Branching Narratives: Character Decision Points Detection

Alexey Tikhonov
Inworld.AI
Berlin, Germany
altsoph@gmail.com

Abstract

This paper presents the Character Decision Points Detection (CHADPOD) task, a task of identification of points within narratives where characters make decisions that may significantly influence the story's direction. We propose a novel dataset based on Choose Your Own Adventure (a registered trademark of Chooseco LLC) games graphs to be used as a benchmark for such a task. We provide a comparative analysis of different models' performance on this task, including a couple of LLMs and several MLMs as baselines, achieving up to 89% accuracy. This underscores the complexity of narrative analysis, showing the challenges associated with understanding character-driven story dynamics. Additionally, we show how such a model can be applied to the existing text to produce linear segments divided by potential branching points, demonstrating the practical application of our findings in narrative analysis.

Keywords: NLP, narrative analysis, CYOA, agency

1. Introduction

Modern Large Language Models (LLMs) are state-of-the-art in a lot of Natural Language Processing (NLP) tasks. However, areas related to the analysis and generation of texts with complex and rich semantic structures remain underexplored. This includes the tasks of analyzing and generating long, engaging, and rich narratives (van Stegeren and Theune, 2019). While modern models can sometimes produce innovative plot twists, they generally create less imaginative scenarios and rhetoric compared to human-authored texts (Begus, 2023).

The traditional machine learning approach to this problem starts from the data collection with necessary annotations. In the narrative analysis field, there are a number of datasets available, such as WikiPlots[1] with 112,936 story plots extracted from English Wikipedia, the MPST dataset with 14K movie plot synopses (Kar et al., 2018), and the DYPLODOC dataset, which includes synopses of 13K TV shows, 21K seasons, and over 300K episodes (Malysheva et al., 2021). However, these plain-text synopses offer limited assistance when the goal is to analyze high-level narrative structures.

Philosophers and linguists make a lot of attempts to conceptualize and formalize concepts of plot, narrative arcs, character development, conflict, and so on (Shklovsky, 1925). One of the fundamental principles in drama and narrative construction is the concept of character agency, which posits that a character's decisions and actions drive the plot forward. Aristotle, in his work *Poetics*, highlights sudden plot twists, or *peripeteia*, especially those tied to *anagnorisis*—the moment when a character comes to a significant realization or discovery that affects subsequent choices. Borges, in his short story *The Garden of Forking Paths*, explores the idea of multiple possible worlds through the metaphor of a labyrinth, representing an infinite number of potential narratives and outcomes based on characters' actions: "*your ancestor <...> believed in an infinite series of times, in a growing, dizzying net of divergent, convergent and parallel times. This web of time—the strands of which approach one another, bifurcate, intersect or ignore each other through the centuries—embrace every possibility*." Propp introduces the concept of functions—recurring, typical actions that move the narrative forward: "*a tale often attributes identical actions to various personages; this makes possible the study of the tale according to the functions of its dramatis personae <...> a function <...> cannot be defined apart from its place in the course of narration*". Gustav Freytag, in (Freytag and MacEwan, 1968), describes *Freytag's Pyramid*, a typical plot structure identifying five pivotal plot points: Opportunity, Change of Plans, Point of No Return, Major Setback, Climax. Aarseth in his *Cybertext* book proposes the term *ergodic literature* to define open, dynamic texts, with which the reader must perform specific actions to generate a literary sequence.

One may argue that we still cannot clearly define what we aim to analyze, and this slows progress in the analysis and generation of narrative structures (Yamshchikov and Tikhonov, 2023). However, the NLP community continues to seek improvements in narrative processing (Fan et al., 2019), by setting subtasks for the formal identification of important plot elements. For instance, in (Tikhonov and Yamshchikov, 2022) the task is to identify "Chekhov's guns"—narrative objects that significantly impact the plot's development; (Papalampidi et al., 2019) introduce a Turning Point Identifica-

[1] https://github.com/markriedl/WikiPlots

Games and NLP @LREC-COLING-2024, pages 70–75
20-25 May, 2024.

tion task—to directly identify Freytag's points in the text, and (Li et al., 2023) proposes a task to extract action models from narrative texts automatically.

In this paper, we propose using characters' decision-making moments to analyze and formalize narrative structure. We introduce a new NLP task—Character Decision Points Detection (CHADPOD). This task focuses on identifying moments in the narrative where characters make decisions that significantly determine the plot's direction. We believe that highlighting such moments will improve our understanding of traditional text plots and open possibilities for working with nonlinear and interactive narrative structures (Juul, 2005; Murray, 2006).

This work contributes by:

1. Proposing a formalization of the Character Decision Points Detection (CHADPOD) task.
2. Introducing a Character Decision Points dataset.
3. Demonstrating the effectiveness of modern models in identifying Character Decision Points (CDPs).
4. Offering an interpretation of CDPs and their relation to the related task of turning points.

2. CHADPOD task

In NLP research, analysts and creators frequently utilize Gamebook genre games, also widely known as Choose Your Own Adventure[2] (CYOA) books, named after one of the earliest popular series in this genre. These sources are crucial for studying nonlinear narratives, alongside interactive fiction games.

For instance, the data from such sources has been used to train systems that generate suggestions for people writing short stories (Clark and Smith, 2021). Another study employs CYOA as a medium for training generative agents to enforce temporal constraints (Rothkopf et al., 2024). In the MACHIAVELLI paper (Pan et al., 2023), authors use a collection of CYOA games to create a game environment for training text agents. Some researchers[3] explore them to analyze a variety of narrative macro-structures.

In this work, we introduce the CHADPOD task, which focuses on identifying narrative points where a character makes a choice that determines the further course of the story. We utilize CYOA game graphs to create a new CHADPOD benchmark, consisting of 1,462 binary classification tasks, with 731 tasks in each class. Each task comprises two text segments—a prefix and a postfix. The positive class includes narrative points where a character makes a choice that significantly influences the story's direction. The negative class consists of randomly segmented texts (we take a continuous text from a single node and split it at some random point between sentences), as well as text points where a character takes some action, but it does not significantly affect the story's progression.

3. Data

In this section, we describe the process of constructing the CHADPOD dataset.

We use the MACHIAVELLI dataset (Pan et al., 2023), which consists of 134 Choose-Your-Own-Adventure games, as our input data.

For each available game, we analyze its graph and extract triplets of the form:

⟨*node1*; *action*; *node2*⟩

where *node1* is the text before the action, *action* is the choice made by the player, and *node2* is the text following the action (see Figure 1).

Next, we filter the triplets—removing exact duplicates, retaining only those with descriptions sufficiently long to provide enough context – both in *node1* and *node2* (to do so we used simple heuristics – at least 4 sentences, at least 50 characters), removing texts that are dialogue segments (dialogues are a very specific type of narrative that should be analyzed separately, see for example (Zhou et al., 2023)), and removing texts with unusual characters. As positive examples of branching points, we only select triplets for which the graph from node1 has more than one possible action, thus excluding scenarios like ⟨*node1*, "1 year later…", *node2*⟩. The remaining 731 examples make up the positive class.

Then we form the negative class from two components—using the division of texts from the same games (nodes) at random points as easy negatives, and the above-described cases when there is exactly one action emanating from node1 in the graph as hard negatives.

Finally, we divide the data[4] into 3 game-wise splits, ensuring that there are no overlaps between the splits in terms of games, thereby eliminating test set leakage. The statistics of the resulting split are presented in Table 1.

4. Task validation

Experiments To validate the usefulness of our dataset, we trained several models for the CHADPOD tasks. We used the DeBERTa model (He

[2]It is a registered trademark of Chooseco LLC.

[3]https://heterogenoustasks.wordpress.com/2015/01/26/standard-patterns-in-choice-based-games/

[4]The data is available through Google Drive. The password is CHADPOD.

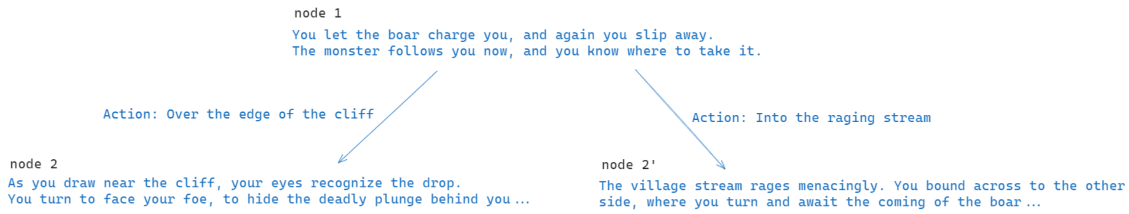

Figure 1: Example of branching in CYOA data, shortened for the simplicity.

Table 1: Data Splits

Class	**Train**	**Dev**	**Test**
Positives	511	110	110
Negatives	256	55	55
Hard Negatives	255	55	55
Total	1022	220	220

et al., 2021) as a strong baseline, known for its state-of-the-art performance in many text classification tasks [5]. Additionally, we included older but widely used models such as BERT (Devlin et al., 2019) and ALBERT (Lan et al., 2020) as weaker baselines. We chose accuracy as the metric due to our data being class-balanced. The training was conducted on a GPU RTX 4090 on the Vast.ai platform, with each model trained with a batch size of 4 and a learning rate of 5.5×10^{-6} until accuracy on a validation split began to decline. A full training run for one task required at most 30 minutes. We also added results for GPT-3.5-turbo[6] and GPT-4-turbo[7] tested in a zero-shot manner with hyperparamteres (temperature, probability threshold) obtained by a grid search on the validation set.

The results are presented in the Table 2.

Table 2: Test Accuracy of Models on CHADPOD

model	**test acc**	**size**
DeBERTa-v3-large	**89%**	340M
DeBERTa-v3-base	85%	110M
ALBERT-v2-base	84%	11M
BERT-base	79%	110M
GPT-4-turbo, 0-shot	62%	unknown
GPT-3.5-turbo, 0-shot	55%	unknown

As seen, the task is solvable but remains quite complex for simpler and smaller models. The presented results are on a test dataset without overlap with the training set in terms of games, minimizing the risk of overfitting. As for LLMs, it seems that using them in a 0-shot manner is not a silver bullet for this task, though results could likely be improved through fine-tuning or prompt engineering. Analysis of the confusion matrix revealed that LLMs underperform on the positive class, leading to a high number of false negatives.

Comparison with Turning Points One may notice that the CHADPOD task is significantly similar to the Turning Points Identification task proposed in (Papalampidi et al., 2019). In this section, we conduct a comparative analysis, demonstrating that despite similar formulations, the tasks differ fundamentally.

Recall that in (Papalampidi et al., 2019), the TRIPOD dataset consists of manually annotated short plot synopses (averaging 35 sentences) of 99 screenplays with sentence-level turning points annotations, where turning points are defined as the 5 classic pivot moments formulated in Freytag's Pyramid.

We transformed the TRIPOD dataset to our format, taking contexts around the indicated turning points as positive examples and random divisions of the same synopses where there were no turning points as negative examples. The final dataset used all available non-overlapping contexts with at least 3 sentences before and after the split point, resulting in 255 positive and 209 negative examples.

Applying our DeBERTa-v3-large based model to these examples yielded the metrics provided in Table 3.

Table 3: Performance on Adapted TRIPOD Dataset

Metric	**Value**
Accuracy	40%
Balanced Accuracy	41%
F1-Score	41%

These results indicate that the semantics of the tasks significantly differ (recall that the model's accuracy on an isolated test set was 89%).

One might suggest that the main difference between these tasks lies in the scale (turning points

[5] https://huggingface.co/altsoph/chadpod

[6] https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates

[7] https://openai.com/blog/new-models-and-developer-products-announced-at-devday

are just 5 key moments in the plot's macrostructure) and in that Freytag's turning points do not necessarily imply character agency. Contrarily, they can be exclusively formed by external events, leaving characters without a choice.

5. Text Segmentation Study

In this part, we demonstrate how the obtained binary classification model can be used for segmenting text into linear segments separated by potential branching points in the narrative.

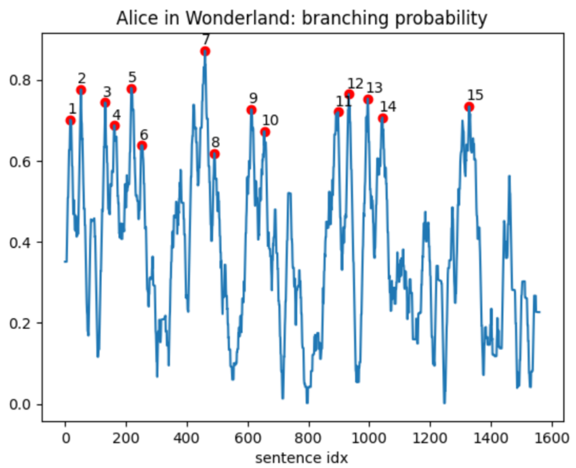

Figure 2: Most probable branching points in the text of Alice in Wonderland.

For our experiments, we utilized the text of *Alice's Adventures in Wonderland* by Lewis Carroll, as it is in the public domain. We employed a sliding window of 10 sentences with a step of 1 sentence and calculated the probability of branching at any given possible point. To reduce noise, we applied a convolution with a linear kernel of width 25, and then on the resulting sequence, we identified local maximums on segments lying above the threshold $TH1 = 0.5$, preserving only peaks above the threshold $TH2 = 0.6$ to obtain 15 main branching points. Indeed, the parameters of such a heuristic can be adjusted to change the sensitivity of the approach.

In the Figure 2 one can see 15 most probable points of branching in the given text. To analyse them and gain understanding if these points really correspond to the important decisions of the chraracter, we used GPT-4-turbo model to assess selected points and propose potential alternatives to the character's action. We refer to the Table 4 for details of these points and alternatives.

Despite the subjectivity of such analysis, it is worth saying that all identified points, except maybe for 1 and 4, correspond to moments when a character makes a decision or performs an action that influences the subsequent development of events.

6. Discussion

This study contributes to the evolving field of Natural Language Processing (NLP) by addressing the nuanced task of detecting Character Decision Points (CDPs) within narrative texts. Through the development and validation of the CHADPOD task, our findings highlight the complexity and potential of leveraging Large Language Models (LLMs) for narrative analysis, particularly in identifying moments of narrative branching that may be important to story development.

The performance of various models on the CHADPOD task, especially the high results of the DeBERTa model, demonstrates the feasibility of detecting narrative branching points with high accuracy. However, the underperformance of smaller and simpler models, as well as zero-shot tests of GPT-3.5 and GPT-4, illustrates the challenges of the task. We suggest these challenges are not solely due to model capacity but also reflect the sophisticated understanding of narrative structure.

The application of our binary classification model to text segmentation, as demonstrated in the analysis of Alice's Adventures in Wonderland, showcases the practical utility of our approach. This illustrative study can be a bit speculative without ground truth labeling, since ChatGPT is able to generate plausible alternatives for any requested point in text. However, subjectively, most of the detected branching points (demonstrated in Table 4) correspond to the turning points of text there the character makes impactful decisions. (this problem can also be approached as a direct segmentation task, as in, for example, (Koshorek et al., 2018); we leave these experiments for future work.)

Our results suggest several directions for future research. First, expanding the dataset to include a broader range of narratives, could enhance the model's understanding of diverse narrative structures. Second, exploring more granular classifications of decision points, such as presented in Syd Field's book *Screenplay* (with 6 key points) or the one based on Vogler's interpretation of Campbell's monomyth (with 12 such points) could offer finer insights into narrative dynamics. Third, using the CHADPOD data can help to construct a macro-assessment of characters' agency within a text, i.e., an assessment that enables comparing different texts in terms of how much the development of the text is determined by the characters' choices.

7. Bibliographical References

E.J. Aarseth. 1997. *Cybertext: perspectives on ergodic literature*. Johns Hopkins Univ Pr.

Aristotle. 2014. *Aristotle collection*. Annotated Classics.

Nina Begus. 2023. Experimental narratives: A comparison of human crowdsourced storytelling and ai storytelling.

Jorge Luis Borges. 1964. *Labyrinths: selected stories & other writings*, volume 186. New Directions Publishing.

Elizabeth Clark and Noah A. Smith. 2021. Choose your own adventure: Paired suggestions in collaborative writing for evaluating story generation models. In *Proceedings of NAACL 2021*, pages 3566–3575, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL 2019*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. *arXiv preprint arXiv:1902.01109*.

G. Freytag and E.J. MacEwan. 1968. *Freytag's Technique of the Drama: An Exposition of Dramatic Composition and Art*. Johnson Reprint.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Jesper Juul. 2005. Half-real: Video games between real rules and fictional worlds.

Sudipta Kar, Suraj Maharjan, A. Pastor López-Monroy, and Thamar Solorio. 2018. Mpst: A corpus of movie plot synopses with tags.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Ruiqi Li, Leyang Cui, Songtuan Lin, and Patrik Haslum. 2023. Automated action model acquisition from narrative texts.

Anastasia Malysheva, Alexey Tikhonov, and Ivan P. Yamshchikov. 2021. *DYPLODOC: Dynamic Plots for Document Classification*. IOS Press.

Janet H Murray. 2006. Toward a cultural theory of gaming: Digital games and the co-evolution of media, mind, and culture. *Popular Communication*, 4(3):185–202.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark.

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

Vladimir Propp. 1968. Morphology of the folktale, trans. *Louis Wagner, 2d. ed.*

Raven Rothkopf, Hannah Tongxin Zeng, and Mark Santolucito. 2024. Enforcing temporal constraints on generative agent behavior with reactive synthesis.

Viktor Shklovsky. 1925. Theory of prose (b. sher, trans.). *Champaign, IL: Dalkey Archive Press. Original work published*.

Alexey Tikhonov and Ivan P. Yamshchikov. 2022. Actionable entities recognition benchmark for interactive fiction.

Judith van Stegeren and Mariët Theune. 2019. Narrative Generation in the Wild: Methods from NaNoGenMo. In *Proceedings of the Second Workshop on Storytelling*, pages 65–74, Florence, Italy. Association for Computational Linguistics.

Ivan Yamshchikov and Alexey Tikhonov. 2023. What is wrong with language models that can not tell a story? In *Proceedings of the The 5th Workshop on Narrative Understanding*, pages 58–64, Toronto, Canada. Association for Computational Linguistics.

Wei Zhou, Xiangyu Peng, and Mark Riedl. 2023. Dialogue shaping: Empowering agents through npc interaction.

Table 4: Branching Points and Alternatives in *Alice's Adventures in Wonderland*

No.	Main Decision Point	Alternatives
1	Alice falls down the well	- Tries to look for something to grab onto to stop - Attempts to fly or float by flapping her arms
2	Alice uses the little golden key to open the small door	- Tries breaking one of the doors with a chair - Climbs onto the table
3	Alice cries a pool of tears and falls there	- Calls out for help - Eats something to change her size
4	Alice thinks she is Mable and continues to cry	- Searches for someone who remember her - Insists that she is not Mabel
5	Alice tells Mice about dogs and scares it	- Stops talking about pets - Apologizes to the Mouse
6	Alice decides to join the Caucus-race	- Suggests a different activity - Objects to the Caucus-race
7	Alice decides to look on top of the mushroom	- Eats a flower - Goes back to the puppy
8	Alice agrees to return to the Caterpillar	- Ignores it and walks away - Loses her temper with the Caterpillar
9	Alice knocks on the door	- Searches for another entrance - Returns to the wood
10	Alice tries to calm down the cook	- Leaves the room - Organizes a cleanup effort
11	Alice decides to leave the tea-party forever	- Stays despite rudeness - Invites the Dormouse to leave
12	Alice stays standing on the arrival of King and Queen	- Lies down like the gardeners - Starts clapping
13	Alice decides to talk with the Cat about the game	- Ignores the Cat and plays alone - Attempts to leave the croquet ground
14	Alice argues with the Duchess	- Agrees with the Duchess - Changes the subject
15	Alice follows the Gryphon to the trial	- Returns to the Mock Turtle - Stays to listen to the soup song