Rethinking Scene Segmentation. Advancing Automated Detection of Scene Changes in Literary Texts

Svenja Guhr[1,2] **Huijun Mao**[2] **Fengyi Lin**[2]
[1]*fortext lab*, Technical University of Darmstadt, Germany
[2]Literary Lab, Stanford University, USA
{sguhr, huijunm, linfy}@stanford.edu

Abstract

Automated scene segmentation is an ongoing challenge in computational literary studies (CLS) to approach literary texts by analyzing comparable units. In this paper, we present our approach to text segmentation using a classifier that identifies the position of a scene change in English-language fiction. By manually annotating novels from a 20th-century US-English romance fiction corpus, we prepared training data for fine-tuning transformer models, yielding promising preliminary results for improving automated text segmentation in CLS.

1 Introduction

Segmenting literary prose into meaningful units, such as events, plots, or scenes, opens up new possibilities for comparative analysis by focusing on smaller units rather than entire texts. However, automating this process remains a significant challenge in CLS. While many computational approaches depend on pre-segmented texts due to input size limitations, standardized methods for segmentation are still lacking. As a result, heuristic approaches, such as dividing texts into equal-sized units or relying on chapter boundaries, are often used – even though chapter divisions typically reflect editorial choices rather than coherent narrative structures, and especially popular fiction and serialized novels often play with cliff hangers that extend a key action beyond chapter boundaries (Pethe et al., 2020; Bartsch et al., 2023; Stiemer et al., 2025).

Drawing from their established use in dramatic texts and film studies, scenes have emerged as useful units for segmenting literary prose. Defined by consistency in time, place, and characters, scenes "center around a particular action" (Gius et al., 2019). This internal coherence allows them to function as self-contained, meaningful units that can be systematically compared to other scenes within a narrative or a text corpus. For instance, consider a novel in which an initial scene takes place in a supermarket where one of the characters is depicted grocery shopping. This is followed by a new scene set in a kitchen where two characters are cooking and talking. Each scene can be analyzed independently in terms of its temporal and spatial dimensions. By segmenting a text into such discrete units, we enable systematic comparative investigations of character constellations, spatial patterns, and thematic developments. For example, after identifying all the scenes that take place in a supermarket, one could compare the recurring characters in those scenes and analyze their actions in that specific space.

The automation of scene annotation was first approached by Gius et al. (2019), whose definition served as the basis for the Shared Task of Scene Segmentation (STSS) of German prose (Zehe et al., 2021b). This initiative included the development of scene detection guidelines (Gius et al., 2021) and the creation of German-language training datasets with manually annotated scenes to support automated methods. The most effective approach, developed by Kurfalı and Wirén (2021), utilized a BERT-based model with weighted cross-entropy and the IOB2 scheme, focusing on identifying scene boundaries rather than full segments.

Our goal is to make a first attempt at developing a scene recognition classifier for US-English fiction. We build on the winning team's approach in the German shared task, but use more recent language models and an approximation strategy that works by predicting scene changes that occur in six-sentence segments[1]. Since the submission of this paper in January 2025, we have learned of an independent but comparable approach developed by Zehe et al. (2025). Their work, focusing on German texts, extends the earlier scene segmentation project (Zehe et al., 2021b), which was

[1]Code is available at: https://github.com/literarylab/scene_segmentation.

Proceedings of LaTeCH-CLfL 2025, pages 79–86

Corpus "Men Made in America"	
female authors	47
romance novels	50
words in total	5,5 Mio.
manually annotated texts	10
in words	572,907
scene changes in gold annotation	795

Table 1: Corpus metadata.

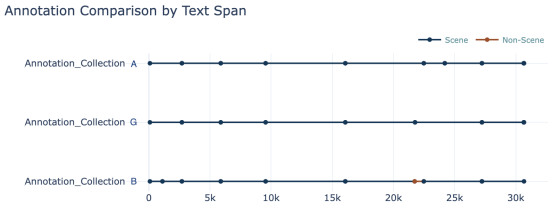

Figure 1: Comparison of two independent annotations (A+B, 0.35γ) with the gold annotation (G) in the middle. Visualization created with GitMA (Vauth et al., 2022) to demonstrate the gold annotation creation process.

paused in 2022 after the completion and evaluation of the shared task at KONVENS 2021. To evaluate their inter-annotator agreement and the performance of their automation, they introduce a new metric, namely a "relaxed F1 score" (Zehe et al., 2025, 5), which allows a tolerance of three sentences for the detected position of a scene change in the manual and automated annotations. The authors argue that fluid scene changes, which cannot be precisely positioned in the text even by human annotators, usually occur within a window of three sentences. Accordingly, the relaxed F1 score gives better scoring results that reflect the performance of the human annotators and the models (Zehe et al., 2025, 5). These findings are consistent with our observation that scene change transitions can span up to three sentences, which led to our decision to use a six-sentence segment approach for the prediction process.

2 Method

2.1 Manual Annotation

Referring to the scene annotation guidelines from Gius et al. (2021), we manually annotated 20% of a corpus of thematically cohesive romance novels from the Harlequin series "Men Made in America" (1982–2002) for scene changes (Table 1 for more information). As already recognized in Zehe et al. (2021a), genre fiction proved easier to annotate than high-brow literature due to its block-style and inherently scenic writing style. The homogeneous corpus consists of 50 novels (each 250 pages – between ca. 40,000 and 75,000 words) written solely by female authors, with each novel telling the romantic story of a couple in one of the 50 United States of America.

As a group of four experts and four trained student annotators from literary studies, we manually annotated ten novels with two annotators per novel[2]. Our inter-annotator agreement[3] (Table 2), ranging from 0.31 to 0.53 Mathet's γ[4], was lower than in Zehe et al. (2021a), who reported an agreement of 0.7 for the annotation of German novels by two trained expert annotators. However, differences in segment length preferences and inclusion or exclusion of chapter headings sparked valuable discussions and resulted in compromise gold annotations. Evaluating annotation quality highlighted the benefits of "collective intelligence" as described by Baledent et al. (2022, 2947), where annotators' errors are mutually offset – such as one favoring shorter segments and the other preferring fewer longer ones. By involving a third annotator to create gold annotations based on the independent annotations by two annotators, the results struck a balanced compromise, mitigating the effects of lower inter-annotator agreement (Figure 1).

This process highlights the interpretive nature of scene segmentation, for a task for which there is no ground truth data, especially when time, place, and character information remain vague. Instead, a negotiated consensus ensures that gold annotations represent a balanced compromise among annotators. Annotators review the entire text, identify scene change positions, highlight the relevant text, and label it as either "scene" or "non-scene." Initial comparisons revealed that scenes are more frequent than non-scenes in novels, with notable variation in the distribution and length of segments depending on the novel (mean segment length: 869.10 words; standard deviation: ±799.47 words; minimum: 69.63; maximum: 1668.57), reflecting dif-

[2]The manual annotation process utilizes the software CATMA 7.1 (Evelyn Gius et al., 2024), which facilitates collaborative annotation and comparison of annotations.

[3]The inter-annotator agreement has been calculated using the Python package GitMA by Vauth et al. (2022).

[4]Mathet's γ is further explained in Mathet et al. (2015) and Zehe et al. (2021a, 3172).

ferences in narrative style. Chapter markers were observed to sometimes signal scene changes, but not as a consistent pattern, as cliffhangers in some novels break this convention. These findings underscore the value of defining scene changes as a semantically meaningful segmentation unit in literary studies, as opposed to relying solely on chapter boundaries. Consistent with Zehe et al. (2021b, 15), scene changes were frequently triggered by temporal shifts (e.g., "two hours later"), spatial transitions, or changes in character configurations. The main consequence of calculating inter-annotator agreement, engaging in discussions, and creating gold annotations was that we decided to include embedded scenes and short non-scenes within larger annotated segments. We also decided to treat temporally parallel actions presented from different perspectives in successive narrative units, but representing the same narrative time and space, as sub-scenes combined into a larger single annotated segment. Drawing on the terminology and analytical framework of film analysis, we refer to these interconnected narrative units as "sequences" (Cutting, 2014, 70–71). In this context, the boundaries of these cohesive narrative units – each of which may consist of multiple smaller segments – were selected and prepared as training data for the automation of their detection in the text.

2.2 Automation Approaches

To automate scene segmentation, we investigated two approaches: (1) using a generative model and (2) fine-tuning a pre-trained custom model.

(1) In our first approach (in November 2024), we provided the novel text (either the entire novel at once or pre-segmented in chapters) along with the scene annotation guidelines from Gius et al. (2021) to several large language models (LLMs), including ChatGPT 4 and 4-o, Claude 3.5 (Sonnet and Haiku), Gemini Pro, and Llama 3.2. However, none of these models produced satisfactory results, as anecdotally noted in the following: For example, ChatGPT 4-o frequently misinterpreted a single conversation scene, dividing it into multiple discrete scenes, likely due to shifts in the speaking character. Additionally, some LLM approaches produced an excessive number of short scene segments, suggesting a tendency to over-annotate rather than accurately detect meaningful boundaries, possibly as a strategy to generate more results without a clear understanding of the underlying structure. Although our findings remain anecdotal due to the lack of a detailed quantitative evaluation, initial experiments showed significant issues with accurate scene boundary detection, leading us to explore alternative approaches. These observations are in line with prior research on LLM performance, which has shown that these models can exhibit signs of misclassifying or overgeneralizing based on their pre-training data (Bamman et al., 2024). Additionally, LLMs struggle with long-context sequences, getting lazy especially in complex real-world scenarios that require them to understand the entire input (Li et al., 2024). Accordingly, we suggest that current LLMs are not yet equipped to effectively process and reason over long, context-rich sequences, which is crucial for tasks like scene segmentation[5]. Given these failures, it became clear that relying on generative models for this task was not yet appropriate.

(2) Consequently, we shifted to fine-tuning a transformer-based pre-trained model for detecting scene change points within a text, which allowed us to derive the desired scene segments. To approximate scene change positions, we pre-processed the manually annotated novels by automatically splitting them into six-sentence passages (after removing typographical elements such as "***" or chapter indications to avoid bias). We chose the passage size based on the aforementioned observation that scene changes often occur gradually over a few sentences, and that annotators' decisions about scene boundaries typically vary by about ±3 sentences, making six sentences a reasonable segment length for the approximation task. Automatically extracted from the manual annotations using regular expressions, the passages are binary labeled as containing a scene change (1) or not (0). We have fine-tuned two transformer-based models to this binary classification task: BERT (Devlin et al. (2019), model version from 2023) and the Universal Sentence Encoder (USE by Cer et al. (2018), model version from 2023 with total parameters: 470,928,387 (1.75 GB) and trainable parameters: 1,538 (6.01 KB)). Although BERT is widely used in most NLP tasks, we found USE to achieve better performance in our specific case. BERT is de-

[5] In a brief trial with DeepSeek in February 2025 (DeepThink R1 (DeepSeek-AI, 2025)), we found that the model detected fewer scene changes than human annotators, but the locations of scene changes in a short test set all overlapped with human annotations. However, we are currently waiting for secure local API access to the LLMs to perform a qualitative experiment on our copyrighted data.

Author (Date)	Title	Cohen's k	Mathet's γ	Scene changes in gold annot.	words
Ferrarella (2000)	*Found: His Perfect Wife* AK	0.49	–	59	65,421
Broadrick (1986)	*Deceptions* CA	0.3	–	69	42,605
Stuart (1984)	*Tangled Lies* HI	0.3	0.31	88	69,124
Palmer (1985)	*Love By Proxy* IL	0.3	–	76	42,063
Campbell (1987)	*Pros and Cons* MA	0.38	–	91	75,527
Webb (2000)	*Warrior's Embrace* MS	0.21	0.34	118	58,903
McKenna (1984)	*Too Near the Fire* OH	0.78	0.41	39	43,319
Leonard (2000)	*Cowboy Be Mine* TX	0.52	0.53	65	62,421
Neggers (1989)	*Finders Keepers* VT	0.51	–	90	52,869
Cassidy (1997)	*Midnight Wishes* WY	0.39	–	100	60,655
10 Romance novels	from "Men Made in America"	ø0.4	ø0.4	795	572,907
20 translated novels	reuse from Zehe et al. (2025)	–	ø0.7	1,250	597,659
30 novels	total training set	–	–	2,045	1,170,566

Table 2: Inter-annotator agreement between two expert human annotators of manual annotations. A visual comparison of the agreement and its relation to the IAA scores can be found in Figure 1 demonstrating an agreement of 0.35γ.

signed to capture the bidirectional context of words within a sentence, making it particularly effective for token-level tasks such as question answering and named entity recognition. In contrast, USE generates fixed-size vector embeddings that represent entire sentences, making it well-suited for semantic similarity and sentence-level tasks. Given that scene detection typically involves analyzing larger segments of text rather than individual words, we hypothesize that USE's sentence-level embeddings provide a more effective representation for this task.When comparing the fine-tuned BERT and USE models in an initial model selection trial, we observed an increase in F1 score of approximately 0.2 for both the balanced training and validation test sets (Table 3), supporting the decision to focus on USE.

For the final training of the model, we combined the ten manually annotated texts from the romance novel corpus (see Table 1) with an automatically generated translation of 20 novels from the training corpus of the shared task described in Zehe et al. (2021b) and Zehe et al. (2025). Furthermore, we upsampled the scene change annotations to provide an equal distribution of the classes and avoid model bias (using random oversampling). Accordingly, for the automation task, the majority baseline dropped from 0.87 to 0.5 in the internal test set.

3 Evaluation and Error Analysis

For the evaluation, we compiled a test set using the final five manually annotated scenes from each of the ten romance novels in the original corpus. These last five scenes were previously excluded from the training set, resulting in a total of 50 scenes. Like the training data, they were segmented into six-sentence segments (0.8 majority baseline). This approach ensured that the test set remained sufficiently similar to the data of interest, namely our US romance novel corpus, while still providing enough variation to assess the model's generalization ability.

The evaluation on the unseen test set reveals that the model is more prone to overlooking a scene change than to mistakenly identifying one where none exists, as there are many more false negatives than false positives. Through an examination of individual examples, we identify several factors that influence the model's predictions: 1) segment length, 2) characters and pronouns, 3) ambiguity in manual annotations. First, we find that the model is more likely to make errors when processing longer inputs. Specifically, by calculating the average segment length, we observed that the biggest difference was between correct cases and false positives, indicating that the model is more likely to detect a scene change in longer segments. Second, we identify character names as a key factor influencing the model's predictions, particularly in cases where errors occur. We recognize that false positive and false negative cases are governed by different aspects of character mentions. In false positive cases, the model misinterprets a continuous scene as a scene change due to the introduction of new characters, which incorrectly signals a break. Conversely, in false negative cases, actual scene breaks are mistaken for continuity because the model recognizes recurring names or pronouns across scenes, leading to incorrect predictions. Finally, we also identify a third group of errors where the reasoning

	(first trial)		(final training)	
Model Performance	**BERT**	**USE**	**USE**	**test set**
Accuracy				0.83
Training	0.92	0.94	0.81	
Validation	0.92	0.95	0.81	
F1				0.5
Training	0.48	0.69	0.66	
Validation	0.48	0.71	0.65	
Precision				0.59
Training	0.47	0.74	0.72	
Validation	0.46	0.76	0.72	
Recall				0.44
Training	0.50	0.65	0.62	
Validation	0.50	0.67	0.61	
Loss				–
Training	0.31	0.16	0.43	
Validation	0.31	0.17	0.42	

Table 3: First trial: Performance comparison of two Transformer models (best epoch) indicating the validation results during the initial training process leading to the decision to use USE as the main model: BERT en_uncased and Universal Sentence Encoder (USE) fine-tuned on four manually annotated training texts (before upsampling). Final training: Performance of the best epoch of the USE model fine-tuned on 20 manually annotated training texts (after upsampling). The last column contains the evaluation results on the independent test set.

behind the human annotator's decision to mark a scene change is unclear, making it difficult to determine the correct interpretation. This is of particular interest given the low agreement among human annotators in manual scene change annotation, suggesting the absence of ground truth for this task for US-English texts.

4 Conclusion and Outlook

In conclusion, the evaluation results and the error analysis[6] are promising, but the current approach only approximates scene change positions within six-sentence segments. To enhance precision, we started developing a sentence-wise prediction model that identifies the first sentence of a six-sentence segment previously predicted with a high probability of bearing a scene-change. However, the task is still far from being solved and with our contribution we want to reopen the discussion on scene segmentation, and add a new perspective to the discourse on meaningful literary text segmentation for CLS.

[6]A detailed analysis of the errors can be found in the Appendix A.

Limitations

The study has several limitations that warrant further investigation: Regarding generalizability, while the segmentation approach may be applicable to other popular fiction genres similar to those found in our annotated corpus, we do not expect it to perform well on more complex, highbrow literary texts. The structural and stylistic differences between such texts and the corpus used in this study pose a challenge for direct transferability (see also Zehe et al. (2021b)).

Another limitation is that our study focuses only on segment boundary detection, without distinguishing between scenes and non-scenes. While this classification is part of the full task as defined by Zehe et al. (2021b), our approach does not account for their distinction, nor for the detection of nested scene structures, where scenes exist within other scenes. Addressing this aspect would require a more hierarchical segmentation approach, which remains an open direction for future research.

Additionally, due to differences in language and test sets, our results are not directly comparable to those reported by Zehe et al. (2021b). This discrepancy should be considered when interpreting our findings in relation to prior work.

Ethics Statement

Our experiments are conducted on an extended version of an existing dataset consisting exclusively of fictional texts, including romance novels, which are subject to copyright restrictions. The scene segmentation task is independent of the specific content of these texts and focuses solely on structural analysis. We do not identify any ethical concerns related to this task or its potential applications. The models presented in this study are intended purely for the analysis of fictional narratives.

Acknowledgments

Many thanks to our student annotators / further project members Mallen Clifton, Agnes Hilger, Jessica Monaco, Alexander J. Sherman, and Ellen Yang who supported the project with their close reading, manual scene annotations, and discoveries of unconventional scene structures. Furthermore, we are very grateful for the extended training data provided by the automatic translations of the manually scene-annotated training texts from the German scene segmentation project, which opens up

new possibilities for the task beyond language barriers. Finally, we thank the anonymous reviewers for their thoughtful comments.

References

Anaëlle Baledent, Yann Mathet, Antoine Widlöcher, Christophe Couronne, and Jean-Luc Manguin. 2022. Validity, Agreement, Consensuality and Annotated Data Quality. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2940–2948, Marseille, France. European Language Resources Association. Read_Status: New Read_Status_Date: 2024-12-09T19:52:13.016Z.

David Bamman, Kent K. Chang, Lucy Li, and Naitian Zhou. 2024. On Classification with Large Language Models in Cultural Analytics. In *Proceedings of the Computational Humanities Research Conference 2024*, volume 3834 of *CEUR Workshop Proceedings*, pages 494–527, Aarhus, Denmark. CEUR.

Sabine Bartsch, Evelyn Gius, Marcus Müller, Andrea Rapp, and Thomas Weitin. 2023. Sinn und Segment. Wie die digitale Analysepraxis unsere Begriffe schärft. *Zeitschrift für digitale Geisteswissenschaften*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder.

James E. Cutting. 2014. Event segmentation and seven types of narrative discontinuity in popular movies. *Acta Psychologica*, 149:69–77. TLDR: Using a sample of 24 movies, results suggest that there are at least four different signatures of narrative shifts to be found in popular movies - general patterns across time, patterns of historical change, genre- specific patterns, and film-specific patterns.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Evelyn Gius, Jan Christoph Meister, Malte Meister, Marco Petris, Dominik Gerstorfer, and Mari Akazawa. 2024. CATMA.

Evelyn Gius, Fotis Jannidis, Markus Krug, Albin Zehe, Andreas Hotho, Frank Puppe, Jonathan Krebs, Nils Reiter, Natalie Wiedmer, and Leonard Konle. 2019. Detection of Scenes in Fiction. In *Book of Abstracts*, Utrecht.

Evelyn Gius, Carla Sökefeld, Lea Dümpelmann, Lucas Kaufmann, Annekea Schreiber, Svenja Guhr, Nathalie Wiedmer, and Fotis Jannidis. 2021. Guidelines for Detection of Scenes.

Murathan Kurfalı and Mats Wirén. 2021. Breaking the Narrative: Scene Segmentation through Sequential Sentence Classification. In *Proceedings of the Shared Task on Scene Segmentation*, pages 49–53.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context LLMs Struggle with Long In-context Learning.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The Unified and Holistic Method Gamma () for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, 41(3):437–479.

Charuta Pethe, Allen Kim, and Steve Skiena. 2020. Chapter Captor: Text Segmentation in Novels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8373–8383, Online. Association for Computational Linguistics.

Haimo Stiemer, Hans Ole Hatzel, Chris Biemann, and Evelyn Gius. 2025. Pause im Text. Zur Exploration semantisch konditionierter Sprechpausen in Hörbüchern. In *DHd2025*, pages 275–278, Bielefeld. Zenodo.

Michael Vauth, Malte Meister, Hans Ole Hatzel, Dominik Gerstorfer, and Evelyn Gius. 2022. GitMA.

Albin Zehe, Elisabeth Fischer, and Andreas Hotho. 2025. Assessing the state of the art in scene segmentation. *Proceedings of the NAACL 2025*.

Albin Zehe, Leonard Konle, Lea Katharina Dümpelmann, Evelyn Gius, Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, Annekea Schreiber, and Nathalie Wiedmer. 2021a. Detecting Scenes in Fiction: A new Segmentation Task. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3167–3177, Online. Association for Computational Linguistics.

Albin Zehe, Leonard Konle, Svenja Guhr, Lea Katharina Dümpelmann, Evelyn Gius, Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, and Annekea Schreiber, editors. 2021b. *Proceedings of the Shared Task on Scene Segmentation*, volume 3001 of *CEUR Workshop Proceedings*. CEUR, KONVENS 2021, Düsseldorf.

A Appendix: Detailed Error Analysis

In this section, we conduct an in-depth analysis of the prediction errors from the six-sentence USE model (see the confusion matrix in Figure 2).

We begin with an overview of the test data. Among the 493 test cases, the model made 403 correct predictions and 90 incorrect ones. Of these 90

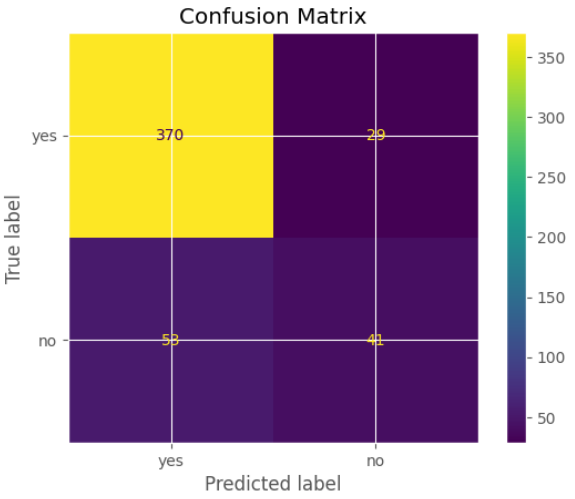

Figure 2: Confusion matrix indicating the predictions on the test set: [[370 29] [53 41]].

errors, 74 were false negatives, and 16 were false positives. This suggests that the model is more prone to overlooking a scene change than mistakenly identifying one when none exists. Through an examination of individual examples, we identify several factors that influence the model predictions: 1) segment length, 2) density of characters, 3) pronoun usage. We will analyze each factor and explore how they manifest in both false positive and false negative cases.

A.1 Length

We compute the average scene segment length for correctly predicted cases, incorrectly predicted cases[7], false positives, and false negatives, as shown in Table 4. Our analysis reveals that incorrect cases tend to have a higher average length than correct ones, suggesting that the model is more prone to errors when processing longer inputs. Additionally, the biggest difference in average length is observed between correct cases and false positives, as shown in Figure 3. With a gap of approximately 150 words, this suggests that the model is more likely to detect a scene change in longer segments.

A.2 Characters

We identify character names as a key factor influencing the model's predictions, particularly in cases where errors occur. To investigate this, we calculate the number of character mentions in each

[7]Correctly predicted cases: true positives and true negatives. Incorrectly predicted cases: false positives and false negatives.

Category	Average segment length	Annot. differing from gold label
correct	337	0
incorrect	425	87
false Negative	411	74
false Positive	488	150

Table 4: Segment length comparison across different categories. The second column calculates the average segment length for each category, while the third column shows the difference in average length between each category and the correct category.

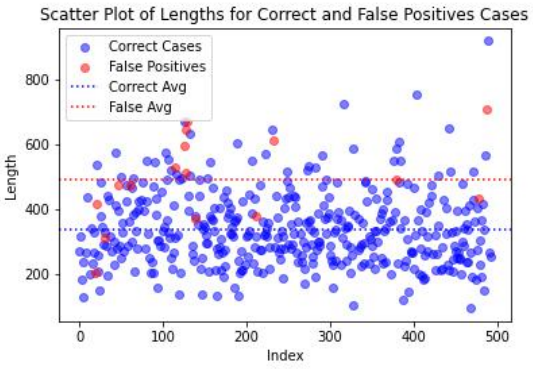

Figure 3: Scatter plot of lengths for correct and false positive cases.

scene segment and compare averages across correct, incorrect, false positive, and false negative cases, as shown in Table 5. Our findings show that, on average, incorrect cases contain slightly more character mentions than correct ones. Notably, the largest difference is observed between correct cases and false positives. As shown in Figure 4, correct cases typically include an average of 1.5 character names, while false positives feature more than 2.5. This suggests that scenes with multiple characters pose a challenge for the model's predictions.

Manually looking into each prediction, we recognize that false positive and false negative cases are governed by different aspects of character mentions. In false positive cases, the model misinter-

Category	Average character count	Annot. differing from gold label
correct	1.67	0
incorrect	2.29	0.62
false negative	2.20	0.53
false positive	2.69	1.02

Table 5: Average character count comparison across different categories. The second column calculates the average character counts for each category, while the third column shows the difference in average character counts between each category and the correct category.

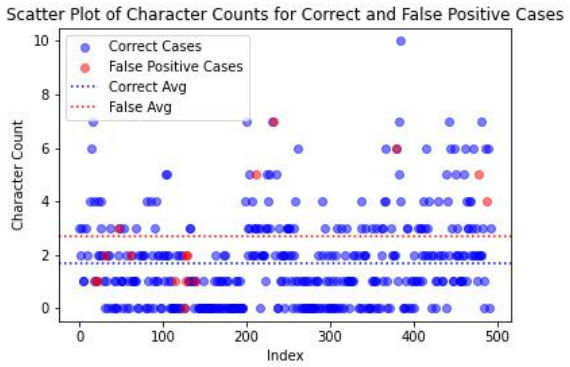

Figure 4: Scatter plot of character counts for correct and false positive cases.

prets a continuous scene as a scene change due to the introduction of a new character, which incorrectly signals a break. Conversely, in false negative cases, actual scene breaks are mistaken for continuity because the model recognizes recurring names or pronouns across scenes, leading to incorrect predictions.

Specifically, the following passage shows an example of when the model predicts a scene change when there is no scene change:

> From the exterior there was no sign of the fire, although there was still work to be done on the inside. The charges against **Abby** had been dropped, and all the loose ends of **Greg's** death and Rusty's betrayal had been tied up. Both **Rusty** and **Richard** had continued to maintain that they'd had nothing to do with the tack under Blackheart's saddle blanket or the hay bale that had nearly killed **Abby**. **Abby** had chalked the incidents up to the hazards and accidents of ranch life. As the wagon drew closer to the dragon tree, all thoughts left **Abby's** head. Beneath the tree, next to the preacher, stood **Luke** and **Cody**.
> (from *Midnight Wishes* by Cassidy)

As for the false negative cases, we recognize that the character patterns can be categorized into two subtypes: 1) same character names; 2) ambiguity of pronouns.

The following passage is an example of the first subcategory where the same character name appears both before and after the scene change. In this instance, the model incorrectly predicts continuity when a scene change actually occurs. We hypothesize that this character continuity misleads the model, resulting in incorrect predictions.

> I think he could have been a real cowboy if he'd tried harder, don't you?" Cody asked. Abby squeezed her son's shoulders sympathetically, unable to speak around the lump of emotion in her throat. Dawn brought a nightmare sight. In the early glow of morning light the full extent of the damage to the house was evident. Abby sat on the bench next to the barn, staring at the gaping black hole that marred the exterior of her home.
> (from *Midnight Wishes* by Cassidy)

The second subcategory arises in situations where similar pronouns appear in both scenes. In such cases, the model may associate the pronouns with the same individuals, leading it to predict continuity when there is actually a scene change. The following passage is an example where a scene change occurs, but the model predicts otherwise. The pronouns indicate the presence of both a male and a female in the first scene, as well as in the second. Despite the time marker "after six" at the beginning of the second scene, we observe that in many cases involving pronoun ambiguity, the model seems to prioritize character continuity over time markers when making its predictions.

> Her head was full of ideals about the world as it should be, and his was full of knowledge about the way it really was. A computer couldn't have picked a man more different from her. HE UNLOCKED THE DOOR and came in shortly after six in the evening. He looked at her apprehensively. She was sitting in the white armchair, watching the rain. She looked paler than usual, and he had a sudden desire to go to her, draw her to her feet and take her in his arms. Except, he thought, that was probably just what she didn't want.
> (from *Pros and Cons* by Campbell)

A.3 Human Annotator

We identify a third group of ambiguous errors, where the reasoning behind the human annotator's decision to mark a scene change is unclear. The following paragraph serves as an example, where the human annotator indicates a scene change, but we cannot identify one, and the model predicts no scene change. It's possible that the scene change occurs at the beginning or end of the segment, but without additional context, it remains uncertain. As a result, we have created a separate error group for these cases, where the ambiguity arises from the lack of clear justification for the scene change, making it difficult to determine the correct interpretation.

> All the time. Nobody to tell her what to do. Not that Carl ever had. Still, she would be all by herself in that big house, keeping her own schedule, marching to her own drummer. Something twisted inside her. The plain fact was this: There was only one tune she wanted to march to, and that was the tune of love.
> (from *Warrior's Embrace* by Webb)