Highlighting exact matching via marking strategies for ad hoc document ranking with pretrained contextualized language models*

Lila Boualili[1] Jose G.Moreno[1] Mohand Boughanem[1]
(1) IRIT, Université de Toulouse III, Toulouse, France
lila.boualili@irit.fr, jose.moreno@irit.fr, mohand.boughanem@irit.fr

RÉSUMÉ

Les modèles de langue pré-entraînés (MLPs) à l'instar de BERT se sont révélés remarquablement efficaces pour le classement ad hoc. Contrairement aux modèles antérieurs à BERT qui nécessitent des composants neuronaux spécialisés pour capturer les différents aspects de la pertinence entre la requête et le document, les MLPs sont uniquement basés sur des blocs de "transformers" où l'attention est le seul mécanisme utilisé pour extraire des signaux à partir des interactions entre les termes de la requête et le document. Grâce à l'attention croisée du "transformer", BERT s'est avéré être un modèle d'appariement sémantique efficace. Cependant, l'appariement exact reste un signal essentiel pour évaluer la pertinence d'un document par rapport à une requête de recherche d'informations, en dehors de l'appariement sémantique. Dans cet article, nous partons de l'hypothèse que BERT pourrait bénéficier d'indices explicites d'appariement exact pour mieux s'adapter à la tâche d'estimation de pertinence. Dans ce travail, nous explorons des stratégies d'intégration des signaux d'appariement exact en utilisant des "tokens" de marquage permettant de mettre en évidence les correspondances exactes entre les termes de la requête et ceux du document. Nous constatons que cette approche de marquage simple améliore de manière significative le modèle BERT vanille de référence. Nous démontrons empiriquement l'efficacité de notre approche par le biais d'expériences exhaustives sur trois collections standards en recherche d'information (RI). Les résultats montrent que les indices explicites de correspondance exacte transmis par le marquage sont bénéfiques pour des MLPs aussi bien BERT que pour ELECTRA. Nos résultats confirment que les indices traditionnels de RI, tels que la correspondance exacte de termes, sont toujours utiles pour les nouveaux modèles contextualisés pré-entraînés tels que BERT.

MOTS-CLÉS : Deep Learning, Modèles de Languge Pré-entraînés, Classement Ad hoc, Appariement Exact.

KEYWORDS: Deep Learning, Pre-trained Language Models, Ad hoc Ranking, Exact Matching.

*. Article publié dans Information Retrieval Journal: Lila Boualili, Jose G. Moreno, and Mohand Boughanem. "Highlighting exact matching via marking strategies for ad hoc document ranking with pretrained contextualized language models." Inf Retrieval J 25.4 (2022): 414-460. https://doi.org/10.1007/s10791-022-09414-x