

JaCoText: A Pretrained Model for Java Code-Text Generation

Jessica López Espejel, Mahaman Sanoussi Yahaya Alassan, Walid Dahhane, El Hassane Ettifouri

Abstract—Pretrained transformer-based models have shown high performance in natural language generation task. However, a new wave of interest has surged: automatic programming language generation. This task consists of translating natural language instructions to a programming code. Despite the fact that well-known pretrained models on language generation have achieved good performance in learning programming languages, effort is still needed in automatic code generation. In this paper, we introduce JaCoText, a model based on Transformers neural network. It aims to generate java source code from natural language text. JaCoText leverages advantages of both natural language and code generation models. More specifically, we study some findings from the state of the art and use them to (1) initialize our model from powerful pretrained models, (2) explore additional pretraining on our java dataset, (3) carry out experiments combining the unimodal and bimodal data in the training, and (4) scale the input and output length during the fine-tuning of the model. Conducted experiments on CONCODE dataset show that JaCoText achieves new state-of-the-art results.

Keywords—Java code generation, Natural Language Processing, Sequence-to-sequence Models, Transformers Neural Networks.

I. INTRODUCTION

When developing software, programmers use both natural language (NL) and programming language (PL). While the latter is the core component of every project, natural language is used to write documentation (ex: JavaDoc) to describe different classes, methods and variables. Documentation is usually written by experts and aims to provide a comprehensive explanation of the source code to every person who wants to use/develop the project.

In the last years, the automation of programming code generation from natural language has been studied using various techniques [1], [2], [3], [4] of artificial intelligence (AI). Leveraging AI increases programmers productivity because it helps them automatically generate code for simple tasks, while allowing them to tackle only the most difficult ones.

After the big success of Transformers Neural Network [5], it has been adapted to many Natural Language Processing (NLP) tasks such as question answering [6], [7], [8], text translation [9] and automatic summarization [10], [11]. Some of the most popular models are GPT [12], [13], BERT [6], BART [1], and T5 [14]. One of the main factors of success of these models is that they were trained on very large corpora. Recently, there has been an increasing interest in programming code generation. Therefore, the scientific community based its

research on proposing systems that are based on pretrained transformers. For instance, CodeGPT and GPT-adapted [15] are based on GPT2 [13], PLBART [1] is based on BART, and CoText [2] follows T5. Note that these models have been pretrained on bimodal data (containing both PL and NL) and on unimodal data (containing only PL).

Programming language generation is more challenging than standard text generation. This is because PLs contain stricter grammar [16] and syntactic [17] rules. Fig. 1 shows an example of an input sequence received by our model (in NL), the output of the model (in PL) and the target code (also called gold standard or reference code).

	nl - code
Source input	Text: sets the value of the maptype property Env: concode_field_sep MapType mapType concode_field_sep MapType getMapType
Gold standard code	<pre>void function (MapType arg0) { this . mapType = arg0 ; }</pre>
Output	<pre>void function (MapType arg) { this . mapType = arg ; }</pre>

Fig. 1 Example of a code generated by our model in comparison with the corresponding gold standard code

In this paper, we present JaCoText, a pretrained model based on Transformers [5]. First, we initialize our model from pretrained weights of CoText-1CC and CoText-2CC, instead of performing a training from scratch. Later, we conduct an additional pretraining step using data that belongs to a specific programming language (Java in our case). Moreover, unlike works that based their pretraining on CodeSearchNet [18] such as CodeBERT [19] and CoText [2], we use more java data in the pretraining stage of our model, as [13] and [14] have shown that Transformers neural network improves its performance significantly from increasing the amount of pretraining data. Furthermore, we carry out experiments to measure the impact of the input and output sequences length on code generation task. Finally, we test the unimodal data and study its impact on the model's performance. This study is crucial to evaluate the model in the pretraining stage.

We highlight our main findings in the state-of-the-art below:

- T5 has shown the best performance in language generation tasks.
- Models initialized from previous pretrained weights achieve better performance than models trained from scratch [15], [2].

- Models such as SciBERT [20], and BioBERT [21] have shown the benefits to pretrain a model using data related to a specific domain.
- Increased data implies better training performance [13], [14]. This finding is intuitive since a large and diversified dataset helps improving the model’s representation.
- The input and output sequence length used to train the model matters in the performance of the model [22].
- The objective learning used during the pretraining stage gives the model some benefits when learning the downstream tasks [14], [23], [24].

II. JACoTEXT

In this section, we describe the core component of JaCoText to achieve state-of-the-art results.

A. Fine-tuning

We fine-tune our models based on two criteria:

a) *Sequence Length*: After analyzing in detail the outputs generated by previous works, we observed that some of the code sequences produced by the models were incomplete compared to the target ones. Consequently, we tokenized the training and validation sets with SentencePiece model [25]. We then computed the largest sequence data, and used its length for both the inputs and the targets.

b) *Number of steps*: Since we increased the length of sequences in our model, we increased the number of fine-tuning steps. According to [14], a way to improve the model’s performance is by increasing the number of steps in the training.

We apply both criteria initializing the fine-tuning from CoText checkpoints 2CC and 1CC, respectively. CoText-1CC is pretrained on unimodal data (only code), and CoText-2CC is pretrained on bimodal data (both code and natural language). Results of these experiments are shown in Table III

B. Additional Pretraining

Authors of [14] made some important observations that we support in our work: (1) for some specific tasks, the way to improve the model’s performance is to pretrain it with a dataset that belongs to a specific domain, (2) additional pretraining can improve the performance of a model, (3) a low number of epochs when pretraining a model leads to higher scores in generation tasks. Besides, other works such as CodeGPT-adapted [15] and CoText [2] show that models initialized with pretrained weights achieve better results than models trained from scratch.

Based on the previous highlighted points, we carried out additional pretraining on unimodal java data (Fig. 2). We initialized JaCoText-B-1CC-PL and JaCoText-B-2CC-PL models from pretrained weights of CoText-1CC and CoText-2CC [2], respectively. We trained the previous models on only-code sequences. We follow the same procedure for both $T5_{base}$ and $T5_{large}$. The input of the encoder is a noisy Java code. The input of the decoder is the original Java code with one position offset.

Briefly, once the model is initialized from T5 weights (previously pretrained on C4 dataset), we further pretrain it on CodeSearchNet and our java dataset. Later, we use the final checkpoints to initialize the fine-tuning on CONCODE dataset.

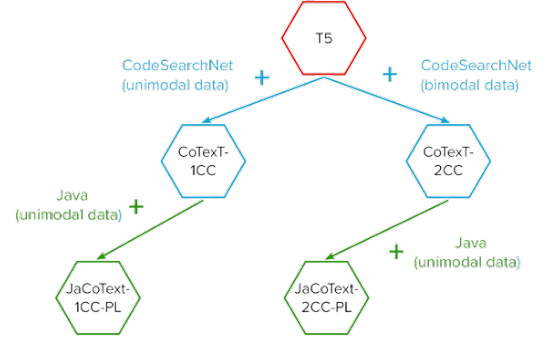


Fig. 2 JaCoText model, best viewed in color

III. EXPERIMENTAL SETUP

A. Architecture

JaCoText uses the same architecture as $T5$ [14], which is based on Transformers [5]. On the one hand, $T5_{base}$ consists of 12 layers in both the encoder and the decoder, with model dimension of 768 and 12 heads (approx. 220M parameters). On the other hand, $T5_{large}$ has 24 layers in both the encoder and the decoder, with model dimension of 1024 and 16 heads (approx. 770M parameters).

B. Code Generation Dataset

To perform our experiments in Java code generation task, we used CONCODE [26], a dataset that contains context of a real world Java programming environment. CONCODE aims to generate Java member functions that have class member variables from documentation. Table I describes CONCODE dataset.

TABLE I
A SUMMARY OF CONCODE DATASET

Category	Size		
	Train	Val	Test
# Text-Code lines	100K	2K	2K

C. Additional Pretraining Dataset

For the additional pretraining, we used our Java dataset. Originally, it consists of 812,008; 40,468, and 51,210 samples in the training, validation, and test sets, respectively. We deleted the problematic samples in the three sets (2974 in the training set, 235 in the validation set, and 161 in the test set). We use the rest of samples (900,316) from the three sets to pretrain our model.

D. Evaluation Metrics

To evaluate our models, we used the three metrics described below.

a) *BLEU*: [27] is a metric based on n-gram precision computed between the candidate and the reference(s). N-gram precision penalizes the model if: (1) there are words that appear in the candidate but not in any of the references, or (2) if a word appears more times in the candidate than in the maximum reference count. However, the metric fails if the candidate does not have the appropriate length. Following [1] and [2] we use the corpus-level BLEU score in the code generation task.

b) *CodeBLEU*: [28] works via n-gram match, and it takes into account both the syntactic and semantic matches. The syntax match is obtained by matching between the code candidate and code reference(s) sub-trees of abstract syntax tree (AST). The semantic match considers the data-flow structure.

c) *Exact Match (EM)*: is the ratio of the number of predictions that match exactly any of the code reference(s).

E. Baselines

We compare our model with four state-of-the-art Transformer-based models.

- CodeGPT, CodeGPT-adapted [15] are based on GPT-2 model [13]. The difference between both models is that CodeGPT is trained from scratch on CodeSearchNet dataset [18], while CodeGPT-adapted is initialized from GPT-2 pretrained weights.
- PLBART [1] uses the same architecture than $BART_{base}$ [29]. Additionally, PLBART uses three noising strategies: token masking, token deletion and token infilling.
- CoText [2] uses the same architecture than $T5_{base}$. It is trained on both unimodal and bimodal data using CodeSearchNet Corpus [18], and GitHub Repositories.

IV. RESULTS AND DISCUSSION

Firstly, we study the performance of T5 model on the Java generation task. We directly fine-tune on CONCODE dataset three types of T5: $T5_{base}$, $T5_{large}$, and $T5_{3B}$. The best parameters we used are highlighted in Table III

TABLE II
RESULTS OBTAINED WHEN FINE-TUNING DIRECTLY FROM T5 MODELS

Parameter # steps	Metrics		
	BLEU	EM	CodeBLEU
T5-base			
45000	34.03	20.45	36.73
60000	34.08	20.30	37.00
T5-Large			
45000	34.00	20.30	36.98
60000	36.23	21.05	38.84
T5-3B			
45000	32.65	21.60	35.47
60000	35.68	21.65	38.37
90000	36.28	22.50	38.97
120000	38.11	22.20	40.81

Best results are in bold.

Table II provides the scores of each type of T5 models directly after the fine-tuning using CONCODE dataset. In all cases, the score improves as the number of steps increases.

Unsurprisingly, the most sophisticated $T5_{3B}$ model gets the best results, followed by $T5_{large}$ and $T5_{base}$, while $T5_{3B}$ takes more time to converge.

Table III provides results obtained when varying the number of steps and the length of input and output sequences while fine tuning CoText-2CC and CoText-1CC checkpoints on CONCODE dataset. Results show that using 60000 steps provides better results than using 45000 steps in the fine-tuning as noted in [2]. In addition, by using the largest code sequence length, we outperform the BLEU and EM scores obtained by [2] (highlighted in *italic*). Results vary slightly, almost undetectable. However, CoText-1CC performs better using BLEU and CodeBLEU, while CoText-2CC achieves better results using the EM metric.

Varying the number of steps and augmenting the length of both the input and target in the fine-tuning provide the first step to improve results on Java code generation task. The second step consists in using the additional pretraining from CoText weights following [15]. After additional pretraining, we fine-tune the model using the best parameter values from Table III

Table IV provides fine-tuning results after performing the additional pretraining using our Java dataset. The models are initialized with JaCoText-B weights when they are trained following the $T5_{base}$ architecture, and with JaCoText-L weights when they are trained following $T5_{large}$. As we mentioned previously, the additional training using our Java dataset is initialized from CoText weights. However, the training of JaCoText-L-1CC-PL and JaCoText-L-2CC-PL models started from $T5_{large}$ weights (previously trained on C4 [14] dataset). We trained $T5_{large}$ on CodeSearchNet dataset, and later on our Java dataset during 200,000 steps each and using unimodal data (PL only). Finally, we fine-tune the model on CONCODE dataset for 45,000 steps.

Results show that JaCoText achieves state-of-the-art results. Unsurprisingly, JaCoText-L models get the highest scores using the three metrics, because $T5_{large}$ has a more sophisticated architecture. In addition, it is noteworthy to mention that in both architectures, *base* and *large*, the best results are obtained with models that were pretrained on bimodal data. This finding proves that training models with bimodal data performs better than with unimodal data.

Finally, Fig. 3 shows the improvements of our model JaCoText-B-2CC-PL with an additional training using our Java dataset. For a fair comparison, the three models are fine-tuned for 60,000 steps, and they all follow the $T5_{base}$ architecture.

V. RELATED WORK

Early interesting approaches mapped natural language to source code using regular expressions [31] and database queries [32], [33]. Most recently, neural networks have proven their effectiveness to automatically generate source code from different general-purpose programming languages like Python [17] and Java [2]. Simultaneously, large-scale datasets have surged in order to facilitate tackling the problem. These datasets include CONCODE [26], CONALA [34], and CodeSearchNet [18].

TABLE III
RESULTS WHEN VARYING THE NUMBER OF STEPS, AND THE INPUT AND OUTPUT SEQUENCE LENGTH DURING CoText-2CC AND CoText-1CC FINE-TUNING

Parameters		CoText 2CC / 1CC		
input / target length	# steps	BLEU	EM	Code BLEU
256 / 256	45000	36.51 / 37.40	20.10 / 20.10	39.49 / 40.14
256 / 256	60000	36.22 / 36.00	20.85 / 20.20	38.88 / 38.62
256 / 379	60000	36.60 / 36.45	20.10 / 20.10	39.23 / 39.20
379 / 379	45000	37.08 / 37.33	21.50 / 21.25	39.80 / 39.85
379 / 379	60000	37.46 / 37.66	21.45 / 21.40	39.94 / 40.03
200 / 200	45000	34.61 / 34.79	19.65 / 19.30	37.64 / 37.60
200 / 200	60000	35.17 / 35.23	19.10 / 18.20	38.12 / 38.19

TABLE IV
RESULTS WITH ADDITIONAL PRETRAINING USING OUR JAVA DATASET

Model	BLEU	EM	CodeBLEU
[30]	24.40	10.05	29.46
CodeGPT	28.69	18.25	35.52
CodeGPT-Adp	32.79	20.10	35.98
PLBART	36.69	18.75	38.52
T5-base	32.74	18.65	35.95
CoText-2CC	36.51	20.10	39.49
CoText-1CC	37.40	20.10	40.14
JaCoText-B-1CC-PL	38.65	21.85	41.19
JaCoText-B-2CC-PL	39.07	22.15	41.53
JaCoText-L-1CC-PL	39.67	22.30	42.19
JaCoText-L-2CC-PL	39.87	22.45	42.49

Best results are in bold.

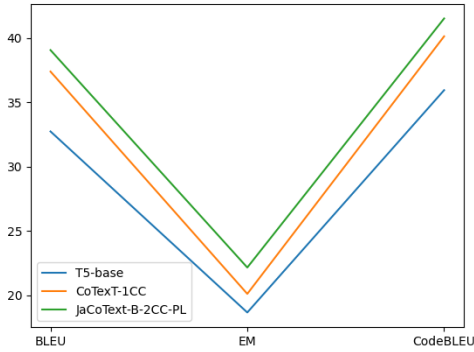


Fig. 3 Improvement of our model through the additional training

Reference [17] used a BiLSTM encoder, and an RNN decoder to generate syntactically valid parse trees. Inspired by the grammar-aware decoder, [26] used Bi-LSTMs encoder to compute the contextual representations of the NL, and an LSTM-based RNN decoder with two-step attention mechanism followed by a copying mechanism to map NL with the source code.

Recently, models based on Transformers [5] and originally intended for the generation of natural language have been of a great benefit for automatic code generation. PLBART uses the same model architecture as $BART_{base}$ [29]. Unlike $BART_{base}$, PLBART stabilizes the training by adding a normalization layer on the top of both the encoder and the decoder, following [35]. Similarly to PLBART, CoText (Code and Text Transfer Transformer) [2] is an encoder-decoder model, and it follows $T5_{base}$ [14] architecture.

Moreover, encoder-only models such as RoBERTa-(code) [15] inspired by RoBERTa [36], and decoder-only models like

CodeGPT and CodeGPT-adapted have achieved competitive results in the state of the art. Similarly to CodeGPT and CodeGPT-adapted, RoBERTa-(code) is pretrained on CodeSearchNet dataset. Unlike RoBERTa-(code), CodeGPT is pretrained on CodeSearchNet from scratch, and CodeGPT-adapted is pretrained starting from pretrained weights of GPT-2 [13]. Both CodeGPT and CodeGPT-adapted follow the same architecture and training objective of GPT-2.

VI. CONCLUSION

We present JaCoText, a set of T5-based [14] pretrained models designed to generate Java code from natural language. We evaluate the performance of three architectures: $T5_{base}$, $T5_{large}$, and $T5_{3B}$ to generate Java code. We follow the recommendations proposed by [2], [14], [15] to improve the performance of T5 model on Java code generation. Some takeaways from these experiments are: (1) pretraining the model using a dataset designed to tackle a specific task is beneficial, (2) additional pretraining can improve the performance of the model, and (3) using a low number of epochs in the pretraining helps improving the final performance.

Our models achieve state-of-the-art results on the Java code generation task. We prove that, each modification in our models, such as the additional training, allows JaCoText to have better comprehension of the java programming language. In the future, it would be interesting to explore other neural network models performance, and improve the programming language syntax through the decoding algorithm. In addition, since in this paper we focus our work on additional training using code only, we leave additional training using bimodal data for future work.

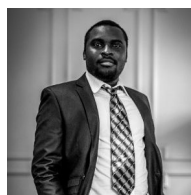
REFERENCES

- [1] Wasi Ahmad and Saikat Chakraborty and Baishakhi Ray and Kai-Wei Chang, *Unified Pretraining for Program Understanding and Generation*. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021.
- [2] Long Phan and Hieu Tran and Daniel Le and Hieu Nguyen and James Annibal and Alec Peltokian and Yanfang Ye, *CoText: Multi-task Learning with Code-Text Transformer*. Proceedings of the 1st Workshop on Natural Language Processing for Programming, 2021.

- [3] Daya Guo and Shuo Ren and Shuai Lu and Zhangyin Feng and Duyu Tang and Shujie Liu and Long Zhou and Nan Duan and Alexey Svyatkovskiy and Shengyu Fu and Michele Tufano and Shao Kun Deng and Colin B. Clement and Dawn Drain and Neel Sundaresan and Jian Yin and Daxin Jiang and Ming Zhou, *GraphCodeBERT: Pre-training Code Representations with Data Flow*. 9th International Conference on Learning Representations, 2021.
- [4] Xu Frank F. and Alon Uri and Neubig Graham and Hellendoorn Vincent Josua, *A Systematic Evaluation of Large Language Models of Code*. Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming, 2022.
- [5] Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N Gomez and Lukasz Kaiser and Illia Polosukhin, *Attention is All you Need*. Advances in Neural Information Processing Systems, 2017.
- [6] Jacob Devlin and Ming-Wei Chang and Kenton Lee and Kristina Toutanova, *BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 2018.
- [7] Daniel Khashabi and Snigdha Chaturvedi and Michael Roth and Shyam Upadhyay and Dan Roth, *Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences*. NAACL, 2018.
- [8] Christopher Clark and Kenton Lee and Ming-Wei Chang and Tom Kwiatkowski and Michael Collins and Kristina Toutanova, *BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 2019.
- [9] Yasmin Moslem and Rejwanul Haque and John Kelleher and Andy Way, *Domain-Specific Text Generation for Machine Translation*. arXiv, 2022.
- [10] Jingqing Zhang and Yao Zhao and Mohammad Saleh and Peter J. Liu, *PEGASUS: Pretraining with Extracted Gap-sentences for Abstractive Summarization*, 2019.
- [11] Peter J. Liu and Yu-An Chung and Jie Ren, *SummAE: Zero-Shot Abstractive Text Summarization using Length-Agnostic Auto-Encoders*, 2019.
- [12] Alec Radford and Karthik Narasimhan, *Improving Language Understanding by Generative Pretraining*. 2018.
- [13] Alec Radford and Jeffrey Wu and Rewon Child and David Luan and Dario Amodei and Ilya Sutskever, *Language Models are Unsupervised Multitask Learners*. 2019.
- [14] Colin Raffel and Noam Shazeer and Adam Roberts and Katherine Lee and Sharan Narang and Michael Matena and Yanqi Zhou and Wei Li and Peter J. Liu, *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Journal of Machine Learning Research, 2020.
- [15] Shuai Lu and Daya Guo and Shuo Ren and Junjie Huang and Alexey Svyatkovskiy and Ambrosio Blanco and Colin B. Clement and Dawn Drain and Daxin Jiang and Duyu Tang and Ge Li and Lidong Zhou and Linjun Shou and Long Zhou and Michele Tufano and Ming Gong and Ming Zhou and Nan Duan and Neel Sundaresan and Shao Kun Deng and Shengyu Fu and Shujie Liu, *CodeXGLUE: A Machine Learning Benchmark Dataset for Code Understanding and Generation*. CoRR, 2021.
- [16] Maxim Rabinovich and Mitchell Stern and Daniel Klein *Abstract Syntax Networks for Code Generation and Semantic Parsing*, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1), 2017.
- [17] Pengcheng Yin and Graham Neubig, *A Syntactic Neural Model for General-Purpose Code Generation*. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 2017.
- [18] Hamel Husain and Ho-Hsiang Wu and Tiferet Gazit and Miltiadis Allamanis and Marc Brockschmidt, *CodeSearchNet Challenge: Evaluating the State of Semantic Code Search*. CoRR, 2019.
- [19] Zhangyin Feng and Daya Guo and Duyu Tang and Nan Duan and Xiaocheng Feng and Ming Gong and Linjun Shou and Bing Qin and Ting Liu and Daxin Jiang and Ming Zhou, *CodeBERT: A Pretrained Model for Programming and Natural Languages*, 2020.
- [20] Iz Beltagy and Kyle Lo and Arman Cohan, *SciBERT: A Pretrained Language Model for Scientific Text*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.
- [21] Jinhyuk Lee and Wonjin Yoon and Sungdong Kim and Donghyeon Kim and Sunkyu Kim and Chan Ho So and Jaewoo Kang, *BioBERT: a pretrained biomedical language representation model for biomedical text mining*. Bioinformatics, 2020.
- [22] Iz Beltagy and Matthew E. Peters and Arman Cohan, *Longformer: The Long-Document Transformer*. arXiv:2004.05150, 2020.
- [23] Kaitao Song and Xu Tan and Tao Qin and Jianfeng Lu and Tie-Yan Liu, *MASS: Masked Sequence to Sequence Pretraining for Language Generation*. International Conference on Machine Learning, 2019.
- [24] Luca Di Liello and Matteo Gabburo and Alessandro Moschitti, *Efficient pretraining objectives for Transformers*, 2021.
- [25] Taku Kudo and John Richardson, *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.
- [26] Srinivasan Iyer and Ioannis Konstas and Alvin Cheung and Luke Zettlemoyer, *Mapping Language to Code in Programmatic Context*. EMNLP, 2018.
- [27] Kishore Papineni and Salim Roukos and Todd Ward and Wei-Jing Zhu, *Bleu: a Method for Automatic Evaluation of Machine Translation*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002.
- [28] Shuo Ren and Daya Guo and Shuai Lu and Long Zhou and Shujie Liu and Duyu Tang and M. Zhou and Ambrosio Blanco and Shuai Ma, *CodeBLEU: a Method for Automatic Evaluation of Code Synthesis*, 2020.
- [29] Mike Lewis and Yinhan Liu and Naman Goyal and Marjan Ghazvininejad and Abdelrahman Mohamed and Omer Levy and Veselin Stoyanov and Luke Zettlemoyer, *BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation, Translation, and Comprehension*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
- [30] Daya Guo and Duyu Tang and Nan Duan and Ming Zhou and Jian Yin, *Coupling Retrieval and Meta-Learning for Context-Dependent Semantic Parsing*. Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL), 2019.
- [31] Nicholas Locascio and Karthik Narasimhan and Eduardo DeLeon and Nate Kushman and Regina Barzilay, *Neural Generation of Regular Expressions from Natural Language with Minimal Domain Knowledge*. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016.
- [32] Xiaojun Xu and Chang Liu and Dawn Song, *SQLNet: Generating Structured Queries From Natural Language Without Reinforcement Learning*, 2017.
- [33] Victor Zhong and Caiming Xiong and Richard Socher, *Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning*, 2017.
- [34] Pengcheng Yin and Bowen Deng and Edgar Chen and Bogdan Vasilescu and Graham Neubig, *Learning to Mine Aligned Code and Natural Language Pairs from Stack Overflow*. Association for Computing Machinery, 2018.
- [35] Yinhan Liu and Jiatao Gu and Naman Goyal and Xian Li and Sergey Edunov and Marjan Ghazvininejad and Mike Lewis and Luke Zettlemoyer, *Multilingual Denoising Pretraining for Neural Machine Translation*. Transactions of the Association for Computational Linguistics, 2020.
- [36] Yinhan Liu and Myle Ott and Naman Goyal and Jingfei Du and Mandar Joshi and Danqi Chen and Omer Levy and Mike Lewis and Luke Zettlemoyer and Veselin Stoyanov, *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. CoRR, 2019.



Jessica López Espejel is a deep learning researcher at Novelis Research and Innovation Lab. Her research is focused on Automatic Code Generation and Transformers Neural Networks. She holds a Ph.D. in Natural Language Processing from Sorbonne Paris Nord University and CEA-LIST (2021). Email: jlopezespejel@novelis.io.



Mahaman Sanoussi Yahaya Alassan works as a researcher at Novelis Research and Innovation Lab. His research focuses on semantic text classification, information retrieval, named entity extraction. He obtained his Ph.D. in NLP from Paris Nanterre University in 2017. Email: syahaya@novelis.io



Walid Dahhane is the CTO and Co-founder of Novelis. He is an Enterprise architect, a specialist in microservices and Smart Automation architectures, and a doctor in AI & NLP. He is in charge of the IS Urbanisation and Cybersecurity and manages the activities around the business solutions. Email: wdahhane@novelis.io



El Hassane Ettifouri holds a Ph.D. in software engineering and Artificial Intelligence. He is an Associate and the Head of Novelis Research and Innovation Lab. His research focuses on Artificial Intelligence, Natural Language Processing, and Computer Vision. He was professor in ENSAO and SupMTI engineering schools, and was also the founder of the ZeroCouplage Framework. Email: eettifouri@novelis.io