Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

DÉfi Fouille de Texte 2024

Théo Charlot Elisabeth Sisarith Nicolas Stucky Rémi Ilango Nicolas Gouget
Hreshvik Sewraj Xavier Pillet
Laboratoire des Sciences du Numérique de Nantes (LS2N), adresse, 44000 Nantes, France
{prenom.nom}@etu.univ-nantes.fr

RÉSUMÉ

Cet article présente une série d'expériences sur la tâche de réponse à des questions à choix multiples de DEFT2024. En s'appuyant sur le corpus FrenchMedMCQA, nous avons mis en œuvre plusieurs approches, incluant des techniques de Récupération augmenté de modèle de langue pré entraîné (REALM).

ABSTRACT

This article presents a series of experiments on the DEFT2024 multiple-choice question answering task. Based on the FrenchMedMCQA corpus, we implemented several approaches, including Retrieval-Augmented Language Model Pre-Training (REALM) techniques.

MOTS-CLÉS : Questions à choix multiples, DEFT, REALM.

KEYWORDS: Multiple-choice questions, DEFT, REALM.

1 Introduction

Afin de confronter différentes méthodes sur des thématiques diverses, la campagne d'évaluation francophone DÉfi Fouille de Textes (DEFT) est organisée chaque année.
Cette édition 2024 est la continuité de celle de l'année précédente, qui consistait à répondre automatiquement à des questionnaires à choix multiples dans le domaine pharmaceutique.
L'objectif est d'explorer d'autres méthodes et de les comparer à celles de 2023.

Dans cet article, nous allons continuer dans le même esprit que l'édition de l'année précédente, en lui implémentant la Récupération augmenté de modèle de langue pré entraîné, ou *Retrieval-Augmented Language Model Pre-Training* (REALM).

2 Corpus

Le corpus FrenchMedMCQA (Labrak *et al.*, 2022) répertorie un total de 3105 questions à choix multiples sur le domaine de la pharmacie. Chaque question est représentée par un identifiant et 5 réponses possibles. Les données sont en français et sont réparties en 3 ensembles : 70% de questions pour l'entraînement, 10% pour la validation et 20% pour le test. En plus de ces fichiers, l'utilisation

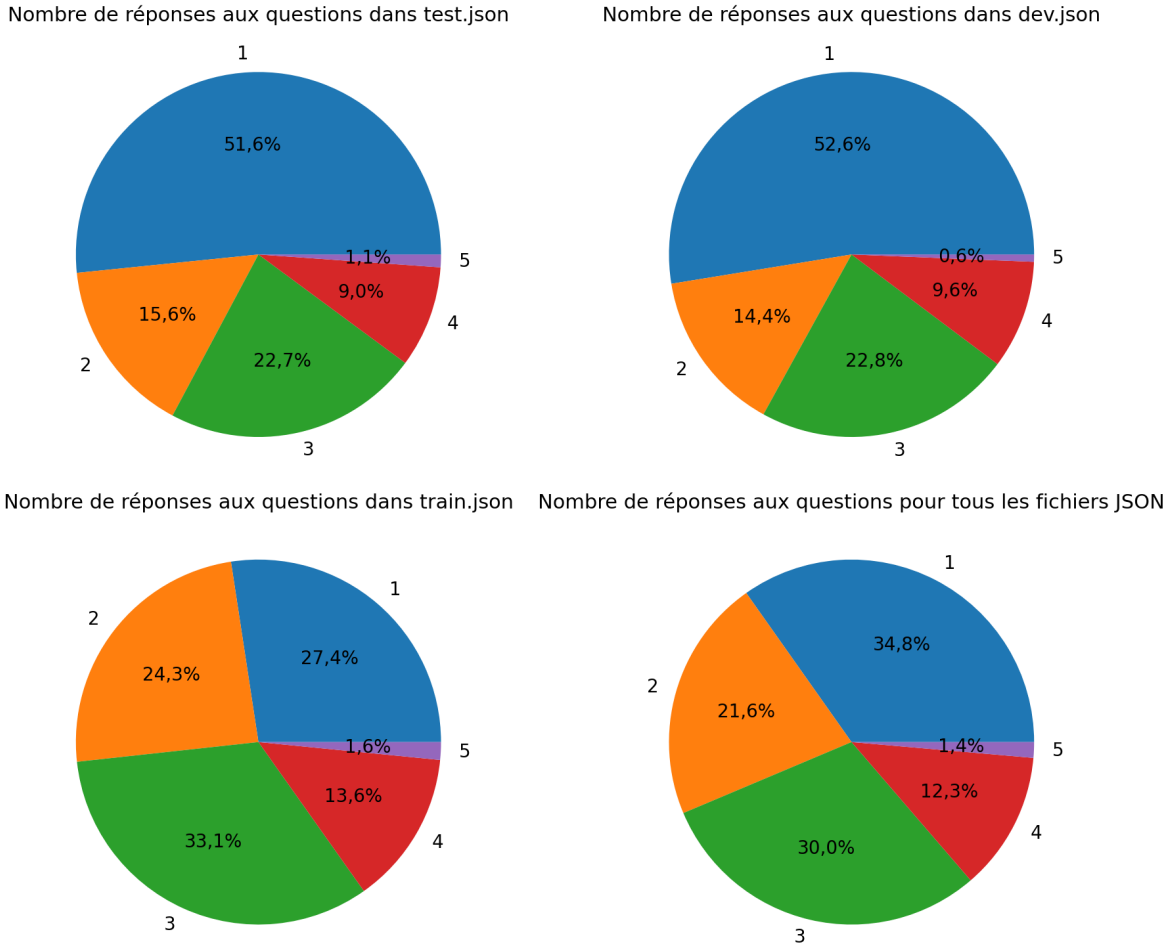

FIGURE 1 – Illustration de la répartition du nombre de bonnes réponses par fichiers. Les différentes parties des camemberts représentent la répartition du nombre de réponses. Si on prend l'exemple du camembert du test, dans 51,6% des cas il n'y a qu'une seule réponse uniquement, dans 22,7% des cas, il y a 3 réponses uniquement.

de deux autres sources de données est autorisée : NACHOS (Labrak *et al.*, 2022) et Wikipedia [1].

Ces graphiques illustrent que les jeux de données possèdent globalement plus de réponses uniques que de réponses à plusieurs choix. Par exemples les réponses à 5 choix corrects sont très rares (moins de 2% du corpus). On observe aussi que les proportions entre les différents nombres de réponses sont relativement respectées entre les différents jeux de données, sauf pour le jeu d'entraînement (train), ce qui pourrait biaiser le modèle et expliquer de moindres performances globales sur le jeu de test.

Nous pouvons observer que les questions où 5 réponses sont possibles sont les moins récurrentes.

3 Présentation de la tâche

La tâche principale de cette édition de DEFT est d'identifier l'ensemble des réponses correctes d'une question donnée parmi cinq réponses possibles de manière automatique.

1. https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal

Pour réaliser cette tâche, nous devrons proposer des systèmes de moins de 3 milliards de paramètres.

À cela s'ajoute l'évaluation des systèmes. Nous utiliserons les métriques Exact Match Ratio (EMR), correspondant au taux de réponses parfaitement justes, et Hammering Score, le taux de réponses justes parmi l'ensemble des réponses de références.

La tâche annexe est identique à la tâche principale, sans la limite de taille pour les systèmes. Nous n'étudierons cependant pas ce cas ici.

4 Systèmes

4.1 Architecture générale

Nous avons choisi de mettre en œuvre une Récupération augmenté de modèle de langue pré entraîné (REALM), en prolongeant les projets du *DEFT 2023*. La REALM permet d'améliorer considérablement la qualité des réponses générées par les modèles de langage en intégrant des informations pertinentes récupérées à partir de grandes bases de données. Cela permet non seulement d'enrichir le contenu généré, mais aussi d'accroître la précision et la pertinence des informations fournies.

En outre, la mise en œuvre de la REALM s'aligne parfaitement avec nos objectifs de recherche, qui visent à explorer des approches pour l'amélioration de la génération de texte automatisée.

Le Récupération augmenté de modèle de langue pré entraîné (REALM) (Lewis *et al.*, 2020), combine des modèles de récupération d'information et des modèles de génération de langage pour produire des réponses de meilleure qualité. Cette approche résout les limitations des modèles de génération de langage seuls en leur fournissant un accès à des informations pertinentes pour une tâche spécifique.

Les modèles de récupération d'information utilisent diverses techniques pour identifier les documents pertinents. La recherche par mots-clés récupère les documents en fonction des termes spécifiés dans la requête, tandis que la recherche sémantique, basée sur des modèles comme *Word2Vec* (Mikolov *et al.*, 2013) ou *BERT* (Devlin *et al.*, 2018), identifie les documents en fonction de leur similitude sémantique avec la requête. Les méthodes d'apprentissage supervisé, où des modèles sont entraînés à classer les documents en fonction de leur pertinence par rapport à la requête, sont également courantes. Ici, nous prenons 3 documents de NACHOS.

Dans le cadre du REALM, ces deux composants travaillent ensemble de manière synergique. Les documents récupérés fournissent un contexte que les modèles de génération de langage utilisent pour produire des réponses plus précises et informatives. Cette synergie permet d'améliorer la qualité des résultats générés en exploitant les informations pertinentes des documents récupérés.

L'un des avantages majeurs du REALM est sa capacité à fournir des réponses contextuellement riches. Cela est particulièrement bénéfique pour les tâches de question-réponse, la génération de textes longs et la vérification des faits, où des informations supplémentaires peuvent grandement améliorer la qualité des réponses.

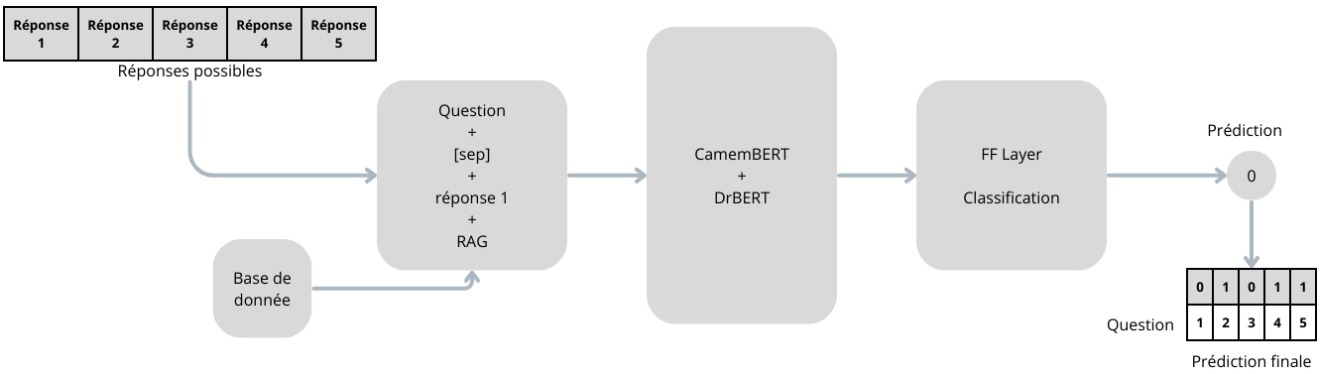

FIGURE 2 – Schéma de l'architecture générale

Le module de classification est composé d'une couche linéaire feedforward et d'une couche de projection pour formater à une sortie binaire.

4.2 Implémentation d'un REALM

Nous avons appliqué un Récupération augmenté de modèle de langue pré entraîné (REALM) en utilisant une combinaison de techniques de récupération d'information basées sur TF-IDF et de classification de séquence avec un modèle de type Transformer.

Nous avons utilisé les 100 000 premières phrases du corpus Nachos et extrait les 3 phrases les plus similaires à la question. Pour palier au problème de la taille maximale de l'entrée de nos modèles nous la tronquons si elle dépasse la limite.

Le *TF-IDF* (Term Frequency - Inverse Document Frequency) est utilisé pour représenter les documents et les requêtes en tant que vecteurs de caractéristiques. Cette méthode permet de pondérer l'importance des mots dans les documents en fonction de leur fréquence d'apparition dans l'ensemble des documents. La *similarité cosinus* est employée pour mesurer la similarité entre les vecteurs de la requête et les documents, permettant ainsi de classer les documents par pertinence par rapport à la requête.

BERT (Bidirectional Encoder Representations from Transformers) ou ses variantes, comme RoBERTa (Liu *et al.*, 2019), sont utilisés pour répondre aux questions en contexte. Ces modèles sont pré-entraînés sur de grandes quantités de données textuelles et sont particulièrement efficaces pour les tâches de questions réponses. Une pipeline de classification de séquences est utilisée pour répondre aux questions, en y ajoutant le contexte des phrases les plus pertinentes récupérées par la phase de TF-IDF.

Les performances du modèle sans l'utilisation de la REALM sont nettement inférieures. Nos tests ont montré que, sans la REALM, les performances sur notre jeu de développement étaient environ 5 fois moins bonnes. Cela démontre l'importance cruciale de cette méthode pour obtenir des résultats de haute qualité.

4.3 ReFT - Representation Fine-Tuning

Dans l'optique d'utiliser le moins de ressources possibles, notre première approche se base sur l'affinage de représentations (ReFT) (Wu *et al.*, 2024) qui au lieu de modifier les poids initiaux du modèle modifie des matrices additionnelles de dimensions nettement inférieures. ReFT se montre être plus économe en terme de paramètres par rapport à LoRA (Hu *et al.*, 2021) de l'ordre de 15 à 65 fois moins de paramètres (Wu *et al.*, 2024). Nous utilisons ici DrBERT (Labrak *et al.*, 2022) comme modèle de base car c'est un modèle qui est pré-entraîné sur des corpus de données médicales en français. ReFT permet d'affiner le modèle en modifiant uniquement 18 444 paramètres au lieu des 110 millions de paramètres du modèle de base ce qui permet d'affiner en 3 minutes le modèle avec le corpus donné sur une machine personnelle [2].

4.4 ExBERT

Notre deuxième approche utilise l'architecture de mixture d'experts (Shazeer *et al.*, 2017) qui s'est démarquée notamment avec Mixtral 8x7b (Jiang *et al.*, 2024). Cette architecture permet durant l'inférence de charger en mémoire et d'exécuter qu'une partie des paramètres du modèle grâce à un routeur sélectionnant l'expert le plus adapté à traiter un jeton. Plus récemment s'est vu démocratiser la création de modèles basés sur une architecture mixture d'experts réutilisant des modèles déjà entraînés pour créer les divers experts et n'entraînant que le routeur à bien sélectionner les experts durant l'inférence (Sukhbaatar *et al.*, 2024). Nous utilisons cette technique pour créer notre modèle à partir de DrBERT (Labrak *et al.*, 2023) ainsi que CamemBERT-bio (Touchent *et al.*, 2023).

5 Résultats

Système	Hamming	EMR
REFT	37,38	**1,60**
REFT-REALM	39,09	5.13
ExBERT - 1 expert	38,33	0,64
ExBERT	**42,20**	1,48

TABLE 1 – Résultats Corpus de validation

Système	Hamming	EMR
REFT	30,77	**4,40**
ExBERT	**45,85**	2,73

TABLE 2 – Résultats obtenus

2. GPU NVIDIA RTX 4070ti

Système	Hamming	EMR
multilabel-classification	33,27	12,22
nli	33,67	**14,15**
instruction-seq2seq	**35,96**	13,67

TABLE 3 – Résultats par l'équipe ALMAnaCH pour DEFT 2023

En comparant aux résultats de l'équipe ALMAnaCH [3] de DEFT 2023, nous réussissons à obtenir de meilleurs scores sur la métrique de Hamming avec la méthode ExBERT. Nous ne parvenons par contre pas à approcher les résultats de l'année dernière sur la métrique d'Exact Match Ratio, en obtenant au mieux 4,4% avec REFT.

Notre méthode ExBERT montre un meilleur score de Hamming mais une moins bonne performance en Exact Match Ratio. Le modèle arrive donc à prédire une grande quantité de paires question réponse mais n'arrive pas à prédire l'entièreté des réponses à une question. En contrôlant la sortie du modèle, nous nous sommes rendus compte que la sortie était erronée et prédisait l'entièreté des réponses à une question donnée, expliquant ainsi le score de Hamming élevé. Une explication à cela peut être un trop faible nombre d'itérations d'affinage, qui est ici de trois itérations, ou bien un problème d'implémentation de nos méthodes.

6 Conclusion

Bien que nous n'ayons pu montrer les qualités de ces approches, nous pensons qu'elles méritent d'être tout de même explorées de manière plus approfondie, notamment la mixture d'expert à partir de modèles pré-existants. Celle-ci soulève plusieurs questions dont celle de la combinaison de modèles aux langues distinctes permettant ou non le transfert des connaissances, par exemple, d'un modèle français spécialisé dans le domaine médical à un modèle japonais spécialisé dans un autre domaine. Cela permettrait, en somme, de réutiliser au mieux les modèles pré-entraînés déjà existants.

Références

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2021). Lora : Low-rank adaptation of large language models. *arXiv preprint arXiv :2106.09685*.

JIANG A. Q., SABLAYROLLES A., ROUX A., MENSCH A., SAVARY B., BAMFORD C., CHAPLOT D. S., DE LAS CASAS D., HANNA E. B., BRESSAND F., LENGYEL G., BOUR G., LAMPLE G., LAVAUD L. R., SAULNIER L., LACHAUX M.-A., STOCK P., SUBRAMANIAN S., YANG S.,

3. https://talnarchives.atala.org/ateliers/2023/DEFT/480058.pdf

ANTONIAK S., SCAO T. L., GERVET T., LAVRIL T., WANG T., LACROIX T. & SAYED W. E. (2024). Mixtral of experts.

LABRAK Y., BAZOGE A., DUFOUR R., DAILLE B., GOURRAUD P.-A., MORIN E. & ROUVIER M. (2022). FrenchMedMCQA : A French multiple-choice question answering dataset for medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, p. 41–46, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (2023). DrBERT : Un modèle robuste pré-entraîné en français pour les domaines biomédical et clinique. In C. SERVAN & A. VILNAT, Éds., *18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, p. 109–120, Paris, France : ATALA. HAL : hal-04130214.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W.-T., ROCKTÄSCHEL T. *et al.* (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. volume 33, p. 9459–9474.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized BERT pretraining approach. *CoRR*, **abs/1907.11692**.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SHAZEER N., MIRHOSEINI A., MAZIARZ K., DAVIS A., LE Q., HINTON G. & DEAN J. (2017). Outrageously large neural networks : The sparsely-gated mixture-of-experts layer.

SUKHBAATAR S., GOLOVNEVA O., SHARMA V., XU H., LIN X. V., ROZIÈRE B., KAHN J., LI D., TAU YIH W., WESTON J. & LI X. (2024). Branch-train-mix : Mixing expert llms into a mixture-of-experts llm.

TOUCHENT R., ROMARY L. & DE LA CLERGERIE E. (2023). Camembert-bio : a tasty french language model better for your health.

WU Z., ARORA A., WANG Z., GEIGER A., JURAFSKY D., MANNING C. D. & POTTS C. (2024). Reft : Representation finetuning for language models. *arXiv preprint arXiv :2404.03592*.