High Quality Neural Machine Translation

BNP Paribas CIB AI Lab develops its own automated translation engine: come & help us build & improve one of the most iconic AI tools today! The tool is available and used by more than 200 000 employees within the group providing high quality translation that can be finetuned to their needs.

Since 2015, almost all the translation competition winners used neural machine translation algorithms, which became the state-of-the-art in almost all the tasks of the field.

The data science team developed a deep expertise on these state-of-the-art architectures [1][2]. However, they come with some drawbacks [3]:

- They are **highly sensitive to data quality, computationally intensive** and **less performant on out-of-domain data**

The goal of the internship will be to challenge and adapt Google's latest advances of their translation system [4] into the BNPP pipeline:

- **Simplifying the Transformer architecture** by using a RNN Decoder [5]
- **Improving data quality** using a Seq2Seq based denoiser fine-tuned on high quality data [6]

You will also be involved in our internal research, building specialized model for our internal clients and adapting approaches that do not require parallel data:

- Using **backtranslation** that is a common technique to **generate synthetic data** based on monolingual corpora [7]
- **Gathering specialized data** from a large set of crawled data using multilingual sentence embedders like **LASER** [8]
- Using **reranking systems** allowing to provide the **translation that can be ranked** according to a business need [9]

The internship will allow you to develop cutting edge natural language processing and deep learning processes. The target is to improve our current models in production.

References:

[1] Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, Wu & AL, CoRR, vol. abs/1609.08144 (2016)

[2] Attention Is All You Need, Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, NIPS 2017

[3] Six Challenges for Neural Machine Translation, Koehn P. & Knowles R., 2017 In Proceedings of the First Workshop on Neural Machine Translation.

[4] https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html

[5] Accelerating Transformer Decoding via a Hybrid of Self-attention and Recurrent Neural Network, Chengyi Wang, Shuangzhi Wu, Shujie Liu

[6] Denoising Neural Machine Translation Training with Trusted Data and Online Data Selection, Wei Wang, Ta ro Watanabe, Macduff Hughes, Tetsuji Nakagawa, Ciprian Chelba

[7] Understanding Back-Translation at Scale Sergey Edunov, Myle Ott, Michael Auli, David Grangier

[8] https://github.com/facebookresearch/LASER

[9] Discriminative Reranking for Neural Machine Translation, Ann Lee Michael Auli Marc'Aurelio Ranzato

Practical Information:
Supervisor: MATHEY Alexis, Data Scientist, alexis.mathey@bnpparibas.com
Location: Grands Moulins de Pantin, 9 Rue du Débarcadère, 93500 Pantin (Possible partial work from home)
Compensation: Competitive salary
How to apply: Please submit your application at paris.cib.analytics.consulting.careers@bnpparibas.com