Master Internship offer - Spring 2023
Personalized data-to-text neural generation

Laure Soulier, Christophe Gravier

1 Introduction

Information

Supervisors: laure.soulier@isir.upmc.fr, christophe.gravier@univ-st-etienne.fr
Localization: Saint-Étienne (LaHC) or Saclay (LISN), France
Duration: 6 months, between February and August 2023.
Stipend: around 573,30 euros / month[1]

Expected profile: Master or engineering degree in Computer Science or Applied Mathematics related to machine learning/natural language processing. The candidate should have a strong scientific background with good technical skills in programming, and be fluent in reading and writing English.

How to apply? Send a CV, a motivation letter and Master records to `laure.soulier@isir.upmc.fr` and `christophe.gravier@univ-st-etienne.fr`. Recommendation letters would be appreciated. Interviews will conducted as they arise and the position will be filled as soon as possible – the latest application date is set to 15th January.

Context

Based on prior works at Jacobs University Bremen in Germany and University of Montréal [2], a new novel neural architecture "transformer" (fully based on attention) had been devised in 2017 in a key paper from Google Brain [18]. The main idea of the attention mechanism is to alienate the limitations of training neural architecture for machine translation, that is the need to predict tokens until the $n-1$ one, in order to predict the $n-th$ word of a sequence (so-called recurrent networks) – thereby allowing parallel training on GPUs of (very) large NLP neural models. The attention mechanism removes the recurrent paradigm in the trained predictor, and instead try to learn the weights of surroundings tokens (i.e. word), depending on the token being processed at a given time. This paper is the building block of many NLP contributions nowadays (the "transformer" paper is cited $28,403$ as of September 2021!).

The transformer architecture led to very large language models such as BERT [5] or RoBERTa [6], which are able to solve tasks such as text classification [16], question answering [19], etc. A tremendously exciting task is text generation, that is the ability to leverage such language models to create NLP systems

[1] Standard internship stipend in France – Computed on Government Website: `https://www.service-public.fr/simulateur/calcul/gratification-stagiaire`, new law to be published that should make it higher in 2023 but actual figures are yet unknown.

that can generate free text – a long-lasting goal in the field of Artificial Intelligence. Among these models, GPT3 [3] is probably the most impressive and creative.

Besides common limitations of such systems [7, 14], a key observation is that the text is generated in a left-to-right fashion - which is called *auto-regressive*. It is therefore not trivial to control on the generator (ie. set constraints as presence/absence of a token for instance). It is even harder to control the way the model express itself, that is to say the style in which it should generate text. The major way to control this is actually to use existing style annotated corpus and create generative models that learn to perform style transfer [1, 4, 8] (the problem is therefore cast as a domain transfer issue). A critical issue being how to evaluate style transfer system for text generation [9, 17].

Objectives

In this internship we are interested in a special case of text generation, which is data-to-text generation. In this setting, the task is to generate sentences in natural language based on structured or semi-structured data. To provide a data to text example, a famous academic dataset is made of statistics of baseball games paired with human written summary of the game [11], that we ultimately want to the system to learn to generate. Beyond this toy example, data to text is of the utmost practical interests in many scenarios such as finance, . . . This task is a special case in text generation and comes with its own specific challenges. The data to text models are prone to hallucinations, that is generating grammatically correct but irrelevant and out of the blue sentences [12]. Moreover, the inputs being structured or semi-structured data, this calls for alternative solution to encode w.r.t. standard texts inputs made of sequence of tokens arranged as sentences.

The objective of the internship is to develop a neural data to text system able to personalize the text generation. Based on a previous work, we will first focus on a movie dataset in which we dispose of movie tabular description (the data), and reviews. The objective will be to personalize review for a given user. Secondly, while there exists studies on how to evaluate data to text generator [10, 13], to the best of our knowledge none consider style transfer/text personalization for text generation besides [15]. As such, finding means to perform style transfer evaluation for data to text generators is fully part of the internship, on top of finding neural solution to perform style transfer aware data to text. The evaluation we seek has to be automatic or semi automatic. For inspiration, a great example of a semi-automatic technique (for the task of summarisation and not data-to-text) is [20].

The workplan proposed to the student are as follows :

1. Literature review on data to text generation and author style transfer/personalization.
2. Become familiar with the work of the previous trainee, i.e. exploring the created dataset and the baselines already explored
3. Pursue the work by proposing novel models and enhancing the evaluation protocol.
4. Conduct experiments on the proposed solution and evaluation schemes with respect to baseline systems.
5. If the internship leads to publish work, we will provide support to go present your work in a conference.

Recommendation for applicants

If you want to know more about the direction of this research and this internship, you may consider reading first the following articles:

- On style transfer: Xiang Ao et al. "PENS: A Dataset and Generic Framework for Personalized News Headline Generation". In: *The Annual Meeting of the Association for Computational Linguistics (ACL)*. Aug. 2021. URL: https://www.microsoft.com/en-us/research/publication/pens-a-dataset-and-generic-framework-for-personalized-news-headline-generation/

- On evaluating style transfer: Remi Mir et al. "Evaluating Style Transfer for Text". 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 495–504. DOI: 10.18653/v1/N19-1049. URL: https://aclanthology.org/N19-1049

- On semi-automatic evaluation of text generators: Shiyue Zhang and Mohit Bansal. "Finding a Balanced Degree of Automation for Summary Evaluation". In: *The 2021 Conference on Empirical Methods in Natural Language Processing*. 2021

References

[1] Xiang Ao et al. "PENS: A Dataset and Generic Framework for Personalized News Headline Generation". In: *The Annual Meeting of the Association for Computational Linguistics (ACL)*. Aug. 2021. URL: https://www.microsoft.com/en-us/research/publication/pens-a-dataset-and-generic-framework-for-personalized-news-headline-generation/.

[2] Dzmitry Bahdanau et al. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[3] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: (2020). arXiv: 2005.14165 [cs.CL].

[4] Kunal Chawla et al. "Semi-supervised Formality Style Transfer using Language Model Discriminator and Mutual Information Maximization". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2340–2354. DOI: 10.18653/v1/2020.findings-emnlp.212. URL: https://aclanthology.org/2020.findings-emnlp.212.

[5] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[6] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL].

[7] Li Lucy et al. "Gender and Representation Bias in GPT-3 Generated Stories". In: *Proceedings of the Third Workshop on Narrative Understanding*. Virtual: Association for Computational Linguistics, June 2021, pp. 48–55. DOI: 10.18653/v1/2021.nuse-1.5. URL: https://aclanthology.org/2021.nuse-1.5.

[8] Eric Malmi et al. "Unsupervised Text Style Transfer with Padded Masked Language Models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 8671–8680. DOI: 10.18653/v1/2020.emnlp-main.699. URL: https://aclanthology.org/2020.emnlp-main.699.

[9] Remi Mir et al. "Evaluating Style Transfer for Text". 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 495–504. DOI: 10.18653/v1/N19-1049. URL: https://aclanthology.org/N19-1049.

[10] Laura Perez-Beltrachini et al. "Analysing Data-To-Text Generation Benchmarks". In: *Proceedings of the 10th International Conference on Natural Language Generation*. Santiago de Compostela, Spain: Association for Computational Linguistics, Sept. 2017, pp. 238–242. DOI: 10.18653/v1/W17-3537. URL: https://aclanthology.org/W17-3537.

[11] Laura Perez-Beltrachini et al. "Bootstrapping Generators from Noisy Data". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1516–1527. DOI: 10.18653/v1/N18-1137. URL: https://aclanthology.org/N18-1137.

[12] Clément Rebuffel et al. "A Hierarchical Model for Data-to-Text Generation". In: *Advances in Information Retrieval*. Ed. by Joemon M. Jose et al. Cham: Springer International Publishing, 2020, pp. 65–80. ISBN: 978-3-030-45439-5.

[13] Clément Rebuffel et al. "Data-QuestEval: A Referenceless Metric for Data to Text Semantic Evaluation". In: *arXiv preprint arXiv:2104.07555* (2021).

[14] Timo Schick et al. "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 2339–2352. DOI: 10.18653/v1/2021.naacl-main.185. URL: https://aclanthology.org/2021.naacl-main.185.

[15] Sandeep Subramanian et al. "Multiple-Attribute Text Style Transfer". In: *CoRR* abs/1811.00552 (2018). arXiv: 1811.00552. URL: http://arxiv.org/abs/1811.00552.

[16] Chi Sun et al. "How to fine-tune bert for text classification?" In: *China National Conference on Chinese Computational Linguistics*. Springer. 2019, pp. 194–206.

[17] Craig Thomson et al. "A Gold Standard Methodology for Evaluating Accuracy in Data-To-Text Systems". In: *Proceedings of the 13th International Conference on Natural Language Generation*. Dublin, Ireland: Association for Computational Linguistics, Dec. 2020, pp. 158–168. URL: https://aclanthology.org/2020.inlg-1.22.

[18] Ashish Vaswani et al. "Attention is all you need". In: *Proc. of NIPS 2017*. 2017, pp. 5998–6008.

[19] Wei Yang et al. "End-to-End Open-Domain Question Answering with BERTserini". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 72–77. DOI: 10.18653/v1/N19-4013. URL: https://aclanthology.org/N19-4013.

[20] Shiyue Zhang et al. "Finding a Balanced Degree of Automation for Summary Evaluation". In: *The 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.