StoryDB: Broad Multi-language Narrative Dataset

Alexey Tikhonov
Yandex
Berlin, Germany
altsoph@gmail.com

Igor Samenko
ICT RAS
Novosibirsk, Russia

Ivan P. Yamshchikov
LEYA Laboratory
Yandex and
Higher School of Economincs
St. Petersburg, Russia
ivan@yamshchikov.info

Abstract

This paper presents StoryDB — a broad multi-language dataset of narratives. StoryDB is a corpus of texts that includes stories in 42 different languages. Every language includes 500+ stories. Some of the languages include more than 20 000 stories. Every story is indexed across languages and labeled with tags such as a genre or a topic. The corpus shows rich topical and language variation and can serve as a resource for the study of the role of narrative in natural language processing across various languages including low resource ones. We also demonstrate how the dataset could be used to benchmark three modern multilanguage models, namely, mDistillBERT, mBERT, and XLM-RoBERTa.

1 Introduction

Stories are central to human culture and communication. However, it seems that stories are easier said than generated. Despite incredible recent progress in natural language processing generation of longer texts is still a challenge (van Stegeren and Theune, 2019; Rashkin et al., 2020). Ostermann et al. (2019) present a machine comprehension corpus for the end-to-end evaluation of script knowledge with 50% of the questions in the corpus that require script knowledge for the correct answer. The authors demonstrate that though the task is not challenging to humans, existing machine comprehension models fail to perform well on the data, even if they make use of a commonsense knowledge base.

Partially, this challenge could be attributed to the lack of adequate memory models. Longer texts demand better memory mechanisms and possible ways to construct such mechanisms are discussed in the literature for the last 25 years. Long short-term memory networks (Hochreiter and Schmidhuber, 1997), Neural Turing Machines (Graves et al., 2014), memory networks (Weston et al., 2014) and many other architectures try to tackle this problem. Attempts to introduce some form of memory in transformers, such as (Guo et al., 2019) or (Burtsev and Sapunov, 2020), could be regarded as the next steps in this long line of work.

There are some interesting recent attempts to generate long texts using some form that makes such longer text feasible for a human reader. For example, Agafonova et al. (2020) generate a diary of a neural network. Yet the generation of a narrative is still challenging. For a detailed review of earlier approaches to narrative generation, we address the reader to (Kybartas and Bidarra, 2016). Even modern models for narrative generation rely heavily on some form of expert knowledge or some type of hierarchical structure of the narrative. For example, Fan et al. (2019) first generate the predicate-argument structure of the text, then generate a surface realization of the predicate-argument structure, finally replace the entity placeholders with context-sensitive names and references. Fan et al. (2019); Ammanabrolu et al. (2020) propose a hierarchical generation framework that first plans a storyline, and then generate a story based on it. present a technique for preprocessing textual story data into event sequences. Xu et al. (2018) develop a model that first generates the most critical phrases, called skeleton, and then expands the skeleton to a complete and fluent sentence. Similarly, Martin et al. (2018) provide a mid-level of abstraction between words and a sentence to minimize event sparsity and present a technique for automated story generation whereby the problem is decomposed into the generation of successive events and the generation of natural language sentences from events. Finally, Brahman et al. (2020) develop an approach, where the user provides the model with such mid-level sentence abstractions in the form of cue phrases during the generation process.

However, we should take into consideration that modern Natural Language Processing (NLP) is fun-

damentally an experimental discipline, so the lack of dedicated data could be another bottleneck for the development of narrative generation. This paper tries to amend this problem.

Unfortunately, the majority of available narrative datasets deal with some constrained form of a short plot that is usually called *scenario*. These scenarios are centered around common activities, i.e. going grocery shopping or taking a shower. These narrative datasets available in the literature are also extremely small and could not be used with the most advanced modern NLP models. Regneri et al. (2010) collect 493 event sequence descriptions for 22 behavior scenarios. Modi et al. (2016) present InScript dataset that consists of 1,000 stories centered around 10 different scenarios. Wanzare et al. (2019) provide 200 scenarios and attempt to identify all references to them in a collection of narrative texts. Mostafazadeh et al. (2016) present a corpus of 50k five-sentence commonsense stories. Finally, there is an MPST dataset that contains 14K movie plot synopses, (Kar et al., 2018), and WikiPlots[1] that contains 112 936 story plots extracted from English language Wikipedia. Recently Malysheva et al. (2021) provided a dataset of TV series along with an instrument for narrative arc analysis. These datasets are useful yet as well as a vast majority of the narrative datasets they are only available in English.

This paper provides a large multi-language dataset of stories in natural language. The stories have a cross-language index and every story and character are cross-linked if they occur in different languages. Additionally, the texts have tags such as a genre or a topic. This is the first story dataset of such magnitude that we know of. We hope that a large dataset of long storylines could be used for various aspects of narrative research as well as to facilitate experiments with end-to-end narrative generation.

2 Data

StoryDB is motivated by several interesting experiments that used WikiPlots — one of the larger English datasets of narratives available for all-purpose narrative research that we have mentioned earlier. Seeing various applications that Wikiplots dataset found in the NLP community, we believe, that StoryDB would be even more useful due to multiple languages, advanced filtering that guarantees higher quality of obtained data, and genre tagging. To improve reproducibility and make StoryDB usable as Wikipedia is further updated we publish the data as well as the code for the filtering pipeline[2]. The stories that form StoryDB are extracted from any Wikipedia article that contains a sub-header that contains the word "plot" (e.g., "Plot", "Plot Summary", etc.) in a corresponding language.

2.1 Dataset structure

The dataset consists of several index files and includes a directory `plots`. Every file in the directory has a similar structure. Two first letters of the filename stand for the ISO 639-1[3] code of the language for the texts presented in the file. For example, `hy_plots.tsv` contains 4 861 plots in Armenian language. The file `simple_plots.tsv` contains stories in Simple English. Every entry in the plots file has a similar structure and includes the following fields:

- ID — the unique number of a plot that is the same across every language in the dataset;
- Lang — the language of this particular entry;
- Link — a link to the Wikipedia page containing the plot;
- Title — the title of the story;
- Text — the text of the story;
- Categories — the categories that Wikipedia assigns to this story.

One can navigate across plot files using StoryDB's Index file `plot_matrix.tsv`. The rows of the file stand for languages. If a given plot is available in a given language then the title of this plot stands in the corresponding cell of the `plot_matrix.tsv`. For example, if "Wee free men" is available in Simple English it could be found by its title in the corresponding `simple_plots.tsv`. StoryDB also includes `plot_rake.tsv` that contains keywords extracted with RAKE algorithm (Rose et al., 2010) for every story.

Finally, the files `ID_lang_tag.tsv` and `ID_tag_average.tsv` include information about tags that correspond to the given story. We discuss tagging procedure in detail later.

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

[2] https://drive.google.com/drive/folders/1RCWk7pyvIpubtsf-f2pIsfqTkvtV80Yv

[3] https://en.wikipedia.org/wiki/ISO_639-1

2.2 Preprocessing

Our motivation is to provide a dataset of storylines for various languages including the low-resource ones. Roughly speaking, we want to be sure that every story that ends up in StoryDB is a legitimate storyline description in the corresponding natural language. Thus we are more interested in the precision of the dataset rather than in the recall. To guarantee a higher quality of the obtained stories we implemented several heuristical filters that we briefly describe here.

English Wikipedia is an order of magnitude bigger than any other Wikipedia both in terms of users and in terms of admins[4]. This makes the English list of storylines to be the most extensive one. We regard it as the least noisy one and use it as a reference source for the filtering procedure. We exclude every page that includes a plot yet has no plot section in English Wikipedia for the same entry.

If Wikipedia in language X has a page with title A and this page is also available in language Y under title B, we list such pair of stories as `[language_X, title_A, language_Y, title_B]`". Every entry in this list is an edge in a graph of stories. Every vertex in this graph has a corresponding name `language, title`. Unlike connected stories from different languages that usually contain similar storylines, the stories listed under the same name in the same language might differ significantly. Say, two stories in language X `[language_X, title_A]` and `[language_X, title_B]` are both linked to one story in another language Y `[language_Y, title_C]`. To avoid such ambiguities we exclude fully connected components that contain more than one entry in the same language. Obtained list of stories ends up in the resulting matrix of stories to navigate the dataset. We experimented with various filtering procedures and found this combination to produce a sufficiently rich dataset with a minimum amount of duplicates.

StoryDB is also equipped with a catalog of characters. If a given character that has an individual Wikipedia page is mentioned in a story, its description in the original language is saved into the corresponding tsv-file alongside the ID of a story and the language of the description.

[4]https://meta.wikimedia.org/wiki/List_of_Wikipedias

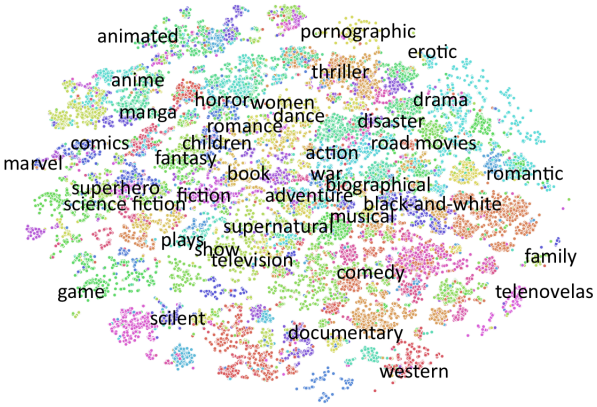

Figure 1: t-SNE visualisation for plots in StoryDB clustered according to their tags. Figure shows centroids of the tags with higher variance across the dataset.

2.3 Tagging

We annotate the resulting stories using meta-information on categories from Wiki API[5]. For every plot, we list all translated categories marked in every language in which this plot is available. Then we search these category lists for substrings that include tags from the manually created list of tags[6]. This allows us to proved language-specific tags for every language, that are listed in `ID_lang_tag.tsv`. For example, Czech version of Black Night has tags `action; crime; drama; superhero; comics;` and `thriller`, while the same story in Persian has no tag `comics`, but has additional tags `neo-noir; psychological; epic;` and `screenplays;`.

File `ID_tag_average.tsv` includes the scores of the tags available for every story. The scores are calculated as follows: we count the number of times that a given tag is associated with a given story. Then we divide this number over the total number of languages in which the story is represented. The obtained space of tags could be useful for narrative exploration. Every story becomes a vector with every coordinate on the interval $[0, 1]$. Figure 1 shows a t-SNE visualisation of this space (Van der Maaten and Hinton, 2008) alongside the centroids of the more distinctive tags.

2.4 StoryDB

Figure 2 shows the relative size of the datasets in every language presented in StoryDB. English

[5]https://www.mediawiki.org/w/api.php?action=help&modules=query%2Bcategories

[6]The list of tags is published with the dataset.

heavily dominates followed by Italian, French, Russian, and German.

There are more than 20 languages that have three thousand or more stories available, including such languages as Finnish, Hungarian, or Persian. Table 1 summarises some of the resulting parameters of the obtained dataset.

Story DB	
Number of languages	42
Median # of stories in a language	2 772
Maximal # of stories in a language	63 756
Minimum # of stories in a language	568

Table 1: Some resulting parameters of the StoryDB.

3 Evaluation

We have used three modern transformer-based architectures for the evaluation:

- mBERT[7] (Devlin et al., 2018) — a multilanguage version of BERT;
- mDistilBERT[8] (Sanh et al., 2019) — a distilled version of multi-language BERT;
- XLM-Roberta[9] (Conneau et al., 2020) — a model that is two times larger than BERT in terms of the number of parameters.

These models are the most widely used multilanguage models to date. The results of the experiments are publicly available at Weights and Biases[10], see (Biewald, 2020). The evaluation was performed on ten largest languages in StoryDB, namely: English — 'en', French — 'fr', Italian — 'it', Russian — 'ru', German — 'de', Dutch — 'nl', Ukrainian — 'uk', Portuguese — 'pt', Polish — 'pl', and Spanish — 'es'.

We evaluated three tasks:

- Task A. Multilabel classification for tags on a multilanguage corpus of plots;
- Task B. Multilabel classification for tags in cross-lingual learning;
- Task C. Multilabel classification for tags in cross-lingual learning with a corpus of overlapping plots that occur in every language.

Let us now describe every task in detail.

[7]https://huggingface.co/bert-base-multilingual-cased

[8]https://huggingface.co/distilbert-base-multilingual-cased

[9]https://huggingface.co/xlm-roberta-base

[10]https://wandb.ai/altsoph/storydb _eval.task1
https://wandb.ai/altsoph/storydb_eval.task3
https://wandb.ai/altsoph/storydb_eval.task3

	Hamming Score	Multilabel Accuracy
mDistillBERT	0.47	0.31
mBERT	0.50	0.33
XLM-RoBERTa	0.50	0.33

Table 2: Task A. Hamming score and multilabel accuracy for the vector of predicted tags on a validation set. Training data consists of sixteen thousand plots in ten languages, with one tenth of the dataset in every language.

3.1 Task A

We have sampled the ten most frequents tags out of StoryDB (tag 'film' was the most frequent yet was excluded as a somewhat redundant one). These tags were: 'drama', 'comedy', 'television', 'fiction', 'series', 'action', 'thriller', 'black-and-white', 'science fiction', 'horror'. These ten tags form a vector, where every dimension corresponds to one particular tag. '1' encodes the presence of the tag and '0' stands for the absence of it.

For every language out of the top ten in StoryDB, we have sampled 2000 plots such that every plot has at least one tag out of the list of the ten most popular tags. In Task A the plots were sampled randomly for every language, so there is some overlap between languages. On average, 2% of the plots in one language reoccur in another one. It is important to note that the set of tags for a given plot might differ across languages and one plot could have several tags simultaneously. Thus, multilabel classification is a natural evaluation task under these circumstances.

Since the dataset is not balanced with respect to tags we used the binary cross-entropy loss[11] over the vector of tags. Table 2 and Table 3 sum up the results of three models on a multilanguage dataset of plots. Further details across languages and tags are available online[12].

3.2 Task B

Now let us do a similar setup yet train every model on one language in StoryDB and test its accuracy

[11]https://pytorch.org/docs/stable/generated/torch.nn.BCELoss.html

[12]https://wandb.ai/altsoph/storydb_eval.task1

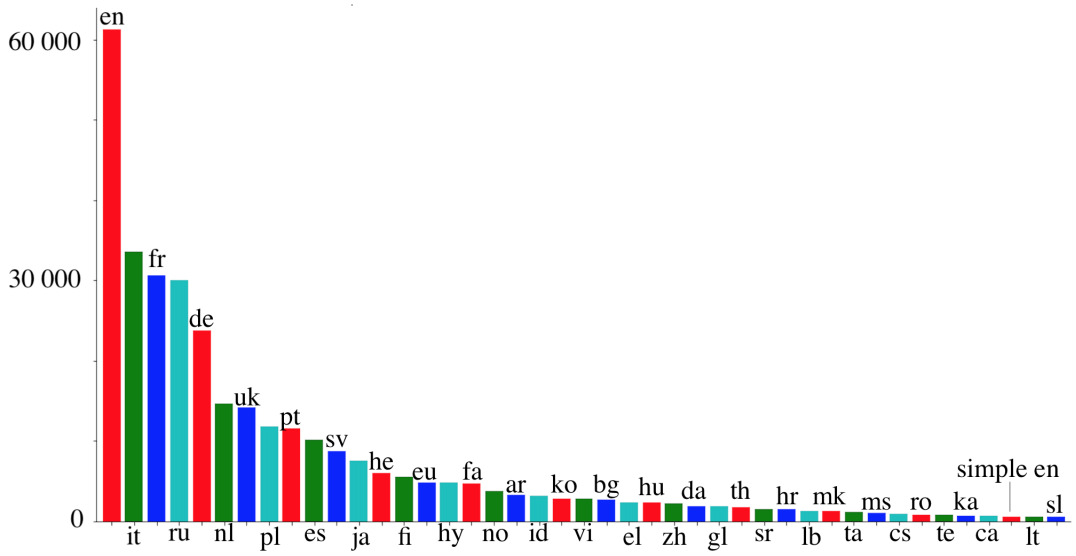

Figure 2: Number of stories in every language that has more that five hundred entries in StoryDB.

	mD-BERT	mBERT	XLM-R
Comedy	0.69	0.67	0.69
Action	0.67	0.70	0.67
Fiction	0.78	0.80	0.81
Thriller	0.67	0.63	0.64
Horror	0.70	0.76	0.75
Drama	0.73	0.74	0.74
Series	0.77	0.78	0.78
Television	0.74	0.76	0.76
Science Fiction	0.78	0.80	0.81
Black and White	0.68	0.65	0.62

Table 3: Task A. AUC-ROC for binary tag classifiers on a validation set. Training data consists of sixteen thousand plots in ten languages, with one tenth of the dataset in every language.

on another language. The parameters of the training datasets and labels are the same as in Task A above but every model is trained on one dataset and is then tested on other languages. Table 4 show the performance of mBERT, yet mDistillBERT and XLM-RoBERTa demonstrate similar behavior. The detailed results could be found online[13].

Table 4 demonstrates that if we train the model on one language and validate it on the other the quality of the multilabel tag classification drops. This drop varies across languages and tends to be smaller for the languages that belong to the same language family.

3.3 Task C

The last validation is similar to Task B, yet now we sample plots that overlap in every language. This limits us to 1500 plots in six languages that we split into train and test. Now every plot occurs in every language. Table 5 shows the model manages to recover certain tags in one language after the pre-training on the other. Table 5 shows the performance of XLM-RoBERTa, yet mDistillBERT and mBERT demonstrate similar behavior. The performance of the models tends to be better on overlapping plots if we compare it to Task B. The detailed results could be found online[14].

The multilabel accuracy for tag prediction declines further yet it can neither be attributed to specific lexical properties of a particular language nor any form of overlap of plots across languages.

[13] https://wandb.ai/altsoph/storydb_eval.task2
[14] https://wandb.ai/altsoph/storydb_eval.task3

	en	de	nl	fr	it	es	pt	ru	uk	pl
en	0.36	0.16	0.14	0.16	0.10	0.17	0.15	0.13	0.12	0.15
de	0.15	0.40	0.16	0.18	0.12	0.16	0.20	0.19	0.18	0.21
nl	0.20	0.33	0.41	0.20	0.22	0.32	0.29	0.31	0.25	0.30
fr	0.16	0.20	0.16	0.51	0.13	0.18	0.18	0.16	0.14	0.19
it	0.19	0.30	0.24	0.21	0.21	0.26	0.28	0.27	0.24	0.30
es	0.23	0.24	0.20	0.22	0.18	0.45	0.27	0.22	0.22	0.23
pt	0.15	0.21	0.17	0.22	0.10	0.19	0.44	0.19	0.14	0.23
ru	0.12	0.21	0.16	0.13	0.12	0.20	0.22	0.45	0.16	0.20
uk	0.10	0.20	0.16	0.14	0.09	0.19	0.19	0.23	0.25	0.20
pl	0.19	0.27	0.19	0.21	0.11	0.20	0.24	0.24	0.20	0.48

Table 4: Task B. Multilabel accuracy for the vector of predicted tags by mBERT. Training data consists of one thousand six hundred plots in one language. Every row shows validation accuracy of a model pretrained on the corresponding language and validated on the plots in a language from the corresponding column.

	en	de	nl	fr	it	es
en	0.29	0.16	0.12	0.12	0.10	0.08
de	0.27	0.31	0.18	0.19	0.15	0.11
nl	0.34	0.30	0.32	0.17	0.13	0.17
fr	0.22	0.20	0.16	0.27	0.15	0.10
it	0.34	0.29	0.23	0.25	0.25	0.16
ru	0.25	0.21	0.18	0.18	0.18	0.13

Table 5: Task C. Multilabel accuracy for the vector of predicted tags by XLM-RoBERTa across the dataset of plots withour cross-language overlaps. Training data consists of one thousand two hundred plots in one language. Every row shows validation accuracy of a model pretrained on the corresponding language and validated on the plots in a language from the corresponding column.

This series of evaluation tasks demonstrates two crucial properties of StoryDB:

- StoryDB could be used to work with narrative structures on the most abstract cross-lingual level;
- StoryDB allows controlling for various cross-lingual similarities of plots during ablation experiments with models of narrative.

4 Discussion

We believe that a broad multilanguage dataset of narratives can facilitate several areas of narrative research.

- Cross-cultural research of narrative structure. StoryDB provides possibilities to compare the structure of narrative in various languages. Since StoryDB includes every story in its original language and is equipped with a universal system of tags it is a natural source for such cross-cultural research.
- Classification of narratives. StoryDB includes an extensive amount of narratives for various languages alongside their genre tags. This allows to develop new methods for narrative classification as well as extensively test the ones that already exist, see for example (Reiter et al., 2014).
- Quantitative research of the narrative structure. (y Pérez, 2007) represents a story as a cluster of emotional links and tensions between characters that progress over storytime. StoryDB includes the description of the plots alongside the key characters. Such information could be insightful for a deeper quantitative understanding of narrative as a by-product of character interaction.
- Summarization of narrative. Parallel corpora in different languages contain similar descriptions of the narrative that could vary in terms of details and length. That makes StoryDB a useful tool for potential narrative summarization research such as (Barros et al., 2019).
- End-to-end narrative generation. StoryDB is the first dataset of narratives that we know of that contains narrative descriptions in various natural languages.

5 Conclusion

This paper presents StoryDB — a broad multilanguage dataset of narratives. We describe the construction of the dataset, provide the code for the

whole pipeline, list the parameters of the resulting dataset, and briefly discuss several areas of natural language processing research, where StoryDB could be useful for the community.

We hope that StoryDB could be broadened as more plot descriptions are added to various languages. These considerations make StoryDB a flexible resource that would be relevant for the NLP community as the subfield of quantitative narrative research moves on.

References

Yana Agafonova, Alexey Tikhonov, and Ivan P Yamshchikov. 2020. Paranoid transformer: Reading narrative of madness as computational approach to creativity. *Future Internet*, 12(11):182.

Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J Martin, and Mark O Riedl. 2020. Story realization: Expanding plot events into sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7375–7382.

Cristina Barros, Elena Lloret, Estela Saquete, and Borja Navarro-Colorado. 2019. Natsum: Narrative abstractive summarization through cross-document timeline generation. *Information Processing & Management*, 56(5):1775–1793.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Faeze Brahman, Alexandru Petrusca, and Snigdha Chaturvedi. 2020. Cue me in: Content-inducing approaches to interactive story generation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 588–597.

Mikhail S Burtsev and Grigory V Sapunov. 2020. Memory transformer. *arXiv preprint arXiv:2006.11527*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. 2019. Star-transformer. 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 1315–1325.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Sudipta Kar, Suraj Maharjan, A Pastor López-Monroy, and Thamar Solorio. 2018. Mpst: A corpus of movie plot synopses with tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ben Kybartas and Rafael Bidarra. 2016. A survey on story generation techniques for authoring computational narratives. *IEEE Transactions on Computational Intelligence and AI in Games*, 9(3):239–253.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Anastasia Malysheva, Alexey Tikhonov, and Ivan P Yamshchikov. 2021. Dyplodoc: Dynamic plots for document classification. *arXiv preprint arXiv:2107.12226*.

Lara J Martin, Srijan Sood, and Mark Riedl. 2018. Dungeons and dqns: Toward reinforcement learning agents that play tabletop roleplaying games. In *INT/WICED@ AIIDE*.

Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. 2016. Inscript: Narrative texts annotated with script information. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3485–3493.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. Mcscript2. 0: A machine comprehension corpus focused on script events and participants. In *Proceedings of the Eighth Joint Conference on Lexical*

and Computational Semantics (SEM 2019)*, pages 103–117.

Rafael Pérez y Pérez. 2007. Employing emotions to drive plot generation in a computer-based storyteller. *Cognitive Systems Research*, 8(2):89–109.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plotmachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295.

Michaela Regneri, Alexander Koller, and Manfred Pinkal. 2010. Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 979–988.

Nils Reiter, Anette Frank, and Oliver Hellwig. 2014. An nlp-based cross-document approach to narrative structure discovery. *Literary and Linguistic Computing*, 29(4):583–605.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, pages 1–20.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Judith van Stegeren and Mariët Theune. 2019. Narrative generation in the wild: Methods from nanogenmo. In *Proceedings of the Second Workshop on Storytelling*, pages 65–74.

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.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4306–4315.