Adaptation de Domaine Simple pour la Recherche Parcimonieuse*

Mathias Vast[1,2,**], Yuxuan Zong[2], Basile Van Cooten[1], Benjamin Piwowarski[2] and Laure Soulier[2]

[1]Sinequa
[2]Sorbonne Université, CNRS, ISIR, F-75005 Paris, France

Résumé

Dans le cadre de la Recherche d'Information (RI), l'apprentissage des modèles repose fortement sur l'approche "pré-entraîner puis affiner". Malgré ses très bons résultats, cette méthode nécessite d'avoir accès à un jeu de données labellisées ce qui complique son application à de nouveaux domaines ou langues, en particulier si ceux-ci sont faiblement fournis. Cet article propose une solution simple au transfert d'un modèle de recherche parcimonieux, SPLADE, vers des domaines sans données labellisées.

Mots-clés

Adaptation de domaine sans entraînement, Recherche parcimonieuse

1. Introduction

En Recherche d'Information (RI) et plus généralement en Traitement Automatique du Langage Naturel (TALN), les meilleurs modèles sont entraînés selon la procédure "pré-entraînement puis raffinage" [1, 2, 3]. Cependant, le manque de données étiquetées dans certaines langues ou domaines rend cette méthode inopérante et contraint à entraîner ces modèles sur un jeu de données parfois très éloigné du domaine cible provoquant de lourdes dégradations de leurs performances [4, 5]. Il devient donc important de développer des méthodes d'adaptation permettant de résoudre ce problème. Parmi elles, on retrouve notamment la génération de données synthétiques [6, 7, 8, 9], de nouvelles méthodes de pré-entraînement [10, 11, 12] ou encore les méthodes dites *Parameter-Efficient Fine-Tuning* [13, 14, 15, 16]. De son côté, Artetxe et al. [17] propose de faire une distinction au sein même des paramètres du modèle entre un sous-ensemble dédié à l'apprentissage de la tâche, P_{task}, et un autre dédié à l'apprentissage du domaine, P_{domain}, pour lequel uniquement des données non-étiquetées existent. Dans la continuité de ce travail, nous considérons ici un scénario d'adaptation inter-domaines sur la tâche de RI où nous ne disposons pas de données étiquetées pour le domaine cible.

CORIA 2024: COnférence en Recherche d'Information et Applications, 3 et 4 avril 2024, La Rochelle, France
*Ce papier a été accepté comme papier court à la conférence *European Conference on Information Retrieval* 2024 sous le titre *Simple Domain Adaptation for Sparse Retrievers*.
**Auteur correspondant
mathias.vast@sinequa.com (M. Vast); yuxuan.zong@isir.upmc.fr (Y. Zong); vancoten@sinequa.com (B. V. Cooten); benjamin.piwowarski@isir.upmc.fr (B. Piwowarski); laure.soulier@isir.upmc.fr (L. Soulier)
0009-0007-4612-717X (M. Vast); 0009-0002-0376-1369 (Y. Zong); 0000-0001-6792-3262 (B. Piwowarski); 0000-0001-9827-7400 (L. Soulier)

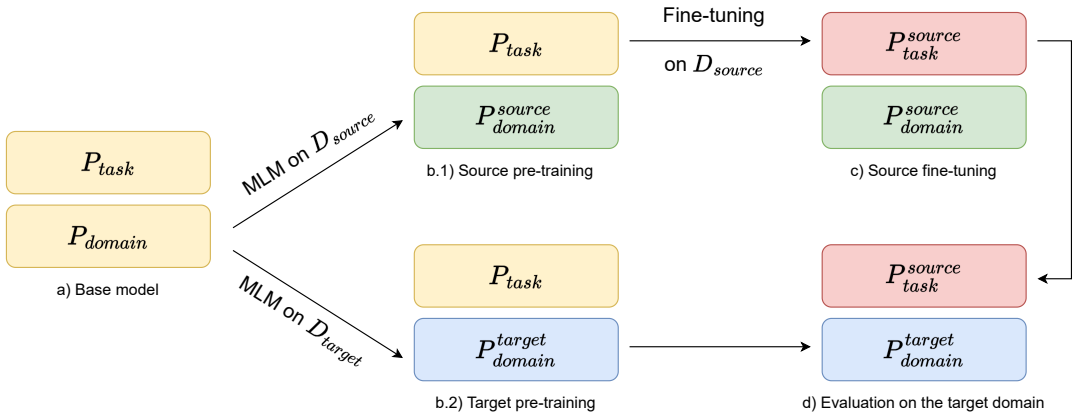

Figure 1: Illustration du processus d'adaptation de domaine

2. Contribution

Notre processus d'entraînement est très similaire à celui d'Artetxe et al. [17] à la différence que nous réalisons un pré-entraînement supplémentaire sur le domaine source afin de spécifier P_{domain} avant de raffiner P_{task} sur ce même domaine[1]. La figure 1 résume le processus complet.

3. Expériences

Nous initialisons SPLADE avec un modèle $BERT_{base}$. Contrairement à Artetxe et al. [17], nous expérimentons avec plusieurs variantes incluant k couches en plus de la couche de plongements au sous-ensemble de paramètres P_{domain}. Afin de vérifier l'efficacité de notre méthode, nous comparons ces variantes à BM25 [18] et SPLADE (sans apprentissage) raffiné sur le domaine source, i.e. MSMARCO [19], ainsi que deux ablations (sans pré-entraînement d'aucune sorte et sans pré-entraînement sur la source). Les résultats montrent que notre approche d'adaptation permet de gagner en moyenne entre 0,7 et 1,4 points en nDCG@10 par rapport au transfert sans-apprentissage. Les ablations démontrent que l'absence d'une partie de notre procédure peut nuire à la performance du modèle jusqu'à 1,2 points en nDCG@10. Les différentes variantes que nous avons testées montrent aussi que réduire l'apprentissage du domaine à la couche de plongement est sous-optimal et que l'on peut observer des gains jusqu'à $k = 4$ couches.

4. Conclusion

Cet article présente une nouvelle approche de l'adaptation inter-domaines en Recherche d'Information, soulignant l'importance d'un pré-entraînement et d'un raffinage prenant en compte les spécificités du modèle sous-jacent. En se concentrant à la fois sur les domaines source et cible, notre méthode optimise la performance du modèle à travers divers domaines avec des données étiquetées limitées. Notre recherche contribue aux travaux en cours en RI et plus largement en TALN, offrant une solution viable aux défis posés par le manque de données spécifiques au domaine et par la complexité croissante des modèles dans ces domaines.

Remerciements

Ce travail est soutenu par le projet ANR ANR-23-IAS1-0003.

[1]Le code source est disponible librement : https://git.isir.upmc.fr/mat_vast/cross_domain_adaptation

References

[1] T. Formal, C. Lassance, B. Piwowarski, S. Clinchant, Splade v2: Sparse lexical and expansion model for information retrieval, 2021. URL: https://arxiv.org/abs/2109.10086. doi:10.48550/ARXIV.2109.10086.

[2] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6769–6781. URL: https://aclanthology.org/2020.emnlp-main.550. doi:10.18653/v1/2020.emnlp-main.550.

[3] R. Nogueira, K. Cho, Passage re-ranking with bert, 2019. URL: http://arxiv.org/abs/1901.04085.

[4] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models, 2021. URL: http://arxiv.org/abs/2104.08663. doi:10.48550/arXiv.2104.08663.

[5] X. Zhang, A. Yates, J. Lin, A little bit is worse than none: Ranking with limited training data, in: N. S. Moosavi, A. Fan, V. Shwartz, G. Glavaš, S. Joty, A. Wang, T. Wolf (Eds.), Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing, Association for Computational Linguistics, Online, 2020, pp. 107–112. URL: https://aclanthology.org/2020.sustainlp-1.14. doi:10.18653/v1/2020.sustainlp-1.14.

[6] R. Nogueira, W. Yang, J. J. Lin, K. Cho, Document expansion by query prediction, ArXiv abs/1904.08375 (2019). URL: https://api.semanticscholar.org/CorpusID:119314259.

[7] M. Li, E. Gaussier, Domain adaptation for dense retrieval through self-supervision by pseudo-relevance labeling, 2022. arXiv:2212.06552.

[8] K. Wang, N. Thakur, N. Reimers, I. Gurevych, Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval, Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2022). URL: http://dx.doi.org/10.18653/v1/2022.naacl-main.168. doi:10.18653/v1/2022.naacl-main.168.

[9] L. Gao, X. Ma, J. Lin, J. Callan, Precise zero-shot dense retrieval without relevance labels, 2022. arXiv:2212.10496.

[10] C. Lassance, H. Dejean, S. Clinchant, An Experimental Study on Pretraining Transformers from Scratch for IR, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, volume 13980, Springer Nature Switzerland, Cham, 2023, pp. 504–520. URL: https://link.springer.com/10.1007/978-3-031-28244-7_32. doi:10.1007/978-3-031-28244-7_32, series Title: Lecture Notes in Computer Science.

[11] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8342–8360. URL: https://aclanthology.org/2020.acl-main.740. doi:10.18653/v1/2020.acl-main.740.

[12] K. Krishna, S. Garg, J. Bigham, Z. Lipton, Downstream datasets make surprisingly good

pretraining corpora, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 12207–12222. URL: https://aclanthology.org/2023.acl-long.682. doi:10.18653/v1/2023.acl-long.682.

[13] J. Zhan, Q. Ai, Y. Liu, J. Mao, X. Xie, M. Zhang, S. Ma, Disentangled Modeling of Domain and Relevance for Adaptable Dense Retrieval, 2022. URL: http://arxiv.org/abs/2208.05753, arXiv:2208.05753 [cs].

[14] R. Litschko, I. Vulić, G. Glavaš, Parameter-efficient neural reranking for cross-lingual and multilingual retrieval, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 1071–1082. URL: https://aclanthology.org/2022.coling-1.90.

[15] V. Pal, C. Lassance, H. Déjean, S. Clinchant, Parameter-efficient sparse retrievers and rerankers using adapters, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 16–31.

[16] W. L. Tam, X. Liu, K. Ji, L. Xue, X. Zhang, Y. Dong, J. Liu, M. Hu, J. Tang, Parameter-Efficient Prompt Tuning Makes Generalized and Calibrated Neural Text Retrievers, 2022. URL: http://arxiv.org/abs/2207.07087. doi:10.48550/arXiv.2207.07087.

[17] M. Artetxe, S. Ruder, D. Yogatama, On the cross-lingual transferability of monolingual representations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, p. 4623–4637. URL: http://arxiv.org/abs/1910.11856. doi:10.18653/v1/2020.acl-main.421, arXiv:1910.11856 [cs].

[18] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at trec-3, in: Text Retrieval Conference, 1994. URL: https://api.semanticscholar.org/CorpusID:3946054.

[19] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, Ms marco: A human generated machine reading comprehension dataset., CoRR abs/1611.09268 (2016). URL: http://dblp.uni-trier.de/db/journals/corr/corr1611.html#NguyenRSGTMD16.