NLP pour l'analyse d'articles scientifiques extraction des données par ChatBot

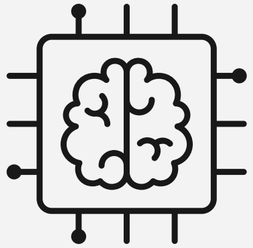
End-to-end Open-Domain Question Answering

Encadré par:

Mr.Thierry Bertin

Réalisé par:

Oumayma EDDOUKS
SALIMA EDDOUKS



Objectifs:

- **Avancement du projet :**

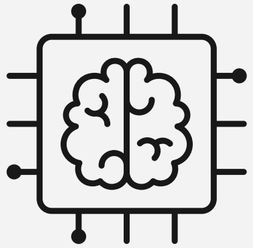
  Mise en place du modèle pré-entraîné

  Méthodologie de prétraitement des données

  Entraînement du modèle

- **Analyse des difficultés :**

  Limitations de l'API Cohere : contraintes sur le nombre de générations de questions-réponses par minute et leur impact sur la taille de notre ensemble de données .



Énoncé du problème:

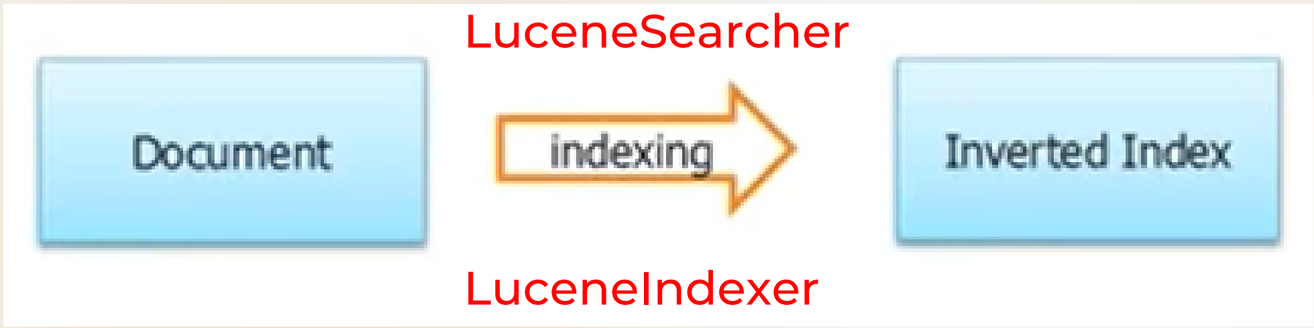
Question Answering:
La recherche en question-réponse (QA) vise à construire des systèmes capables de répondre aux questions posées en langage naturel par les humains. Elle est couramment utilisée pour créer des applications client conversationnelles, incluant les chatbots.

Closed-domain:
Les réponses à une question donnée sont récupérées dans un domaine spécifique.

Open-domain:
La tâche de question-réponse sur des ensembles de données en domaine ouvert tels que Wikipedia consiste à répondre à des questions qui peuvent couvrir n'importe quel sujet, ce qui rend encore plus difficile pour le système de QA de trouver la bonne réponse.

Indexer

- L'indexation des moteurs de recherche collecte, analyse et stocke des données pour faciliter une récupération d'informations rapide et précise.



Retriever-Reader

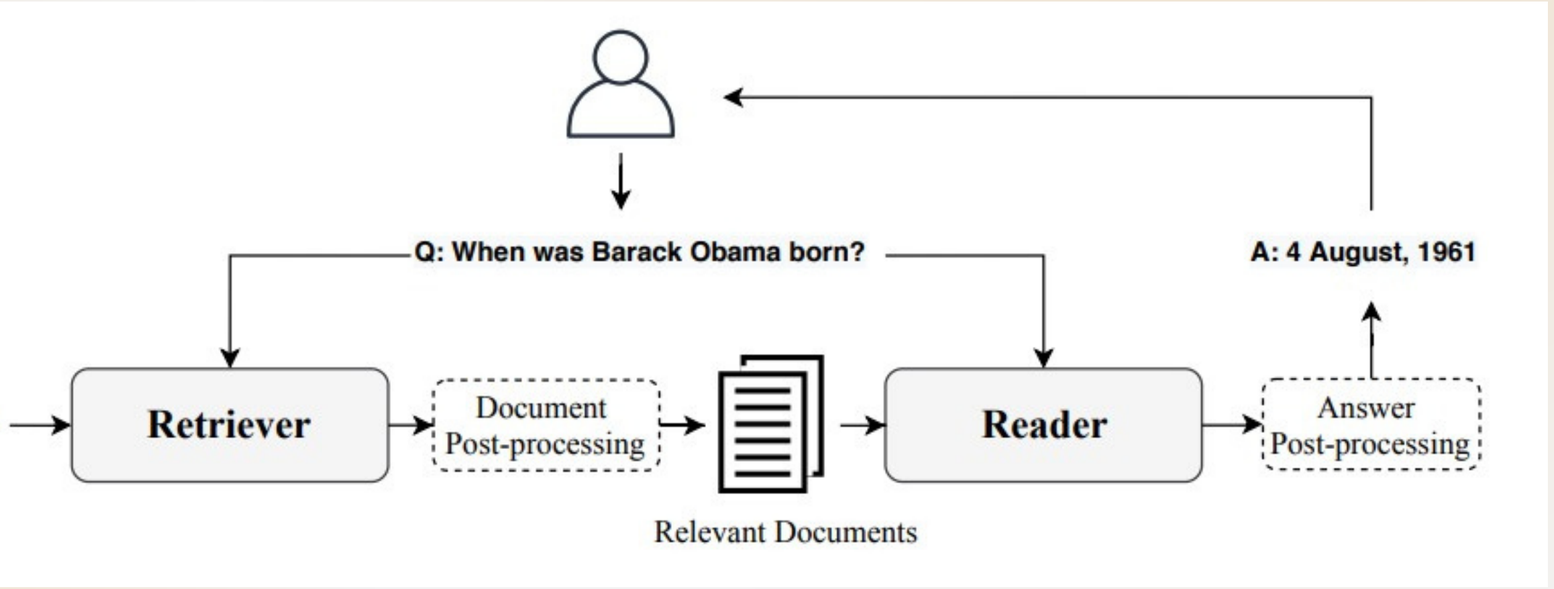
Retriever

responsable de collecter les documents pertinents à partir d'un corpus donné par rapport au texte de la question.

Reader

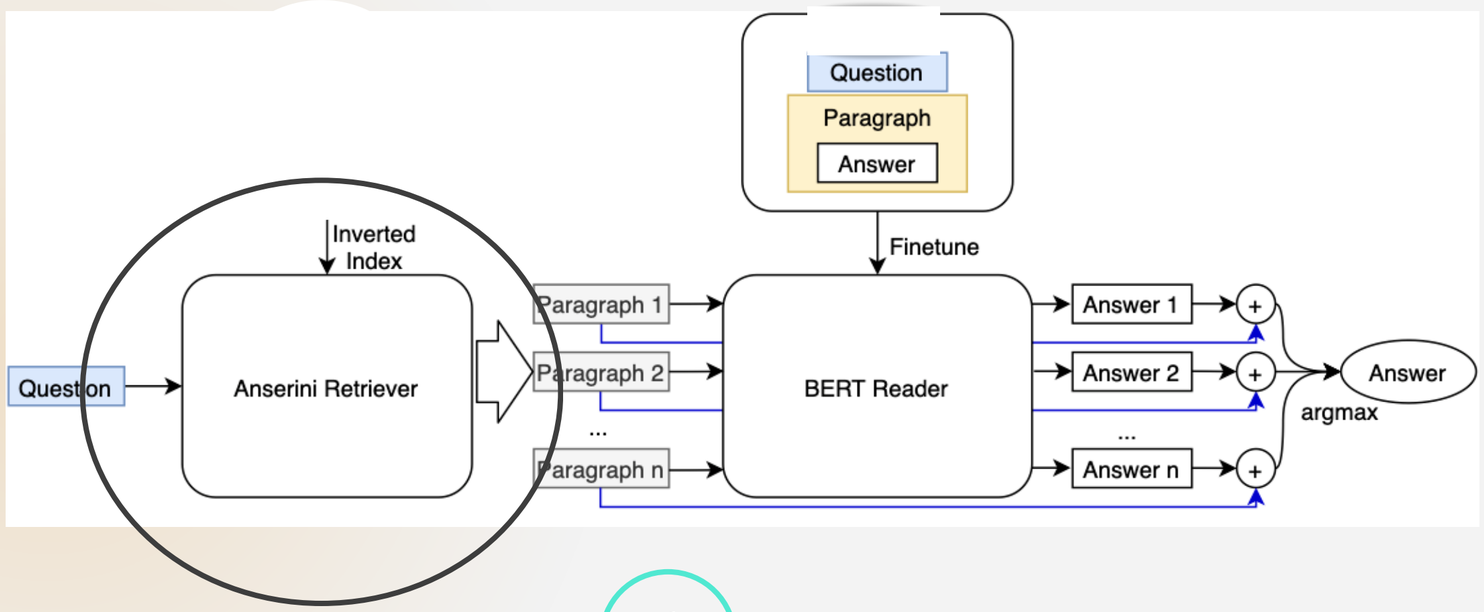
Le but est d'identifier la réponse finale à partir des documents reçus.

unstructed Documentsnd

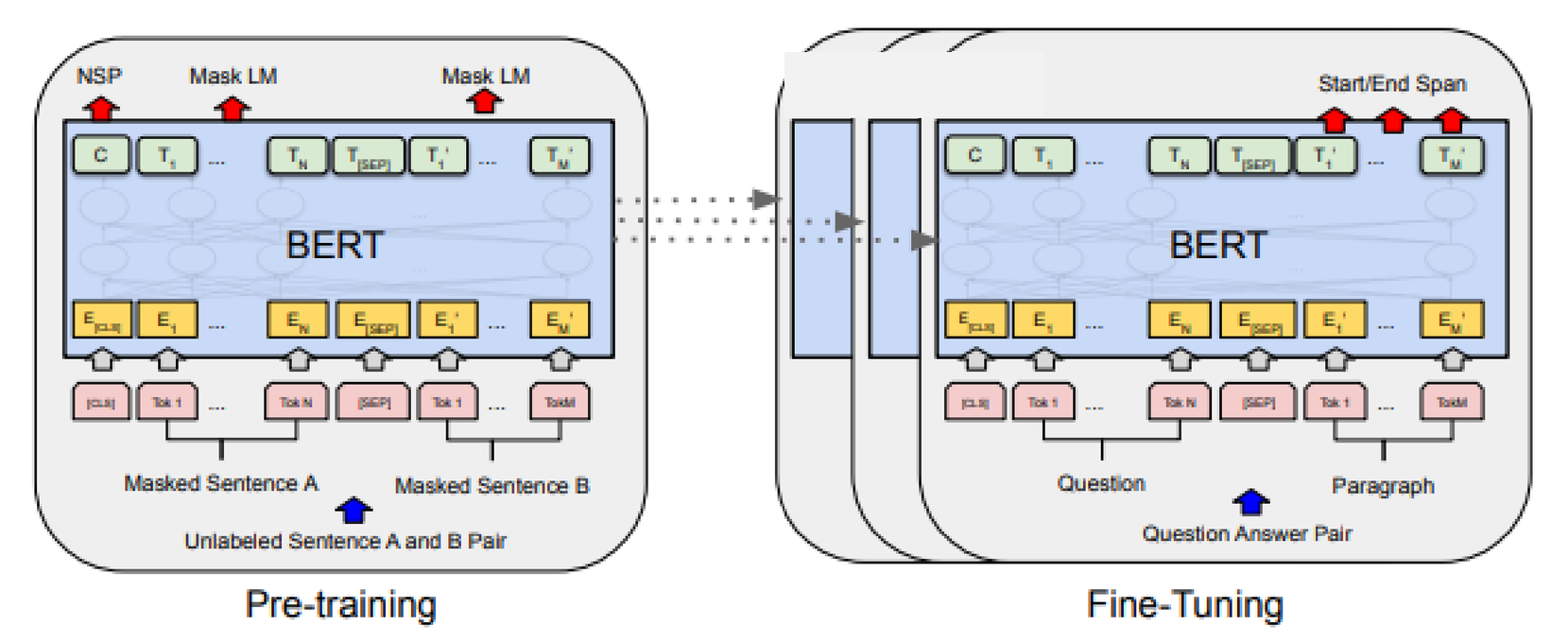


BERTserini: Anserini retriever

Le récupérateur : une boîte à outils closed source pour la recherche d'informations qui s'appuie sur Lucene. Nous l'avons implémenté en utilisant Pyserini, une boîte à outils Python pour la recherche reproductible en matière de recherche d'informations.

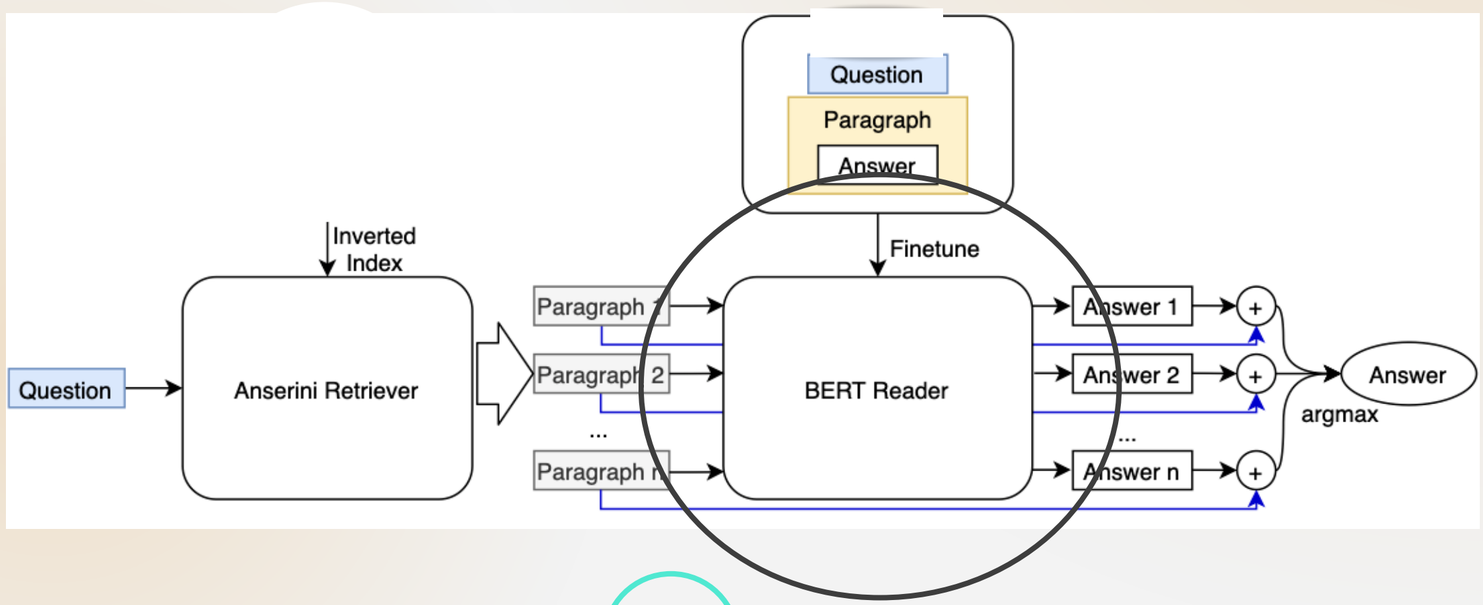


pre-training and fine-tuning procedures for BERT



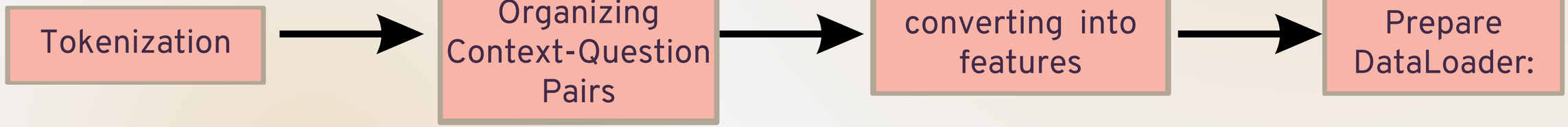
BERTserini: BERT reader

BERT est un cadre d'apprentissage automatique pour le traitement du langage naturel (NLP). Le modèle BERT est pré-entraîné en utilisant du texte provenant de Wikipedia, puis affiné avec notre Dataset, un ensemble de données de questions et réponses.

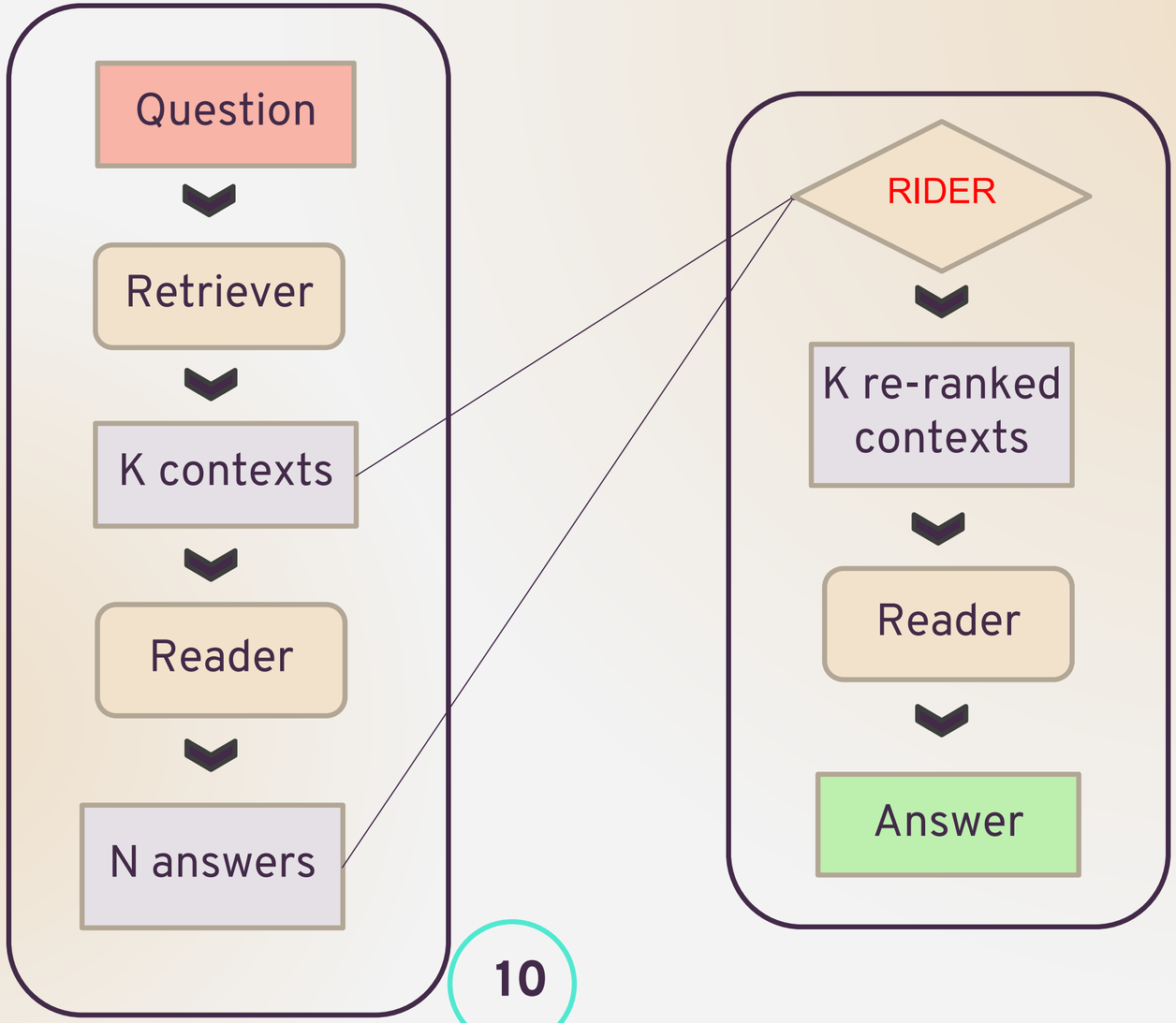


BERTserini: BERT reader

Data preprocessing



Extensions – RIDER re-ranking



Resultats

Question

What is the title of the paper by Eric Moulines and Francis R. Bach?

Top 5 segments

Document 1205, score: 13.489500045776367 pp. 1097–1105, 2012. Maas, Andrew L, Daly, E.....

Document 1201, score: 6.656000137329102 averaging. Previously in Moulines & Bach (2011)....

Document 1054, score: 4.189700126647949 Language Models are Unsupervised Multitask.....

Document 3, score: 4.09509992599487 3 Under review as a conference paper at.....

Document 35, score: 3.8008999824523926 Under review as a conference paper at ICLR.....

Response

The answer is: Non-asymptotic analysis of stochastic approximation algorithms for machine learning

Experiments – Dataset & Metrics

Exact-Match (EM):
C'est une métrique binaire qui mesure chaque paire question-réponse. Si les prédictions correspondent exactement aux réponses correctes, alors EM = 1, sinon EM = 0.

F1-score:
C'est une métrique courante pour les problèmes de classification. Elle exploite les potentiels de rappel et de précision en calculant leur moyenne harmonique

Conclusions et futures perspectives