Semantic Planning for Multilingual Fiction Generation

Nayla Escribano

HiTZ Basque Center for Language Technologies - Ixa NLP Group, University of the Basque Country UPV/EHU

Abstract
Recent approaches to fiction generation make use of generative Large Language Models to create fluent narratives, but they still struggle in other tasks, such as keeping coherence in rather long stories or respecting logical relations between events. Furthermore, most work focuses in English fiction, due to its prevalence in linguistic resources. In this PhD thesis we hypothesise that planning on semantic information may not only help models in such tasks, as stated by previous work, but also allow for fiction generation in other languages as an intermediate representation of stories. To that end, we collected a preliminary version of a dataset of high-quality human-written stories with extensive metadata. This new dataset will be used to test the influence of semantic planning on multilingual fiction generation and improve relevant story attributes such as coherence, logicality or likability in different languages.

Keywords
Fiction Generation, Multilingualism, Generative Models, Knowledge-based Methods

1. Introduction

Recent advances pushed by the development of generative Large Language Models (LLMs) make us wonder what are their true capabilities at generating fictional texts in different languages. Thus far, conditioning methods such as control mechanisms or plot-based planning have proven better coherence along generated stories than just prompting state-to-the-art generative LLMs [1, 2, 3, 4, 5]. Furthermore, this kind of conditioning usually aimes at imitating the writing process performed by professional writers or helping them at this task [3, 6, 4, 7], leading us to a more natural way of creating new stories. On the other hand, efforts to improve fiction generation rely on English-written data and, to the best of our knowledge, there are no proposals for multilingual fiction generation.

In this context, we define semantic planning as the creation of plans built upon semantic information (such as events, semantic roles, named entities, and temporal or causal relations) that enables guided surface realisation. Thus, for the fiction generation task, semantic planning should create skeletons based on semantic information to later build narrative texts. Inspired by recent work on cross-lingual transfer [8], we hypothesise that fiction generation in languages other than English can benefit from using semantic planning as multilingual bridges, given the lack of semantic and fiction resources in most languages.

Doctoral Symposium on Natural Language Processing from the Proyecto ILENIA, 28 September 2023, Jaén, Spain.

nayla.escribano@ehu.eus (N. Escribano)
0000-0003-3761-7947 (N. Escribano)

The present thesis project addresses the research question *How does semantic planning affect multilingual fiction generation?* To perform our experiments we created PromptStories[1], a large collection of up to 200 k high-quality human-written English prompt-story pairs with extensive metadata. Table 1 shows an example of a prompt-story pair from PromptStories. This dataset will allow us to test the capabilities of generative LLMs at creating narrative texts, study the influence of semantic planning as an intermediate representation of stories and analyse its effect on multilingual fiction generation.

Prompt	A group of men burst into your house dressed in what looks like Viking armour. In gruff voices, they inform you they are here to serve your dog who they believe is the reincarnation of Fenrir. Your dog is a four pound Chihuahua called Mr Wiggles.
Date	16/3/2019
Score	10318
Story	There was a polite knock at the front door. Drying my hands, I left the kitchen and slung the towel over my shoulder and opened the large inner front door and pushing the frenzied, barking Mr. Wiggles. On the other side of the screen door stood twenty or so people in strange armor. A tall man in chainmail, furs, and a rounded skullcap stepped forward. He spoke, but I did not understand a word he said. Mr. Wiggles jumped into view, resuming his wild, frenzied barking. They all immediately fell to one knee, crying out a single word in unison. “Fenrir!” I looked to them. Then to Mr. Wiggles, who was still barking. I looked back to them. “I uh. I don’t want any?” I closed the door. [...]
Date	17/3/2019
Score	1704

Table 1
Example of a prompt-story pair from PromptStories. “Date” refers to the posting date and “Score” to the result of upvotes minus downvotes.

2. Related Work

2.1. Story Datasets

Story collections are useful resources for training and evaluating fiction generation models. ROCStories [9] gathers 40 k English five-sentence stories full of causal and temporal commonsense relations from everyday life events to test story understanding, while [10] propose the CaTeRS scheme to annotate causal and temporal relations in 320 of these stories. Recently, the validation split of ROCStories has been translated by professionals to 10 other languages to create the multilingual XStoryCloze dataset, but it contains less than 2 k examples due to their experimental objectives [11]. [12, 4] prefer using plots from Wikipedia to learn how to generate coherent stories. However, stories collected in these datasets are not written in a natural narrative fashion, due to their specific purposes.

[1] This dataset is still in a preliminary phase. For the moment, a sample of the texts is available at https://github.com/ixa-ehu/PromptStories.

On the other hand, the STORIUM dataset [6] gathers 6 k stories that have been extensively annotated by collaboratively writing in a gamified framework, whereas StoryWars [7] offers 40 k stories extracted from another online collaborative storytelling platform to investigate different understanding and generation tasks. WritingPrompts [1] collects 300 k prompt-story pairs written in English by reddit-users[2] to train and evaluate story generation methods, but lacks relevant metadata related to the scraped texts that could inform us about the posting date, the quality perceived by users, and so on. Although [13] do consider such metadata to design a new story evaluation method, it was not possible to retrieve their dataset from the corresponding sites[3]. For this reason, we propose PromptStories as an extension of WritingPrompts that includes such information, gathering up to 200 k new prompt-story pairs.

Moreover, XStoryCloze is the only one of the previous datasets to present multilingual parallel stories, but this data is too scarce for our purposes. Although there exist multiple collections of fictional texts in several languages, these are usually too diverse in style, length and language coverage. PromptStories, on the contrary, contains a large amount of human-written short stories that we plan to translate from English to Spanish and Basque primarily, thus creating a large parallel story dataset.

2.2. Fiction Generation

Plot planning has been used since early attempts to fiction generation, such as the novel writer from [14], TALE-SPIN [15] or UNIVERSE [16]. These first approaches relied on hand-crafted rules to build plots from closed worlds of possible events. In order to overcome the limitations of manually created worlds, later work focused on statistical ways to extract possible events from existing stories [17, 18]. Nonetheless, the recent development of neural networks has motivated new methods to design plots for automatic fiction generation. These approaches usually make use of control mechanisms [19, 20, 1, 3, 4, 21, 22] and/or more fine-grained knowledge-based plot planning [12, 2, 6].

[1] test the capabilities of hierarchical control to maintain the relevance of generated stories to their corresponding prompts on the WritingPrompts dataset. In a later experiment, [2] show that applying Semantic Role Labelling (SRL) and coreference-based entity anonymization to decompose stories into action sequences and entity mentions improves the diversity and coherence of generated events and entities. In this work we propose to use the last approach, where event-based plot planning not only helps keeping coherence, but also allows for generating stories in different languages by translating the original stories and projecting their annotations.

2.3. Story Evaluation

As an open-ended generation task, the evaluation of human or automatic stories remains unsolved. Referenced metrics do not capture the complex characteristics of creative generation, where several outputs may correspond to a single input and many specific attributes could be considered (fluency, coherence, logicality, creativity, likability, etc.) [1, 6, 23, 24, 13]. To address this problem, some recent works try to evaluate stories using new unreferenced metrics, by

[2] https://www.reddit.com/r/WritingPrompts/

[3] We did neither receive an answer to our request.

selecting the appropriate story among machine-generated negative samples [23, 24] or assessing general quality after rating automatically created comments on concrete aspects about the story [13]. These techniques are backed by alleged correlation with human judgements, which are still the most reliable but expensive evaluation method. In this project we plan to test the last existing evaluation approaches and compare them with in-house human evaluations to later use them in our experiments.

3. Research Proposal

The **Main Research Question (MRQ)** of this thesis project is the following one: *How does semantic planning affect multilingual fiction generation?* To articulate the project work, we can divide this MRQ in smaller Research Questions (RQs). RQs present the same experimental structure (prepare the dataset, train models and evaluate them) and focus on different evaluation objectives.

3.1. RQ1: Are current models able to create good stories?

Testing the capabilities of current generative LLMs seems to be a proper initial step, given that they are widely used for story generation either in storytelling systems with control mechanisms or to create the surface realisation in those with knowledge-based plot planning. Indeed, RQ1 tackles the *fiction generation* part in MRQ. To answer this research question, we reduce fiction generation to short story generation (ranging ~100 to ~3000 words in length) as it constitutes a more manageable framework than working with larger fiction. On the other hand, we let humans answer the difficult question of *what is a good story?* by collecting a dataset of stories rated by their own readers. Thus, this dataset will allow us to evaluate the performance of state-of-the-art generative LLMs by comparing their zero-shot and fine-tuned results on a manual evaluation to check whether fine-tuning on stories rated as good by readers improves the story generation capabilities of these models.

3.2. RQ2: Does semantic planning improve specific story attributes?

Several fiction generation systems use knowledge-based methods to improve different attributes like intra-story coherence or logicality between events, among others. RQ2 mostly involves the *semantic planning* part of MRQ. For this reason, we will test different event representation schemes on a small sample of our dataset to find the most appropriate for story semantic planning, and we will use this scheme to annotate our prompt-story pairs. Then, we will prepare a human evaluation to compare stories generated by prompting our fine-tuned model from RQ1 and those from a model fine-tuned on our semantically-annotated dataset. Participants will be asked to evaluate specific story attributes, such as language fluency, intra-story coherence, logicality between events, creativity and relevance based on the prompt or likability from the user perspective.

3.3. RQ3: May this semantic planning help generating stories in languages other than English?

Finally, RQ3 involves the *multilingual* part of MRQ. Given the lack of parallel semantically-annotated corpora of short stories and the weak performance of semantic labelling in other languages (specially for under-resourced ones), we plan to translate our original English prompt-story pairs and project their annotations from RQ2 to the target languages. Because of evaluation availability, we wish to apply this translation and projection to Spanish first and, if feasible, to Basque. Similarly to the previous research questions, we will fine-tune multilingual generative LLMs both on raw and semantically annotated stories, and we will compare the human evaluations of these new models with the zero-shot setting. Furthermore, we will also analyse the accumulated error from this annotate-translate-and-project method.

4. Experimental Setup

4.1. Data Collection

We collected the data in PromptStories from the WritingPrompts subreddit as in [1] and filtered out undesired texts such as removed or moderator posts. Our preliminary dataset contains 200 k prompt-story pairs from 2019 to the beginning of 2023 along with relevant metadata like creation date, score received by users and so on, which allow us to select the best stories. The dataset has been split in train, dev and test sets (80%, 10% and 10% of prompt-story pairs each), but we may prepare smaller subdatasets to experiment on stories with a minimum score, the best *n* stories per a unique prompt, and so on.

4.2. Annotation, Translation and Projection

Before starting experiments on multilingual fiction generation, we need to create a common frame for all the languages that we wish to study. To that end, we will annotate our English dataset with semantic information in order to build event plans that represent stories as in [2]. Then, we will automatically translate our prompt-story pairs to the desired languages and use word-aligning projections to project those events to the translated texts. This annotate-translate-and-project techique has proven to be a solid method for cross-lingual sequence-labelling tasks in zero-resource settings and facilitates the annotation of under-resourced languages [25, 26, 8].

We wish to test different state-of-the-art models for each of these three subtasks by comparing their performance on a human evaluation of a small sample of prompt-story pairs. These are the considered models[4]:

- Event extraction: AllenNLP SRL [27] and AllenNLP SRL_BERT [28].
- Translation: M2M100 [29], NLLB200 [30] and DeepL[5].
- Projection: SimAlign [25] and AWSoME [26].

[4] For very large models such as M2M100 and NLLB200, we will compare the most appropriate versions according to our processing capabilities.

[5] https://www.deepl.com/

4.3. Fiction Generation

Once we have created our semantically annotated parallel story dataset, we can proceed to study our specific research questions. We will first test the performance of different state-of-the-art generative LLMs on the RQ1 zero-shot setting and choose the most appropriate one for story generation. The selected model will be used for analysing RQs 1-3 by fine-tuning it on raw and annotated prompt-story pairs for each of the studied languages. Following recent progress in developing generative LLMs, for this task we will consider LLaMA [31] or similar models and specific ones designed for storytelling, such as MPT-StoryWriter [32].

5. Conclusion

We present this thesis project on studying how semantic planning affects multilingual fiction generation. To that end, we propose an analysis frame based on previous work that also explores new tasks such as using planning on semantic information to act as a bridge at creating stories in languages other than English. We explain our future experiments to investigate the main research question and present a new dataset under development for these experiments.

Acknowledgments

Nayla Escribano is funded by the Basque Government PhD grant "Programa Predoctoral de Formación de Personal Investigador No Doctor del Departamento de Educación del Gobierno Vasco".

References

[1] A. Fan, M. Lewis, Y. Dauphin, Hierarchical neural story generation, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 889–898. URL: https://aclanthology.org/P18-1082. doi:10.18653/v1/P18-1082.

[2] A. Fan, M. Lewis, Y. Dauphin, Strategies for structuring story generation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2650–2660. URL: https://aclanthology.org/P19-1254. doi:10.18653/v1/P19-1254.

[3] D. Ippolito, D. Grangier, C. Callison-Burch, D. Eck, Unsupervised hierarchical story infilling, in: Proceedings of the First Workshop on Narrative Understanding, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 37–43. URL: https://aclanthology.org/W19-2405. doi:10.18653/v1/W19-2405.

[4] H. Rashkin, A. Celikyilmaz, Y. Choi, J. Gao, PlotMachines: Outline-conditioned generation with dynamic plot state tracking, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4274–4295. URL: https://aclanthology.org/2020.emnlp-main.349. doi:10.18653/v1/2020.emnlp-main.349.

[5] L. J. Martin, Neurosymbolic Automated Story Generation (Thesis Dissertation), Georgia Institute of Technology, 2021. URL: http://hdl.handle.net/1853/64643.

[6] N. Akoury, S. Wang, J. Whiting, S. Hood, N. Peng, M. Iyyer, STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6470–6484. URL: https://aclanthology.org/2020.emnlp-main.525. doi:10.18653/v1/2020.emnlp-main.525.

[7] Y. Du, L. Chilton, Storywars: A dataset and instruction tuning baselines for collaborative story understanding and generation, 2023. arXiv:2305.08152.

[8] I. García-Ferrero, R. Agerri, G. Rigau, Model and data transfer for cross-lingual sequence labelling in zero-resource settings, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6403–6416. URL: https://aclanthology.org/2022.findings-emnlp.478.

[9] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, J. Allen, 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, Association for Computational Linguistics, San Diego, California, 2016, pp. 839–849. URL: https://aclanthology.org/N16-1098. doi:10.18653/v1/N16-1098.

[10] N. Mostafazadeh, A. Grealish, N. Chambers, J. Allen, L. Vanderwende, CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures, in: Proceedings of the Fourth Workshop on Events, Association for Computational Linguistics, San Diego, California, 2016, pp. 51–61. URL: https://aclanthology.org/W16-1007. doi:10.18653/v1/W16-1007.

[11] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru, S. Shleifer, P. S. Koura, V. Chaudhary, B. O'Horo, J. Wang, L. Zettlemoyer, Z. Kozareva, M. Diab, V. Stoyanov, X. Li, Few-shot learning with multilingual generative language models, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 9019–9052. URL: https://aclanthology.org/2022.emnlp-main.616.

[12] L. Martin, P. Ammanabrolu, X. Wang, W. Hancock, S. Singh, B. Harrison, M. Riedl, Event representations for automated story generation with deep neural nets, Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018). URL: https://ojs.aaai.org/index.php/AAAI/article/view/11430. doi:10.1609/aaai.v32i1.11430.

[13] H. Chen, D. Vo, H. Takamura, Y. Miyao, H. Nakayama, StoryER: Automatic story evaluation via ranking, rating and reasoning, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 1739–1753. URL: https://aclanthology.org/2022.emnlp-main.114. doi:10.18653/v1/2022.emnlp-main.114.

[14] S. Klein, J. F. Aeschlimann, D. F. Balsiger, S. L. Converse, C. Court, M. Foster, R. Lao, J. D. Oakley, J. Smith, Automatic novel writing: A status report, 1973. URL: http://digital.library.wisc.edu/1793/57816.

[15] J. R. Meehan, Tale-spin, an interactive program that writes stories, in: Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI'77, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1977, p. 91–98.

[16] M. Lebowitz, Story-telling as planning and learning, Poetics 14 (1985) 483–502. doi:10.1016/0304-422X(85)90015-4.

[17] N. McIntyre, M. Lapata, Learning to tell tales: A data-driven approach to story generation, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Association for Computational Linguistics, Suntec, Singapore, 2009, pp. 217–225. URL: https://aclanthology.org/P09-1025.

[18] M. O. Riedl, Story planning: Creativity through exploration, retrieval, and analogical transformation, Minds & Machines 20 (2010) 589-–614. doi:https://doi.org/10.1007/s11023-010-9210-2.

[19] N. Peng, M. Ghazvininejad, J. May, K. Knight, Towards controllable story generation, in: Proceedings of the First Workshop on Storytelling, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 43–49. URL: https://aclanthology.org/W18-1505. doi:10.18653/v1/W18-1505.

[20] J. Xu, X. Ren, Y. Zhang, Q. Zeng, X. Cai, X. Sun, 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, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4306–4315. URL: https://aclanthology.org/D18-1462. doi:10.18653/v1/D18-1462.

[21] K. Yang, Y. Tian, N. Peng, D. Klein, Re3: Generating longer stories with recursive reprompting and revision, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 4393–4479. URL: https://aclanthology.org/2022.emnlp-main.296.

[22] K. Yang, D. Klein, N. Peng, Y. Tian, Doc: Improving long story coherence with detailed outline control, 2023. arXiv:2212.10077.

[23] J. Guan, M. Huang, UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 9157–9166. URL: https://aclanthology.org/2020.emnlp-main.736. doi:10.18653/v1/2020.emnlp-main.736.

[24] S. Ghazarian, Z. Liu, A. S M, R. Weischedel, A. Galstyan, N. Peng, Plot-guided adversarial example construction for evaluating open-domain story generation, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4334–4344. URL: https://aclanthology.org/2021.naacl-main.343. doi:10.18653/v1/2021.naacl-main.343.

[25] M. Jalili Sabet, P. Dufter, F. Yvon, H. Schütze, SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1627–1643. URL: https://aclanthology.org/2020.

findings-emnlp.147. doi:10.18653/v1/2020.findings-emnlp.147.

[26] Z.-Y. Dou, G. Neubig, Word alignment by fine-tuning embeddings on parallel corpora, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 2112–2128. URL: https://aclanthology.org/2021.eacl-main.181. doi:10.18653/v1/2021.eacl-main.181.

[27] L. He, K. Lee, M. Lewis, L. Zettlemoyer, Deep semantic role labeling: What works and what's next, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 473–483. URL: https://aclanthology.org/P17-1044. doi:10.18653/v1/P17-1044.

[28] P. Shi, J. Lin, Simple bert models for relation extraction and semantic role labeling, 2019. arXiv:1904.05255.

[29] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, A. Joulin, Beyond english-centric multilingual machine translation, 2020. arXiv:2010.11125.

[30] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No language left behind: Scaling human-centered machine translation, 2022. arXiv:2207.04672.

[31] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[32] M. N. Team, Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. URL: www.mosaicml.com/blog/mpt-7b.