STUDENT INTERNSHIP

Topic: Scientific and clinical data extraction for therapeutic target identification
Duration: 3 to 6 months start on February
Location: Oncodesign HQ – Dijon
Benefits: Monthly indemnity + meal Ticket

Our Company

OPM is a technological company specialized in precision medicine. OPM's mission is to bring innovative therapeutic and diagnostic solutions to treat therapeutic resistance and metastasis evolution. The patient is at the center of our reflection, of our unique innovative model, and our investments. For OPM "our collective success is paramount", there can be no value creation without exchange, without dialogue. The value creation resulting for us from reciprocity, i.e. balanced and fair exchanges at all levels, whether between internal collaborators, or with our partners, therapists, patients, experts and investors.

Context

In the past decades, the volume of scientific research publications has exponentially increased. For instance, during the last 2 years (2020 - up to this date), more than 500,000 scientific articles about oncology have been listed in the PubMed search engine. Moreover, in the last years, researchers have strongly increased the submission to pre-print repositories. this deluge of information is leading to information overload, which creates serious problems as research nowadays requires thinking and deciding faster than ever before; however, with this information profusion, this is now impossible.

Natural language processing (NLP) has gained popularity with the adoption of neural networks, the progress in language comprehension, text summarization and has been driven by a new generation of models. In recent years, transformer and attention-based architectures have beaten previous state-of-the-art results. Recent work on BERT's domain adaptation for the scientific NLP includes SciBERT [1] (trained on a 1.2 million Scopus article), BIOBERT (trained on PubMed abstract and free articles) [2], and CLINICALBERT (on clinical text) [3]. Different pre-trained and benchmark datasets are now available, allowing the development of a possible application. However, to our knowledge, no models have incorporated scientific, biological, and clinical text.

As part of our research and therapeutic target validation activities, we are currently pursuing our knowledge building efforts within our OncoSNIPER platform. To select the ideal target candidate, many different information is required (clinical information, biological information, potential side-effects) which have to be gathered and analyzed in order to prioritize a potential target. However, extract relevant information from large data source is not a trivial task and artificial intelligence approaches are required.

We are currently developing and evaluating deep learning models that are conducing different NLP tasks: text summarization, speech-to-text, named entity extraction and relation extraction. Our work is focused in extract relevant information for therapeutic target discovery and this internship is focused in developing pipeline for extract information from large corpora of text

18 rue Jean Mazen
21000 Dijon France
Tél. +33 (0)380 788 260
info@oncodesign.com
SA - Capital social : 2 343166,00 € - SIRET : 89222676200011 - RCS Dijon : 892226762 - Code APE : 7211Z - TVA intracommunautaire : FR24892226762

Missions & activities of the internship

- Evaluating, adapting and implementing different models on text summarization, speech-to-text, named entity extraction and relation extraction on scientific and clinical text.
- Fine-tuning of the models on different sources
- Development of different strategies of evaluation of NLP models (quality, accuracy, potential bias)
- Interfacing and integrating these models into OncoSNIPER to populate the integrated graph database.

Keywords

Python, NLP, deep learning

Student expected background/Knowledge

M2 or last year into engineer school with specialty/knowledge in Computer Science / Bioinformatics / Statistics Biology with knowledge in programming (R/Python). Knowledge and/or 1st experience on NLP with big data is a must.

References

1. Beltazy I. et al. *SciBERT: A Pretrained Language Model for Scientific Text*. arXiv. 2019
2. Lee J. *BioBERT: a pre-trained biomedical language representation model for biomedical text mining*. arXiv. 2019
3. Alsentzer E. *Publicly Available Clinical BERT Embeddings*. arXiv. 2019

How apply?

Contact: Thierry Billoué – Chief Human Resources Officer – Oncodesign Precision Medicine
Send your application (resume & motivation letter) under ref "TheraTargID" to tbilloue@oncodesign.com