Master Internship offer - Spring 2022
Personalized data-to-text neural generation

Laure Soulier, Christophe Gravier

1 Introduction

Information

Supervisors: laure.soulier@lip6.fr, christophe.gravier@univ-st-etienne.fr
Localization: Paris or Saint-Étienne, France
Duration: 6 months, between February and August 2022.
Stipend: 573.30 euros / month

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@lip6.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 [19]. 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 [7], which are able to solve tasks such as text classification [17], question answering [20], etc. A tremendously exciting task is text generation, that is the ability to leverage such language models to create NLP systems

\footnote{Standard internship stipend in France – Computed on Government Website: \url{https://www.service-public.fr/simulateur/calcul/gratification-stagiaire} new law to be published that should make it higher in 2022 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 [8, 15], 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, 9] (the problem is therefore cast as a domain transfer issue). A critical issue being how to evaluate style transfer system for text generation [10, 18].

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 [12], that we ultimately want to teach 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 [13]. 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 that is preserving or recreating different expression styles. We are first looking for neural based solution that can alter the text generator to be sensible to author phraseology. Secondly, while there exists studies on how to evaluate data to text generator [11, 14], to the best of our knowledge none consider style transfer for text generation besides [10]. 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 [21].

The workplan proposed to the student are as follows :

1. Literature review on data to text generation and author style transfer/personalization.
2. Devise neural solutions to encode author styles in data to text generators. A starting point is a prior work at LIP6 on linking two neural networks to perform both text classification profile learning [6].
3. Propose an evaluation scheme for style transfer data to text generation in a automatic or semi automatic way.
4. Conduct experiments on the proposed solution and evaluation schemes with respect to baseline systems. Both LIP6 and Lab. Hubert Curien have clusters of GPUs for you to access and run your evaluations.
5. If the internship leads to publish work, 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:


References


