Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

IR-SenTransBio: Modèles Neuronaux Siamois pour la Recherche d'Information Biomédicale

Safaa Menad[1]
(1) Univ. Rouen Normandie, LITIS UR4108, 76000, Rouen
safaa.menad1@univ-rouen.fr

RÉSUMÉ

L'entraînement de modèles transformeurs de langages sur des données biomédicales a permis d'obtenir des résultats prometteurs. Cependant, ces modèles de langage nécessitent pour chaque tâche un affinement (fine-tuning) sur des données supervisées très spécifiques qui sont peu disponibles dans le domaine biomédical. Dans le cadre de la classification d'articles scientifiques et les réponses aux questions biomédicales, nous proposons d'utiliser de nouveaux modèles neuronaux siamois (sentence transformers) qui plongent des textes à comparer dans un espace vectoriel. Nos modèles optimisent une fonction objectif d'apprentissage contrastif auto-supervisé sur des articles issus de la base de données bibliographique MEDLINE associés à leurs mots-clés MeSH (Medical Subject Headings). Les résultats obtenus sur plusieurs benchmarks montrent que les modèles proposés permettent de résoudre ces tâches sans exemples (zero-shot) et sont comparables à des modèles transformeurs biomédicaux affinés sur des données supervisés spécifiques aux problèmes traités. De plus, nous exploitons nos modèles dans la tâche de la recherche d'information biomédicale. Nous montrons que la combinaison de la méthode BM25 et de nos modèles permet d'obtenir des améliorations supplémentaires dans ce cadre.

ABSTRACT

IR-SenTransBio : Siamese Neural Networks for Biomedical Information Retrieval

Training transformers models on biomedical data has yielded promising results. However, these language models require fine-tuning on very specific supervised data for each task, which are not widely available in the biomedical domain. In the context of document classification and question answering in the biomedical domain, we propose to use new Siamese neural models (sentence transformers) that embed texts to be compared in a vector space. The proposed models optimize an objective self-supervised contrastive learning function on articles from the MEDLINE bibliographic database associated to their MeSH (Medical Subject Headings) keywords. The obtained results on several benchmarks show that the proposed models can solve these tasks without examples (zero shot) and are comparable to biomedical transformers fine-tuned on supervised data specific to the problem at hand. Moreover, our models are exploited in biomedical information retrieval task. We show that the combination of BM25 and our models improves biomedical information retrieval.

MOTS-CLÉS : Modèles de Langage · Transformeurs · Apprentissage Contrastif · Modèles Neuronaux Simaois · Apprentissage sans Exemple · Apprentissage auto-supervisé · Recherche d'Information · Classification de Documents · Réponses aux Questions · Textes Biomédicaux.

KEYWORDS: Language Models · Transformers · Contrastive Learning · Siamese Neural Networks · Zero-shot Learning · Self-supervised Learning · Information Retrieval . Document Classification ·

Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 131–142.

1 Introduction

Le développement de modèles transformeurs pré-entraînés, tels que BERT (Bidirectional Encoder Representations from Transformers) (Devlin *et al.*, 2019), a permis d'améliorer les performances du traitement automatique du langage (TAL). L'abondance de données biomédicales disponibles, comme les articles scientifiques, a aussi rendu possible l'entraînement de ces modèles sur des corpus de textes (p. ex. documents, dossiers cliniques de patients) pour des applications biomédicales de prédiction (Alsentzer *et al.*, 2019; Lee *et al.*, 2020; Liu *et al.*, 2021). Ces modèles de langage nécessitent cependant un affinement (fine-tuning) pour chaque tâche sur des données supervisées très spécifiques et rarement disponibles, ce qui limite fortement leur usage en pratique. Comme la plupart des tâches de TAL biomédical (p. ex. extraction de relations, classification de documents, questions-réponses) peuvent se réduire au calcul d'une mesure de similarité sémantique entre deux textes (p. ex. catégorie/résumé d'un article, requête/résultats, question/réponse), nous proposons ici de construire un nouveau modèle transformeur siamois (sentence transformeur) IR-SenTransBio (Information Retrieval-Sentence Transformer in Biomedical data) pré-entraîné qui plonge des paires de textes sémantiquement liés (longs et courts) dans un même espace de représentation vectoriel. En plus d'être applicable à plusieurs types de tâches de TAL, un modèle siamois a aussi l'avantage de permettre de gagner du temps lors de son utilisation en précalculant les représentations vectorielles des textes. Par exemple, en recherche documentaire, un modèle siamois peut permettre de précalculer et d'indexer les représentations vectorielles des textes du corpus ciblé pour n'en calculer que la représentation des requêtes lorsqu'elles sont soumises au moteur, contrairement aux modèles transformeurs "affinés" qui prennent en entrée la combinaison de toutes les paires de textes à comparer. Grâce à ce modèle, nous souhaitons : i) éviter les coûts engendrés par l'étiquetage des données, les calculs d'entraînement et d'affinement ; et ii) réduire considérablement ceux de la prédiction en proposant un modèle auto-supervisé de référence directement applicable à un large éventail de tâches biomédicales.

Dans ce contexte, nous comparons plusieurs modèles transformeurs siamois que nous avons entraînés sur des paires de textes formées, d'une part, de résumés du corpus d'articles biomédicaux PubMed [1], et d'autre part, des mots-clés MeSH (Medical Subject Headings) [2] qui leur sont associés. Nous utilisons une fonction objectif d'apprentissage contrastif auto-supervisé. Étant donnée une paire de textes (résumé, mots-clés), le modèle doit prédire laquelle, parmi un ensemble d'autres paires de textes échantillonnées au hasard, lui est réellement associée dans PubMed. Nous montrons ensuite expérimentalement sur plusieurs benchmarks biomédicaux que sans affinement pour une tâche spécifique, notre meilleur modèle siamois pré-entraîné permet, sans exemples d'apprentissage (zero shot), de classer des documents et de répondre à des questions, et cela avec des résultats comparables aux modèles transformeurs biomédicaux ou encore généralistes affinés sur des données supervisées spécifiques aux problèmes traités. En dernier lieu, nous mettons en pratique ces modèles en les exploitant pour la recherche d'information biomédicale. Pour évaluer notre approche, nous l'appliquons sur deux corpus biomédicaux, à savoir TREC-COVID (Voorhees *et al.*, 2021) et NFCorpus (Boteva *et al.*, 2016). Nous montrons aussi que la combinaison de la méthode BM25 avec nos modèles permet d'améliorer les performances de recherche d'information.

1. https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/
2. Le MeSH https://www.nlm.nih.gov/mesh/ est un thésaurus spécialisé du domaine biomédical composé de 30 000 descripteurs utilisés pour l'indexation d'articles PubMed.

Cet article est structuré de la manière suivante : dans la section 2, nous proposons une revue de la littérature qui explore les performances des transformeurs dans la tâche de recherche d'information (RI). La section 3 présente en détail les modèles transformeurs pré-entraînés et leur application dans les modèles siamois. Nous décrivons nos propres modèles siamois dans la section 4. Les résultats obtenus sur des benchmarks de référence sont exposés dans la section 5. Dans la section 6, nous détaillons la tâche de recherche d'information biomédicale et analysons les performances de nos modèles dans ce domaine. Enfin, nous concluons et ouvrons des perspectives de recherche dans la section 7.

2 Travaux similaires

L'efficacité des modèles de langage dans la tâche de recherche d'information a été abordée par plusieurs études. Ces travaux se sont intéressés à évaluer les performances des modèles existants dans la recherche, la classification et la récupération d'informations pertinentes. Des approches basées sur des modèles tels que BERT, Sentence-BERT et d'autres transformeurs ont été explorées pour améliorer la précision et la pertinence des résultats.

Dans (Nguyen *et al.*, 2022), les auteurs créent un nouveau corpus en langue vietnamienne et proposent un système de question-réponse à deux étapes pour ce corpus. Le système récupère des documents pertinents en se basant sur la question d'entrée à l'aide d'un moteur de recherche basé sur le TF-IDF. Dans la deuxième étape, le système utilise un modèle S-BERT affiné pour identifier la phrase la plus pertinente dans les documents récupérés et pour extraire la réponse à la question.

Dans (Yang *et al.*, 2020), les auteurs utilisent également les bi-encodeurs pour la tâche de la RI et montrent que cette approche permet d'obtenir de très bons résultats sur différents benchmarks dans un domaine général.

(Tinn *et al.*, 2021) propose l'application des transformeurs tels que BERT et GPT sur des tâches de TAL biomédicales comme la reconnaissance d'entités nommées et l'extraction des relations. Les résultats qu'ils ont obtenu montrent l'avantage de réajustement de ces modèles sur ces tâches.

Dans (Soni & Roberts, 2020), les auteurs évaluent les performances de plusieurs modèles pré-entraînés et de modèles affinés sur une tâche de question-réponse clinique en utilisant l'ensemble de données MedQuAD. Ils comparent les performances de ces modèles à plusieurs méthodes de base et analysent l'impact de la sélection de l'ensemble de données sur les performances. Ils cherchent à identifier les ensembles de données de pré-entraînement et d'affinement les plus efficaces pour cette tâche.

Les auteurs de (Chakraborty *et al.*, 2020) proposent un modèle pré-entraîné sur le corpus BREATHE (corpus médical) pour la tâche de question-réponse. Ils montrent que l'entraînement de ces modèles sur des données biomédicales leur permettent de dépasser les modèles de base comme BERT.

3 Les Transformeurs

Les transformeurs sont des réseaux neuronaux basés sur le mécanisme d'auto-attention multi-têtes améliorant l'efficacité de l'apprentissage des modèles de grande taille. Il est composé d'un encodeur qui transforme le texte d'entrée en vecteur, et d'un décodeur qui transforme ce vecteur en texte en

sortie. Le mécanisme d'attention fournit de meilleures performances grâce à la modélisation des liens entre les éléments d'entrée et de sortie. Un modèle de langage pré-entraîné (MLP) est un réseau neuronal entraîné sur une grande quantité de données non annotées de manière non supervisée. Le modèle est ensuite transféré pour une tâche de TAL cible (downstream task), où un ensemble de données annotées plus petit et spécifique à la tâche est utilisé pour affiner le MLP permettant ainsi de construire le modèle final capable d'exécuter la tâche cible (ajustement d'un MLP).

3.1 Modèles pré-entraînés

Les modèles de langage pré-entraînés, tels que BERT, ont conduit à des gains impressionnants dans de nombreuses tâches de TAL. Les travaux existants traitent des données généralistes. Dans les tâches de TAL biomédicales, le pré-entraînement sur les textes de PubMed par exemple a permis d'obtenir de meilleures performances (Beltagy *et al.*, 2019; Lee *et al.*, 2020; Peng *et al.*, 2019a). L'approche standard de pré-entraînement d'un modèle biomédical débute avec un modèle généraliste et poursuit le pré-entraînement en utilisant un corpus biomédical. Par exemple, BioBERT (Lee *et al.*, 2020) utilise pour cela les résumés extraits de PubMed et les articles en texte intégral de PubMed Central (PMC). BlueBERT (Peng *et al.*, 2019b) utilise à la fois le texte de PubMed et les notes cliniques MIMIC-III (Medical Information Mart for Intensive Care) (Johnson *et al.*, 2016). SciBERT (Beltagy *et al.*, 2019) constitue une exception, le pré-entraînement est réalisé à partir de zéro, en utilisant la littérature scientifique.

3.2 Modèles siamois

Les transformeurs de paires de phrases (sentence-transformers) sont des modèles développés pour la tâche de calcul d'un score de similarité entre deux phrases (p. ex.calcul de similarité sémantique entre phrases, recherche d'informations, reformulation de phrases etc). Ces transformeurs sont basés sur deux architectures : i) les cross-encodeurs, qui traitent la concaténation de la paire ; et ii) les modèles siamois bi-encodeurs, qui encodent en vecteur chacun des éléments de la paire. Par exemple, Sentence-BERT (Reimers & Gurevych, 2019) est un bi-encodeur basé sur BERT permettant de générer des plongements de phrases sémantiquement significatifs à utiliser dans des comparaisons de similarité textuelle. Pour chaque entrée (voir figure 1), le modèle produit un vecteur de taille fixe (u et v). La fonction objectif est choisie de façon à ce que l'angle entre les deux vecteurs u et v soit d'autant plus faible que les entrées sont similaires. Plus précisément, la fonction objectif utilise le cosinus de l'angle : $cos(u, v) = \frac{u.v}{||u||||v||}$, si $cos(u, v) = 1$ alors les phrases sont similaires et si $cos(u, v) = 0$ alors les phrases n'ont aucune relation sémantique. D'autres modèles de plongement de phrases ont été développés (Gao *et al.*, 2021; Wang *et al.*, 2021; Cohan *et al.*, 2020). Parmi eux MiniLM-L6-v25 [3] qui est un bi-encodeur basé sur une version simplifiée de MiniLM (Wang *et al.*, 2020). Ce modèle, rapide et de petite taille, a permis d'obtenir de bonnes performances sur différentes tâches pour 56 corpus (Muennighoff *et al.*, 2022).

3. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

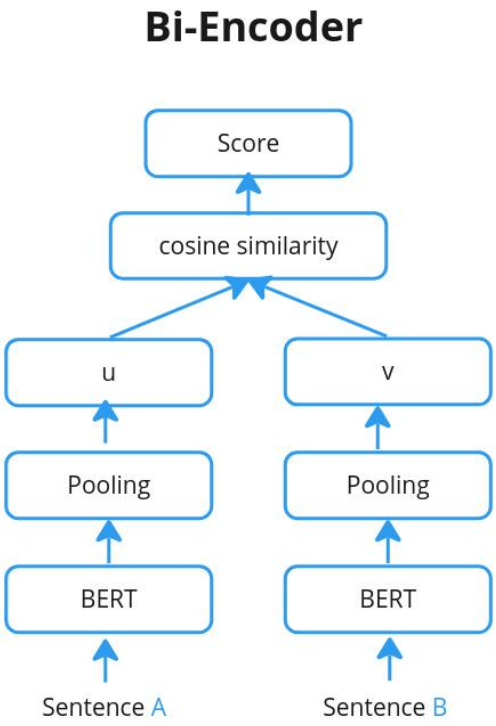

FIGURE 1 – Un encodeur siamois.

4 Modèles de Langage Proposés

Les transformeurs siamois donnent de bons résultats dans des domaines généralistes, mais pas dans les domaines de spécialité, comme le domaine biomédical (Muennighoff *et al.*, 2022). Nous proposons ici de nouveaux modèles siamois pré-entraînés sur le corpus PubMed. Les transformeurs siamois ont été initialement conçus pour transformer des phrases (de taille similaire) en vecteurs. Nous proposons dans notre approche de transformer dans le même espace vectoriel les termes MeSH, les titres et les résumés des articles PubMed en entraînant un modèle de transformeur siamois sur ces données que nous avons préparées. Nous voulons nous assurer qu'il y a une correspondance dans cet espace vectoriel entre le texte court et le texte long. Nous avons donc entraîné nos modèles avec des paires d'entrées (titre, terme MeSH) et (résumé, terme MeSH). À partir de ces données, nous avons construit deux types de modèles : le premier type est nos propres transformeurs siamois (BioSTransformers) construits à partir d'un transformeur pré-entraîné sur des données biomédicales et le second est un transformeur siamois déjà pré-entraîné sur des données généralistes (BioS-MiniLM).

BioSTransformers. Pour ce type, nous nous sommes inspirés du modèle Sentence-BERT (Reimers & Gurevych, 2019) en remplaçant BERT par d'autres transformeurs. Nous avons utilisé des transformeurs qui ont été entraînés sur des données biomédicales (bio-transformeurs). Pour construire les transformeurs siamois (BioSTransformers), nous avons ajouté une couche de pooling et modifié la fonction objectif. Ensuite, nous les avons entraîné sur nos données. La couche de pooling calcule le vecteur moyen des vecteurs de sortie du transformeur (token embeddings). Les deux textes en entrée passent successivement dans le transformeur produisant deux vecteurs u et v en sortie du pooling, qui sont par la suite utilisées par la fonction objectif. Parmi les bio-transformeurs disponibles, nous avons

sélectionné les modèles les plus performants : BlueBERT (Peng *et al.*, 2019b), PubMedBERT (Gu *et al.*, 2022) et BioELECTRA (Kanakarajan *et al.*, 2021). Ces modèles ont été entraînés sur PubMed, à part BlueBERT qui a également été entraîné sur des notes cliniques.

BioS-MiniLM. Pour ce modèle nous avons adapté un transformeur siamois pré-entraîné sur des données généralistes. Plusieurs modèles généraux de sentence-transformer déjà pré-entraînés sont disponibles [4]. Ils diffèrent en taille, vitesse et performance. Parmi ceux qui obtiennent les meilleures performances, nous avons affiné MiniLM-L6-v2 (voir section 3) qui a été pré-entraîné sur 32 corpus généralistes (Reddit comments, S2ORC, WikiAnswers etc.).

Fonction objectif. Pour un transformeur de paires de phrases classique on dispose de données supervisées sous forme de triplets (phrase 1, phrase 2, score de similarité entre les deux phrases). Dans notre cas cependant, nous ne disposons d'aucun score pour les résumés ou les titres et leurs termes MeSH correspondants. Nous considérons donc qu'un résumé, un titre et les termes MeSH associés au même article (identifié par un PMID) sont similaires (le score est égal à 1) et inversement, qu'un résumé ou un titre avec des termes MeSH qui ne sont pas associés au même article ne sont pas similaires (le score est égal à 0).

Nous utilisons une fonction objectif d'apprentissage contrastif auto-supervisé basée sur la fonction de perte de classement négatif multiple (Henderson *et al.*, 2017) dite MNRL (Multiple Negative Ranking Loss) dans le package Sentence-Transformers [5]. Cette fonction permet à un modèle d'apprendre à partir de données non étiquetées en utilisant la comparaison de paires d'exemples similaires et différents. Elle vise à maximiser la similarité entre les représentations de deux exemples similaires et à minimiser la similarité entre les représentations de deux exemples différents. La MNRL n'a besoin que des paires positives en entrée (le titre ou le résumé et un terme MeSH associé à l'article dans notre cas). Pour une paire positive (titre_i ou résumé_i, MeSH_i), la MNRL considère que chaque paire (titre_i ou résumé_i, MeSH_j) avec $i \neq j$ dans le même batch est négative. Comme un article peut être associé à plusieurs termes MeSH, nous avons fait en sorte dans la génération des batchs qu'un résumé (ou un titre) associé à un terme MeSH dans PubMed ne soient jamais considérés comme une paire négative.

5 Expérimentations et Résultats

5.1 Expérimentations

Dans un premier temps, pour tester les différents transformeurs et la fonction objectif à retenir, nous n'avons utilisé que les titres et nous avons réduit le nombre de termes MeSH. Au total 1 402 termes MeSH et 3,79 millions de paires (titre, MeSH) ont été sélectionnées. Pour la validation nous avons utilisé 18 940 articles avec leurs titres et termes MeSH.

Dans un second temps, une fois sélectionnés les modèles transformeurs et la fonction objectif MNRL, nous avons évalué nos modèles BioSTransformers et BioS-MiniLM sur les paires (titre, MeSH) et (résumé, MeSH) générés à partir de tous les termes MeSH de PubMed. Ayant constaté qu'il n'était pas nécessaire d'utiliser toutes les paires des 35 millions d'articles de PubMed, nous avons sélectionné

4. https://huggingface.co/sentence-transformers
5. https://www.sbert.net/docs/package_reference/losses.html#multiplenegativesrankingloss

6,75 millions de paires pour le fine-tuning. Un total de 18 557 articles sert à la validation.

Les deux tâches de TAL, ainsi que les données utilisées sont décrites ci-après :

1. La classification de documents : le corpus Hallmarks of Cancer (HoC) est constitué de 1 852 résumés de publications PubMed annotés manuellement par des experts selon une taxonomie qui est composée de 37 classes. Chaque phrase du corpus se voit attribuer zéro à plusieurs classes (Hanahan & Weinberg, 2000) ;
2. Questions-réponses (QA) :
   (a) PubMedQA est un corpus pour les réponses aux questions spécifiques à la recherche biomédicale. Il contient un ensemble de questions, ainsi qu'un champ annoté indiquant si le texte contient la réponse à la question de recherche (Jin *et al.*, 2019) ;
   (b) BioASQ est un corpus qui contient plusieurs tâches de QA avec des données annotées par des experts, y compris des questions oui/non, de liste et de résumés. Nous nous concentrons sur le type de questions oui/non (tâche 7b) (Nentidis *et al.*, 2019).

Nous considérons les deux tâches précédentes comme un problème de similarité de textes : pour chaque requête nous considérons les k résultats les plus proches, k étant le nombre de résultats attribués à la requête par l'expert (gold standard). La similarité entre la requête et les résultats est mesurée par la similarité cosinus entre le vecteur de la requête et les vecteurs des résultats. Dans une tâche de classification, la requête est la catégorie et les résultats sont les documents classés dans cette catégorie. Dans une tâche de questions-réponses, la requête est la question et les résultats sont une réponse.

5.2 Résultats

Nos modèles sont évalués selon le score F1 utilisé dans les benchmarks HoC, PubmedQA et BioASQ dans (Gu *et al.*, 2022). Les résultats obtenus par nos modèles transformeurs siamois sans exemple (sans fine-tuning) sont synthétisés dans le Tableau 1, avec en gras les meilleurs scores.

Corpus/modèle	BioS-MiniLM	S-BioELECTRA	S-PubMedBERT	S-BlueBERT
HoC	0,492	**0,499**	0,489	0,468
PubMedQA	0,649	0,675	**0,729**	0,652
BioASQ	0,747	0,694	**0,751**	0,713

TABLE 1 – Résultats d'évaluation de nos modèles sur différents benchmarks selon le F1 score.

Le Tableau 2 résume les résultats obtenus sur les mêmes tâches par des modèles affinés spécifiquement à ces tâches (Gu *et al.*, 2022). Pour chaque benchmark, ces modèles sont affinés avec les données supervisées disponibles dans chaque cas. Ces résultats montrent que les modèles que nous proposons permettent de réaliser ces tâches avec des résultats comparables à des modèles biomédicaux affinés sur des données supervisés spécifiques aux problèmes traités, mais que nous n'avons pas utilisés dans notre approche sans exemple.

Pour le benchmark HoC, les résultats obtenus par notre meilleur modèle S-BioELECTRA sont très en dessous des résultats obtenus par PubMedBERT+aff (0,499 vs. 0,823). En effet, les modèles de (Gu *et al.*, 2022) ont été affinés spécifiquement pour chaque tâche, notamment la classification des documents, en modifiant l'architecture du modèle et en ajoutant des couches spécifiques pour

Corpus/modèle	BERT +aff	RoBERTa +aff	BioBERT +aff	SciBERT +aff	ClinicalBERT +aff	BlueBERT +aff	PubMedBERT +aff
HoC	0,802	0,797	0,815	0,812	0,807	0,805	**0,823**
PubmedQA	0,516	0,528	**0,602**	0,574	0,491	0,484	0,558
BioASQ	0,744	0,752	0,841	0,789	0,685	0,687	**0,876**

TABLE 2 – Résultats d'évaluation des modèles affinés (+aff) spécifiquement à ces tâches sur différents benchmarks selon le F1 score.

chaque cas. En revanche, pour le benchmark PubMedQA, les résultats obtenus par notre meilleur modèle S-PubMedBERT dépassent les résultats obtenus par BioBERT+aff (0,729 vs. 0,602). Enfin, pour le benchmark BioASQ, les résultats obtenus par notre meilleur modèle S-PubMedBERT sont comparables aux résultats obtenus par les modèles affinés même si PubMedBERT+aff donne de meilleurs résultats (0,751 vs. 0,876). Tout cela a été obtenu sans réadapter l'architecture de nos modèles pour chaque tâche et sans les affiner sur les données spécifiques aux benchmarks cités.

Les modèles BioSTransformers obtiennent de meilleurs résultats que le BioS-MiniLM, cela s'explique par le fait que le modèle BioS-MiniLM a été pré-entraîné sur des données généralistes, tandis que les autres modèles ont été pré-entraînés sur des données biomédicales spécialisées. Cela démontre l'importance de la phase de pré-entraînement.

6 Recherche d'Information Biomédicale

Après avoir évalué nos modèles sur deux tâches de NLP dans le domaine biomédical, dans cette section, nous allons les appliquer sur la tâche de la RI. Bien que les méthodes neuronales ont surpassé les approches traditionnelles de la RI telles que TF-IDF (Term Frequency - Inverse Document Frequency) et BM25 dans des domaines généralistes, elles demeurent cependant insuffisantes dans le domaine biomédical. Dans cette partie, nous proposons d'améliorer la recherche d'informations biomédicale avec nos modèles siamois proposés.

Expérimentations

Nous considérons la tâche de la RI comme une recherche de proximité entre les représentations d'une requête et des documents. Nous utilisons nos modèles pour plonger les deux entrées et calculer le score de similarité. Nous testons nos modèles sur deux corpus biomédicaux TREC-COVID (Voorhees *et al.*, 2021) et NFCorpus (Boteva *et al.*, 2016).

TREC-COVID. Le jeu de données TREC-COVID est une collection d'articles scientifiques liés à la COVID-19. Il a été créé dans le cadre de la tâche de RI de la conférence Text REtrieval Conference (TREC) COVID, qui visait à soutenir la communauté scientifique dans sa réponse à la pandémie de COVID-19 par des systèmes de RI efficaces. Il est composé de 171 000 articles et de 50 requêtes.

NFCorpus. NFCorpus est un ensemble de données d'extraction de texte intégral en anglais pour la RI biomédicale qui concerne la nutrition. Il contient un total de 3 244 requêtes en langage naturel

(extraites du site NutritionFacts.org) avec 169 756 jugements de pertinence extraits automatiquement pour 9 964 documents médicaux provenant principalement de PubMed.

Corpus/modèle	BM25	BioS-MiniLM	BM25+BioS-MiniLM	BM25+S-BERT
TREC-COVID	0,616	0,478	0,616	**0,656**
NFCorpus	**0,300**	0,282	0,283	0,262

TABLE 3 – Résultats d'évaluation de nos modèles sur les données biomédicales selon NDCG@10.

Le Tableau 3 recense les résultats obtenus avec nos modèles et des modèles de base selon la métrique d'évaluation NDCG@10 (Wang *et al.*, 2013).

Il Compare les performances de nos modèles avec le modèle BM25+S-BERT utilisé comme un modèle de reclassement qui donne les meilleurs résultats avec TREC-COVID dans le benchmark BEIR [24]. Dans BM25+S-BERT et BM25+BioS-MiniLM, les 100 meilleurs résultats donnés par la fonction de classement BM25 sont réordonnés en utilisant le modèle S-BERT [6].

D'après le tableau nous pouvons déduire que BM25 (baseline) obtient de bons résultats sur les deux jeux de données biomédicales. Cependant, cette méthode se base sur une recherche lexicale qui dépend de la fréquence des termes et ne capte pas la sémantique des phrases. Notre modèle ne dépasse pas la BM25. Cela peut s'expliquer par le fait que notre modèle est entraîné sur la récupération de documents à la base de mots clés très spécifiques (MeSH) et non pas de requêtes en langage naturel général. Cependant, sur le corpus TREC-COVID, BM25 et notre méthode peuvent se compléter : leur utilisation conjointe permet de légèrement améliorer les résultats et de bénéficier de l'avantage des deux méthodes. BM25+S-BERT obtient mieux car le modèle S-BERT a été pré-entraîné sur le corpus MSMARCO [7], un corpus volumineux spécialement conçu pour l'entraînement des modèles dédiés à la tâche de recherche d'information. Contrairement à notre modèle qui a été pré-entraîné sur un corpus généraliste qui n'est pas dédié à la RI. Le principal avantage de notre méthode est qu'elle permet de calculer les représentations des documents et de les indexer auparavant. Le modèle calcule la représentation de la requête soumise au moteur et récupère directement les documents les plus pertinents dans un temps réduit.

7 Conclusion

Dans cette étude, nous avons proposé de nouveaux modèles neuronaux siamois pour créer des représentations vectorielles de textes. Ces représentations sont ensuite utilisées pour calculer la similarité entre les textes. Nous avons testé nos modèles sur les deux tâches de classification et de question-réponse dans le domaine biomédical et nous avons montré qu'ils permettaient d'obtenir des résultats comparables à ceux des modèles qui ont été spécifiquement conçus pour ces tâches et entraînés sur des données correspondantes. Nous avons ensuite appliqué nos modèles sur la tâche de recherche d'information biomédicale, en utilisant deux corpus couramment utilisés dans ce domaine. Nous avons montré que l'utilisation de nos modèles en combinaison avec une méthode lexicale traditionnelle BM25 conduit à de meilleurs résultats et permet de bénéficier de l'avantage des deux types de méthodes. Ces résultats comparables et encourageants nous permettent d'envisager d'étendre notre approche en entraînant nos modèles sur des données biomédicales contenant des mots clés différents

6. sentence-transformers/msmarco-distilbert-base-v4
7. https ://github.com/microsoft/MSMARCO-Passage-Ranking

de ceux actuellement utilisés. Nous prévoyons également d'ajuster nos modèles spécifiquement pour la tâche de RI en incluant des étapes supplémentaires dans le processus d'entraînement. Enfin, nous étendrons la RI à d'autres données et à la recherche de dossiers patients pour permettre la constitution de cohortes dans les études cliniques.

Remerciements

Ce travail de thèse est fait sous la supervision de Fatima Lina SOUALMIA (TIBS, LITIS, Université de Rouen Normandie) et Saïd ABDEDDAIM (TIBS, LITIS, Université de Rouen Normandie).

Références

ALSENTZER E., MURPHY J., BOAG W., WENG W.-H., JINDI D., NAUMANN T. & MCDERMOTT M. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, p. 72–78, Minneapolis, Minnesota, USA : Association for Computational Linguistics. DOI : 10.18653/v1/W19-1909.

BELTAGY I., LO K. & COHAN A. (2019). SciBERT : A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3615–3620.

BOTEVA V., GHOLIPOUR D., SOKOLOV A. & RIEZLER S. (2016). A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval : 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, p. 716–722 : Springer.

CHAKRABORTY S., BISONG E., BHATT S., WAGNER T., ELLIOTT R. & MOSCONI F. (2020). Biomedbert : A pre-trained biomedical language model for qa and ir. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 669–679.

COHAN A., FELDMAN S., BELTAGY I., DOWNEY D. & WELD D. S. (2020). Specter : Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 2270–2282.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, p. 4171–4186.

GAO T., YAO X. & CHEN D. (2021). Simcse : Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 6894–6910.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2022). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, **3**(1), 1–23. DOI : 10.1145/3458754.

HANAHAN D. & WEINBERG R. A. (2000). The hallmarks of cancer. *Cell*, **100**(1), 57–70.

HENDERSON M., AL-RFOU R., STROPE B., SUNG Y.-H., LUKÁCS L., GUO R., KUMAR S., MIKLOS B. & KURZWEIL R. (2017). Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv :1705.00652*.

JIN Q., DHINGRA B., LIU Z., COHEN W. & LU X. (2019). PubMedQA : A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 2567–2577.

JOHNSON A. E., POLLARD T. J., SHEN L., LEHMAN L.-W. H., FENG M., GHASSEMI M., MOODY B., SZOLOVITS P., ANTHONY CELI L. & MARK R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, **3**(1), 1–9.

KANAKARAJAN K. R., KUNDUMANI B. & SANKARASUBBU M. (2021). BioELECTRA : Pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, p. 143–154, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.bionlp-1.16.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2020). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240.

LIU F., SHAREGHI E., MENG Z., BASALDELLA M. & COLLIER N. (2021). Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4228–4238.

MUENNIGHOFF N., TAZI N., MAGNE L. & REIMERS N. (2022). Mteb : Massive text embedding benchmark. *arXiv preprint arXiv :2210.07316*.

NENTIDIS A., BOUGIATIOTIS K., KRITHARA A. & PALIOURAS G. (2019). Results of the seventh edition of the BioASQ challenge. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, p. 553–568 : Springer.

NGUYEN N. T.-H., HA P. P.-D., NGUYEN L. T., VAN NGUYEN K. & NGUYEN N. L.-T. (2022). Spbertqa : A two-stage question answering system based on sentence transformers for medical texts. In *Knowledge Science, Engineering and Management : 15th International Conference, KSEM 2022, Singapore, August 6–8, 2022, Proceedings, Part II*, p. 371–382 : Springer.

PENG Y., YAN S. & LU Z. (2019a). Transfer learning in biomedical natural language processing : An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, p. 58–65.

PENG Y., YAN S. & LU Z. (2019b). Transfer learning in biomedical natural language processing : An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, p. 58–65, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-5006.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3982–3992, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

SONI S. & ROBERTS K. (2020). Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 5532–5538.

TINN R., CHENG H., GU Y., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2021). Fine-tuning large neural language models for biomedical natural language processing. *arXiv preprint arXiv :2112.07869*.

VOORHEES E., ALAM T., BEDRICK S., DEMNER-FUSHMAN D., HERSH W. R., LO K., ROBERTS K., SOBOROFF I. & WANG L. L. (2021). Trec-covid : constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, p. 1–12 : ACM New York, NY, USA.

WANG K., REIMERS N. & GUREVYCH I. (2021). Tsdae : Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics : EMNLP 2021*, p. 671–688.

WANG W., WEI F., DONG L., BAO H., YANG N. & ZHOU M. (2020). Minilm : Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, **33**, 5776–5788.

WANG Y., WANG L., LI Y., HE D. & LIU T.-Y. (2013). A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, p. 25–54 : PMLR.

YANG Y., CER D., AHMAD A., GUO M., LAW J., CONSTANT N., ABREGO G. H., YUAN S., TAR C., SUNG Y.-H. *et al.* (2020). Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 87–94.