Adapting Transformer Language Models for Application in Computational Creativity: Generating German Theater Plays with Varied Topics

Lukas Wertz, Jonas Kuhn
University of Stuttgart
Institute for Natural Language Processing (IMS)
lukas.wertz@ims.uni-stuttgart.de

Abstract

Pre-trained transformer language models have been shown to generate human-like quality texts of different styles. In this study, we generate short drama dialogues in the style of German theater plays and adapt their content to various different topics using a simple fine-tuning scheme. We show that the generations keep the dramatic play structure while adapting large parts of their content to a target topic, effectively creating scenes from theater plays about a variety of subjects. We experiment with hyperparameters to find fitting fine-tuning configurations for various topic datasets as well as highlight how the generations adapt to the topics in a qualitative analysis. Our findings present a useful tool for computer assisted or fully autonomous creative writing. Furthermore, we motivate and explore the use of transformer language models in the context of computational creativity, highlighting the need for constrained and controlled language generation.

Introduction

This paper reports on a set of pilot experiments that we conducted in preparation of a possible integration of AI generated elements in an actual theater production. The output produced by recent transformer language models such as GPT-2 is often intriguingly natural. Yet when applying such models for language generation in a specific computational creativity context, there are typically additional constraints on the desired model output: in our pilot scenario, the generated text was for instance (i) supposed to follow the structural characteristics of dramatic text; in addition (ii), the text was supposed to revolve around a specific domain content. We argue that such constraints to the application of pre-trained language models are not a peculiarity arising from our application scenario, but reflect a general challenge that an integration of recent model types from Natural Language Processing (NLP) research into a computational creativity scenario faces.

Our preliminary experimental results on adapting transformer language models for creative language generation are thus not only informative for scenarios with a similar constellation of training and tuning resources; by reporting on experience from our pilot study we also hope to make a contribution to an open-ended (and presumably long-term) process of identifying suitable workflows and methodological set-ups for interdisciplinary work in the broader field of computational creativity research.

Motivation and Background

Many scenarios in which the acts of a human (or a group of humans) are commonly described as creative involve language production. Writing a novel, a poem, or a theater play is taken to involve creativity; but many other uses of language may as well. Take the example of giving a quick-witted response to an unpleasant interview question or to some remark that is considered inappropriate. Since language is ubiquitous in human activity – and it is comparatively easy to collect samples of language output and for instance process text corpora with the computer – it comes as no surprise that a lot of research on human or machine creativity targets creativity manifested in (some aspects of) text(s).

What is problematic however, in particular when the goal is to develop a systematic understanding of the processes underlying creativity, is the following: Language production (and hence text as its output) is a massively multi-layered phenomenon.

The multi-layered character of text. A highly diverse collection of knowledge spheres and contextual factors play together in the production of any element of text. Hence, pinpointing the role of a particular knowledge source in an account of creative behavior based on textual evidence is very hard. Since humans can effortlessly handle the network of cross relations among levels of language and text, a decision at one level will have consequences in multiple other levels in human-generated text. For instance, different ways of describing a certain action ("she warned him/she gave him a heads-up/she drew his attention to the fact that ...") may be truth-conditionally equivalent, but connotations, conventions in a particular text genre, domain-specific jargon, script knowledge about (culture-specific) scenarios etc. can make specific alternatives appear humorous, sarcastic, arrogant, mildly impolite, etc. Some of the most aesthetically appealing examples of creative language use keep most cross-level relations aligned with what is to be expected from conventions etc., but then break expectations (Takala 2005; Raby 2010) at a possibly subtle, but effective point. Creative language use thus plays with the reader's/audience's (mostly unconscious) knowledge about typical cross-dependencies of levels of language and text.

Consequences for computational creativity research. The rich interrelations between levels and connotations of language elements poses considerable challenges to systematic generative research. Controlled experiments manipulating certain text elements can easily be disrupted by side effects at entirely different text levels that for instance cause human readers to find passages unnatural.

For a long time, important subfields of computational creativity such as story generation (Gatt and Krahmer 2018; Gervas 2009), had therefore adopted the strategy of focusing on a particular text level for evaluation (and systematic description), e.g., the plot level. The surface realization of a formal plot description as an actual story does not aim to reach the aesthetic sophistication of human writing. Advances in the fields underline that this strategy of focusing on particular levels is effective for developing a better systematic understanding of particular elements of creative writing (without drawing into question that they interact freely in actual human creative performance) (Lehnert 1981; Holmes 1985; Papalampidi, Keller, and Lapata 2019).

Transformer language models. The developments in Natural Language Processing research on language modeling of the past 5-10 years call for a new assessment of the situation: transformer language models using hundreds of billions of parameters (Brown et al. 2020) and trained on gigantic collections of text apparently display a generative behavior that reflects many of the dependencies across text levels and relevant knowledge spheres. In typical application examples of completing a short text prompt, the models' choices in text production quite often seem to adhere to what for a human writer would be traced back to an intuition regarding connotations, genre convention and the other knowledge spheres listed above. Therefore, it is little wonder that the new generation of language models are finding many applications in a creative writing context (Bena and Kalita 2020; Ammanabrolu et al. 2019).

The solution? One might feel inclined to conclude that with transformer language models, the challenge from the multi-layered character of text for research on creativity has been overcome: The models are capable of generating stretches of text that are indistinguishable from human text. However, no matter whether one wants to employ such a model to enhance human creativity (in a co-creative scenario) or to use algorithmic models of creativity to advance our understanding of the processes underlying human creativity – the plain task of eloquently completing a given text prompt provides too little control. The language generator can "wander off" freely from the starting point and may take arbitrary "turns", most of which can be traced back to some explainable connection after the fact. But what is missing is even a slight element of goal-orientation. With all the difficulties in defining creativity, there is a far-reaching consensus that it not only involves an element of originality/novelty, but the product of creativity also needs to have a value of some kind (creativity as the production of "something original and worthwhile" (Sternberg, Sternberg, and Mio 2012)). This second element is not within the scope of the computational model of the process when the language model can wander off freely. For systematic research into creativity, this precludes the testing of specific hypotheses regarding creative processes (beyond the class of hypotheses that addresses only the role that experience and exposure to text collections that reflect certain conventions). In a pure creativity enhancement scenario, the inspiring effect of prompt-based generation alone may carry quite far, depending on the human maker's readiness to weed out fruitless output. But here too, exerting control over certain dimensions of the generated output could make the use of language models considerably more effective.

Desideratum. To make progress in the integration of transformer language models into computational creativity research and applications, we can hence identify a goal for the next years: a model architecture and methodological should be developed that is (i) based on current transformer language models with their ability to replicate the cross-level coherence of human language production, and (ii) at the same time allows for a constraining of several important dimensions of the generated text.

This paper reports on experimental work aiming to contribute to this goal. We start out with a pre-trained transformer language model and aim to constrain its generative behavior both in terms of text structure and in terms of the content domain that the text output is about. In a generalizable methodological set-up, it should be possible to characterize the two dimensions of constraining separately (i.e. the method should not only be applicable when there is a sufficiently large dataset for model tuning that happens to combine the two dimensions).

The computational work we report on in this paper grew out of pilot experiments conducted to have some tangible input for brainstorming sessions regarding the integration of AI generated elements in an actual theater production.

Related Work

From a computational point of view, automatic language generation has long been tackled as a task that would employ a pipeline architecture. First, knowledge structures such as dependency graphs or tables are used to form a plan of the events to describe (planning step). Then the the appropriate language is inserted via automatic grammars or slot-filling mechanisms (realization step). Such systems employ character goals (Meehan 1977), author goals (Dehn 1981) or underlying discourse states, (McKeown 1985) among other approaches (Callaway and Lester 2002) (Gervás et al. 2019). In recent years however, powerful language modeling approaches based on transformer deep neural networks (Vaswani et al. 2017) such as GPT-2 (Radford et al. 2019), GPT-3 (Brown et al. 2020) or T5 (Raffel et al. 2020) have shown to generate text of near human quality without the need of underlying knowledge structures (`https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html`, `https://openai.com/blog/better-language-models/`). In order to achieve this

level of knowledge, these language models are typically trained on millions of documents. However, while these models easily produce sophisticated text on their own, controlling the output can be difficult. One approach is to employ *Conditional Language Modeling* (Keskar et al. 2019) which prepends specific codes to text data. Using these codes during inference allows the language model to draw the generations from the part of the training data to which the code belongs. Other approaches apply the information at the self attention layer of the transformer network (Ziegler et al. 2019). The approach presented in this paper also shares similarities with (Dathathri et al. 2020)[1] who influence the gradient of the language model with keyword lists or a separate classifier in order to guide the language generation towards certain topics. Similarly (Pascual et al. 2021) use a simple method of guiding the language model towards semantically similar words of a desired topic. In many ways, the challenge in natural language generation today lies in reconnecting pre-trained language model with underlying knowledge structures (Chen et al. 2020a; 2020b; Peng et al. 2021).

Experiments

In our experiment, we build a text generation model using the GPT-2[2] transformer model from OpenAI (Radford et al. 2019). GPT-2 learns to predict the most probable next word in a sequence of words on a large collection of text. Pre-trained GPT-2 models have seen millions of documents and have been shown to produce several paragraphs of text almost indistinguishable from a human author (see **Related Work**). We leverage this learning power to first train a model that generates German theater plays. While these generations are already of fairly high quality, we inject these generated theater plays with new topics by further fine-tuning the GPT-2 generation models with additional data. We refer to this additional data as *topic corpus*. Our goal is to produce generations which keep the formal structure of a theater play (i.e. a sequence of speech acts that are exchanged between 2 or more characters) but change the topic of the play to that of the topic corpus. This way we attempt to constrain and guide the generation model towards a specific topic without changing the underlying structure of the learned language. We believe that by utilizing German language, our experiments have the additional merit of demonstrating the effectiveness of language modeling on languages other than English, which has already been extensively researched.

Datasets

The **Quadrama Corpus** (`https://quadrama.github.io/index.en`) is a machine readable collection of German theater plays from the late 18th and 19th centuries. The corpus contains many detailed annotations such as character relations, however we mostly make use of the plain text along with the annotations of the surface realizations of characters. We extract the text of each play token by token and mark the characters to which the text belongs by putting their names in capitalized letters followed by a colon (":"). We add a special token ($< |scene\_end| >$) at the end of every scene, which is later used in the generations. To form the plain text training corpus all scenes of all plays are concatenated into one text file. The final concatenated dataset contains 687 plays with a total of almost 14 million words. We refer to this dataset as *Quadrama*.

For fine-tuning the drama generation on a specific topic we use a variety of corpora, all in German language. We refer to each of these datasets as **topic corpus** throughout the experiments:

German Recipes Dataset. A collection of cooking recipes from a German recipe website available from *Kaggle*: `https://www.kaggle.com/sterby/german-recipes-dataset`. We concatenate all recipes to form the corpus, which we refer to as *recipe* corpus in the experiments. The *recipe* corpus contains 12190 recipes consisting of around 1.4 million words.

Horror Fanfiction. We create a small collection of stories categorized under *horror* from German website `https://fanfiction.de`. It should be noted that these stories do not contain popular media characters (as is common for fanfiction) but are entirely original. We concatenate all chapters of all stories into a single file to form the *horror-fanfiction* corpus. The corpus consists of 948 chapters with approximately 1 million words. **Expert Interviews**. This corpus contains a set of concatenated journalist interview transcriptions. The interview topics revolve around modern day issues concerning business, technology and role of artificial intelligence. We concatenate all interviews, including interviewer questions into a single file. In the experiments, we refer to this corpus as *expert-interview* corpus. This is our smallest corpus, containing 1242 utterances from 14 interviews and consisting of around 91000 words. description

Evaluation

Our goal in the evaluation is to get an idea of how well the generations adapt to the content of the desired topic while keeping the structure of theater plays. While we curate a number of generations for a qualitative analysis, we also devise a simple automatic evaluation scheme which uses a combination of three statistical properties of the generations with regards to the topic corpus. We preprocess each topic corpus by filtering stopwords and punctuation, lowercasing and stemming all words, creating a set of content words D we use to represent each topic corpus.

Given a collection of generations G, we first calculate the number of generated words that appear in D. For each $g \epsilon G$ we count how many of the generated tokens appear in D and average the count over all generations. We assume that the generations are thematically closer to the topic corpus when they use a higher number of content words. We refer to this

[1] We experimented with the approach in preliminary experiments but found the method to be difficult to tune and easily lead to repetitive generations.

[2] We choose to use GPT-2 over a newer model such as GPT-3 or T5 because of the relatively small size of GPT-2 (compared to its successors) and the high number of available pre-trained language models including a model trained on German texts.

measure as *content word frequency* (1).

$$content - word - frequency : \frac{\sum_{g \epsilon G} |\{w | w \epsilon G \wedge w \epsilon D\}|}{|G|} \quad (1)$$

$$topic{-}corpus{-}coverage : \frac{\sum_{g \epsilon G} \sum \frac{\{count(w)_C | w \epsilon g \wedge w \epsilon D\}}{|C|}}{|G|} \quad (2)$$

In addition to how many content words are used in each generation, we are also interested how frequent these words are in the topic corpus. For every $w \epsilon D$ we calculate the percentage of how often it appears in the set of tokens of the topic corpus *C*. We refer to this score as *corpus coverage*. For every *g*, we sum the corpus coverage of all content words that appear in *g* and then average over the whole generation set *G*, yielding a score we refer to as *topic corpus coverage*. We report *topic corpus coverage* as a percentage from 0 to 1. (2)

While the former two scores estimate the degree how much the generation model adapts to the topic corpus, we also want to make sure that we are not losing the text structure of theater plays. The nature of plays entails, that there are characters present in the text who speak the dialogue. We verify that this property holds in the generations by making use of the *Quadrama* annotations. In the *Quadrama* corpus, characters are written in capitalised letters followed by a colon (":"). Therefore, we can count how many speakers we find in a generation with simple surface matching. In the **Results** section we refer to this score as *number of speakers*.

Our quantitative evaluation approach gives us the possibility to investigate a large amount of generations automatically. In particular, we can verify to what extent the generation adapts its content words to the domain corpus. Overall, we omit an analysis of readability. Manual inspection of around 100 generated texts shows That quality of the generations is generally close to human level and the desired drama style is often difficult to read, even in the original drama corpus. We also decide not to evaluate coherence of the generated text, as that is not the focus of our experiment. We do however perform a qualitative analysis of two examples per domain corpus in the Section **Handpicked Examples**. We highlight both a successful topic adaption as well as a generation, where the play structure has been lost or the topic has not been integrated.

Setup

First, we fine-tune a pre-trained German GPT-2 model[3] on the *Quadrama* dataset (Section **Datasets**) for 3 epochs using *ADAM* optimizer with a starting learning rate of $5e^{-5}$. The resulting model is capable of generating drama text with consistent grammar in a very distinct language style (Figure 5). In order to incorporate domain specific content into the generated plays, we perform fine-tuning again using one of the topic corpora (see Section **Datasets**) for a single epoch with different learning rates. In particular, we investigate 3 learning rates: $5e^{-5}$, $5e^{-6}$ and $5e^{-7}$. We find that training with a learning rate higher than $5e^{-5}$ leads to overfitting and repetitive, stale generations.

It is common practice to provide a piece of text that the generation model then attempts to complete. This is also called **cue**. In the experiments we use very short pieces of text since we want the generations to be mostly dependent on the generation model. We also find that generating without a cue, the fine-tuned models will generally stick to the drama style language instead of incorporating the new information. As such, we provide a cue to the model that starts with a drama character and is then followed by one or two words from the topic corpus. We select words with a generally high frequency in the topic corpus (Section **Evaluation**) to serve as the generation model cue. For each learning rate, we fine-tune the *Quadrama*-model on the topic corpus and output 100 generations, using sampling decoding with a top *k* of 50. We generate until the $< |scene\_end| >$ token is reached up to a maximum of 100 tokens. For each topic corpus, we compare the output of the adapted generation models to the base *Quadrama*-model.

Results

Statistical Analysis

Figure 1 illustrates the results of the statistical evaluation across all generation experiments. Starting with the *recipe* topic corpus, we see that the fine-tuned generation model achieves significantly higher topic term frequency and corpus coverage when using the larger two learning rates ($5e^{-5}$, $5e^{-6}$). When using a learning rate of $5e^{-5}$, the model scores 30 words relevant to the topic in each generation on average, which is double the amount compared to using the *Quadrama only* model which was not fine-tuned on the *recipe* corpus. Similarly corpus coverage more than triples when using the larger two learning rates from 0.05 for the *Quadrama only* model to around 0.15 for the fine-tuned model. This signifies that the fine-tuned generations contain words which span around 15% of the *recipe* corpus. However, looking at the number of speakers we see that the improvements come at the cost of the play structure. Without fine-tuning on the *recipe* corpus, the model achieves 4 speakers per generation on average. This number decreases to 1 when using the larger two learning rates. We therefore assume that the play structure has been lost in most generations. We find that using the smallest considered learning rate $5e^{-7}$ yields the best compromise between play structure and topic integration. The fine-tuned model achieves on average 20 topic words which span around 7% of the topic corpus while keeping the average number of speakers around 3.

For the *horror-fanfiction* corpus, we find the overall best compromise between topic adaption and theater play structure when using the learning rate $5e^{-6}$. While the larger learning rate yields a higher number of topic words per generation it also decreases the number of speakers to an average of 1 per generation. The smallest learning rate preserves the number of speakers well at around 4 but

[3] using the language modeling code and *anonymous-german-nlp/german-gpt2* model freely available on huggingface: `https://huggingface.co`

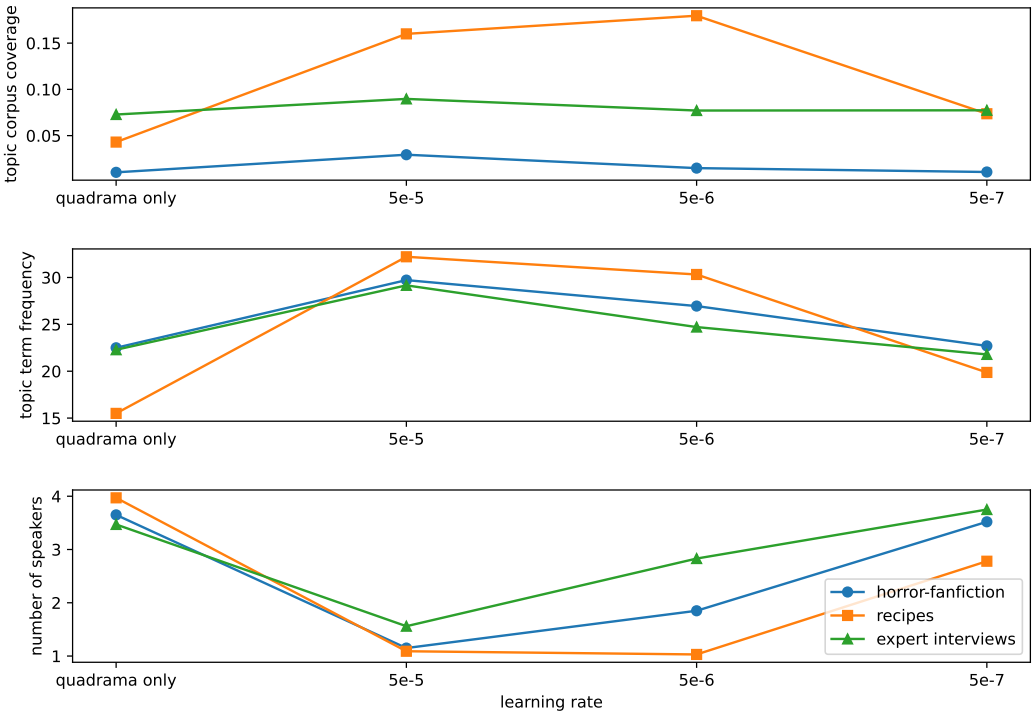

Figure 1: Statistical analysis of all generation models. Plots show, top to bottom *topic corpus coverage*, *topic term frequency* and *number of speakers* averaged over 100 generations from the generation model. The generation models considered are trained only on the *Quadrama* corpus (*quadrama only*) or received an additional fine-tuning step for 1 epoch with the listed learning rate on the x-Axis ($5e^{-5}$, $5e^{-6}$, $5e^{-7}$) on the respective topic corpus.

hardly affects the number of topic words (around 23, same as *Quadrama only* model) or their coverage the topic corpus (around 0.01, same as *Quadrama only* model).
Lastly, we generate theater plays with technical or business related topics by fine-tuning with the *expert-interviews* corpus. We find that the experiment behaves similarly to using the *horror-fanfiction* corpus. For the smallest learning rate $5e^{-7}$, the frequency of topic words per generation nor the coverage of the topic corpus improves upon using the *Quadrama only* model (around 23 and 0.8 respectively). The play structure is present, as can be seen by an average of 4 speakers per generation but we can assume that the topic is not integrated. Overall, we see that topic corpus coverage is not improved by any of the fine-tuned models, as it is already quite high for the *Quadrama only* model at around 0.08 compared to the other topic corpora. However, we find that when using the largest learning rate $5e^{-5}$ there are still slightly less than 2 speakers per generation on average while there is an improvement of 10 topic words per generation from around 15 to around 25 compared to the *Quadrama only* model. Therefore, we assume that the play structure is still present in many of the generations, if only with 2 speakers.

Handpicked Examples

We inspect all generated texts manually and select samples that illustrate the effect of topic integration in Figure 2, 3 and 4. We choose what we find to be the best fitting generations for the model trained only on the *Quadrama* corpus on the left. The highest quality generation for the model fine-tuned on the topic corpus is found on the right. The cue given to the model is marked in **bold**, phrases and words pertaining to the topic corpus (selected regarding to topic corpus words, see Section **Evaluation**) are marked in red and pointed brackets (<>).
Figure 2 shows that without fine-tuning on the topic corpus, the generation model accommodates the recipe domain to an extent. The characters aptly react to the presence of food items in the cue and mention **Zwiebeln**(*onions*) and **Pfeffer**(*pepper*), as well as **Salat**(*salad*) a number of times. The main language style of the generation is still that of a theater play, the food items seem to replace some of the subjects that are talked about. Overall, the presence of food items in the context seems to confuse the generation model which leads to the scene being generally difficult to understand and nonsensical in nature. Interestingly, the characters in this generation seem to actively move away from the topic of *salad*, as one character exclaims: **Wir haben uns zu sehr mit Salat beschäftigt.**(*We have concerned ourselves too much with salad*). The effect we see when fine-tuning on the *recipe-*

PRINZESSIN: Zuerst Tomaten<sauce, dann Zwiebeln>.
JAKOB: Nun? So?
SICHEL: Warum denn für < Salat>?
JAKOB: Dazu. Man kann jeden mit <Salat und Pfeffer>aufdecken. Und wenn man auch nichts Besseres bietet, als <Salat>, das macht mir sehr viel Spaß.
SICHEL: Wir haben uns zu sehr mit <Salat> beschäftigt.
JAKOB: Mir scheint, es freut mich.
SICHEL: Ist es denn möglich, daß sie dir...

PRINZESSIN: Zuerst Tomaten<sauce und Pfeffer> setzen, <dann die Gurken dazu und dann den Knoblauch dazu bis zum Rand fein hacken, bis sie den Salat putzen und damit die Gurken> schmieren!
FRAU VON WOLFF: So! So! Jetzt <erst die Gurken zusammen>pflegen und dann <den grünen Salat> in Topftöpfen festbinden <mit Pfeffer und Essig und den Wein eingießen und mit Salz und Pfeffer würzen>.
FRAU VON WOLFF: Aber lieber <das Salz ins Wasser>.

Figure 2: Generations for the *recipe* topic corpus. The left column shows generations with the model trained only on the *Quadrama*-corpus. The right column shows generations of models fine-tuned on the topic corpus.

PRINZESSIN: Im Schatten will ich hinscheiden.
Lulu.: Im <Schatten> will ich hinscheiden? Ich sehe des Vaters <Tod>!
EUSENLEITNER: Mit aller Kraft.
Lulu.: Und doch ist er mir entrissen – Und doch steht mein Mann –
TOLOMEO: Was soll das heißen? –
Lulu.: Er hat mich verlassen, aber ich darf nichts von ihm hören, wie das <Sterben eines Kindes>.
TOLOMEO: Wozu sollt ich hinauß?

PRINZESSIN: Im Schatten des Lichts, ist es <ein Dämon>?
DER KARAMUKAIN: Es ist <eine furchtbare Finsternis>. Ich wachte darauf, sah die Erde mit den Sinnen an. <Der Anblick dieses Untiers hat mich verwirrt>. Doch <im Schatten> des Lichts, ist es <ein Dämon>! Ich fühle mich nicht dort, fühle mich hier <gefangen>. Versteh mich nicht weiter von diesen Augen und auch nur von diesen verrannten Fühlen und von dem Glauben bei der Ankunft des Herrn.
(fine-tuned on *horror-fanfiction*, learning rate $5e-6$)

Figure 3: Generations for the *horror-fanfiction* topic. The left column shows generations with the model trained only on the *Quadrama*-corpus. The right column shows generations of models fine-tuned on the *horror-fanfiction* corpus.

corpus is much more pronounced: The characters dialogue essentially changes to recipes as the two characters in the fine-tuned example seem to actively talk about the preparation of a meal. The whole dialogue is structure in the style of a recipe with several ingredients being mentionend, such as **Knoblauch**(*garlic*), **Gurken**(*cucumber*) and **Essig**(*vinegar*). In addition, both characters also reference methods of preparation for these ingredients, such as **Salat putzen**(*clean the salad*) or **Wein eingießen**(*pour in Wine*). There is also still a degree of interaction between the speakers as the second character picks up the *cucumber* and salad mentioned by the first character and furthers the cooking instructions to now include seasoning. There are some incoherences: **Topftöpfen** would mean something like *potpots*, which does not have a clear meaning. Also **Tomatensauce und Pfeffer setzen** (*put tomato sauce and pepper*) is not a valid expression since the presence of the verb **setzen** would be highly confusing to a native German speaker in this context. In general though, incoherences seem particularly noticeable here as the recipe style dialogue contains explicit instructions that are easily understood and leave little room for interpretation compared to a more poetic style of language.

We illustrate two of the generations for the *horror* topic in Figure 3. Without additional fine-tuning on the *horror-fanfiction* corpus, the generation model already produces words that can be considered relevant to the topic, such as **Tod**(*Death*) or **sterben**(*to die*). However most of the language clearly sticks to the drama style. The word **hinscheiden**(*pass away*) for example is much more poetic and more typical of drama language than what we find in the topic corpus. The generation after fine-tuning on the *horror-fanfiction* corpus clearly adopts a more prosaic style, leading to a long text of the second character. We see, that the play structure is not as strong as it was before fine-tuning on the topic corpus. The language itself however reads much more like a scary novel using words such as **Finsternis**(*darkness*), **Dämon**(*demon*) or **Untier**(*monster,beast*). While the theater texts already contain a good amount of dramatic or even scary terms which lead to some adaption even without fine-tuning we find that the fine-tuned generations tend to fit the style of the topic corpus much better. Though this often leads to the characters talking in long monologues, effectively removing the theater play structure. As we already inferred from the statistical analysis, adapting to the *horror-fanfiction* corpus and keeping the drama language structure seems more difficult compared to for example the *recipe* corpus. A possible reason for this is that the chapters in the *horror-fanfiction* corpus tend to be a lot longer than a single recipe and as such, the generation model is incentivised to produce longer paragraphs of text. Another big difference is of course the author perspective which for the *recipe* corpus is fixed to an imperative, instructional style and in the *expert-interview* corpus is exclusively first person dialogue. Both of these language style naturally fit easily into theater plays, which consist mostly of discourse between several participants. The third person narration employed in many chapters of the *horror-fanfiction* corpus however is very rarely used in drama text. As such, we suspect the generation model has a harder time integrating much of the text in the corpus, causing many

PRINZESSIN: Der Zukunftsboom treibt noch immer den Geist des <Kapitals> herauf. Wenn es aber auf der weiten Welt dem Proletariat so lange nachhängt, so hat das keinen Grund : die Arbeiterschaft gewinnt die höchste Stellung.
BARBARA: Warum ist der <Weltverbesserer> so bieder und so tiefsinnig, mein Fürst?
ZERBINO: Sein Bruder ist vielleicht krank. Sein Vater hat ihn gesund und hat ihm alle Gesundheit verloren. Seit zwei Weltkriegen hat sein Vater über ihn zu sprechen, wie ...

PRINZESSIN: Der Zukunftsboom wird die <europäische Gesellschaft> bald reformieren können. Aber Sie haben hier eine Chance.
REDLICH: Ich habe hier <ein enormes Marktpotential> erkannt. Ich glaube auch, dass die <Industrie das Thema der Zukunft durchaus antizipieren> konnte, so dass man das nicht zu einer <Diskussion von Chancen oder Herausforderungen führt>. Und das ist auch die Grundlage dafür, dass wir es jetzt auch direkt mit den <ökonomischen, sozialen und politischen Dimensionen des Wandels> zu tun haben. Wir haben diese Chance nämlich in unserer <internationalen Debatte> ...
(fine-tuned on *expert-interviews*, learning rate $5e-5$)

Figure 4: Generations for the *expert-interview* topic. The left column shows generations with the model trained only on the *Quadrama*-corpus. The right column shows generations of models fine-tuned on the *expert-interview* corpus.

PRINZESSIN: Euer Hoheit, daß ich mich empfehle. Ein andres Exempel.
HERR VON BIEDERLING.: Ich bitte, auf wen ich mich beziehe :
DER JUNGE: Ich muß es zu ihm sagen, will dem Herrn von Biederling sagen, da ich dem Herrn auf dem Ball nicht allein zuhören soll.
FRAU VON MAUER.: Mein Herr, ich darf mich noch auf die Sache einigen.
DER DRITTE: Zum Glück kann ich mir meine Untersagung zur Unzeit überlegen, Ihr habt ja vor der Verlobung in der Galerie mit Herrn von Biederling gestanden.
HERR VON MAUER.: Ich weiß, daß Ihr Euch heute noch auf das Wort einläßt, weil –
DER JUNGE: Sie ist zu der Aufseherin gekommen, die in der Galerie wartet.

Figure 5: Generation from the German GPT-2 model fine-tuned on the *Quadrama* corpus. The cue given to the generation model is marked in **bold**.

generated texts to trail off into narrations rather than theater plays.
Figure 4 illustrates generation results using the *expert-interview* corpus. Again, we find that the model can adapt to the topic without seeing the topic corpus, albeit within the confines of its play context. The scene generated without fine-tuning on the topic corpus yields a conversation about politics, mentioning words like **Kapital**(*capital*),**Arbeiterschaft**(*working class*), **Proletariat** and **Weltkrieg**(*world war*), which are all topics that can reasonably occur in theater plays. Though these terms are technical and relate to finance and politics, they do not reflect the topics of the *expert-interview* corpus which deals more with modern day businesses and computer technology. After fine-tuning on the *expert-interview* corpus we find that the generation incorporates much more modern terms, such as **Marktpotential**(*market potential*) and **internationale Debatte**(*international debate*) which are not very typical of theater plays thus demonstrating a degree of topic integration that was not present before. It should be noted that this is the only experiment where we picked generations using the largest learning rate of $5e^{-5}$. While for the other two topics, this learning rate caused the play structure from the generations to be lost, here we can still find many generations with at least 2 speakers. This might well be because the *expert-interviews* corpus consists of dialogue-style language and as such, causes the model to retain this dialogue structure after fine-tuning.

Discussion

Selection of high quality Generations

First, we should note that there are many generations which do not exhibit the favourable properties of the ones shown in Section **Handpicked Examples**. Some generations do not include the topic at all, despite fine-tuning and the generation cue. Other generations that fit the desired topic stray to far from the structure of a theater play and as such do not introduce any speakers into the scene. In order to find high quality results manual inspection of the generations is necessary. We do find however, that the provided statistical analysis is helpful in selecting good generations. While we checked all generations when curating the best results we ended up finding the most promising generated scenes from the models that offered the best compromise between the number of speakers and the frequency of topic words. In addition, we believe that our approach works well in an assisted creative writing setting where the author has more control over text that is generated line by line. This way, the generation model can be used like a tool that inspires creative output, in our case theater plays with possibly unusual topics.

Quality of generated Scenes

We find that many of the generations lack coherence overall. Many of the spoken dialogues, while grammatically correct, are very hard or impossible to make sense of. We investigate a generated example fine-tuned only on the *Quadrama* corpus in Figure 5. While the general style of language is

PRINZESSIN: Euer Hoheit!
KÖNIG: Ja Prinzessin. Was gibt es?
PRINZESSIN: Nun, ich will die ganze Welt aufbieten, und Euer Hoheit wollen mich nicht in den Krieg stürzen.
KÖNIG: Ich bin ein alter Narr, und ich bin ein ehrlicher Mann ; ich habe mich mit den alten Menschen in Verbindung gesetzt.
PRINZESSIN: Das wäre ein Unglück, wenn Ihr mich nicht in den Krieg ziehen lassen würdet.

Figure 6: Generation from the German GPT-2 model fine-tuned on the *Quadrama* corpus. This generation was created in tandem with the generation model. Text pieces provided by the author are marked in bold, the remaining text is automatically generated.

very evocative of a classic German theater play, the actual content of the scene is harder to follow. We find some plot points in the scene though: Someone was waiting in a gallery (**Galerie**) before an engagement (**Verlobung**) and now an attendant (**Aufseherin**) is waiting there. It is not clear who is speaking to whom however, which makes constructing a narrative near impossible. We also find that the generation model greatly benefits from a longer context. Figure 6 shows a scene created by alternating between the human author and the generation model. We find that in Figure 6, existing characters are repeated more consistently. In addition, the princess (**PRINZESSIN**) character states a desire to go to a war in both of her generated passages, displaying a coherence that is not present in Figure 5.

Interestingly, we also find that after fine-tuning on a topic corpus, the generations generally show more coherence when they actually adapt to the topic and are easier to understand. This effect can also be observed in the generations presented in Section **Handpicked Examples**, for example in Figure 2, where the generation without fine-tuning on the topic corpus seems confused by the presence of particular words. We assume that the reason for this lies primarily in the fact, that the words which are relevant to the topic are very rare in the *Quadrama*-corpus and as such, the generation model is less certain on how to continue the generation. This can lead either to the generated words become more and more random or to the generation model starting to ignore the topic words. It should also be noted however, that the language present in the *Quadrama* corpus is generally very complex and often hard to understand even for native speakers. Theater plays employ a very distinct language style and often obscure details of characters motivations, actions and intentions within the dialogue. In addition, many plays in the corpus are more than on hundred years old and use a vocabulary that is very different to modern language. This is a possible reason why the GPT-2 generation model replicates the language style but struggles with generating a coherent narrative. Apart from providing longer contexts to the generation model, another possible way to possibly improve overall cohesiveness would be to use more data for more robust fine-tuning, avoiding possible overfitting. Another approach is to tackle the decoding process of the language model. There are decoding strategies that reportedly improve the coherence in generated content ((Holtzman et al. 2019)) and those methods will likely improve results in our experiments as well.

Conclusion

Across all experiments, we find that our fine-tuning approach can achieve integration of the desired topic without losing the structure of theater plays. In particular, we show that the generation models incorporate words and concepts that were not present before the fine-tuning on the respective topic corpus. Furthermore, we illustrate that these concepts are integrated into dialogue spoken by at least two characters, creating a mixture between the theater play structure and the respective topic. While there is still room for improvement, particularly in the coherence of generated texts and the fairly high selection effort, we conclude from our results that our approach generally achieves its goal of injecting a new topic into the existing language structure. Furthermore, our approach does not require abundant data or specialised annotations. Apart from the corpus of theater plays, topic corpora similar to the ones presented in Section **Datasets** can be easily acquired from openly available sources. In addition, we also show that such a topic corpus does not need to be particularly large. The smallest topic corpus we use is the *expert-interviews* corpus with less than one hundred thousand words and we still see a strong effect there. This is useful in practice, as training a transformer language model on too little data can quickly lead to overfitting and consequently causes uninteresting, often repetitive generations.

We propose to further experiment with different ways of encoding to improve the readability and coherence of the generations. We also encourage the use of our fine-tuning approach in creative writing settings, be it fully automatic or in co-operation with the generation model in order to try out unusual combinations of topics.

Author Contributions

Lukas Wertz - Idea, Revision of Full Paper, Related Work and References, Experiments, Discussion/Analysis
Jonas Kuhn - Introduction, Motivation and Background

Acknowledgements

This research was funded and supported by the Institute of Natural Language Processing (Institut für Maschinelle Sprachverarbeitung: **IMS**) at the University of Stuttgart.

References

Ammanabrolu, P.; Broniec, W.; Mueller, A.; Paul, J.; and Riedl, M. 2019. Toward automated quest generation in text-adventure games. In *Proceedings of the 4th Workshop on Computational Creativity in Language Generation*, 1–12. Tokyo, Japan: Association for Computational Linguistics.

Bena, B., and Kalita, J. 2020. Introducing aspects of creativity in automatic poetry generation.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Callaway, C. B., and Lester, J. C. 2002. Narrative prose generation. *Artificial Intelligence* 139(2):213–252.

Chen, W.; Chen, J.; Su, Y.; Chen, Z.; and Wang, W. Y. 2020a. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7929–7942. Online: Association for Computational Linguistics.

Chen, W.; Su, Y.; Yan, X.; and Wang, W. Y. 2020b. KGPT: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8635–8648. Online: Association for Computational Linguistics.

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2020. Plug and play language models: A simple approach to controlled text generation.

Dehn, N. 1981. Story generation after tale-spin. In *IJCAI*, volume 81, 16–18. Citeseer.

Gatt, A., and Krahmer, E. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.* 61(1):65–170.

Gervas, P. 2009. Computational approaches to storytelling and creativity. *AI Magazine* 30(3):49.

Gervás, P.; Concepción, E.; León, C.; Méndez, G.; and Delatorre, P. 2019. The long path to narrative generation. *IBM Journal of Research and Development* 63(1):8:1–8:10.

Holmes, D. I. 1985. The analysis of literary style–a review. *Journal of the Royal Statistical Society. Series A (General)* 148(4):328.

Holtzman, A.; Buys, J.; Forbes, M.; and Choi, Y. 2019. The curious case of neural text degeneration. *CoRR* abs/1904.09751.

Keskar, N. S.; McCann, B.; Varshney, L. R.; Xiong, C.; and Socher, R. 2019. Ctrl: A conditional transformer language model for controllable generation.

Lehnert, W. G. 1981. Plot units and narrative summarization. *Cognitive Science* 5(4):293–331.

McKeown, K. R. 1985. Discourse strategies for generating natural-language text. *Artificial Intelligence* 27(1):1–41.

Meehan, J. R. 1977. Tale-spin, an interactive program that writes stories. In *Ijcai*, volume 77, 91–98.

Papalampidi, P.; Keller, F.; and Lapata, M. 2019. Movie plot analysis via turning point identification.

Pascual, D.; Egressy, B.; Meister, C.; Cotterell, R.; and Wattenhofer, R. 2021. A plug-and-play method for controlled text generation. *ArXiv* abs/2109.09707.

Peng, X.; Li, S.; Wiegreffe, S.; and Riedl, M. 2021. Inferring the reader: Guiding automated story generation with commonsense reasoning.

Raby, G. 2010. Improvisation and devising: The circle of expectation, the invisible hand, and RSVP. *Canadian Theatre Review* 143(1):94–97.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8):9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21(140):1–67.

Sternberg, R. J.; Sternberg, K.; and Mio, J. 2012. *Cognitive psychology*. Cengage Learning Press.

Takala, T. 2005. Creativity as disruptive adaptation – a computational case study. In *HI'05 Sixth International Roundtable Conference Computational and Cognitive Models of Creative Design*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is all you need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ziegler, Z. M.; Melas-Kyriazi, L.; Gehrmann, S.; and Rush, A. M. 2019. Encoder-agnostic adaptation for conditional language generation. *arXiv preprint arXiv:1908.06938*.