Shared Task on Scene Segmentation@KONVENS 2021

Albin Zehe[*°] **Leonard Konle**[*°] **Svenja Guhr**[†] **Lea Dümpelmann**[•] **Evelyn Gius**[†]
Andreas Hotho[°] **Fotis Jannidis**[°] **Lucas Kaufmann**[°]
Markus Krug[°] **Frank Puppe**[°] **Nils Reiter**[‡] **Annekea Schreiber**[†]
[*]Equal Contribution [°]University of Würzburg [•]University of Heidelberg
[†]TU Darmstadt [‡]University of Cologne

Contact: zehe@informatik.uni-wuerzburg.de

Abstract

This paper describes the Shared Task on Scene Segmentation[1] STSS@KONVENS 2021: The goal is to provide a model that can accurately segment literary narrative texts into *scenes* and *non-scenes*. To this end, participants were provided with a set of 20 contemporary dime novels annotated with scene information as training data. The evaluation of the task is split into two tracks: The test set for Track 1 consists of 4 in-domain texts (dime novels), while Track 2 tests the generalisation capabilities of the model on 2 out-of-domain texts (highbrow literature from the 19th century). 5 teams participated in the task and submitted a model for final evaluation as well as a system description paper, with the best-performing models reaching F1-scores of 37 % for Track 1 and 26 % for Track 2. The results show that the task of scene segmentation is very challenging, but also suggest that it is feasible in principle. Detailed evaluation of the predictions reveals that the best-performing model is able to pick up many signals for scene changes, but struggles with the level of granularity that actually constitutes a scene change.

1 Introduction

The objective of this shared task is to develop a model capable of solving the task of scene segmentation, as discussed by Gius et al. (2019) and formally introduced by Zehe et al. (2021). According to their definition, a scene can be understood as "a segment of a text where the story time and the discourse time are more or less equal, the narration focuses on one action and space and character constellations stay the same". The task of scene segmentation is therefore a kind of text segmentation task applicable specifically to narrative texts (e.g., novels or biographies): These texts can be seen as a sequence of segments, where some of the segments are *scenes* and some are *non-scenes*. The goal of scene segmentation is to provide both the borders of the segments as well as the classification of each segment as a scene or non-scene. Solving this task advances the field of computational literary studies: the texts of interest in this field are often very long and can therefore not easily be processed with NLP methods. Breaking them down into narratologically motivated units of meaning like scenes would enable processing these units (semi-)individually and then aggregating the results over the entire text. In addition, a segmentation into scenes allows plot- and content-based analyses of the texts.

2 Background: Scene Segmentation

This section provides an overview of the task of scene segmentation according to the definition by Zehe et al. (2021). For the full motivation and description, we refer to this paper.

From a narratological point of view, a scene can be defined by reference to a set of four dimensions: *time*, *space*, *action* and *character constellation*. Using these dimensions, a *scene* is a segment of the *discours* (presentation) of a narrative which presents a part of the *histoire* (connected events in the narrated world) such that (1) time is equal in *discours* and *histoire*, (2) place stays the same, (3) it centers around a particular action, and (4) the character constellation is equal. All of these conditions are not absolute but rather relative, that is, small changes in either of them do not necessarily lead to a scene change but can rather be seen as indicators.

Casting this definition as a machine learning task, we receive as input a (narrative) text and want to develop a model that (a) splits the text into a sequence of segments and (b) labels each of these segments either as a scene or as a non-scene. Depending on the realisation of the changes, there are

[1] https://go.uniwue.de/stss2021

strong or weak boundaries. Segments separated by a weak boundary can be aggregated into one segment, while segments with hard boundaries need to be considered separately.

3 Related Work

The related work for scene segmentation has been covered in much detail by Zehe et al. (2021). For completeness, we reproduce their overview here with only minor adaptation:

Segmentation tasks have been discussed in NLP for a while, mostly with the goal of identifying regions of news or other non-fictional texts discussing certain topics. The task of topic segmentation is then to identify points in the text where the topic under discussion changes. Early work to this end uses similarity of adjacent text segments (such as sentences or paragraphs) with a manually designed similarity metric in order to produce the resulting segments. One of the most well known systems of this manner is TextTiling (Hearst, 1997), which was applied to science magazines. Similarity based on common words (Choi, 2000; Beeferman et al., 1999) was superseded with the introduction of Latent Dirichlet Allocation (Blei et al., 2003), which allowed to segment the text into coherent text snippets with similar topic distributions (Riedl and Biemann, 2012; Misra et al., 2011). This procedure was extended by the integration of entity coherence (John et al., 2016) and Wanzare et al. (2019) have used it on (very short) narrative texts in an attempt to extract scripts. Recently, many approaches making use of neural architectures deal with the detection and classification of local coherence (e. g. Li and Jurafsky, 2016; Pichotta and Mooney, 2016; Li and Hovy, 2014), which is an important step for a text summarization of high quality (Xu et al., 2019). Text segmentation using neural architectures was conducted on Chinese texts and it was shown that recurrent neural networks are able to predict the coherence of subsequent paragraphs with an accuracy of more than 80 % (Pang et al., 2019). Lukasik et al. (2020) compare three BERT based architectures for segmentation tasks: Cross-Segment BERT following the NSP Pretraining-Task and fine-tuned on segmentation, a Bi-LSTM on top of BERT to keep track of larger context and an adaption of a Hierarchical BERT network (Zhang et al., 2019).

Some work has been done on segmenting narrative texts, but aiming at identifying topical segments – which, as we have pointed out above, is different from scene segmentation. With a set of hand-crafted features, Kauchak and Chen (2005) achieve a WindowDiff score (Pevzner and Hearst, 2002) of about 0.5, evaluated on two novels. Kazantseva and Szpakowicz (2014) have annotated the novel *Moonstone* with topical segments, and presented a model to create a hierarchy of topic segments. They report about 0.3 WindowDiff score. Recently, Pethe et al. (2020) have introduced the task of chapter segmentation, which is similar to scene segmentation in that they both focus on narrative texts. However, it aims at detecting chapters, which are based on structural information like headers, whereas scenes are defined by features of the told story not directly connected to structural information. Notably, our dataset contains some scenes that cross chapter boundaries, since our characteristics of scenes are entirely independent of such formal markers. Most closely related to our task are the papers by Reiter (2015), who documents a number of annotation experiments, and Kozima and Furugori (1994), who present lexical cohesiveness based on the semantic network Paradigme (Kozima and Furugori, 1993) as an indicator for scene boundaries and evaluates their approach qualitatively on a single novel. However, neither of them provide annotation guidelines, annotated data or a formal definition of the task.

A related area of research is discourse segmentation, where the goal is also to find segments that are not necessarily defined by topic, and are also assigned labels in addition to the segmentation. There are annotated news corpora in this area featuring fine-grained discourse relations between relatively small text spans (Carlson et al., 2002; Prasad et al., 2008). Although larger structures have been discussed in literature (Grosz and Sidner, 1986), no annotated corpora have been released.

4 Shared Task on Scene Segmentation - STSS

The Shared Task on Scene Segmentation was organised as one of the shared tasks of KONVENS 2021.[2] There were a total of 8 registrations, out of which 5 teams submitted a model for the final evaluation as well as a system description paper. The task was split into two tracks, with the first one evaluating on in-domain data and the second one on out-of-domain data. The test data was kept back for the entire duration of the task and trained

[2] https://konvens2021.phil.hhu.de/

models were submitted to the organisers as Docker images for the final evaluation.

4.1 Data

Trial Data A single text, "Der kleine Chinesengott" by Pitt Strong (aka Elisabeth von Aspern) was released as trial data before the actual training set, in order to show the format of the dataset and enable participants to start working on their implementation as soon as possible.

Training Data The training data consisted of 20 annotated dime novels, which include the 15 texts from Zehe et al. (2021) as well as 5 new texts that were annotated according to the same guidelines. The texts are given in the appendix in Table 3, along with detailed dataset statistics (Table 4). Since the texts are protected by copyright, they could not be distributed directly. Instead, participants were asked to register for a German ebook shop[3] and received the books as a gift on this website, along with standoff annotations and a script to merge the epub files with the annotations.

Evaluation Data

Track 1 The first subset of the evaluation data, used in Track 1 of the shared task, consisted of 4 texts from the same domain as the training set, that is, dime novels. Detailed statistics for this dataset are available in Table 5.

Track 2 The second evaluation set, used for Track 2, consisted of out-of-domain data, specifically 2 high-brow literature novels. This set, presented in detail in Table 6, was chosen to investigate how well the submitted approaches were able to deal with texts that are assumed to differ strongly from the training data in writing style.

4.2 Evaluation Metrics

Evaluating scene segmentation is a somewhat challenging problem in itself. Zehe et al. (2021) use two evaluation metrics, F1-score and Mathet's γ (Mathet et al., 2015), arguing that γ is the more suitable measure for scene segmentation: F1-score only counts a scene boundary as correct if it is predicted at exactly the right position, while an offset of one sentence would already count as a complete miss. On the other hand, γ tries to align the predicted boundaries with the gold boundaries and score both the fit of the alignment as well as the classification into scene and non-scene. However, since the γ measure itself requires the user to specify certain parameters and it is not immediately obvious how to set these parameters in our context, the main evaluation in this shared task is based on the exact F1-score. More precisely, we represent the segmentation produced by each system as a list of boundary predictions: Each sentence in the text is labelled as either `NOBORDER`, `SCENE-TO-SCENE`, `SCENE-TO-NONSCENE` or `NONSCENE-TO-SCENE`. For example, a sentence that starts a new scene after a segment that is classified as a non-scene would be labelled as `NONSCENE-TO-SCENE`. This classification can then directly be compared to the gold standard annotations.

The classes in this scheme are highly imbalanced, with `NOBORDER` making up the vast majority of the labels. Therefore, for our main evaluation, we exclude the label `NOBORDER` and build micro-averaged scores between the other classes. We chose to use micro-averaging despite the class imbalance since the minority classes are not more important to the classification and therefore micro-averaged scores lead to a better representation of the overall classification performance.

For informative reasons, we also report the γ score of the approaches.

4.3 Submitted Systems

This section provides an overview of the approaches to scene segmentation submitted by the participants of the Shared Task.

Kurfali and Wirén (2021) apply the sequential sentence classification system proposed by Cohan et al. (2019) to the scene-segmentation task. This system is based on BERT, but uses a customised input format, where each sentence of the input sequence is separated by BERT's special token "[SEP]". After passing a sequence through BERT, the output of those "[SEP]" tokens is fed into a multi-layer perceptron to predict a label for its preceding sentence. While the original system utilises a mean-squared-error loss, Kurfali and Wirén (2021) implement weighted cross-entropy to deal with the class imbalance in the scene dataset and make use of the IOB2 scheme instead of simple classification with categories.

The system submitted by Gombert (2021) builds on the idea to use sentences functioning as scene borders as feature vectors for the prediction of

[3] https://www.ebook.de/de/

scene borders. For this purpose, first a sentence embedding space is learned in a twin BERT training setup. The model separates sentences functioning as scene borders from sentences within scenes. In a second step, a gradient-boosted decision tree ensemble is fed with feature vectors from the sentence embeddings generated by the model.

The system submitted by Barth and Dönicke (2021) focuses on the manual design of vectors covering different sets of features for scene segmentation. The first set consists of general linguistic features like tense, POS tags, etc. The other sets focus on features crucial for the scene segmentation task, explicitly encoding temporal expressions as well as entity mentions. These feature vectors are then used as input to a random forest classifier.

The system of Hatzel and Biemann (2021) casts the problem of scene segmentation as a kind of next-sentence-prediction: It focuses on the "[SEP]" tokens which appear in between two subsequent sentences in the input representation for a BERT model, and uses their embedding representation from a German BERT model. In addition to the BERT-embeddings, the authors add manual features capturing changes in the character constellation that are derived from a German adaptation of the coarse-to-fine co-reference architecture (Lee et al., 2018). This final representation is fed into a fully connected layer with a softmax activation function in order to detect scene changes. Since this approach predicts too many scenes in close proximity, they evaluate different ways to suppress neighbouring scenes for their final prediction. Specifically, they use a cost function which punishes very short scenes harshly.

The team Schneider et al. (2021) present the "Embedding Delta Signal" as a method for both scene segmentation and topic segmentation. They focus on context change in documents using a sliding window method that compares cluster assignments of word embeddings using the cosine distance measure and detect scene changes by searching for local maxima in the signal. In a further step, they distinguish between different scene types using a simple support vector machine approach with hyper-parameter search. They use an additional evaluation method, intersection over union of predicted and actual scenes, arguing that this measure is more suitable because it punishes scene boundaries that are in the vicinity of the gold annotations less severely than the F1-score.

4.4 Evaluation of the Automatic Systems for Scene Segmentation

In the following, we present and discuss the performance of the submitted systems in our shared task. All results are summarised in Table 1.

The most successful system on Track 1 was the one proposed by Kurfali and Wirén (2021), reaching an F1-score of 37 % on the evaluation set for Track 1. For Track 2, their model was somewhat less successful, reaching an F1-score of 17 %, which still corresponds to the second place. On Track 2, the system proposed by Gombert (2021) performed best, with an F1-score of 26 % (16 % on Track 1). All results for both systems, with evaluation for all border classes on individual texts, can be found in the appendix in Tables 7 and 8. Overall, these results show that scene segmentation is a very challenging, but not impossible task. Especially the winning system is capable of finding 51 % of all annotated scene boundaries in the in-domain data, which is a promising score. The bigger issue of this system at the moment seems to be the precision (29 %), indicating that many of the boundaries the systems predicts are wrong. We provide an analysis of what leads to these results in the next section.

Interestingly, all systems except the one from Kurfali and Wirén (2021) actually performed better on the out-of-domain evaluation set of Track 2 than on the (in-domain) dime novels of Track 1. However, it must also be noted that the scores are overall somewhat low and the differences should therefore not be overinterpreted. We can also see that the ranking according to the γ measure would be rather similar to the F1-score-ranking. However, there are also differences in the ranking, for example the system submitted by Hatzel and Biemann (2021) would have been ranked higher in both tracks according to γ. This shows that the selection of a fitting evaluation measure for scene segmentation is indeed important.

4.5 Additional Evaluation

Addressing the fact that our F1-score is a very unforgiving measure, since only exact matches are counted as correct scene boundaries, we performed some additional evaluation on the predictions by the different systems.

As a first step, we noticed that some of the systems had a tendency to predict multiple short scenes in the vicinity of a hand-annotated scene change. Therefore, we conducted an additional eva-

System	Track 1				Track 2			
	Prec.	Rec.	F1	γ	Prec.	Rec.	F1	γ
Kurfali and Wirén (2021)	**0.29**	**0.51**	**0.37**	**0.19**	0.14	**0.22**	0.17	**0.31**
Gombert (2021)	0.22	0.13	0.16	0.09	**0.39**	0.20	**0.26**	0.17
Barth and Dönicke (2021)	0.06	0.11	0.07	0.06	0.13	0.12	0.12	0.08
Hatzel and Biemann (2021)	0.02	0.03	0.02	0.12	0.08	0.17	0.11	0.13
Schneider et al. (2021)	0.01	0.02	0.02	0.05	0.06	0.03	0.04	0.06

Table 1: Micro avg. Precision, Recall, F1-score and Mathet's γ for all submissions in both tracks of the STSS

luation where we merged scenes that were less than 5 sentences long to the preceding or following scene, if this led to a correctly predicted scene (e.g., if the beginning of the short scene was a gold scene boundary and the end of the next scene was a gold scene boundary, the two scenes were merged). This improved some of the scores by up to 3 percentage points in F1-score. Note that this is not a "valid" evaluation, since the decision whether to merge to the preceding or following scene is taken based on the gold standard. However, it does show that correct handling of short scenes would have some positive influence on the results.

Additionally, we analysed whether we could determine especially "important" scene boundaries more reliably. To this end, the existing annotations of Track 1 were re-edited: Annotators were asked to identify *strong* and *weak* boundaries between the previously annotated scenes, depending on how they judged the importance of each boundary. A strong boundary is one that must be set in any annotation, while a weak boundary is one that may be omitted based on the desired level of granularity. Note that we did not collect any additional scene annotations, but only categorised the existing ones further. We did not see significant changes in the performance when considering only *strong* boundaries. In particular, the recall was not consistently higher than for all scene boundaries.

5 Manual Error Analysis

In this section, we provide a deeper analysis of the prediction errors that the best-performing system on Track 1 (Kurfali and Wirén, 2021) makes. To this end, we manually analyse the predictions and potential error sources on two texts from Track 1:

- *Hochzeit wider Willen* (*Wedding against will*, the text with the best γ score)
- *Bomben für Dortmund* (*Bombs for Dortmund*, the text with the second worst γ score; we decided not to use the text *Die Begegnung*, which has the worst γ score, since it was a very hard text even for the annotators)

Table 2 compares the manual to the automatic annotations for these texts. The analysis reveals that the following factors have a particular influence on the predictions: (a) Length of the detected scenes or granularity of scene detection in general (b) explicit markers of time and space changes, (c) changes in character constellation (entrance and exit of characters, especially protagonists), (d) naming and description of newly introduced characters (full name plus verb sequence), as well as (e) end of dialog passages. We provide a brief overview of the problematic factors here and refer to Appendix C for a detailed analysis with specific examples.

5.1 Analysis of Markers

First, we investigate how the markers used in our definition of scenes influence the system's decisions regarding scene borders.

Time Markers The system clearly seems to have identified time markers as an important signal for scene changes. Many false positives (scene borders annotated by the system, but not by the human annotators) start with temporal markers, especially the word "as". Overall, the system appears to have overgeneralised the impact of temporal markers, seeing every mention of time in the text as a strong signal for a scene change.

Location Markers A similar issue arises with the presence of location markers: the system is very sensitive to changes in action space, often producing false positives at the mention of locations. According to our annotation guidelines, only significant location changes induce a scene border while, for example, moving through rooms in a house is not necessarily cause enough for a scene change.

	Bomben für Dortmund	*Hochzeit wider Willen*
total length (tokens)	28830	26042
longest gold standard scene (tokens)	1920	1474
longest correctly predicted scene (tokens)	967	1401
annotated segments by winner system	78	98
annotated segments in gold standard	45	60

Table 2: Information on the two sample texts for the following error analysis comparing the output of the winner system with the gold standard annotations.

Changes in Character Constellation Another marker that our scene definition takes into account is the character constellation. We find that the model is capable of identifying the introduction of a new character, often accompanied with the character's full name as well as a short description, as a marker for a new scene. However, once again the system seems to struggle with the importance of character constellation changes, showing a tendency to start a scene for every introduction.

Dialogue Passages Dialogue passages are not explicitly part of our scene definition, however it is reasonable to assume that they can be valuable markers for scenes: for one, dialogues appear almost exclusively in scenes, rarely in non-scenes. Additionally, a new scene usually does not start in the middle of a dialogue passage. The model seems to have picked up this fact, since it has a tendency to predict scene changes on the end of dialogue passages. While this can be a valid marker, it again leads to false positives in the system's output.

5.2 General Issues of the Model's Output

Here, we attempt to extract a generalisation of the specific issues described before. They can be grouped into two major categories: issues with scene length and issues with the granularity of markers.

Scene Length One of the most general problems was that the system predicts very short scenes in succession, often caused by the occurrence of multiple markers within a few sentences. In our manual annotations, very short passages are usually not considered as separate scenes, but rather as part of the preceding or following scene. The system does not appear to have learned this and therefore often predicts multiple very short scenes in succession.

Granularity of Markers An issue that was noticeable for any of the markers discussed above is the system's apparent inability to infer the importance of a scene change marker. Many false positive predictions are caused by small changes in time, place or character constellation that were not considered as significant enough for a scene change by the annotators. In some cases, the model's decision to predict a scene change is perfectly reasonable and can be seen as a more fine-grained scene segmentation than the one agreed on in our annotations (cf. Section 6). In other cases, however, the oversensitivity of the system is clearer, as for example with the temporal marker "as" (see above).

6 Discussion

In this section, we briefly discuss the results of the shared task along with possible next steps towards the improvement of automatic scene segmentation.

The winning systems for both tracks (Kurfali and Wirén, 2021; Gombert, 2021) are based on BERT variants, showing that, as for many other NLP-tasks, pre-trained Transformer models are very valuable for scene segmentation. However, the results also reinforce our belief that scene segmentation cannot be solved by BERT alone, but requires a deeper understanding of the text. Some of the submissions of the shared task explore alternative ways of approaching scene segmentation, either adapting methods from co-reference resolution (Hatzel and Biemann, 2021), handcrafting features that are assumed to be helpful for scene segmentation (Barth and Dönicke, 2021), or using differences in the text over time to derive scene change candidates (Schneider et al., 2021).

The two most consistent sources of errors in the most successful model are the granularity of scene change markers and the length of scenes. Both of these problems should be – at least in part – addressable by introducing additional constraints or signals to the model. For the scene length, it seems promising to make the model aware of the length of the current scene, which could prevent it

from predicting many short scenes, or to use global information about the scene boundaries. A possible approach to this has been used by Pethe et al. (2020) for the related task of chapter segmentation and was also applied in this shared task by Hatzel and Biemann (2021) with some success.

For the problem of granularity, the model could be given access to explicit information regarding the scale of the markers. For example, information from knowledge graphs about the scale of temporal markers or location changes could be useful (e.g., a minute is much less relevant than a month; a different room is much less relevant than a different country). Character changes appear to be more challenging in this regard, since the model needs to be able to judge the importance of a character for the current scene. This might be achieved by applying co-reference resolution to the texts and building a local character network, representing how many interactions each character has with others in the neighbouring text, how often they are mentioned, etc. Although a somewhat boring solution, using more training data might also enable the model to learn the granularity of markers, at least for location and temporal markers. A possible step in this direction is to use the related task of chapter segmentation (Pethe et al., 2020), for which a large amount of weakly labelled training data is available, for pre-training and then fine-tuning the resulting model for scene segmentation. While chapters and scenes are different in principle (cf. Section 3), they may be similar enough to make this pre-training step promising. On the other hand, it might be interesting to explore the scene segmentations provided by the model further. Our annotations represent *our* understanding of a scene, however other applications may require a more fine- or coarse-grained definition. To this end, it seems promising to optimise a model for recall (i.e., detect as many annotated scene borders as possible) in a first step and then filter these candidates for the desired level of granularity in a second step.

One of the most surprising results of the shared task is the fact that most models perform better on the out-of-domain high-brow literature than the in-domain dime novels. This is in stark contrast to our previous intuition, for two reasons: First, the training data consists of dime novels, which should lead to a model that is better suited to this type of texts. Secondly, from a literary perspective, we expected the high-brow literature to be more challenging to understand and therefore the scene segmentation to be more difficult. However, the more implicit style of writing in high-brow literature may actually be helpful for the models here. While dime novels often present explicit references to characters, locations or the passing of time, high-brow literature may use these references much more sparsely, making them more reliable markers of scene changes. Although the number of data points is too low to make a reliable statement, the higher precision of predictions from Gombert (2021) on the high-brow texts compared to the dime novels (cf. Table 8) might point in a similar direction.

Finally, we also see that the choice of evaluation measure is important, as F1-score and γ lead to different rankings in both tracks. For this shared task, we have decided to use the exact F1-score as the main measure, however this decision is not final. As already discussed before, measures that take into account the proximity of predicted to gold standard scenes, like γ, are equally valid, albeit more difficult to interpret. Schneider et al. (2021) propose a third potentially useful measure, intersection over union. While this measure would have to be adapted to be able to handle both non-scenes and scenes, this is also a promising direction.

7 Conclusion

In this paper, we have summarised the results of the Shared Task on Scene Segmentation, where the objective was to develop a method for automatic scene segmentation in literary narrative texts. To this end, we provided a training set of 20 dime novels and evaluated the submitted systems on two tracks, one with in-domain data and one with out-of-domain data in the form of high-brow literature. Overall, our shared task has received five submissions with very different approaches to scene segmentation. While none of these systems were capable of solving the task completely, especially the best performing systems for each track yielded promising results, with F1-scores of 37 % on Track 1 and 26 % on Track 2, respectively. These results show that scene segmentation remains challenging, but also that it is not an impossible task. In manual analysis, we discovered that the models are capable of picking up many important markers for scene boundaries, but sometimes still struggle to draw the correct conclusions from these markers.

Acknowledgements

We would like to thank all participants for their submissions. We are especially happy about the wide range of completely different and orthogonal approaches, opening great possibilities for future work on this challenging task!

References

Florian Barth and Tillmann Dönicke. 2021. Participation in the konvens 2021 shared task on scene segmentation using temporal, spatial and entity feature vectors. In *Shared Task on Scene Segmentation*.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34(1-3):177–210.

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

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. Rst discourse treebank, ldc2002t07. Technical report, Philadelphia: Linguistic Data Consortium.

Freddy YY Choi. 2000. Advances in domain independent linear text segmentation. *arXiv preprint cs/0003083*.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. 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 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Evelyn Gius, Fotis Jannidis, Markus Krug, Albin Zehe, Andreas Hotho, Frank Puppe, Jonathan Krebs, Nils Reiter, Nathalie Wiedmer, and Leonard Konle. 2019. Detection of scenes in fiction. In *Proceedings of Digital Humanities 2019*.

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

Sebastian Gombert. 2021. Twin bert contextualized sentence embedding space learning and gradient-boosted decision tree ensembles for scene segmentation in german literature. In *Shared Task on Scene Segmentation*.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Hans Ole Hatzel and Chris Biemann. 2021. Applying coreference to literary scene segmentation. In *Shared Task on Scene Segmentation*.

Marti A Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Adebayo Kolawole John, Luigi Di Caro, and Guido Boella. 2016. Text segmentation with topic modeling and entity coherence. In *International Conference on Hybrid Intelligent Systems*, pages 175–185. Springer.

David Kauchak and Francine Chen. 2005. Feature-based segmentation of narrative documents. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 32–39, Ann Arbor, Michigan. Association for Computational Linguistics.

Anna Kazantseva and Stan Szpakowicz. 2014. Hierarchical topical segmentation with affinity propagation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 37–47, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Hideki Kozima and Teiji Furugori. 1993. Similarity between words computed by spreading activation on an english dictionary. *In Proceedings of the European Association for Computational Linguistics*.

Hideki Kozima and Teiji Furugori. 1994. Segmenting narrative text into coherent scenes. *Literary and Linguistic Computing*, 9(1):13–19.

Murathan Kurfali and Mats Wirén. 2021. Breaking the narrative: Scene segmentation through sequential sentence classification. In *Shared Task on Scene Segmentation*.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Jiwei Li and Eduard Hovy. 2014. A model of coherence based on distributed sentence representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2039–2048.

Jiwei Li and Dan Jurafsky. 2016. Neural net models for open-domain discourse coherence. *arXiv preprint arXiv:1606.01545*.

Michal Lukasik, Boris Dadachev, Gonçalo Simões, and Kishore Papineni. 2020. Text segmentation by cross segment attention.

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.

Hemant Misra, François Yvon, Olivier Cappé, and Joemon Jose. 2011. Text segmentation: A topic modeling perspective. *Information Processing & Management*, 47(4):528–544.

Yihe Pang, Jie Liu, Jianshe Zhou, and Kai Zhang. 2019. Paragraph coherence detection model based on recurrent neural networks. In *International Conference on Swarm Intelligence*, pages 122–131. Springer.

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.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Comput. Linguist.*, 28(1):19–36.

Karl Pichotta and Raymond J Mooney. 2016. Learning statistical scripts with lstm recurrent neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Rashmi Prasad, Alan Lee, Nikhil Dinesh, Eleni Miltsakaki, Geraud Campion, Aravind Joshi, and Bonnie Webber. 2008. Penn Discourse Treebank Version 2.0 LDC2008T05. Web download, Linguistic Data Consortium, Philadelphia.

Nils Reiter. 2015. Towards Annotating Narrative Segments. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 34–38, Beijing, China. Association for Computational Linguistics.

Martin Riedl and Chris Biemann. 2012. Topictiling: a text segmentation algorithm based on lda. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42. Association for Computational Linguistics.

Felix Schneider, Björn Barz, and Joachim Denzler. 2021. Detecting scenes in fiction using the embedding delta signal. In *Shared Task on Scene Segmentation*.

Lilian Diana Awuor Wanzare, Michael Roth, and Manfred Pinkal. 2019. Detecting everyday scenarios in narrative texts. In *Proceedings of the Second Workshop on Storytelling*, pages 90–106, Florence, Italy. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Discourse-aware neural extractive model for text summarization. *arXiv preprint arXiv:1910.14142*.

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. 2021. 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.

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

A Dataset Information

	Title	Author	Publisher	Series	Genre	EAN
Training	Die hochmütigen Fellmann-Kinder	Patricia Vandenberg	Martin Kelter	Sophienlust Classic	Familienroman	9783740950484
	Der Sohn des Kometen	Hugh Walker	Pabel Moewig Verlag	Mythor	Fantasy	9783845397535
	Als der Meister starb	Wolfgang Hohlbein	Bastei Lübbe	Der Hexer	Fanstasy	9783838721675
	Der Turm der 1000 Schrecken	Jason Dark	Bastei Lübbe	John Sinclair	Horror	9783838727868
	Bezaubernde neue Mutti	Regine König	Martin Kelter	Fürstenkinder	Adelsroman	9783740965716
	Immer wenn der Sturm kommt	O. S. Winterfield	Bastei Lübbe	Die schwarzen Perlen	Fantasy	9783732517695
	Hetzjagd durch die Zeit	Frank Rehfeld	Bastei Lübbe	Dino-Land	Abenteuer	9783732535200
	Lass Blumen sprechen	Verena Kufsteiner	Bastei Lübbe	Das Berghotel	Heimatroman	9783732591732
	Ein Weihnachtslied für Dr. Bergen	Marina Anders	Bastei Lübbe	Notärztin Andrea Bergen	Arztroman	9783732557905
	Prophet der Apokalypse	Manfred Weinland	Bastei Lübbe	2012	SciFi	9783838713625
	Die Abrechnung	Frank Callahan	Bastei Lübbe	Skull Ranch	Western	9783732597314
	Tausend Pferde	G.F. Unger	Bastei Lübbe	G.F. Unger Sonder-Edition	Western	9783732596249
	Verschmäht	Hedwig Courths-Mahler	Bastei Lübbe	Hedwig Courths-Mahler	Liebesroman	9783732502929
	Wechselhaft wie der April	Andreas Kufsteiner	Bastei Lübbe	Der Bergdoktor	Arztroman	9783732591725
	Wir schaffen es - auch ohne Mann	Friederike von Bucher	Bastei Lübbe	Toni der Hüttenwirt	Heimatroman	9783740941093
	Ein sündiges Erbe	Jack Slade	Bastei Lübbe	Lassiter	Western	9783732596881
	Griseldis	Hedwig Courths-Mahler	Bastei Lübbe	Hedwig Courths-Mahler	Liebesroman	9783732522033
	Deus Ex Machina	Jana Paradigi & Ramon M. Randle	Bastei Lübbe	Maddrax	SciFi	9783732584017
	Die Widows Connection	Jerry Cotton	Bastei Lübbe	Jerry Cotton	Krimi	9783732596546
	Widerstand zwecklos	Roma Lentz	Bastei Lübbe	Silvia-Gold	Liebesroman	9783732586875
Eval 1	Im Bann der Vampire	Emily Blake	Romantruhe	Dunkelwelt der Anderen	Horror	9783864734816
	Die Begegnung	Alfred Bekker	Bastei Lübbe	Bad Earth	SciFi	9783732548767
	Hochzeit wider Willen	Diana Laurent	Bastei Lübbe	Fürsten-Roman	Liebe	9783838751405
	Bomben für Dortmund	Peter Heben	Bastei Lübbe	Der Bundesbulle	Krimi	9783732538201
Eval 2	Effi Briest	Theodor Fontane	TextGrid Repository	-	-	-
	Aus guter Familie	Gabriele Reuter	Projekt Gutenberg	-	-	-

Table 3: Detailed information on the training and test dataset.

Title	Number of Segments	Percentage of Scenes	Number of Sentences	Number of tokens	Avg. Scene Length (Tokens)
Bezaubernde neue Mutti	44	1.00	2 742	26 771	608.43
Widerstand zwecklos	37	0.84	1 896	21 125	570.95
Tausend Pferde	104	0.99	3 353	38 016	365.20
Der Turm der 1000 Schrecken	58	1.00	2 862	24 657	425.03
Die Widows Connection	66	0.82	2 748	26 831	406.53
Ein sündiges Erbe	43	0.98	3 081	26 363	613.09
Immer wenn der Sturm kommt	82	0.87	3 288	30 245	368.48
Wechselhaft wie der April	73	0.96	2 826	23 939	327.92
Lass Blumen sprechen	54	0.96	2 526	24 301	449.59
Prophet der Apokalypse	65	0.98	2 277	25 502	391.91
Verschmäht	77	1.00	2 823	32 285	419.10
Der Sohn des Kometen	41	0.95	2 250	26 393	643.29
Ein Weihnachtslied für Dr. Bergen	70	0.96	2 546	25 378	362.16
Die hochmütigen Fellmann-Kinder	79	0.94	2 993	30 852	390.54
Als der Meister starb	82	0.98	4 503	62 920	767.32
Hetzjagd durch die Zeit	52	0.96	2 310	28 807	553.94
Wir schaffen es - auch ohne Mann	53	0.87	3 250	25 649	483.66
Griseldis	75	0.91	2 729	33 913	452.17
Deus Ex Machina	43	1.00	2 371	25 886	602.00
Die Abrechnung	56	1.00	2 439	23 541	420.38
mean	62.70	0.95	2 790.65	29 168.70	481.08
std	17.56	0.06	555.57	8 883.18	118.72
min	37	0.82	1 896	21 125	327.92
25%	50	0.93	2 422	25 197.75	391.57
50%	61.50	0.96	2 745	26 378	437.31
75%	75.50	0.99	3 015	30 396.75	578.71
max	104	1.00	4 503	62 920	767.32

Table 4: Statistics for the training dataset

Title	Number of Segments	Percentage of Scenes	Number of Sentences	Number of tokens	Avg. Scene Length (Tokens)
Im Bann der Vampire	25	0.96	1 826	19 511	780.44
Bad Earth	38	0.87	2 098	22 592	594.32
Hochzeit wider Willen	60	0.82	2 665	26 042	433.98
Bomben für Dortmund	45	0.98	3 244	28 830	640.67
mean	42	0.91	2 458.25	24 243.75	612.35
std	14.58	0.08	629.73	4 057.69	142.82
min	25	0.82	1 826	19 511	433.98
25%	34.75	0.86	2 030	21 821.75	554.23
50%	41.50	0.91	2 381.50	24 317	617.49
75%	48.75	0.96	2 809.75	26 739	675.61
max	60	0.98	3 244	28 830	780.44

Table 5: Statistics for the evaluation dataset of Track 1

Title	Number of Segments	Percentage of Scenes	Number of Sentences	Number of tokens	Avg. Scene Length (Tokens)
Aus guter Familie	220	1.00	6 312	74 517	331.53
Effi Briest	220	0.85	6 849	98 037	443.21
mean	220	0.92	6 580.50	86 277	387.37
std	0	0.11	379.72	16 631.15	78.97
min	220	0.85	6 312	74 517	331.53
25%	220	0.88	6 446.25	80 397	359.45
50%	220	0.92	6 580.50	86 277	387.37
75%	220	0.96	6 714.75	92 157	415.29
max	220	1.00	6 849	98 037	443.21

Table 6: Statistics for the evaluation dataset of Track 2

B Detailed Results

	prec.	rec.	f1	supp.
S-S	0.39	0.64	0.48	22
S-NS	0.00	0.00	0.00	1
NS-S	0.00	0.00	0.00	1
micro avg	0.33	0.58	0.42	24
macro avg	0.13	0.21	0.16	24
weighted avg	0.36	0.58	0.44	24

(a) Im Bann der Vampire

	prec.	rec.	f1	supp.
S-S	0.21	0.52	0.29	27
S-NS	0.00	0.00	0.00	5
NS-S	0.00	0.00	0.00	5
micro avg	0.21	0.38	0.27	37
macro avg	0.07	0.17	0.10	37
weighted avg	0.15	0.38	0.22	37

(b) Die Begegnung

	prec.	rec.	f1	supp.
S-S	0.37	0.84	0.52	37
S-NS	0.14	0.09	0.11	11
NS-S	0.00	0.00	0.00	11
micro avg	0.33	0.54	0.41	59
macro avg	0.17	0.31	0.21	59
weighted avg	0.26	0.54	0.34	59

(c) Hochzeit wider Willen

	prec.	rec.	f1	supp.
S-S	0.30	0.55	0.39	42
S-NS	0.00	0.00	0.00	1
NS-S	0.00	0.00	0.00	1
micro avg	0.28	0.52	0.37	44
macro avg	0.10	0.18	0.13	44
weighted avg	0.29	0.52	0.37	44

(d) Bomben für Dortmund

	prec.	rec.	f1	supp.
S-S	0.09	0.13	0.10	219
S-NS	0.00	0.00	0.00	0
NS-S	0.00	0.00	0.00	0
micro avg	0.08	0.13	0.10	219
macro avg	0.03	0.04	0.03	219
weighted avg	0.09	0.13	0.10	219

(e) Aus guter Familie

	prec.	rec.	f1	supp.
S-S	0.20	0.45	0.28	152
S-NS	0.00	0.00	0.00	33
NS-S	0.00	0.00	0.00	33
micro avg	0.20	0.31	0.24	218
macro avg	0.07	0.15	0.09	218
weighted avg	0.14	0.31	0.19	218

(f) Effi Briest

Table 7: Evaluation of the winner system from Track 1 (Kurfali and Wirén, 2021) on all texts in the test set and all border types (`SCENE-TO-SCENE`, `SCENE-TO-NONSCENE`, `NONSCENE-TO-SCENE`).

	prec.	rec.	f1	supp.
S-S	0.25	0.09	0.13	22
S-NS	0.00	0.00	0.00	1
NS-S	0.00	0.00	0.00	1
micro avg	0.22	0.08	0.12	24
macro avg	0.08	0.03	0.04	24
weighted avg	0.23	0.08	0.12	24

(a) Im Bann der Vampire

	prec.	rec.	f1	supp.
S-S	0.15	0.07	0.10	27
S-NS	0.00	0.00	0.00	5
NS-S	0.00	0.00	0.00	5
micro avg	0.15	0.05	0.08	37
macro avg	0.05	0.02	0.03	37
weighted avg	0.11	0.05	0.07	37

(b) Die Begegnung

	prec.	rec.	f1	supp.
S-S	0.33	0.43	0.38	37
S-NS	1.00	0.09	0.17	11
NS-S	0.00	0.00	0.00	11
micro avg	0.33	0.29	0.31	59
macro avg	0.44	0.17	0.18	59
weighted avg	0.40	0.29	0.27	59

(c) Hochzeit wider Willen

	prec.	rec.	f1	supp.
S-S	0.20	0.10	0.13	42
S-NS	0.00	0.00	0.00	1
NS-S	0.00	0.00	0.00	1
micro avg	0.19	0.09	0.13	44
macro avg	0.07	0.03	0.04	44
weighted avg	0.19	0.09	0.12	44

(d) Bomben für Dortmund

	prec.	rec.	f1	supp.
S-S	0.43	0.19	0.26	219
S-NS	0.00	0.00	0.00	0
NS-S	0.00	0.00	0.00	0
micro avg	0.42	0.19	0.26	219
macro avg	0.14	0.06	0.09	219
weighted avg	0.43	0.19	0.26	219

(e) Aus guter Familie

	prec.	rec.	f1	supp.
S-S	0.39	0.31	0.34	152
S-NS	0.00	0.00	0.00	33
NS-S	0.00	0.00	0.00	33
micro avg	0.36	0.22	0.27	218
macro avg	0.13	0.10	0.11	218
weighted avg	0.27	0.22	0.24	218

(f) Effi Briest

Table 8: Evaluation of the winner system from Track 2 (Gombert, 2021) on all texts in the test set and all border types (`SCENE-TO-SCENE`, `SCENE-TO-NONSCENE`, `NONSCENE-TO-SCENE`).

C Detailed Manual Analysis

C.1 Time markers

As a starting point for the error analysis, the actual markers for scene changes, as known from the guidelines (Gius et al., 2021), were considered separately: changes of narrated time, place, action and character constellation. Thereby, an overgeneralisation of time markers could be detected in the output of the winner system. It is noticeable that many annotated scenes start with formulations like "as", "it was five over", "at this moment", "three minutes elapsed", indicating changes in the time of the narrative. Especially many falsely annotated scene changes (false positives) begin with the temporal indicator "as". The following passage shows an example of a wrong scene change indication triggered by the temporal conjunction "as", marking a change in the narrated time. According to the gold standard, this passage does not include a scene change.

> 'If there really was something to the call, the colleagues in the radio patrol car might still be able to catch the man who had buzzed me out of my sleep. I got up and went to take a shower. A cold one would have been best now. But I wasn't brave enough to do that yet. It was five over. Fat Peter Steiner, the owner of the bar Steinkrug, had by all appearances put not only rat poison but also a strong sleeping pill in the grain. **As** I got dressed and was about to leave the apartment, the phone rang again. 'Mattek?' 'Speaking.' 'Did you get the message through to the alarm center?' 'Yes." (German original text in Figure 1)

In this example, the short reflective passage containing the first-person-narrator's thoughts about the night before interrupts the narrated action, which is resumed with the words "**As** I got dressed and was about to leave [...]". The temporal conjunction "as" could have caused the system to indicate a scene change, whereas according to the gold standard there is no scene change. This indication of a scene change may have resulted from overgeneralisation of the system. The use of temporal markers as indicators of probable scene changes is often successful, but risks an over-sensitive system.

However, not only temporal conjunctions seem to trigger the system to indicate scene changes, but also multi word expressions containing information on the narrated time, as can be seen in the following example, again from *Bomben für Dortmund*, in which a new scene was indicated differently to the gold standard annotation. Looking at this example, the question of granularity arises that will be encountered in the later subsection C.6.

> 'There was no need to hurry. Regarding the station, we had everything under control. Sure, there were loopholes to escape, but someone who didn't even suspect being expected had no reason to look for them and use them. I simply assumed that Jutta Speißer didn't have the faintest idea that we knew practically everything about her. **Two, maybe three minutes passed.** 'Can you hear me, Hermann?' I had hidden the walkie-talkie under my leather jacket so that I could talk into it if I lowered my head a little. 'Yes." (German original text in Figure 2)

Nevertheless, there are also many passages containing temporal markers that the system correctly indicated as new scenes. Another example from *Bomben für Dortmund* shows how it detected the scene change without requiring the temporal marker to be at the beginning of the sentence.

> "Maybe,' I said, '[...]. One devilish lady, one big bastard, and the third bomb we know about. We're going back to the station. Lampert and Blechmann will report there.' The feeling of being watched faded **when she left the train** in Brackel and the long, skinny man who had caught her attention on the train was no longer behind her. But she had quickly calmed down.' (German original text in Figure 3)

Although this correct detection of the scene change could also be related to the simultaneous occurrence of a change of the space of action, which will be discussed in more detail in the next section.

C.2 Change of action space

Another possible overgeneralisation of the system could be its hypersensitivity to descriptions of the action space, since many scene changes annotated by the system happen to be accompanied by references to changes in the action space at the beginning of a new scene.

The following passage is an example from *Bomben für Dortmund* of a correctly annotated scene change followed by an indication of a change in the action space.

> 'I nodded to DAB. 'Give me your walkie-talkie, DAB. Get one from another officer.' He didn't expect anything from it, it was clear from his face, but he gave me his walkie-talkie. **I disappeared from track one and walked through the underpass to the stairs leading up to track three.** There was no sign of Tin Man. Nor was there any sign of the person he had described. That meant they must already be upstairs. I stopped in the middle of the stairs, lit a fresh cigarette and waited.' (German original text in Figure 4)

In addition to many true positive scene changes that the system recognises as in the previous sample passage, there are also many false positives that can be interpreted as the result of the system's overgeneralisation. One example can be found in the following sample passage, in which the main characters do not move but an action outside of the scene setting is described that probably triggered the annotation of a wrong scene change within that scene. There is no scene change according to the gold standard.

> "You'd make a great cop chick,' I said, 'I used to be in the Scouts.' **Outside, in the small reception hall,** someone pounded on the bell as I did. 'I don't have time now, have to take care of the guests and sell rooms, or I'll be out of a job. At nine?' 'You bet!'" (German original text in Figure 5)

Since the system generally tends towards fine-grained scene segmentation, it is not surprising that it often annotates scene changes too much in addition to some actual scene changes. The following passage shows an example of fine-grained scene annotation by the system. In the passage, the main characters move from the hotel reception to the kitchen in the next room. For the manual annotation process, this change of action space is a prototypical example of the application of the container principle defined in the annotation guidelines (Gius et al., 2021, 4). This principle is used to summarise short scenes without clear scene change indicators, e.g., when the characters remain the same and the change from one action space to another is described, while the settings are close to each other and often in the same building, as it is the case in this sample passage. Nonetheless, this scene change could be reasonable with the goal of more finely granulated scene annotation. These considerations have inspired us to look more closely at the distinction between weak and strong boundaries, which we analyse in subsection 4.5.

> "Free choice. You won first prize with me.' 'What would the second have been?' 'A washing machine.' 'I'd rather have the first, to be honest. I finish at eleven.' 'Then the choice of fine venues is very limited.' 'Pull strings,' she said. I followed her into the small, white-tiled kitchen, where breakfast was also made for the guests. The sight of her made me look forward to the evening.' (German original text in Figure 7)

Another recurring phenomenon that often triggers a change of scene is a character entering or exiting a scene. As in the following example from *Bomben für Dortmund* that contains a collection of typical verbal phrases, the exiting and reentering of a character is introduced by the indication of a character's movement from one to another location by the phrases 'to leave', 'to go back to', 'to turn into', and 'to disappear into'. However, according to the gold standard, the scene change should be displayed before the beginning of the sentence 'It was still raining cats and dogs' to mark the beginning of the new scene outside the restaurant. Probably due to oversensitivity, the winning system annotated two scene changes instead of only one as in the gold standard, also missing the actual position of the scene change.

> 'She flushed the toilet as a cover, **left** the cabin and washed her hands. Then, in front of the large, clean mirror, she fixed her frayed hair, which had nothing to be fixed. Then she **went back to** the restaurant. She drank the rest of the ouzo left in the glass, smiled at Dimitri, the owner, secretly wished him all known and unknown venereal diseases, preferably all at once, left an appropriate tip and **left** the restaurant. It was still raining cats and

dogs. In the reflection of some lanterns, the rain looked like many cords next to each other, which did not tear and did not come to an end. It was just before seven when she **turned into** Karl Marx Street, crossed it and **disappeared into** Rubel Street. Out of the street stood the green Sierra.' (German original text in Figure 6)

As has become clear in this subsection, characters and their entrances and exits play a significant role in automatic annotation as markers of a likely scene change. In the following subsection, we will discuss another phenomenon related to characters that often coincides with scene change annotations, namely changes in character constellation.

C.3 Change in Character Constellation

Another marker that frequently occurs at the beginning of automatically detected scenes is the introduction of a new character with the respective full name as well as the accompanying description of the character, its state or an action (presented as a combination of full name plus verb sequence). It can be concluded that the system has learned that this combination occurs frequently at scene beginnings. However, the following two examples (German original text in the appendices 8 and 9) show that this is not always the case.

The first passage is an example for the correct detection by the winner system of a new scene beginning with an introduction of a new character from *Bomben für Dortmund*.

'At nine I had an appointment with Marlies. Lohmeyer couldn't ruin it for me. Jutta Speißer ate stifado and drank Cypriot Aphrodite wine. For Dimitri, the owner of the Greek restaurant 'Akropolis' on Karl-Zahn-Straße, she was a new, welcome guest.' (German original text in Figure 8)

The second passage is an example of scene annotation differing from the gold standard, that contains the introduction and description of three new characters.

'Baldwein started the green Sierra. He slowly steered the vehicle past the post office and drove in the direction of Hoher Wall. Although Police Sergeant Werner Okker had not been drinking last night, because of the duties he had to fulfill as now officially to his he looked bad. He was sitting at the counter of the Steinkrug. His angular, broad shoulders slumped forward in a tired manner. He seemed to be visibly struggling to lift his beer glass. Susanne Steiner stood behind the bar. Large, coarse-boned, Nordic. A girl who had grown up in the pub milieu. She had long, brunette hair and a decidedly beautiful face with full, sensual lips. Peter Steiner, her father, who was standing next to her at the tap, was not at all like her. He was around sixty. A former tusker.' (German original text in Figure 9).

According to the gold standard, there is only one scene change in the text before „Although Police Sergeant Werner Okker [...]", which was also detected by the automatic system. In addition to this, however, another scene change was indicated at the introduction of the new character Peter Steiner. It is noticeable that the constructions around the introduction of the character Susanne Steiner and the character Peter Steiner are similar in structure, but the sentence introducing Susanne Steiner was not recognized as the beginning of a new scene.

Another example of an incorrectly marked new scene which coincides with the introduction of a new character can be found in *Hochzeit wider Willen*. According to the gold standard, there is no scene change in the following passage.

'It was a warm morning at the beginning of August, the sun was shining golden in the breakfast room of the town palace. Here the Hohenstein family had gathered for the first meal of the day. Fürst Heinrich, head of the family and chairman of the Hohenstein Bank, a traditional house in the Frankfurt financial center, was talking lively with his elder son Bernhard.' (German original text in Figure 10)

Since similar constructions occur in the text *Hochzeit wider Willen* and can be found at the beginning of scenes detected by the system (like in Figure 10), it could be determined that this is not a singular phenomenon that occurs specifically in the text *Bomben für Dortmund*.

C.4 Dialogue passages

It is also noticeable that the end of an automatically detected scene is often accompanied by the end of a dialogue passage, which is then followed by a descriptive passage that represents the beginning of a new scene.

The following example from *Hochzeit wider Willen* shows a passage which the winner system segmented into four different scenes, indicating a scene change after every ending of a dialogue passage followed by a descriptive passage without any dialogues. However, according to the gold standard, there is only one scene change in the passage before „Prince Frederik appeared in his office a little later than usual that morning“.

> 'Frederik gazed pensively into his coffee. **'Well, someday I'll get myself a lovely wife and a few offspring, but I still have a bit of a reprieve. Let's say ten to fifteen years ...' 'You've got a lot of nerve.'** The princess laughed and stood up. **'You don't really believe that.' 'Oh yes I do,'** he murmured and smiled narrowly. **'I know that.'** Prince Frederik appeared in his office a little later than usual that morning. Carina Böttiger, his secretary, was used to this and also knew what state her boss was in on such days. The petite blonde with the sky-blue eyes had strong coffee and aspirin ready. She brought both together with the signature folder. [...]. **'You look lovely today,'** Frederik noted, glancing at her dress. He eyed her rather thoughtfully for a moment, and she pretended not to notice, just thanking him artfully for the compliment and asking if there was anything else she could do for him. **'No, that was all for the moment.'** He gave her back the signature folder. **'When Herr von Solm comes, send him right through. I have something else to discuss with him.'** He noticed her slightly s üffisant look, so he clarified: **'Something business-related.'** Carina laughed slightly and **left** the executive room. The fact that Frederik had noticed her new dress made her happy. Until now, she had always believed that he hardly had an eye for her. But she didn't want to get any ideas about that either. After all, it seemed clear that this man was out of her reach. And she was really too good for a brief fling with the ladies' man.' (German original text in Figure 12).

One possible interpretation of this regular annotation of a scene change as a separation of dialogue and descriptive passages could be that the system recognises these passages as different writing styles, leaving the actual reasons for scene changes unconsidered.

C.5 Scene Length

However, the most common error, which was also the easiest to spot, was in the output of scenes that are only one to three sentences long as in the example from *Hochzeit wider Willen*

> 'One could see from the mother's face that this was not necessarily the case. But Hedwig sensed that she would not receive any more information from Carina. 'My little princess ...' That was what she had called Carina as a child. None of them could have imagined that she would ever become a real princess. And if the young woman was honest, she still couldn't quite believe it now. A little later, the bride and groom left for the airport. Ewald Böttiger asked his wife: Ẃhat did you have to talk about for so long? Everyone was waiting for you'. 'I'm not sure if Carina married the right guy[...].' (German original text in Figure 11)

In the manual scene annotation following the guidelines by Gius et al. (2021), the decision was made to append very short scenic passages to the appropriate preceding or following scene in the sense of the container principle. In the given example, however, there is no scene change at all, because it is only a description of the exit of the characters, which takes place within the scene at the bride's parents' house.

C.6 Granularity of Scenes

With respect to the length of the individual passages that should be detected as scenes, there is also the question of how granular the segmentation into scenes should be without becoming too small-scale. The following passage from *Hochzeit wider Willen*

is an example of a small-scale, granular scene segmentation choice by the winning system, in which three scenes were indicated while there are only two according to the gold standard.

> ’Prince Frederik was quite pleased with himself. Carina had swallowed his excuse whole. She had thus given him a free pass, so to speak, to finally go back to living the way he liked. And he was determined to do so immediately ... The very next evening, Frederik called his wife to let her know that it was getting late. He was supposedly waiting for the conclusion of a lucrative business deal. Carina did not suspect anything - yet. When she asked him the next morning when he had come home, he did not tell her the truth.’
> (German original text in Figure 13)

> ’Prince Frederik was quite pleased with himself. Carina had swallowed his excuse whole. She had thus given him a free pass, so to speak, to finally go back to living the way he liked. And he was determined to do so immediately ... The very next evening, Frederik called his wife to let her know that it was getting late. He was supposedly waiting for the conclusion of a lucrative business deal. Carina did not suspect anything - yet. When she asked him the next morning when he had come home, he did not tell her the truth.’
> (German original text in Figure 14)

In this text passage, the choice of the automatic system to recognize another scene is not an implausible one. On the contrary, the system’s decision can be justified, but the small-scale granularity of scene annotation should be avoided in view of the overall goal of the segmentation task, in which a text is to be segmented into units of meaning in terms of content, which should exceed a minimum token length for their further use. The system was more precise than the gold standard.

C.7 German original text of the sample passages

’Falls an dem Anruf wirklich etwas dran war, konnten die Kollegen im Funkstreifenwagen den Mann vielleicht noch stellen, der mich aus dem Schlaf gebimmelt hatte. Ich stand auf und ging unter die Dusche. Eine kalte wäre jetzt am besten gewesen. Aber dazu war ich noch nicht mutig genug. Es war fünf vorbei. Der fette Peter Steiner, der Wirt vom Steinkrug, hatte allem Anschein nach nicht nur Rattengift, sondern auch ein starkes Schlafmittel in den Korn gepanscht. **Als** ich mich angezogen hatte und die Wohnung verlassen wollte, läutete das Telefon erneut. ’Mattek?’ ’Am Apparat’. ’Haben Sie die Meldung an die Alarmzentrale durchgegeben?’ ’Ja.”

Figure 1: Example from *Bomben für Dortmund* of a wrong scene change indication triggered by the temporal conjunction ’als’ marking a change in the narrated time.

’Es bestand kein Grund zur Eile. Was den Bahnhof anging, so hatten wir alles unter Kontrolle. Sicher gab es Schlupflöcher zum Entkommen, aber jemand, der nicht einmal ahnte, dass er erwartet wurde, hatte auch keinen Grund, danach zu suchen und sie zu benutzen. Ich ging einfach davon aus, dass Jutta Speißer nicht den blassesten Schimmer davon hatte, dass wir praktisch alles über sie wussten. **Zwei, vielleicht drei Minuten verstrichen.** ’Kannst du mich hören, Hermann?’ Ich hatte das Walkie-talkie so unter der Lederjacke verborgen, dass ich hineinsprechen konnte, wenn ich den Kopf etwas senkte. ’Ja.”

Figure 2: Example from *Bomben für Dortmund* of an annotated scene change with a temporal marker that deviates from the gold standard. Its legitimacy could be a matter of granularity.

”Vielleicht’, sagte ich. ’[...]. Eine teuflische Lady, einen großen Schweinehund und die dritte Bombe, von der wir wissen. Wir fahren ins Revier zurück. Lampert und Blechmann werden sich dort melden.’ Das Gefühl, beobachtet zu werden, schwand, **als sie in Brackel den Zug verließ** und der lange, dürre Mann nicht mehr hinter ihr war, auf den sie im Zug aufmerksam geworden war. Aber sie hatte sich schnell wieder beruhigt.’

Figure 3: Example from *Bomben für Dortmund* of a correctly indicated scene change probably caused by the temporal marker ’als’.

’Ich nickte DAB zu. ’Gib mir dein Walkie-talkie, DAB. Hol dir eins von einem anderen Beamten.’ Er

versprach sich nichts davon, das war ihm deutlich anzusehen, aber er gab mir sein Walkie-talkie. **Ich verschwand von Gleis eins und lief durch die Unterführung bis zur Treppe, die nach Gleis drei hinaufführte.** Von Blechmann war nichts zu sehen. Von der Person, die er beschrieben hatte, ebenfalls nicht. Das hieß, sie mussten schon oben sein. Ich blieb mitten auf der Treppe stehen, zündete mir frische Zigarette an und wartete.

Figure 4: Correctly annotated scene change following an indication of a change of action space from *Bomben für Dortmund*.

”Du wärst eine prima Polizistenbraut’, sagte ich. ’Ich war mal bei den Pfadfindern.’ **Draußen, in der kleinen Empfangshalle,** hämmerte jemand auf die Glocke, wie ich es getan hatte. ’Ich habe jetzt keine Zeit mehr, muss mich um die Gäste und Zimmer verkaufen, sonst bin ich meinen Job los. Um neun?’ ’Worauf du dich verlassen kannst!”

Figure 5: Incorrectly annotated scene change following a location description from *Bomben für Dortmund*.

’Sie spülte zur Tarnung, verließ die Kabine und wusch sich die Hände. Anschließend richtete sie sich vor dem großen, sauberen Spiegel die ausgefransten Haare, an denen es nichts zu richten gab. Dann ging sie ins Restaurant Sie trank den Rest Ouzo, der sich noch im Glas befand, lächelte Dimitri, den Besitzer, an, wünschte ihm insgeheim alle bekannten und unbekannten Geschlechtskrankheiten, am liebsten auf einmal, ein angemessenes Trinkgeld liegen und verließ das Restaurant. Es regnete noch immer in Strömen. Im Widerschein einiger Laternen sah der Regen aus wie viele sich nebeneinanderbefindliche Bindfäden, die nicht rissen und kein Ende nahmen. Es war kurz vor sieben, als sie in die Karl-Marx-Straße einbog, sie kreuzte und in der Rubelstraße verschwand. Ausgangs stand der Grüne Sierra.’

Figure 6: Example from *Bomben für Dortmund* of a wrong scene change triggered by markers that could indicate that a character is leaving the plot space, which in this case is not the case.

’Freie Auswahl. Du hast mit mir den ersten Preis gewonnen.’ ’Was wäre der zweite gewesen?’ ’Eine Waschmaschine.’ ’Der erste ist mir, ehrlich gesagt, lieber. Ich mache um elf Schluss.’ ’Dann ist die Auswahl der feinen Lokalitäten sehr begrenzt.’ ’Lass deine Beziehungen spielen’, sagte sie. Ich folgte ihr in die kleine, weißgekachelte Küche, in der auch das Frühstück die Gäste gemacht wurde. Ihr Anblick ließ mich auf den Abend hoffen.’

Figure 7: Example from *Bomben für Dortmund* of a scene that begins with characters changing action space, where this change was handled according to the container principle defined in the annotation guidelines (Gius et al., 2021), namely not indicating a new scene, but annotated as a new scene by the winner system. Nonetheless, this scene change could be reasonable with the goal of more finely granulated scene annotation.

’Um neun hatte ich eine Verabredung mit Marlies. Die konnte Lohmeyer mir nicht kaputtmachen. Jutta Speißer aß Stifado und trank zypriotischen Aphrodite-Wein. Für Dimitri, den Besitzer des griechischen Restaurants ’Akropolis’ in der Karl-Zahn-Straße, war sie ein neuer, willkommener Gast.’

Figure 8: Example from *Bomben für Dortmund* for the correct detection of a new scene beginning with an introduction of a new character.

’Baldwein startete den grünen Sierra. Langsam lenkte er das Fahrzeug am Postgiroamt vorbei und fuhr in Richtung Hoher Wall. Obgleich Polizeimeister Werner Okker gestern Nacht nicht getrunken hatte, wegen der Pflichten, die er als nun offiziell Verlobter seiner Verlobten gegenüber zu erfüllen hatte, sah er schlecht aus. Er saß am Tresen vom Steinkrug. Die eckigen, breiten Schultern waren müde nach vorn abgefallen. Es schien ihm sichtlich Mühe zu bereiten, sein Bierglas zu heben. Susanne Steiner stand hinter der Theke. Groß, grobknochig, nordisch. Ein Mädchen, das im Kneipenmilieu großgeworden war. Sie hatte langes, brünettes Haar und ein ausgesprochen schönes Gesicht mit vollen, sinnlichen Lippen. Peter Steiner, ihr Vater, der neben ihr am Zapfhahn stand, war ihr überhaupt nicht ähnlich. Er war um die Sechzig herum. Ein ehemaliger Hauer.’

Figure 9: Example of an incorrect scene change indication from *Bomben für Dortmund*.

’Es war ein warmer Morgen Anfang August, die Sonne schien golden in das Frühstückszimmer des Stadtpalais. Hier hatte sich die Fürstenfamilie Hohenstein zur ersten gemeinsamen Mahlzeit des Ta-

ges versammelt. Fürst Heinrich, Familienoberhaupt und Vorstand der Hohenstein-Bank, eines traditionsreichen Hauses am Frankfurter Finanzplatz, unterhielt sich angeregt mit seinem älteren Sohn Bernhard.'

Figure 10: Example of an introduction of a new character from *Hochzeit wider Willen* incorrectly marked as a new scene (there is no scene change in the passage according to the gold standard).

'Man sah der Mutter an, dass dies nicht unbedingt der Fall war. Doch Hedwig spürte, sie würde von Carina keine weiteren Auskünfte erhalten. [...] 'Meine kleine Prinzessin So hatte sie Carina als Kind genannt. Keiner von ihnen hätte sich wohl vorstellen können, dass sie jemals eine wirkliche Prinzessin werden würde. Und wenn die junge Frau ehrlich war, konnte sie es jetzt noch immer nicht so ganz fassen. Wenig später fuhr das Brautpaar zum Flughafen. Ewald Böttiger fragte seine Frau: 'Was hattet ihr denn noch so lange zu bereden? Alle haben auf euch gewartet.' 'Ich bin mir nicht sicher, ob Carina den Richtigen geheiratet hat."

Figure 11: Example of an one-sentence scene from *Hochzeit wider Willen*.

'Frederik blickte sinnend in seinen Kaffee. **'Na ja, irgendwann werde ich mir eben ein liebes Frauchen und ein paar Sprösslinge zulegen, aber ein bisschen Galgenfrist bleibt mir ja noch. Sagen wir mal zehn bis fünfzehn Jahre 'Du hast Nerven.'** Die Prinzessin musste lachen und erhob sich. **'Das glaubst du doch wohl nicht im Ernst.' 'Oh doch'**, murmelte er und lächelte schmal. **'Das weiß ich.'** Prinz Frederik erschien **an diesem Morgen** etwas später als sonst in seinem Büro. Carina Böttiger, seine Sekretärin, war das gewohnt und wusste auch, in welchem Zustand ihr Chef an solchen Tagen war. Die zierliche Blondine mit den himmelblauen Augen hielt starken Kaffee und Aspirin bereit. Beides brachte sie zusammen mit der Unterschriftenmappe. [...] **'Sie sehen heute hübsch aus'** , stellte Frederik mit einem Blick auf ihr Kleid fest. Er musterte sie einen Moment lang ziemlich nachdenklich, und sie tat so, als merke sie es gar nicht, bedankte sich nur artig für das Kompliment und fragte, ob sie sonst noch etwas für ihn tun könne. **'Nein, das war im Moment alles.'** Er gab ihr die Unterschriftenmappe zurück. **'Wenn Herr von Solm kommt, schicken Sie ihn gleich durch. Ich habe noch etwas mit ihm zu besprechen.'** Er bemerkte ihren leicht süffisanten Blick und stellte deshalb klar: **'Etwas Geschäftliches.'** Carina lächelte leicht und **verließ** das Chefzimmer. Dass Frederik ihr neues Kleid bemerkt hatte, machte sie glücklich. Bislang hatte sie immer geglaubt, dass er kaum einen Blick für sie hatte. Doch sie wollte sich darauf auch nichts einbilden. Schließlich schien es klar, dass dieser Mann außerhalb ihrer Reichweite war. Und für eine kurze Affäre mit dem Frauenliebling war sie sich wirklich zu schade.'

Figure 12: Example from *Hochzeit wider Willen* regarding the separate annotation of dialogue and descriptive passages by the system.

'Prinz Frederik war ganz zufrieden mit sich selbst. Carina hatte seine Ausrede glatt geschluckt. Damit hatte sie ihm sozusagen selbst den Freifahrtschein ausgestellt, um endlich wieder so zu leben, wie es ihm gefiel. Und er war fest entschlossen, dies auch umgehend zu tun **Bereits am nächsten Abend** meldete Frederik sich telefonisch bei seiner Frau und ließ sie wissen, dass es spät wurde. Angeblich wartete er auf den Abschluss eines lukrativen Geschäfts. Carina schöpfte noch keinen Verdacht. **Als sie ihn am nächsten Morgen** fragte, wann er heimgekommen sei, sagte er ihr nicht die Wahrheit.'

Figure 13: Granularity example from *Hochzeit wider Willen* of three automatically detected scenes differing to gold standard.

'Prinz Frederik war ganz zufrieden mit sich selbst. Carina hatte seine Ausrede glatt geschluckt. Damit hatte sie ihm sozusagen selbst den Freifahrtschein ausgestellt, um endlich wieder so zu leben, wie es ihm gefiel. Und er war fest entschlossen, dies auch umgehend zu tun . . . **Bereits am nächsten Abend** meldete Frederik sich telefonisch bei seiner Frau und ließ sie wissen, dass es spät wurde. Angeblich wartete er auf den Abschluss eines lukrativen Geschäfts. Carina schöpfte noch keinen Verdacht. **Als sie ihn am nächsten Morgen** fragte, wann er heimgekommen sei, sagte er ihr nicht die Wahrheit.'

Figure 14: Granularity example from *Hochzeit wider Willen* of two manually detected scenes in the gold standard.