# The Power of Selecting Key Blocks with Local Pre-ranking for Long Document Information Retrieval

*Article publié dans ACM Transactions on Information Systems 41 (3), pages 1-35, janvier 2023*

Minghan Li[1] Diana Nicoleta Popa[2] Johan Chagnon[3] Yagmur Gizem Cinar[4]
Eric Gaussier[1]
(1) Université Grenoble Alpes, France
(2) Telepathy Labs, Switzerland
(3) University of Wollongong, Australia
(4) Amazon, United Kingdom

RÉSUMÉ

Les réseaux neuronaux profonds et les modèles fondés sur les transformeurs comme BERT ont envahi le domaine de la recherche d'informations (RI) ces dernières années. Leur succès est lié au mécanisme d'auto-attention qui permet de capturer les dépendances entre les mots indépendamment de leur distance. Cependant, en raison de sa complexité quadratique dans le nombre de mots, ce mécanisme ne peut être directement utilisé sur de longues séquences, ce qui ne permet pas de déployer entièrement les modèles neuronaux sur des documents longs pouvant contenir des milliers de mots. Trois stratégies standard ont été adoptées pour contourner ce problème. La première consiste à tronquer les documents longs, la deuxième à segmenter les documents longs en passages plus courts et la dernière à remplacer le module d'auto-attention par des modules d'attention parcimonieux. Dans le premier cas, des informations importantes peuvent être perdues et le jugement de pertinence n'est fondé que sur une partie de l'information contenue dans le document. Dans le deuxième cas, une architecture hiérarchique peut être adoptée pour construire une représentation du document sur la base des représentations de chaque passage. Cela dit, malgré ses résultats prometteurs, cette stratégie reste coûteuse en temps, en mémoire et en énergie. Dans le troisième cas, les contraintes de parcimonie peuvent conduire à manquer des dépendances importantes et, *in fine*, à des résultats sous-optimaux. L'approche que nous proposons est légèrement différente de ces stratégies et vise à capturer, dans les documents longs, les blocs les plus importants permettant de décider du statut, pertinent ou non, de l'ensemble du document. Elle repose sur trois étapes principales : (a) la sélection de blocs clés (c'est-à-dire susceptibles d'être pertinents) avec un pré-classement local en utilisant soit des modèles de RI classiques, soit un module d'apprentissage, (b) l'apprentissage d'une représentation conjointe des requêtes et des blocs clés à l'aide d'un modèle BERT standard, et (c) le calcul d'un score de pertinence final qui peut être considéré comme une agrégation d'informations de pertinence locale. Dans cet article, nous menons tout d'abord une analyse qui révèle que les signaux de pertinence peuvent apparaître à différents endroits dans les documents et que de tels signaux sont mieux capturés par des relations sémantiques que par des correspondances exactes. Nous examinons ensuite plusieurs méthodes pour sélectionner les blocs pertinents et montrons comment intégrer ces méthodes dans les modèles récents de RI.

MOTS-CLÉS : Modèles de langue, modèles neuronaux, recherche d'information dans les documents longs.

KEYWORDS: BERT-based language models, long-document neural information retrieval.