Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3346–3353
Marseille, 20-25 June 2022

Exploring Text Recombination for Automatic Narrative Level Detection

Nils Reiter♠, Judith Sieker◊, Svenja Guhr♣, Evelyn Gius♣, Sina ZarrieSS◊
♣Technical University of Darmstadt, Residenzschloss, Marktplatz 15, 64283 Darmstadt
♠University of Cologne, Albertus-Magnus-Platz, 50931 Cologne, nils.reiter@uni-koeln.de
◊University of Bielefeld, Universitätstr. 25, 33615 Bielefeld, sina.zarriess@uni-bielefeld.de

Abstract

Automatizing the process of understanding the global narrative structure of long texts and stories is still a major challenge for state-of-the-art natural language understanding systems, particularly because annotated data is scarce and existing annotation workflows do not scale well to the annotation of complex narrative phenomena. In this work, we focus on the identification of narrative levels in texts corresponding to stories that can be embedded (stories within stories) or otherwise coordinated within narratives. Lacking sufficient pre-annotated training data, we explore a solution to deal with data scarcity that is common in machine learning: the automatic augmentation of an existing small data set of annotated samples with the help of data synthesis. We present a workflow for narrative level detection, that includes the operationalization of the task, a model and a data augmentation protocol for automatically generating narrative texts annotated with breaks between narrative levels. Our experiments suggest that narrative levels in long text constitute a challenging phenomenon for state-of-the-art NLP models, but generating training data synthetically does improve the prediction results considerably.

Keywords: computational literary studies, narrative levels, training data induction

1. Introduction

The availability of annotated resources is still a major bottleneck for natural language processing research that is concerned with the analysis of long texts, such as book-length stories or narratives. To comprehend stories, one must not only be able to recognize elements at the phrase or sentence level, e.g., entities, characters, direct speech, but also global structural elements, such as plot arcs or narrative levels. The latter (to be introduced more thoroughly below) commonly represents stories within a story, e.g., told by a character. The analysis of stories in terms of these structures is essential in research areas ranging from journalistic writing to storytelling in social media to interviews in the social sciences and literary fiction. While corpora of raw texts covering these domains and types of narratives are available, annotated corpora of narratives are small and cover only some of the relevant phenomena: While the ProppLearner corpus by Finlayson (2015) contains annotations of plot elements in folktales, Bamman et al. (2020) have published a data set with annotated coreference chains on 2000-word samples of stories. Both corpora are potentially helpful for processing long narrative texts, but there is no way to combine them in a fruitful way. Hence, automatizing the process of understanding the global narrative structure of long texts/stories is still a major challenge for state-of-the-art natural language understanding systems, particularly because annotated data is scarce.

This scarcity of annotated data, however, is not easy to fix, because the common annotation workflow (hire annotators, ask them to annotate according to guidelines) scales very badly to longer texts: i) Annotating longer texts makes the reading experience much more important, as the annotation density (i.e., raw number of annotations per token) is much lower compared to established annotations tasks in NLP. However, annotation tools are usually desktop or web applications that are not optimized for pleasurable reading, which will impact annotation quality negatively. ii) Even if this problem would be overcome, actually achieving inter-subjective annotations is still difficult, because annotations of narrative structures depend on reading concentration, memory and attention. While this is certainly true for all annotation tasks, linguistic tasks such as syntactic annotation allow a ‘mental reset’ after each sentence. Annotating narrative levels requires attention over hundreds of pages of texts. Even reading the text in question can often not done in a single session, but is spread over days. iii) In addition, popular texts and plots may already be known to annotators, possibly even unconsciously. Thus, it will be impossible to control whether annotators annotate purely based on the text or mix in their memory of a TV show they saw years ago in which one sub plot was similar or inspired on the current text. We therefore argue that producing annotations for longer texts is not just a question of funding and/or motivation, but that serious conceptual challenges need to be overcome.

As an alternative, we therefore explore a common solution to deal with data scarcity in machine learning: the automatic increase of an existing small data set of annotated samples. Image data for training object recognition systems in computer vision, for instance, is very commonly augmented through simple automatic techniques for cropping, tilting and transforming a given set of labeled images and pairing the resulting, manipulated images with the label of their original (Howard, 2013; Szegedy et al., 2015). Data augmentation has also been explored for some standard NLP tasks on sentences or short texts (Wang et al., 2018; Wei and Zou, 2019; Liu et al., 2020), where the augmentation technique typically manipulates a given word sequence on the token level, e.g., by deleting or swapping tokens, or back-translating the sequence. To the best of our knowledge, data augmentation has not yet been explored for tackling data scarcity issues in tasks dealing with longer texts or narratives.

We will first give some background on modeling narratives in general, narrative levels in particular as well as recent attempts at annotating them, and discuss data augmentation techniques in Section 2. We then describe our approach to construct a workflow for narrative level detection, including the operationalization of the task, the model and the data synthesis Section 3. Section 4 describes our experiments and Section 5 discusses our findings.

2. Background

2.1. Modeling narrative structure

Structural properties of narrative texts have received little, but continuous attention in recent years: Piper et al. (2021) provide a comprehensive overview, geared towards computational linguists. Focusing on longer discourses, Ouyang and McKeown (2014) have proposed a system to detect discourse relations as in the Penn Discourse Treebank in narratives and Papalampidi et al. (2020) have published work on the summarization of screenplays using narrative structures.

Narrative segments are discussed by Reiter (2015) from an annotation perspective, and a system for the automatic detection of narrative segments, in this case defined as narrative scenes, is described by Zehe et al. (2021), using German dime novels as a corpus. The system achieves an F1-score of 0.24. Chapters, another segmentation criterion, are targeted by Pethe et al. (2020). They employ a BERT-based model and achieve an F1-score of 0.453 on full-length English novels. These two criteria for segmentation differ strongly: a scene, following the definition by Zehe et al. (2021), is a textual segment with a continuity of narrated time, place, action and character constellation, comparable to a movie scene. Chapters are not necessarily homogeneous with respect to the plot: They are not units of content, but are influenced by publishing and stylistic preferences, and used for organizing a text. Most strikingly, cliffhangers are a commonly appearing stylistic device that separate the plot, often used at chapter boundaries. Such a cliffhanger would probably not be annotated as a scene boundary.

2.2. Narrative Levels

Narrative levels are a third way of segmenting a text and a well documented phenomenon in the scholarly occupation with narratives (Bal, 1997, 43 ff.). The most crucial criterion for a narrative level is that a level forms a story on its own – most often, this is a story told by a character of another story. Thus, narrative levels do not form a flat segmentation as scenes and chapters, but can be nested and thus hierarchically organized.

> [. . .] "With joy and goodly gree," answered Shahrazad, "if this pious and auspicious King permit me." "Tell on," quoth the King who chanced to be sleepless and restless and therefore was pleased with the prospect of hearing her story. So Shahrazad rejoiced; and thus, on the first night of the Thousand Nights and a Night, she began with the
>
> ‡**Tale of The Trader and the Jinni.**
> It is related, O auspicious King, that there was a merchant of the merchants who had much wealth, and business in various cities. [. . .]

Figure 1: Excerpt of Arabian Nights, showing the beginning of an embedded narrative (‡). Translation by Richard Francis Burton, published in 1885 and available via archive.org.

> "Off with her head!" the Queen shouted at the top of her voice. Nobody moved.
> "Who cares for you?" said Alice, (she had grown to her full size by this time.) "You're nothing but a pack of cards!"
> At this the whole pack rose up into the air, and came flying down upon her: she gave a little scream, half of fright and half of anger, and tried to beat them off, ‡ and found herself lying on the bank, with her head in the lap of her sister, who was gently brushing away some dead leaves that had fluttered down from the trees upon her face. "Wake up, Alice dear!" said her sister; "Why, what a long sleep you've had!" "Oh, I've had such a curious dream!" said Alice [. . .].

Figure 2: Passage from Carroll's *Alice in Wonderland*, Chapter XII, showing a narrative level change (‡)

One of the most well known examples is the book *Arabian Nights* (often also referred to as *One Thousand and One Nights*, originally written during the Islamic Golden Age), in which the female protagonist Shahrazad tells a story every night to avoid being executed by her husband. Because he is interested in the continuation of each story, he spares her life. Figure 1 shows the beginning of one embedded story as an example. A few characteristics of embedded narratives can be discerned directly: in the frame story in the beginning of the segment, Shahrazad and the king are speaking through direct speech, marked with quotation symbols. The embedded story itself, however, even though clearly narrated by Shahrazad, is not marked with quotation symbols, showing that it has a different status than the utterances before.

Often, embedded stories are narrated by a character of the frame story, as first described by (Genette, 1980).

There are, however, narrative levels that are differentiated in other ways: (Ryan, 1991) introduced ontologically different story worlds as a marker for embedded narratives, often in the context of dreams or otherwise different universes. One such example can be found in Figure 2, showing a segment of the well known story *Alice in Wonderland* by Lewis Carroll. The figure shows a passage with a change from the second narrative level – the embedded story in which Alice finds herself in Wonderland – back to the first narrative level: the frame story in which the narrator tells the story of Alice waking up from her dream. Interestingly, the narrative level change (marked in bold) takes place within a sentence, without any kind of formal or typographic marking.

The complexities of narrative levels thus lie in the fact that they are related to both the *how* of the narration (i.e., the textual representation of the story) as well as the *what* of the narration (i.e., the contents of the story).

2.3. Systematic Analysis of Narrative Texts through Annotation

Given the many different structural and plot-wise ways of initiating level changes, it is not surprising that the annotation of narrative levels is a major challenge even for human annotators. This was demonstrated in the shared task SANTA (Gius et al., 2019; Gius et al., 2021, Systematic Analysis of Narrative Texts through Annotation). This shared task was aimed at developing guidelines for the manual annotation of narrative levels. Participants were asked to select their preferred theoretical basis for narrative levels (e.g., Ryan (1991) or Genette (1980)) and develop guidelines that implement this theoretical framework. The guidelines were then evaluated in an annotation experiment with multiple student annotators who annotated in parallel. This shared task could attract seven participant teams from different disciplines (computational linguistics, linguistics, literary studies, digital humanities). The final version of each of these guidelines as well as a conceptual comparison of the guidelines can be found in Gius et al. (2021).

2.4. Data augmentation

Data augmentation is a widely used technique to boost the performance of data-hungry machine learning methods, which has found some attention in natural language processing (NLP). Typical set-ups for data augmentation in NLP are applications where some initial annotated training data of texts paired with labels or, e.g., translations is given, such that $D_{train} = \{x_i, y_i\}$ (Kumar et al., 2020). Based on D_{train}, common augmentation protocols generate a synthetic version of the data $D_{syn} = \{\hat{x_i}, y_i\}$, where $\hat{x_i}$ is a perturbed version of some original $x_i \in D_{train}$ that carries the same label as the original data point. Procedures for obtaining the perturbed data points range from random word replacement and swapping operations (Wang et al., 2018; Wei and Zou, 2019) to more controlled, grammar-based or linguistically motivated augmentation methods (Jia and Liang, 2016; Andreas, 2020). The general idea is to produce data that is “good enough” (Andreas, 2020) for training a machine learning model despite its obviously lower quality as compared to carefully annotated data sets.

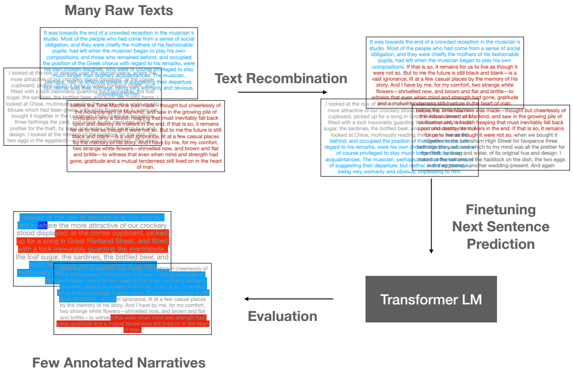

Figure 3: Text recombination for narrative level detection

3. Approach

The approach we want to explore in this paper follows the data augmentation method described above in general, but differs in some details from the previous methods as we are interested in discovering the structure of narratives and need to generate data with text *segmentations* rather than labels. Figure 3 shows a visual representation of our approach, which will be explained textually below.

3.1. Task definition

The task of automatically detecting narrative levels in prose texts can be cast in different ways: content-wise, a narrative level forms a unit on its own. It may have a unique and level-specific story world, which is populated with characters, locations and taking place at a certain time. As such, the task would be a **unitizing task**. In contrast to well known unitizing tasks such as named entity recognition (NER), the units in this case are much longer and can be discontinuous. Automatically detecting an entire narrative level as a unit would require machinery to cope with a potentially long discourse and untangle the level-forming narrative constituents (characters, locations, events, . . .).

As an alternative operationalization, we can operationalize the problem as a **segmentation task**. The goal is then to identify boundaries of narrative levels in the text. For segmentation, a set of candidate boundaries is often extracted, which are then subjected to a binary classification. Since narrative levels can either stand independently side by side or have a hierarchical nature with subordinate and super-ordinate levels, the boundaries must be sub-classified to distinguish the beginning and end of a subordinate or super-ordinate level. As decisions are made on individual boundaries, such a segmentation approach cannot incorporate knowledge about the text as a whole: the fact that, for instance, a character no longer appears in a narrative level cannot be taken into account.

From a pure output perspective, it does not matter which approach was chosen, as annotations of boundaries can be converted into units and vice versa. Technically, however, segmentation approaches are far more common and easier to implement than proper unitizing approaches, and even well known unitizing tasks such as NER are usually implemented as a segmentation problem (using the BIO scheme). We therefore follow the segmentation approach for our experiments.

3.2. Data Set Construction

In order to generate training data automatically, we use the ELTeC-eng (Burnard and Odebrecht, 2021) corpus as source material and selected its 38 shortest Anglophone texts (between 14 002 and 68 607 words long). These were first randomly split into a training and test sub set (70/30). From these, we generate 700 training and 300 test texts as follows: each of the generated texts consists of multiple base texts, whose number has been sampled from a normal distribution with $\mu = 3$ and $\sigma = 1$. The generated texts thus mostly contain between one and five individual stories, although there are some outliers with more stories. As they have nothing in common content-wise, any story but the first can be considered an embedded story.[1]

We first experiment, how well these 'narrative level'-boundaries can be detected automatically and use the synthetic texts to select hyper parameters. With the best performing hyper parameters we then evaluate on the annotations created (and published) in the shared task SANTA Gius et al. (2021), in which 13 texts were annotated according to the seven different guidelines described above. We evaluate on each guideline separately.

3.3. Break Prediction

For modeling narrative level detection as a text segmentation task, we exploit a standard state-of-the-art transformer language model, i.e., BERT (Devlin et al., 2019a). Neural language modeling frameworks can be widely and successfully used to fine-tune specific models for various NLP tasks, but have also proven useful in more analysis-oriented work aimed at evaluating and probing the linguistic representations and processing capacities "emerging" in large-scale language models (Belinkov and Glass, 2019). Many of the existing probing tasks for language models were inspired by theoretical linguistic and psycholinguistic research (Belinkov and Glass, 2019) and are designed to test the model's knowledge of specific syntactic or semantic phenomena as, e.g., agreement (Linzen et al., 2016) or negation (Ettinger, 2020). Here, most studies exploit BERT directly as a language model, i.e., as a predictor of words and word probabilities in context.

Next to masked language modeling on the word level, BERT is also pre-trained on a sentence-level prediction task called next-sentence prediction (NSP, Devlin et al. (2019a)). For NSP, the model is trained on pairs of sentences drawn from text and randomly paired sentences from different texts. Its task is to predict whether the pair is an actually occurring or a random sentence pair. Devlin et al. (2019a) found that pre-training the model on NSP is an important prerequisite for obtaining high performance when fine-tuning on various inference tasks such as question answering or textual entailment. Later work on related transformer architectures has found that NSP pre-training does not improve or even impairs performance on certain downstream tasks, so that this pre-training task is sometimes removed from the training set-up in subsequent variants of the original BERT (Liu et al., 2019).

In this paper, we use the original BERT model with its NSP head to predict narrative level boundaries in text. Unfortunately, the exact data sampling procedure for training NSP is not described in (Devlin et al., 2019a), but we expect that NSP pre-training is similar to our text recombination protocol. Therefore, as a first step, we directly evaluate a pre-trained BERT model for NSP on narrative level prediction. This allows us to test to what extent a narrative level boundary has a left and right context that is clearly different because the story being told has changed. Furthermore, we use our recombined text data to fine-tune BERT's NSP head for the task of detecting breaks in narrative texts. This allows to test to what extent the off-the-shelf language models might need to be fine-tuned to the domain of narratives.

To sum up, in our experiments, we investigate whether (i) an off-the-shelf BERT model is able to detect break points in a narrative without being explicitly tuned to particular annotation categories and (ii) fine-tuning on synthetically recombined narrative text supports the off-the-shelf model to discover breaks between narrative levels in annotated narratives.

4. Experiments

4.1. Evaluation

An evaluation of level detection approaches can be performed on segment boundaries or units. We opt for a segmentation evaluation here, in line with our operationalization. Concretely, we use three metrics: precision and recall are calculated as usual, counting only exactly matching character positions as true positives. A precision of 50 thus indicates that one half of the predicted boundaries are at correct locations. In addition, we employ boundary similarity (Fournier, 2013), which is based on a boundary edit distance and distinguishes the edit operations addition/deletion (for missing or spurious boundaries), transposition (for boundaries that are 'close enough' to be moved) and substitution (for boundaries of the wrong type; not used in our case). We

[1] An obvious caveat at this point is that the transitions from one story to the next are not softened whatsoever. The stories are just concatenated.

FT	Precision	Recall	Boundary sim.
No	$2.61\pm\ 1.8$	55.41 ± 25.2	$2.51\ \pm\ 1.8$
Yes	32.39 ± 36.09	25.68 ± 27.12	$19.20\ \pm24.25$

Table 1: Best evaluation results on artificially created boundaries after hyperparameter selection. Scores are averaged over 300 texts. $\pm$ designates the standard deviation, boundary similarity (Fournier, 2013) is calculated with a transposition window of $n_t = 100$ characters. The table shows results without and with finetuning (FT) on a training data set.

particularly look at the impact of the fine-tuning on the synthetic training data.

4.2. Data and Set-up

For the automatic prediction of narrative level boundaries, we make use of the pre-trained BERT model (Devlin et al., 2019b) in its base uncased variant as provided by the huggingface library[2]. The texts are tokenized using BERT's pre-trained tokenizer, and we add a `CLS`-token to the beginning of the sequence, a `SEP`-token to separate the parts of the sequence, and a `SEP`-token at the end. Each paragraph break is considered a candidate boundary, for which we take a fixed token window to the left and right of the candidate as input. We evaluate the window sizes 54, 154, and 254 tokens in both directions. The shortest window size yields the best results and is used for all subsequent experiments. If there are less tokens available, we pad the input sequences. The output of the pre-trained NSP head is transformed into the probability that this paragraph break is also a narrative level break.

4.3. Results

Figure 4 shows the probabilities for the boundaries of two texts predicted by the model, with the true boundaries marked in red. We can see that enlarging the context window leads to a lower number of predicted boundaries. It seems that positive indicators can be found rather locally, while a larger context offers indicators against a break. We also see a pretty clear distinction between boundaries and non-boundaries: paragraphs before and after a predicted boundary have an almost-zero probability of being a boundary. Finally, we can see the aforementioned effect – that larger context windows lead to fewer boundaries – is not caused by simply predicting a subset of boundaries: some of the boundaries predicted with the 154-token window are not predicted at all with the 54-token window. More systematic evaluation reveals that larger window sizes yield considerable less performance, we therefore focus on the 54-token-windows in all following experiments.
Table 1 shows quantitative evaluation scores on the synthetic test corpus with and without fine-tuning on the synthetic training corpus. As we can see, base performance (without fine-tuning) particularly achieves a low precision. Generally, fine-tuning on our synthetic data set yields a better performance, although neither precision nor recall are totally satisfying.
Table 2 shows the precision and recall scores achieved on the ground truths created in the SANTA shared task (Gius et al., 2021). Since the data set consists of two annotations (by two annotators) for each of the seven different guidelines, we give both results for each guideline.
As expected, the performances show a slightly different pattern on the synthetic and ground truth data sets. Generally, the recall is lower on the latter. Since the test data consists of real narrative level boundaries, the difference in content before and after the boundaries is potentially smaller than in the synthetic data, making this an expected result. We do see, however, that finetuning on synthetically created data has a positive effect on precision for almost all of the guidelines. The gain in precision almost always outweighs the loss in recall. Furthermore, results in Table 2 show that the absolute performance of the narrative level detection and the relative increase achieved by finetuning on synthetically recombined texts depends to a considerable extent both on the underlying guideline and the annotator doing the annotation. For instance, while the performance and the effect of finetuning is stable for both guideline 7 (i.e. precision increases from 7.69 to 17.95), annotations based on guideline 4 show a marked effect of finetuning for only one of the annotators (i.e. precision increases from 12.27 to 33.33). This is in line with our main argument, stated in Section 1, that serious conceptual challenges will need to be overcome to scale existing annotation workflows to global, structural aspects of long, narrative text.

5. Discussion

Our experiments suggest that narrative levels in long text constitute a challenging phenomenon for state-of-the-art NLP models. Thus, despite the fact that BERT's next sentence prediction head is pre-trained on a supposedly similar task (distinguishing random from actually occurring sentence or paragraph pairs), we find that its off-the-shelf performance is very low in our setting, even on synthetically generated narrative breaks. This supports observations made in other work suggesting that BERT's NSP head might not be stable (Liu et al., 2019). Note, however, that (Pethe et al., 2020) achieve satisfactory performance with finetuning BERT's NSP head on the task of detecting chapter boundaries. An obvious direction for future work is to investigate and, potentially, modify the pre-training protocols of recent transformer language models to improve their performance in representing and detecting various elements of narrative texts. This should also involve a more systematic exploration of the length of the narrative text's snippets as we, somewhat counterintuitively, found that

[2] `https://huggingface.co/`

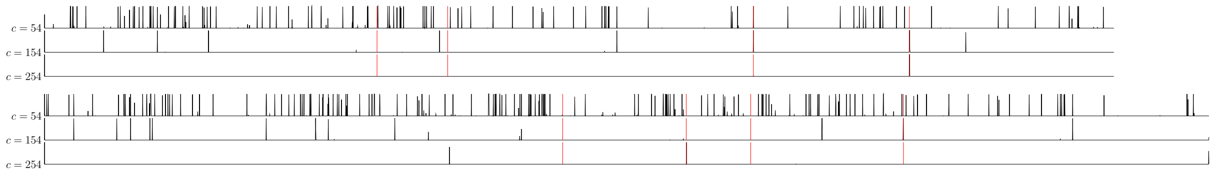

Figure 4: Predicted break probabilities in two randomly selected texts for different context windows. Red lines indicate true boundaries.

	Without finetuning		With finetuning		Gain by finetuning	
Guideline	Precision	Recall	Precision	Recall	Precision	Recall
1	14.36	9.37	25.64	5.89	11.28	−3.48
	14.36	14.27	21.79	7.72	7.44	−6.54
2	11.41	6.75	17.95	4.59	6.54	−2.17
	7.56	4.58	14.10	5.22	6.54	0.64
4	12.27	11.08	33.33	10.76	21.06	−0.32
	7.69	7.13	10.26	2.98	2.56	−4.15
5	12.18	9.79	17.95	7.40	5.77	−2.39
	15.13	9.70	10.26	2.75	−4.87	−6.95
6	15.69	14.43	23.85	14.47	8.15	0.04
	18.36	9.03	21.79	3.90	3.44	−5.13
7	7.69	19.23	17.95	13.46	10.26	−5.77
	7.69	19.23	17.95	12.09	10.26	−7.14
8	8.08	11.54	17.95	9.10	9.87	−2.44
	9.87	13.41	15.38	5.40	5.51	−8.01

Table 2: Prediction results for narrative level boundaries without and with fine-tuning of the BERT model on synthetic data. The predictions are evaluated on the annotations by two different annotators for each guideline.

the most robust model is the one that predicts breaks in narrative levels based on a context window of 54 words to the left and right of the break.

Next to being a challenge for automatic text analysis, narrative levels point to some important limitations of standard resource creation workflows in Digital Humanities. To date, these workflows heavily rely on annotation that is done *post hoc* on complex text by expert annotators and typically involves several cycles of narrowing down the annotation guidelines, training annotators and further steps to achieve a high annotator agreement. The generative workflow we have explored in this paper might offer a complementary method for those phenomena where extensive manual annotation simply is too costly and too difficult to set up in a rigorous way. In our workflow, no post-hoc annotation is needed, but we use an algorithm that produces an artificial, recombined text with annotations of breaks between narrative levels. We believe that this offers an interesting and different way to draw on expert knowledge in an ML pipeline for the Digital Humanities. In our setting, the expert's role is not to formulate her knowledge in terms of annotation decisions or feature design for the detection algorithm, but in terms of text generation rules for data augmentation protocols. Thus, human experts base the identification of narrative levels on properties such as changes of speaker, characters, time and/or ontological space in the fictional world. In future work, we plan to look at formulating these regularities as generation rules leading to more plausible synthetic narratives that the ones we have obtained through simple text recombination in this work.

6. Conclusions

Annotated reference data is both crucial for model development and thus large-scale text analysis, and at the same time very hard to produce in long texts and/or over long textual distances. As this is a common problem, a lot of research in machine learning has explored methods that avoid the need for annotated data in some way, e.g., semi-supervised/unsupervised learning, data augmentation, self training, ... Unsupervised learning has been used in the context of digital humanities applications in the form of standard language modeling pipelines (or word embeddings), but data augmentation has not been systematically explored.

In this paper, we have explored the use of data augmentation and synthesis to circumvent the data scarcity, focusing on the long-text problem of recognizing narrative level detection. Our results show that while narrative level detection is a hard task, generating training data synthetically does improve the prediction results considerably. An obvious next step is a more controlled synthesizing of data: in the current form, we have randomly combined texts with full prose stories, without any kind of transition or embedding. A more natural embedding of a narrative level might yield even better results.

7. Bibliographical References

Andreas, J. (2020). Good-enough compositional data augmentation. In *Proceedings of ACL*.

Bal, M. (1997). *Narratology: Introduction to the Theory of Narrative*. University of Toronto Press, 2 edition.

Bamman, D., Lewke, O., and Mansoor, A. (2020). An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France, May. European Language Resources Association.

Belinkov, Y. and Glass, J. (2019). Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 04.

Burnard, L. and Odebrecht, C. (2021). English novel corpus (eltec-eng) april 2021 release (v1.0.1).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019a). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May. arXiv: 1810.04805.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019b). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ettinger, A. (2020). What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Finlayson, M. A. (2015). ProppLearner: Deeply annotating a corpus of Russian folktales to enable the machine learning of a Russian formalist theory. *Digital Scholarship in the Humanities*, 32(2):284–300, 12.

Fournier, C. (2013). Evaluating text segmentation using boundary edit distance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1702–1712, Sofia, Bulgaria, August. Association for Computational Linguistics.

Genette, G. (1980). *Narrative Discourse – An Essay in Method*. Cornell University Press, Ithaca, New York.

Gius, E., Reiter, N., and Willand, M. (2019). A Shared Task for the Digital Humanities. Special Issue of *Cultural Analytics*, November.

Gius, E., Willand, M., and Reiter, N. (2021). On organizing a shared task for the digital humanities – conclusions and future paths. *Journal of Cultural Analytics*, 6(4), 12.

Howard, A. G. (2013). Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*.

Jia, R. and Liang, P. (2016). Data recombination for neural semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Berlin, Germany, August. Association for Computational Linguistics.

Kumar, V., Choudhary, A., and Cho, E. (2020). Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China, December. Association for Computational Linguistics.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, R., Xu, G., Jia, C., Ma, W., Wang, L., and Vosoughi, S. (2020). Data boost: Text data augmentation through reinforcement learning guided conditional generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9031–9041, Online, November. Association for Computational Linguistics.

Ouyang, J. and McKeown, K. (2014). Towards automatic detection of narrative structure. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4624–4631, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Papalampidi, P., Keller, F., Frermann, L., and Lapata, M. (2020). Screenplay summarization using latent narrative structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1920–1933, Online, July. Association for Computational Linguistics.

Pethe, C., Kim, A., and Skiena, S. (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, November. Association for Computational Linguistics.

Piper, A., So, R. J., and Bamman, D. (2021). Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empiri-*

cal Methods in Natural Language Processing, pages 298–311, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Reiter, N. (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, July. Association for Computational Linguistics.

Ryan, M.-L. (1991). *Possible worlds, artificial intelligence, and narrative theory*. Indiana University Press, Bloomington.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Wang, X., Pham, H., Dai, Z., and Neubig, G. (2018). SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium, October-November. Association for Computational Linguistics.

Wei, J. and Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China, November. Association for Computational Linguistics.

Zehe, A., Konle, L., Dümpelmann, L. K., Gius, E., Hotho, A., Jannidis, F., Kaufmann, L., Krug, M., Puppe, F., Reiter, N., Schreiber, A., and Wiedmer, N. (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, April. Association for Computational Linguistics.