Intégration de Vocabulaire Spécifique au Domaine dans une Méthode d'Ordonnancement de Documents Biomédicaux Basée sur BERT.

Maël Lesavourey[1], Gilles Hubert[1]

[1]IRIT, 118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9

Abstract

Ce résumé présente un aperçu de la méthode utilisée lors de la campagne d'évaluation BioASQ 11B sur l'ordonnancement de documents biomédicaux et détaillée dans les « working notes » de la campagne [1]. Nous avons proposé une méthode en 2 phases : la première basée sur les sacs de mots et BM25, développée avec Pyserini, et la seconde basée sur une implémentation de CEDR, un modèle basé sur BERT. Nous présentons une stratégie pour incorporer des connaissances biomédicales dans de tels modèles afin d'améliorer leur compréhension du contexte, l'idée étant qu'un terme du domaine est porteur d'une information spécifique.

Keywords

biomedical document ranking, information retrieval, thesaurus-based knowledge, BERT cross-encoder, multi-stage retrieval

1. Introduction

Le nombre de publications scientifiques dans le domaine biomédical est en constante augmentation. Ceci permet aux chercheurs d'accéder à une plus grande quantité de savoirs mais en même temps il devient de plus en plus difficile d'avoir une navigation précise lorsqu'on interroge ces sources de connaissances. De nombreuses initiatives voient le jour pour faire face à ce problème et les travaux de Recherche d'Information Biomédicale (RIBio) se multiplient [2], notamment autour de la génération automatique de résumé et l'ordonnancement de documents [3]. Ces dernières années, les Modèles de Langage Pré-entraînés (MLPs) ont permis des avancées significatives sur ces sujets par leur capacité à capturer les relations sémantiques entre les termes d'un document. Cependant des spécificités liées au domaine comme le lexique (noms, abréviations, symboles) et la polysémie limitent les performances des MLPs sur la litérature biomédicale [4, 5].

Dans ce contexte, nous proposons d'étudier une méthode d'incorporation de connaissances du domaine biomédical dans les systèmes d'ordonnancement de documents basés sur BERT (Bidirectional Encoder Representations from Transformers) [6]. Plus précisément nous souhaitons modifier la séquence d'entrée grâce à un balisage des termes biomédicaux avec l'intuition

CORIA 2024: COnférence en Recherche d'Information et Applications, 03–04 Avril, 2024, La Rochelle, France

mael.lesavourey@irit.fr (M. Lesavourey); gilles.hubert@irit.fr (G. Hubert)

https://www.irit.fr/~Gilles.Hubert/ (G. Hubert)

CEUR Workshop Proceedings (CEUR-WS.org)

qu'un terme référencé dans un thesaurus sera porteur d'une information plus importante pour comprendre le contexte de la requête et du document considéré.

2. Méthodologie

Les spécificités du vocabulaire biomédical ne sont pas les seules limites des approches classiques pour l'ordonnancement de documents. En effet, la séquence en entrée de BERT est au maximum de 512 « tokens » et le calcul de pertinence se résume souvent à une réduction de dimension sur le plongement du « token » [CLS] en omettant les plongements des mots en sortie. Nous avons implémenté une version de Contextualized Embeddings for Document Ranking (CEDR) [7] permettant de pallier ces deux limites. De plus, il fournit un score de référence pour évaluer notre méthode.

Afin d'incorporer des connaissances biomédicales à ce modèle, nous avons utilisé le thesaurus MeSH[1] (Medical Subject Headings), un vocabulaire contrôlé permettant d'indexer des publications du domaine. Nous proposons d'intégrer ce vocabulaire en modifiant la séquence en entrée de modèle par une stratégie de balisage. Cette stratégie a été utilisée dans d'autres contextes, notamment par [8] pour mettre en évidence les correspondances exactes entre les mots d'une requête et d'un document. Nous avons orienté le balisage vers les termes référencés dans le thésaurus ainsi que leurs synonymes pour obtenir une correspondance souple (« soft-match »).

3. Évaluation / Résultats

Pour évaluer notre système, nous nous sommes appuyés sur l'initiative BioASQ[2] [3] qui propose une campagne d'évaluation annuelle sur des tâches de RIBio. La tâche B [9] consiste, pour une requête donnée, à retourner les 10 articles les plus pertinents pour y répondre.

Pour cette tâche, nous avons mis en place un système en deux étapes [10]. Un premier bloc basé sur BM25 [11] et implémenté avec Pyserini [12] permet de créer une liste candidate de 500 documents. Le second a pour objectif de réordonner cette liste de documents à l'aide des méthodes décrites dans la section 2.

Les résultats de cette première implémentation étaient moyens (Classement système : 16/27, Classement équipe : 4/8) mais prometteurs car la version « balisée » du modèle donne régulièrement des résultats supérieurs à la version classique. Les performances de cette méthode d'incorporation de connaissances restent cependant assez instables ce qui va dans le sens des remarques de [13] sur l'efficacité de la correspondance lorsqu'elle est comparée à des baselines plus fortes.

Pour plus de détails, nous invitons le lecteur à se référer à l'article original [1] qui décrit notre participation à la campagne d'évaluation BioASQ 11B.

[1] https://www.nlm.nih.gov/mesh/meshhome.html
[2] http://www.bioasq.org/

References

[1] M. Lesavourey, G. Hubert, BioASQ 11B: Integrating Domain Specific Vocabulary to BERT-based Model for Biomedical Document Ranking. (2023).

[2] L. Tamine, L. Goeuriot, Semantic Information Retrieval on Medical Texts: Research Challenges, Survey, and Open Issues, ACM Comput. Surv. 54 (2021). URL: https://doi.org/10.1145/3462476. doi:10.1145/3462476.

[3] A. Nentidis, A. Krithara, G. Paliouras, E. Farre-Maduell, S. Lima-Lopez, M. Krallinger, BioASQ At CLEF2023: The Eleventh Edition Of The Large-Scale Biomedical Semantic Indexing And Question Answering Challenge, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer-Verlag, Berlin, Heidelberg, 2023, p. 577–584. URL: https://doi.org/10.1007/978-3-031-28241-6_66. doi:10.1007/978-3-031-28241-6_66.

[4] J. Tan, J. Hu, S. Dong, Incorporating entity-level knowledge in pretrained language model for biomedical dense retrieval, Computers in Biology and Medicine 166 (2023) 107535.

[5] Q. Xie, P. Tiwari, S. Ananiadou, Knowledge-enhanced Graph Topic Transformer for Explainable Biomedical Text Summarization, IEEE Journal of Biomedical and Health Informatics (2023) 1–12. doi:10.1109/JBHI.2023.3308064.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[7] S. MacAvaney, A. Yates, A. Cohan, N. Goharian, CEDR: Contextualized Embeddings for Document Ranking, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1101–1104. URL: https://doi.org/10.1145/3331184.3331317. doi:10.1145/3331184.3331317.

[8] L. Boualili, J. G. Moreno, M. Boughanem, MarkedBERT: Integrating Traditional IR Cues in Pre-Trained Language Models for Passage Retrieval, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1977–1980. URL: https://doi.org/10.1145/3397271.3401194. doi:10.1145/3397271.3401194.

[9] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, BioASQ-QA: A manually curated corpus for Biomedical Question Answering, Scientific Data 10 (2023) 170.

[10] R. F. Nogueira, W. Yang, K. Cho, J. Lin, Multi-Stage Document Ranking with BERT, CoRR abs/1910.14424 (2019). URL: http://arxiv.org/abs/1910.14424. arXiv:1910.14424.

[11] S. Robertson, H. Zaragoza, et al., The probabilistic relevance framework: BM25 and beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389.

[12] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, R. Nogueira, Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 2356–2362. URL: https://doi.org/10.1145/3404835.3463238. doi:10.1145/3404835.3463238.

[13] J. Lin, R. F. Nogueira, A. Yates, Pretrained transformers for text ranking: BERT and beyond, CoRR abs/2010.06467 (2020). URL: https://arxiv.org/abs/2010.06467. arXiv:2010.06467.