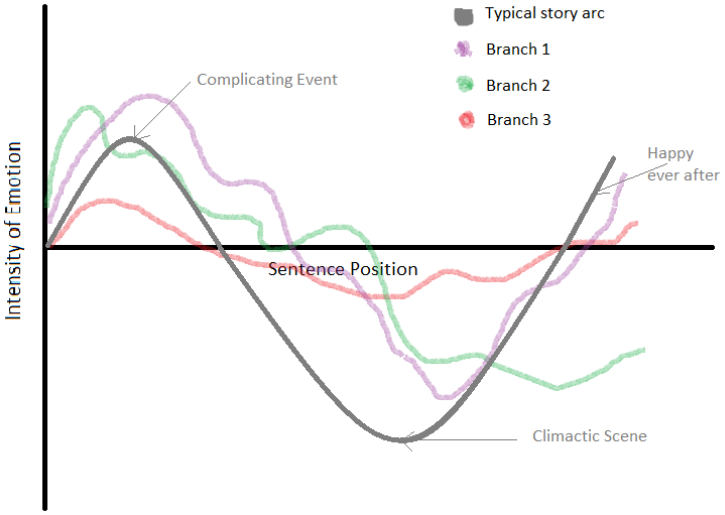

FIGURE 4.3: The bold grey line shows the typical emotion arc for linear narratives. Six common shapes of such emotional arcs have been identified in previous work [182]. The other lines show illustrative sample emotion arcs in IDN. Branch 3 has no emotionally intense regions, indicating that that branch might be bland. Branch 2 has a different trajectory compared to the rest indicating better variety.

4.5.1 Emotion Detection : in game affect type and intensity, variety

Out of the datasets that could be gathered from recent literature reviews on emotion detection [5] [2], there are 5 which include story-like text (3 have literary text and 2 are based on dialogues from TV shows) but only 2 of these have annotations for the reader's emotion. A majority of the work on emotion detection focus on emotion expressed in the text and not the reader's emotion. Some work on reader emotion exists that could serve as useful starting points. These include work on prediction of feelings aroused by news articles [94] (uses attention based HNNs and achieves an accuracy of 57.95% on SinaEmotion 2017), some models that have been trained and tested on reader emotion as well as writer emotion [264] (adversarial network based model proposed in reports r-value about 0.4,0.2 and 0.2 for valence, arousal and dominance respectively for reader emotion on fiction in EmoBank Dataset [34]), and a dataset of news articles that are annotated with additional emotional information including cause of emotion and the experiencer of the emotion [27]. Further, emotion expressed in the text can also be useful feedback since it has narratological significance as illustrated in [182] in which they use a lexicon based emotion detection method to plot emotional arcs over novels. They then cluster the arcs and find that there are 6 common shapes to stories as first theorised by Vonnegut [129]. Sentiment information is also used in other tasks like story cloze [44] where they learn sentiment trajectories from a corpus of stories and use that information to predict the ending of a given story [48]. [208] also used emotional arcs to automatically assess narrative quality of student

writing. In context of authoring feedback, emotion detection could provide feedback in the following ways:

Feedback 1 - Emotion Arcs : Plotting emotional arcs (reader or writer emotion) over each possible story like in [182] to identify bland branches or making sure there enough fluctuation and/or making sure that the shape of the story in each branch is as intended or fits narratological conventions.
Feedback 2 - Emotion Types Distribution : Producing different kinds of visualizations indicating fine grained emotion type and/or intensity by performing emotion classification using any of the available datasets [34, 4] and approaches [5, 2] so that the author can ensure variety of emotion in the experience.

4.5.2 Next Sentence Prediction : Uncertainty, Expectation, Continuity

Next sentence / event prediction can be used to calculate uncertainty and expectation of all sentences in any branch given all the sentences so far or a subset of the sentences. Since NSP can be used to get probability of sentences following another one[70], they can be used as an indication of sudden breaks in continuity as well. While most of existing research on Next Sentence Prediction are not performed on long narrative text [213, 44], large story corpora like Project Gutenberg[2] are available on which these methods can be applied or can help with domain adaptation. However, when it comes to accuracy, while ML models perform as well as humans in some MRC benchmarks, on tasks like story cloze[44] which is most similar to our use cases, machine performance is not as good as humans [213].

Most papers tackling NSP and story comprehension use the Story cloze dataset [44]. The task is to predict the correct ending to common sense stories. For example, [48] tackles this by using sentiment trajectories, checking for topical consistency and applying event sequence prediction. But these are very small simple stories and IDN stories are substantially bigger. Very few have looked at larger story texts. One paper that does apply NSP to longer narrative text (short stories) is [244]. They test various models of suspense based on uncertainty in short stories. They use a dataset of short stories taken from writing prompts [77] that are annotated per sentence for suspense as [Big decrease, decrease, same, increase, big increase] and use a hierarchical model based on GPT [178] for next sentence prediction. Suspense is modelled as surprise or low probability of current event given the context so far, sudden changes or distance from last sentence embedding, reducing variation in possible outcomes or entropy reduction, and reducing uncertainty of outcome or reduction of mean of uncertainty, all weighted according to emotional arousal. The best model agreed with human annotated suspense (Spearman's rho - 0.698) to a comparable degree to human annotators agreeing with each other (rho - 0.614).

[2]http://www.gutenberg.org

NSP could therefore be applied to generate feedback on uncertainty and expectation in the following ways:

Feedback 3 - Uncertainty Arcs : Calculating predictability of events or sentences along branches using methods like in [244] so that the author can identify boring, predictable branches, manipulate events to set up for suspense and surprise and make sure the uncertainty trajectory is as intended.
Feedback 4 - Debugging : In IDN debugging normally refers to issues with the interactive logic (for example, unreachable nodes [156]), however NSP provides a method for addressing the logic within the stories themselves. This can be through direct contradiction detection like in [68] or based on predicted probability of co-occurrence like in [244]. Highly unexpected events or sentences might actually be bugs created through unintentional combination of content units by the story engine. Such sentences could be highlighted to make sure that unlikely or unexpected sentences or events were intended that way by the author.

4.5.3 Summarisation: Several Dimensions

A recent survey of automatic text summarization techniques [74] talks about summarization applied to different use cases and domains. It only cites one paper, [111] about story summarization in which they produce indicative summaries that will help the reader decide whether or not to read the full story. They also say that complete plot summarization is a much harder task. However, there has been some later work in plot summarization like [49] which introduces a dataset for summarization of novel chapters and [173] which tries to summarise screenplays.

While there is some plot summarization work on novels [133, 246] and movies [173, 87] and a paper on harvesting data for query based summarization from Fandom [3] [101], most work on summarization is mostly focused on domains like news and academic papers. It might take some experimentation to see if the same approaches work for the narrative domain as well. There is also work on game log summarization [52, 18] and a dataset of transcripts and summaries of the show Critical Role [181] (a video recording of role playing game sessions). These are closest in domain to IDN text. State of the art models for abstractive summarization achieve Rouge scores of up to 44.79 on CNN/Daily Mail dataset [219]. No papers on benchmarks or comparisons between existing plot summarization approaches could be found, so determining state of the art in this area is not straightforward, but evaluations conducted separately in each of the papers indicate potential.

Summaries can be extractive (showing extracts from the original text) or abstractive (generating new text concisely paraphrasing the original text)[74]. Abstractive

[3]Fandom - https://www.fandom.com/

summaries are often more readable than extractive summaries but generative models used for abstractive summarisation are prone to hallucinations (producing inaccurate information)[128] since new text needs to be generated. On the hand, since extractive summarisation simply selects excerpts from the original text to reflect back to the author, it is less prone to generating inaccurate summaries.

Application of ATS as authoring feedback could look like:

Feedback 5 - Branch-wise summaries: Summaries of different readings (generated like in [156]) of the IDN. It provides a summary of all important events in each branch showing the author potential experiences the user could have and allowing them to visualize many of the UX dimensions especially those under Drama. It could also be used to show the author a summary of the potential routes a reader might have taken to a given point in the story, which could be useful when writing a new node.

4.6 Discussion

This chapter focuses on bridging two research communities - NLP and IDN, by mapping theoretical problems that are of interest to both communities. The mappings shown in table 4.2 show that there is a lot of untapped potential in applying NLP to generate automatic feedback to assist authoring that is worth investigating. In practice, the performance of NLP models affects their utility. Many of these techniques will also require modifications and adaptations to fit the IDN domain and specific use cases within it. This work also expands on some of these theoretical mappings to reveal the potential of integrating these methods into an authoring tool to provide author feedback, giving concrete examples of possible feedback items. Although the potential feedback items listed are just examples (there are many ways of integrating an emotion detection model or a summarization model into an authoring tool and there are many ways of presenting them to the author), they are concrete starting points bringing us a step closer to implementing intelligent narrative feedback.

No related NLP literature could be found for 24 of the UX dimensions. Two dimensions (usability and aesthetics) are not related to text and it is unlikely that these can be estimated using NLP. This also includes dimensions for which, while it might be possible to estimate them automatically, it is not immediately obvious how they can be modelled in terms of properties of the text. This is because in case of these concepts it is not straightforward to see what properties of the source text cause the desired effects in the user. Concepts like presence, suspension of disbelief and degree of focus have complex causal relationships with system and content properties as well the other external factors, making them more difficult to model. Work in media psychology and audience studies like [91], [56] and [245] discuss some of these concepts and the nature of their relationships with each other and the source in more

detail. These concepts need to be studied more closely before they can be modelled using NLP.

For the NLP community, the work presented therefore brings to light new directions of research that have use cases within IDN. The dimensions for which no mappings could be found present completely new tasks that have not yet addressed by the NLP community - like modelling and estimating concepts like Dissonance or Presence from text. For the UX dimensions where mappings exist, NLP assisted IDN authoring provides a new use case and domain. These differences mean that existing techniques may need to be modified or adapted before they are useful. For example, generating indicative summaries to jog the author's memory is different from generating an indicative summary that helps a reader decide if they want to read the original source. IDN text is also slightly different from other types of text — it may look like a movie script or a novel or a mixture of both, has instruction-like text such as those found in games, and represents not one linear text, but a complex network of text, with multiple alternative routes that a reader might take.

From the IDN perspective, the work presented in this paper brings us closer to implementing intelligent narrative feedback and helping authors. The mappings point out many opportunities that are yet unexplored in terms of directly or indirectly leveraging existing NLP technologies to generate automatic feedback and provides worked examples that can be used as starting points to investigate to what extent existing NLP techniques can help authoring.

A limitation of this study is that the NLP review was done in an exploratory manner. Note that this means that the proportion of papers discovered for each UX dimension or whether a paper was discovered at all for a particular dimension as shown in the tables 4.2 and 4.1 may not reflect reality of how much NLP work actually exits for that UX dimension as closely as a systematic review could have done. However, these tables do show examples of work that does exist for the UX dimensions for which mappings could be easily found, revealing opportunities that could be pursued and acting as starting points for collaboration between the two fields.

4.6.1 Answering the Research Questions

Answering RQ1.b, this chapter points out many different ways in which NLP could potentially be applied to automatically generate intelligent feedback that can give insight into different aspects of User Experience. The mappings show areas of NLP research that can be leveraged to generate feedback that can give insight into the corresponding UX dimension. 3 of NLP areas are investigated further to get 5 concrete forms of feedback that can theoretically be implemented to gain insight into the corresponding UX dimension. However, deeper research is required to assess the

feasibility of implementing this type of feedback and identify the unique challenges and nuances of the domain (IDN) as use case (assisting authoring). Rest of this thesis is dedicated to exploring this for one of the identified forms of feedback - feedback in the form of summaries, by investigating challenged of applying existing summarisation techniques to IDN data and how they can be adapted to suit the domain better.

Using insights from this chapter and the last, we can now attempt to answer RQ1 - *What type of feedback has the potential to be both useful to IDN authors and feasible to generate using NLP techniques?* Automatic text summarisation, specifically generic summarisation, maps to the most number of UX dimensions. While comparative summarisation could be especially useful for IDNs, this is not as well researched within the NLP community as generic summarisation. While no existing NLP work could be found focusing on IDN summarisation, existing approaches may be applied to individual playthroughs of an IDN. These summaries can provide a concise view of possible experiences through the interactive narrative to the author, allowing them to reflect on how different affective and dramatic elements are built up and relieved over the course of the story. Providing a overview of all the ways in which a player could have reached a particular node, the author can also ensure overall continuity and understanding of the story along all the paths by making sure that all the necessary information has been set up and revealed on all the paths. Additionally, making the writing process intuitive to authors is an important concern in authoring tool design [225] and feedback in the form of summaries could be more intuitive to authors than metrics and graphs.

In the context of assisting authoring, the aim is to produce summaries that can act as a recap for the authors rather than providing a summary of the narrative to a new reader. Recaps are commonly presented as extractive summaries, for example, when providing recaps of previous episodes to viewers at the beginning of TV shows. Additionally, with extractive summarisation, there is less risk of producing inaccurate summaries. Therefore, this thesis focuses on extractive summarisation since it can present the author with potentially useful feedback without giving a false sense of certainty. In case of errors in NLP performance, the author can adjust the level of detail.

In this way feedback in the form of extractive summaries for individual playthroughs of the narrative has the potential to be both useful and feasibile. For these reasons, feedback in the form of extractive summaries was chosen for deeper investigation. The next chapter will further investigate the feasibility of applying NLP techniques to the IDN domain from a more practical perspective. Further investigation and validation of the usefulness of these summaries in the context of assisting IDN authoring will be addressed in future work described in section 7.3.2.

While summarisation is well researched within the NLP community, several challenges remain:

1. No datasets are readily available for studying IDN summarisation.
2. While there is some work methods for narrative summarisation, not much investigation has gone into summarisation approaches for interactive narratives.
3. Evaluating summaries for any domain can be challenging since there can be many valid summaries for the same text and the decision regarding what is important enough to include in a summary is subjective and can vary with context.

Chapter 5 addresses the first challenge and 6 addresses the second one. This work uses a mixed evaluation strategy using both quantitative and qualitative methods, but leave optimising the evaluation strategy for the domain for future work discussed in section 7.3.3.

4.7 Conclusion

The non linear nature of interactive narratives makes it hard for the author to picture how the user will experience the story they are writing. This problem intensifies as the size and complexity of the work increases. NLP techniques can be applied to run automatic analyses over the authored content to generate intelligent feedback that can provide the author with insight into potential user experiences offered by the interactive narrative piece. While automatic feedback to assist authoring has been proposed before [156, 217] and discussed at a high level, it is not immediately clear what exactly this feedback needs to be. Generating such intelligent feedback to assist authoring using NLP techniques has also not been explored before.

By reviewing available NLP literature related to each UX dimension, this work maps these UX dimensions to corresponding areas of NLP research, outlining ways in which existing NLP work might be applied to generate feedback and exploring what this feedback might look like. In total 15 NLP tasks were shown to map to 24 of the UX dimensions. Three of these mappings are elaborated on with 5 examples of how specifically this might be done: plotting emotional arcs, visualising emotion type and intensity, revealing the predictability of events, debugging internal story logic, and branch-wise summarization. The contribution at this stage is not to quantify or validate any one specific approach, but rather to set out the overall landscape, and draw attention to the very rich set of possibilities that arise from bringing IDN and NLP together. This gives insight into RQ 1.2 and together with insights from the previous chapter, we can see which types of feedback have the potential to be both

useful to IDN authors and feasible to generate using NLP techniques, answering RQ 1. While this work shows theoretical mappings between NLP tasks and IDN requirements that could be exploited, further research is required to understand practicality of this approaches and how they would need to be modified to work better for IDN text. Therefore, the rest of this thesis will investigate one of these forms of feedback - branch-wise summaries in more detail to understand the challenges of applying standard techniques to IDN data and how they can be better adapted for the domain.

NLP Area	NLP Sub-area	Related UX Codes	References (General)	References (In Domain)
Machine Reading Comprehension (MRC)	Causality	Effectance, Control	in common sense stories[188]	films [105]
	Contradiction detection, plausibility	Logical Consistency, Perceived Realism	lit review [213] contradiction [68] plausibility [139]	
	Next sentence prediction	Uncertainty, Expectation, Continuity	lit review [213] ending prediction [48] news stories [44]	short stories[244]
	Question Answering	Logical consistency, Narrative Understanding, Perceived Realism, Game Understanding	lit review [213]	NarrativeQA[119] mysteries, fairy tales[6]
	Coherence	Continuity	local coherence [154] discourse coherence[207]	
	Intention recognition	Objectives, Activities, Manipulation	desire fulfilment[47] dataset [180] lit review[98]	
Automatic Text Summarization (ATS)	Generic summarization	Uncertainty, Expectation, curiosity, closure, desirable outcomes, in game affect type and intensity, continuity, Narrative understanding, eudaimonic appreciation	lit review[74]	screenplays[173] novel chapters[133]
	Query based summarization, controllable summarization	Logical consistency, perceived realism, Narrative understanding,	lit review[74] controllable summarization [163]	Fandom dataset [101]
	Comparative summarization, Multidocument summarization	Autonomy, Effectance, personalization, control, variety, novelty, eudaimonic appreciation	lit review[74] multi-document summarization [9] comparative summarization [102]	
Text Classification (TC)	Emotion detection	In game affect type, intensity, variety, novelty	lit review [5, 2] in news articles [94]	novels [182] in fairy tales [4]
	Sentiment Classification	desired outcomes, identification	lit review[149]	sentiment networks[131]
	Empathy detection	Identification, desired outcomes	in mental health [203] in news stories [35]	
	Character profiling and role identification	Identification, variety, desired outcomes, novelty		in folk tales[232] fiction [81]
	Semantic similarity	Variety, Novelty	metrics for semantic similarity[60]	
	Topic modelling	Themes, Variety, Novelty	lit review [114] LDA [25]	19th century lit[106] novels [246]

TABLE 4.1: Relevant NLP Research

		ATS			TC						MRC					
		Generic Summarisation	Comparative Summarisation	Query Based Summarisation	Emotion Detection	Character Profiling	Topic Modelling	Sentiment Classification	Semantic Similarity	Empathy Detection	Question Answering	Intention Recognition	Next Sentence Prediction	Contradiction	Causality	Coherence
Drama	Variety		[102]		[182]	[81]	[106]		[60]							
	Novelty		[102]		[182]	[81]	[106]		[60]							
	Desired outcomes	[173]				[81]		[131]		[35]						
	Uncertainty	[173]											[244]			
	Expectation	[173]											[244]			
	Curiosity	[173]														
	Closure	[173]														
	Themes						[106]									
Cognition	Logical Consistency			[101]							[119]			[68]		
	Perceived Realism			[101]							[119]			[139]		
	Narrative Understanding	[173]		[101]							[119]					
	Game Understanding										[119]					
	Perceived Understanding															
	Ambiguity															
	Challenge															
	Storification															
Agency	Effectance		[102]												[188]	
	Control		[102]												[188]	
	Manipulation											[180]				
	Autonomy		[102]													
	Personalization		[102]													
	Usability															
Affect	in game affect intensity	[173]			[182]											
	in game affect type	[173]			[182]											
	at game affect intensity															
	at game affect type															
Immersion	Identification					[81]		[131]		[35]						
	Continuity	[173]											[70]			[154]
	Presence															
	Suspension of disbelief															
	Degree of focus															
	Object of focus															
	Aesthetics															
	Safety															
Rewards	Eudaimonic appreciation	[173]	[102]													
	Sense of reward															
	Learning															
	Interest															
Motivation	Objectives											[180]				
	Activities											[180]				
	to continue															
	to replay															
	to interact															
	Reinforcement															
Dissonance	Interactivity															
	Narrativity															
	Dissonance															

TABLE 4.2: This table summarises how tasks from three NLP research areas - Automatic Text Summarisation (ATS), Text Classification (TC) and Machine Reading Comprehension (MRC) map to UX dimensions in IDN.

Chapter 5

IDN-Sum - A New Dataset for Extractive Text Summarisation of IDN

5.1 Introduction

Automatic summarization has often 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. In this chapter, the first dataset for IDN Summarisation, IDN-Sum is created and applicability of standard summarisation techniques to this dataset is explored. The work outlined in this chapter was published as part of Proceedings of The Workshop on Automatic Summarization for Creative Writing at COLING 2022[185].

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 [156],

[1] www.fandom.com

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 5.3.1), 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. A manual analysis of the quality of the dataset is also presented in section 5.6.1. Novel contributions of this chapter 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 approaches on IDN-Sum and qualitative analysis of the predictions made by them to gain insight into RQ 2.a *How well do standard summarisation approaches work on this domain (IDN text)?*

5.2 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 [168]. Summarisation datasets for narratives include datasets with novel chapters and corresponding human written summaries from online guides, [49] [133], extractive summaries that read like telegraphs[150], stories and summaries from Wattpad[257], transcripts and summaries of movies[87], transcripts of TV shows [173, 51] and subtitles [10]. Papers on game summarisation are few and usually involve game logs from online games like *DOTA* [18, 52] or commentary from sports[193]. However, IDN text is typically more similar to movie scripts or novels than game logs. The critical role dataset [181] 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 [75]. Supervised methods include different RNNs and Tranformers, using pretrained models such as Bert for summarisation [162, 145]. Variations of BertSum[145], SummaRuNNer[169], MatchSum[260], Discobert[247], HiBert[258], Banditsum[72] and neusum[262] are among the most commonly used baselines for extractive summarisation in the past

IDN	N=1	N=125	N=1250
Before The Storm	42.78%	93.29%	93.49%
Wolf Among Us	30.76%	77.31%	77.40%

TABLE 5.1: Coverage of Choice Points in the first N data points of the dataset.

three years. However, most of these were designed for short documents (CNN/DM). Longformer[19] is an adpatation of BertSum for longer documents. There is also some research on summarisation approaches that are specific to the narrative domain, primarily looking at movie and TV show transcripts [87, 228, 173] however, they generally summarise at a scene-level. Recently Large Language Models (LLMs)[33] have excelled in a variety of NLP Tasks including summarisation, although as generative models, they lend themselves more readily to abstractive summarisation. Some work on applying them to extractive summarisation shows traditional fine-tuning methods outperforming ChatGPT (GPT -3.5) on several datasets [253].

5.3 Methodology for IDN Dataset Creation

5.3.1 Data Collection

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 [156]. The script accounts for the following types of game mechanics surrounding choices when generating the playthroughs:

1. Choices that have immediate consequences. These are the most common type of choice - effect relationship where the game provides immediate feedback to the choice made by the player.
2. Choices that have delayed consequences : These are choices whose effects are not revealed until later in the game.
3. Boolean combination of choices : Consequence is determined by a boolean combination of multiple choices made by the player at different points in the game.
4. Cumulative choices : Consequence is determined by the value of an internal counter variable which multiple choices influence.

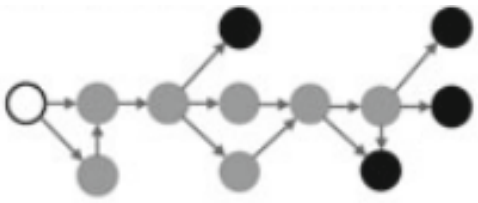

FIGURE 5.1: Gauntlet IDN Structure based on Ashwell's standard patterns[157]

5. Multiple Optional Interactions : Different options at the same choice point can be explored if the choice point. This was implemented only cases where the choice points was marked as "Optional Interactions" in the transcript.
6. Scene Order Choices: Choices that effect the order in which scenes are shown.
7. Scene level Choices : Choices that effect if entire scenes get shown.

To ensure that the ReaderBot generates a wide variety of playthroughs through the game, it keeps track of all the choice combinations it has generated previously. In each step, it makes the choice that creates a choice combination that has minimum overlap with the combinations generated so far. This was done instead of a tree-based search since the number of possible pathways is very large. The choices taken for each playthrough are provided along with the simulated playthrough in the dataset. Table 5.1 shows the percentage of choice points covered by the first N playthroughs in the dataset. [2]Note that the ReaderBot simulates playthroughs through the game capturing major aspects of the game, but it does not perfectly mimic it (for example, not all game mechanics in the game are reflected in the ReaderBot). More details regarding the Readebot's limitations are described on the Github page[3].

Both games have what are referred to as a Gauntlet structure [157] which means the story changes in some places based on player choices but sticks to an overall common storyline making a gauntlet shape as shown in figure 5.1. 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. However, even though only two games are included, together, they cover 6 out of 8 choice idioms described in [243] including blind choices, dilemma, false choices, flavour choices, delayed effect and puzzle choices.

[2]The comparatively lower percentage for Wolf Among Us is due to multiple choice options indicating the default option having the same label. Since the sampling algorithm tries to maximize variance they were skipped.

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

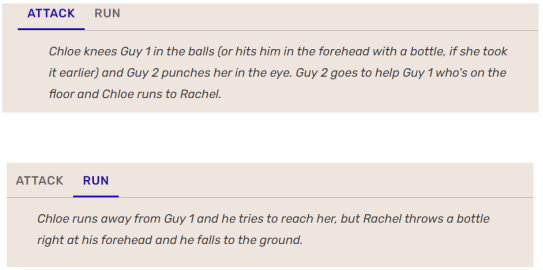

FIGURE 5.2: Example of Choices shown on Fandom

5.3.2 Automatic Annotation

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. An example can be found in figure 5.2. This html page is parsed and different playthroughs are then generated by a ReaderBot[156] 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[4]. Through the authors' own inspection, the summaries were found to be of good quality.

There is only one human authored abstractive summary per episode. This overall abstractive plot summary is taken from Fandom and extractive summaries for each playthrough is produced using the TransformerSum[5] library. This library follows the method used in [169] 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 sentences (similar to CNN/DM) but also longer target lengths of 9 and 27 sentences, since for narrative datasets the source text and reference summaries are much longer. For IDN and CRD3, target length of 81 is also generated since the reference summaries for the these datasets are considerably larger than 27. The human authored abstractive summary for each episode is also provided along with the dataset so that annotations can be generated using any alignment algorithm.

5.4 Dataset Characteristics and Comparison

Table 5.2 compares **IDN-Sum (IDN)** with several other narrative datasets. The Novel Chapter dataset from [133] is included since it contains narrative elements like plot

[4]stats taken from https://community.fandom.com/wiki/Special:Statistics
[5]https://github.com/HHousen/TransformerSum

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 5.2: 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

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 5.3: 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

but is not as structurally different from the CNN/DM as the screenplay datasets. **Scriptbase (SB),graphsum** was chosen for comparison because the IDN text that is generated by the ReaderBot is very similar to screenplays. **Critical Role Dataset (CRD3),critical** 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 [168] 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).

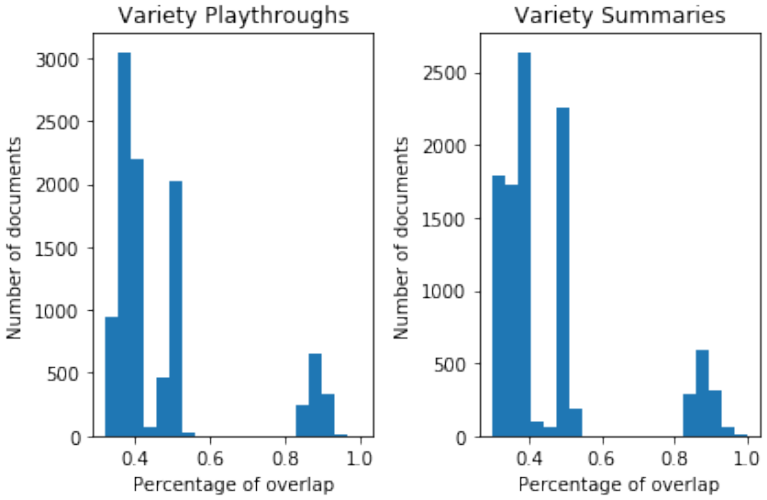

FIGURE 5.3: Variety in IDN Dataset

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 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 5.3. 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 5.3 and 5.4. Fig 5.3 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. This forms a trimodal distribution. 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. There is lesser overlap with playthroughs from the other two episodes. For comparison, a similar graph is shown from ScriptBase which contains screenplays that are entirely unrelated to each other in fig 5.4. In this case, all data points have only a small overlap and forms a normal distribution. Examples of the data are shown in Appendix A.1.

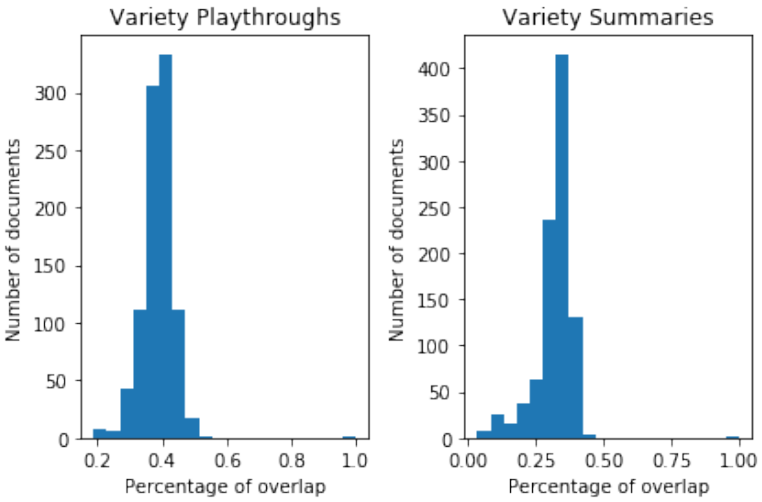

FIGURE 5.4: Variety in Scriptbase Dataset

5.5 Baseline Experiments

5.5.1 Methods

Baseline models used in this paper represent a sample of commonly used baselines for extractive text summarisation as of January 2022 when this experiment was performed to contextualise the dataset using well known summarisation techniques. The baselines were chosen so that they include two simple baselines, Random-N and LEAD-N, a commonly used graph-based unsupervised method, TextRank [155] and neural network based methods (Pretrained Language Model(PLM) based approaches BertSum[145], Longformer[19] and an RNN based sequence model, SummaRuNNer[169]). Out of the popular baselines mentioned in section 5.2, SummaRuNNer was included because it was the most easily extendable to longer documents whereas most of the later transformer based approaches are architecturally constrained to a given sequence length. This is also a widely used baseline for extractive summarisation [95, 93]. Pretrained Language Models (and more recently in 2023, Large Language Models (LLMs)) have been showing remarkable performance across various NLP tasks. BertSum was included to represent PLM based approaches since this was the most popular baseline used in extractive summarisation papers[192, 95, 130] and state of art approaches like MATCHSUM [260] build on this base model. However since BERT (and most open source LLMS) have a sequence limit of 512 tokens, a variation of it for longer documents, Longformer was included so that a more recent model designed for longer documents is also included as a baseline. At the time of experimentation, the largest token limit for which a pretrained Longformer model for extractive summarisation was available was 4098. Narrative summarisation models mentioned in section 5.2 work at a scene level and hence return huge summaries for complete narratives/IDN's, so these methods are not included. Note

that LLMs are not included in this chapter since they emerged after the completion of this experiment. The next chapter includes baselines that use LLMs. Other recent advancements that can be applied to IDN summarisation are discussed in Future Work section 7.3.1.

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[171] 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 [169]. 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. Results on two variations of SummaRuNNer are reported - 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, results on LongFormer for these is reported 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[260] and DiscoBert[247] uses an underlying Bert model, they suffer from this limitation as well and hence, were not included as baselines.

5.5.2 Experiment Setup

For TextRank, Gensim[6] library's implementation was used. For BertSum, TransformerSum library's[7] implementation of BertSum and LongFormer was used. At the time of running experiments, this implementation of LongFormer supported upto 4096 tokens as input. For SummaRuNNer, the implementation from hpzao [8] was used.

The first three episodes of *Wolf Among Us* was used as the training set, the last two episodes of *Wolf Among Us* was used as the validation set and *Before the Storm* was used as the test set. Using a different game for the test set ensures that there is no data

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

Dataset_ Length	RAND-N	LEAD-N	Text Rank (TR)	Bert Sum (BS)	Summa RuN-Ner (SR)	Long For-mer (LF)	Summa RuN-Ner 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 5.4: 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.

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 5.5: ROUGE1 F1 scores against automatically aligned extractive summary

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 default 100). SummaRuNNer originally truncated documents at 100 sentences. 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 are reported 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 the target length of 3 sentences 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.

5.5.3 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[9]. ROUGE1 F1 score is calculated against the human authored abstractive summary with porter stemming (as commonly done in papers such as [3]) for all models and datasets and is compared in Table 5.4. Table 5.5 shows ROUGE1 scores computed against automatically aligned extractive summaries. ROUGE2 F1 scores are shown in Appendix B.1. The best and worst summaries (according to ROUGE-1) from the best model were also analysed qualitatively. The qualitative investigations help assess aspects of quality that cannot be captured by the ROUGE scores.

5.6 Results

Table 5.4 shows the performance of the baseline models. SummaRuNNer appears to scale 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). A manual inspection of sample summaries was performed and the results of this analysis are discussed below.

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

5.6.1 Quality of aligned extractive summaries

The ROUGE1 F1 scores of the automatically aligned extractive summary overlap to human authored summary is shown in table 5.3. 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 for SB and IDN, 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 5.3 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 [133]. 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.

In the case of IDN summaries, while they capture the overall plot well, the choices made for the main choice points in different branches are not always obvious. This is mainly because the variation is not captured in a lot of depth in the human-written abstractive summary and when it is captured, they are heavily paraphrased and condensed. By running the automatic alignment, branch-specific events are captured in the extractive summary for each document in some cases, but many were missed. However, while these summaries have room for improvement, they provide a reasonable starting point for investigation into IDN summarisation. An example oracle summary for IDN is shown in Appendix A.2.

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 5.6: Analysis of best and worst ROUGE1 scoring generated summaries by SRL model. 'relevant' shows ratio 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 ratio of sentences in ground truth abstractive summary that match sentences in the generated summary.

5.6.2 Quality of Summaries from Best Model

Automatic metrics to evaluate summarisation is known to have many limitations [76]. 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 5.6. 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 5.6. 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 5.4. 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 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 5.5 shows an example of a high quality snippet of model summary and fig 5.6 shows and

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 5.5: 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 5.6: Example of low quality extract

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.

An example is shown in Appendix A.3 . Overall, it is hard to discern aspects of user experience discussed in the previous chapters and how they might vary between playthroughs from these summaries due to the low coverage. The summary quality would need to improve significantly in terms of both relevance and coverage for it to be useful in terms of assisting IDN authors.

5.7 Challenges

5.7.1 Document Length and Long Range Dependencies

One major limitation when it comes to narrative datasets compared to CNN-DM is the document length. Most state-of-the-art summarisation approaches take advantage of transformer-based Pretrained Language Models. However, most pretrained models

have a context limit of 512 tokens. This limits us from directly applying these to IDN-Sum. PLMs adapted for long documents like Longformer have context lengths up to 4096 which comes close to the document length for novel chapters, but is still much smaller than the other narrative datasets. Applying simple methods of handling this like splitting the documents into chunks of 200 sentences and then summarising each chunk and putting them together, the ROUGE-1 score for Longformer can be improved to 0.45. But since the models can only see one block of text at time when summarising this way, it cannot account for long range dependencies between sentences and events that are common in narrative text. However, this is a fast moving area with many different techniques to handle this being applied and investigated[123]. As of 2023, commercial LLMs like GPT-4 are capable of considering larger context windows. Many open-source LLMs are also limited by the context window, however LLMs like MPT[224] are capable of handling longer context windows. However, processing narrative-length documents with long-range dependencies remains a challenging problem [144]. Many of the long-range dependencies that IDNs would have are related to choice points and consequences, and unlike traditional narratives, these are clearly marked with a "CHOICE" tag. In the next chapter, a training strategy that leverages this knowledge is experimented with where the model is trained to give more attention to these parts of the text when deciding which sentences to include in the summary. In Chapter 7, other approaches for handling long documents and long-range dependencies (not specific to interactive narratives) are discussed.

5.7.2 Nature of Interactive Narrative Text and Low Variation in Data

In narratives, the model needs to understand causal and temporal relationships between events and the overall development of the plot. Sentences describing the main events (events that drive a narrative thread forward) as well side events related to them tend to be included in the summary and events not related to the main events tend to be excluded [259]. This means that the models could benefit from considering each sentence in the context of the sentences that describe major plot points to decide whether they should be included in the overall summary or not. In interactive narratives, in addition to the narratives elements like plot and characters, interactive elements like choice making also influence which sentences are included in the summary. Therefore, considering sentences in the context of text surrounding interaction points (choices made by the player and their consequences) could help interactive narrative summarisation.

Given enough examples of different interactive narratives, supervised models combined with approaches for handling long documents as discusses in the last section may be able to make such connections and figure out these patterns. However,

while our datasets contain 10K playthrough documents, it only covers 2 IDN games. While parallel research on game playing AI can help making readerbots easier, currently, manually implementing readerbot for each game with unique mechanics is not easy. Therefore, we explicitly leverage domain knowledge regarding the importance of choice points through rationale based learning in the next chapter to improve summary quality.

Additionally while our dataset contains 10k docs, it contains only 26k unique sentences which is much lesser than the other narrative datasets. Most of the variation after first 125 playthroughs is in combination of content units rather than new content. This makes models prone to overfitting and sensitive to hyperparameters when training on IDN data. SummaRuNNer improves to 0.4778 for IDN with more frequent validation rounds (30 steps) and early stopping (if no improvement is observed in 5 validation rounds). The next chapter experiments with focussing training on regions of text that are most different across playthroughs - the text around choice points and their consequences.

5.7.3 Oracle Summaries

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[133] might do better in this regard. The next chapter experiments with using an additional supervision signal in the form of rationales indicating proximity of words to choice points and their TF-IDF scores to guide training. Further experimentation with other alignment techniques and extract granularity is left for future work.

5.7.4 Evaluation Strategy

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 Table 5.5.

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 [51] where new entity centric evaluation metrics are proposed.

Finding a good evaluation metric to assess summary quality is a known challenge, even in case of the CNN/DM dataset[76]. 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 [135]. 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 this type of evaluation. This research follows a mixture of automatic evaluation using ROUGE score and qualitative human evaluation with examples of generated summaries to showcase the more intuitive aspects following evaluation approach used in recent extractive summarisation research[223, 192]. However, the manual nature of the qualitative analysis limits the amount and scale at which this can be done.

5.8 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.

The dataset has 1250 playthroughs per episode and 8 episodes overall, but the code and JSONs for the ReaderBot has also be made available on GitHub[10]. 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. For games having a similar branching gauntlet structure like Life is Stange I, II and III, only small modifications to the script will be required to account for additional game mechanics and changes in parsing the html page for the script, but for games with completely different narrative design, a similar approach may be used, but it will require designing and implementing a corresponding ReaderBot. This work also reports and analyses performance of some standard baseline approaches quantitatively and qualitatively.

This chapter applied standard summarisation approaches to each linear playthrough of the IDN. But developing methods to create an overall summary of the entire IDN will take further exploration and research. This includes requirements like capturing the important differences between different playthroughs which is a very significant aspect of summarising interactive narratives. 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[9] has been studied, IDN documents have a narrative structure and elements (plot, protagonist, emotions, etc) which influence the relative importance of sentences. Unlike narrative text, in addition to these elements, IDN also has an interactive structure (choices, consequences, etc) which also influence the relative importance of sentences. The nature of differences between documents is different from domains like academic papers where comparative summarization has

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

been studied[102]. 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.

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 4) model architecture and training methodology. Hopefully, this dataset can help aid future research in these directions.

5.9 Conclusion

In this chapter, the first summarisation dataset for interactive narratives is presented. This was done by collecting fan made transcripts and abstractive summaries from Fandom and generating simulated playthroughs by assuming different combinations of choices. Standard summarisation approaches were then applied to this dataset to answer RQ 2.1 - *How well do standard summarisation approaches work on IDN data?*. 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. This dataset is intended to facilitate future research into improved annotation methods, evaluation strategies, and summarisation approaches for interactive digital narratives.

Chapter 6

Rationale-based Learning for IDN Summarisation

6.1 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[173], 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 is addressed in this chapter.

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 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. This chapter investigates leveraging this knowledge regarding the importance of choices to enhance IDN summarisation.

To incorporate this knowledge into the training process, this chapter explores for the first time, choice-focussed rationale-based learning for extractive summarisation of IDN. Our approach is motivated by the text classification model of [110], 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, this chapter explores 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 chapter 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. To the best of our knowledge, this work is the first to explore using rationale-based learning with automatically generated rationales for any type of narrative summarisation. This workhas been published at LREC/ COLING 2024.

The novelty of this approach is in the formulation of the data and training objectives for this unique domain (IDN). In this chapter, the efficacy of this approach on variants of the classic SummaRuNNer model and Google's flan-t5 model is investigated. The 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 work are as follows:

1. A novel method using word and sentence level rationales applied to existing models 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 limitations to guide future researchers in the area of IDN summarisation.

Related work is reviewed in section 6.2 before outlining the rationale-based training approach and the models trained in section 6.3. Section 6.4 reports results from automatic and manual evaluations and analysis of variability of generated summaries across different playthroughs of the same interactive narrative, which is discussed and concluded in section 6.5 and section 6.6.

6.2 Related Work

Previous studies on extractive summarisation have focussed on various techniques including RNN-based models [169], language model-based methods [145] and graph-based methods [9]. However, these methods are most commonly trained and tested on datasets like news [103] and academic articles [97]. While some approaches for summarisation of traditional narratives have been explored, like using GCNs for screenplay summarisation [135] and taking turning point information into account [173], summarisation of interactive narratives has not been explored in much depth. IDN-Sum [185] presented in the previous chapter is a dataset introduced for studying interactive narrative extractive summarisation and is used for the experiments in this chapter. 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 [83]. This has been applied in a variety of NLP tasks including Text Classification [11, 55], Natural Language Inference [37, 210] and Sentiment Analysis [261]. Both local explanations [84] and global explanations have been applied to guide training [143] in this way. Rationales are incorporated into training through various means including supervised attention [110], which is the approach used in this chapter. However, in this chapter, the effectiveness of choices as rationales is investigated in the novel context of summarising IDNs as well as different kinds of explanations applied at both word and sentence levels.

6.3 Method

6.3.1 Choice Focussed Rationales

Information regarding the importance of choices in IDN summarisation is introduced into the training process through rationales that indicate the proximity of words and sentences to choice points. In IDN-Sum dataset[185] used in these experiments, 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 [110], 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: $\alpha + \alpha_1 + \alpha_2 = 1$,
L = Total Loss,
L_l = Cross-entropy loss calculated for the output of the model against the target labels,
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.

6.3.2 Base Models

While this training approach could theoretically be applied to any model with an attention layer, introducing supervised attention to multi-head attention introduces additional layers of complexity (eg. how many and which attention heads do we align with the rationales). Therefore this chapter first tests this approach on simple attention layers, saving other attention types for future research.

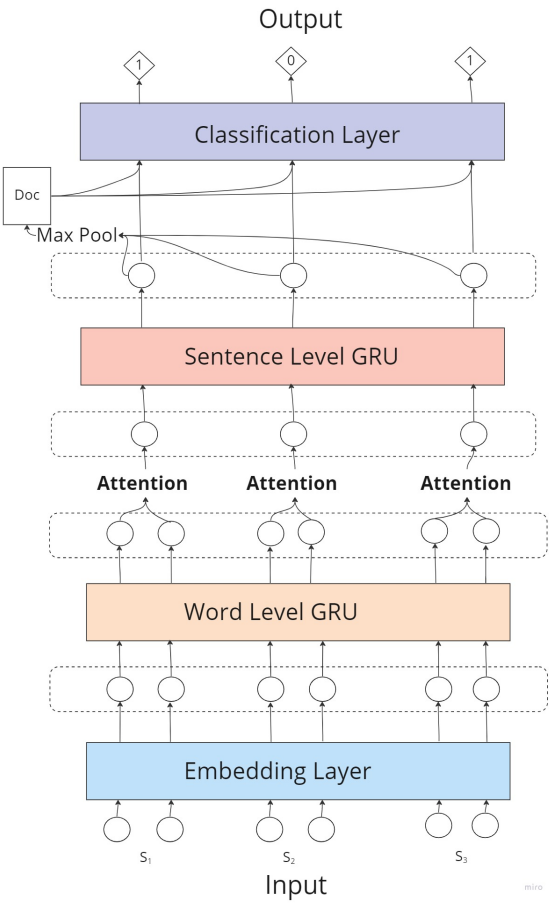

FIGURE 6.1: SummaRuNNer modified to use attention instead of max pooling at word level (wordonlyAttnRNN).

In the experiments in this chapter, models based on SummaRuNNer and Google's flan t5 model are utilized. SummaRuNNer is an RNN-based model for extractive summarisation with simple attention layers added to it. SummaRuNNer was chosen as one of the base models because of its superior performance on the IDN-Sum dataset, outperforming even more recent models like Longformer[19] on this dataset[185] 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. Additionally, the heirarchical architecture of this model lends itself readily exploration of rationales at both word and sentence level. The model referred to as RNN, in this chapter, represents the original architecture used in SummaRuNNer, modified to truncate documents at 3000 sentences instead of 100.

Originally, 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

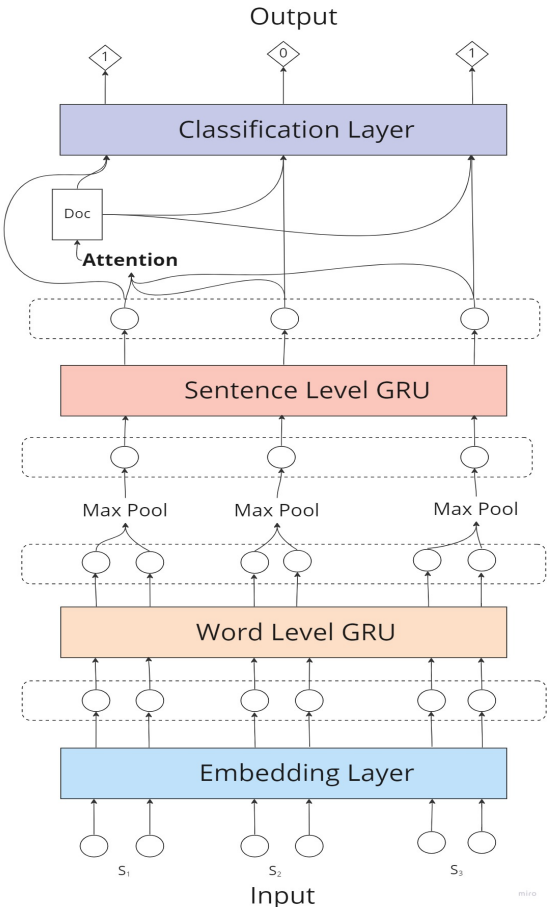

FIGURE 6.2: SummaRuNNer modified to use attention instead of max pooling at sentence level (sentonlyAttnRNN).

pool is replaced with attention layers at different levels in the following three ways, inspired by Hierarchical Attention Networks (HAN) [250] 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 6.1. 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 6.2. The third modified architecture is **AttnRNN**, modelled after Hierarchical Attention Networks [250], which uses attention at both the word and sentence level and is illustrated in Figure 6.3. 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** represents AttnRNN trained with both rationales. All these models have approx 1.7M parameters.

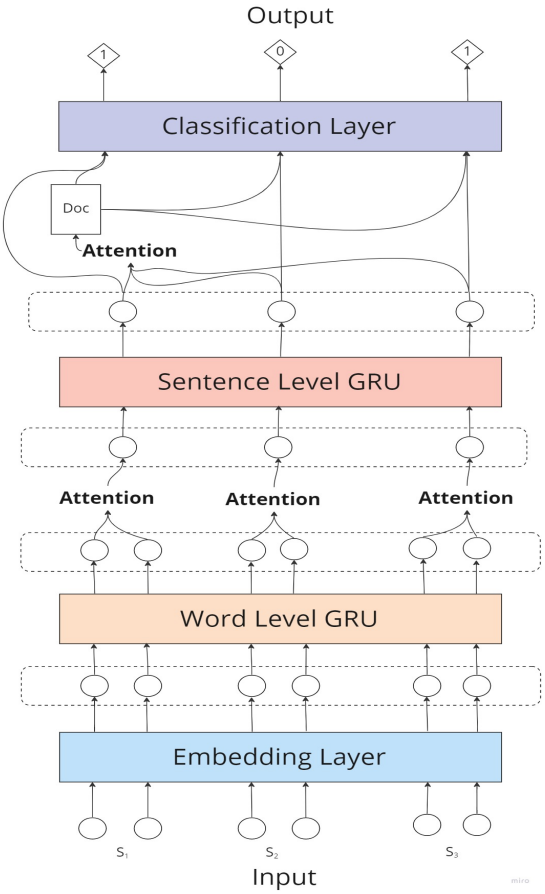

FIGURE 6.3: SummaRuNNer modified to use attention instead of max pooling at both word and sentence level(AttnRNN).

Given the recent success of Large Language Models (LLMs), this approach is also tested on variants of Google's flan T5 base model[57]. Google's flan series of LLMs are pretrained, instruction tuned Encoder-Decoder models. This was chosen as a base model for this experiment since even smaller versions of these models show superior performance on a wide range of NLP tasks. The experiments in this chapter uses the base version of the model which has 250M parameters. The decoder layers were replaced with a simple attention layer and a classification head as shown in figure 6.4 to adapt the model to extractive summarisation and enable application of rationale based learning. Layer norm was added to stabilize training. A significant limitation of many pretrained Language models (PLMs) including flan T5 models is their fixed context length (512 tokens) whereas IDN-Sum has an average document length of 22,900 tokens. Therefore documents were chunked and processed in chunks of 25 sentences at a time both during training and inference. This model is referred to as **Google flan-t5-base Encoder** in Table 5.4 and the version of this model trained with word level rationales is referred to as **Google flan-t5-base Encoder + rationale**. Sentence level rationale is not applied to this model since unlike the SummaRuNNer model, sentence level attention is not applied to create a document representation.

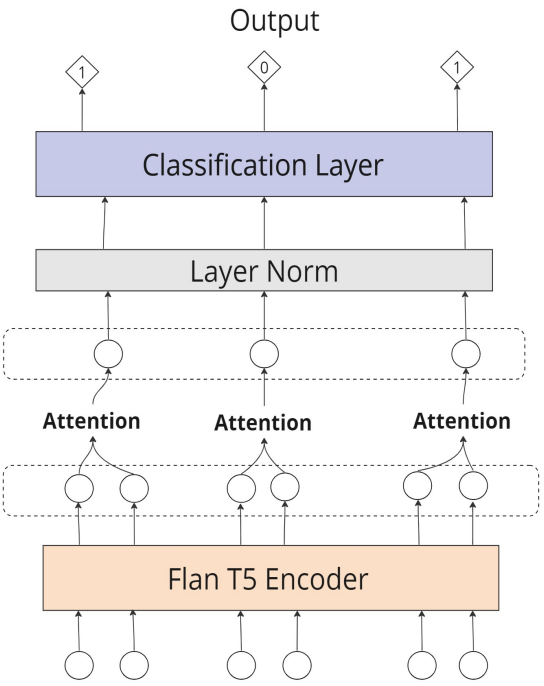

FIGURE 6.4: Flan T5 Encoder with an attention layer and classification head for extractive summarisation.

6.3.3 Experiment Set Up

6.3.3.1 Dataset

The IDN-Sum dataset [185] 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.)

6.3.3.2 Models

Implementations provided on Github [1] were used as the starting point for the modifications described in section 6.3.2 for SummaRuNNer. 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, the performance of a few different variants of LLM-based approaches using Google's flan-t5-base model[57] (instruction tuned LLM with 250m parameters) is also shown for comparison. The pretrained model was downloaded from Huggingface[3]. Summaries are generated 25 sentences at a time, to fit the context window and strung together at the end to get the final summary. LLMs lend themselves more readily for abstractive summarisation. Therefore, a version of the model where the decoder is replaced with an attention layer and classification head as described in section 6.3.2 is also used. The source code for this model will be made available on GitHub[4]. This version of the model is trained with and without rationales since it has a simple attention layer which can be aligned with rationales using the rationale based learning approach. ROUGE scores for the original model in a zero shot setting and under PEFT fine tuning is also reported for comparison even though they do not produce summaries that are strictly extractive. For the zero shot setting, 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:"* For the fine tuned version (Google flan-t5-base FT), this model was fine tuned on the training data from IDN-Sum using Low Rank Adaptors (LORA). Zero shot performance for Flan-t5-large (0.8B parameters) is also reported for comparison. Refer to Appendix C for more training details including the full list of hyperparameters and hardware details.

[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] https://github.com/AshwathyTR/IDN_SR

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

[4] https://github.com/AshwathyTR/idn_flan_exps

6.3.3.3 Evaluation

ROUGE-1(R1), ROUGE-2(R2), and ROUGE-L(RL) were used to compare the performance of different summarisation approaches. 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 against the branch-wise extractive summaries and ROUGE scores calculated with and without the stop word filter is be shown in Appendix D.

Some studies rely solely on ROUGE for comparing summarisation approaches[260, 251, 65]. 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[222], examples of the model-generated summaries and reference summaries are provided for human evaluation in Appendix E and perform a qualitative analysis to compare and illustrate intuitive aspects of quality that the ROUGE-based evaluation is unable to capture. The manual analyses were performed by one person (the author) who has background in NLP and IDN research but is not an IDN author.

The 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. The variation between summaries generated by the model for different playthroughs through the same episode is analysed 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, a manual fault analysis is also performed to understand the limitations of the 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.

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
Google flan-t5-base (zero-shot)	0.46577	0.46519 - 0.46637	0.11833	0.11800 - 0.11866	0.41051	0.40997 - 0.41112
Google flan-t5-base FT	0.37972	0.37885 - 0.38049	0.08504	0.08479 - 0.08530	0.32296	0.32208 - 0.32376
Google flan-t5-base Encoder	0.44885	0.44794 - 0.44969	0.10322	0.10280 - 0.10364	0.40304	0.40209 - 0.40393
Google flan-t5-base Encoder + rationale	0.47444	0.47398 - 0.47490	0.11660	0.11612 - 0.11707	0.42987	0.42940 - 0.43031
Google flan-t5-large (zero-shot)	0.49386	0.49291 - 0.49484	0.13068	0.13038 - 0.13099	0.43159	0.43064 - 0.43252

TABLE 6.1: 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).

6.4 Results

6.4.1 Automatic Evaluation

Table 6.1 shows the ROUGE scores calculated against the human-authored abstractive summary. 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 calculated against the automatically aligned extractive summary is also provided in Appendix D.

The rationale-based models outperform those trained without rationales both in case of both SummaRuNNer and Flan-t5. This indicates that choice focussed rationale-based learning can improve the performance of summarization models for IDN. The SummaRuNNer based model that incorporated rationales at the sentence level (sentonly AttnRNN + rationale) shows the best performace when ROUGE score

[5]https://github.com/AshwathyTR/IDN_SR/tree/master/outputs

is 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 original RNN model. Flan T5 encoder only model trained with rationales shows around 6% improvement compared to the same model trained without.

6.4.2 Human Evaluation

The best and worst scoring summaries from Episode 1 of Before the Storm from each model were reviewed manually to get an understanding of subjective aspects of quality that automatic metrics are unable to capture. All the output summaries are made available on GitHub and examples of summaries from each model are shown in Appendix E for reference.

For the flan-t5 models, despite the instruction to generate extractive summaries, in the zero shot setting, both flan-t5 models (base and large) tended to paraphrase the sentences from the original text and produced many hallucinations. Fine-tuning helps generate summaries that are extractive most of the time, leading to fewer hallucinations, however this is not always the case. The encoder only models on the other hand, produce fully extractive summaries by design.

In case of SummaRuNNer variants, 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 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 slightly 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 (sentonly AttnRNN + rationale) with the best ROUGE score from each episode were also analysed qualitatively to get inisght into their potential utility. To someone reading the summary without any other context, the summary gives a vague, fragmented view of what is going on. 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. For example, from a sample in the first episode, the following extracts were included from the first scene:

"chloe price, standing on train tracks and wearing a black hoodie, flicks her lighter a few times and lights up her cigarette. a train begins to approach her. after a few moments, the guy she ran into earlier and his friend come to confront her. rachel takes chloe's hand again and they run towards the entrance to the show. frank sees them and chloe stops, looking at the guys behind him. the men leave and frank looks back to see that rachel and chloe are gone. if she attacked the skeevy guys, she will now have a bruise under her eye."

Through these extracts, the summary conveys that there was a fight with two skeevy guys, Chloe had the option to attack them and that Chloe and Rachel were chased by them and Frank saw them. The information that Frank intervened to save them is only very subtly hinted at. Someone looking for a recap from the summary may be reminded of this from the extracts, but this information is not directly conveyed. Additionally, context information, for example, information setting up characters, setting and scene transitions, is not included, reducing the readability of the summary and giving an incomplete, fragmented view of the plot to someone not already familiar with the story. Some important plot elements like Chloe and Rachel having a fight at the junkyard later in the story are completely missed and the summary has no extracts related to this event.

Even with increased attention to choice points, there is still room for improvement in the coverage of choices. In episode 1, some choices that the player (as Chloe) takes are represented in the summary (for example, her response when asked if Miranda was feeling true love, and some of her answers in the Two Truths and a Lie game). However, her responses to many of the choice points are unclear from the summary (including whether she attacked the skeevy guys, whether she was compassionate towards her mother, stepped in when Drew was bullying Nathan, or whether she felt something more than friendship for Rachel) and these are important aspects of the overall narrative.

6.4.3 Variability Analysis

Table 6.2 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
Flan-t5-base Encoder only	66.95
Flan T5 Encoder only + rationale	65.68

TABLE 6.2: Average number of overlapping sentences for every pair of summaries from each episode for each model (out of a total of 81 sentences).

6.4.4 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 6.3.

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

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 6.3. Redundant sentences, sentences having incomplete information and irrelevant sentences are more prevalent than unclear sentences, but these three errors are similarly prevalent.

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

6.5 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 [115], turning point identification [172] 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 only one type of IDN was available to use in the study. The effectiveness of this approach on other types of IDN is yet to be

determined. This chapter has tested rationale based learning using choice based rationales on two types of models (SummaRuNNer and Flan T5 variants) with supervised attention applied to simple attention mechanisms that were added to these models. While this approach can be theoretically be applied to other model architectures and other types of attention, testing them empirically is outside the scope of this thesis.

The results reported are for single runs with specified hyperparameters. While default values were for most hyperparameters in case of SummaRuNNer, it is worth noting that IDN-Sum has many differences from datasets like CNN-DM on which 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.

6.6 Conclusion

Choices are central to interactive narratives and this chapter has explored choice focussed self-supervised rationale-based learning at the word and sentence level to improve IDN extractive text summarisation. This experience developing better summaries for IDNs using rationale based learning 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. Answering RQ 2.2, this chapter shows one way of adapting existing summarisation appraoches to better suit IDN data. 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.

Chapter 7

Conclusions

7.1 Summary

Interactive Digital Narratives (IDNs) are narratives that allow players to interact with the story. Authoring Interactive Digital Narratives (IDN) is challenging because past a certain size it becomes hard to keep track of the user's experience along all the possible story permutations. Providing intelligent feedback to aid authoring has been proposed as a way to speed up authoring, give the author more control, and to enable the authoring of more complex interactive narratives.

However, there is little research investigating what concrete feedback items would be useful for interactive digital narrative (IDN) creators. This motivates the first Research Question *"What type of feedback shows potential in terms of both impact and feasibility?"*. Since ensuring a good experience for the user is often emphasised as the primary concern of authoring[164], this is used as the starting point of investigation into potentially useful forms of intelligent feedback to assist authoring. However, UX in IDN is a broad concept that is described and used in many ways. This motivates the first part of RQ 1 - *"What aspects of user experience is the author interested in?"*. To answer this question, Chapter 3 performs a systematic literature review to make a list of concrete feedback items of interest related to the most emphasised concern of authoring - the effect of the interactive narrative on the user. 47 User Experience (UX) dimensions in the IDN literature are identified that could serve as useful feedback items, covering 8 categories - Agency, Cognition, Immersion, Affect, Drama, Rewards, Motivation and Dissonance. This list combines and untangles how different IDN researchers have interpreted and expressed interest in the complex idea of UX in the past decade and gives us insight into what concrete aspects of UX might be useful to estimate via automated feedback.

Natural Language Processing (NLP) provides us with an opportunity to generate intelligent feedback such as those identified above. This motivates the second part of

RQ 1 - *"How can NLP techniques be applied to generate feedback that can give insight into these aspects of player experience?*. In chapter 4, evidence for this is provided by mapping User Experience (UX) dimensions in IDN to NLP tasks which could be used to gain insight into them automatically. 24 UX dimensions were found to map to 14 NLP tasks. 5 specific examples are then derived from these mappings to show how they could be applied to authoring tools: visualising emotion (intensity and types), calculating predictability of events, debugging internal story logic and branch-wise summarization. This work reveals new opportunities for research, highlighting an unexplored problem space for NLP researchers and acting as a signpost for the future integration of NLP into authoring tools for IDN.

Together the outcomes of these experiments answer RQ 1 - *"What type of feedback shows potential in terms of both impact and feasibility?"*. The codebook from Chapter 3 shows which concrete aspects of UX have the potential to be useful to authors if shown as feedback (impact) and the mappings and example feedback items from Chapter 4 show NLP techniques that might be used to estimate them automatically (feasibility). One of these potential forms of feedback, specifically, branchwise summarisation, is then chosen for further investigation since automatic summarisation is well researched within the NLP community and feedback in this form can give insight to many important UX dimensions.

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. This motivates the first part of RQ 2 - *"How well do standard summarization approaches work for this domain (IDN text)?"*. In chapter 5, 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.

Choices and consequences to choices are often central to determining relative importance of sentences in IDNs, however, existing extractive summarisation approaches do not take this into consideration while summarising. This motivates the second part of RQ 2 - *"How can these be modified to work well for IDN?"* In chapter 6, we explore using explanation guided training based on supervised attention to place special focus on words and sentences surrounding choice points when creating summaries. We experiment with using word level and sentence level explanations indicating proximity of words and sentences to choice points. Our results show that

such explanations can improve performance of models, at the same time, producing summaries that have more variability across different playthroughs of the same interactive narrative. This suggests that choice point annotations can act as effective explanations for extractive summarisation of IDN.

Together, by showing types of potential feedback items that could be generated, assessing how well standard approaches work when applied to generating one of these types of feedback (automatically generated summaries) and modifying these approaches to account for choice-making, a type of interaction common in IDNs, these findings shed light on the overarching research objective - *"How can NLP be used to generate intelligent feedback to assist authoring of IDN?"*

7.2 Impact and Applications

1. UX Codebook : The Codebook shown in chapter 3 brings together and untangles different interpretations of UX in the IDN literature, resulting in a list of 47 feedback items. This provides insight into the relative interest and usefulness of modeling different dimensions of UX in the IDN community and offers a starting point for generating automated feedback for IDN authors to assist in authoring interactive narratives. It also helps in giving a broader and more complete understanding of UX for IDN. It can act as a reference for researchers and practitioners working in the field of interactive narrative design (IDN) to explore how these dimensions can be used in evaluation to give a more complete understanding of UX for IDNs. This was published at ICIDS 2020[184].

2. UX-NLP Mapping heatmap : Chapter 4 provides a framework for mapping opportunities in NLP to requirements in IDN, unlocking many avenues for research by connecting NLP and IDN. It bridges the research communities of NLP and IDN by mapping theoretical problems that are of interest to both communities and revealing the untapped potential in applying NLP to generate automatic feedback to assist authoring in IDN. It provides a new use case and domain for the NLP community. It also highlights a lack of related NLP literature for some UX dimensions, illustrating the need for further research in these areas. This work is under review at IEEE Multimedia.

3. Case studies and sample feedback items : Chapter 4 lists concrete examples of possible feedback items that can be generated using NLP methods and highlights the need for modifications and adaptations of NLP techniques to fit the IDN domain and specific use cases. This brings the IDN community closer to implementing intelligent narrative feedback and helping authors and highlights the opportunities for directly or indirectly leveraging existing NLP technologies to generate automatic feedback for IDN authors.

4. IDN summarisation dataset and open source scripts to simulate playthroughs of the games from the transcripts on Fandom. : The dataset is released as open source for future researchers to train and test their own approaches for IDN text. The dataset can be used to investigate summarization approaches for interactive and game narratives. It can also be used to study how summarization models respond to small changes in text and target summary and new NLP problems such as comparative plot summarization. This was published as part of Proceedings of The Workshop on Automatic Summarization for Creative Writing at COLING 2022.

5. Baseline performance of standard summarisation approaches on IDN data. Chapter 5 applies standard summarization approaches to each linear playthrough of the IDN and the analyses their performance quantitatively and qualitatively. This reveals several challenges some of which are common to long, narrative texts in general and some specific to the interactive narrative dataset, encouraging further investigation into interactive narrative summarisation. This was published as part of Proceedings of The Workshop on Automatic Summarization for Creative Writing at COLING 2022.

6. Novel IDN Summarisation approach : Chapter 6 is one of the first to apply explanation-based learning to interactive narrative summarization. The results show that using rationales in training can lead to improved performance in terms of ROUGE scores. The analysis of error types in the model generated summaries provides insights into the specific areas where the model can improve. The work also provides a comparison between models with and without rationales, and with different level of rationales. The results of this study can be used to guide future research on interactive narrative summarization and explanation-based learning in other tasks as well. The ability to control how much emphasis is placed on choices while generating summaries through the use of different values for alpha and window sizes can be used to generate summaries with more variability, highlighting the important differences between each branch of the interactive narrative. Understanding the most common error types for different models (incomplete information errors, irrelevant information errors, redundant sentences, and lack of coverage of important information in the reference summary) can help guide future research and development in the field of interactive narrative summarization. Fault analysis and qualitative analysis outlines in this chapter also illustrate limitations of the approach. This work has been published at LREC-COLING 2024.

7.3 Future Work

7.3.1 Improvements to Narrative and Interactive Narrative Summarisation

In Chapter 5 some commonly used standard summarisation techniques were applied to get a baseline performance of IDN summarisation. Rather than trying to find the best possible model architecture for IDN, this thesis has focused on investigating narrative rationales that could be applied to any attention based architecture to improve summarisation. NLP techniques for automatic summarisation is a fast moving research area with new techniques for generic summarisation, long document summarisation and narrative summarisation being proposed and experimented with - many of which could work well on IDN text as well. Chapter 5 includes an popular unsupervised approach (Textrank), RNN based appraoch (SummaRuNNer) and Language Model based approaches (BertSum and Longformer). A number of graph based and reinforcement learning-based approaches have also been applied to summarisation which are not included as baselines in this thesis. Some promising alternate avenues of exploration for IDN summarisation are outlined below:

7.3.1.1 Graph based approaches

Considering the extensive use of graphs to capture long term relationships for plot summarisation[87] and use of graph based architectures for summarising long texts[135] and for multidocument summarisation[9], graph based Graphs have been used in several ways. [10] uses sentence similarity graph to get centrality and use that as an indicator of salience. [228] uses character networks to identify important characters and use their presence in a scene as indicator of salience. [87] uses a bi-partite graph of characters and scenes with edges indicating sentiment,interactions and presence of a character in a scene. These show that graphs are useful in capturing saliency information but all of these are used along with heuristics based methods. Graphs have also been used for screenplay summarisation at scene level [135]. Graph based neural networks have not yet been experimented with for sentence level narrative summarization.

7.3.1.2 LLMs

In the past year, LLMs have shown remarkable performance on NLP tasks across the board. While LLM baselines are not included in Chapter 5, baselines using smaller versions of Flan-T5 LLM is included and discussed in Chapter 6. Since this model has a context window of 512 tokens, the document had to be summarised in chunks of 25 sentences at a time. The missing context resulting from this appraoch could explain

why the RNN based model performed better than the LLM in these experiments. More sophisticated ways of handling long documents like Sliding Window Attention[] and using more recent LLMs with larger context windows like MPT Storywriter[224] could help improve performance. LLMs have also been applied to zero shot identification of important narrative events in TV series[183].

7.3.1.3 Improving training data

In addition to experimenting with new model architectures, findings in chapter 5 suggests that improvements in the alignment step where extractive labels are derived from the abstractive summary (for example, by using weighted ROUGE instead of ROUGE as a metric for alignment) could give significant performance improvements. This is an important line of investigation given that the qualitative analysis of the oracle summaries conducted in Chapter 5 reveal several shortcomings and any supervised training or fine-tuning approach applied to this data will be limited by the quality of the oracle labels. Additionally, the dataset could be augmented with reference summaries of different sizes from additional sources (like Wikipedia) to allow models to better produce summaries with different granularity.

7.3.1.4 Further investigation of Rationale based learning for Narratives

In this thesis, rationale based training is only tested on a simple attention layer. Future work will adapt and test this approach on other attention-based models with different types of attention. The experiment is also limited by the availability of interactive narrative datasets. Future work will expand the number of IDNs covered in the dataset and further investigate generalizability of these findings across different types of IDN. Future work will also look at testing this approach for narrative and interactive narratives summarisation with different types of rationales.

7.3.1.5 Abstractive Summarisation

This thesis had chosen to focus on extractive summarisation since abstractive summarisation tends to produce hallucinations. Recently LLMs have shown remarkable success in various abstractive summarisation tasks. While hallcuinations are still a concern, the quality of generated text has come a since the decision to focus on extractive summarisation was made (in 2021). Abstractive summaries can be more readable and compact than extractive summaries, so this is also an interesting direction to pursue. Instruction tuned LLMs can even produce specialised summaries that highlight the emotional trajectory of a character or pick out a specific narrative thread or topic to focus the summary on. Commercial LLMs like GPT-4 already have

large context lengths, perform remarkably well at summarisation and is capable of performing tasks like this. It is worth noting, however, that processing such requests for thousands of IDN playthroughs would be expensive. However, smaller open source LLMs are quickly catching up.

7.3.2 Author Evaluation / Iterative Design

One limitation of this work is that utility of generated summaries for IDN has not been analysed systematically beyond the manual qualitative evaluation and discussion in section 6.4.2. The earlier phases of this research focused on identifying promising areas for applying NLP techniques to generate authoring feedback. Insights from this phase guided the focus for the later stages of this work - one of these types of feedback (summarisation) was chosen was further exploration and the later phases then focused on making resources available to investigate how well standard NLP techniques work on this domain (IDN text) and investigating one way in which standard practices can be adapted to better suit the domain. Investigating how making such feedback available to authors affects the writing process in practice falls outside the scope of this thesis but is a natural future direction for this work.

This line of investigation involves showing the summaries generated by the models discussed above to real IDN authors and interviewing them to gather feedback on which aspects need to be improved. While qualitative analyses conducted in thesis suggests that summaries could serve as a recap for authors, conducting interviews and task based evaluation with IDN authors can help us understand to what extent they give usable insight into the UX dimensions like uncertainty, expectation, curiosity, closure, desirable outcomes, in game affect type and intensity, continuity, narrative understanding, eudaimonic appreciation. Interviews will explore the authors' perspectives on the utility of the summaries in terms of giving insight into listed UX dimensions. Such feedback from authors can help guide future research and refine the understanding of what a useful summary looks like in the context of IDN authoring. For task based evaluations, the author will be presented with a partially completed IDN work and asked to complete it with the help of automatically generated summaries and without will be conducted. They will then answer questions on how summary based feedback changes their experience of authoring. This can give insight into how authoring feedback affect the authoring process.

7.3.3 Evaluation strategies for IDN Summarisation

The main bottleneck to quick experimentation and iteration of different summarisation approaches to find the most suitable one for IDN summarisation, is the limitations of the evaluation strategy. Rouge scores are based on word overlap, and

extensive paraphrasing of the summaries makes it hard to judge the results solely based on this. Several other metrics have been proposed[76] and while they cannot fully replace manual evaluation, computing additional metrics (for example, BertScore is another commonly used metric which measures semantic similarity rather than keyword overlap) can give additional perspective to the quality of the summaries. The manual fault and qualitative analyses, while insightful are time-consuming and resource-intensive and hence limited in that only a few summaries can be analysed this way. Additionally, the large document lengths and summaries in IDN-Sum makes large-scale human evaluation challenging. Recent NLP papers have also explored using LLMs for different types of automated evaluation[167]. An evaluation protocol using LLM designed for IDN summarisation can offer an alternate perspective into the results and speed up development through faster iteration. A common strategy currently used to evaluate plot summarisation involves checking how many questions about the plot human participants can answer by just looking at the summary. In future work, IDN-Sum dataset will be augmented with a list of important plot and choice points which can facilitate this type of evaluation.

7.3.4 Investigating Other Forms of Feedback

This thesis chose to focus on extractive summarisation based on the state of NLP appraoches at the time the review outlined in Chapter 4 was performed. However, with NLP techniques improving drastically with LLMs throughout the past few years, many other forms of feedback items discussed in Chapter 4 can now be feasibly generated using LLMs. For example, Question Answering using LLMs over long contexts and multiple documents using Retreival Augmented Generation (RAG) is being researched heavily[]. This could be applied to authoring feedback where the author is allowed to ask specific questions about given branches.

7.4 Final Remarks

While some general trends in NLP and Creative AI are geared towards Artificial General Intelligence (AGI) and Generative AI, this thesis has attempted to focus instead on an application of AI to empower creators by providing authoring feedback rather than replacing them. Applying NLP techniques to generate authoring feedback can open up new possibilities for IDN creation by allowing the author to keep better track of the complex non linear story space of IDNs. To this end, this research has mapped requirements expressed in the IDN research community to techniques being investigated in the NLP community and made resources available to further investigate one NLP problem (Automatic Text Summarisation) in the underexplored domain of Interactive Digital Narratives. One avenue of inquiry to improve IDN

summarisation (using rationale based learning with choice based rationales) was also explored using this dataset and has yielded encouraging results. The hope is that this research can become a foundation stone for future mixed-initiative approaches.

Appendix A

IDN-Sum Dataset Excerpts

Examples of the data are shown in this appendix. Appendix A.1.1 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.1.2 and appendix A.1.3 respectively. The complete summaries can be seen in the github page.

A.1 Data Example

A.1.1 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 ass-child. 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.1.2 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.1.3 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 .

A.2 Example of Automatically Aligned Extractive Summary

chloe takes another couple hits from her cigarette before letting it fall in between the tracks . she jumps out of the way of the train at the last second , watching it go by , then takes off her hood and looks at the sawmill across from her . if i 'm gon na get inside , i 'll have to get through that door . [ex] s0 : chloe can try to walk past the bouncer to the door . [ex] chloe : this is the old mill , right ? [ex] : sc : s0 : chloe manages to go upstairs and see the band . after a few moments , the guy she ran into earlier and his friend come to confront her . frank sees them and chloe stops , looking at the guys behind him . after a few moments she stops smoking , puts the ashtray away and sits up on her bed . (looking at her photo with william and max) better

change clothes . [ex] s0 : chloe goes to open the door but stops . how drunk was i last night ? i can use it to call mine , then figure out where the hell i left it . chloe follows the sound and finds her phone on the bathroom floor , under a towel , beside the toilet . [ex] s0 : chloe goes back to joyce and david 's room and takes her purse . no more dawdling , i need to talk to you ! she then goes downstairs . i made you breakfast ! and you will say thank you . you 'll need to bring him his keys from the ashtray . chloe looks back to see if joyce is not looking and quickly puts the money in her mom 's purse . [ex] : sc : s0 : chloe leaves the house and sees david . you 're going to be late . [ex] chloe : (looks at the tempest poster on a noticeboard nearby) whatever . [ex] s0 : choice : firewalk show . we 're called pisshead ? skip gets his phone and plays the demo to chloe . what do you think of this hypothesis : that you 'll be in your seat by the time chemistry class begins today ? this conversation occurs only if chloe sat on the crate on the stage . principal wells approaches chloe and she gets up from the crate , jumping off the stage and landing in front of him . it 's part of the tabletop game we play . [ex] chloe : (taking the dvd) we 're at the end of the campaign , so it 'll only take like ... twenty minutes ? in other words , chloe price . it 's an honor to fight alongside you . [ex] : sc : s0 : chloe is about to climb the stairs , but gets knocked down by nathan prescott , who is being chased by drew north . samantha myers comes and stands next to chloe . chloe approaches the school entrance , but rachel , dressed in a costume , opens the door first from the other side . [ex] : sc : : sc : s0 : chloe and rachel enter the hayden and dana rehearse under mr. keaton 's supervision . the question is : are miranda 's feelings of instant passion for ferdinand just inexperience and dramatic circumstances or ... has she actually just met the love of her life ? first she pulls out a photo of a young rachel with her father . there 's rachel 's belt . [ex] chloe : rachel amber ? rachel starts running after the passing train . [ex] chloe : if you had n't have shown up ... [ex] rachel : rachel moves to sit on the floor of the train carriage . well , that 's too bad , because it 's true . [ex] s0 : choice : lie . after they finish listening to the music , both girls take out their earbuds and chloe puts them away . got a knife on you ? she 's right there . [ex] s0 : chloe uses rachel 's nail file to unscrew the " martin lewis prescott " dedication plate . she throws the plate to the ground and takes the quarter from inside the viewfinder . you find some people for us to spy on , and then you and i will act out what they 're saying and thinking . the girls see a man and a woman , meeting under the oak tree . at chloe 's mark , the man and the woman start kissing . the honor student wants to show the school delinquent how to party ? they have a bottle of wine . the man and woman get up and go over to rachel . [ex] s0 : chloe tries to snatch the wine , but the couple notices her . what are you waiting for ? we 'll keep watch while you go . this woman needs help from someone who actually knows what they 're doing . [ex] s0 : chloe follows rachel as they leave the park . rachel 's been acting kind of standoffish ever since we left the park . rachel sits on a crate , looking upset , still holding her wine . okay , i still have no idea what 's going on with rachel , but apparently she gets smashy when she 's

angry . chloe : a real friendship . [ex] s0 : chloe looks at one of the objects around her . [ex] s0 : chloe looks at another object in the car . a truck appears outside the window and crashes into the left side of william 's car . [ex] : sc : s0 : chloe wakes up in her father's wrecked car . then she goes towards the train tracks and starts walking back to the overlook , as a raven flies overhead . i thought i could catch him , or something ... [ex] rachel : [ex] s0 : chloe gives her the lighter . rachel sets the photo on fire and lets it fall into the trash can . they both look at the fire and david puts his arm around her . then he notices steph , mikey and drew , hanging at the picnic table far from him . the woman seen kissing james is sitting on a bench at the overlook park , looking at the fire .

A.3 Example Summary from Best Model (SummaRuNNer)

chloe picks it up and walks away from him . [ex] s0 : chloe can try to take a t - shirt from the vendor 's car . [ex] s0 : chloe approaches the vendor 's car . the vendor goes talk to the truck driver . [ex] s0 : chloe takes the shirt out of the car . [ex] : sc : s0 : chloe sees the crowd and tries to push through it . chloe flips them off and walks away from them . alright , mosh pit is a no go . they run downstairs and rachel frees her hand from chloe 's . she rolls on her side and picks up her ashtray , then she puts the ashtray below her chest and starts smoking . after a few moments she stops smoking , puts the ashtray away and sits up on her bed . seeing firewalk live ... i said breakfast ! oh , can you grab my cellphone too ? she then goes downstairs . you used to love to learn . just let me know so i can stop fighting with blackwell to keep you on scholarship . money 's tight enough as it is . and you will say thank you . but i love you . [ex] chloe : i love you , too . the car , too . she goes to his side and he starts talking . [ex] david : chloe , is that a black eye ? then , both he and chloe get into the car . burnin ' the midnight oil again . william does n't answer . a truck crashes into the left side of the car , hitting william , and then everything goes black . you 're going to be late . so how about the tempest ... blackwell theater at its most pretentious . people have been taking me so seriously since i won the beacon 's young artist award for my photography . [ex] s0 : choice : vortex club . [ex] chloe : the story is about how relationships only work if people are willing to lie to each other . [ex] principal wells : chloe price ... is that a black eye ? do i have to initiate a search of your person in order to establish the veracity of these allegations , miss price ? you will meet me in my office after school for a formal reprimand . what else have you got to do before class ? the heavyset orc sergeant still remains . the orc clutches his groin , never to father children again . mr. keaton , sorry to interrupt , but does this look better ? i think it 's in my bag over there . [ex] s0 : chloe looks into rachel 's backpack . first she pulls out a photo of a young rachel with her father . that 's the price of valor . rachel climbs into a carriage on the train , then helps chloe as she joins her . it 's nice weather

. and then the other person has to guess which is which . you say three things [ex] rachel : rachel takes chloe 's marker and writes " rachel amber " on the floor of the train carriage with her right hand and then repeats the same successfully with her left hand . you 're hella mysterious , chloe price . which brings me to your alleged cat allergy . i bet it 's hard to impress chloe price . when your dad is the district attorney , i guess lying is ... something you 're used to . [ex] chloe : i know who to call if i need to get out of a ticket , then . [ex] chloe : let me know if you need an accomplice . luckily , we 've got some high - tech surveillance equipment right here . rachel inserts the coin into the viewfinder , then starts hitting it , but it does n't work . i 'd love to get it working for her . got a knife on you ? [ex] chloe : i guess you could stab someone with a nail file ... but you 've been on me for three hours ! stealing a dedication plate takes ... persistence . [ex] s0 : choice : not to brag . rachel brings chloe to the picnickers . the man and woman get up and go over to rachel . [ex] woman : did n't you used to be a lifeguard ? there 's a ranger station on the other side of the park . we 'll keep watch while you go . you 'd better run away before it gets you too . chloe approaches the car and starts hitting its hood with her fists and crying . then she falls on her back and continues crying on the ground . what is going on ? [ex] s0 : chloe looks at one of the objects around her . the car stops next to rachel , who is looking at chloe with wide eyes . [ex] : sc : s0 : chloe wakes up in her father 's wrecked car . [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 ... frank and his friend are hanging out next to his rv at the old mill . then he notices steph , mikey and drew , hanging at the picnic table far from him . the woman seen kissing james is sitting on a bench at the overlook park , looking at the fire . while smoking a cigarette , she starts smiling mysteriously .

Appendix B

IDN-Sum Further Analysis

B.1 ROUGE2 F1 Scores against human authored abstractive summaries

Table B.1 shows ROUGE2 scores computed against human authored abstarctive summaries.

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 B.1: ROUGE2 F1 scores against human authored abstractive summary

B.2 Trends and Outliers

The following trends and outliers were observed in the results:

1. Long version of Summarunner outperforms all other models except in case of novel dataset at length 81.
2. Some trends that can be observed in the R scores against human authored abstractive summary cannot be observed in R scores against aligned extractive summary. For example, TR outperforms SR (truncated at 100) in most cases when looking at scores against abstractive summaries but SR (truncated) performs significantly better for Novels and IDN when looking at scores against extractive summary.
3. In many cases, especially in case of IDN, at target 81, only SRF shows any improvement from random.
4. LF does improve on BS in most cases (when looking at r scores against extractive reference). However, this does not seem to be the case for novel dataset for target lengths 3, 37 and 81 and crd3 in case of target length 9.

The first two of these trends are investigated further below. The remaining will be addressed as part of future work.

B.2.1 Best model

Summarunner seems to scale for longer documents and the version of it that truncates documents at 3000 sentences (SRF) outperforms the other models in all cases except for novel dataset for target length 81, for which BertSum performs the best. The issue is low precision because of many irrelevant sentences as can be seen in table []. Qualitative analysis of a random sample of predictions also reveal a considerable amount of irrelevant sentences compared to BertSum which generates much shorter summaries due to the 512 token limit. Because of the token limit, BertSum only generates summaries that are, on average, 19 sentences long, even though the target length is set to 81 sentences. 81 sentences is potentially too long for this dataset, considering that the abstractive reference is only 20 sentences long on average and the original text is only approximately 200 sentences on average (refer Table 5.2). This is also reflected in the oracle summaries for novels, where the the average summary length for automatically aligned extractive summaries is 27, even when maximum length is set to 81. This is not the case for other datasets. The average summary lengths for BertSum and Oracle are shown in Table []. The alignment method employed, stops adding sentences to the oracle summary if R score cannot be

Length	3	9	27	81
Recall	0.43	0.43	0.60	0.83
Precision	0.37	0.48	0.45	0.29
F1	0.38	0.35	0.47	0.38

TABLE B.2: Precision and recall for Summarunner (full version) novel dataset calculated against automatically aligned extractive summaries (oracle summaries)

Length	3	9	27	81
Recall	0.24	0.39	0.49	0.48
Precision	0.38	0.43	0.42	0.44
F1	0.27	0.38	0.42	0.41

TABLE B.3: Precision and recall for BertSum on novel dataset calculated against automatically aligned extractive summaries (oracle summaries)

Dataset	avg len (BS)	avg len (O)
novel3	2.99	3
novel9	8.91	8.71
novel27	18.65	19.08
novel81	19.39	27.69
crd3_3	3.12	2.96
crd3_9	9.15	8.8
crd3_27	26.57	25.45
crd3_81	31.84	69.91
sb3	3.97	2.99
sb9	12.93	8.97
sb27	36.93	26.46
sb81	48.84	71.09
idn3	3.57	3
idn9	11.79	9
idn27	36.72	27
idn81	53.09	80.98

TABLE B.4: Average length of generated summaries for BertSum (BS) and Oracle (O)

improved further. All other models generate summaries of the maximum length. BertSum hence generates summaries that are of a length closer to the oracle summaries in case of novels, but this is just an artefact of the 512 token limit.

B.2.2 Reference Summaries

When looking at R scores against abstractive ref, TR performs better than SR (truncated). While this is not the case for novel27 and novel81, QA had revealed that selected sentences were often irrelevant even when R scores were high because of matching on stop words and common words. Calculating R scores without stopwords shows tr performs better even for these. This is expected since TR has access to the

entire document without truncation. But when considering R scores against extractive summaries SR performs better for novel and IDN datasets. However since SR only considers first 100 sentences, TR produces more useful summaries if analysed qualitatively. To understand error classes and reasons for discrepancy with rouge scores, we conduct a more detailed analysis. Since the difference is most obvious in case of idn target len 3, we choose this for analysis. We randomly sample 50 instances from corpus and the predictions from SR, TR and Oracle for comparison. Since the difference is emphasised only in r scores against extractive, we also sample 10 instances from oracle. In the first pass through the samples, recurring error classes were identified. In a second pass through the samples, the frequency of these errors were counted to get table B.5.

The remainder of the section goes through each of the top five error classes and discusses them in more detail. An example of a summary without any of these error classes is given below for reference:

> *[EX] S0 : CHOICE: (Didn't help Nathan) [EX] CHLOE : If "involved" means not sticking out my neck for Blackwell's richest ass-child. [EX] :SC: S0 : Rachel and Chloe leave Principal Wells' office followed by Joyce. [EX] S0 : CHOICE: (Said Rachel was a friend) [EX] CHLOE : (thinking) What's with everyone trying to pawn clothes off on me?*

The most frequent types of errors that were seen in the first pass through the predictions were as follows:

1. **Sentences that lack sufficient context** Some sentences were very short, or lacked enough context to convey any useful information, even when there is some text overlap or were picked up from parts of the text that were also discussed in human authored abstractive summary. This can be seen in sentences like *"in every way that matters ... [ex] s0 : rose rests her hand on james ' shoulder and he places his hand on hers "*. The remainder of that dialogue will inform the reader of important information : *"...Rose is my wife and your mother. But the woman you saw at the overlook...her name is Sera. Your...birth mother."* But this is broken up into separate extracts because there is another sentence in between. The summary was marked to have this error if it contained any sentences like this. TextRank and Summarunner both have a moderate amount of this error. Oracle does not exhibit this error at target length 3. It mostly picks out narration like sentences instead of sentences from inside utterances. However, an examination of oracle summaries at higher target lengths shows that some such sentences are included in the summary. This suggests that single sentence-level extracts may not be a good choice for screenplay-like text. However, since these sentences still contain

Error Classes	TR	SR	Oracle	Example
Includes sentences that lack sufficient context to be useful or understandable.	med (25/50)	med (37/50)	low (0/10)	hell is empty - script [ex] : sc : : sc : s0 : chloe and rachel sit on the couch in the living room . in every way that matters ... [ex] s0 : rose rests her hand on james ' shoulder and he places his hand on hers .
All narrative elements mentioned in human written summary missed.	low (10/50)	low (1/50)	low (0/10)	Frank looks at the guys and back to Chloe and Rachel. [EX] CHLOE : (thinking) Rachel looks awesome... [EX] CHLOE : (thinking) Rachel looks so happy here...
All interactive elements mentioned in the human written summary missed.	med (22/50)	high (48/50)	high (10/10)	She jumps out of the way of the train at the last second , watching it go by , then takes off her hood and looks at the sawmill across from her . Chloe approaches the school entrance , but Rachel , dressed in a costume , opens the door first from the other side . After they finish listening to the music , both girls take out their earbuds and Chloe puts them away .
Summary contains irrelevant sentences.	med (35/50)	low (2/50)	low (2/10)	[EX] CHOICE: RUN S0 : Chloe runs away from Guy 1 and he tries to reach her, but Rachel throws a bottle right at his forehead and he falls to the ground. [EX] S0 : CHOICE: (Didn't see the photo Rachel posted) [EX] CHLOE : Who exactly is talking shit about me? [EX] CHLOE : (thinking) Rachel looks awesome...
No extracts from middle/end of text.	low (0/50)	high (50/50)	low (0/10)	she takes a deep breath , then takes the cigarette out of her mouth and breathes out the smoke . she jumps out of the way of the train at the last second , watching it go by , then takes off her hood and looks at the sawmill across from her . alright , it 's not gonna take any more than an hour to do it .

TABLE B.5: Frequency of Top 5 Error Types in TextRank (TR) Summarunner (SR) and Oracle for IDN dataset, target length 3.

some amount of keyword overlap, this error will not be reflected in ROUGE scores.

2. **All narrative elements missed**

Some summaries contained none of the events or plot points that were covered by the human authored abstractive summary. This can be seen in the corresponding example where the summary contains no useful plot information. Summary was marked to contain this error only if all the plot elements in the human authored were missed. While missing narrative elements is relatively low in all cases, TextRank summaries exhibit this error more often than Summarunner and Oracle. Note that these numbers do not reflect the number of story elements that get picked up, but the number of summaries that do not have any plot elements at all. Summarunner picks up at least one plot element in most cases, but it is usually many sentences about one plot element from the beginning of the text because of document truncation (error 5).

3. **All interactive elements missed**

   Some summaries contained none the interactive elements (such as decision points, tasks set out by the game that the player needs to do) that were covered by the human authored abstractive summary. The reference summary provided above, in comparison contains several "CHOICE" points illustrating which choice was taken at that decision point. Summary was marked to contain this error only if all the interactive elements in the human authored were missed. TextRank picks up interactive elements a lot more often whereas compared to Summarunner. However since and Oracle also does not pick these up, the Rouge scores against oracle summaries will not reflect this. Additionally, since Summarunner is trained against the Oracle summaries, its performance in this aspect could be improved by improving the training data using a better automatic alignment algorithm.

4. **Summary contains irrelevant sentences**

   The sentence is considered irrelevant if the topic covered by it does not appear in the human authored summary. In case of the example shown, the choice around not seeing the photo and Chloe asking who was talking about her was not mentioned in human authored summary. Whereas in case of the example summary without any of the error classes, all the sentences relate to topics discussed in the human authored summary. Summary was marked to have this error if it contained at least one irrelevant sentence. TextRank also has many summaries that contain at least one irrelevant sentence, whereas Summarunner and Oracle summaries rarely do. Again, note that these numbers show number of summaries that have at least one irrelevant sentence and does not show number of irrelevant sentences within a summary.

5. **No extracts from middle or end of text**

   Due to document truncation at 100 sentences in case of Summarunner, all of its predictions have this error. Since TextRank and Oracle do not truncate

documents before summarising, neither of them have this error. Because of this error, overall, TR summaries are more useful and informative than the ones from SR. However, SR summaries do a good job of picking important information if it is contained within the first 100 sentences. In many cases, at least one out of three extracts in the Oracle are from the beginning of the text. The SR summaries get a high ROUGE for these instances, increasing its overall score.

SR does better than TR in terms of not picking irrelevant sentences and mostly being able to capture at least one story element. However TR does better than SR in terms of variety - by picking up both story and interactive elements and not being limited to the beginning of the text. TR also has fewer incoherent sentences. It also does better in terms of overall quality, according to human judgement. However many of the cases where TR does better is not reflected in the scores due to limitations of the ROUGE metrics, and short comings in the aligned extractive summary (not picking up interactive elements). Short comings in the aligned extractive summary also gives insight into why this trend can be observed in the scores against it, but not in the scores against human authored abstractive summaries.

Appendix C

Additional training and model details

C.1 Training details for SummaRuNNer variants

Full list of parameters:
embed_dim = 100
embed_num = 100
pos_dim = 50
pos_num = 3000
seg_num = 10
hidden_size = 200
lr = 1e-3
batch_size = 1
epochs = 5 with early stopping if no improvement was observed within 5 validation rounds.
seed = 66
report_every = 30
seq_trunc = 50
topk = 81

C.2 Training Details for Flan variants

C.2.1 Flan-T5-base

Training was performed using LORA fine tuning. Due to limitation of context length, training was performed in 25 sentence chunks. To avoid empty predictions, chunks

Model	num_params
RNN	1726201
AttnRNN	1727001
sentonlyAttn	1726601
wordonlyAttn	1726601
Flan T5 base (LORA fine tuning)	1769472
Flan T5 base Encoder only (LORA fine tuning)	612098

TABLE C.1: Number of trainable parameters in each model. Rationale based training does not introduce any additional parameters. Note that Flan variants show number trainable parameters under LORA not total number of parameters.

which had no positive examples were excluded. Hyperparameters used are listed below:
learning_rate=1e-4,
gradient_accumulation_steps=16,
per_device_train_batch_size=4,
per_device_eval_batch_size=1,
num_train_epochs=3,
save_steps=500,
eval_steps = 500,
save_strategy = "steps",
warmup_steps = 10000,
weight_decay = 0.01
input_max_length = 128
output_max_length = 384

C.2.2 Flan-T5-base Encoder only

Training was performed using LORA fine tuning. Due to limitation of context length, training was performed in 25 sentence chunks. To account for label imbalance , positive examples were weighted more (double) during loss computation and chunks which had no positive examples were excluded. Hyperparameters used are listed below:
evaluation_strategy="steps",
learning_rate=1e-4,
gradient_accumulation_steps=16,
per_device_train_batch_size=4,
num_train_epochs=3,
save_steps=500,
eval_steps=500,
save_strategy = "steps",

warmup_steps = 10000,
weight_decay = 0.01

C.2.3 LORA Config

r=16, lora_alpha=32, target_modules=["q", "v"],
modules_to_save=['sent_attention','classifier','layer_norm'], lora_dropout=0.05

C.3 Infrastructure

The models were trained on compute cluster containing Nvidia Tesla V100 and Nvidia GTX 1080Ti graphics cards. Only 1 GPU was used for training each model.

C.4 Hyperparameter optimisation method

Optimal hyperparameters were found through manual tuning within range : 2-8 sentences and 20-80 words for ws and 0.25 - 0.99 for alpha.

Appendix D

Further analysis of summaries trained with choice based rationale

Table D.1 shows ROUGE scores of the model summaries calculated against the automatically generated extractive summaries from IDNSum rather than the abstractive summaries shown in Table 6.1 It is worth noting that when considering these ROUGE scores, no improvement is observed when rationales were introduced and attention models trained without rationale seem to perform best. To investigate this further, ROUGE scores with stop word filter turned on were also calculated against the human authored abstractive summaries and automatically aligned extractive summaries. This is shown in table D.2. Here, we again see that rationale-based models perform better in all cases. This suggests that the better ROUGE scores shown by the attention models without rationales are due to keyword

Model	R1(ext)	R2(ext)	RL(ext)
RNN	0.60399	0.27518	0.59484
sentonly AttnRNN	0.61649	0.27374	0.60744
sentonly AttnRNN +rationale	0.60082	0.29412	0.59181
wordonly AttnRNN	0.61835	0.28871	0.61041
wordonly AttnRNN +rationale	0.61178	**0.30197**	0.60336
AttnRNN	**0.61975**	0.25674	**0.61095**
AttnRNN +rationale	0.61073	0.29291	0.60227
Google flan-t5-base (zero-shot)	0.50939	0.15510	0.45585
Google flan-t5-base FT	0.58108	0.22058	0.52768
Google flan-t5-base Encoder	0.58912	0.23545	0.55075
Google flan-t5-base Encoder + rationale	0.61301	0.25956	0.57839
Google flan-t5-large (zero-shot)	0.51102	0.13795	0.45901

TABLE D.1: ROUGE scores against automatically generated extractive summary. Rationales do not seem to show an improvement here in case of SummaRuNNer models. Refer Table D.2 for further analysis of why.

Model	R1(abs)	R2(abs)	R1(ext)	R2(ext)
RNN	0.32315	0.03680	0.45616	0.20193
sentonly AttnRNN	0.32272	0.03865	0.45859	0.19988
sentonly AttnRNN + rationale	0.33955	0.03969	0.46052	0.21759
wordonly AttnRNN	0.33317	0.03699	0.47849	0.21426
wordonly AttnRNN + rationale	**0.34771**	**0.04240**	0.49499	**0.23106**
AttnRNN	0.31796	0.03191	0.45885	0.17821
AttnRNN + rationale	0.34716	0.04234	**0.50062**	0.22407

TABLE D.2: Rouge Scores calculated against the human-authored abstractive summary (abs) and automatically aligned extractive summary (ext) with the stop word filter turned on for Summarunner variants. Results show rationale-based models performing better in all cases indicating that the higher rouge scores for non-rationale-based models in table D.1 are due to overlap on insignificant words.

overlap on insignificant words while the rationale based models perform better when considering significant words.

Appendix E

Model Outputs

This Appendix shows an example of model output from each of models discussed in Chapter 6. The human authored abstractive summary for the IDN is also included for reference.

E.1 Human authored abstractive summary

the episode starts with a hooded chloe price smoking a cigarette and standing at the railroad tracks , waiting for the train to come . after its passage , she takes off her hood and goes towards the old mill , ignoring the " no trespassing " sign after jumping over a fence . in order to get inside the mill , she starts an argument with the bouncer , in which she can win and be allowed inside or have to use the backdoor to get in if she fails . after entering , she can interact with people , objects , graffiti and even steal a shirt and some money (or not) . if she steals the money , she 'll have the option of buying weed from frank bowers or save the money for later . after this , she 'll try to go through the crowd in order to see the band that 's playing and ends up bumping into two skeevy guys , the taller one being somewhat aggressive towards her . however , she 'll manage to see the band by going upstairs , even with the floor there being rotten . she enjoys the music for a while until the two guys from before appear and confront her . she 's saved by rachel amber , in her very first appearance , and will have the first major choice of the game : attack one of the guy s or run without doing nothing . after running downstairs , the two girls stop and look at frank , who notices what 's going to happen and quickly stops the guys , causing both of them to get angry and leave the mill . rachel pulls chloe along with her , and they enjoy the firewalk show for the rest of the night . the next morning , chloe is shown waking up in her room , at the price household . she sits up and takes her red ' oregon ' ashtray and starts smoking a cigarette (or weed , if she bought it from frank earlier with the stolen money) . chloe can look at a photo of her and max caulfield as kids along with

her dad , william price and can also look at her diary . after being called by her mother , joyce price , for breakfast downstairs , she 'll get up from her bed and will be able to interact with many objects around her room . she 'll change clothes before leaving her room (the player gets to choose her outfit) . due to having drunk too much , chloe notices her phone is missing . she then goes to her mother 's room to call her phone with her mother 's phone . there , she can interact with another object before using her mother 's phone . by doing this , chloe finds her phone in the bathroom , and just after she takes it , her mother asks her from downstairs to bring her purse along with her phone , causing her to go back to the room . when she finally goes downstairs , she can interact with the objects around the living room and even get some information that can be used later on . she 'll then talk with her mother about several topics , and at the end of the conversation , she 'll have to choose between being comprehensive towards joyce or saying how she actually feels . depending on her choice , joyce will be either kind or tough towards her . 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 . after leaving the house and going to david , he 'll ask her to get the tools he needs to fix his car in the garage . after this , she gets in the car and david will try to start a conversation with her . she can either start a fight with him or listen to what he has to say . after the talk , chloe will fall asleep and have a weird dream about being in a car with her dad , going to pick up her mom from the grocery store . the dream abruptly ends with a truck crashing through william 's car . when she wakes up , she 's already at blackwell academy . there , she can talk to eliot hampden (and choose whether or not she wants to watch the tempest play with him) , victoria chase (with the option of sabotaging her homework if doing so) , skip matthews (with the option of listening to the demo of his band , pisshead , and giving him your opinion on it if doing so) , principal wells (only if she sits on a crate on the stage and with the option of starting a backtalk challenge with him if doing so) , michelle grant , mikey north and steph gingrich (in order to get her dvd and with the additional option of playing a tabletop game with them) , and other students . afterwards , she 'll go towards the school entrance , but she 'll be interrupted by drew north and nathan prescott , who are starting a fight . a student called samantha myers urges chloe to do something to help nathan , and chloe can either backtalk drew and defend nathan or just ignore them , which will cause samantha to either thank chloe or be upset with her for not helping . when the fight is over , she can finally enter the school , and just as she opens the door , rachel appears on the other side and pulls her along with her to the drama lab where travis keaton is rehearsing with dana ward and hayden jones . rachel will ask for her opinion on miranda 's love for fernando , both portrayed by dana and hayden respectively in " the tempest " play , and chloe can choose whether to say it 's true love or not . after the class is over , everyone will leave the room , except for chloe and rachel . rachel will change to normal clothes and ask chloe to get her belt from her bag and bring it to her . after doing this , they 'll have a short conversation before rachel invites chloe to skip

school , and they end up in a train carriage where , after finding some crates to sit on , they play the game two truths and a lie . chloe can either cheat or follow the game rules . after their game , chloe will have the option of sharing or not her earbuds with rachel during their trip . upon arriving at overlook park , they 'll play another game using the viewfinders to spy on people around the park . however , the viewfinder that they intend to use it broken . chloe asks rachel for something sharp like a knife and she gives her a nail file , which chloe uses to unscrew a deducation plate from a park bench and then uses it to break open the viewfinder , allowing them to use it for free . when they get a closer view of the last couple available , a man and a woman under a tree , rachel gets distressed when they start kissing and puts an end to their game , telling chloe she needs to get drunk . they 'll then go to the other side of the park , where a couple is having a picnic and have a bottle of wine on their table . rachel approaches the couple and starts acting sick , throwing herself to the ground and pretending to be in need of resuscitation . chloe can encourage the man to help her , either succeeding or failing on doing so resulting in the man " saving " rachel 's life or t he woman seeing through their ploy . whatever the outcome is , the two girls will get the wine . after this , they 're shown walking on the train tracks . chloe invites rachel to explore a junkyard nearby and rachel lets her explore on her own . after a long conversation between the two , no matter what choices the player has made so far , rachel will leave , but not before chloe tries to convince her to stay . chloe has the option to say that they have a real friendship or something more . once rachel leaves , chloe gets angry and breaks everything around her . she then has a meltdown upon seeing her late father 's car . she 'll fall asleep and have another dream about her father , this time , with him advising / warning her on her relationship with rachel , in which chloe will see rachel outside of the car , who will then catch fire . when she wakes up , it 's already night and she goes back to the overlook where she finds rachel . 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 . rachel takes out a photo of her as a kid with her father , asks chloe for her lighter which she uses it burn the photo , throwing it into a nearby trash bin , and starting a wildfire by kicking the bin into a tree nearby . rachel then screams , increasing the fire 's intensity . the episode ends showing each character 's reaction to the wildfire seen in the sky .

E.2 Summary from SummaRuNNer (RNN)

: sc : s0 : chloe price , standing on train tracks and wearing a black hoodie , flicks her lighter a few times and lights up her cigarette . she takes a deep breath , then takes the cigarette out of her mouth and breathes out the smoke . a train begins to approach her . chloe takes another couple hits from her cigarette before letting it fall in between the tracks . she jumps out of the way of the train at the last second , watching it go by ,

then takes off her hood and looks at the sawmill across from her . this place is awesome . [scoffs] meaning you . yeah , your problem . those guys need to get a room . man 1 then slaps man 2 . i really get it now , i — i do . it 's not a bad fake , kid . the bouncer throws her id on the ground . chloe picks it up and walks away from him . or can something around here help me convince him ? [ex] s0 : chloe can try to walk past the bouncer to the door . he holds out his arm to block her and she turns around exasperation . i heard firewalk is playing here tonight . just follow the lights and the sound . [ex] chloe : (thinking) still a dick . that guy 's a dick . chloe releases the parking brake and the car slides down . the vendor goes talk to the truck driver . [ex] s0 : chloe spots a box with money near the shirts . [ex] : sc : s0 : chloe sees the crowd and tries to push through it . look at that getup . what are you even doing here ? hard to get to the stage . i could definitely use something to take the edge off . after a few moments , the guy she ran into earlier and his friend come to confront her . guy 2 goes to help guy 1 who 's on the floor and chloe runs to rachel . they look at each other and notice guy 2 helping guy 1 to get up . they run downstairs and rachel frees her hand from chloe 's . rachel takes chloe 's hand again and they run towards the entrance to the show . frank sees them and chloe stops , looking at the guys behind him . he then jumps in front of the guys . rachel blows them a kiss and pulls chloe by the hand , who also blows them a kiss and flips them off . in front of the stage , rachel and chloe dance together . the night ends with chloe making one last pose before going back to dancing . [ex] : sc : : sc : s0 : chloe 's alarm clock starts playing music and she wakes up . she rolls on her side and picks up her ashtray , then she puts the ashtray below her chest and starts smoking . after a few moments she stops smoking , puts the ashtray away and sits up on her bed . daily rituals are important , even when they involve writing unread letters to friends who 've forgotten you ... i smell like cigarettes and beer . okay , mom 's phone is probably in her room . i can use it to call mine , then figure out where the hell i left it . she then gets her mom 's phone from the nightstand and unlocks it . how can mom look at this every day and not see what a tool she 's dating ? chloe follows the sound and finds her phone on the bathroom floor , under a towel , beside the toilet . you can put my purse on the dining table . might still have time for breakfast if you hurry . i know what time you came home last night . just let me know so i can stop fighting with blackwell to keep you on scholarship . but sometimes we need to make more room in our hearts for new people . okay , david 's waiting . chloe looks back to see if joyce is not looking and quickly puts the money in her mom 's purse . the car , too . chloe throws the keys to david and he catches them in time , putting them in his back pocket . better just get the socket wrench and get this over with . she goes to his toolbox , leans down and opens it . she then takes the socket wrench . he frowns at her and holds out his hand , and she gives him the socket wrench . he takes it and goes back to fixing his car . he takes the toolbox from the ground and walks towards a table in the corner . burnin ' the midnight oil again . blackwell theater at its most pretentious . skip gets his phone and

plays the demo to chloe . after it ends he puts his phone in his back pocket . well ... the prescotts have made an extremely generous donation to the school , which is good , but instead of going to support more science and mathematics , it 's all being dedicated to the arts . principal wells approaches chloe and she gets up from the crate , jumping off the stage and landing in front of him . rachel climbs into a carriage on the train , then helps chloe as she joins her . they come across the american rust salvage yard . chloe picks the bat up from the ground and looks around angrily . rachel puts her hand on the glass , chloe puts hers on the other side . a truck appears outside the window and crashes into the left side of william 's car . she gets out of it and leans on the hood one last time . then she goes towards the train tracks and starts walking back to the overlook , as a raven flies overhead . as the garbage inside the can starts burning , she takes a step back . frank stares in shock at the fire and smoke in the distance . james amber and principal wells are talking to a police officer at the blackwell parking lot . the woman seen kissing james is sitting on a bench at the overlook park , looking at the fire .

E.3 Summary from Sentonly SummaRuNNer trained without rationales (sentonly AttnRNN)

: sc : s0 : chloe price , standing on train tracks and wearing a black hoodie , flicks her lighter a few times and lights up her cigarette . a train begins to approach her . she then walks down toward the mill . the bouncer throws her id on the ground . chloe picks it up and walks away from him . the pitbull does n't bark at chloe . the vendor goes talk to the truck driver . [ex] : sc : s0 : chloe sees the crowd and tries to push through it . you 're trying too hard . after a few moments , the guy she ran into earlier and his friend come to confront her . they look at each other and notice guy 2 helping guy 1 to get up . the men leave and frank looks back to see that rachel and chloe are gone . she rolls on her side and picks up her ashtray , then she puts the ashtray below her chest and starts smoking . after a few moments she stops smoking , puts the ashtray away and sits up on her bed . how can mom look at this every day and not see what a tool she 's dating ? chloe follows the sound and finds her phone on the bathroom floor , under a towel , beside the toilet . oh , can you grab my cellphone too ? she then goes downstairs . you used to love to learn . i used to think drugs were lame , too . money 's tight enough as it is . he 's a good man . and you will say thank you . try not to kill each other . the car , too . she goes to his side and he starts talking . he takes it and goes back to fixing his car . then , both he and chloe get into the car . william does n't answer . i know what a spark plug does , jerkwad . a truck crashes into the left side of the car , hitting william , and then everything goes black . eliot sees her , puts down the book he 's reading , and approaches her . stopped any gang wars lately ? so i went to the mill last night , caught firewalk live . potion would n't have

worked . you ’re asking me ? you are an elf barbarian . we ’re supposed to kill the dur - dude . upon arriving at the training ground you are spotted by a heavyset orc , who immediately shouts and points . there are a dozen raiders on the training field , all of whom raise their weapons and charge ! the heavyset orc sergeant still remains . the orc clutches his groin , never to father children again . what about the loot ? my dad lost his job at the shipyard when your dad closed it down . mr. keaton , sorry to interrupt , but does this look better ? first she pulls out a photo of a young rachel with her father . a rhetorical question ? rachel climbs into a carriage on the train , then helps chloe as she joins her . wish max were here , so i could ask . rachel moves to sit on the floor of the train carriage . second , i was born in new york , the land of fashion and broadway , to which i will one day return when my heinous exile here in arcadia bay comes to an end . rachel takes chloe ’s marker and writes " rachel amber " on the floor of the train carriage with her right hand and then repeats the same successfully with her left hand . but i ’ve passed by your locker a few times , and i ’ve seen that old photo of a cat you keep in there . luckily , we ’ve got some high - tech surveillance equipment right here . i ’d love to get it working for her . but you ’ve been on me for three hours ! stealing a dedication plate takes ... persistence . at chloe ’s mark , the man and the woman start kissing . oh , honey , i think we used the vibrating bed for too long . last i checked , you ’re supposed to be chloe price . or we could go try to find a liquor store instead ? the man and woman get up and go over to rachel . talk about committing to a performance . [ex] s0 : chloe tries to snatch the wine , but the couple notices her . there ’s a ranger station on the other side of the park . you ’d better run away before it gets you too . guess we ’re leaving now . but i want to find out . " burning the midnight oil " song is still playing on the radio . william turns it off and looks at chloe . the raven suddenly appears on the hood of the car , and almost immediately disappears . a truck appears outside the window and crashes into the left side of william ’s car . then she goes towards the train tracks and starts walking back to the overlook , as a raven flies overhead . i do n’t know how to talk about this . so when i saw he got a text from an unknown number ... asking him to meet ... frank and his friend are hanging out next to his rv at the old mill . they both look at the fire and david puts his arm around her . the three of them look at the fire . james amber and principal wells are talking to a police officer at the blackwell parking lot . the woman seen kissing james is sitting on a bench at the overlook park , looking at the fire . while smoking a cigarette , she starts smiling mysteriously .

E.4 Summary from Sentonly SummaRuNNer trained with rationales (sentonly AttnRNN + rationale)

: sc : s0 : chloe price , standing on train tracks and wearing a black hoodie , flicks her lighter a few times and lights up her cigarette . a train begins to approach her . chloe

releases the parking brake and the car slides down . after a few moments , the guy she ran into earlier and his friend come to confront her . guy 2 goes to help guy 1 who 's on the floor and chloe runs to rachel . they look at each other and notice guy 2 helping guy 1 to get up . frank sees them and chloe stops , looking at the guys behind him . frank looks at the guys and back to chloe and rachel . he then jumps in front of the guys . the men leave and frank looks back to see that rachel and chloe are gone . if she attacked the skeevy guys , she will now have a bruise under her eye . she rolls on her side and picks up her ashtray , then she puts the ashtray below her chest and starts smoking . after a few moments she stops smoking , puts the ashtray away and sits up on her bed . i can use it to call mine , then figure out where the hell i left it . how can mom look at this every day and not see what a tool she 's dating ? chloe follows the sound and finds her phone on the bathroom floor , under a towel , beside the toilet . she then goes downstairs . you can put my purse on the dining table . [ex] s0 : choice : slip money in joyce 's purse (stole vendor 's money and did n't buy weed from frank) he takes the toolbox from the ground and walks towards a table in the corner . 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 . alone , you have fought your way through the raider camps , seeking their warlord leader , duurgaron the unscarred . to your left , the raiders ' training ground . upon arriving at the training ground you are spotted by a heavyset orc , who immediately shouts and points . there are a dozen raiders on the training field , all of whom raise their weapons and charge ! the orc clutches his groin , never to father children again . rachel takes chloe 's hand and pulls her into the building . first she pulls out a photo of a young rachel with her father . rachel has her back turned to chloe and is wearing jeans and a bra . to tell the truth , i went to bed last night wishing it never had to end . a rhetorical question ? rachel puts more makeup on chloe 's bruise . when she 's done , the bruise is no longer visible . rachel climbs into a carriage on the train , then helps chloe as she joins her . rachel moves to sit on the floor of the train carriage . second , i was born in new york , the land of fashion and broadway , to which i will one day return when my heinous exile here in arcadia bay comes to an end . rachel takes chloe 's marker and writes " rachel amber " on the floor of the train carriage with her right hand and then repeats the same successfully with her left hand . so new york 's on the bucket list ? but i 've passed by your locker a few times , and i 've seen that old photo of a cat you keep in there . hate to break it to you , but chloe price is n't exactly renowned throughout arcadia bay as a bastion of trust and empathy . rachel smiles , takes an earbud from chloe and puts it in . after they finish listening to the music , both girls take out their earbuds and chloe puts them away . this game involves spying on people from afar . luckily , we 've got some high - tech surveillance equipment right here . i admit , it was really dumb to lock the keys in the car . stealing a dedication plate takes ... persistence . chloe uses the plate to pry open the viewfinder . she throws the plate to the ground and takes the quarter from inside the viewfinder . she approaches rachel and holds out the quarter

triumphantly . the girls see a man and a woman , meeting under the oak tree . at chloe 's mark , the man and the woman start kissing . oh , honey , i think we used the vibrating bed for too long . last i checked , you 're supposed to be chloe price . they have a bottle of wine . or we could go try to find a liquor store instead ? there 's a ranger station on the other side of the park . rachel and chloe walk down a train track . rachel is drinking the wine that the two of them stole from the picnickers and chloe is balancing on the rails . they come across the american rust salvage yard . [ex] s0 : chloe scans the area , and looks almost relieved when she finds a baseball bat leaning against one of the old rusted cars . " burning the midnight oil " song is still playing on the radio . william turns it off and looks at chloe . the raven suddenly appears on the hood of the car , and almost immediately disappears . [ex] s0 : chloe sees rachel , walking towards the oak tree , as william 's car passes it by . a truck appears outside the window and crashes into the left side of william 's car . she gets out of it and leans on the hood one last time . then she goes towards the train tracks and starts walking back to the overlook , as a raven flies overhead . so when i saw he got a text from an unknown number ... asking him to meet ... rachel sets the photo on fire and lets it fall into the trash can . as the garbage inside the can starts burning , she takes a step back . its burning contents fall out towards the oak tree , setting it on fire . rachel starts screaming loudly , and at the same time a gust of wind comes from behind her , spreading the fire to the entire tree . frank and his friend are hanging out next to his rv at the old mill . his friend is on the phone and frank is drinking a bottle of beer . frank stares in shock at the fire and smoke in the distance . they both look at the fire and david puts his arm around her . [ex] s0 : nathan is sitting at the fountain , looking through his picture book . then he notices steph , mikey and drew , hanging at the picnic table far from him . the three of them look at the fire . james amber and principal wells are talking to a police officer at the blackwell parking lot . the woman seen kissing james is sitting on a bench at the overlook park , looking at the fire .

E.5 Summary from wordonly SummaRuNNer trained without rationales (wordonly AttnRNN)

: sc : s0 : chloe price , standing on train tracks and wearing a black hoodie , flicks her lighter a few times and lights up her cigarette . she takes a deep breath , then takes the cigarette out of her mouth and breathes out the smoke . a train begins to approach her . chloe takes another couple hits from her cigarette before letting it fall in between the tracks . she jumps out of the way of the train at the last second , watching it go by , then takes off her hood and looks at the sawmill across from her . she then walks down toward the mill . the bouncer throws her id on the ground . chloe picks it up and walks away from him . he holds out his arm to block her and she turns around exasperation . the pitbull does n't bark at chloe . just follow the lights and the sound .

[ex] s0 : chloe bends down and pets the pitbull . you looking to get beat ? chloe releases the parking brake and the car slides down . the vendor goes talk to the truck driver . [ex] s0 : chloe spots a box with money near the shirts . [ex] : sc : s0 : chloe sees the crowd and tries to push through it . studs ? you 're trying too hard . after a few moments , the guy she ran into earlier and his friend come to confront her . they look at each other and notice guy 2 helping guy 1 to get up . frank sees them and chloe stops , looking at the guys behind him . the men leave and frank looks back to see that rachel and chloe are gone . the night ends with chloe making one last pose before going back to dancing . if she attacked the skeevy guys , she will now have a bruise under her eye . she rolls on her side and picks up her ashtray , then she puts the ashtray below her chest and starts smoking . after a few moments she stops smoking , puts the ashtray away and sits up on her bed . okay , mom 's phone is probably in her room . i can use it to call mine , then figure out where the hell i left it . i think i saw mom 's purse in her room . [ex] s0 : chloe goes back to joyce and david 's room and takes her purse . oh , can you grab my cellphone too ? [ex] s0 : chloe slips joyce 's phone into her purse and leaves the room . she then goes downstairs . you can put my purse on the dining table . you used to love to learn . you 'll need to bring him his keys from the ashtray . [ex] s0 : choice : slip money in joyce 's purse (stole vendor 's money and did n't buy weed from frank) the car , too . she goes to his side and he starts talking . [ex] david : chloe , is that a black eye ? she goes to his toolbox , leans down and opens it . he takes it and goes back to fixing his car . then , both he and chloe get into the car . [ex] s0 : chloe looks at the purse beside her . [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 . out of the car , chloe . [ex] s0 : chloe opens the door , gets out of the car and stands holding the door looking at david . i 'd rather have my eyes gouged out with rusted forks . after it ends he puts his phone in his back pocket . if i had known the celestial avenger was bloodied , i would have totally given him my potion . you stand at a three - way crossing . to your left , the raiders ' training ground . rachel has her back turned to chloe and is wearing jeans and a bra . a rhetorical question ? rachel starts running after the passing train . both girls take off running . rachel climbs into a carriage on the train , then helps chloe as she joins her . rachel moves to sit on the floor of the train carriage . rachel takes chloe 's marker and writes " rachel amber " on the floor of the train carriage with her right hand and then repeats the same successfully with her left hand . i admit , it was really dumb to lock the keys in the car . stealing a dedication plate takes ... persistence . chloe uses the plate to pry open the viewfinder . the girls see a man and a woman , meeting under the oak tree . oh , honey , i think we used the vibrating bed for too long . or we could go try to find a liquor store instead ? i think it 's contagious . you 'd better run away before it gets you too . the raven suddenly appears on the hood of the car , and almost immediately disappears . a truck appears outside the window and crashes into the left side of william 's car . she gets out of it

and leans on the hood one last time . then she goes towards the train tracks and starts walking back to the overlook , as a raven flies overhead . so when i saw he got a text from an unknown number ... asking him to meet ... plus you came along with me , no questions asked . my mom might skip grounding and just go straight to the death penalty . they both look at the fire and david puts his arm around her . [ex] s0 : nathan is sitting at the fountain , looking through his picture book . the three of them look at the fire . the woman seen kissing james is sitting on a bench at the overlook park , looking at the fire . while smoking a cigarette , she starts smiling mysteriously .

E.6 Summary from wordonly SummaRuNNer trained with rationales (wordonly AttnRNN + rationale)

: sc : s0 : chloe price , standing on train tracks and wearing a black hoodie , flicks her lighter a few times and lights up her cigarette . chloe picks it up and walks away from him . the pitbull does n't bark at chloe . she reaches the shirt and the vendor slaps her hand away . chloe releases the parking brake and the car slides down . the vendor goes talk to the truck driver . [ex] : sc : s0 : chloe sees the crowd and tries to push through it . a man in crowd elbows chloe backward and she bumps into a man , spilling his beer . [ex] s0 : chloe looks at the stairwell near the entrance . after a few moments , the guy she ran into earlier and his friend come to confront her . they look at each other and notice guy 2 helping guy 1 to get up . frank sees them and chloe stops , looking at the guys behind him . he then jumps in front of the guys . rachel blows them a kiss and pulls chloe by the hand , who also blows them a kiss and flips them off . the men leave and frank looks back to see that rachel and chloe are gone . she rolls on her side and picks up her ashtray , then she puts the ashtray below her chest and starts smoking . after a few moments she stops smoking , puts the ashtray away and sits up on her bed . okay , mom 's phone is probably in her room . i can use it to call mine , then figure out where the hell i left it . chloe follows the sound and finds her phone on the bathroom floor , under a towel , beside the toilet . [ex] s0 : chloe slips joyce 's phone into her purse and leaves the room . she then goes downstairs . you 'll need to bring him his keys from the ashtray . the car , too . she goes to his side and he starts talking . he takes it and goes back to fixing his car . [ex] s0 : david goes to the garage and puts back the socket wrench inside his toolbox . he takes the toolbox from the ground and walks towards a table in the corner . then , both he and chloe get into the car . chloe looks at the socket wrench in front of her . a truck crashes into the left side of the car , hitting william , and then everything goes black . out of the car , chloe . [ex] s0 : chloe opens the door , gets out of the car and stands holding the door looking at david . i 'd rather have my eyes gouged out with rusted forks . so i went to the mill last night , caught firewalk live . if i had known the celestial avenger was bloodied , i would have totally given him my potion . rachel

takes chloe ’s hand and pulls her into the building . chloe and rachel are left alone . rachel has her back turned to chloe and is wearing jeans and a bra . rachel starts running after the passing train . rachel climbs into a carriage on the train , then helps chloe as she joins her . rachel moves to sit on the floor of the train carriage . rachel takes chloe ’s marker and writes " rachel amber " on the floor of the train carriage with her right hand and then repeats the same successfully with her left hand . his name was bongo . he was a gift from my dad . hate to break it to you , but chloe price is n’t exactly renowned throughout arcadia bay as a bastion of trust and empathy . after they finish listening to the music , both girls take out their earbuds and chloe puts them away . i admit , it was really dumb to lock the keys in the car . stealing a dedication plate takes ... persistence . chloe uses the plate to pry open the viewfinder . she throws the plate to the ground and takes the quarter from inside the viewfinder . at chloe ’s mark , the man and the woman start kissing . oh , honey , i think we used the vibrating bed for too long . last i checked , you ’re supposed to be chloe price . or we could go try to find a liquor store instead ? the girls run to the parking lot . rachel and chloe walk down a train track . rachel is drinking the wine that the two of them stole from the picnickers and chloe is balancing on the rails . they come across the american rust salvage yard . rachel ’s been acting kind of standoffish ever since we left the park . chloe smashes the mannequin ’s head off . after some random smashing , chloe hits the truck ’s tailgate . there she sees her fathers wrecked car and drops the bat . william turns it off and looks at chloe . the raven suddenly appears on the hood of the car , and almost immediately disappears . [ex] s0 : chloe sees rachel , walking towards the oak tree , as william ’s car passes it by . [ex] s0 : chloe looks at one of the objects around her . a truck appears outside the window and crashes into the left side of william ’s car . she gets out of it and leans on the hood one last time . then she goes towards the train tracks and starts walking back to the overlook , as a raven flies overhead . my mom might skip grounding and just go straight to the death penalty . as the garbage inside the can starts burning , she takes a step back . its burning contents fall out towards the oak tree , setting it on fire . frank and his friend are hanging out next to his rv at the old mill . frank stares in shock at the fire and smoke in the distance . they both look at the fire and david puts his arm around her . [ex] s0 : nathan is sitting at the fountain , looking through his picture book . then he notices steph , mikey and drew , hanging at the picnic table far from him . james amber and principal wells are talking to a police officer at the blackwell parking lot . the woman seen kissing james is sitting on a bench at the overlook park , looking at the fire . while smoking a cigarette , she starts smiling mysteriously .

E.7 Summary from SummaRuNNer with both sentence and word level attention trained without rationales (AttnRNN)

: sc : s0 : chloe price , standing on train tracks and wearing a black hoodie , flicks her lighter a few times and lights up her cigarette . she takes a deep breath , then takes the cigarette out of her mouth and breathes out the smoke . a train begins to approach her . she then walks down toward the mill . the bouncer throws her id on the ground . chloe picks it up and walks away from him . the pitbull does n't bark at chloe . you looking to get beat ? the vendor goes talk to the truck driver . [ex] : sc : s0 : chloe sees the crowd and tries to push through it . [ex] s0 : chloe tries to leave , but the guy steps in her way . you 're trying too hard . after a few moments , the guy she ran into earlier and his friend come to confront her . they look at each other and notice guy 2 helping guy 1 to get up . the men leave and frank looks back to see that rachel and chloe are gone . if she attacked the skeevy guys , she will now have a bruise under her eye . she rolls on her side and picks up her ashtray , then she puts the ashtray below her chest and starts smoking . after a few moments she stops smoking , puts the ashtray away and sits up on her bed . [ex] s0 : chloe goes to her drawer and gets changed . i can use it to call mine , then figure out where the hell i left it . oh , can you grab my cellphone too ? [ex] s0 : chloe slips joyce 's phone into her purse and leaves the room . she then goes downstairs . you used to love to learn . david 's had some hard times , too , you know . unless he tries to give me advice . [ex] s0 : choice : slip money in joyce 's purse (stole vendor 's money and did n't buy weed from frank) she goes to his side and he starts talking . [ex] david : chloe , is that a black eye ? [ex] s0 : david goes to the garage and puts back the socket wrench inside his toolbox . then , both he and chloe get into the car . william does n't answer . chloe sees the said family photo with david replacing william . [ex] s0 : chloe hears a horn three times and approaches william in panic . [ex] s0 : chloe opens the door , gets out of the car and stands holding the door looking at david . i 'd rather have my eyes gouged out with rusted forks . stopped any gang wars lately ? so i went to the mill last night , caught firewalk live . if i had known the celestial avenger was bloodied , i would have totally given him my potion . here 's a character sheet . you stand at a three - way crossing . to your left , the raiders ' training ground . " the raiders could have some good loot at the training ground . [ex] s0 : chloe walks behind the dressing screen . a rhetorical question ? now about that eye ... that 's a hell of a battle scar . both girls take off running . rachel climbs into a carriage on the train , then helps chloe as she joins her . i think we should play two truths and a lie . but i 've passed by your locker a few times , and i 've seen that old photo of a cat you keep in there . after they finish listening to the music , both girls take out their earbuds and chloe puts them away . i 'd love to get it working for her . [ex] s0 : chloe tries opening the viewfinder with the nail file . stealing a dedication plate takes ... persistence . chloe uses the plate to pry open the viewfinder . oh , honey , i think we used the vibrating bed for too long . or we could go try to find a liquor

store instead ? talk about committing to a performance . [ex] s0 : chloe tries to snatch the wine , but the couple notices her . i think it 's contagious . you 'd better run away before it gets you too . the girls run to the parking lot . i could use a drink after trying to keep up with you . [ex] chloe : i 've heard that actors are moody , but , wow , rachel . i know i 'm not the easiest person to be around . i asked you to leave me alone . i guess it 's easier to be alone if you decide it 's a choice . then she falls on her back and continues crying on the ground . the raven suddenly appears on the hood of the car , and almost immediately disappears . [ex] s0 : chloe looks at one of the objects around her . a truck appears outside the window and crashes into the left side of william 's car . she gets out of it and leans on the hood one last time . then she goes towards the train tracks and starts walking back to the overlook , as a raven flies overhead . [ex] chloe : the ones who were making out ? [ex] i 've felt like my dad 's been lying about something for a while . so when i saw he got a text from an unknown number ... asking him to meet ... plus you came along with me , no questions asked . my mom might skip grounding and just go straight to the death penalty . they both look at the fire and david puts his arm around her . the three of them look at the fire . the woman seen kissing james is sitting on a bench at the overlook park , looking at the fire .

E.8 Summary from SummaRuNNer with both sentence and word level attention trained without rationales (AttnRNN + rationale)

chloe releases the parking brake and the car slides down . hard to get to the stage . after a few moments , the guy she ran into earlier and his friend come to confront her . frank sees them and chloe stops , looking at the guys behind him . he then jumps in front of the guys . she rolls on her side and picks up her ashtray , then she puts the ashtray below her chest and starts smoking . she then goes downstairs . he takes the toolbox from the ground and walks towards a table in the corner . then , both he and chloe get into the car . so i went to the mill last night , caught firewalk live . wait , you went to the mill last night ? rachel takes chloe 's hand and pulls her into the building . rachel has her back turned to chloe and is wearing jeans and a bra . when she 's done , the bruise is no longer visible . rachel climbs into a carriage on the train , then helps chloe as she joins her . rachel moves to sit on the floor of the train carriage . so , which is the lie ? rachel takes chloe 's marker and writes " rachel amber " on the floor of the train carriage with her right hand and then repeats the same successfully with her left hand . he was a gift from my dad . after they finish listening to the music , both girls take out their earbuds and chloe puts them away . i admit , it was really dumb to lock the keys in the car . chloe uses the plate to pry open the viewfinder . she throws the plate to the ground and takes the quarter from inside the viewfinder . she approaches rachel and holds out the quarter triumphantly . the girls see a man and a woman ,

meeting under the oak tree . at chloe 's mark , the man and the woman start kissing . oh , honey , i think we used the vibrating bed for too long . last i checked , you 're supposed to be chloe price . rachel brings chloe to the picnickers . or we could go try to find a liquor store instead ? rachel starts breathing heavily and collapses to the ground . there 's a ranger station on the other side of the park . the girls run to the parking lot . rachel takes the bottle from chloe and starts drinking , then offers it to chloe . rachel and chloe walk down a train track . rachel is drinking the wine that the two of them stole from the picnickers and chloe is balancing on the rails . they come across the american rust salvage yard . i know i 'm not the easiest person to be around . acknowledging her request , she stands up and takes the bat from chloe and examines it . i asked you to leave me alone . rachel turns away and heads back towards the tracks . but i want to find out . chloe picks the bat up from the ground and looks around angrily . chloe smashes the mannequin 's head off . after some random smashing , chloe hits the truck 's tailgate . there she sees her fathers wrecked car and drops the bat . chloe approaches the car and starts hitting its hood with her fists and crying . then she falls on her back and continues crying on the ground . " burning the midnight oil " song is still playing on the radio . william turns it off and looks at chloe . the raven suddenly appears on the hood of the car , and almost immediately disappears . in the next shot david is sitting in the driver 's seat , but in a moment he is replaced by william . this time she turns her head in chloe 's direction . the car stops next to rachel , who is looking at chloe with wide eyes . rachel puts her hand on the glass , chloe puts hers on the other side . suddenly rachel catches on fire . a truck appears outside the window and crashes into the left side of william 's car . she gets out of it and leans on the hood one last time . then she goes towards the train tracks and starts walking back to the overlook , as a raven flies overhead . rachel stands under the oak tree , crying , while chloe silently approaches her from behind . so when i saw he got a text from an unknown number ... asking him to meet ... rachel takes a picture of her and her father out of her pocket . my mom might skip grounding and just go straight to the death penalty . rachel sets the photo on fire and lets it fall into the trash can . as the garbage inside the can starts burning , she takes a step back . after a moment of hesitation , rachel kicks the trash can over . its burning contents fall out towards the oak tree , setting it on fire . rachel starts screaming loudly , and at the same time a gust of wind comes from behind her , spreading the fire to the entire tree . rachel is breathing heavily and crying . then she lets out another scream , and another gust of wind comes blowing at the fire . both girls look on in shock as the fire starts spreading to other trees . frank and his friend are hanging out next to his rv at the old mill . his friend is on the phone and frank is drinking a bottle of beer . frank stares in shock at the fire and smoke in the distance . they both look at the fire and david puts his arm around her . then he notices steph , mikey and drew , hanging at the picnic table far from him . the three of them look at the fire . james amber and principal wells are talking to a police officer at the blackwell parking lot . james looks at the fire . the

woman seen kissing james is sitting on a bench at the overlook park , looking at the fire . while smoking a cigarette , she starts smiling mysteriously .

E.9 Summary from flan-t5-base (zero shot)

chloe price is standing on train tracks and wearing a black hoodie. The bouncer throws her id on the ground and chloe picks it up and walks away from him. The bouncer and chloe are going to the firewalk concert. The vendor is selling a shirt for twenty bucks. chloe is in the crowd and tries to push through it. chloe is trying to see a band but it's too packed to get to the stage chloe is a good friend of guy 1. rachel and chloe dance together in front of the guys. The narrator's mother is in the room. chloe goes back to joyce and david's room and takes her purse. Joyce and chloe are having a conversation before school. chloe is worried about her grades and her mother's disregard for her rules. David is a good man and he's a good man. chloe and joyce are talking about their mother and dad. chloe takes the keys from the ashtray and puts them in his back pocket. chloe and david are talking about the spark plug. David finishes fixing the car and closes the hood. He takes the toolbox from the ground and walks towards a table in the corner. chloe is hurting for her ex. chloe is in panic and is unable to get out of the car. chloe and eliot are going to see the tempest in blackwell theater tomorrow. Skip and chloe are in a band Skip and chloe are at a house party. grant's chemistry class starts today. grant's sincerity is overwhelming. grant's chloe's generosity is overwhelming. grant's new donors disagree with her. People are taking victoria seriously since she won the beacon's young artist award for her photography. chloe price chloe and blackwell are discussing their drug use and tolerance issues. evan and steph are discussing the first encounter of the celestial avenger. chloe and steph are playing a game of chess The elf barbarian is a famous hero in the kingdom of avernon. elamon is a wizard and he is battling the undead. The orc sergeant is a heavyset orc sergeant and he is armed with a warhammer. chloe is about to climb the stairs but gets knocked down by nathan prescott chloe and nathan are talking about their dad's loss at the shipyard. chloe and rachel enter the hayden and dana rehearse under mr. keaton's supervision rachel and chloe are having trouble with their affections. chloe and rachel are left alone after the first period improv with freshmen. chloe looks into rachel's backpack and finds her belt. chloe and rachel are going to see the show live. rachel and chloe are going to seattle. chloe and rachel are having a conversation about the weather and the weather. rachel and chloe are having a party. chloe and rachel are on the train in arcadia bay. rachel and chloe are talking about their feelings about the truth and how they should cheat. chloe is allergic to cats and has seen an old photo of a cat she kept in her locker. rachel is impressed with chloe price's ability to lie. chloe is unsure of her own identity and is not sure who to call if she needs to get out of a ticket. chloe and rachel are sharing a music playlist. rachel

and chloe are playing a game of improvisation. chloe and rachel are discussing how to use a viewfinder. chloe is angry at the man for not coming to her party. rachel and chloe are playing a game of spying on people. chloe and rachel are ditching school for hours. The man and woman are going to try to save a lifeguard from drowning in the park. The girl is in trouble and she is waiting for help. The man is going to check her pulse and he will keep watch. The woman is going to the ranger station and she will keep watch. chloe and rachel are drinking and s0 is trying to keep up with them. rachel and chloe are exploring the junkyard. chloe is alone and wants to be alone. rachel is leaving and chloe is waiting for her. rachel and chloe are talking about their relationship. chloe is crying and hitting the car hood with her fists and crying on the ground. chloe is crying and rachel is crying. chloe and rachel are upset about their dad's death. rachel owes chloe an apology. rachel and chloe are going to leave this place forever. nathan and steph look at the fire.

E.10 Summary from flan-t5-large (zero shot)

chloe is standing on train tracks and lights up her cigarette. she jumps out of the way of a train and runs toward the sawmill. Chloe is a fake. She is trying to get in a club. chloe is going to see firewalk tonight. Chloe is looking for a shirt and the vendor is selling it for twenty bucks. She approaches the vendor's car and he slaps her hand away. Chloe is in a crowd and bumps into a man. Chloe is trying to see a band at a concert but it's too packed. chloe spilled beer on guy 1 and his friend rachel and chloe dance together in front of the stage Chloe is looking for her mom's phone. She finds it in her room. Chloe takes joyce's purse and phones her mom. chloe and david used to love to learn and think drugs were lame. Joyce made breakfast for them. Chloe is angry with her mother Joyce for not letting her go to school. She is also worried about her grades and disregard for rules. chloe is lonely and her father is a good man. Joyce expects her to listen to David when he takes her to school today. chloe's mom asks her to bring david his keys from the ashtray. chloe throws the keys to david and they drive away. chloe is trying to get the socket wrench from david's toolbox. chloe and david are fixing a car. chloe is tired of staying out late and drinking. david wants her to change and put someone else first. chloe is in the car with her dad and her brother david when a truck crashes into the left side of the car, hitting william and then everything goes black. chloe and eliot are going to the tempest tomorrow. skip asked justin williams'mom to move her mercedes out of handicapped parking. chloe went to the mill last night to catch firewalk live. s0 and chloe are talking about a party at the old mill up north. Chloe will be in her seat by the time chemistry class begins today. Grant is surprised by Chloe's sincerity. victoria is being courted by people who are interested in her. a student is attacked by two men on stage chloe is accused of drug use and principal wells has it out for her. evan and steph discuss the importance of

fire in their game. Chloe is a gamer. She wants to join a game with steph. steph and mikey are two famous heroes in the kingdom of avernon. elamon is an elf barbarian who has been trained by steph ergel's acid blast is used to kill the orc sergeant. chloe is about to climb the stairs when she gets knocked down by nathan prescott who is being chased by drew north. nathan is upset with his father for trying to buy off the coach. chloe and rachel enter the school entrance dressed in a costume rachel apologizes to mr. keaton for interrupting her performance. dana, hayden and mr. keaton leave the school. chloe and rachel are left alone. Chloe gives rachel her belt. Chloe and rachel are dressed in costumes. Chloe likes rachel's costume. rachel and chloe are going on a field trip. Chloe and rachel are running after a train. They get on a train and head north. Chloe is hanging out with Rachel. She used to have a friend named Max, but she left for northern pastures. Chloe and Rachel are on a train. rachel is on a train with chloe. rachel is a california girl who was born and raised in new york. Chloe is mysterious and has a cat allergy. Rachel thinks she is allergic to cats. rachel is impressed with chloe's work. Chloe and Rachel are listening to music together. Chloe and Rachel are playing a game of improvisation. Chloe is not an actor. chloe is trying to open a viewfinder with her bare hands. rachel suggests using a nail file. Chloe steals a dedication plate from rachel. Chloe and Rachel are playing a game of spying on people. chloe is having an affair with a man and woman on vibrating beds The man and woman are trying to save a life of a picnicker. Chloe is in trouble and needs help. The man will check her pulse. Chloe and Rachel are running away from a contagious virus. rachel and chloe explore an American rust salvage yard. rachel is acting standoffish since they left the park. chloe is angry with rachel and wants to be left alone. Chloe is leaving and wants to make a real friendship with Rachel. chloe is angry at rachel for not being her friend. chloe wakes up in william's car crying and screaming. chloe wakes up in her father's wrecked car Chloe and Rachel are talking about their father. rachel and chloe are having an intense and new relationship. rachel and chloe are going to leave the place they are in and go to another one. a man kisses a woman in the park and looks at the fire

E.11 Summary from flan-t5-base (fine-tuned)

a train begins to approach her. [ex] : sc : man 2 : i understand, i understand, really, an — and it won't ever happen again, i swear! [ex] : sc : chloe : what's her name? [ex] : sc : chloe : hey, are you selling any —?. [ex] : sc : vendor : if you don't have twenty bucks, beat it. [ex] : sc : s0 : chloe sees the crowd and tries to push through it. [ex] s0 : chloe looks at the stairwell near the entrance. [ex] : sc : s0 : chloe manages to go upstairs and see the band. [ex] s0 : rachel and chloe smile at them. [ex] chloe : (thinking) i smell like cigarettes and beer. [ex] s0 : chloe goes back to joyce and david's room and takes her purse. i made you breakfast! [ex] chloe : i 'll enjoy that,

huh?. is that the band you mentioned last week? [ex] s0 : choice : dad was a good man. [ex] chloe : (thinking) bringing david his keys is about the most humiliating thing mom could ask me to do. [ex] : sc : s0 : chloe leaves the house and sees david. [ex] chloe : (thinking) what's david doing, leaving stuff in our garage anyway? [ex] s0 : david goes to the garage and puts back the socket wrench inside his toolbox. [ex] : sc : chloe : dad, turn it up! [ex] s0 : chloe looks at the purse beside her. [ex] s0 : choice : (didn't see the photo rachel posted) [ex] chloe : who exactly is talking shit about me? [ex] s0 : choice : firewalk show. after it ends he puts his phone in his back pocket. [ex] chloe : you don't think more money should be spent in the arts? [ex] : sc : samantha : hey, chloe. [ex] : sc : s0 : this conversation occurs only if chloe sat on the crate on the stage. [ex] s0 : chloe stays silent. [ex] : sc : s0 : choice : (first encounter) [ex] mikey : if i had known the celestial avenger was bloodied, i would have totally given him my potion. [ex] chloe : dream life over real life, that's my motto. 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. i have heard of you in the castle. [ex] chloe : you see why i haven't really needed a partner. [ex] : sc : s0 : chloe is about to climb the stairs, but gets knocked down by nathan prescott, who is being chased by drew north. [ex] s0 : nathan grabs onto his book but drew moves it away. [ex] s0 : chloe approaches the school entrance, but rachel, dressed in a costume, opens the door first from the other side. i mean... does she really mean that? i have first period improv with a class full of freshmen, now. [ex] s0 : chloe looks into rachel's backpack. [ex] s0 : choice : (attacked skeevy guys) [ex] rachel : i hoped you 'd say that. [ex] : sc : s0 : chloe and rachel nearby the water tower at the train tracks. [ex] s0 : choice : (brought the belt to rachel) [ex] chloe : it's nice rachel we're having. [ex] chloe : i owe you, that's for sure. [ex] s0 : chloe moves to join rachel on the floor of the train carriage. [ex] chloe : (thinking) something about myself... should i start things off with a lie, or with the truth? [ex] rachel : which brings me to your alleged cat allergy. [ex] rachel : when your dad is the district attorney, i guess lying is... something you're used to. [ex] s0 : chloe smiles, takes an earbud from chloe and puts it in. luckily, we've got some high - tech surveillance equipment right here. [ex] chloe : i guess you could stab someone with a nail file... oh, right. we might have some leftover food when we're done. [ex] : sc : : sc : s0 : choice : statue [ex] chloe : they totally stole my third - grade art project. i'm totally seeing double. [ex] s0 : rachel approaches the picnickers. there's a ranger station on the other side of the park. [ex] s0 : chloe grabs the bottle and both girls run away. they come across the american rust salvage yard. [ex] s0 : rachel lets out a very frustrated groan, returning to her seat. [ex] rachel : i thought chloe price doesn't need friends. i don't want to be alone anymore. [ex] : sc : s0 : chloe wakes up in william 's car again. [ex] : sc : s0 : rachel stands under the oak tree, crying, while chloe silently approaches her from behind. i love him, and i never want to see his fucking face again. [ex] rachel : [taking the photo] chloe, i owe you an apology. [ex] chloe : so if i came to you tomorrow and told you to pack your

bags... [ex] rachel : i'm serious. the three of them look at the fire.

E.12 Summary from flan-t5-base (Encoder only) trained without rationales

i heard firewalk is playing here tonight . she reaches the shirt and the vendor slaps her hand away . hard to get to the stage . after a few moments , the guy she ran into earlier and his friend come to confront her . guy 2 goes to help guy 1 who 's on the floor and chloe runs to rachel . chloe stops and looks back to see if the guys are coming . wakey , wakey , eggs and bakey ! how drunk was i last night ? i said breakfast ! i can use it to call mine , then figure out where the hell i left it . no more dawdling , i need to talk to you ! you wanted to talk ? why ca n't we just have some pleasant conversation before school ? is that the band you mentioned last week ? just let me know so i can stop fighting with blackwell to keep you on scholarship . if he 's kind enough to share his experience , i expect you to listen ... [ex] chloe : like i 'd let him get within fifteen feet of me . unless he tries to give me advice . no point in putting off the inevitable torture of driving to school with david . you know what a spark plug does ? he takes the toolbox from the ground and walks towards a table in the corner . you 've enjoyed enough of a vacation from having a father figure . so there 's some things i want to be real clear about ... [ex] s0 : choice : ignore and endure . chloe sees the said family photo with david replacing william . eliot sees her , puts down the book he 's reading , and approaches her . do you need to go to the nurse or something ? stopped any gang wars lately ? i did n't know you were into music like that . if pisshead came on the radio , i 'd turn that shit up . i recently made the case that stem programs should receive more support , but apparently our new donors disagree with me . what are you reading ? you know blackwell has a zero - tolerance policy . knobcone pine cones , for example , which require temperatures above 350 degrees to open . what else have you got to do before class ? i am elamon , wizard of the third circle , foremost advisor to king tiberius , and sworn defender of avernon . " i have heard of you in the castle . king tiberius owes you his life . the raiders could have some good loot at the training ground . where do you wish to go ? there are a dozen raiders on the training field , all of whom raise their weapons and charge ! there 's a sweet and sour kind of smell as the flesh melts off their bones like warm candle wax . guess nathan prescott made the shitlist . you do n't have your dad try to buy off the coach . how much financial aid does your deadbeat dad need again ? you ca n't be a part of the team and be into this stupid crap at the same time . keaton : hayden ! you 've had weeks to be off book ! i have first period improv with a class full of freshmen , now . chloe and rachel are left alone . why does it have to end ? rachel climbs into a carriage on the train , then helps chloe as she joins her . it 's a game where each person offers up three facts about themselves , two of which are the truth and one of which is

... [ex] chloe : a lie ? russia , greece ... kathmandu ? he was a gift from my dad . after they finish listening to the music , both girls take out their earbuds and chloe puts them away . luckily , we 've got some high - tech surveillance equipment right here . how about a nail file ? wonder if i can find something sturdier to use ? i admit , it was really dumb to lock the keys in the car . no , i 'm not — what kind of food ? the honor student wants to show the school delinquent how to party ? or we could go try to find a liquor store instead ? talk about committing to a performance . yeah , i guess i 'd remember something like that . uh , i do n't remember how to do cpr , waahh . " there 's a ranger station on the other side of the park . you can do this ... i ca n't do it ! this woman needs help from someone who actually knows what they 're doing . i could use a drink after trying to keep up with you . rachel is drinking the wine that the two of them stole from the picnickers and chloe is balancing on the rails . they come across the american rust salvage yard . i do n't exactly have tons of experience with the whole friendship thing . but i want to find out . burning the midnight oil " song is still playing on the radio . then she goes towards the train tracks and starts walking back to the overlook , as a raven flies overhead . so when i saw he got a text from an unknown number ... asking him to meet ... i thought i could catch him , or something ... [ex] rachel : chloe , i love my dad . it started raining , and i fell and broke my arm three miles from the car . i still remember the smell of his coat , and how calm he was , and the sound of his voice , and ... [ex] s0 : choice : comfort her . we have all the time in the world to figure out whatever this is . his friend is on the phone and frank is drinking a bottle of beer . frank stares in shock at the fire and smoke in the distance . the three of them look at the fire .

E.13 Summary from flan-t5-base (Encoder only) trained with rationales

chloe takes another couple hits from her cigarette before letting it fall in between the tracks . she jumps out of the way of the train at the last second , watching it go by , then takes off her hood and looks at the sawmill across from her . because trust with me is earned in actions , not words . i heard firewalk is playing here tonight . she reaches the shirt and the vendor slaps her hand away . hard to get to the stage . chloe stops and looks back to see if the guys are coming . frank sees them and chloe stops , looking at the guys behind him . frank looks at the guys and back to chloe and rachel . the night ends with chloe making one last pose before going back to dancing . she rolls on her side and picks up her ashtray , then she puts the ashtray below her chest and starts smoking . another day in paradise , right ? i can use it to call mine , then figure out where the hell i left it . she then gets her mom 's phone from the nightstand and unlocks it . and do n't think you can get out of school just because you missed the bus again . but sometimes we need to make more room in our hearts for new people . but

what 's the point in getting along if it means pretending everything 's fine when it 's not ? you 'll need to bring him his keys from the ashtray . unless he tries to give me advice . no point in putting off the inevitable torture of driving to school with david . he takes it and goes back to fixing his car . he takes the toolbox from the ground and walks towards a table in the corner . then , both he and chloe get into the car . chloe sees the said family photo with david replacing william . eliot sees her , puts down the book he 's reading , and approaches her . principal wells approaches chloe and she gets up from the crate , jumping off the stage and landing in front of him . your words alone have convinced me of your guilt . you will meet me in my office after school for a formal reprimand . what else have you got to do before class ? you are an elf barbarian . 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 . alone , you have fought your way through the raider camps , seeking their warlord leader , duurgaron the unscarred . as you enter the final camp , bloodied and weary , you see your fellow hero approaching from the opposite direction . i am elamon , wizard of the third circle , foremost advisor to king tiberius , and sworn defender of avernon . " elamon narrows his eyes at the elf in front of him and says , " i am here to defeat duurgaron the unscarred in the name of king tiberius . i have heard of you in the castle . king tiberius owes you his life . every raider suddenly starts screaming and writhing in pain . there 's a sweet and sour kind of smell as the flesh melts off their bones like warm candle wax . the orc clutches his groin , never to father children again . guess nathan prescott made the shitlist . you do n't have your dad try to buy off the coach . how much financial aid does your deadbeat dad need again ? [ex] s0 : rachel starts spinning , showing off her costume . i have first period improv with a class full of freshmen , now . chloe and rachel are left alone . first she pulls out a photo of a young rachel with her father . rachel has her back turned to chloe and is wearing jeans and a bra . why does it have to end ? you 're on a freaking train with rachel freaking amber . it 's a game where each person offers up three facts about themselves , two of which are the truth and one of which is ... [ex] chloe : a lie ? and then the other person has to guess which is which . [ex] chloe : all the bad girls do it . he was a gift from my dad . car first , embarrassing number of moving violations second . after they finish listening to the music , both girls take out their earbuds and chloe puts them away . this is one i learned in theater class . it always looks so easy in the movies . she throws the plate to the ground and takes the quarter from inside the viewfinder . yet we 've been ditching now for hours and we have n't even gotten wasted yet . the honor student wants to show the school delinquent how to party ? talk about committing to a performance . there 's a ranger station on the other side of the park . this woman needs help from someone who actually knows what they 're doing . rachel takes the bottle from chloe and starts drinking , then offers it to chloe . i could use a drink after trying to keep up with you . rachel is drinking the wine that the two of them stole from the picnickers and chloe is balancing on the rails . they come across the american

rust salvage yard . i do n’t exactly have tons of experience with the whole friendship thing . but i want to find out . there she sees her fathers wrecked car and drops the bat . burning the midnight oil " song is still playing on the radio . this time she turns her head in chloe ’s direction . she gets out of it and leans on the hood one last time . then she goes towards the train tracks and starts walking back to the overlook , as a raven flies overhead . so when i saw he got a text from an unknown number ... asking him to meet ... i thought i could catch him , or something ... [ex] rachel : chloe , i love my dad . i still remember the smell of his coat , and how calm he was , and the sound of his voice , and ... [ex] s0 : choice : comfort her . its burning contents fall out towards the oak tree , setting it on fire . his friend is on the phone and frank is drinking a bottle of beer . frank stares in shock at the fire and smoke in the distance . they both look at the fire and david puts his arm around her .

References

[1] Espen J Aarseth. *Cybertext: Perspectives on ergodic literature*. JHU Press, 1997.

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

[3] Sanchit Agarwal, Nikhil Kumar Singh, and Priyanka Meel. 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, 2018. .

[4] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 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, October 2005. Association for Computational Linguistics. URL https://aclanthology.org/H05-1073.

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

[6] Prithviraj Ammanabrolu, Wesley Cheung, Dan Tu, William Broniec, and Mark Riedl. Bringing stories alive: Generating interactive fiction worlds. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 16(1):3–9, Oct. 2020. URL https://ojs.aaai.org/index.php/AIIDE/article/view/7400.

[7] Nantheera Anantrasirichai and David Bull. Artificial intelligence in the creative industries: a review. *Artificial intelligence review*, pages 1–68, 2022.

[8] Sergi Andreu and Monica Villanueva Aylagas. Neural synthesis of sound effects using flow-based deep generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 2–9, 2022.

[9] Diego Antognini and Boi Faltings. 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, November 2019. Association for Computational Linguistics. . URL https://aclanthology.org/D19-5404.

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

[11] Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. Marta: Leveraging human rationales for explainable text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35 (7):5868–5876, May 2021. . URL https://ojs.aaai.org/index.php/AAAI/article/view/16734.

[12] Hadi Askari, Anshuman Chhabra, Muhao Chen, and Prasant Mohapatra. Assessing llms for zero-shot abstractive summarization through the lens of relevance paraphrasing. *arXiv preprint arXiv:2406.03993*, 2024.

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

[14] Sasha Azad, Jennifer Wellnitz, Luis Garcia, and Chris Martens. Anthology: a social simulation framework. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 224–231, 2022.

[15] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[16] David Bamman and Noah A Smith. New alignment methods for discriminative book summarization. *arXiv preprint arXiv:1305.1319*, 2013.

[17] David Bamman, Brendan O'Connor, and Noah A Smith. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, 2013.

[18] 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. Bardic: Generating multimedia narratives for game logs. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 13(2): 154–161, Jun. 2021. URL https://ojs.aaai.org/index.php/AIIDE/article/view/12987.

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

[20] Sebastian Hurup Bevensee, Kasper Alexander Dahlsgaard Boisen, Mikael Peter Olsen, Henrik Schoenau-Fog, and Luis Emilio Bruni. Project aporia – an exploration of narrative understanding of environmental storytelling in an open world scenario. In David Oyarzun, Federico Peinado, R. Michael Young, Ane Elizalde, and Gonzalo Méndez, editors, *Interactive Storytelling*, pages 96–101, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-34851-8.

[21] Sebastian Hurup Bevensee, Kasper Alexander Dahlsgaard Boisen, Mikael Peter Olsen, Henrik Schoenau-Fog, and Luis Emilio Bruni. Aporia – exploring continuation desire in a game focused on environmental storytelling. In David Oyarzun, Federico Peinado, R. Michael Young, Ane Elizalde, and Gonzalo Méndez, editors, *Interactive Storytelling*, pages 42–47, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-34851-8.

[22] Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland, May 2022. Association for Computational Linguistics. . URL https://aclanthology.org/2022.acl-long.269.

[23] Michal Bída, Martin Černý, and Cyril Brom. 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, 2013. Springer-Verlag. ISBN 9783319027555. . URL https://doi.org/10.1007/978-3-319-02756-2_11.

[24] Oloff C. Biermann, Ning F. Ma, and Dongwook Yoon. From tool to companion: Storywriters want ai writers to respect their personal values and writing strategies. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, DIS '22, page 1209–1227, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393584. . URL https://doi.org/10.1145/3532106.3533506.

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

[26] 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. *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, Online,

July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.nuse-1.0.

[27] Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 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, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.194.

[28] Adi Botea and Vadim Bulitko. Tiered state expansion in optimization crosswords. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 79–86, 2022.

[29] Jeanne H. Brockmyer, Christine M. Fox, Kathleen A. Curtiss, Evan McBroom, Kimberly M. Burkhart, and Jacquelyn N. Pidruzny. The development of the game engagement questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, 45(4):624–634, 2009. ISSN 0022-1031. . URL https://www.sciencedirect.com/science/article/pii/S0022103109000444.

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

[31] Luis Emilio Bruni and Sarune Baceviciute. Narrative intelligibility and closure in interactive systems. In Hartmut Koenitz, Tonguc Ibrahim Sezen, Gabriele Ferri, Mads Haahr, Digdem Sezen, and Güven Çatak, editors, *Interactive Storytelling*, pages 13–24, Cham, 2013. Springer International Publishing. ISBN 978-3-319-02756-2.

[32] Luis Emilio Bruni, Sarune Baceviciute, and Mohammed Arief. Narrative cognition in interactive systems: Suspense-surprise and the p300 erp component. In Alex Mitchell, Clara Fernández-Vara, and David Thue, editors, *Interactive Storytelling*, pages 164–175, Cham, 2014. Springer International Publishing. ISBN 978-3-319-12337-0.

[33] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[34] Sven Buechel and Udo Hahn. 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,

April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2092.

[35] Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 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, October-November 2018. Association for Computational Linguistics. . URL https://aclanthology.org/D18-1507.

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

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

[38] Tara Capel and Margot Brereton. What is human-centered about human-centered ai? a map of the research landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. . URL https://doi.org/10.1145/3544548.3580959.

[39] Elin Carstensdottir, Erica Kleinman, and Magy Seif El-Nasr. Player interaction in narrative games: structure and narrative progression mechanics. In *Proceedings of the 14th international conference on the foundations of digital games*, pages 1–9, 2019.

[40] Elin Carstensdottir, Nathan Partlan, Steven Sutherland, Tyler Duke, Erika Ferris, Robin M. Richter, Maria Valladares, and Magy Seif El-Nasr. 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, 2020. Association for Computing Machinery. ISBN 9781450367080. . URL https://doi.org/10.1145/3313831.3376527.

[41] Ashley Castleberry and Amanda Nolen. Thematic analysis of qualitative research data: Is it as easy as it sounds? *Currents in Pharmacy Teaching and Learning*, 10:807–815, 6 2018. ISSN 1877-1297. .

[42] Marc Cavazza and R. Michael Young. *Introduction to Interactive Storytelling*, pages 377–392. Springer Singapore, Singapore, 2017. ISBN 978-981-4560-50-4. . URL https://doi.org/10.1007/978-981-4560-50-4_55.

[43] Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1662–1675, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. . URL https://aclanthology.org/N18-1150.

[44] Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL https://aclanthology.org/P08-1090.

[45] M Charity and Julian Togelius. Aesthetic bot: interactively evolving game maps on twitter. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 18–25, 2022.

[46] Megan Charity, Ahmed Khalifa, and Julian Togelius. Baba is y'all: Collaborative mixed-initiative level design. In *2020 IEEE Conference on Games (CoG)*, pages 542–549, 2020. .

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

[48] Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 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, September 2017. Association for Computational Linguistics. . URL https://aclanthology.org/D17-1168.

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

[50] Laifu Chen and Minh Le Nguyen. Sentence selective neural extractive summarization with reinforcement learning. In *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–5, 2019. .

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

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

[53] Anshuman Chhabra, Hadi Askari, and Prasant Mohapatra. Revisiting zero-shot abstractive summarization in the era of large language models from the perspective of position bias. *arXiv preprint arXiv:2401.01989*, 2024.

[54] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. . URL https://aclanthology.org/D14-1179.

[55] Seungtaek Choi, Haeju Park, Jinyoung Yeo, and Seung-won Hwang. 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, November 2020. Association for Computational Linguistics. . URL https://aclanthology.org/2020.emnlp-main.543.

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

[57] 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. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

[58] John Joon Young Chung, Shiqing He, and Eytan Adar. The intersection of users, roles, interactions, and technologies in creativity support tools. In *Designing Interactive Systems Conference 2021*, pages 1817–1833, 2021.

[59] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[60] Davide Colla, Enrico Mensa, and Daniele P. Radicioni. Novel metrics for computing semantic similarity with sense embeddings. *Knowledge-Based Systems*, 206:106346, 2020. ISSN 0950-7051. . URL https://www.sciencedirect.com/science/article/pii/S0950705120305025.

[61] Michael Cook. Puck: a slow and personal automated game designer. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 232–239, 2022.

[62] Seth Cooper. Sturgeon: tile-based procedural level generation via learned and designed constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 26–36, 2022.

[63] Flávio Coutinho and Luiz Chaimowicz. On the challenges of generating pixel art character sprites using gans. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 87–94, 2022.

[64] Daniel Cox. We make how we learn: The role of community in authoring tool longevity. In *The Authoring Problem: Challenges in Supporting Authoring for Interactive Digital Narratives*, pages 65–72. Springer, 2023.

[65] Peng Cui, Le Hu, and Yuanchao Liu. Enhancing extractive text summarization with topic-aware graph neural networks. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5360–5371, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. . URL https://aclanthology.org/2020.coling-main.468.

[66] Edirlei Soares de Lima, Bruno Feijó, Simone Barbosa, Antonio L. Furtado, Angelo Ciarlini, and Cesar Pozzer. Draw your own story: Paper and pencil interactive storytelling. In Junia Coutinho Anacleto, Sidney Fels, Nicholas Graham, Bill Kapralos, Magy Saif El-Nasr, and Kevin Stanley, editors, *Entertainment Computing – ICEC 2011*, pages 1–12, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-24500-8.

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

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

[69] Pierre Le Pelletier de Woillemont, Rémi Labory, and Vincent Corruble. Automated play-testing through rl based human-like play-styles generation. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 146–154, 2022.

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

[71] Andrea Di Pastena, Dennis Jansen, Brian de Lint, and Amanda Moss. "the link out". In Rebecca Rouse, Hartmut Koenitz, and Mads Haahr, editors, *Interactive Storytelling*, pages 206–216, Cham, 2018. Springer International Publishing. ISBN 978-3-030-04028-4.

[72] Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 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, October-November 2018. Association for Computational Linguistics. . URL https://aclanthology.org/D18-1409.

[73] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[74] Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679, 2021. ISSN 0957-4174. . URL https://www.sciencedirect.com/science/article/pii/S0957417420305030.

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

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

[77] Angela Fan, Mike Lewis, and Yann Dauphin. 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, July 2018. Association for Computational Linguistics. . URL https://aclanthology.org/P18-1082.

[78] Rachelyn Farrell, Mira Fisher, and Stephen G Ware. Salience vectors for measuring distance between stories. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 95–104, 2022.

[79] Lucas N Ferreira, Lili Mou, Jim Whitehead, and Levi HS Lelis. Controlling perceived emotion in symbolic music generation with monte carlo tree search. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 163–170, 2022.

[80] Rui Figueiredo and Ana Paiva. "i'm sure i made the right choice!" - towards an architecture to influence player's behaviors in interactive stories. In Mei Si, David Thue, Elisabeth André, James C. Lester, Theresa Jean Tanenbaum, and Veronica Zammitto, editors, *Interactive Storytelling*, pages 152–157, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-25289-1.

[81] Lucie Flekova and Iryna Gurevych. 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, September 2015. Association for Computational Linguistics. . URL https://aclanthology.org/D15-1208.

[82] Jonas Freiknecht and Wolfgang Effelsberg. Procedural generation of interactive stories using language models. In *Proceedings of the 15th International Conference on the Foundations of Digital Games*, FDG '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450388078. . URL https://doi.org/10.1145/3402942.3409599.

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

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

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

[86] F.A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: continual prediction with lstm. In *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, volume 2, pages 850–855 vol.2, 1999. .

[87] Philip John Gorinski and Mirella Lapata. 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, May–June 2015. Association for Computational Linguistics. . URL https://aclanthology.org/N15-1113.

[88] Alex Goslen, Dan Carpenter, Jonathan Rowe, Roger Azevedo, and James Lester. Robust player plan recognition in digital games with multi-task multi-label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 105–112, 2022.

[89] Daniel Green. Novella: An authoring tool for interactive storytelling in games. In Rebecca Rouse, Hartmut Koenitz, and Mads Haahr, editors, *Interactive Storytelling*, pages 556–559, Cham, 2018. Springer International Publishing. ISBN 978-3-030-04028-4.

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

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

[92] Arne Grinder-Hansen and Henrik Schoenau-Fog. The elements of a narrative environment. In Hartmut Koenitz, Tonguc Ibrahim Sezen, Gabriele Ferri, Mads Haahr, Digdem Sezen, and Güven Çatak, editors, *Interactive Storytelling*, pages 186–191, Cham, 2013. Springer International Publishing. ISBN 978-3-319-02756-2.

[93] Nianlong Gu, Elliott Ash, and Richard Hahnloser. MemSum: Extractive summarization of long documents using multi-step episodic Markov decision processes. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6507–6522, Dublin, Ireland, May 2022. Association for Computational Linguistics. . URL https://aclanthology.org/2022.acl-long.450.

[94] Xinyu Guan, Qinke Peng, Xintong Li, and Zhibo Zhu. 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, 2019. .

[95] Yong Guan, Shaoru Guo, Ru Li, Xiaoli Li, and Hongye Tan. Frame semantic-enhanced sentence modeling for sentence-level extractive text summarization. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4045–4052, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. . URL https://aclanthology.org/2021.emnlp-main.331.

[96] Anisha Gupta, Dan Carpenter, Wookhee Min, Jonathan Rowe, Roger Azevedo, and James Lester. Enhancing multimodal goal recognition in open-world games with natural language player reflections. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 37–44, 2022.

[97] Vivek Gupta, Prerna Bharti, Pegah Nokhiz, and Harish Karnick. 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, August 2021. Association for Computational Linguistics. . URL https://aclanthology.org/2021.acl-srw.30.

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

[99] Hákon Jarl Hannesson, Thorbjørn Reimann-Andersen, Paolo Burelli, and Luis Emilio Bruni. Connecting the dots: Quantifying the narrative experience in interactive media. In Henrik Schoenau-Fog, Luis Emilio Bruni, Sandy Louchart, and Sarune Baceviciute, editors, *Interactive Storytelling*, pages 189–201, Cham, 2015. Springer International Publishing. ISBN 978-3-319-27036-4.

[100] Charlie Hargood, David E. Millard, Alex Mitchell, and Ulrike Spierling. *The Authoring Problem: An Introduction*, pages 1–13. Springer International Publishing, Cham, 2022. ISBN 978-3-031-05214-9. . URL https://doi.org/10.1007/978-3-031-05214-9_1.

[101] Benjamin Hättasch, Nadja Geisler, Christian M. Meyer, and Carsten Binnig. Summarization beyond news: The automatically acquired fandom corpora. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6700–6708, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.827.

[102] Lei He, Wei Li, and Hai Zhuge. 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, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1098.

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

[104] Andrew Hoyt, Matthew Guzdial, Yalini Kumar, Gillian Smith, and Mark O Riedl. Integrating automated play in level co-creation. *arXiv preprint arXiv:1911.09219*, 2019.

[105] Zhichao Hu and Marilyn Walker. 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, August 2017. Association for Computational Linguistics. . URL https://aclanthology.org/W17-5540.

[106] Matthew L. Jockers and David Mimno. Significant themes in 19th-century literature. *Poetics*, 41(6):750–769, 2013. ISSN 0304-422X. . URL https://www.sciencedirect.com/science/article/pii/S0304422X13000673. Topic Models and the Cultural Sciences.

[107] Joey Donald Jones. Authorial burden. In *The Authoring Problem: Challenges in Supporting Authoring for Interactive Digital Narratives*, pages 47–63. Springer, 2023.

[108] Nic Junius, Michael Mateas, Noah Wardrip-Fruin, and Elin Carstensdottir. Playing with the strings: Designing puppitor as an acting interface for digital games. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 18(1):250–257, Oct. 2022. . URL https://ojs.aaai.org/index.php/AIIDE/article/view/21970.

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

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

[111] Anna Kazantseva and Stan Szpakowicz. Summarizing short stories. *Computational Linguistics*, 36(1):71–109, 2010. . URL https://aclanthology.org/J10-1003.

[112] Nazanin Yousefzadeh Khameneh and Matthew Guzdial. World models with an entity-based representation. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 215–222, 2022.

[113] Muhammad Junaid Khan, Syed Hammad Ahmed, and Gita Sukthankar. Transformer-based value function decomposition for cooperative multi-agent reinforcement learning in starcraft. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 113–119, 2022.

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

[115] Evgeny Kim, Sebastian Padó, and Roman Klinger. 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, August 2017. Association for Computational Linguistics. . URL https://aclanthology.org/W17-2203.

[116] Barbara Kitchenham. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26, 2004.

[117] Sofia Kitromili and María Cecilia Reyes. Understanding the process of authoring. In *The Authoring Problem: Challenges in Supporting Authoring for Interactive Digital Narratives*, pages 17–30. Springer, 2023.

[118] Erica Kleinman, Karina Caro, and Jichen Zhu. From immersion to metagaming: Understanding rewind mechanics in interactive storytelling. *Entertainment Computing*, 33:100322, 2020. ISSN 1875-9521. . URL https://www.sciencedirect.com/science/article/pii/S1875952117301167.

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

[120] Hartmut Koenitz. Towards a theoretical framework for interactive digital narrative. In *Interactive Storytelling: Third Joint Conference on Interactive Digital Storytelling, ICIDS 2010, Edinburgh, UK, November 1-3, 2010. Proceedings 3*, pages 176–185. Springer, 2010.

[121] Hartmut Koenitz. Towards a specific theory of interactive digital narrative. In *Interactive digital narrative*, pages 91–105. Routledge, 2015.

[122] Hartmut Koenitz, Teun Dubbelman, and Christian Roth. An educational program in interactive narrative design. In *Interactive Storytelling: 12th International Conference on Interactive Digital Storytelling, ICIDS 2019, Little Cottonwood Canyon, UT, USA, November 19–22, 2019, Proceedings 12*, pages 22–25. Springer, 2019.

[123] Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Comput. Surv.*, 55(8), dec 2022. ISSN 0360-0300. . URL https://doi.org/10.1145/3545176.

[124] Lobke Kolhoff and Frank Nack. How relevant is your choice? In Rogelio E. Cardona-Rivera, Anne Sullivan, and R. Michael Young, editors, *Interactive Storytelling*, pages 73–85, Cham, 2019. Springer International Publishing. ISBN 978-3-030-33894-7.

[125] Max Kreminski, Melanie Dickinson, Noah Wardrip-Fruin, and Michael Mateas. Loose ends: a mixed-initiative creative interface for playful storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 120–128, 2022.

[126] Max Kreminski, Noah Wardrip-Fruin, and Michael Mateas. Authoring for story sifters. In *The Authoring Problem: Challenges in Supporting Authoring for Interactive Digital Narratives*, pages 207–220. Springer, 2023.

[127] K Yu Kristen, Matthew Guzdial, Nathan R Sturtevant, Morgan Cselinacz, Chris Corfe, Izzy Hubert Lyall, and Chris Smith. Adventures of ai directors early in the development of nightingale. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 70–77, 2022.

[128] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, November 2020. Association for Computational Linguistics. . URL https://aclanthology.org/2020.emnlp-main.750.

[129] Vonnegut Kurt. Shapes of stories. URL https://www.youtube.com/watch?v=oP3c1h8v2ZQ.

[130] Jingun Kwon, Naoki Kobayashi, Hidetaka Kamigaito, and Manabu Okumura. Considering nested tree structure in sentence extractive summarization with pre-trained transformer. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4039–4044, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. . URL https://aclanthology.org/2021.emnlp-main.330.

[131] Vincent Labatut and Xavier Bost. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys (CSUR)*, 52(5):1–40, 2019.

[132] Vincent Labatut and Xavier Bost. Extraction and analysis of fictional character networks: A survey. *ACM Comput. Surv.*, 52(5), September 2019. ISSN 0360-0300. . URL https://doi.org/10.1145/3344548.

[133] Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown. 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, July 2020. Association for Computational Linguistics. . URL https://aclanthology.org/2020.acl-main.453.

[134] Eric W Lang and R Michael Young. Rpgpref: a planning heuristic that uses playstyle preferences to model player action and choice. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 129–136, 2022.

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

[136] Tinea Larsson, Jose Font, and Alberto Alvarez. Towards ai as a creative colleague in game level design. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 137–145, 2022.

[137] Lee T Lemon and Marion J Reis. *Russian formalist criticism: Four essays*, volume 405. U of Nebraska Press, 1965.

[138] Boyang Li, Beth Cardier, Tong Wang, and Florian Metze. 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, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1520.

[139] Zhongyang Li, Tongfei Chen, and Benjamin Van Durme. Learning to rank for plausible plausibility. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4818–4823, Florence, Italy, July 2019. Association for Computational Linguistics. . URL https://aclanthology.org/P19-1475.

[140] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

[141] Grace Lin and Marilyn Walker. All the world's a stage: Learning character models from film. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 7(1):46–52, Oct. 2011. . URL https://ojs.aaai.org/index.php/AIIDE/article/view/12431.

[142] Zhiyu Lin, Rohan Agarwal, and Mark Riedl. Creative wand: a system to study effects of communications in co-creative settings. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 45–52, 2022.

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

[144] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

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

[146] Owen Lockwood and Mei Si. A review of uncertainty for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 155–162, 2022.

[147] Stephanie M Lukin, Kevin Bowden, Casey Barackman, and Marilyn A Walker. Personabank: A corpus of personal narratives and their story intention graphs. *arXiv preprint arXiv:1708.09082*, 2017.

[148] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[149] Alhassan Mabrouk, Rebeca P. Díaz Redondo, and Mohammed Kayed. Deep learning-based sentiment classification: A comparative survey. *IEEE Access*, 8: 85616–85638, 2020. .

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

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

[152] Marcel Marti, Jodok Vieli, Wojciech Witoń, Rushit Sanghrajka, Daniel Inversini, Diana Wotruba, Isabel Simo, Sasha Schriber, Mubbasir Kapadia, and Markus Gross. Cardinal: Computer assisted authoring of movie scripts. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, page 509–519, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450349451. . URL https://doi.org/10.1145/3172944.3172972.

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

[154] Mohsen Mesgar and Michael Strube. 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, October-November 2018. Association for Computational Linguistics. . URL https://aclanthology.org/D18-1464.

[155] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[156] David Millard, Charlie West-Taylor, Yvonne Howard, and Heather Packer. The ideal readerbot:: Machine readers and narrative analytics. In *NHT?18, July 2018, Baltimore, USA*. ACM, July 2018. URL https://eprints.soton.ac.uk/422387/.

[157] David E Millard. Strange patterns: Structure and post-structure in interactive digital narratives. *The Authoring Problem: Challenges in Supporting Authoring for Interactive Digital Narratives*, pages 147–169, 2023.

[158] Matthew K. Miller and Regan L. Mandryk. 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, 2016. Association for Computing Machinery. ISBN 9781450342483. . URL https://doi.org/10.1145/2992154.2992185.

[159] Alex Mitchell and Kevin McGee. Supporting rereadability through narrative play. In Mei Si, David Thue, Elisabeth André, James C. Lester, Theresa Jean Tanenbaum, and Veronica Zammitto, editors, *Interactive Storytelling*, pages 67–78, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-25289-1.

[160] Giulio Mori, David Thue, and Stephan Schiffel. Em-glue: a platform for decoupling experience managers and environments. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 266–274, 2022.

[161] Christopher Moser and Xiaowen Fang. Narrative control and player experience in role playing games: Decision points and branching narrative feedback. In Masaaki Kurosu, editor, *Human-Computer Interaction. Applications and Services*, pages 622–633, Cham, 2014. Springer International Publishing. ISBN 978-3-319-07227-2.

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

[163] Rajdeep Mukherjee, Hari Chandana Peruri, Uppada Vishnu, Pawan Goyal, Sourangshu Bhattacharya, and Niloy Ganguly. 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, 2020. Association for Computing Machinery. ISBN 9781450380164. . URL https://doi.org/10.1145/3397271.3401269.

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

[165] John T Murray. *Telltale hearts: encoding cinematic choice-based adventure games*. University of California, Santa Cruz, 2018.

[166] John T Murray and Anastasia Salter. Mapping the unmappable: Reimagining visual representations of interactive narrative. In *The Authoring Problem: Challenges in Supporting Authoring for Interactive Digital Narratives*, pages 171–190. Springer, 2023.

[167] Ben Naismith, Phoebe Mulcaire, and Jill Burstein. Automated evaluation of written discourse coherence using GPT-4. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada, July 2023. Association for Computational Linguistics. . URL https://aclanthology.org/2023.bea-1.32.

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

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

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

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

[172] Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 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, November 2019. Association for Computational Linguistics. . URL https://aclanthology.org/D19-1180.

[173] Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. Screenplay summarization using latent narrative structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1920–1933, Online, July 2020. Association for Computational Linguistics. . URL https://aclanthology.org/2020.acl-main.174.

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

[175] Bronwin Patrickson. Multi-user interactive drama: The macro view - three structural layers. In Mei Si, David Thue, Elisabeth André, James C. Lester, Theresa Jean Tanenbaum, and Veronica Zammitto, editors, *Interactive Storytelling*, pages 317–321, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-25289-1.

[176] Bronwin Patrickson. Multi-user interactive drama: A micro user drama in process. In Mei Si, David Thue, Elisabeth André, James C. Lester, Theresa Jean Tanenbaum, and Veronica Zammitto, editors, *Interactive Storytelling*, pages 199–206, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-25289-1.

[177] Marieke MM Peeters, Jurriaan van Diggelen, Karel Van Den Bosch, Adelbert Bronkhorst, Mark A Neerincx, Jan Maarten Schraagen, and Stephan Raaijmakers. Hybrid collective intelligence in a human–ai society. *AI & society*, 36:217–238, 2021.

[178] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. URL http://openai-assets.s3.amazonaws.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

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

[180] Elahe Rahimtoroghi, Jiaqi Wu, Ruimin Wang, Pranav Anand, and Marilyn Walker. 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, August 2017. Association for Computational Linguistics. . URL https://aclanthology.org/W17-5543.

[181] Revanth Rameshkumar and Peter Bailey. 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, July 2020. Association for Computational Linguistics. . URL https://aclanthology.org/2020.acl-main.459.

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

[183] Alison Reboud, Ismail Harrando, Pasquale Lisena, and Raphaël Troncy. Stories of love and violence: zero-shot interesting events' classification for unsupervised tv series summarization. *Multimedia Systems*, pages 1–19, 2023.

[184] Ashwathy T. Revi, David E. Millard, and Stuart E. Middleton. A systematic analysis of user experience dimensions for interactive digital narratives. In Anne-Gwenn Bosser, David E. Millard, and Charlie Hargood, editors, *Interactive*

Storytelling, pages 58–74, Cham, 2020. Springer International Publishing. ISBN 978-3-030-62516-0.

[185] Ashwathy T. Revi, Stuart E. Middleton, and David E. Millard. 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, October 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.creativesumm-1.1.

[186] Ashwathy T Revi, Stuart E Middleton, and David E Millard. Rationale-based learning using self-supervised narrative events for text summarisation of interactive digital narratives. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13557–13585, 2024.

[187] David L Roberts and Charles L Isbell. 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)*, 2007.

[188] Melissa Roemmele and Andrew Gordon. An encoder-decoder approach to predicting causal relations in stories. In *Proceedings of the First Workshop on Storytelling*, pages 50–59, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. . URL https://aclanthology.org/W18-1506.

[189] Christian Roth. The 'angstfabriek' experience: Factoring fear into transformative interactive narrative design. In Rogelio E. Cardona-Rivera, Anne Sullivan, and R. Michael Young, editors, *Interactive Storytelling*, pages 101–114, Cham, 2019. Springer International Publishing. ISBN 978-3-030-33894-7.

[190] Christian Roth and Hartmut Koenitz. 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, 2016. Association for Computing Machinery. ISBN 9781450345217. . URL https://doi.org/10.1145/2983298.2983302.

[191] Christian Roth, Christoph Klimmt, Ivar E. Vermeulen, and Peter Vorderer. The experience of interactive storytelling: Comparing "fahrenheit" with "façade". In Junia Coutinho Anacleto, Sidney Fels, Nicholas Graham, Bill Kapralos, Magy Saif El-Nasr, and Kevin Stanley, editors, *Entertainment Computing – ICEC 2011*, pages 13–21, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[192] Qian Ruan, Malte Ostendorff, and Georg Rehm. HiStruct+: Improving extractive text summarization with hierarchical structure information. In

Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1292–1308, Dublin, Ireland, May 2022. Association for Computational Linguistics. . URL https://aclanthology.org/2022.findings-acl.102.

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

[194] Rushit Sanghrajka and R. Michael Young. Evaluating reader comprehension of plan-based stories containing failed actions. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 18(1):179–188, Oct. 2022. . URL https://ojs.aaai.org/index.php/AIIDE/article/view/21962.

[195] Rushit Sanghrajka, Daniel Hidalgo, Patrick Chen, and Mubbasir Kapadia. Lisa: Lexically intelligent story assistant. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 13(1):221–227, Jun. 2021. URL https://ojs.aaai.org/index.php/AIIDE/article/view/12956.

[196] Rushit Sanghrajka, R. Michael Young, and Brandon Thorne. Headspace: Incorporating action failure and character beliefs into narrative planning. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 18(1):171–178, Oct. 2022. . URL https://ojs.aaai.org/index.php/AIIDE/article/view/21961.

[197] Anurag Sarkar and Seth Cooper. tile2tile: learning game filters for platformer style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 53–60, 2022.

[198] Henrik Schoenau-Fog. Hooked! – evaluating engagement as continuation desire in interactive narratives. In Mei Si, David Thue, Elisabeth André, James C. Lester, Theresa Jean Tanenbaum, and Veronica Zammitto, editors, *Interactive Storytelling*, pages 219–230, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-25289-1.

[199] Henrik Schoenau-Fog, Luis Emilio Bruni, Faysal Fuad Khalil, and Jawid Faizi. First person victim: Developing a 3d interactive dramatic experience. In Ruth Aylett, Mei Yii Lim, Sandy Louchart, Paolo Petta, and Mark Riedl, editors, *Interactive Storytelling*, pages 240–243, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-16638-9.

[200] Henrik Schoenau-Fog, Luis Emilio Bruni, Faysal Fuad Khalil, and Jawid Faizi. Authoring for engagement in plot-based interactive dramatic experiences for learning. In Zhigeng Pan, Adrian David Cheok, Wolfgang Müller, Ido Iurgel, Paolo Petta, and Bodo Urban, editors, *Transactions on Edutainment X*, pages 1–19, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-37919-2.

[201] Magy Seif El-Nasr, David Milam, and Tony Maygoli. Experiencing interactive narrative: A qualitative analysis of façade. *Entertainment Computing*, 4(1):39–52, 2013. ISSN 1875-9521. . URL https://www.sciencedirect.com/science/article/pii/S187595211200016X.

[202] David Servan-Schreiber, Axel Cleeremans, and James McClelland. Learning sequential structure in simple recurrent networks. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1988. URL https://proceedings.neurips.cc/paper_files/paper/1988/file/9dcb88e0137649590b755372b040afad-Paper.pdf.

[203] Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 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, November 2020. Association for Computational Linguistics. . URL https://aclanthology.org/2020.emnlp-main.425.

[204] Yotam Shibolet, Noam Knoller, and Hartmut Koenitz. A framework for classifying and describing authoring tools for interactive digital narrative. In Rebecca Rouse, Hartmut Koenitz, and Mads Haahr, editors, *Interactive Storytelling*, pages 523–533, Cham, 2018. Springer International Publishing. ISBN 978-3-030-04028-4.

[205] Samuel Shields, Ross Mawhorter, Edward Melcer, and Michael Mateas. Searching for balanced 2d brawler games: successes and failures of automated evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 189–198, 2022.

[206] Victor Socas-Guerra and Carina S. González-González. User attention in nonlinear narratives: A case of study. In Francisco Cipolla-Ficarra, Kim Veltman, Miguel Cipolla-Ficarra, and Andreas Kratky, editors, *Communicability, Computer Graphics and Innovative Design for Interactive Systems*, pages 104–111, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33760-4.

[207] Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 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, August 2014. Dublin City University and Association for Computational Linguistics. URL https://aclanthology.org/C14-1090.

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

Association for Computational Linguistics. . URL https://aclanthology.org/2020.nuse-1.12.

[209] Ulrike Spierling and Nicolas Szilas. Authoring issues beyond tools. In Ido A. Iurgel, Nelson Zagalo, and Paolo Petta, editors, *Interactive Storytelling*, pages 50–61, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-10643-9.

[210] Joe Stacey, Yonatan Belinkov, and Marek Rei. 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, 2022.

[211] Ingibergur Stefnisson and David Thue. Mimisbrunnur: Ai-assisted authoring for interactive storytelling. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 14(1):236–242, Sep. 2018. URL https://ojs.aaai.org/index.php/AIIDE/article/view/13046.

[212] Alexander D Stoneman, Josh Aaron Miller, and Seth Cooper. Effects of player-level matchmaking methods in a live citizen science game. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 199–206, 2022.

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

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

[215] Tomi "bgt" Suovuo, Natasha Skult, Tapani N. Joelsson, Petter Skult, Werner Ravyse, and Jouni Smed. *The Game Experience Model (GEM)*, pages 183–205. Springer International Publishing, Cham, 2020. ISBN 978-3-030-37643-7. . URL https://doi.org/10.1007/978-3-030-37643-7_8.

[216] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

[217] Neil Suttie, Sandy Louchart, Ruth Aylett, and Theodore Lim. Theoretical considerations towards authoring emergent narrative. In Hartmut Koenitz, Tonguc Ibrahim Sezen, Gabriele Ferri, Mads Haahr, Digdem Sezen, and Güven

Çatak, editors, *Interactive Storytelling*, pages 205–216, Cham, 2013. Springer International Publishing. ISBN 978-3-319-02756-2.

[218] Ivo Swartjes and Mariët Theune. Iterative authoring using story generation feedback: Debugging or co-creation? In Ido A. Iurgel, Nelson Zagalo, and Paolo Petta, editors, *Interactive Storytelling*, pages 62–73, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-10643-9.

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

[220] Nicolas Szilas and Ioana Ilea. Objective metrics for interactive narrative. In Alex Mitchell, Clara Fernández-Vara, and David Thue, editors, *Interactive Storytelling*, pages 91–102, Cham, 2014. Springer International Publishing. ISBN 978-3-319-12337-0.

[221] Nicolas Szilas and Ioana Ilea. Objective metrics for interactive narrative. In *Interactive Storytelling: 7th International Conference on Interactive Digital Storytelling, ICIDS 2014, Singapore, Singapore, November 3-6, 2014, Proceedings 7*, pages 91–102. Springer, 2014.

[222] Peggy Tang, Kun Hu, Rui Yan, Lei Zhang, Junbin Gao, and Zhiyong Wang. OTExtSum: Extractive Text Summarisation with Optimal Transport. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1128–1141, Seattle, United States, July 2022. Association for Computational Linguistics. . URL https://aclanthology.org/2022.findings-naacl.85.

[223] Peggy Tang, Junbin Gao, Lei Zhang, and Zhiyong Wang. Efficient and interpretable compressive text summarisation with unsupervised dual-agent reinforcement learning. In Nafise Sadat Moosavi, Iryna Gurevych, Yufang Hou, Gyuwan Kim, Young Jin Kim, Tal Schuster, and Ameeta Agrawal, editors, *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 227–238, Toronto, Canada (Hybrid), July 2023. Association for Computational Linguistics. . URL https://aclanthology.org/2023.sustainlp-1.17.

[224] MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. URL www.mosaicml.com/blog/mpt-7b.

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

[226] Mariët Theune, Jeroen Linssen, and Thijs Alofs. Acting, playing, or talking about the story: An annotation scheme for communication during interactive digital storytelling. In Hartmut Koenitz, Tonguc Ibrahim Sezen, Gabriele Ferri, Mads Haahr, Digdem Sezen, and Güven Çatak, editors, *Interactive Storytelling*, pages 132–143, Cham, 2013. Springer International Publishing. ISBN 978-3-319-02756-2.

[227] Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A Fox. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*, 2020.

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

[229] Milo N.R. Utsch, Gisele L. Pappa, Luiz Chaimowicz, and Raquel O. Prates. A new non-deterministic drama manager for adaptive interactive storytelling. *Entertainment Computing*, 34:100364, 2020. ISSN 1875-9521. . URL https://www.sciencedirect.com/science/article/pii/S1875952119300849.

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

[231] Josep Valls-Vargas, Jichen Zhu, and Santiago Ontañón. 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, 2017. Association for Computing Machinery. ISBN 9781450353199. . URL https://doi.org/10.1145/3102071.3106362.

[232] Josep Valls-Vargas, Jichen Zhu, and Santiago Ontañón. Toward automatic role identification in unannotated folk tales. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 10(1):188–194, Jun. 2021. URL https://ojs.aaai.org/index.php/AIIDE/article/view/12732.

[233] Renske van Enschot, Iris Boogaard, Hartmut Koenitz, and Christian Roth. The potential of interactive digital narratives. agency and multiple perspectives in last hijack interactive. In Rogelio E. Cardona-Rivera, Anne Sullivan, and R. Michael Young, editors, *Interactive Storytelling*, pages 158–169, Cham, 2019. Springer International Publishing. ISBN 978-3-030-33894-7.

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

[235] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need.

In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

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

[237] Attada Venkataramana, K Srividya, and R Cristin. Abstractive text summarization using bart. In *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, pages 1–6. IEEE, 2022.

[238] Ivar E. Vermeulen, Christian Roth, Peter Vorderer, and Christoph Klimmt. Measuring user responses to interactive stories: Towards a standardized assessment tool. In Ruth Aylett, Mei Yii Lim, Sandy Louchart, Paolo Petta, and Mark Riedl, editors, *Interactive Storytelling*, pages 38–43, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-16638-9.

[239] Anton Vinogradov and Brent Harrison. Using multi-armed bandits to dynamically update player models in an experience managed environment. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 207–214, 2022.

[240] Claudia Volpetti, K. Vani, and Alessandro Antonucci. 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, 2020. Association for Computing Machinery. ISBN 9781450376426. . URL https://doi.org/10.1145/3383972.3383988.

[241] Mirjam Vosmeer and Ben Schouten. Interactive cinema: Engagement and interaction. In Alex Mitchell, Clara Fernández-Vara, and David Thue, editors, *Interactive Storytelling*, pages 140–147, Cham, 2014. Springer International Publishing. ISBN 978-3-319-12337-0.

[242] Tyrone Vriesede and Frank Nack. Storystream: Unrestricted mobile exploration of city neighbourhoods enriched by the oral presentation of user-generated stories. In Mei Si, David Thue, Elisabeth André, James C. Lester, Theresa Jean Tanenbaum, and Veronica Zammitto, editors, *Interactive Storytelling*, pages 231–242, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-25289-1.

[243] Peter Mawhorter Michael Mateas Noah Wardrip and Fruin Arnav Jhala. Towards a theory of choice poetics. In *In Proceedings of the 9th International Conference on the Foundations of Digital Games*, 2014.

[244] David Wilmot and Frank Keller. 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, July 2020. Association for Computational Linguistics. . URL https://aclanthology.org/2020.acl-main.161.

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

[246] Zongda Wu, Li Lei, Guiling Li, Hui Huang, Chengren Zheng, Enhong Chen, and Guandong Xu. A topic modeling based approach to novel document automatic summarization. *Expert Systems with Applications*, 84:12–23, 2017. ISSN 0957-4174. . URL https://www.sciencedirect.com/science/article/pii/S0957417417303020.

[247] Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online, July 2020. Association for Computational Linguistics. . URL https://aclanthology.org/2020.acl-main.451.

[248] Divakar Yadav, Jalpa Desai, and Arun Kumar Yadav. Automatic text summarization methods: A comprehensive review, 2022.

[249] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. . URL https://doi.org/10.1145/3313831.3376301.

[250] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 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, June 2016. Association for Computational Linguistics. . URL https://aclanthology.org/N16-1174.

[251] Ruifeng Yuan, Shichao Sun, Zili Wang, Ziqiang Cao, and Wenjie Li. Separating context and pattern: Learning disentangled sentence representations for low-resource extractive summarization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7575–7586, Toronto, Canada, July 2023. Association for Computational Linguistics. . URL https://aclanthology.org/2023.findings-acl.479.

[252] Nelson Zagalo, Sandy Louchart, and Maria T. Soto-Sanfiel. Users and evaluation of interactive storytelling. In Ruth Aylett, Mei Yii Lim, Sandy Louchart, Paolo Petta, and Mark Riedl, editors, *Interactive Storytelling*, pages 287–288, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-16638-9.

[253] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193*, 2023.

[254] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193*, 2023.

[255] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR, 2020.

[256] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.

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

[258] Xingxing Zhang, Furu Wei, and Ming Zhou. 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, July 2019. Association for Computational Linguistics. . URL https://aclanthology.org/P19-1499.

[259] Chao Zhao, Faeze Brahman, Kaiqiang Song, Wenlin Yao, Dian Yu, and Snigdha Chaturvedi. NarraSum: A large-scale dataset for abstractive narrative summarization. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 182–197, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. . URL https://aclanthology.org/2022.findings-emnlp.14.

[260] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online, July 2020. Association for Computational Linguistics. . URL https://aclanthology.org/2020.acl-main.552.

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

[262] Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 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, July 2018. Association for Computational Linguistics. . URL https://aclanthology.org/P18-1061.

[263] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G. Michael Youngblood. Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8, 2018. .

[264] Suyang Zhu, Shoushan Li, and Guodong Zhou. 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, July 2019. Association for Computational Linguistics. . URL https://aclanthology.org/P19-1045.