Real-time Passage Retrieval System

Passage retrieval is a common Natural Language Processing (NLP) task, which consists of proposing a list of ordered passages from a large corpus, which answer the given query, expressed in natural language.

Classical IR approaches such as **TF-IDF** and **BM25** [1] work at syntax-level and often fail to capture the vocabulary misalignment between query and passages. **While Learning to Rank** [2] leverages machine-learning techniques to address this issue, their models are shallow and fitted over handcrafted IR features.

With the development of pre-trained transformer-based Language Models (LM) such as T5 [3], it is possible to employ large-scale neural models to bridge the gap. On the **MS MARCO** dataset [4] for **Passage Retrieval task**, these LM-based methods are unarguably dominating the leaderboard.

Most successful solutions adapt a “funnel” approach, where

- a **retriever** searches efficiently in a large corpus and returns a reasonable number of candidates with high recall,
- a **ranker** then scores the candidates with more elaborated (and time-consuming) models, usually based on language models

The entire system must take no more than **hundreds of milliseconds** to be useful in production.

Techniques such as **Question Generation** [5] can also be used to further boost the performance.

As part of the CIB Analytics Consulting team, your role will be to **implement a state-of-the-art LM-based IR system addressing the following issues**

1. Most SOTA models use extremely large models and/or ensemble, which could be slow and thus not production ready
2. Our internal datasets are intrinsically **close-domain**, whereas most SOTA models are evaluated on Wikipedia-like corpus which is open-domain
3. The implemented model must be seamlessly integrated into the existing **Elasticsearch / Apache Lucene** based system

References:

[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.

[2] Tie-Yan Liu. 2009. Learning to Rank for Information Retrieval. Found. Trends Inf. Retr. 3, 3 (March 2009), 225-331

[3] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

[4] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng. 2018. MS MARCO: A Human-Generated MAchine Reading COmprehension Dataset

[5] Hangbo Bao *et al*, UNILMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training

Practical Information:

Supervisor: CASANOVA Pierre-Yves, Data Scientist, pierre-yves.casanova@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