Université Mohammed Premier OUJDA

Faculté des Sciences

Centre d'Etudes Doctorales Sciences et Techniques

N° d'ordre : 826/23

THÈSE

Pour l'obtention du grade de :

DOCTEUR EN SCIENCES

Formation Doctorale : Mathématiques et Informatique

Spécialité : Informatique

Présentée par :

Zakaria KADDARI

TITRE

Amélioration des performances des systèmes "Question Answering" dans le domaine biomédical

Soutenue le samedi 13/05/2023 devant le jury :

Pr. Mohammed RAHMOUNE	ENSAO	Président
Pr. El Habib NFAOUI	Faculté des Sciences de Fès	Rapporteur
Pr. Ilhame El FARISSI	ENSAO	Rapporteur
Pr. Ilham SLIMANI	Faculté des Sciences d'Oujda	Rapporteur
Pr. Mohammed BOURHALEB	ESTO	Examinateur
Pr. Jamal BERRICH	ENSAO	Co-directeur de thèse
Pr. Toumi BOUCHENTOUF	ENSAO	Directeur de thèse

DÉDICACES

Ce travail est dédié à la mémoire de mon défunt père, que Dieu lui garde dans son vaste paradis, qu'il apprécie cet humble geste comme preuve de reconnaissance de la part d'un fils qui a toujours prié pour le salut de son âme.

A ma très chère mère et à mon très cher défunt père,

Aucune dédicace ne saurait exprimer mon respect, mon amour éternel et ma considération pour les sacrifices que vous avez consenti pour mon instruction et mon bien-être, je vous remercie pour tout le soutien et l'amour que vous me portez depuis mon enfance et j'espère que votre bénédiction m'accompagne toujours.

A ma très chère épouse Sara,

Tu es la femme de ma vie, mon âme-sœur et la lumière de mon chemin.

Sans toi, cette thèse n'aurait jamais pu voir le jour.

Aucune dédicace ne pourrait exprimer mon amour et mon attachement à toi.

A ma très chère fille Houyam et mon très cher fils Firas

Puisse Dieu vous protéger et vous procurer santé et longue vie.

A mes très chers frères Amine et Hicham

Puisse ce travail témoigner de ma profonde affection et de ma sincère estime.

A tous mes amis et à mes chers professeurs de mon cursus scolaire

Merci beaucoup pour votre aide et pour votre présence dans ma vie.

REMERCIEMENTS

En premier lieu, je tiens à remercier mon directeur de thèse, Monsieur Toumi BOUCHENTOUF, pour son encadrement avisé, sa méthode de travail efficace, ses conseils et ses remarques toujours très pertinentes, et pour tous ces efforts qui ont amené à réussir cette thèse.

Ma reconnaissance aussi au professeur Mohammed RAHMOUN qui m'a donné l'occasion d'intégrer le laboratoire LaRSA (Laboratoire des Recherches en Sciences Appliquées) à l'Ecole Nationale Sciences Appliquées d'Oujda.

Je tiens à remercier vivement les professeurs Nfaoui El Habib, El Farissi Ilhame et Slimani Ilham d'avoir accepté de rapporter cette thèse. J'adresse aussi mes sincères remerciements à tous les membres du jury qui m'ont fait l'honneur d'accepter de lire et d'expertiser mon travail.

Je souhaite remercier aussi tous mes collègues au laboratoire LaRSA. Je leur exprime ma profonde sympathie.

RÉSUMÉ

La recherche en médecine et biologie se développe rapidement, ce qui se traduit par une augmentation considérable des articles de recherche biomédicale. Trouver des informations pertinentes dans cette littérature en pleine expansion devient de plus en plus difficile pour les chercheurs et les professionnels de la santé. Les moteurs de recherche classiques tels que PubMed, bien que très utiles, se contentent de renvoyer une liste de documents contenants probablement la réponse à une question. De l'autre côté, les systèmes question-réponse, en anglais « Question Answering » (QA) peuvent extraire, et même synthétiser des réponses précises à des questions formulées naturellement. Malheureusement, lorsqu'ils sont appliqués directement à la littérature biomédicale, ces modèles QA donnent souvent des résultats insatisfaisants. Notre problématique de recherche en cette thèse est donc l'amélioration des systèmes QA dans le domaine biomédical.

Nous nous sommes focalisés sur trois axes de recherche qui s'inscrivent dans notre problématique. (1) Nouvelles approches pour le QA biomédical, (2) La construction de nouveaux ensembles de données (datasets) QA biomédical, et (3) Couplage d'un moteur de recherche classique avec un modèle QA biomédical. Afin de répondre à ces axes de recherche, nos contributions sont : L'introduction d'une nouvelle approche QA biomédical pour les questions de types factoïde et liste. Notre méthode proposée a donné des résultats de pointe sur plusieurs datasets du challenge BioASQ ; la compétition de référence dans le domaine de QA biomédical. Une autre contribution est l'exploitation de l'apprentissage par transfert pour les questions de types Oui/Non et résumé. Aussi, la construction du premier dataset QA biomédical français. Enfin, La proposition d'une démarche de couplage d'un moteur de recherche classique avec un modèle QA biomédical, que nous avons aussi exploité dans le contexte de la pandémie de COVID-19 pour créer un moteur de recherche biomédical spécial au COVID-19.

Mots-clés : Moteur de recherche, Question Answering, QA, QA biomédical, Dataset, BioASQ, Apprentissage par transfert, COVID-19

ABSTRACT

Research in medicine and biology is growing rapidly, resulting in a dramatic increase in biomedical research articles. Finding relevant information in this rapidly expanding literature is becoming increasingly difficult for researchers and healthcare professionals. Traditional search engines such as PubMed, while very useful, simply return a list of documents that probably contain the answer to a question. On the other hand, Question Answering (QA) systems can retrieve and even synthesize specific answers to naturally formulated questions. Unfortunately, when applied directly to biomedical literature, these QA models often yield unsatisfactory results. Our research problem in this thesis is therefore the improvement of QA systems in the biomedical domain.

We have focused on three research axes that fit our problematic: (1) New approaches for biomedical QA, (2) Construction of new biomedical QA datasets, and (3) Coupling a classical search engine with a biomedical QA model. In order to address these research directions, our contributions are the introduction of a new biomedical QA method for factoid and list questions. Our proposed method yielded state of the art results on several datasets of BioASQ; the most important competition in the field of biomedical QA. Another contribution is the experimentation with transfer learning for Yes/No and summary type questions. Also, The construction of the first French biomedical QA dataset. Finally, the introduction of an a approach for coupling a classical search engine with a biomedical QA model, which we also used in the context of the COVID-19 pandemic to create a special biomedical search engine for COVID-19.

Keywords: Search engine, Question Answering, QA, Biomedical QA, Dataset, BioASQ, Transfer learning, COVID-19

Liste des figures

Liste des tableaux

Liste des abréviations

IA	Intelligence Artificielle
ML	Machine Learning
DL	Deep Learning
NLP	Natural Language Processing
QA	Question Answering
BQA	Biomedical Question Answering
IR	Information Retrieval
NER	Named Entity Recognition
NLG	Natural Language Generation
BERT	Bidirectional Encoder Representations from Transformers
SOTA	State-Of-The-Art
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional LSTM
BIDAF	Bi-Directional Attention Flow
SQuAD	Stanford Question Answering Dataset
RNN	Recurrent Neural Network
DCN	Dynamic Coattention Networks
SOP	Sentence Order Prediction
NSP	Next Sentence Prediction
PLM	Pre-trained Language Model
Acc	Accuracy
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
SAcc	Strict Accuracy
LAcc	Lenient Accuracy
MRR	Mean Reciprocal Rank
ROUGE	Recall-Oriented Understudy for Gist Evaluation
MAP	Mean Average Precision
DME	Dossiers Médicaux Électronique
PMC	PubMed Central
UMLS	Unified Medical Language System
BMKC	Biomedical Knowledge Comprehension
BEST	Biomedical Entity Search Tool
MeSH	Medical Subject Headings
NIH	National Institutes of Health
KB	Knowledge Base
RDF	Resource Description Framework
MRC	Machine Reading Comprehension

QE	Question Entailment
VQA	Visual QA
GCN	Graph Convolutional Networks
NLM	National Library of Medicine
NLI	Natural Language Inference
CV	Computer Vision
CNN	Convolutional Neural Networks
FFN	Feed-Forward Network
BioNER	Biomedical NER
GPU	Graphics Processing Unit
GAN	Generative Adversarial Network
GNN	Graph Neural Network
TF-IDF	Term Frequency-Inverse Document Frequency
BOW	Bag Of Words
DPR	Dense Passage Retrieval

Table des matières

Introduction générale

L'intelligence artificielle (IA) est l'intelligence dont font preuve les machines, par opposition à l'intelligence naturelle dont font preuve les animaux et les humains. La recherche sur l'IA a été définie comme le domaine d'étude des agents intelligents, c'est-à-dire tout système qui perçoit son environnement et prend des mesures qui maximisent ses chances d'atteindre ses objectifs. Les applications de l'IA comprennent les moteurs de recherche Web avancés (par exemple, Google), les systèmes de recommandation (utilisés par YouTube, Amazon et Netflix), la compréhension de la parole humaine (comme Siri et Alexa), les voitures à conduite autonome (par exemple, Tesla), et bien d'autres. Au cours des premières décennies du XXIe siècle, l'apprentissage automatique, en anglais « Machine Learning » (ML) hautement mathématique et statistique a dominé le domaine de l'IA, et s'est avéré très efficace, contribuant à résoudre de nombreux problèmes difficiles dans l'industrie et le monde universitaire.

L'un des sous-domaines les plus importants de l'intelligence artificielle est le traitement automatique de langage naturel, en anglais Natural Language Processing (NLP). C'est aussi un domaine interdisciplinaire qui combine la linguistique, les sciences cognitives et l'informatique. La principale préoccupation de l'NLP est de permettre aux ordinateurs de comprendre et de produire du langage aussi proche que le font les humains. Au cours des trois dernières années, les modèles basés sur l'apprentissage automatique tels que BERT [1] et GPT-3 [2], ont fait des progrès impressionnants dans l'état de l'art de nombreuses tâches NLP.

La réponse aux questions, en anglais « Question Answering » (QA) est une discipline informatique à l'intersection des domaines de la recherche d'information, en anglais « Information Retrieval » (IR) et de l'NLP, qui s'intéresse à la construction de systèmes qui répondent automatiquement aux questions posées par les humains avec un langage naturel. Les programmes QA peuvent construire des réponses à partir de l'interrogation d'une base de connaissances structurée ou d'une collection non structurée de documents en langage naturel. Les systèmes QA sont soit à domaine fermé (répondant aux questions d'un domaine spécifique), soit à domaine ouvert (s'appuyant sur des ontologies générales et des connaissances répandues). Ces systèmes sont capables d'extraire la réponse à une question à partir d'un texte donné. Ceci

est utile pour rechercher une réponse dans un document. Selon le système utilisé, la réponse peut être directement extraite du texte ou générée automatiquement. Les systèmes QA sont souvent utilisés pour automatiser la réponse aux questions fréquemment posées en utilisant une base de connaissances (par exemple, des documents) comme contexte. En tant que tels, ils sont utiles pour les assistants virtuels intelligents, utilisés pour le support client des entreprises par exemple. De plus, de nombreux systèmes de recherche augmentent leurs résultats de recherche avec des réponses instantanées, qui fournissent à l'utilisateur un accès immédiat à des informations pertinentes pour sa requête.

Notre problématique de recherche dans cette thèse est l'amélioration des systèmes QA dans le domaine biomédical. La recherche en médecine et biologie se développe rapidement, ce qui se traduit par une augmentation considérable des articles de recherche biomédicale. En moyenne, plus de trois mille articles sont rajoutés à PubMed[1] (la plus grande base de données d'articles de recherche biomédicale) chaque jour. Trouver des informations pertinentes dans cette littérature en pleine expansion devient de plus en plus difficile pour les chercheurs et les professionnels de la santé, ce qui accroît également le fossé entre la recherche et la pratique professionnelle. Ce constat peut être observé dans l'épidémie récente de COVID-19, où les chercheurs biomédicaux sont engagés dans une course contre la montre pour trouver des informations pertinentes sur les traitements possibles ou vaccins efficaces.

Les systèmes IR classiques tels que PubMed, bien que très utiles, renvoient toujours beaucoup plus de résultats de recherche que ce qui est idéalement souhaitable [3]. Il faut donc plus de temps pour évaluer la pertinence des documents renvoyés, puis extraire l'information requise et la synthétiser sous une forme qui peut facilement informer la prise de décision en matière de soins de santé. D'autre part, les systèmes QA ont le potentiel de surmonter les lacunes des systèmes IR classiques et de transformer positivement l'expérience de recherche. En effet, plutôt que de renvoyer des documents entiers, les systèmes QA peuvent extraire, et même synthétiser des réponses précises à des questions formulées naturellement. Malheureusement,

[1] https://pubmed.ncbi.nlm.nih.gov

lorsqu'ils sont appliqués directement à la littérature biomédicale, ces modèles QA donnent souvent des résultats insatisfaisants.

À la lumière de cette vue d'ensemble de la problématique d'amélioration des performances des systèmes QA dans le domaine biomédical, nous nous intéressons dans cette thèse aux axes suivants en cherchant à répondre à un ensemble de questions scientifiques de recherche pour chaque axe.

Axe 1 – Nouvelles approches pour la BQA :

Les modèles QA actuels peuvent désormais trouver des réponses précises à partir d'un texte, d'un passage ou d'un document entier. Malheureusement, lorsqu'ils sont appliqués directement à la littérature biomédicale, ces modèles donnent souvent des résultats insatisfaisants. Cela est dû à un changement de la distribution des mots des corpus généraux vers les corpus biomédicaux et également aux caractéristiques spécifiques de la littérature biomédicale, telles que des documents volumineux, une terminologie complexe propre au domaine et une typologie de questions spécifiques au domaine. Par conséquent, plusieurs modèles et techniques QA biomédicale ont été proposés, et ces modèles sont souvent plus performants que leurs homologues du domaine général. Néanmoins, des modèles plus puissants restent nécessaires pour relever les défis spécifiques du QA biomédicale.

Axe 2 – La construction de nouveaux ensembles de données (datasets) BQA :

Une autre raison de la lenteur du progrès du QA biomédicale par rapport au QA du domaine général est le nombre limité de datasets QA biomédicale, comparé avec le nombre de datasets QA du domaine général. Une autre limitation réside dans le nombre d'instances d'entraînement. Dans le domaine du QA biomédicale, le plus grand dataset annoté, BioASQ [16] contient 3 243 instances, alors que le plus grand dataset QA dans le domaine général, SQuAD v2.0 [5] contient 150 mille instances. En plus, presque la totalité des datasets BQA sont en anglais, ce qui limite la recherche en BQA dans les autres langues.

Axe 3 – Couplage d'un système IR avec un modèle BQA :

Un système IR classique se contente seulement de retourner les documents pertinents pour une question donnée. Souvent exprimée en mots-clés. Alors qu'un modèle QA/BQA retourne directement la réponse à la question. Souvent exprimée en langue naturelle. D'où vient l'intérêt de coupler les moteurs IR classiques avec les modèles QA/BQA. Afin de tirer avantage des performances et particularités offertes par les deux approches IR et QA.

Afin de répondre aux axes de recherche ci-dessus, nos contributions sont :

Axe 1 – Nouvelles approches pour la BQA :

- L'introduction d'une nouvelle méthode BQA pour les questions de type factoïde et liste. Cette méthode a donné des résultats de pointe (State-Of-The-Art (SOTA)) sur plusieurs lots de tests des datasets 10b, 9b, 8b et 7b du challenge BioASQ [4]
- L'exploitation de l'apprentissage par transfert pour les questions de types Oui/Non et résumé.

Axe 2 – La construction de nouveaux ensembles de données (datasets) BQA :

- La construction du premier dataset BQA français
- Le lancement du premier tableau de classement public des modèles BQA français

Axe 3 – Couplage d'un système IR avec un modèle BQA :

- La proposition d'une démarche de couplage d'un moteur IR avec un modèle BQA
- La création d'un moteur de recherche biomédical spécial au COVID-19, basé sur la démarche proposée. Ce moteur de recherche s'inscrit dans le cadre de notre réponse à un appel à projets en relation avec COVID-19 lancé par le Centre National pour la Recherche Scientifique et Technique (CNRST)

Les trois axes de recherche cités ci-dessus représentent la ligne directive du travail présenté dans cette thèse. Le reste des sections sont structurées comme suit :

Chapitre I – État de l'art : Traitement automatique de langage naturel (NLP), et Question Answering (QA) :

D'abord, nous allons dresser l'état de l'art du domaine de traitement automatique de langage naturel. Son historique et ses tâches principales. Puis nous allons aborder la tâche QA avec plus de détails. Nous allons commencer par définir la tâche du QA, puis nous allons énumérer et comparer ses datasets et challenges. Ensuite, nous allons décrire les modèles et systèmes QA de références.

Chapitre II – État de l'art : Biomedical Question Answering (BQA) :

Ce chapitre fait le point sur l'état de l'art de la tâche de QA biomédical. À l'image du premier chapitre, nous allons commencer par définir la tâche de BQA, puis nous allons énumérer et comparer ses datasets et challenges. Ensuite, nous allons décrire les modèles et systèmes BQA de références.

Chapitre III – Nouvelle approche BQA pour les questions de type factoïde et liste :

Nous présentons tout d'abord les principaux modèles et méthodes BQA introduits précédemment qui sont étroitement liés à notre approche. Dans la deuxième section, nous décrivons notre nouvelle approche BQA pour les questions de type factoïde et liste. Dans la troisième section, nous présentons et discutons les résultats des expériences que nous avons réalisées sur le dataset BioASQ.

Chapitre IV – Application de l'apprentissage par transfert pour les questions BQA de types Oui/Non et résumé :

Nous allons citer les méthodes et systèmes BQA état-de-l'art pour les questions de type Oui/Non et résumé. Après ceci, nous allons détailler notre approche pour les deux types de questions, en termes de datasets utilisés pour le transfert d'apprentissage, et les modèles et architectures adoptées. Nous allons terminer la présentation de notre approche avec les résultats obtenus toujours sur le dataset BioASQ.

Chapitre V – FrBMedQA : le premier dataset BQA en langue française :

Au début, nous donnons un aperçu des travaux connexes qui ont été réalisés sur les datasets BQA et sur le QA en français. Dans la troisième section, nous décrivons en détail notre dataset FrBMedQA. Dans la section quatre, nous décrivons les modèles de base ainsi que les modèles neuronaux que nous avons appliqués au dataset, en donnant également les résultats de nos expériences et en les discutant.

Chapitre VI – Couplage d'un moteur de recherche d'information IR avec un modèle BQA :

Dans ce chapitre, nous présentons la méthode que nous proposons pour coupler un modèle BQA avec un système IR afin de former un moteur de recherche biomédical, basé sur des modèles IR et BQA de pointe.

Avec l'arrivé de la pandémie de COVID-19, nous avons décidé d'appliquer notre approche de couplage moteur IR classique avec modèle QA pour construire un moteur de recherche intelligent spécifique aux questions en relation avec la maladie de COVID-19. La deuxième partie de ce chapitre est consacrée à la présentation de ce moteur de recherche.

Chapitre I :
État de l'art : Traitement automatique de langage naturel (NLP) et Question Answering (QA)

1.1 Introduction

L'un des sous-domaines les plus importants de l'intelligence artificielle est le traitement automatique de langage naturel, en anglais « Natural Language Processing » (NLP). C'est aussi un domaine interdisciplinaire qui combine la linguistique, les sciences cognitives et l'informatique, la principale préoccupation de l'NLP est de permettre aux ordinateurs de comprendre et de produire du langage aussi proche que le font les humains. Au cours des trois dernières années, les modèles basés sur l'apprentissage automatique tels que BERT [1] et GPT-3 [2], ont fait des progrès impressionnants dans l'état de l'art de nombreuses tâches NLP.

La réponse automatique aux questions, en anglais « Question Answering » (QA) est une discipline informatique à l'intersection des domaines de la recherche d'information, en anglais « Information Retrieval » (IR) et de l'NLP, qui s'intéresse à la construction de systèmes qui répondent automatiquement aux questions posées par les humains avec un langage naturel. Les programmes QA peuvent construire des réponses à partir de l'interrogation d'une base de connaissances structurée ou d'une collection non structurée de documents en langage naturel. Les systèmes QA sont soit à domaine fermé (répondant aux questions d'un domaine spécifique), soit à domaine ouvert (s'appuyant sur des ontologies générales et des connaissances répandues). Ces systèmes sont capables d'extraire la réponse à une question à partir d'un texte donné. Ceci est utile pour rechercher une réponse dans un document. Selon le système utilisé, la réponse peut être directement extraite du texte ou générée automatiquement. Les systèmes QA sont souvent utilisés pour automatiser la réponse aux questions fréquemment posées en utilisant une base de connaissances (par exemple, des documents) comme contexte. En tant que tels, ils sont utiles pour les assistants virtuels intelligents, utilisés pour le support client des entreprises par exemple. De plus, de nombreux systèmes de recherche augmentent leurs résultats de recherche avec des réponses instantanées, qui fournissent à l'utilisateur un accès immédiat à des informations pertinentes pour sa requête. Les cinq dernières années ont vu une explosion de la recherche dans ce domaine en termes de nouveaux ensembles de données (datasets) [5, 6, 7, 8, 9, 10], et de nouveaux modèles et méthodes [11, 12, 13, 1, 14] basés sur l'apprentissage profond [15], qui ont surpassé tous leurs prédécesseurs.

Le reste du chapitre est organisé comme suit. D'abord, nous allons commencer par introduire le domaine de l'NLP. Nous allons définir ce domaine et donner son historique depuis sa naissance dans les années 1950 jusqu'au présent. Ensuite, nous allons décrire les tâches NLP les plus courantes, en proposant aussi une classification de ces tâches. La deuxième partie du chapitre sera consacré à la tâche de QA. Nous allons commencer par définir cette tâche, puis nous allons énumérer et comparer ses datasets et challenges « benchmarks ». Ensuite, nous allons décrire les modèles et systèmes de références de QA. Enfin, nous allons terminer le chapitre par une conclusion.

1.2 Traitement automatique de langage naturel (NLP)

1.2.1 Définition et historique

Le traitement automatique de langage naturel (NLP) est un sous-domaine interdisciplinaire de la linguistique, de l'informatique et de l'intelligence artificielle qui s'intéresse aux interactions entre les ordinateurs et le langage humain, en particulier à la manière de programmer les ordinateurs pour traiter et analyser de grandes quantités de données en langage naturel. L'objectif est de créer un ordinateur capable de "comprendre" le contenu des documents, y compris les nuances contextuelles de la langue qu'ils contiennent. La technologie peut alors extraire avec précision les informations et les idées contenues dans les documents, ainsi que classer et organiser les documents eux-mêmes. Les défis posés par le traitement automatique de langage naturel concernent souvent la reconnaissance de la parole, la compréhension du langage naturel et la génération du langage naturel.

Le traitement automatique de langage naturel trouve ses racines dans les années 1950. L'approche symbolique était l'approche dominante de cette époque jusqu'au début des années 1990. Le principe de l'approche symbolique est le suivant : étant donné un ensemble de règles (par exemple, un recueil de phrases, avec des questions et des réponses correspondantes), l'ordinateur émule la compréhension du langage naturel (ou d'autres tâches NLP) en appliquant ces règles aux données auxquelles il est confronté. Cette approche n'a connu qu'un succès très limité, car elle ne reposait que sur des règles codées en dur qui ne permettaient pas de saisir la variété du langage humain. À partir de la fin des années 1980, une révolution s'est produite dans le traitement de langage naturel avec l'introduction d'algorithmes d'apprentissage automatique

pour le traitement du langage. Cette approche est désignée sous le nom de l'approche statistique. Les premiers succès notables des méthodes statistiques dans le traitement de langage naturel ont été enregistrés dans le domaine de la traduction automatique. Avec le développement du web, des quantités croissantes de données linguistiques brutes (non annotées) sont devenues disponibles depuis le milieu des années 1990. La recherche s'est donc de plus en plus concentrée sur les algorithmes d'apprentissage non supervisés et semi-supervisés. Ces algorithmes peuvent apprendre à partir de données qui n'ont pas été annotées manuellement avec les réponses souhaitées ou en utilisant une combinaison de données annotées et non annotées. En général, cette tâche est beaucoup plus difficile que l'apprentissage supervisé et produit généralement des résultats moins précis pour une quantité de données d'entrée. Cependant, il existe une énorme quantité de données non annotées (y compris, entre autres, le contenu entier du World Wide Web), qui peut souvent compenser les résultats inférieurs si l'algorithme utilisé a une complexité temporelle suffisamment faible pour être pratique. L'un des principaux inconvénients des méthodes statistiques est qu'elles nécessitent une ingénierie élaborée des caractéristiques. Depuis le début des années 2010, le domaine a donc largement abandonné les méthodes statistiques et s'est tourné vers les réseaux neuronaux pour l'apprentissage automatique. Les techniques populaires comprennent l'utilisation de vecteurs de mots « Word Embeddings » pour capturer les propriétés sémantiques des mots, et une augmentation de l'apprentissage de bout en bout d'une tâche de plus haut niveau (Question Answering par exemple) au lieu de s'appuyer sur un pipeline de tâches intermédiaires distinctes (par exemple, le marquage des parties du discours et l'analyse syntaxique des dépendances). Dans certains domaines, cette évolution a entraîné des changements substantiels dans la façon dont les systèmes de traitement automatique des langues sont conçus, de sorte que les approches basées sur les réseaux neuronaux profonds peuvent être considérées comme un nouveau paradigme distinct du traitement statistique du langage naturel. Par exemple, le terme "traduction automatique neuronale" (NMT) souligne le fait que les approches de traduction automatique basées sur l'apprentissage profond apprennent directement les transformations de séquence à séquence, évitant ainsi le recours à des étapes intermédiaires telles que l'alignement des mots et la modélisation du langage, qui étaient utilisées dans la traduction automatique statistique (SMT).

1.2.2 Tâches courantes de l'NLP

Voici une liste des tâches les plus couramment étudiées dans le domaine du traitement automatique de langage naturel. Certaines de ces tâches ont des applications directes dans le monde réel, tandis que d'autres servent plus souvent de sous-tâches utilisées pour aider à résoudre des tâches plus importantes. Bien que les tâches de l'NLP soient étroitement liées entre elles, elles peuvent être subdivisées en catégories pour plus de commodité. Une division grossière est présentée ci-dessous. La Figure 1 montre une proposition de classification de tâche NLP par catégories.

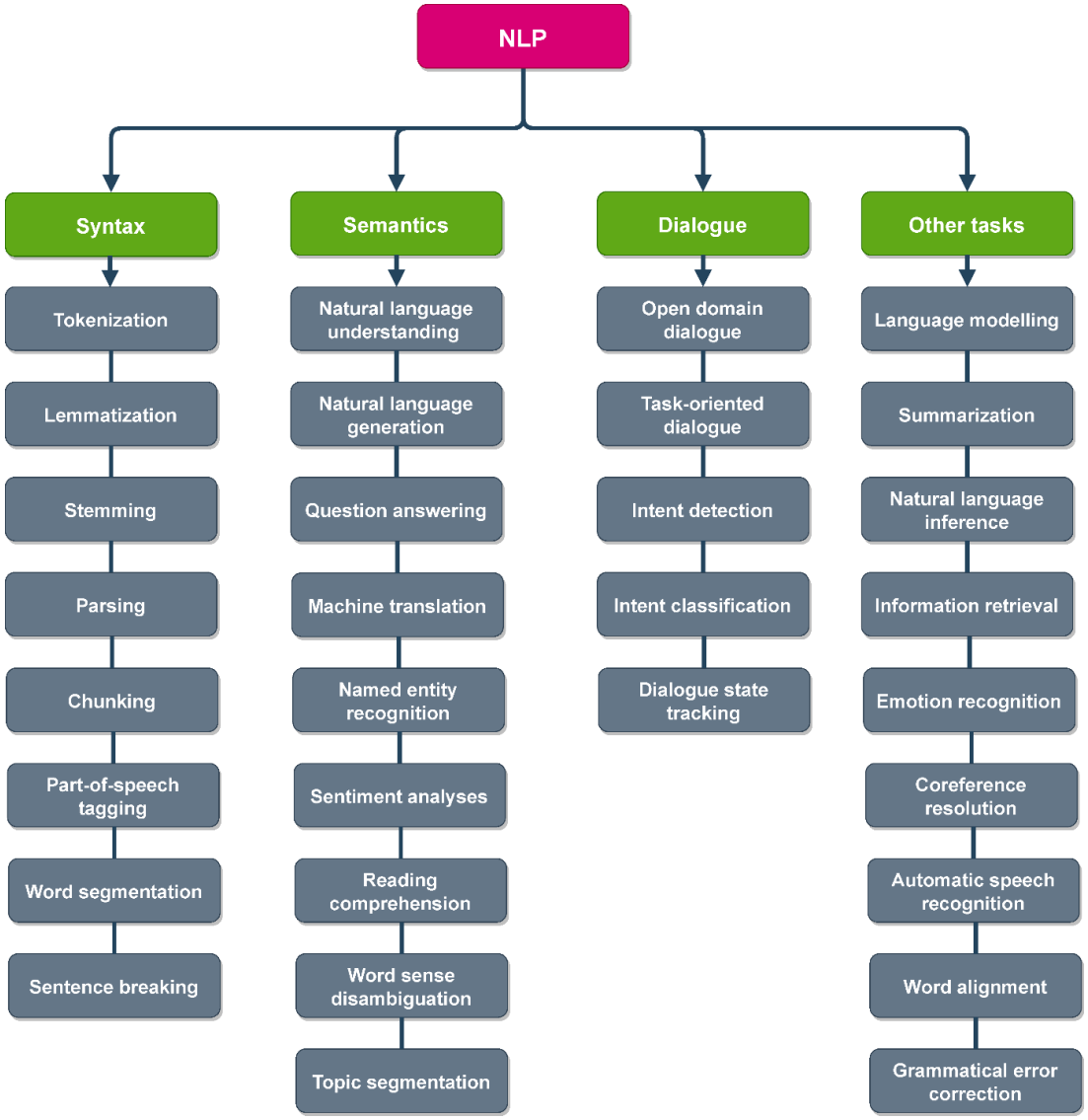

Figure 1 Classification des tâches NLP

1.2.2.1 Tâches syntaxiques

Segmentation des mots « Word segmentation (Tokenization) » : séparer un morceau de texte continu en mots distincts.

Induction de la grammaire « Grammar induction » : générer une grammaire formelle qui décrit la syntaxe d'un langage.

Découpage de phrases « Sentence breaking » : pour un morceau de texte, trouver les limites de la phrase. Les limites des phrases sont souvent marquées par des points ou d'autres signes de ponctuation.

Analyse syntaxique « Parsing » : déterminer l'arbre d'analyse (analyse grammaticale) d'une phrase donnée.

Lemmatisation « Lemmatization » : la tâche consistant à supprimer les terminaisons flexionnelles uniquement et à retourner la forme de base du dictionnaire d'un mot, également connue sous le nom de lemme.

Segmentation morphologique « Morphological segmentation » : séparation des mots en morphèmes individuels et identification de la classe des morphèmes.

Étiquetage de la partie du discours « Part-of-speech tagging » : étant donné une phrase, déterminer la partie du discours « Part-Of-Speech » (POS) pour chaque mot.

Dérivation « Stemming » : processus consistant à réduire les mots dérivés à une forme de base

« Chunking » : il s'agit du processus de division d'un grand morceau de texte en morceaux plus petits (chunks) plus faciles à gérer. Les "chunks" peuvent être définis de différentes manières, en fonction de la tâche NLP spécifique et des informations à extraire du texte.

1.2.2.2 Tâches sémantiques

Reconnaissance d'entités nommées « Named entity recognition » (NER) : étant donné un flux de texte, déterminer quels éléments du texte correspondent à des noms propres, tels que des personnes ou des lieux, et quel est le type de chacun de ces noms (par exemple, personne, lieu, organisation).

Analyse des sentiments « Sentiment analysis » : extraire des informations subjectives généralement à partir d'un ensemble de documents, en utilisant souvent des avis en ligne pour déterminer la "polarité" à propos d'objets spécifiques. Elle est particulièrement utile pour identifier les tendances de l'opinion publique dans les médias sociaux, pour le marketing.

Désambiguïsation du sens des mots « Word-sense disambiguation » (WSD) : de nombreux mots ont plus d'une signification ; nous devons sélectionner la signification qui a le plus de sens dans le contexte. Pour ce problème, on nous donne généralement une liste de mots et de sens associés.

Extraction de relations « Relationship extraction » : étant donné un morceau de texte, identifier les relations entre les entités nommées (par exemple, qui est marié à qui).

Analyse sémantique « Semantic parsing » : étant donné un morceau de texte (typiquement une phrase), produire une représentation formelle de sa sémantique, soit sous forme de graphe, soit en accord avec un formalisme logique.

Étiquetage des rôles sémantiques « Semantic role labelling » : pour une phrase donnée, identifier et désambiguïser les prédicats sémantiques (par exemple, les cadres verbaux), puis identifier et classer les éléments du cadre (rôles sémantiques).

Segmentation et reconnaissance de sujets « Topic segmentation and recognition » : étant donné un morceau de texte, le séparer en segments dont chacun est consacré à un sujet, et identifier le sujet du segment.

Traduction automatique « Machine translation » (MT) : traduire automatiquement un texte d'une langue humaine à une autre. Il s'agit de l'un des problèmes les plus difficiles nécessitant tous les différents types de connaissances que les humains possèdent (grammaire, sémantique, faits sur le monde réel, etc.) pour être résolus correctement.

Génération de langage naturel « Natural language generation » (NLG) : convertir des informations provenant de bases de données informatiques ou d'intentions sémantiques en langage humain lisible.

Compréhension du langage naturel « Natural language understanding » (NLU) : convertir des morceaux de texte en représentations plus formelles, telles que des structures logiques du premier ordre, plus faciles à manipuler par les programmes informatiques. La compréhension du langage naturel implique l'identification de la sémantique voulue parmi les multiples sémantiques possibles qui peuvent être dérivées d'une expression en langage naturel qui prend généralement la forme de notations organisées de concepts en langage naturel.

Réponses automatiques aux questions « Question Answering » (QA) : étant donné une question en langage humain, déterminer sa réponse. Les questions typiques ont une bonne réponse spécifique (telle que "Quelle est la capitale du Canada ?"), mais parfois des questions ouvertes sont également considérées (telles que "Quel est le sens de la vie ?").

Compréhension automatique de la lecture « Machine Reading Comprehension » (MRC) : il s'agit de la capacité d'un système à comprendre et à interpréter un texte écrit. La compréhension de la lecture est une tâche complexe qui exige d'un système qu'il soit capable de comprendre le sens des mots et des phrases, d'identifier les idées principales et les détails complémentaires d'un texte, et de comprendre les relations entre les différentes parties du texte.

1.2.2.3 Tâches de dialogue

Dialogue de domaine ouvert « Open domain dialogue » : tâche dans laquelle un système est capable d'engager une conversation avec un utilisateur humain sur n'importe quel sujet. Dans un système de dialogue à domaine ouvert, le système n'est pas limité à un domaine spécifique ou à un ensemble de sujets, mais plutôt capable de comprendre et de répondre à tout sujet qu'un utilisateur pourrait aborder.

Dialogue axé sur la tâche « Task-oriented dialogue » : le dialogue orienté tâche fait référence à un type de tâche dans lequel un système est conçu pour aider les utilisateurs à accomplir des tâches spécifiques ou à atteindre des objectifs spécifiques. Contrairement au dialogue dans un domaine ouvert, qui est conçu pour une conversation plus générale et qui n'est pas limité à un ensemble spécifique de sujets.

Détection d'intention « Intent detection » : la détection d'intention est une tâche qui consiste à identifier le but ou l'objectif d'une entrée de texte ou de parole d'un utilisateur. Dans le contexte

des systèmes de dialogue orientés tâche ou des « chatbots », la détection d'intention est le processus qui consiste à déterminer ce que l'utilisateur essaie d'accomplir ou l'action qu'il veut que le système entreprenne.

Classification d'intention « Intent classification » : la classification des intentions est le processus qui consiste à déterminer l'intention ou l'objectif d'une entrée textuelle ou vocale d'un utilisateur. La classification des intentions est souvent utilisée pour classer les entrées de l'utilisateur dans des catégories prédéfinies ou "intentions", telles que les demandes d'informations, les demandes d'exécution d'une action ou les déclarations de faits.

Suivi de l'état du dialogue « Dialogue state tracking » : le suivi de l'état du dialogue est le processus qui consiste à garder une trace de l'état actuel d'un dialogue entre un utilisateur et un système, tel qu'un « chatbot » ou un assistant virtuel. Dans les systèmes de dialogue axés sur les tâches, le suivi de l'état du dialogue implique de conserver un enregistrement des objectifs et des préférences de l'utilisateur, ainsi que des actions entreprises par le système en réponse aux entrées de l'utilisateur.

1.2.2.4 Autres tâches

Résumé automatique « Automatic summarization » : produire un résumé lisible d'un morceau de texte. Souvent utilisé pour fournir des résumés du texte d'un type connu, comme les documents de recherche, les articles de la section financière d'un journal.

Correction des erreurs grammaticales « Grammatical error correction » : la détection et la correction des erreurs grammaticales impliquent une grande variété de problèmes à tous les niveaux de l'analyse linguistique (phonologie/orthographie, morphologie, syntaxe, sémantique).

Résolution de coréférences « Coreference resolution » : pour une phrase ou un morceau de texte plus important, déterminer quels mots ("mentions") font référence aux mêmes objets ("entités").

Modélisation du langage « Language modeling » : la modélisation du langage est une tâche qui consiste à prédire la probabilité d'une séquence de mots dans un langage donné. Les modèles

de langage sont utilisés pour comprendre et générer des textes en langage naturel en apprenant les modèles et la structure du langage.

L'inférence en langage naturel « Natural Language Inference » (NLI) : l'inférence en langage naturel est une tâche qui consiste à déterminer la relation entre deux morceaux de texte. L'inférence en langage naturel exige qu'un système soit capable de comprendre le sens du texte et d'utiliser cette compréhension pour déterminer si un élément du texte (l'hypothèse) est vrai, faux ou inconnu compte tenu d'un autre élément du texte (la prémisse).

Recherche d'information « Information retrieval » (IR) : la recherche d'information est le processus de recherche et de récupération d'informations à partir d'une collection de documents ou d'autres sources. Dans le contexte de l'NLP, la recherche d'information se réfère au processus de recherche et de récupération de documents ou d'autres sources textuelles qui sont pertinents pour une requête ou un sujet particulier.

Reconnaissance des émotions « Emotion recognition » : la reconnaissance des émotions est le processus de détection et d'interprétation des émotions humaines. Cela peut se faire par divers moyens, tels que les expressions faciales, le langage corporel ou les indices verbaux.

Reconnaissance automatique de la parole « Automatic speech recognition » (ASR) : la reconnaissance automatique de la parole est une technologie qui permet aux ordinateurs de reconnaître et de transcrire le langage parlé. Elle est utilisée dans un large éventail d'applications, notamment la transcription de la voix en texte, la numérotation vocale et le contrôle vocal des appareils.

Alignement des mots « Word alignment » : l'alignement des mots est le processus qui consiste à déterminer la correspondance entre les mots d'une langue source et leurs traductions dans une langue cible. Il est souvent utilisé en traduction automatique, où il sert à identifier les mots d'une phrase en langue source qui ont le plus de chances d'être traduits par un mot ou une expression spécifique dans la langue cible.

1.3 Question Answering (QA)

1.3.1 Définition

Le QA est un sous-domaine de la IR et de l'NLP qui concerne la construction de systèmes qui répondent automatiquement aux questions posées par les humains dans un langage naturel. La tâche QA peut être formulée comme un problème d'apprentissage supervisé. Étant donné une collection d'exemples d'apprentissage textuels $\{(p_i, q_{i,}, a_i)\}_{i=1}^{n}$, où p est un passage de texte, et q est une question concernant le texte p. La tâche consiste alors à apprendre un prédicteur f qui prend un passage de texte p et une question q correspondante comme entrées et donne la réponse a comme sortie, ce qui pourrait être formulé avec la formule suivante :

$$a = f(p, q) \tag{1.1}$$

Il est nécessaire qu'une majorité de locuteurs natifs soient d'accord pour dire que la question q concerne le texte p, et la réponse a est une réponse correcte qui ne contient pas d'informations non-pertinentes pour cette question. Le tableau 1 ci-dessous, montre un exemple d'une instance d'apprentissage QA tiré du dataset SQuAD v1.0 [5]

Question	In what year was Nikola Tesla born?
Passage de contexte	Nikola Tesla (Serbian Cyrillic: Никола Тесла; 10 July **1856** – 7 January 1943) was a Serbian American inventor, electrical engineer, mechanical engineer, physicist, and futurist best known for his contributions to the design of the modern alternating current (AC) electricity supply system
Réponse	1856

Table 1 Exemple d'une instance QA tiré du dataset SQuAD v1.0

La tâche QA peut être catégorisée selon le type de corpus utilisé, le type de questions, et le type et la source des réponses. Chaque ongle de classification donne lieu à plusieurs sous-catégories. Le type de corpus peut être soit textuel ou multimodal, c'est-à-dire incluant en plus du texte, des ressources visuelles ou sonores. Le type de question peut être soit naturel, respectant la grammaire du langage naturel, cloze (de type remplir le vide), ou sous forme de mots-clés. Tandis que la réponse peut être une phrase naturelle ou un mot à choisir à partir d'une liste

prédéfinie de mots. En ce qui concerne la source des réponses, ces derniers peuvent être extraites directement d'un passage ou bien générées automatiquement, mais toujours en se basant sur un passage de contexte. La Figure 2 montre cette classification de la tâche QA avec une catégorisation selon chaque ongle de classification.

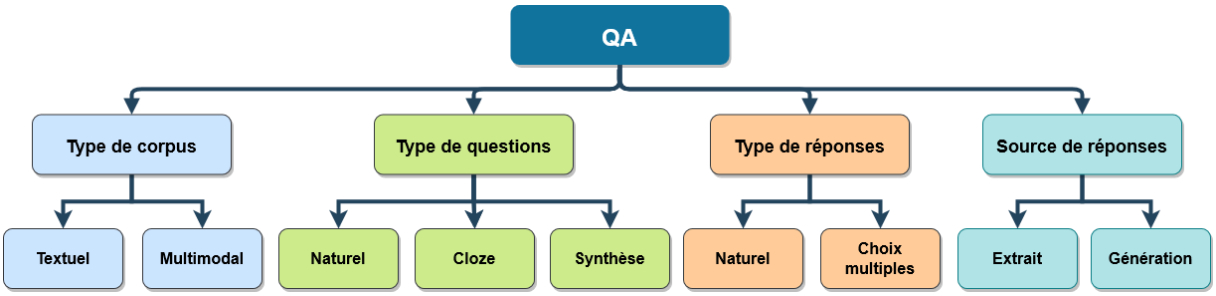

Figure 2 Classification de la tâche QA

1.3.2 DataSets et challenges internationaux

Ces dernières années, un grand nombre de datasets QA à grande échelle ont été créés. Ces datasets ont grandement poussé la recherche en QA. Une grande partie de ces datasets disposent de tableaux de classement public, ce qui leur positionne comme compétitions et challenges internationaux. Comme pour tout type de dataset, les datasets QA peuvent être catégorisées selon plusieurs critères. Comme la taille (en nombre d'instances), la méthode de création, le corpus source, et le type de passage de contexte (paragraphe, document, …etc). Mais le critère le plus important de classification des datasets QA est toujours le thème du dataset. La Figure 3 montre la classification des datasets QA par thème. Ci-dessous, nous allons détailler ces thèmes, tout en citant les datasets les plus connus dans chacun d'eux.

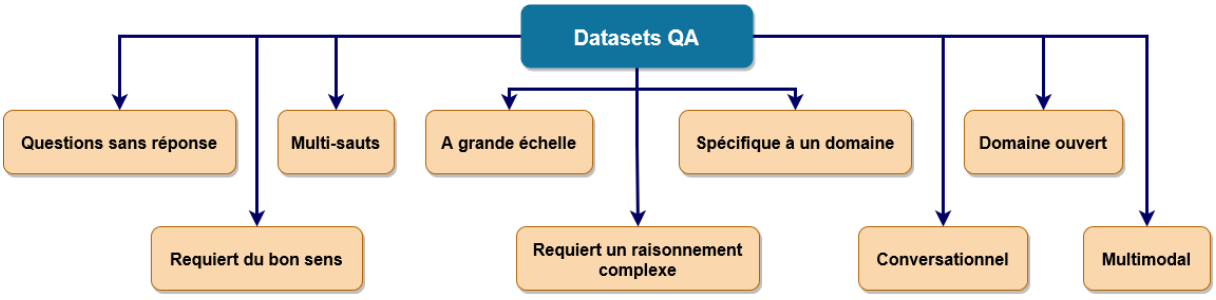

Figure 3 Classification des datasets QA par thème

1.3.2.1 Datasets avec des questions sans réponse

Les datasets QA existants manquent souvent des questions sans réponse, ce qui affaiblit la robustesse des systèmes QA. Par conséquent, lorsque les modèles QA répondent à des questions sans réponse, ils essaient toujours de donner la réponse la plus probable, plutôt que de refuser de répondre à ces questions sans réponse. Ainsi, quelle que soit la réponse retournée par le modèle, les réponses sont forcément fausses. Pour résoudre ce problème, les chercheurs ont proposé de nombreux datasets contenant des questions sans réponse. Dons les plus connus sont : SQuAD 2.0 [16], MS MARCO [17], Natural Questions [18] et WikiQA [19].

1.3.2.2 Datasets multi-sauts

Dans la plupart des datasets QA, la réponse à une question peut généralement être trouvée dans un seul paragraphe ou un seul document. Cependant, dans la compréhension de lecture humaine réelle, lors de la lecture d'un roman par exemple, nous sommes très susceptibles d'extraire des réponses de plusieurs paragraphes. Par rapport au QA d'un seul passage, le QA multi-sauts est plus difficile et nécessite une recherche et un raisonnement multi-sauts sur des passages ou des documents confus. Les datasets QA de type multi-sauts sont HotpotQA [20], NarrativeQA [21], Qangaroo [22], et MultiRC [23].

1.3.2.3 Datasets nécessitant du bon sens

Le langage humain est complexe. Pour répondre à des questions, nous devons souvent faire appel à notre bon sens ou à notre connaissance du monde. De plus, dans le processus du langage humain, de nombreux jeux de mots conventionnels et des mots polysémiques ont été formés. L'utilisation des mêmes mots dans différents scénarios exige également que l'ordinateur ait une bonne maîtrise du sens commun ou de la connaissance du monde. Les tâches classiques du QA consistent généralement à répondre à des questions sur des passages donnés. Dans les datasets QA existants, seule une petite proportion des questions doit être répondue avec des connaissances de sens commun. Afin de construire des modèles QA avec des connaissances du bon sens ou du monde, des datasets axés sur le bon sens ont été créés, comme CommonSenseQA [24], ReCoRD [25], et OpenBookQA [26]

1.3.2.4 Datasets nécessitant un raisonnement complexe

Le raisonnement est une capacité innée des êtres humains, qui peut s'incarner dans la pensée logique, la compréhension de la lecture et d'autres activités. Le raisonnement est également un élément clé de l'intelligence artificielle et un objectif fondamental des systèmes QA. Ces dernières années, le raisonnement a été un sujet essentiel au sein de la communauté QA. Les chercheurs visent à ce que les systèmes QA pourront non seulement lire et apprendre la représentation de la langue, mais aussi comprendre réellement le contexte et répondre à des questions complexes. Afin de progresser vers des systèmes QA à raisonnement complexe, de nombreux datasets ont été créés, tels que DROP [27], RACE [28], et CLOTH [29].

1.3.2.5 Datasets à grande échelle

La taille des premiers datasets QA n'est généralement pas très importante, comme QA4MRE [30], CuratedTREC [31] et MCTest [32]. Avec l'émergence des datasets à grande échelle, la tâche de QA a pu profiter des performances offertes par les réseaux de neurones, notamment le deep learing (DL). Car ce dernier nécessite un très grand nombre d'instances d'entraînement pour pouvoir donner de bons résultats.

1.3.2.6 Datasets spécifiques à un domaine

Un dataset spécifique à un domaine fait référence à un dataset dont le contexte provient d'un domaine particulier, tel que des examens scientifiques, des films, des rapports cliniques...etc. Par conséquent, les modèles de réseaux de neurones entraînés sur ces datasets peuvent être appliqués directement à un certain domaine. Par exemple, CliCR [33] est un dataset QA dans le domaine médical, SciQ [34] est un autre dataset pour les questions d'examens scientifiques sur la physique, la chimie et la biologie. D'autres datasets spécifiques à un domaine sont ReviewQA [35], SciTail [36], et QASPER [37].

1.3.2.7 Datasets multimodal

Lorsque les humains lisent, ils le font souvent de manière multimodale. Par exemple, pour comprendre les informations et répondre aux questions, nous devons parfois lire à la fois les textes et les illustrations, et nous devons également utiliser notre cerveau pour imaginer, reconstruire, raisonner, calculer, analyser ou comparer. Actuellement, la plupart des datasets QA existants ne contiennent que du texte, ce qui présente certaines limites. Certains concepts

complexes ou précis ne peuvent être décrits ou communiqués uniquement par le texte. Par exemple, si nous avons besoin que l'ordinateur réponde à des questions précises concernant la maintenance d'un moteur d'avion, nous devrons peut-être lui fournir l'image du moteur de l'avion.

Le QA multimodale est un domaine interdisciplinaire avec un grand nombre d'applications possibles. Compte tenu de l'hétérogénéité des données, le QA multimodale pose des défis uniques aux chercheurs en NLP, car le modèle doit comprendre à la fois les textes et les images. Ces dernières années, en raison de la disponibilité de données internet à grande échelle, de nombreux datasets QA multimodale ont été créés, tels que TQA [38], RecipeQA [39], COMICS [40], et MovieQA [41]

1.3.2.8 Datasets de domaine ouvert

Le QA de domaine ouvert a été défini à l'origine comme la recherche de réponses dans des collections de documents non structurés. Avec le développement de la recherche en QA, de nombreux datasets QA tendent à être utilisés pour résoudre le QA dans le domaine ouvert. La publication de nouveaux datasets QA tels que MCTest [32], CuratedTREC [31], Quasar [42], et SearchQA [43] favorise grandement la recherche en QA dans un domaine ouvert.

1.3.2.9 Datasets conversationnels

C'est un moyen naturel pour les êtres humains d'échanger des informations par le biais d'une série de conversations. Dans les tâches typiques du QA, les différentes paires de questions et de réponses sont généralement indépendantes les unes des autres. Cependant, dans la communication réelle en langage humain, nous parvenons souvent à comprendre efficacement des informations complexes par le biais d'une série de conversations interdépendantes. De même, dans les scénarios de communication humaine, nous posons souvent des questions de notre propre initiative, afin d'obtenir des informations clés qui nous aident à comprendre la situation. Dans le processus de conversation, nous devons avoir une compréhension approfondie des conversations précédentes afin de répondre correctement aux questions de l'autre ou de poser de nouvelles questions significatives. Par conséquent, dans ce processus, les informations historiques de la conversation deviennent également une partie du contexte.

Ces dernières années, le QA conversationnel est devenu un domaine de recherche prioritaire par la communauté des chercheurs en QA et NLP, et de nombreux datasets connexes ont émergé, tels que CoQA [7], QuAC [8], DREAM [44] et ShARC [45].

1.3.3 Approches de base

Depuis le début de la recherche en QA, plusieurs méthodes, systèmes, et modèles ont été proposés. Depuis les systèmes classiques basés sur les composants comme [46], jusqu'aux modèles de langages comme BERT [1], en passant par d'autres systèmes basés sur diverses techniques et architectures deep learing comme LSTM [47], BiLSTM [48], ou le mécanisme d'attention [49]. Ci-dessous, nous allons présenter les systèmes QA de références qui ont été proposés au fil des années.

1.3.3.1 Match-LSTM

L'un des premiers modèles proposés juste après l'introduction du dataset SQuAD [5] est match-LSTM [50]. L'architecture de ce modèle est basée sur un travail antérieur [51] réalisé par les mêmes auteurs pour l'implication textuelle, en anglais « textual entailment », et sur Pointer Net [52], un modèle de séquence à séquence proposé pour contraindre les jetons de sortie à faire partie des séquences d'entrée. Les auteurs proposent deux façons d'utiliser Pointer Net pour la tâche de QA. Les expériences montrent que les deux modèles sont nettement plus performants que les meilleurs résultats obtenus par Rajpurkar et al [5] en utilisant la régression logistique et des caractéristiques, en anglais « features » créées manuellement.

Une vue d'ensemble des deux modèles proposés est présentée à la Figure 4. Les deux modèles sont composés de trois couches : une couche de prétraitement LSTM [47] qui prétraite le passage de contexte et la question à l'aide des LSTMs. Une couche LSTM de correspondance qui essaie de faire correspondre le passage de contexte à la question. Et une couche Answer Pointer (Ans-Ptr) qui utilise Ptr-Net pour sélectionner un extrait du passage comme réponse. La différence entre les deux modèles réside uniquement dans la troisième couche.

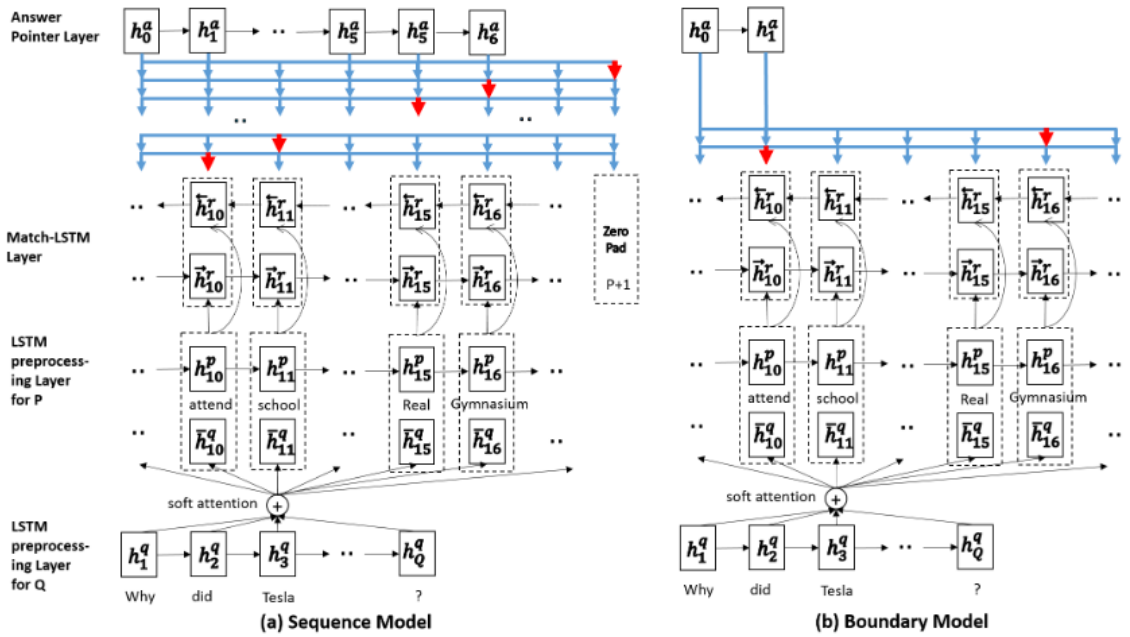

Figure 4 Une vue d'ensemble des deux modèles Match-LSTM proposés : le modèle de séquence et le modèle de frontière

1.3.3.2 Dynamic Coattention Networks

Plusieurs modèles d'apprentissage profond ont été proposés pour la tâche de QA. Cependant, en raison de leur nature à passage unique, ils n'ont aucun moyen de récupérer des maxima locaux correspondant à des réponses incorrectes. Pour résoudre ce problème, Caiming et al. ont proposé le réseau de coattention dynamique (DCN) [53] pour la tâche de QA. Le DCN, illustré dans la Figure 5, fusionne d'abord les représentations co-dépendantes de la question et du document afin de se concentrer sur les parties pertinentes des deux. Ensuite, un décodeur de pointage dynamique itère sur les plages de réponses potentielles. Cette procédure itérative permet au modèle de récupérer des maxima locaux initiaux correspondant à des réponses incorrectes. Sur le dataset SQuAD, un seul modèle DCN améliore l'état de l'art précédent de 71.0% F1 à 75.9%, tandis qu'un modèle d'ensemble DCN obtient 80.4% F1

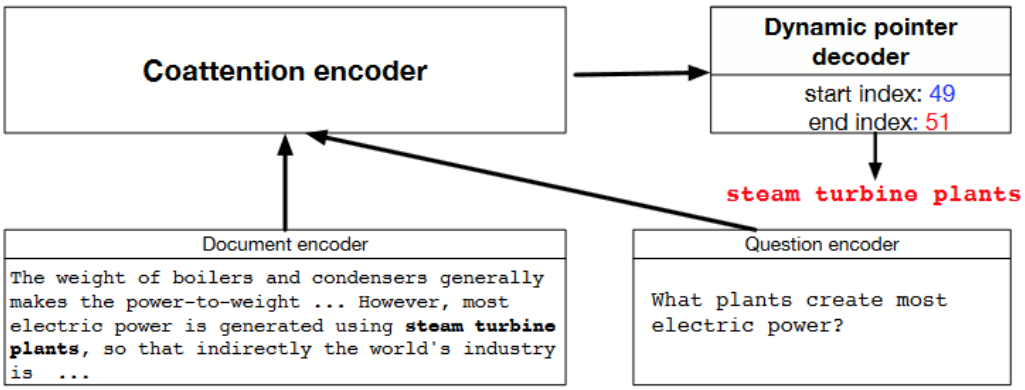

Figure 5 Aperçu du modèle de Coattention Dynamique

1.3.3.3 BiDAF

L'un des modèles QA les plus connus est BIDAF (Bi-Directional Attention Flow) [54]. Il est basé sur une architecture hiérarchique à plusieurs niveaux pour modéliser les représentations du paragraphe contextuel à différents niveaux de granularité (Figure 6). BIDAF inclut des incorporations au niveau des caractères, des mots et du contexte, et utilise un flux d'attention bidirectionnel pour obtenir une représentation du contexte adaptée aux requêtes. Le mécanisme d'attention BIDAF offre les améliorations suivantes par rapport aux paradigmes d'attention proposés précédemment. Premièrement, la couche d'attention n'est pas utilisée pour résumer le paragraphe contextuel en un vecteur de taille fixe. Au lieu de cela, l'attention est calculée pour chaque pas de temps, et le vecteur assisté à chaque pas de temps, ainsi que les représentations des couches précédentes, sont autorisés à passer à la couche de modélisation suivante. Cela permet de réduire la perte d'informations causée par la compression précoce. Deuxièmement, les auteurs ont utilisé un mécanisme d'attention [49] sans mémoire. C'est-à-dire que, bien qu'ils calculent itérativement l'attention dans le temps comme dans Bahdanau et al. [49], l'attention à chaque pas de temps est fonction uniquement de la requête et du paragraphe de contexte au pas de temps actuel et ne dépend pas directement de l'attention au pas de temps précédent. Les auteurs supposent que cette simplification conduit à la division du travail entre la couche d'attention et la couche de modélisation. Elle oblige la couche d'attention à se concentrer sur l'apprentissage de l'attention entre la requête et le contexte, et permet à la couche de modélisation de se concentrer sur l'apprentissage de l'interaction au sein de la représentation du contexte sensible à la requête (la sortie de la couche d'attention). Cela permet également à l'attention à chaque étape temporelle de ne pas être affectée par des assistances incorrectes à

des étapes temporelles précédentes. Les expériences montrent que l'attention sans mémoire donne un avantage clair sur l'attention dynamique. Troisièmement, les auteurs utilisent des mécanismes d'attention dans les deux sens, de requête à contexte et de contexte à requête, qui fournissent des informations complémentaires l'une à l'autre.

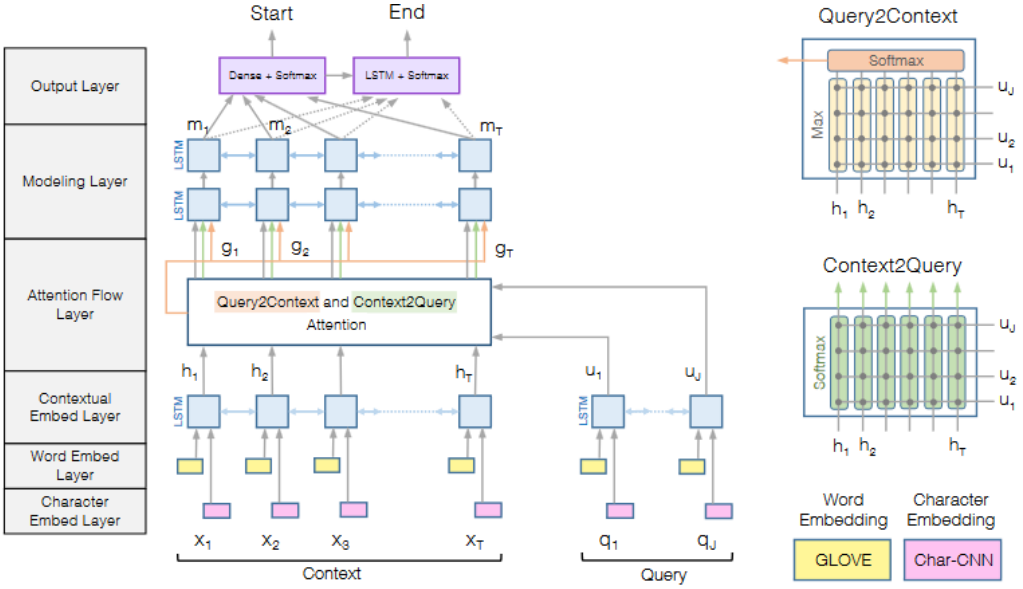

Figure 6 Aperçu du modèle BiDAF (BiDirectional Attention Flow)

Le modèle BiDAF surpasse toutes les approches précédentes sur les datasets hautement compétitifs SQuAD [5] et CNN/DailyMail [55].

1.3.3.4 R-NET

R-NET [56], un réseau d'auto-appariement à portes, illustré à la Figure 7, est un modèle de réseau neuronal de bout en bout proposé pour la tâche de QA. Le modèle se compose de quatre parties : 1) le codeur de réseau récurrent pour construire la représentation des questions et des passages séparément, 2) la couche d'appariement à portes pour faire correspondre la question et le passage, 3) la couche d'appariement automatique pour agréger les informations de l'ensemble du passage, et 4) la couche de prédiction de la réponse basée sur le réseau de pointeurs. Le modèle proposé ajoute une porte supplémentaire aux réseaux récurrents basés sur l'attention [49, 57, 51], pour tenir compte du fait que les mots du passage ont une importance différente pour répondre à une question particulière. Dans [51], les mots d'un passage avec leur contexte de question correspondant pondéré par l'attention sont codés ensemble pour produire

une représentation du passage sensible aux questions. En introduisant un mécanisme à portes, R-NET attribue différents niveaux d'importance aux parties du passage en fonction de leur pertinence par rapport à la question, en masquant les parties non-pertinentes du passage et en mettant l'accent sur les parties importantes.

Ce modèle introduit également un mécanisme d'auto-appariement, qui permet d'agréger efficacement les preuves de l'ensemble du passage pour déduire la réponse. Par le biais d'une couche d'appariement à portes, la représentation du passage consciente de la question qui en résulte encode efficacement les informations de la question pour chaque mot du passage. Cependant, les réseaux récurrents ne peuvent mémoriser qu'un contexte de passage limité en pratique, malgré leurs capacités théoriques. Le candidat à une réponse n'est souvent pas conscient des indices présents dans d'autres parties du passage. Pour résoudre ce problème, les auteurs ont proposé une couche d'auto-appariement pour affiner dynamiquement la représentation du passage avec des informations provenant de l'ensemble du passage. Sur la base de la représentation du passage sensible aux questions, un réseau récurrent basé sur l'attention à portes est utilisé sur le passage contre le passage lui-même, en agrégeant les preuves pertinentes pour le mot actuel du passage à partir de chaque mot du passage. Une couche de réseau récurrent basé sur l'attention et une couche d'auto-appariement enrichissent dynamiquement chaque représentation de passage avec des informations agrégées à partir de la question et du passage, permettant au réseau suivant de mieux prédire les réponses.

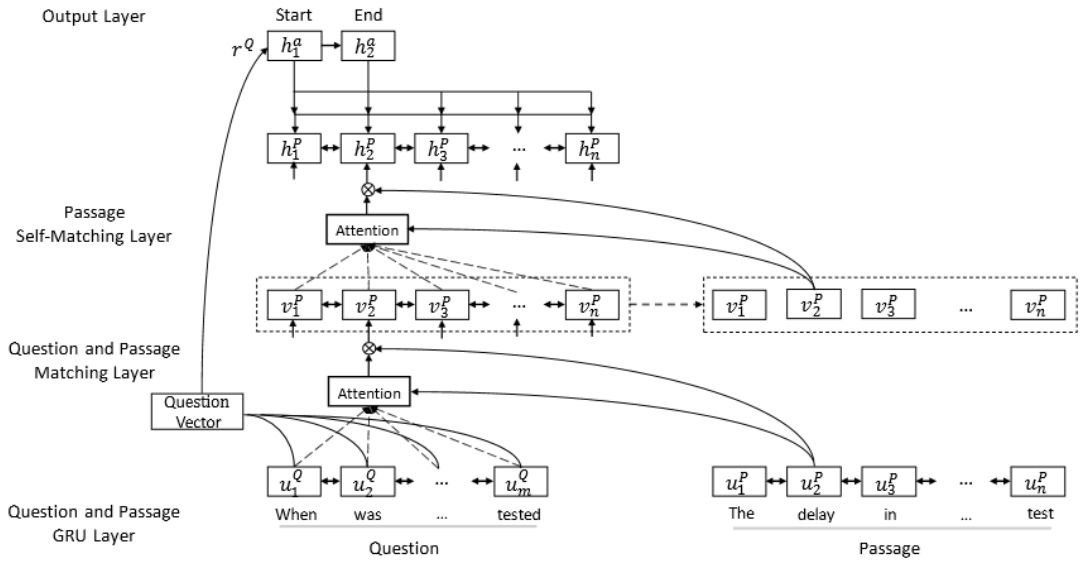

Figure 7 Aperçu de la structure du réseau d'auto-adaptation à porte (R-NET)

R-NET a obtenu des résultats de pointe par rapport à des modèles de bases solides. Un seul modèle permet d'obtenir une précision de 71,3% de correspondance exacte sur SQuAD [5], tandis que le modèle d'ensemble fait passer ce résultat à 75,9%.

1.3.3.5 QANet

Tous les modèles QA décrits précédemment sont principalement basés sur des réseaux de neurones récurrents (RNN) [58] avec le mécanisme d'attention [49]. Malgré leur succès, ces modèles sont souvent lents, tant pour l'apprentissage que pour l'inférence, en raison de la nature séquentielle des RNN. Pour résoudre ce problème, le modèle QANet [59] est introduit. Il utilise exclusivement des convolutions et des auto-attentions comme blocs de construction d'encodeurs qui encodent séparément la question et le contexte. Il apprend ensuite les interactions entre le contexte et la question par des attentions standards [53, 13, 49]. La représentation résultante est encodée à nouveau avec un encodeur sans récurrence avant d'être finalement décodée en fonction de la probabilité que chaque position soit le début ou la fin de l'intervalle de réponse. L'architecture globale de QANet est présentée dans la Figure 8.

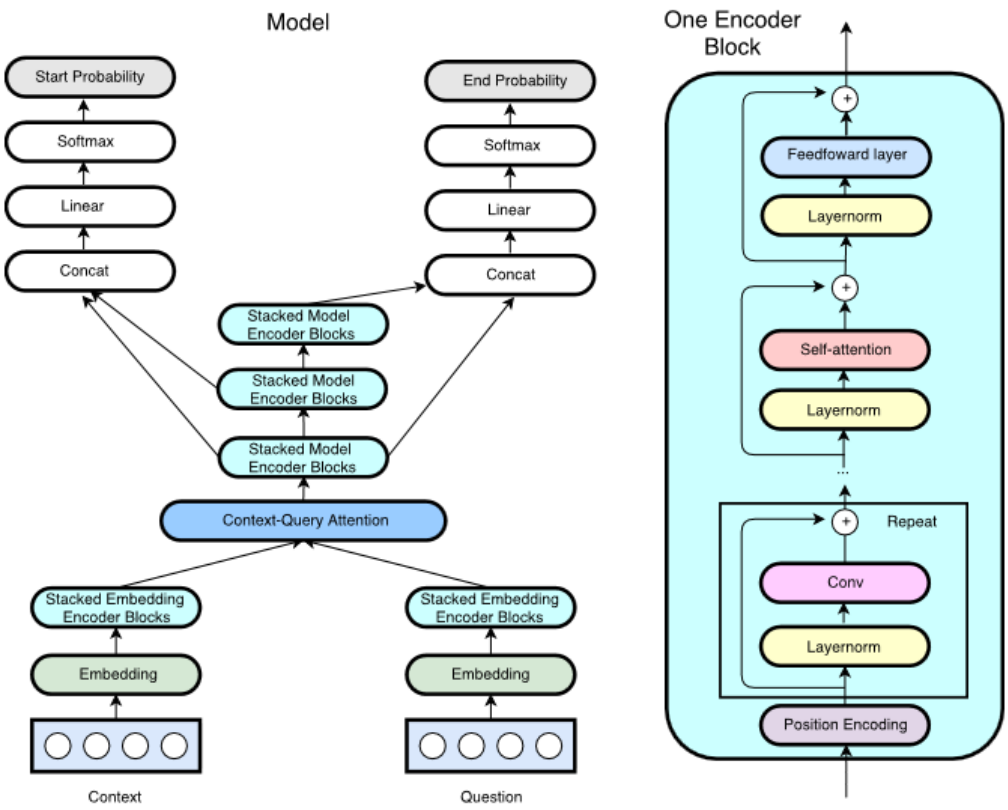

Figure 8 Un aperçu de l'architecture du QANet

La motivation clé derrière la conception de QANet est la suivante : la convolution capture la structure locale du texte, tandis que l'auto-attention apprend l'interaction globale entre chaque paire de mots. L'attention supplémentaire pour les requêtes contextuelles est un module standard pour construire le vecteur contextuel pour chaque position dans le paragraphe contextuel, qui est utilisé dans les couches de modélisation suivantes. La nature « feed-forward » de l'architecture proposée accélère le modèle de manière significative. Les expériences réalisées sur SQuAD montrent que le modèle est 3 à 13 fois plus rapide en apprentissage et 4 à 9 fois plus rapide en inférence. Comme le modèle est rapide, il peut être entraîné avec beaucoup plus de données que les autres modèles. Pour améliorer encore le modèle, les auteurs proposent une technique complémentaire d'augmentation des données pour augmenter les données d'entraînement. Cette technique paraphrase les instances d'entraînement en traduisant les phrases originales de l'anglais vers une autre langue, puis de nouveau vers l'anglais, ce qui non seulement augmente le nombre d'instances d'entraînement mais, diversifie également la formulation des questions-réponses.

Sur le dataset SQuAD, QANet entraîné avec les données augmentées obtient un score F1 de 84,6 sur l'ensemble de test, ce qui est nettement supérieur au meilleur résultat publié.

1.3.3.6 BERT

BERT (Bidirectional Encoder Representations from Transformers) [1] est un modèle de langage pré-entraîné similaire à ELMo et GPT-3 [60, 2]. Ce modèle est à l'origine du plus grand impact sur la recherche en NLP ces dernières années. BERT a obtenu des résultats de pointe sur onze tâches NLP. Souvent avec d'énormes améliorations. Il est conçu pour pré-entraîner des représentations bidirectionnelles profondes à partir de textes non étiquetés en conditionnant conjointement le contexte de gauche et de droite dans toutes les couches. Par conséquent, le modèle BERT pré-entraîné peut être affiné avec une seule couche de sortie supplémentaire afin de créer des modèles de pointe pour un large éventail de tâches, telles que le QA et l'inférence linguistique, sans modification substantielle de l'architecture spécifique à la tâche. L'architecture du modèle BERT est un codeur Transformer bidirectionnel multicouche basé sur l'implémentation originale décrite dans l'article original du Transformer [61].

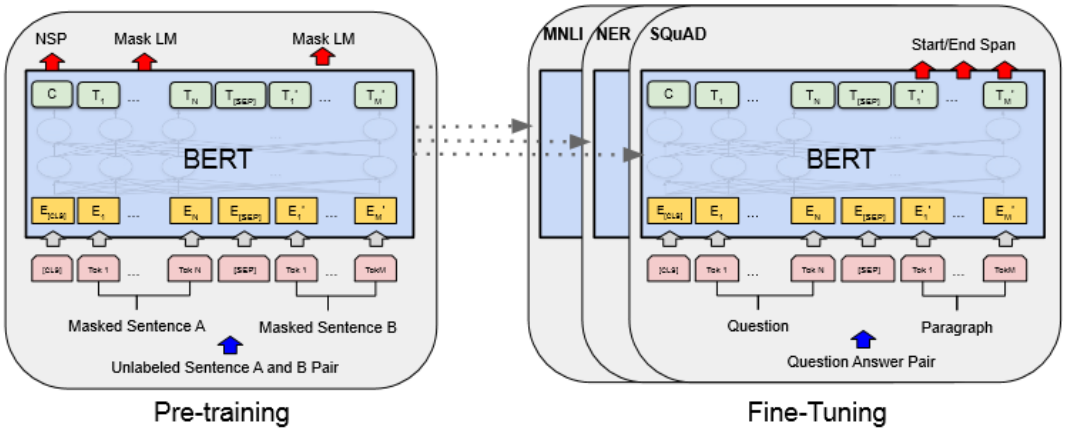

Figure 9 Procédures générales de pré-entraînement et d'adaptation (fine-tuning) de BERT

Le Framework du modèle comporte deux étapes : le pré-entraînement et le réglage fin (fine-tuning). Pendant le pré-entraînement, le modèle est entraîné sur des données non étiquetées au cours de différentes tâches de pré-entraînement. Pour le pré-entraînement, le modèle BERT est d'abord initialisé avec les paramètres pré-entraînés, et tous les paramètres sont affinés en utilisant des données étiquetées provenant des tâches en aval. Chaque tâche en aval dispose de

modèles affinés distincts, même s'ils sont initialisés avec les mêmes paramètres pré-entraînés. La Figure 9 montre l'utilisation de BERT dans le cas du QA.

Sur le dataset SQuAD v1.0 [5], BERT surpasse le meilleur système avec +1.5 F1 ensembliste et de +1.3 F1 en système unique. Sur la version v2.0 [16] de SQuAD, il surpasse le meilleur système avec +5,1 F1.

1.3.3.7 ALBERT

L'augmentation de la taille des modèles lors du pré-entraînement des modèles de langage se traduit souvent par une amélioration des performances dans les tâches en aval. Cependant, à un moment donné, il devient plus difficile d'augmenter la taille du modèle en raison des limitations de mémoire des GPU/TPU et des temps d'apprentissage plus longs. Pour résoudre ces problèmes, un nouveau modèle appelé ALBERT [62] est introduit. Il intègre deux techniques de réduction des paramètres qui lèvent les principaux obstacles à la mise à l'échelle des modèles pré-entraînés. La première est une paramétrisation d'intégration factorisée, en anglais « factorized embedding parameterization ». En décomposant la grande matrice d'intégration du vocabulaire en deux petites matrices, ALBERT sépare la taille des couches cachées de la taille de l'intégration du vocabulaire. Cette séparation permet d'augmenter plus facilement la taille des couches cachées sans augmenter de manière significative la taille des paramètres d'intégration du vocabulaire. La deuxième technique est le partage des paramètres entre les couches. Cette technique empêche le nombre de paramètres de croître avec la profondeur du réseau. Les deux techniques réduisent considérablement le nombre de paramètres pour BERT sans nuire aux performances, améliorant ainsi l'efficacité des paramètres. Une configuration ALBERT similaire à BERT-large a 18 fois moins de paramètres et peut être entraînée environ 1,7 fois plus rapidement. Les techniques de réduction des paramètres agissent également comme une forme de régularisation qui stabilise l'apprentissage et aide à la généralisation.

Pour améliorer encore les performances d'ALBERT, les auteurs ont également introduit une perte auto-supervisée, en anglais « self-supervised loss » pour la prédiction de l'ordre des phrases (SOP). SOP se concentre principalement sur la cohérence inter-phrase et est conçu pour remédier à l'inefficacité [63, 64] de la perte de prédiction de la phrase suivante (NSP) proposée dans le BERT [1] original.

Grâce à ces décisions de conception, les auteurs ont été en mesure de passer à des configurations ALBERT beaucoup plus grandes qui ont toujours moins de paramètres que BERT-large, mais qui atteignent des performances nettement supérieures. ALBERT établit un nouveau résultat de pointe sur SQuAD [5] avec un score F1 de 92,2

1.4 Conclusion

Dans ce chapitre, nous avons introduit le domaine de traitement automatique de langage naturel (NLP), en citant son historique et ses tâches les plus importantes. Ensuite, nous avons présenté l'état de l'art de la tâche NLP de « Question Answering » (QA). Nous avons commencé par une définition formelle de cette tâche, suivie par une classification des systèmes et approches proposées. Ensuite, nous avons décrit les datasets et challenges internationaux de références pour le QA tout en comparons leurs caractéristiques. Enfin, nous avons présenté les approches et les systèmes les plus importants du QA qui ont été proposés au fil des années. Dans le chapitre suivant, nous allons présenter l'état de l'art de la tâche de « Biomedical Question Answering » (BQA), une sous-tâche du QA qui est concernée par les textes du domaine médical.

Chapitre II :
État de l'art : Biomedical Question Answering (BQA)

2.1 Introduction

La recherche en médecine et biologie se développe rapidement, ce qui se traduit par une augmentation considérable des articles de recherche biomédicale. En moyenne, plus de trois mille articles sont rajoutés à PubMed[2] (la plus grande base de données d'articles de recherche biomédicale) chaque jour. Trouver des informations pertinentes dans cette littérature en pleine expansion devient de plus en plus difficile pour les chercheurs et les professionnels de la santé, ce qui accroît également le fossé entre la recherche et la pratique professionnelle. Ce constat peut être observé dans l'épidémie récente de COVID-19, où les chercheurs biomédicaux sont engagés dans une course contre la montre pour trouver des informations pertinentes sur les traitements possibles ou vaccins efficaces.

Une étude [3] sur les pratiques de recherche d'informations des professionnels de la santé a révélé que, pour une tâche de recherche donnée, il faut en moyenne une heure par base de données, trois minutes pour examiner la pertinence de chaque document, ce qui représente un temps de recherche total de quatre heures pour une tâche de recherche. La même étude a également révélé que les besoins des professionnels de santé ne sont pas entièrement pris en charge par les applications actuelles de recherche documentaire. Pour répondre à ces besoins, la recherche dans le domaine de l'exploration de textes biomédicaux [65], la recherche d'informations (IR), la réponse aux questions (QA), et la réponse aux questions biomédicales (BQA) a connu un essor considérable ces dernières années.

Les systèmes IR classiques tels que PubMed, bien que très utiles, renvoient toujours beaucoup plus de résultats de recherche que ce qui est idéalement souhaitable [3]. Il faut donc plus de temps pour évaluer la pertinence des documents renvoyés, puis extraire l'information requise et la synthétiser sous une forme qui peut facilement informer la prise de décision en matière de soins de santé. D'autre part, les systèmes QA ont le potentiel de surmonter les lacunes des systèmes IR classiques et de transformer positivement l'expérience de recherche. En effet,

[2] https://pubmed.ncbi.nlm.nih.gov

plutôt que de renvoyer des documents entiers, les systèmes QA peuvent extraire, et même synthétiser des réponses précises à des questions formulées naturellement.

Les modèles QA actuels peuvent désormais trouver des réponses précises à partir d'un texte, d'un passage ou d'un document entier. Malheureusement, lorsqu'ils sont appliqués directement à la littérature biomédicale, ces modèles donnent souvent des résultats insatisfaisants. Cela est dû à un changement de la distribution des mots des corpus généraux vers les corpus biomédicaux et également aux caractéristiques spécifiques de la littérature biomédicale, telles que des documents volumineux, une terminologie complexe propre au domaine et une typologie de questions spécifique au domaine. Par conséquent, plusieurs modèles et techniques QA biomédicale ont été proposés, et ces modèles sont souvent plus performants que leurs homologues du domaine général. Néanmoins, des modèles plus puissants restent nécessaires pour relever les défis spécifiques du QA biomédical.

Une autre raison de la lenteur des progrès du QA biomédicale par rapport au QA du domaine général est le nombre limité de datasets QA biomédicale, comparé avec le nombre de datasets QA du domaine général. Une autre limitation réside dans le nombre d'instances d'entraînement. Dans le domaine du QA biomédicale, le plus grand dataset annotées BioASQ [4] contient 3 243 instances, alors que le plus grand dataset QA dans le domaine général SQuAD v2.0 [5] contient 150 mille instances.

Les datasets QA biomédicales nécessitent l'annotation d'un expert, ce qui est coûteux et demande beaucoup de travail, contrairement au « crowdsourcing » qui est adopté par la majorité des datasets QA du domaine général. Pour remédier à ces deux limitations, un certain nombre de datasets de type cloze [66] construits automatiquement ont été proposés [67, 68, 69]. Ces datasets ont beaucoup plus d'instances d'entraînement, cependant, ils contiennent principalement des questions factuelles auxquelles il est possible de répondre sans raisonnement. Ce qui n'aide pas à construire des modèles QA puissants et généralisables.

Le reste du chapitre est organisé comme suit. D’abord, nous allons commencer par définir la tâche de BQA, puis nous allons énumérer et comparer ses datasets et challenges « benchmarks ». Ensuite, nous allons décrire les modèles et systèmes de références BQA. Enfin, nous allons terminer le chapitre par une conclusion.

2.2 Biomedical Question Answering (BQA)

2.2.1 Définition

L'acquisition de connaissances biomédicales est une tâche importante dans le domaine de la recherche d'information et de la gestion des connaissances. Les professionnels de la biomédecine ainsi que le grand public ont besoin d'une aide efficace pour accéder à des concepts biomédicaux complexes, les comprendre et les utiliser, par exemple : les médecins doivent connaître les preuves cliniques les plus récentes pour le diagnostic et le traitement des maladies et le grand public est de plus en plus intéressé par l'information sur son propre état de santé sur Internet.

Traditionnellement, les systèmes de recherche d'information « Information retrieval » (IR), c'est-à-dire les moteurs de recherche comme Google et PubMed, sont utilisés pour répondre à ces besoins d'information. Cependant, les systèmes IR classiques ne sont pas encore assez efficaces. Par exemple, une étude récente [3] indique qu'il faut 4 heures d'expertise pour répondre à des requêtes médicales complexes à l'aide de moteurs de recherche. Comparés aux systèmes de recherche qui renvoient généralement une liste de documents pertinents à lire par les utilisateurs, les systèmes QA qui fournissent des réponses directes aux questions des utilisateurs sont plus simples et intuitifs.

La réponse aux questions biomédicales « Biomedical question answering » (BQA) est une sous-tâche de l'NLP et du QA général, qui vise à répondre aux questions dans le domaine biomédical sur la base d'un ou plusieurs passages connexes. Récemment, de nombreuses approches basées sur les réseaux de neurones et les modèles de langage pré-entraîné « Pre-trained language models » (PLM) ont largement amélioré ses performances.

Dans la suite de ce chapitre, nous classons et analysons les systèmes BQA sous trois aspects différents : approches, contenus et formats. Un aperçu de cette classification est présenté brièvement dans la Figure 10 Classification des systèmes BQA.

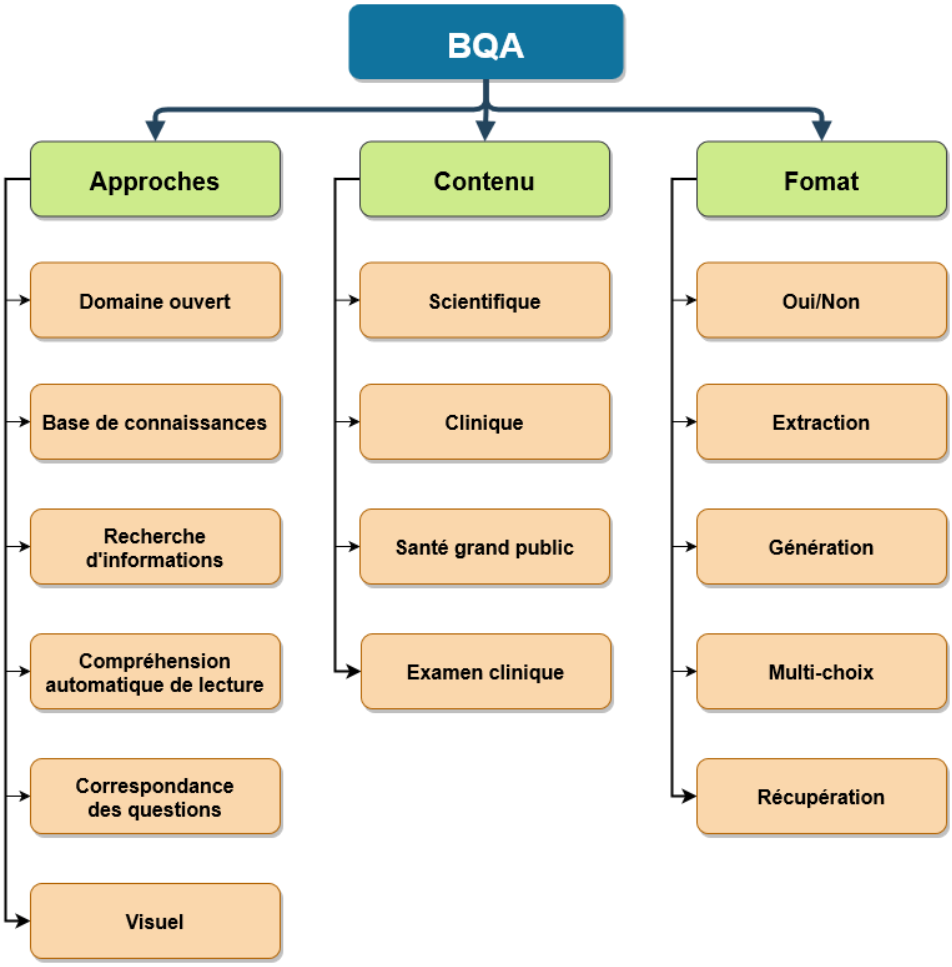

Figure 10 Classification des systèmes BQA

2.2.2 Formats des réponses aux questions

Les réponses des questions BQA se présentent sous différents formats : 1. Oui/Non ; 2. extraites : où les réponses sont extraites de contextes connexes ; 3. générées : où les réponses sont générées à partir de zéro et ne sont ni oui/non ni extraites des contextes ; 4. multi-choix : où toutes les réponses potentielles, qui sont spécifiques aux questions, sont données, et les modèles les classent ou choisissent la meilleure comme réponse ; 5. récupération : les systèmes renvoient une liste classée de documents ou d'extraits d'une collection de documents pré-spécifiée. Les mesures d'évaluation de BQA dépendent largement des formats de réponse. La définition de chaque type de réponse est donnée ci-dessous.

2.2.2.1 Questions de type Oui/Non

Dans le cadre d'une question BQA de type Oui/Non, les réponses sont soit "Oui", soit "Non ", ce qui peut être formellement défini comme suit :

$$A \in \{Yes, No\} \tag{2.1}$$

Où A désigne la réponse. Ce type de questions est souvent modélisé comme une tâche de classification de phrases. Les résultats de ce type de questions peuvent être évalués par la métrique de précision « Accuracy » (Acc)

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.2}$$

Où TP (True Positive) désigne le nombre de questions positives auxquelles il est répondu correctement, et TN (True Negative) le nombre de questions négatives auxquelles il est répondu correctement. FP (False Positive) et FN (False Negative) désignent respectivement le nombre de questions négatives et de questions positives auxquelles on a répondu incorrectement.

Le tableau 2 ci-dessous montre un exemple d'une question de type Oui/Non tiré du dataset BioASQ 10b.

Question	Does MicroRNA-21 (miR-21) contribute to cardiovascular disease?
Passage de contexte	During recent years, additional roles of miR-21 in cardiovascular and pulmonary diseases, including cardiac and pulmonary fibrosis as well as myocardial infarction have been described.
Réponse	Yes

Table 2 Exemple de question de type Oui/Non tiré du dataset BioASQ 10b

2.2.2.2 Questions de type extraction

Dans la BQA extractive, on peut répondre aux questions par des extraits du contexte. Formellement, nous désignons le contexte comme :

$$C = \{t_1, t_2, \dots, t_n\} \tag{2.3}$$

Où t_i est l'i-ème jeton du contexte et n est la longueur du contexte. Une réponse peut être formée comme suit :

$$A = \{t_k, t_{k+1}, \dots, t_{k+m-1}\} \subseteq C \quad (2.4)$$

Où A est un extrait de C qui commence à l'indice k et a une longueur de m mots, qui satisfont à :

$$1 \leq k \leq k + m - 1 \leq n \quad (2.5)$$

Les questions factoïdes sont courantes dans la BQA extractive, où les extraits sont généralement des entités biomédicales. Leurs réponses sont des mots simples ou des phrases courtes. Parfois, les questions ont plus d'une réponse, ce que l'on appelle des questions de type liste. Le tableau 3 ci-dessous montre un exemple d'une question de type factoïde tiré du dataset BioASQ 10b.

Question	Which disease can be treated with Delamanid?
Passage de contexte	This finding suggests that delamanid could enhance treatment options for multidrug-resistant **tuberculosis**
Réponse	tuberculosis

Table 3 Exemple d'une question de type factoïde tiré du dataset BioASQ 10b

Le tableau 4 ci-dessous montre un exemple d'une question de type liste tiré du dataset BioASQ 10b.

Question	Which mutations of SCN5A gene are implicated in Brugada syndrome?
Passage de contexte	Among the cohort of 19 patients, one missense mutation (G400A) in SCN5A was detected in a conserved region. An H558R polymorphism was detected on the same allele. Novel SCN5A mutation (Q55X) associated with age-dependent expression of Brugada syndrome presenting as neurally mediated syncope Novel SCN5A gene mutations associated with Brugada syndrome: V95I, A1649V and delF1617
Réponses	H558R Q55X V95I

Table 4 Exemple d'une question de type liste tiré du dataset BioASQ 10b

Pour le type de question factoïde, trois mesures d'évaluation sont utilisées. La précision stricte (SAcc), la précision indulgente, en anglais « Lenient Accuracy » (LAcc) et le rang moyen réciproque, en anglais « Mean Reciprocal Rank » (MRR). La précision stricte ne prend en compte que la première réponse retournée. Dans le cas de la précision indulgente, les cinq réponses retournées sont prises en compte. Si la réponse correcte est présente dans la liste des cinq réponses retournées, alors la question est considérée comme ayant reçu une réponse. MRR est identique à LAcc, sauf qu'il prend en compte l'ordre de la réponse correcte dans la liste des réponses retournées. Vous trouverez ci-dessous la définition de ces trois métriques.

$$SAcc = \frac{C_1}{n}, \qquad LAcc = \frac{C_5}{n}, \qquad MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{r(i)} \tag{2.6}$$

Où n est le nombre de questions factoïdes. C_1 est le nombre de questions factoïdes auxquelles on a répondu correctement en prenant en compte que le premier élément de chaque liste retournée. C_5 est le nombre de questions factoïdes correctement répondues en prenant en compte l'ensemble des cinq réponses retournées. $r(i)$ est la position de la réponse correcte dans la liste des cinq réponses retournées pour la question i. Si la liste des réponses retournées pour la question i ne contient pas la réponse correcte, alors $r(i) \to \infty$, c'est-à-dire, $\frac{1}{r(i)} = 0$.

Pour les questions de type liste, trois mesures d'évaluation sont adoptées par BioASQ [4], la précision (P), le rappel, en anglais « Recall » (R) et la F-mesure (F_1). Ces trois métriques sont définies comme suit.

$$P = \frac{TP}{TP + FP} \tag{2.7}$$

TP est le nombre d'entités qui sont mentionnées à la fois dans la liste retournée et dans la liste d'or, en anglais « Golden list » contenant les entités correctes ; FP est le nombre d'entités qui sont mentionnées dans la liste retournée, mais pas dans la liste d'or.

$$R = \frac{TP}{TP + FN} \tag{2.8}$$

FN est le nombre d'entités qui sont mentionnées dans la liste d'or, mais pas dans la liste retournée.

$$F_1 = 2 \times \frac{P \times R}{P + R} \quad (2.9)$$

La précision moyenne, le rappel moyen et la F-mesure moyenne sont obtenus en calculant la moyenne de la précision, du rappel et de la F-mesure sur les questions de la liste.

2.2.2.3 Questions de type génération

Dans la BQA générative, les réponses ne peuvent pas être extraites directement depuis le passage de contexte, et doivent donc être générées à partir de celui-ci. La BQA générative est plus robuste en situation réelle, car il n'y a pas de restriction sur le format des réponses et presque tous les datasets BQA peuvent être modélisés et évalués par des méthodes et des mesures BQA générative. Cependant, la BQA générative est considérée comme beaucoup plus difficile que la BQA extractive. Le tableau 5 ci-dessous montre un exemple d'une question de type génération, aussi appelée type résumé, tiré du dataset BioASQ 10b.

Question	What is Targeted Chromatin Capture (T2C)?
Passage de contexte	Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. Here we describe a new technique termed Targeted Chromatin Capture (T2C), to interrogate large selected regions of the genome. T2C provides an unbiased view of the spatial organization of selected loci at superior resolution (single restriction fragment resolution, from 2 to 6 kbp) at much lower costs than Hi-C due to the lower sequencing effort. T2C is an efficient, easy, and affordable with high (restriction fragment) resolution tool to address both genome compartmentalization and chromatin-interaction networks for specific genomic regions at high resolution for both clinical and non-clinical research. Here we describe a new technique termed Targeted Chromatin Capture (T2C), to interrogate large selected regions of the genome. Here we describe a new technique termed Targeted Chromatin Capture (T2C), to interrogate large selected regions of the genome.

Réponse	Targeted Chromatin Capture (T2C) is an efficient, easy, and affordable with high (restriction fragment) resolution tool to address both genome compartmentalization and chromatin-interaction networks for specific genomic regions at high resolution for both clinical and non-clinical research.

Table 5 Exemple d'une question de type génération (ou résumé) tiré du dataset BioASQ 10b

Il existe plusieurs métriques d'évaluation pour les questions BQA génératives, qui mesurent essentiellement la similarité entre deux phrases, c'est-à-dire : la phrase prédite et la ou les références.

Les réponses générées peuvent être notées manuellement, les examinateurs jugeant une réponse en fonction de différentes mesures de qualité, telles que le rappel « Recall » de l'information, la précision de l'information, la répétition de l'information et la lisibilité, selon leurs impressions subjectives. Le score manuel est la méthode de référence pour évaluer la BQA générative et il est défini comme la métrique principale pour la tâche de génération de réponses idéales dans le challenge BioASQ.

Cependant, l'évaluation manuelle des réponses de grands datasets génératives prend beaucoup de temps et son coût est prohibitif. Des mesures d'évaluation automatiques sont donc nécessaires. Ces mesures comparent essentiellement la réponse prédite avec la ou les références. La mesure ROUGE (Recall-Oriented Understudy for Gist Evaluation) [70] est une mesure couramment utilisée.

$$ROUGE_n = \frac{\sum_{ref \in Refs} \sum_{gram_n \in ref} Count(gram_n|cand)}{\sum_{ref \in Refs} \sum_{gram_n \in ref} Count(gram_n|ref)} \quad (2.10)$$

Où $gram_n$ se réfère à tout n-gramme dans une référence ref. $Count(gram_n|ref)$ indique le nombre de fois où $gram_n$ apparaît dans la référence ref, et donc le dénominateur compte les n-grammes dans les références. $Count(gram_n|cand)$ indique combien de fois $gram_n$ apparaît dans la réponse du candidat.

Contrairement à ROUGE qui mesure le rappel « Recall » des candidats, une autre métrique largement utilisée est BLEU [71], s'intéresse davantage à la précision.

$$BLEU_n = \frac{\sum_{gram_n \in cand} Count(gram_n|Refs)}{\sum_{gram_n \in cand} Count(gram_n|cand)} \quad (2.11)$$

où $cand$ est la réponse candidate générée par les systèmes BQA, et $Count_{Refs}(gram_n)$ est le nombre maximum de $Count(gram_n|ref)$ parmi toutes les références. En outre, $Count(gram_n|Refs)$ doit être modifié en $Count_{clip}$, qui désigne le plus petit nombre entre $Count(\cdot\,|Refs)$ et $Count(\cdot\,|cand)$, de sorte que la valeur de $BLEU_n$ ne dépasse pas 1.

2.2.2.4 Questions de type multi-choix

Les questions sont associées à plusieurs réponses candidates. Formellement, pour la question Q_i, la réponse est :

$$A_i \in \mathcal{A}_i = \{A_{i1}, A_{i2}, \dots, A_{in_i}\} \quad (2.12)$$

Où $\mathcal{A}_i$ est un ensemble de n_i réponses candidates. Les systèmes BQA à choix multiples modélisent généralement la tâche comme une tâche de classification où les classes sont les réponses candidates.

La précision est une mesure courante pour évaluer les modèles BQA multichoix. En outre, comme les systèmes à choix multiples renvoient parfois une liste classée des candidats, la précision et le MRR peuvent également être utilisés comme métriques d'évaluation.

2.2.2.5 Questions de type récupération

Les questions de type récupération nécessitent comme réponses une liste classée de documents ou d'extraits d'une collection de documents prédéfinie. Ce type de question est évalué à travers trois métriques :

Précision: La précision est le nombre de documents pertinents retrouvés par une recherche divisé par le nombre total de documents retrouvés par cette recherche.

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \quad (2.13)$$

Au lieu de prendre en compte tous les documents, la précision peut également être évaluée pour un nombre donné de documents récupérés, où le modèle est évalué en ne prenant en compte

que les k premiers documents récupérés. Dans ce cas, la métrique de précision est appelée précision à k ou $P@K$.

Rappel : Le rappel, en anglais « Recall » est le nombre de documents pertinents récupérés par une recherche divisé par le nombre total de documents pertinents existants.

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|} \quad (2.14)$$

Précision moyenne « Mean Average Precision » : Pour un seul besoin d'information, la précision moyenne est la moyenne de la valeur de précision obtenue pour l'ensemble des documents top-k existant après la récupération de chaque document pertinent, et cette valeur est ensuite moyennée sur les besoins d'information. L'équation de la précision moyenne (MAP) est la suivante.

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk}) \quad (2.15)$$

Avec, l'ensemble des documents pertinents pour un besoin d'information $q_j \in Q$ est $\{d_1, \dots, d_{mj}\}$ et R_{jk} est l'ensemble des documents de recherche classés à partir du premier résultat jusqu'au document d_k.

Rang réciproque moyen « Mean Reciprocal Rank » : Le rang réciproque moyen est défini comme la moyenne des rangs inverses pour toutes les requêtes Q

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (2.16)$$

Où $rank_i$ est la position du premier document pertinent pour la requête i

2.2.3 Datasets et challenges internationaux

Sur la base du contenu des questions, nous pouvons classer les datasets BQA en quatre types : 1. **Scientifique** : les questions portent sur des découvertes biomédicales de pointe et les réponses doivent être trouvées dans la littérature scientifique ; 2. **Clinique** : les questions portent sur la prise en charge des patients en clinique. Les réponses peuvent être trouvées dans

les dossiers médicaux électroniques (DME) des patients ou dans la littérature clinique ; 3. **Santé grand public** : où les questions sont posées par le grand public, généralement sur des moteurs de recherche en ligne ; 4. **Examen** : où les paires question-réponse sont tirés des examens de licence médicale. Ci-dessous, nous examinons et comparons les datasets et challenges internationaux BQA les plus importants dans l'ordre chronologique.

La tâche génomique de TREC : Le volet génomique de TREC est l'un des plus grands et des plus longs challenges dans le domaine de la biomédecine. Il s'est étalé de 2003 à 2007. En 2006 et 2007, la piste a ajouté une nouvelle tâche de QA [72, 73]. Le dataset génomiques fourni par TREC était l'un des premiers jeux de données proposés pour la BQA. La tâche consistait à extraire des passages pertinents pour une question à partir d'un corpus de 162 259 documents. Le dataset a été construit à partir d'articles de recherche depuis 49 journaux biomédicaux. Cette première tentative de QA biomédical, bien que très importante, a souffert du nombre limité de questions proposées, 28 en 2006 et 36 questions en 2007. Le petit nombre d'instances questions-passages proposées n'était pas suffisant pour évaluer avec précision les modèles BQA proposés.

BioASQ (2015) : Contrairement aux autres datasets, BioASQ [4] est annoté manuellement par des experts biomédicaux, ce qui est particulièrement important pour la tâche de BQA. Il contient environ 3 mille instances de questions-réponses, beaucoup moins que les autres datasets, en raison du coût des annotations manuelles des experts. Ce dataset est publié dans le cadre d'un défi beaucoup plus vaste intégrant la classification hiérarchique de textes et de passages, le QA, le résumé multi-documents et la de la génération de langage naturel (NLG). Chaque instance du dataset contient des questions, des réponses, des documents connexes et des concepts. Il y a quatre types de questions dans dataset, oui/non, factoïde, liste et résumé. Pour chacun de ces quatre types de questions, une réponse exacte et une réponse idéale sont attendues, à l'exception du type résumé où seule une réponse idéale est attendue. Les modèles ciblant ce dataset doivent alors répondre avec des réponses exactes comme oui ou non pour les questions oui/non, des entités nommées dans le cas des questions factoïde, liste d'entités nommées pour les questions de type liste, et une réponse idéale comme des résumés sous forme de phrase ou de paragraphe pour tous les types de questions.

BioRead (2018) : BioRead [68] est actuellement le plus grand dataset BQA, et l'un des plus grands datasets QA en général. Il a été construit de manière cloze [66] et contient environ 16,4 millions d'instances. Les auteurs du dataset ont également fourni une petite version appelée BioReadLite avec 900 mille instances, pour les chercheurs disposant de ressources informatiques limitées. Pour construire l'ensemble de données, les auteurs ont sélectionné de manière aléatoire 906 mille articles de la section à accès libre de PubMed Central (PMC), puis ils ont appliqué METAMAP [74] qui reconnaît les mots et les phrases se référant aux concepts de l'UMLS (Unified Medical Language System) ou système de langage médical unifié pour annoter les entités biomédicales dans chaque article. Le dataset a été construit en utilisant la stratégie de style cloze suivante : des séquences de 21 phrases ont été extraites des articles. Les 20 premières phrases ont été utilisées comme un passage et la dernière comme une question de style cloze. Une entité biomédicale de la question était ensuite remplacée par un « placeholder ». La tâche proposée par les auteurs du dataset est donc de deviner quelle entité biomédicale du passage peut le mieux remplir le caractère de remplacement (le « placeholder »). Le principal inconvénient du dataset BioRead est qu'il souffre d'un taux de bruit assez élevé. Car plusieurs questions proviennent de la section des références ou incluent des légendes et des notes de bas de page. Cela est dû au fait que les auteurs ont utilisé le texte intégral des articles. Une autre source de bruit provient de METAMAP [74], qui parfois identifie par erreur des entités biomédicales.

BMKC (2018) : Le dataset Biomedical Knowledge Comprehension (BMKC) [69] suit la même méthode cloze [74] que BioRead. Il contient environ 500 mille instances de questions contextuelles réparties en deux sous-ensembles : Biomedical Knowledge Comprehension Title (BMKC T) et Biomedical Knowledge Comprehension Last Sentence (BMKC LS). Ce dataset a été construit en trois étapes. Tout d'abord, les auteurs du dataset ont rassemblé des articles de recherche de PubMed. Ils ont utilisé les résumés (abstracts) comme contextes, les questions, quant à elles, ont été générées selon deux approches différentes. Dans le sous-ensemble BMKC T la question est construite à partir du titre de l'article. Dans le sous-ensemble BMKC LS, la dernière phrase du résumé a été utilisée comme question. Dans la deuxième étape, ils ont extrait des entités pour générer des réponses candidates aux questions de type cloze. Pour ce faire, ils ont utilisé l'extracteur d'entités nommées biomédicales de Biomedical Entity Search Tool

(BEST) [75] ainsi que Medical Subject Headings (MeSH), un thésaurus hiérarchique de vocabulaire biomédical. Grâce à MeSH, ils ont pu regrouper différentes entités nommées biomédicales ayant la même signification en une seule identification MeSH (ID). Ensuite, ils ont remplacé toutes les entités nommées par leurs ID d'entités équivalentes. La dernière étape consistait à filtrer les paires contexte-réponse en suivant deux règles de qualité qu'ils ont définies. Le but ici était de limiter le bruit et de rendre la tâche proposée plus difficile.

MedQA (2018) [76] : Le dataset MedQA se base sur un corpus de médecine clinique et vise à simuler un scénario du monde réel. Les auteurs de ce dataset ont rassemblé une grande collection de textes de type clinique, pour apprendre à lire des textes à grande échelle. Les questions sont tirées d'examens de certification médicaux en Chine, où les médecins sont évalués sur leurs connaissances professionnelles et leur capacité à établir un diagnostic. Les questions de ces examens sont variées et nécessitent généralement une compréhension des concepts médicaux connexes pour y répondre. Un modèle d'apprentissage automatique doit apprendre à trouver des informations pertinentes dans la collection de documents, à raisonner sur ces informations et à prendre des décisions concernant la réponse.

emrQA (2018) [77] : emrQA est un dataset BQA à grande échelle pour les dossiers médicaux électroniques, construit automatiquement en exploitant les annotations d'experts existantes sur les notes cliniques. Le dataset résultant compte 1 million de questions-formes logiques et plus de 400 mille paires de questions-réponses.

MedQuAD (2019) : Le dataset MedQuAD [78] contient 47 457 paires de questions-réponses annotées sur les maladies, les médicaments, les examens médicaux, les procédures et les traitements, extraites de 12 sites web médicaux fiables de la section des ressources médicales du site web du National Institutes of Health (NIH). Chaque page web des sites web décrit un sujet spécifique (par exemple, le nom d'une maladie) et comprend souvent des synonymes du sujet principal. Pour chaque site web, les auteurs de ce dataset ont crawlé le site web, et ont extrait le texte libre de chaque page ainsi que les synonymes du sujet principal. Les paires question-réponse ont ensuite été générées automatiquement à partir de chaque page en se basant sur le titre et la structure de la page. Chaque question a également été annotée avec le sujet et les synonymes associés.

PubMedQA (2019) : Le dataset PubMedQA [79] est le premier dataset BQA qui nécessite du raisonnement, en particulier du raisonnement sur des contenus quantitatifs. Il est collecté à partir des résumés de PubMed et compte mille instances annotées par des experts, 61,2 mille instances non annotées et 211,3 mille instances générées artificiellement. Chaque instance comporte une question, un contexte, une réponse longue et une réponse oui/non/peut-être.

BIOMRC (2020) : Le dataset BIOMRC [67] est une version améliorée de BIOREAD [68]. Il contient 812 mille instances de questions-passage, et également livré avec une petite version de 100 mille instances pour être utilisé par les chercheurs avec une puissance de calcul limitée. Contrairement à BioRead, les auteurs de ce dataset ont utilisé PUBTATOR [80], un référentiel qui fournit environ 30 millions de résumés et leurs titres correspondants de PUBMED en plus des articles en texte intégral depuis le sous-ensemble PMC à accès libre, PUBTATOR fournit également des notations automatiques des concepts biomédicaux. Une autre différence entre la construction de ce dataset et BioRead est l'utilisation de l'outil d'annotations d'entités biomédicales de DNORM [81] qui est plus précis que METAMAP [74]. Ce dataset suit la même stratégie de construction de type cloze que BioRead, mais avec une attention particulière à la réduction du bruit, qui est le principal inconvénient du dataset BioRead. Pour cela, en plus d'adopter DNORM au lieu de METAMAP, les auteurs de BIOMRC ont utilisé uniquement les résumés et les titres des articles comme passages et questions, au lieu d'utiliser le texte intégral comme dans le cas de BioRead, ceci afin d'éviter d'extraire du texte en croisant sections, ou des références, des légendes, des notes de bas de page et des tableaux.

Dans le Tableau 6, nous comparons les datasets présentés précédemment en termes de sources de données, de nombre d'instances d'entraînement, de raisonnement, types de réponses supportées, et la procédure adoptée dans la construction du dataset.

Dataset	Source	Nombre instances	Type de contenu	Types de réponses	Type de construction
TREC Genomics	49 revues biomédicales	64	Scientifique	Factoïde	Annoté manuellement
BioASQ	experts biomédicaux	2K	Scientifique	Yes/no/ Factoïde /liste/ résumé	Annoté manuellement

BioRead	PubMed	16.4M	Scientifique	Réponses à choix multiples	Cloze-style
BMKC	PubMed	842 907	Scientifique	Réponses à choix multiples	Cloze-style
MedQA	Examen national de licence médicale en Chine	235k	Examens	Réponses à choix multiples	Annoté manuellement
emrQA	Notes cliniques	455k	Clinique	Factoïde	Extraction
MedQuAD	Les sites web du NIH	47 457	Santé grand public	Factoïde	NA
PubMedQA	PubMed	273.5K	Scientifique	yes/no/maybe	NA
BIOMRC	PubMed	812k	Scientifique	Réponses à choix multiples	Cloze-style

Table 6 Comparaison des caractéristiques des datasets BQA présentés

2.2.4 Approches de base

Les systèmes BQA peuvent être classés en 6 approches principales, comme le montre la Figure 10 : **1**. **Approche à domaine ouvert** : le cadre où aucun matériel de soutien (par exemple, des documents pertinents, des images) n'est fourni dans la tâche. Elle inclut les systèmes non-neuronaux traditionnels basés sur un pipeline complexe de modules de traitement des questions, des documents et des réponses ; **2**. **l'approche par base de connaissances (KB)** : où les systèmes traduisent explicitement les questions d'entrée en requêtes RDF à rechercher ou utilisent implicitement les connaissances intégrées de certaines KB biomédicales pour obtenir les réponses ; **3**. **l'approche par recherche d'information (IR**) : où les systèmes récupèrent les documents pertinents pour répondre aux questions données ; **4**. **l'approche par compréhension automatique de la lecture (MRC)** : où les systèmes lisent des contextes donnés sur les questions pour prédire les réponses. Les contextes de l'approche MRC peuvent être fournis par l'approche IR ; **5. L'approche QE (Question Entailment)** : les systèmes trouvent d'abord des questions similaires auxquelles on a déjà répondu dans une base de données de paires QA et réutilisent leurs réponses pour répondre à la question donnée ; **6. l'approche VQA (Visual QA)** : les systèmes utilisent des méthodes multimodales pour répondre aux questions avec des images. Ces approches sont décrites en détail ci-dessous, ainsi que les principaux systèmes proposés pour chacune d'elles.

2.2.4.1 L'approche à domaine ouvert

Traditionnellement, les systèmes BQA ouverts se composent de 3 phases principales : **1. le traitement des questions**, où les systèmes déterminent le type de la question et le type correspondant de la réponse attendue et puis forment des requêtes qui sont transmises à certains systèmes de récupération de documents ; **2. Le traitement des documents**, où les systèmes récupèrent les documents pertinents à partir des requêtes générées à l'étape précédente puis extraient les réponses candidates des documents pertinents ; **3. Le traitement des réponses**, où les systèmes classent les réponses candidates en fonction de certains critères. La Figure 11 montre l'architecture générale d'un système BQA traditionnel à domaine ouvert.

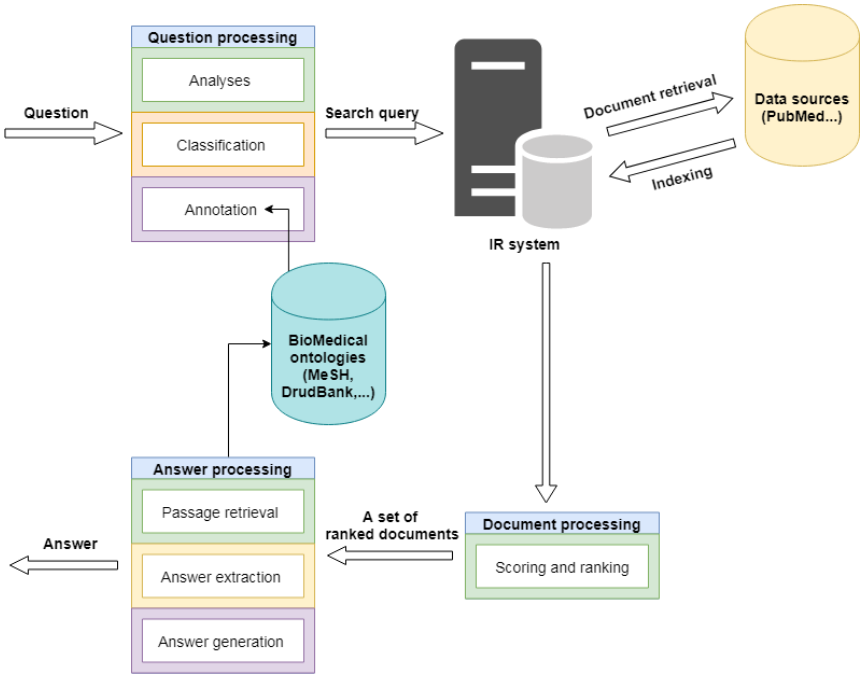

Figure 11 L'architecture générale d'un système BQA traditionnel

Ci-dessous, la description des systèmes BQA traditionnels à domaine ouvert les plus importants, présentés dans l'ordre chronologique.

Niu et al. [82] explorent la méthode PICO (P : Patient/Problème, I : Intervention, C : Comparaison, O : Outcome) dans le projet EPoCare en utilisant des extracteurs simples basés sur des mots-clés. Demner-Fushman et Lin ont étudié plus en détail la BQA sémantique basée sur PICO pour la pratique de la médecine fondée sur les preuves (EBM) dans une série de travaux [83, 84, 85], où l'étape principale consiste à rechercher des articles PubMed annotés

avec des connaissances médicales extraites. Huang et al. [86] étudient la faisabilité de l'utilisation du format PICO pour représenter les questions cliniques et concluent que PICO est principalement axé sur les questions cliniques de type thérapeutique et ne convient pas pour représenter les autres (par exemple : pronostic, étiologie). Pour aborder un plus large éventail de scénarios, la plupart des autres systèmes BQA traditionnels à domaine ouvert acceptent les questions en langage naturel : Le système de réponse aux questions de définition médicale MedQA [87] est le premier système BQA entièrement mis en œuvre qui génère des réponses par résumé extractif pour les questions de définition des utilisateurs à partir de grands corpus de textes. BioSquash [88] est adapté du synthétiseur de domaine général Squash [89] et se concentre sur le résumé de documents biomédicaux dans un scénario QA. Terol et al. [90] utilisent des formes logiques pour la BQA, où l'étape principale consiste à dériver les formes logiques des questions et à les utiliser pour générer les réponses. HONQA [91] est un système BQA bilingue français/anglais, qui se concentre sur la classification supervisée des types de questions et de réponses pour la réponse aux questions. Lin et al. [92] explorent la réponse aux questions sur les événements biomoléculaires avec des entités nommées en utilisant la correspondance syntaxique et sémantique des caractéristiques. Gobeill et al. [93] génèrent 200 questions à partir de bases de données relationnelles biomédicales pour évaluer leur plateforme EAGLi1. Cao et al. [94] proposent le système askHERMES, un système BQA qui effectue plusieurs analyses sémantiques, y compris la classification des sujets de questions et le regroupement de contenu, afin de fournir des résumés extractifs pour les questions cliniques. SemBioNLQA [95] est un système BQA ouvert traditionnel qui peut retourner les quatre types de réponses requises par BioASQ : oui/non, factoïde, liste et résumé.

L'approche BQA traditionnelle à domaine ouvert s'appuie sur des règles élaborées manuellement et sur divers modules ad hoc dans ses pipelines complexes. Bien que ces éléments puissent être nécessaires dans les applications industrielles, ils sont difficiles à développer et à maintenir dans le milieu universitaire. En outre, la plupart des systèmes traditionnels BQA à domaine ouvert n'ont pas été validés sur des datasets à grande échelle. Avec l'introduction de divers ensembles de données BQA axés sur des tâches spécifiques BQA, seuls quelques systèmes BQA à domaine ouvert ont été proposés récemment.

2.2.4.2 L'approche par base de connaissances (KB)

Nous définissons les bases de connaissances biomédicales comme des bases de données qui décrivent les entités biomédicales et leurs relations, qui peuvent généralement être stockées dans des triplets sujet-prédicat-objet RDF. Des efforts considérables ont été déployés pour construire des bases de données biomédicales, notamment des ontologies telles que Medical Subject Headings (MeSH) pour les sujets de textes biomédicaux, Gene Ontology (GO) [96, 97] pour les termes génétiques, International Classification of Diseases (ICD) pour les maladies et Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) [98] pour les termes médicaux. Le système de langage médical unifié (UMLS) construit un métathésaurus qui intègre près de 200 KB biomédicales différentes comme MeSH et ICD.

Divers systèmes basés sur les bases de connaissances biomédicales ont été introduits. Traditionnellement, on peut d'abord convertir les questions du langage naturel en requêtes RDF et les utiliser pour rechercher les réponses dans les bases de données, ce que nous appelons l'approche d'**interrogation explicite des connaissances** « Explicit Knowledge Querying ». Récemment, de nombreux travaux ont tenté d'intégrer les bases de données biomédicales dans les modèles neuronaux par le biais de modules de pré-entraînement ou de raisonnement. Nous les considérons comme l'approche d'**intégration implicite des connaissances** « Implicit Knowledge Integration ».

Interrogation explicite des connaissances : pour utiliser les KBs avec la BQA, une approche évidente consiste à traduire les questions du langage naturel en langage d'interrogation comme SPARQL pour rechercher la réponse dans les KBs. Par exemple, la question en langage naturel "Quelles maladies l'aspirine peut-elle traiter ?" se traduit par la requête "head name : aspirine AND predicate name : traiter AND tail type : maladie". Plusieurs systèmes proposés suivent cette approche. Rinaldi et al. [99] adaptent le système ExtrAns du domaine général [100] au domaine de la génomique. Ils convertissent d'abord les documents en formes logiques minimales (MLF) et les utilisent pour construire une KB pendant la phase hors ligne. Dans la phase en ligne, le système convertit également la question donnée en MLFs par le même mécanisme, et obtient ensuite la réponse en recherchant la KB MLFs construite. Abacha et Zweigenbaum [101, 102] proposent MEANS pour la BQA médicale qui convertit les questions

en requêtes SPARQL avec un pipeline de classification des types de questions, de recherche des types de réponses attendues, de simplification des questions, de reconnaissance des entités médicales, d'extraction des relations sémantiques et de construction de requêtes SPARQL basée sur les entités et les relations sémantiques. Kim et Cohen [103] présentent Linked Open Data Question Answering (LODQA) pour générer des requêtes SPARQL pour SNOMED-CT par des relations prédicat-argument à partir de phrases.

Intégration implicite des connaissances : l'utilisation des KBs en interrogeant directement la base de données RDF est une tâche simple. Cependant, les méthodes de traduction de questions ne peuvent pas extraire des réponses avec des chaînes logiques complexes et impliquent des pipelines complexes. Pour résoudre ce problème, de nombreux travaux ont tenté d'intégrer implicitement les connaissances biomédicales, notamment les entités, les relations et la structure des graphes des bases de données, dans des modèles neuronaux de bout en bout. Plusieurs systèmes proposés suivent cette approche. Li et al. [104] proposent un modèle qui extrait les entités des questions et des options de réponse, et trouve des triplets connexes comme sources de connaissances. Un réseau convolutif de graphes (GCN) [105] peut capturer des informations de sous-graphes multi-sauts et un mécanisme de co-attention est responsable de l'injection de connaissances dans les représentations des questions et des options. Les modèles de langage pré-entraînés sont largement utilisés pour les modèles BQA en tant qu'encodeurs de texte. Récemment, certains modèles de langage enrichis de KB biomédicales ont été proposés : Hao et al. [106] ajoutent une tâche de classification de triplets pour les relations UMLS afin d'entraîner BERT [1] et ALBERT [62]. CODER [107] utilise les synonymes et les relations d'entités de l'UMLS pour obtenir des représentations d'entités médicales inter-linguistiques. Les contextes textuels peuvent être insuffisants pour répondre aux questions biomédicales, et les connaissances biomédicales des bases de données peuvent être utilisées pour les compléter. L'utilisation des bases de connaissances pour améliorer les modèles de langage pré-entraînés, qui sont omniprésents dans les systèmes BQA actuels, est une direction prometteuse à explorer.

2.2.4.3 L'approche par recherche d'information (IR)

La BQA par recherche d'information (RI) désigne l'approche qui utilise les méthodes BQA de la IR pour récupérer des extraits de texte pertinents à partir de certaines collections de

documents pour la question donnée. Les extraits récupérés peuvent être soit directement utilisés comme réponses, soit transmis à des modèles de compréhension automatique de la lecture.

PubMed et PubMed (PMC) Central sont les collections de documents biomédicaux les plus utilisées dans cette approche. Toutes les deux ont été développées et sont maintenues par la bibliothèque nationale de médecine « National Library of Medicine » (NLM) des États-Unis. PubMed fournit un accès gratuit à plus de 30 millions citations de la littérature biomédicale, où chaque citation contient principalement le titre de l'article, des informations sur l'auteur, un résumé et des indices sémantiques. PubMed Central comprend les textes complets de plus de 6 millions d'articles biomédicaux en plus des informations fournies dans les citations PubMed. Des corpus plus spécifiques sont également utilisés dans certains cas afin de filtrer le bruit potentiel dans PubMed et PMC, comme CORD-19 [108] pour le cas de COVID-19.

Ci-dessous la description des systèmes BQA les plus importants qui suivent l'approche de recherche d'information, présentés dans l'ordre chronologique. Nous allons principalement discuter des méthodes BQA IR pour le défi BioASQ [4] tâche B phase A. Qui est le plus grand défi IR BQA.

Wishart [109] reclasse et combine les phrases des documents récupérés pour former les réponses idéales pour la phase B de la tâche BioASQ, et génère des réponses exactes à partir des réponses idéales en fonction du type de question. L'équipe USTB [110] remporte tous les lots dans la phase de récupération de documents, des extraits et de concepts dans BioASQ 5. Elle utilise le modèle de dépendance séquentielle [111], le pseudo feedback de pertinence, le modèle de dépendance séquentielle [112] et le modèle de divergence aléatoire [113]. L'équipe AUEB propose une série de modèles [114, 115, 116] qui remportent la plupart des défis de la phase A de la tâche B depuis BioASQ 6. À BioASQ 6, ils [114] utilisent le modèle de pertinence convolutif récurrent tenant compte de la position [117] et le modèle de correspondance de pertinence profonde [118] pour la récupération de documents, et utilisent le modèle Bi-CNN de base [119] pour la récupération d'extraits. Ils remportent 3/5 et 5/5 lots pour la récupération de documents et d'extraits à BioASQ 6, respectivement. À BioASQ 7, ils [115] combinent le système de récupération de documents et d'extraits en modifiant leur système BioASQ 6 pour qu'il produise également le score de pertinence au niveau de la phrase (c'est-à-dire : extrait)

dans chaque document. Ils ont gagné 4/5 et 4/5 lots pour la récupération de documents et d'extraits dans BioASQ 7, respectivement. Dans BioASQ 8, ils [116] continuent à utiliser ce système et gagnent 2/5 pour l'extraction de documents et 4/5 pour l'extraction d'extraits. Les méthodes d'extraction traditionnelles comme TF-IDF et BM25 ont été aussi bien étudiées et omniprésentes dans l'approche BQA de la RI. Les études futures doivent se concentrer davantage sur le pré-entraînement des méthodes de reclassement basées sur les modèles de langue (LM) [120], et sur la manière de mieux relier les modèles de IR et de MRC pour la BQA.

2.2.4.4 L'approche par compréhension automatique de la lecture (MRC)

La compréhension automatique de la lecture « Machine Reading Comprehension » (MRC) est une tâche QA/BQA bien étudiée, où les systèmes BQA répondent à des questions sur des contextes textuels donnés, par exemple une question spécifique sur les dosages de médicaments dans les DME d'un patient. Les ensembles de données MRC BQA sont généralement spécialisés dans leur contenu et ont un format de réponse prédéterminé, de sorte que la plupart des méthodes MRC BQA développées sur ces datasets sont des modèles neuronaux de bout en bout.

Avec l'introduction des datasets MRC à grande échelle comme SQuAD [5, 16], une variété de modèles MRC neuronaux ont été proposés qui améliorent de manière incrémentale la performance de la tâche d'MRC, comme DCN [53], Bi-DAF [13], FastQA [121]. Les incorporations de mots contextualisées pré-entraînées par des modèles de langage (LM) comme ELMo [60] et BERT [1] montrent des améliorations significatives sur diverses tâches NLP, y compris l'MRC. Des LM pré-entraînés sur des corpus biomédicaux, tels que BioELMo [122], BioBERT [123], SciBERT [124], BERT clinique [125] et PubMedBERT [126], améliorent encore leurs performances dans le domaine. Les expériences de sondage et les analyses de Jin et al. [122] indiquent qu'un meilleur encodage du type d'entité biomédicale et des informations relationnelles conduit à la supériorité des représentations textuelles « embeddings » pré-entraînés spécifiques au domaine. Nous passons dans la suite en revue les systèmes les plus performants de la phase B de la tâche B des défis BioASQ afin de montrer l'évolution des méthodes MRC en BQA.

Les deux premiers défis BioASQ [127, 128] utilisent une base de référence motivée par Watson [129] il s'agit d'un ensemble de fonctions de notation qui classent les concepts

pertinents avec une coercition de type pour répondre aux questions données. Le système Fudan [130] de BioASQ 3B contient trois composants principaux : 1. l'analyse des questions qui extrait principalement les types de réponses sémantiques des questions ; 2. la génération de candidats par PubTator [80] et les outils POS de Stanford ; 3. le classement des candidats basé sur la fréquence des mots. L'équipe de SNU [131] combine directement les passages pertinents récupérés pour générer la réponse idéale et atteindre des performances de pointe. À BioASQ 4B : l'équipe HPI [132] propose un algorithme basé sur LexRank [133] pour générer des réponses idéales, qui n'utilise que des entités nommées biomédicales dans la fonction de similarité. Ils remportent 1/5 lot dans la génération de réponses idéales. À BioASQ 5B : l'équipe de l'UNCC [134] utilise un résumé extractif basé sur le chaînage lexical pour obtenir les meilleurs scores ROUGE pour la génération de réponses idéales.

Au cours des dernières années de BioASQ, l'apprentissage par transfert a gagné plus d'attention, où les modèles sont d'abord pré-entraînés sur des datasets QA de domaine général à grande échelle ou sur des datasets BQA, puis affinés sur l'ensemble d'entraînement de BioASQ. Wiese et al. [135] obtiennent des performances de pointe sur les questions factoïdes et des performances compétitives sur les questions de type liste en transférant le modèle FastQA [121] pré-entraîné par SQuAD à BioASQ. Dhingra et al. [136] montrent une amélioration significative des performances par rapport à l'apprentissage purement supervisé en pré-entraînant le GA-Reader [137] sur un ensemble de données BQA de type cloze à grande échelle généré automatiquement (section 6.1), puis en l'affinant sur BioASQ. Du et al. [138, 139] font des observations similaires avec l'apprentissage par transfert à partir du jeu de données SQuAD. Kang [140] montre que l'apprentissage par transfert à partir des ensembles de données « Natural Language Inference » (NLI) améliore également les performances de BioASQ pour les questions de type oui/non (+5,59 %), factoïde (+0,53 %) et liste (+13,58 %). En général, deux composantes principales sont omniprésentes dans les systèmes les plus performants des derniers défis BioASQ: 1. des modèles de langage pré-entraînés spécifiques au domaine [141], tels que BioBERT [123] ; 2. des ensembles de données QA spécifiques à la tâche qui peuvent pré-entraîner davantage les modèles utilisés, tels que SQuAD [5] pour la QA extractive et PubMedQA [79] pour les questions de type oui/non.

2.2.4.5 L’approche « Question Entailment » (QE)

Harabagiu et Hickl [142] montrent que l'implication textuelle « Question Entailment » peut être utilisée pour améliorer les systèmes QA. De nombreuses nouvelles questions peuvent trouver des questions similaires auxquelles on a déjà répondu, et ces questions peuvent être résolues par l'approche QE. L'approche QE pour la BQA est essentiellement une méthode du plus proche voisin qui utilise les réponses de questions similaires et déjà répondues (par exemple, les questions fréquemment posées, FAQ) pour répondre à la question donnée.

L’approche QE est formellement définie par Abacha et Demner-Fushman [143] comme suit : une question Q_a entraîne une question Q_b si chaque réponse à Q_b est également une réponse correcte à Q_a. L'inférence en langage naturel, en anglais « Natural language inference » (NLI) est une tâche NLP pertinente qui prédit si la relation d'implication, de contradiction ou de neutralité existe entre une paire de phrases. Dans le domaine général, la prédiction de la similarité question-question est un domaine de recherche actif avec des applications potentielles dans la recommandation de questions et la réponse aux questions communautaires [144].

Luo et al. [145] proposent le système SimQ pour retrouver des questions similaires sur la santé des consommateurs sur le web en utilisant les caractéristiques sémantiques annotées UMLS et syntaxiques AQUA-parsed [146] des questions. CMU OAQA [147] utilise une approche avec des réseaux neuronaux récurrents bidirectionnels (RNN) et un mécanisme d'attention pour prédire la similarité des questions ; Abacha et Demner-Fushman [78] utilisent un classificateur de régression logistique basé sur les caractéristiques et des modèles d'apprentissage profond qui passent la concaténation de deux « embeddings » de questions à plusieurs couches ReLU [148] pour reconnaître le QE ; Zhu et al. [149] affinent les modèles de langage pré-entraînés BERT [1] et MT-DNN [150] pour classer les paires de questions et effectuer un apprentissage de transfert à partir de NLI pour améliorer les performances.

Les composants les plus importants pour l'approche QE de BQA sont les datasets de paires question-question (Q-Q) et question-réponse (Q-A), qui sont actuellement limités en termes d'échelle ou de qualité. Pour résoudre ce problème, il convient d'explorer des méthodes de collecte automatique d'ensembles de données Q-Q et Q-A à grande échelle et de haute qualité.

2.2.4.6 L'approche VQA (Visual QA)

L'imagerie médicale est omniprésente et joue un rôle essentiel dans la prise de décision clinique pour les diagnostics et les traitements. Cependant, l'interprétation manuelle des images médicales prend beaucoup de temps et est sujette à des erreurs. Par conséquent, il est utile de répondre automatiquement aux questions des médecins en langage naturel sur les images médicales, ce qui est l'objectif de la réponse visuelle aux questions biomédicales « biomedical Visual Question Answering » (VQA). Étant donné que VQA est une nouvelle variante de la tâche QA qui se situe à l'intersection de la vision par ordinateur « Computer Vision » (CV) et du langage naturel, des méthodes multimodales qui fusionnent les techniques de langage naturel pour comprendre les questions et les techniques de CV sont nécessaires pour capturer les caractéristiques des images dans les datasets biomédicaux VQA.

En général, pour la VQA biomédicale, les images et les textes sont codés séparément sous forme de caractéristiques « features », et un mécanisme de mise en commun « pooling » multimodale est chargé d'obtenir les représentations mixtes pour générer les réponses.

Encodeurs d'images : VGGNet [151] et ResNet [152] sont couramment utilisés pour l'extraction des caractéristiques des images. Yan et al. [153] et Ren et Zhou [154] adoptent la mise en commun de la moyenne globale [155] après VGGNet pour le codage des images, ce qui permet d'éviter le sur-ajustement sur les petits datasets. Pour surmonter la limitation des données d'images, Nguyen et al. [156] appliquent le méta-apprentissage agnostique de modèle (MAML) [157] et l'auto-encodeur de débruitage convolutif (CDAE) [158] pour initialiser les couches CNN sur VQA-Rad, et obtiennent 43,9 et 75,1 de précision sur les questions ouvertes et fermées respectivement.

Encodeurs de texte : Les questions sont généralement encodées par un réseau récurrent (ex. : RNN ou LSTM) ou un modèle de langage pré-entraîné (ex. : BERT) comme les autres méthodes BQA. Le mécanisme de co-attention est utilisé pour trouver les mots et les régions importants afin d'améliorer la représentation textuelle et visuelle. Les réseaux d'attention empilés (SAN) [159] utilisent la représentation textuelle pour interroger plusieurs fois la représentation visuelle afin d'obtenir un raisonnement en plusieurs étapes.

Mise en commun multimodale : elle est cruciale pour combiner les caractéristiques des encodeurs visuels et textuels. Leur concaténation directe peut servir de base de référence. La mise en commun bilinéaire compacte multimodale (MCB) [160], la mise en commun bilinéaire factorisée multimodale (MFB) [161] et la mise en commun d'ordre élevé factorisée multimodale (MFH) [162] sont souvent utilisées pour la fusion de caractéristiques dans la VQA.

Récemment, plusieurs modèles multimodaux pré-entraînés ont été proposés qui utilisent des transformateurs [163, 164] pour générer des représentations visuelles et textuelles dans le domaine général. Ren et Zhou [165] présentent CGMVQA, qui introduit des caractéristiques VGGNet et d'intégration de mots dans un seul transformateur pour la classification ou la génération sur VQA-Med, qui atteint une précision de 64,0, et un score BLEU de 65,9.

Dans le domaine général, la méthode de pointe (SOTA) de VQA [166] effectue un pré-entraînement à partir de grands ensembles de données texte-image et effectue un réglage fin sur des ensembles de données VQA spécifiques. Les ensembles de données VQA biomédicales ont encore moins d'instances, de sorte que les méthodes de pré-entraînement multimodales devraient être explorées pour augmenter les performances VQA biomédicales.

2.3 Conclusion

Nous avons présenté dans ce chapitre l'état de l'art de la tâche de BQA. Nous avons commencé par une définition formelle de BQA, suivi par une classification des systèmes et approches proposées pour cette tâche. Ensuite, nous avons décrit ses datasets et challenges internationaux tout en comparons leurs caractéristiques. Enfin, nous avons présenté les approches et les systèmes les plus importants de la BQA qui ont été proposés au fil des années. Dans le chapitre suivant, nous allons présenter notre propre approche pour la BQA destinée aux types de questions factoïde et liste.

Chapitre III :
Nouvelle approche BQA pour les questions de type factoïde et liste

3.1 Introduction

Au fil des années, plusieurs modèles BQA ont été introduits. Cela va des systèmes traditionnels à base de composants comme [102, 95] aux modèles à base de réseaux de neurones comme [123, 124]. Les modèles BQA de pointe (SOTA) actuels sont tous basés sur l'architecture du transformateur « Transformer » [61]. Ces modèles suivent la même stratégie de pré-entraînement que BERT [1]. Mais, au lieu de texte général, ils sont pré-entraînés sur du texte biomédical. Le modèle BQA le plus connu qui est basé sur l'architecture du transformateur est BioBERT [123], qui a obtenu les résultats SOTA sur de nombreuses tâches NLP biomédicales, y compris la BQA. La couche d'auto-attention dans l'architecture du transformateur joue un rôle central dans les prédictions du modèle. Des études récentes [167, 168, 169] sur le fonctionnement interne de la couche d'auto-attention du transformateur supposent que les jetons avec des scores d'attention plus élevés ont une plus grande probabilité de faire partie de la réponse prédite. Cette corrélation, si elle est prouvée, peut être exploitée pour améliorer les performances du modèle. Cette promesse a motivé notre travail et la méthode que nous proposons.

Dans ce chapitre, nous présentons une nouvelle approche pour la BQA, ciblant spécifiquement les questions de type factoïde et liste. L'idée principale de notre méthode est d'enrichir la couche d'auto-attention du transformateur avec des informations biomédicales et d'entités nommées préalablement extraites de la question et du passage contextuel. Nous l'appelons mécanisme d'enrichissement de l'auto-attention, et nous l'appliquons à BioBERT [123], mais il peut être utilisé avec n'importe quel autre modèle basé sur le transformateur [61]. Tout d'abord, nous marquons toutes les entités biomédicales, les associations et les entités nommées dans la question et le passage contextuel. Ensuite, nous augmentons les scores d'attention entre les entités étiquetées et l'adjectif interrogatif de la question en se basant sur l'hypothèse présentée précédemment sur la corrélation entre les scores d'attention et l'étendue de la réponse prédite. Notre méthode proposée a donné des résultats de pointe (SOTA) sur plusieurs lots des datasets 10b, 9b, 8b et 7b de BioASQ [4].

Le reste du chapitre est organisé comme suit : nous présentons tout d'abord les principaux modèles et méthodes BQA introduits précédemment qui sont étroitement liés à notre approche. Dans la première section, nous décrivons le travail que nous avons effectué en termes de marquage d'entités biomédicales et d'identification de relations. Nous commençons par énumérer les ontologies et les référentiels biomédicaux que nous avons utilisés dans l'outil NER que nous proposons. Nous décrivons ensuite la stratégie de marquage d'entités biomédicales et NER que nous avons adoptée. Dans la deuxième section, nous décrivons notre méthode. Nous commençons par définir formellement la tâche de BQA pour les questions de type factoïde et liste. Ensuite, nous donnons un aperçu et discutons des études importantes sur le fonctionnement interne de l'auto-attention de BERT/BioBERT. Enfin, nous décrivons la méthode que nous proposons. Dans la troisième section, nous présentons et discutons les résultats des expériences que nous avons réalisées sur le jeu de données BioASQ. Enfin, nous terminons le chapitre par une conclusion.

3.1.1 L'architecture du transformer

Le Transformer [61] est un modèle de séquence à séquence (seq-2-seq) qui se compose d'un encodeur et d'un décodeur, chacun étant une pile de *L* blocs identiques. Chaque bloc d'encodage est principalement composé d'un module d'auto-attention à têtes multiples et d'un réseau à action directe « Feed-Forward Network » (FFN). Pour construire un modèle plus profond, une connexion résiduelle [152] est utilisée autour de chaque module, suivie d'un module de normalisation des couches [170]. Par rapport aux blocs encodeurs, les blocs décodeurs insèrent un deuxième module d'attention croisée entre les modules d'auto-attention multi-têtes et les FFN de position. L'architecture globale du transformer est illustrée à la Figure 12.

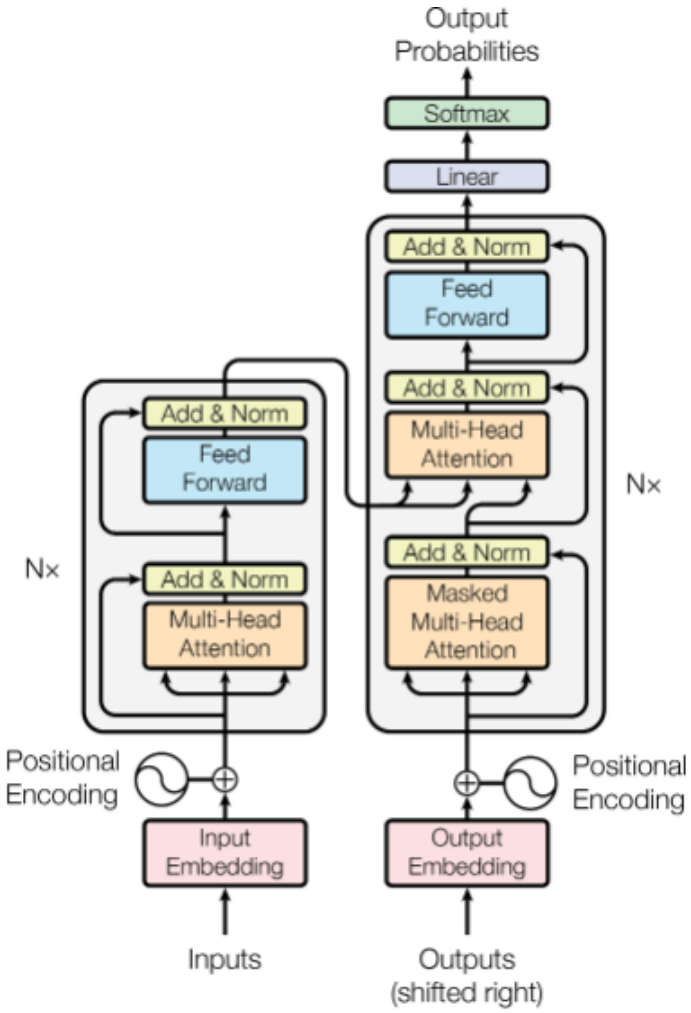

Figure 12 L'architecture globale du Transformer

3.1.2 Les modèles de langues pré-entrainés « Pre-trained Language Models » (PLM)

Les PLMs sont des modèles de langage qui ont été entraînés avec de grands ensembles de données tout en restant agnostiques par rapport aux tâches spécifiques pour lesquelles ils seront employés (Voir la Figure 13). En pratique, pour tirer profit des PLM, les dernières couches de sortie doivent être adaptées à la tâche : cette étape est appelée "réglage fin" « Fine-tuning ».

OpenAI GPT [2], BERT [1], et XLNet [63] sont des exemples de modèles pré-entraînés qui peuvent être adaptés à diverses tâches NLP. Les PLM ont fait l'objet d'une grande attention après que BERT ait obtenu des résultats de pointe sur 11 tâches NLP. Des variantes du modèle

BERT et d'autres PLM peuvent être trouvées dans des référentiels en ligne[3]. Actuellement, des milliers de modèles ont déjà été mis à la disposition de la communauté. Ces modèles peuvent être classés dans les catégories suivantes : a) adaptation à une tâche spécifique et/ou à un domaine spécifique, ou b) optimisation, où l'objectif est d'améliorer le cœur du modèle ou de réduire son coût de calcul.

Alors que BERT atteint d'excellentes performances dans plusieurs sous-tâches NLP, plusieurs chercheurs se sont concentrés sur la création de PLMs spécifiquement adaptés au contexte d'un domaine spécifique donné, généralement en affinant ou en réentraînant complètement BERT sur un autre corpus. Ainsi, des approches ont été proposées pour le langage biomédical [123], les articles scientifiques [124], les notes cliniques [171, 172], entre autres.

Parallèlement, d'autres ont expérimenté avec l'adaptation des PLMs à des tâches qui n'ont pas été évaluées à l'origine par les auteurs de BERT. Ces processus n'impliquent généralement qu'un réglage fin de BERT, ce qui est beaucoup moins coûteux en termes de calcul que le pré-entraînement de BERT sur un autre corpus. Adhikari et al. [173] proposent par exemple de régler finement BERT pour obtenir un modèle capable de classer un document complet, tandis que Lee et Hsiang [174] s'attaquent au problème de la classification des brevets.

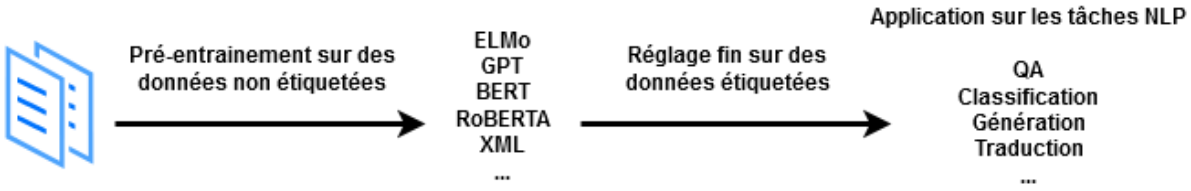

Figure 13 Le mode standard du pré-entrainement et réglage fin des PLMs

3.1.3 Le PLM BioBERT

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) [123] est un modèle de langue pré-entraîné pour le domaine biomédical. Le processus

[3] https://huggingface.co/models

global de pré-entraînement et de réglage fin de BioBERT est illustré à la Figure 14. Tout d'abord, BioBERT est initialisé avec les poids de BERT [1], qui a été pré-entraîné sur des corpus de domaine général (Wikipedia en anglais et BooksCorpus). Ensuite, il est pré-entraîné sur des corpus du domaine biomédical (résumés PubMed et articles PMC en texte intégral). Avec presque la même architecture pour toutes les tâches NLP, BioBERT surpasse largement BERT et les modèles de pointe précédents dans une variété de tâches de fouille de textes biomédicaux lorsqu'il est pré-entraîné sur des corpus biomédicaux. Alors que BERT obtient des performances comparables à celles des modèles de pointe précédents, BioBERT les surpasse de manière significative dans les trois tâches suivantes: reconnaissance d'entités nommées biomédicales « Biomedical Named Entity Necognition » (BioNER) (amélioration du score F1 de 0,62 %), extraction de relations biomédicales « biomedical relation extraction » (amélioration du score F1 de 2,80 %) et BQA (amélioration du MRR de 12,24 %).

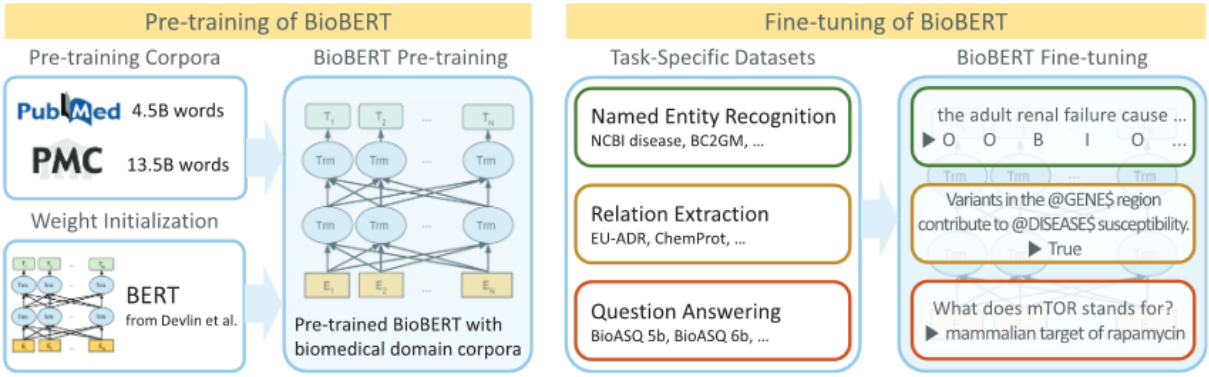

Figure 14 Aperçu du pré-entraînement et de réglage fin de BioBERT

3.1.4 Analyse du mécanisme d'auto-attention de BERT dans le contexte QA

Quatre ans après la sortie de BERT, son fonctionnement interne est toujours étudié sans être totalement compris. Il existe même un domaine d'étude spécial appelé BERTologie, qui s'intéresse à l'explicabilité des prédictions de BERT et au fonctionnement global du modèle. De nombreuses études [169, 167, 168] se concentrent également sur l'explicabilité des prédictions de BERT dans le cas de la tâche de QA. En particulier, la relation entre les poids d'auto-attention de BERT et la prédiction finale. Dans [167], les auteurs ont mené une série d'expériences analytiques pour examiner les relations entre l'auto-attention des têtes multiples de BERT et la performance finale du modèle dans le cas du QA. Ils ont constaté que les poids d'attentions

« Attention » de passage à la question et de compréhension du passage montrent une forte corrélation avec la performance finale du modèle. En particulier, ils ont constaté que les jetons de l'étendue de la réponse ont dans de nombreux cas des valeurs d'attention plus élevées. Ils ont également observé une forte attention sur les adjectifs interrogatifs de la question, comme "pourquoi" ou "où", vers les mots de l'espace de réponse. Un exemple de cette dernière observation est présenté dans la Figure 15.

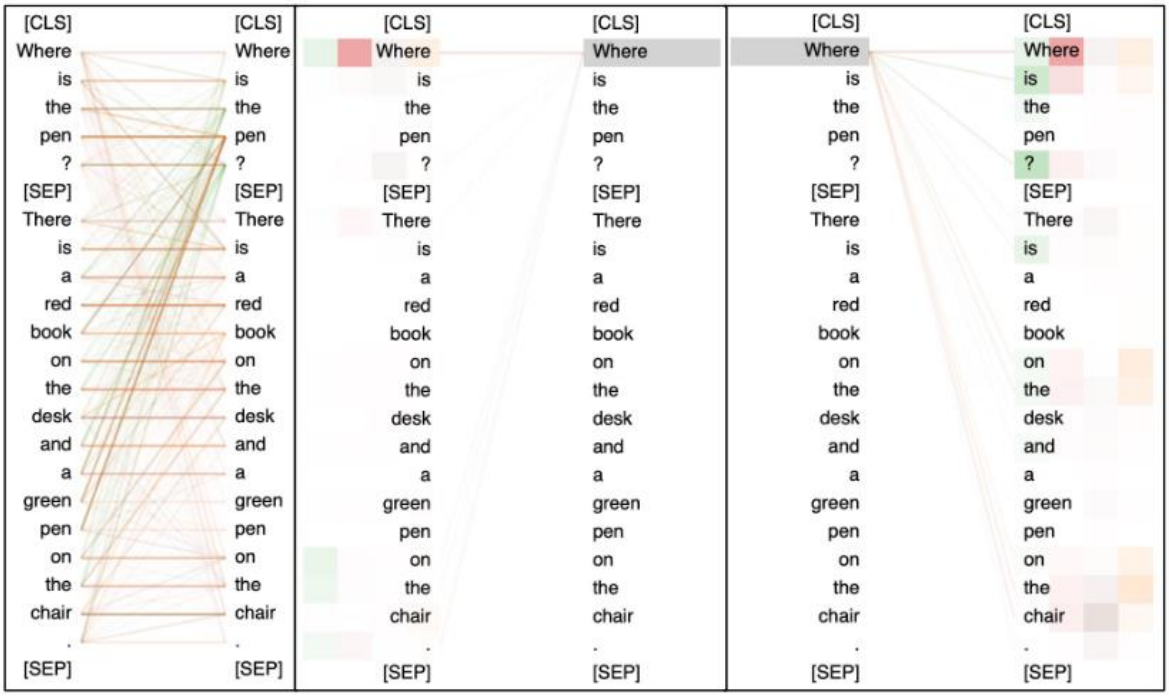

Figure 15 Visualisation des têtes spécifiques dans la couche 12 montrant les valeurs de chaque tête d'attention

Dans [169], les auteurs ont effectué une analyse par couche des états cachés de BERT, également spécifique au QA. Ils ont constaté, parmi de nombreuses autres observations, que les premières couches de BERT sont plus faibles que les autres couches en termes de syntaxe et de sémantique.

Nous basons notre approche sur les deux dernières études mentionnées [167, 169]. Nous supposons que l'enrichissement des couches inférieures de BioBERT [123] avec des informations sémantiques peut améliorer la performance globale du modèle. En particulier, en augmentant les valeurs d'attention entre l'adjectif interrogatif de la question et les entités biomédicales et NER dans le passage contextuel, qui ont en général une plus grande probabilité

d'être la réponse à des questions biomédicales de type factoïde ou liste. Nous expliquons notre approche avec plus de détails dans le reste du chapitre.

3.2 État de l'art : approches d'enrichissement des PLM

Plusieurs méthodes et modèles QA et BQA ont été proposés au fil des années. Nous ne présenterons ici que ceux qui sont liés à la méthode que nous proposons. En plus des méthodes BQA, nous présenterons également les méthodes QA et les méthodes générales de l'NLP liées à notre approche. Les méthodes qui proposent d'enrichir les modèles de langage pré-entraînés (PLM) sous quelque forme que ce soit. Afin d'augmenter leurs capacités ou leurs performances.

Enrichissement des PLM pendant la phase de pré-entraînement : [106] ont introduit une méthode d'apprentissage conjointe pour ajouter des informations de base de connaissances provenant du système de langage médical unifié (UMLS) dans le pré-entraînement des modèles de langue pour les corpus du domaine clinique. Deux modèles pré-entraînés ont été créés par les auteurs, Clinical KB-BERT et Clinical KB-ALBERT, basés respectivement sur BERT [1] et ALBERT [62]. Les modèles obtenus se sont avérés plus efficaces que leurs homologues originaux. En utilisant également les connaissances de l'UMLS, deux autres modèles ont été introduits, KeBioLM [175] et UmlsBERT [176]. Ils exploitent tous deux les connaissances des bases de connaissances UMLS pendant le pré-entraînement. Similaire à ce travail, mais spécifique aux informations relatives aux maladies. [177] ont proposé une nouvelle procédure d'entraînement par l'infusion de connaissances sur les maladies pour les PLMs de type BERT afin d'enrichir ces modèles avec des connaissances sur les maladies. Les expériences réalisées par les auteurs montrent que ces modèles peuvent être améliorés par cette méthode d'infusion de maladies dans presque tous les cas.

Intégration de bases de connaissances (KB) dans les PLM : [178] ont présenté KnowBert, un nouveau modèle qui intègre des KB dans BERT, en améliorant les représentations textuelles d'origine avec des connaissances structurées et élaborées par l'homme à partir de plusieurs KB. Le nouveau modèle a permis d'améliorer la perplexité, la capacité à rappeler les faits et l'amélioration globale de l'extraction des relations, du typage des entités et de la désambiguïsation du sens des mots. Pour la tâche d'inférence en langage naturel (NLI), Gajbhiye et al. [179] ont introduit un nouveau modèle qu'ils ont appelé External Knowledge

Enhanced BERT (ExBERT). Le modèle proposé enrichit la représentation contextuelle de BERT avec des connaissances de bon sens du monde réel provenant de sources de connaissances externes comme ConceptNet [180] et améliore les capacités de compréhension du langage et de raisonnement de BERT. ExBERT a obtenu des résultats SOTA sur les benchmarks SciTail [181] et SNLI [182].

Méthodes appliquées pendant l'auto-attention : [183] ont proposé une méthode d'attention locale consciente de la syntaxe qui peut être intégrée aux PLM, comme BERT [1], pour que le modèle se concentre sur les mots syntaxiquement pertinents. Les expériences menées par les auteurs ont montré qu'une attention plus ciblée sur les mots syntaxiquement pertinents permettait d'obtenir des gains constants par rapport à BERT sur tous les datasets de référence utilisés dans les expériences. Pour les tâches de correspondance textuelle sémantique, [184] ont injecté des connaissances préalables dans le mécanisme d'attention [61] multi-têtes de BERT afin d'améliorer ses performances. Pour ce faire, ils ont construit une matrice de similarité des mots qui est utilisée pendant la phase d'attention multi-têtes pour augmenter l'attention de BERT sur les paires de mots sémantiquement similaires. Les expériences menées par les auteurs ont démontré que le modèle de type BERT proposé et qui est amélioré par la connaissance est capable d'améliorer de manière constante les performances de correspondance sémantique textuelle par rapport au modèle BERT original. Un autre modèle basé sur les transformateurs « Transformer » appelé TOME, présenté par [185], effectue l'attention sur une représentation semi-paramétrique de l'ensemble du corpus textuel de Wikipedia. Le modèle peut récupérer des informations de plusieurs sources sans supervision, agréger les informations dans le transformateur et raisonner sur les informations récupérées. Le modèle proposé a produit de fortes améliorations sur de multiples tâches de vérification de réclamations dans des domaines ouverts et de réponse à des questions basées sur des entités.

Fusionner les représentations originales des PLMs avec des caractéristiques externes : [186] a extrait des caractéristiques externes, telles que les parties de la parole (POS) et les entités nommées, et les a fusionnées avec la représentation textuelle originale codée par BioBERT. La méthode proposée a permis d'améliorer globalement les performances sur trois métriques de la tâche QA pour les questions de type factoïde sur les datasets BioASQ [4] 6b, 7b et 8b. Utilisant

également BioBERT [123], [187] ont proposé un mécanisme pour l'ajuster avec un dataset d'entités nommées afin d'améliorer ses performances en QA. Le cadre proposé a obtenu des performances prometteuses dans les datasets BioASQ 6b et 7b. Dans [188], les auteurs ont contextualisé conjointement le passage de la question et du contexte avec des connaissances provenant de Wikipedia sur les entités présentes dans le passage de contexte et la question. La méthode proposée est basée sur RoBERTa [64] et a permis d'obtenir des améliorations sur plusieurs datasets QA, y compris BioASQ pour la BQA. Pour la tâche de détection de discours haineux, [189] a proposé d'incorporer des caractéristiques lexicales extraites d'un lexique de discours haineux avec BERT pour la détection de langage abusif. De la même manière, mais pour les tâches d'étiquetage de séquences, [190] ont proposé DyLex, une méthode incorporant un lexique dynamique pour améliorer les performances des modèles de type BERT. Le modèle proposé a obtenu des résultats SOTA dans de nombreuses tâches d'étiquetage de séquences.

3.3 Notre approche BQA pour les questions de type factoïde et liste

3.3.1 Identification et marquage des entités et relations biomédicales

3.3.1.1 Référentiels et ontologies utilisés

Puisque notre approche repose sur le marquage des entités biomédicales et l'identification des relations, nous avons commencé par construire notre propre outil de reconnaissance des entités nommées biomédicales (BioNER). Les outils existants tels que MetaMap [74] et scispaCy [191] ne peuvent que marquer les entités biomédicales sans être en mesure d'identifier les relations entre elles (par exemple, les associations gène-maladie ou maladie-médicament), d'où la nécessité de construire notre propre outil BioNER. La première étape de ce processus a été de construire une collection de données sur les maladies, les gènes, les médicaments, les protéines et les enzymes. Ainsi que des associations gène-maladie et médicament-maladie. À cette fin, nous nous sommes appuyés sur une variété d'ontologies et d'ensembles de données biomédicales. Le tableau 7 présente la liste de toutes ces sources. Après avoir téléchargé les ensembles de données (généralement au format XML ou CSV), nous avons extrait uniquement les informations les plus importantes. Le tableau 8 montre les informations extraites par type d'entité biomédicale.

Référentiel/ontologie	Type d'entité	Nombre d'instances
Disgenet [192]	Associations gènes-maladies	1 134 942
Uniprot [193]	Protéines	6 127
Omim [194]	Gènes	26 453
DISEASES [195]	Associations gènes-maladies	543 405
Ctdbase [196]	Associations gènes-maladies	50 599 050
Ncbi gene dataset [197]	Gènes	-
DrugBank [198]	Médicaments et associations maladie-médicaments	500 000
Drugcentral [199]	Médicaments	4714
Brenda [200]	Enzymes	8 282

Table 7 Référentiels et ontologies biomédicaux utilisés dans la construction de notre outil BioNER

Type d'entité	Informations extraites
Maladie	Nom, synonymes, associations de gènes
Gène	Nom, synonymes, symbole, association de maladies, indication, fabricants
Médicament	Nom, synonymes
Protéine	Nom
Enzyme	Nom

Table 8 Informations extraites par type d'entité biomédicale

En utilisant les informations biomédicales extraites, nous avons pu construire une collection unifiée de noms de maladies, de gènes, de médicaments, de protéines et d'enzymes, de symboles, de synonymes et d'associations que nous avons ensuite utilisés pour annoter les entités et associations biomédicales dans le passage contextuel de l'ensemble de données QA utilisé.

3.3.1.2 Marquage d'entités biomédicales

À l'aide de notre outil BioNER, nous avons d'abord annoté tous les noms d'entités biomédicales, les synonymes et les symboles dans le passage contextuel des instances d’entraînement et de

test du dataset BioASQ [4], qui est le jeu de données que nous avons utilisé dans nos expériences. Pour les associations d'entités, nous n'avons considéré que les entités biomédicales dans le passage contextuel qui sont liées à d'autres entités dans la question. Comme une maladie dans le passage du contexte qui est liée à un gène dans la question. Pour chaque entité rencontrée, nous avons déterminé les positions de début et de fin dans le texte. Nous avons ensuite construit un vecteur pour chaque passage contextuel où chaque entité biomédicale correspondante possède un identifiant numérique unique, allant de un pour les maladies à cinq pour les enzymes. Les jetons non biomédicaux ont été mis à zéro. En plus des entités biomédicales, nous avons également annoté les entités NER ordinaires comme les dates, les organisations, les pourcentages, les quantités et les valeurs numériques. Pour ce faire, nous avons utilisé le module NER de Spacy[4]. En fait, un bon nombre de réponses aux questions de BQA sont liées aux dates de sortie des médicaments et des entreprises pharmaceutiques, ou sont de type pourcentage, quantité ou valeurs numériques. Mais nous n'avons appliqué l'annotation NER régulière que lorsque la question contient l'un des mots suivants : "quand", "quelle entreprise", "quel pourcentage", et "combien". Des identificateurs numériques uniques ont également été attribués aux entités NER régulières dans le vecteur construit pour chaque passage contextuel.

[4] https://spacy.io

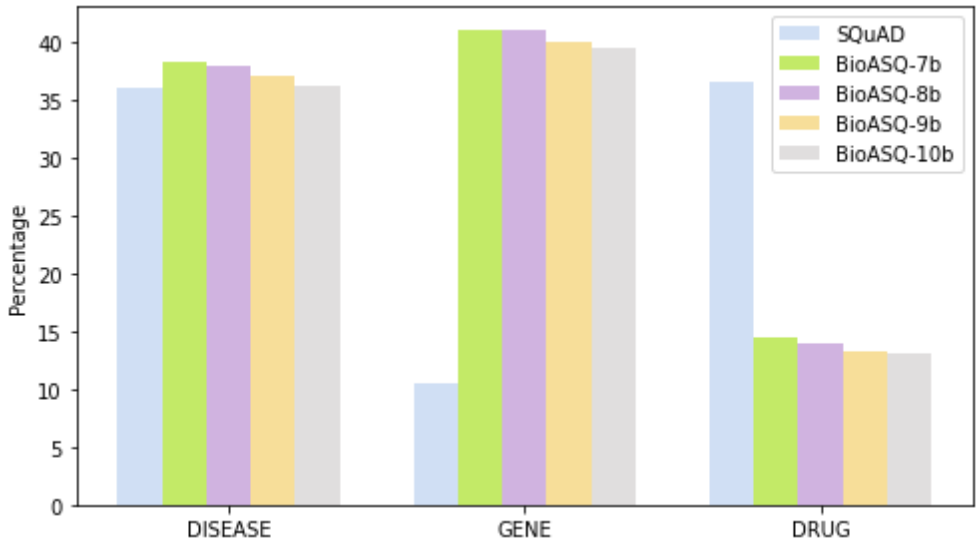

Figure 16 Pourcentage d'entités de types maladie, gène, et médicament par rapport à l'ensemble des entités biomédicales étiquetées dans les datasets utilisés

La Figure 16 montre la distribution des entités de types maladie, gène, et médicament par rapport à toutes les entités biomédicales étiquetées dans les cinq datasets utilisés. Comme on peut le voir sur la figure, bien que SQuAD soit considéré comme un dataset QA du domaine général. Il contient un pourcentage considérable d'entités de types maladie et médicament. Il surpasse même les quatre datasets BioASQ utilisés pour le pourcentage d'entités de type médicament. La distribution statistique des maladies, des gènes et des médicaments est presque la même dans les quatre jeux de données BioASQ. Les autres entités biomédicales comme les protéines et les enzymes sont négligeables par rapport aux entités maladies, gènes et médicaments dans tous les jeux de données.

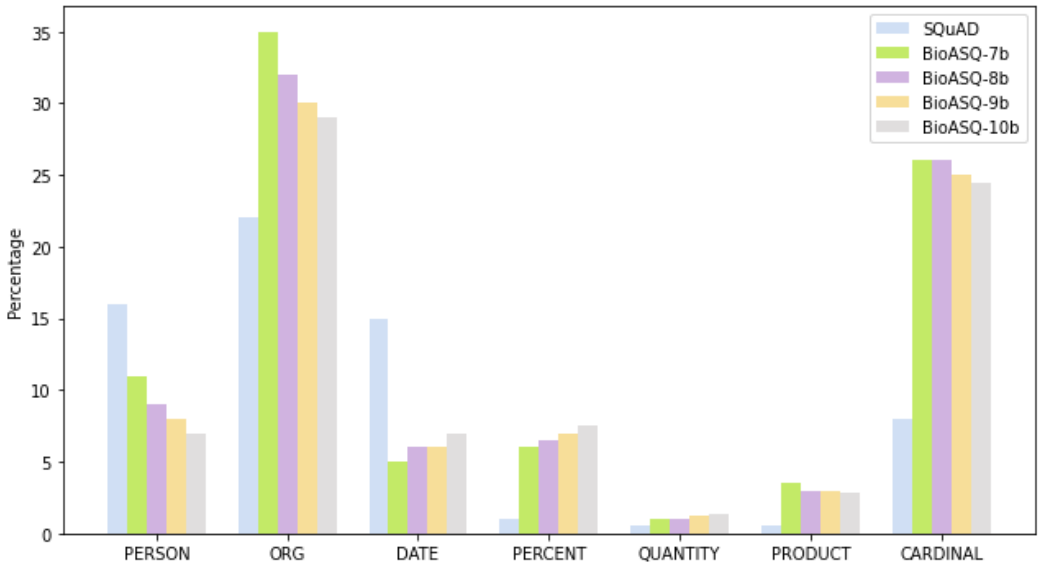

Figure 17 Distribution des entités de type NER dans les datasets utilisés

La Figure 17 montre la distribution des entités de type NER dans les cinq ensembles de données. Les pourcentages sont similaires dans les différents jeux de données BioASQ. SQuAD contient également un pourcentage considérable d'entités PERSON, ORG et DATE.

La Figure 18 montre un exemple de marquage biomédical du passage de contexte « Emapalumab in Children with Primary Hemophagocytic Lymphohistiocytosis. » tiré du dataset BioASQ 10b. Le passage est découpé « tokenization » avec le WordPiece tokenizer de BioBERT. Le chiffre 3 désigne les médicaments, le chiffre 1 désigne les maladies. Nous avons choisi les chiffres arbitrairement, car le but est juste de pouvoir identifier les entités biomédicales et leurs types (maladie, médicament, gène,…etc.). Le vecteur résultant est exploité par la suite dans le calcul des scores d'attention dans les blocs de transformers comme illustré dans la Figure 19

Emapalumab in Children with Primary Hemophagocytic Lymphohistiocytosis.																				
Em	apa	lum	ab	in	Children	with	Primary	He	mo	pha	go	cy	tic	L	ymph	oh	ist	io	cy	tosis	.
3	3	3	3	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0

Figure 18 Exemple de marquage biomédical d'un passage de contexte d'une instance du dataset BioASQ 10b

3.3.2 Définition de la tâche

Comme pour le QA, la tâche BQA peut être divisée en deux catégories : extractive et générative. Dans la BQA extractive, qui est au centre de notre approche, on donne une question $Q = \{q_1, \dots, q_n\}$ où q_i représente les jetons de la question de longueur n, et un passage contextuel $P = \{p_1, \dots, p_m\}$ où p_i représente les jetons de passage contextuel de longueur m. La tâche consiste alors à prédire l'étendue de la réponse $A = \{p_i\}_s^e$ des jetons continus p_i pour la question Q à l'intérieur du passage contextuel P, où s est l'indice du premier jeton de l'étendue de la réponse, et e est l'indice du dernier jeton. Le QA extractive peut également être divisée en deux types : les questions de type factoïde et les questions de type liste. Pour les questions factoïdes, un seul intervalle de réponse doit être prédit, car la question peut être répondue par une seule phrase. Pour les questions de type liste, des réponses multiples doivent être prédites, car la question ne peut être entièrement satisfaite qu'en renvoyant plusieurs phrases (par exemple : Quels sont les symptômes du COVID-19 ?). Notre approche peut être appliquée aux questions de type factoïde et liste.

Comme mentionné précédemment, le modèle que nous proposons est basé sur BioBERT [123], qui est une variante de BERT [1]. La tâche de QA est implémentée dans BERT et donc dans BioBERT comme suit. En donnant la question originale Q et le passage contextuel P, l'entrée du modèle BERT/BioBERT est $X = \{[CLS] \,||\, Q' \,||\, [SEP] \,||\, P' \,||\, [SEP]\}$ où Q' et P' sont la question Q et le passage contextuel P tokenizés par le WordPiece tokenizer de BioBERT. [CLS] est un jeton BERT prédéfini utilisé dans les tâches de classification. Le jeton BERT prédéfini [SEP] est utilisé comme séparateur entre les entrées du modèle. La question et le passage contextuel tokenisés, ainsi que les jetons spéciaux de BERT, sont ensuite concaténés pour former une seule entrée. Le vecteur de représentation caché du $i^{ième}$ jeton d'entrée est noté $h_i \in \mathbb{R}^H$ où H est la taille cachée. Nous désignons par les paramètres entraînables $S \in \mathbb{R}^H$ et $E \in \mathbb{R}^H$ les vecteurs de début et de fin de l'étendue de la réponse. La probabilité que les $i^{ième}$ et $j^{ième}$ jetons soient respectivement les jetons de début et de fin est calculée comme suit :

$$p_i^{start} = \frac{e^{S \cdot t_i}}{\sum_{k=1}^{m} e^{S \cdot t_k}}, \qquad p_j^{end} = \frac{e^{E \cdot t_j}}{\sum_{k=1}^{m} e^{E \cdot t_k}} \tag{3.1}$$

Où $t_i \in \mathbb{R}^H$ et $t_j \in \mathbb{R}^H$ désignent la $i^{ième}$ et la $j^{ième}$ représentation du jeton de la couche finale de BioBERT, m est la taille de la séquence d'entrée (en jetons), et · désigne le produit scalaire. La perte « loss » est définie comme la moyenne arithmétique de $Loss_{start}$ et $Loss_{end}$, la log-vraisemblance « log-likelihood » négative pour les jetons de début et de fin corrects respectivement. Les trois pertes sont définies comme suit :

$$Loss_{start} = -\frac{1}{N}\sum_{k=1}^{N} \log P_{y_s}^{start,k}, \quad Loss_{end} = -\frac{1}{N}\sum_{k=1}^{N} \log P_{y_e}^{end,k} \tag{3.2}$$

$$Loss = (Loss_{start} + Loss_{end})/2 \tag{3.3}$$

Où y_s et y_e désignent les positions correctes des jetons de début et de fin respectivement, et N est la taille du lot.

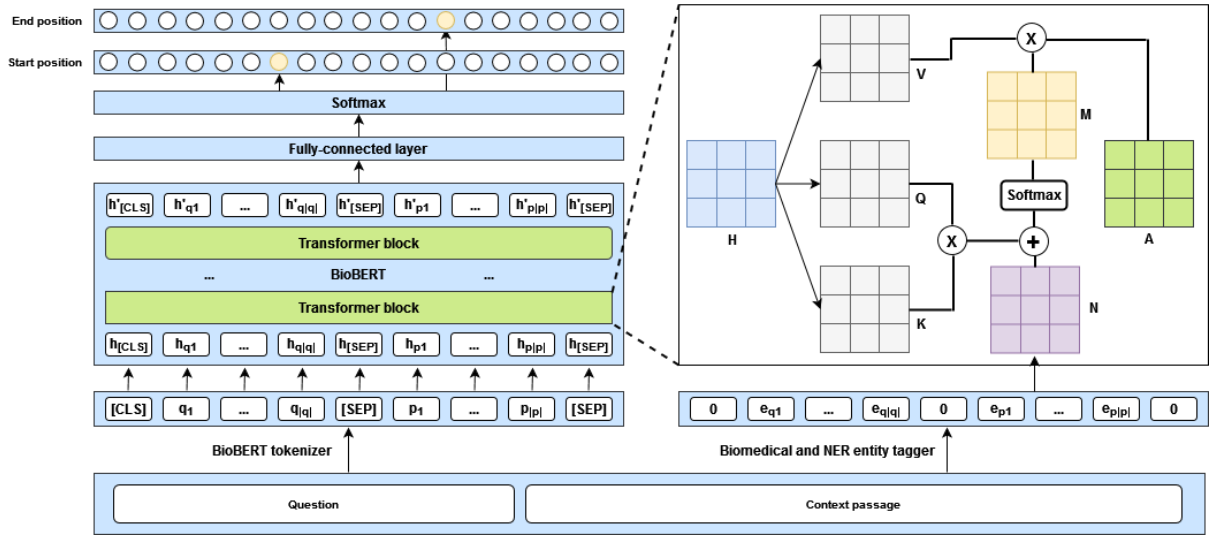

Figure 19 Architecture globale de notre approche BQA pour les questions de type factoïde et liste

3.3.3 Notre nouveau mécanisme d'enrichissement de l'auto-attention

Comme nous l'avons dit précédemment, notre méthode est spécifique aux questions biomédicales de type factoïde et liste. Il s'agit de questions qui nécessitent comme réponse, un nom d'entité biomédicale particulier (ex : maladie, médicament,etc), un nombre ou une expression courte similaire.

Dans notre méthode proposée, nous avons utilisé BioBERT-large finement réglé sur SQuAD [5] que nous avons étendu avec notre nouveau mécanisme d'enrichissement de l'auto-attention. BioBERT [123] est également un modèle de langage pré-entraîné comme BERT, il est également basé sur l'architecture de transformateur [61]. Qui est composé principalement d'un ensemble de couches d'auto-attention multi-têtes empilées avec plusieurs couches denses. Donnant une séquence de vecteurs d'entrée $H = [h_1, h_2, \dots, h_{|H|}]$, avec $H \in \mathbb{R}^{n \times d}$ comme matrice de représentation cachée, où n est la longueur maximale de la séquence d'entrée, et d la taille de la dimension cachée. Tout d'abord, H est transformée en représentation de requête, de clé et de valeur en utilisant trois couches denses.

$$Q = W^Q H, \qquad K = W^K H, \qquad V = W^V H \tag{3.4}$$

Le produit scalaire du vecteur de requête Q et du vecteur clé K est ensuite calculé, suivi d'une normalisation softmax pour obtenir la matrice d'attention $M \in \mathbb{R}^{n \times n}$.

$$M = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{3.5}$$

Où d_k est la dimension de chaque tête d'attention. Enfin, le produit scalaire de la matrice d'attention M et de la représentation du vecteur de valeur V est calculé pour former la représentation finale de l'auto-attention $A = M \cdot V$

Afin de valider notre hypothèse mentionnée dans la partie précédente de cette section. Que l'enrichissement des couches inférieures de BioBERT avec des informations sémantiques peut augmenter la performance globale du modèle. Nous avons utilisé les entités biomédicales et NER étiquetées et les associations pour construire une matrice $N \in \mathbb{R}^{n \times n}$ à partir du passage de contexte et de la question. Où nous avons donné le chiffre 1 à chaque entité biomédicale et NER étiquetée, et zéro pour les autres jetons. La matrice résultante N est ensuite ajoutée au produit scalaire de la représentation de la requête et du vecteur clé lors du calcul de la matrice d'auto-attention M

$$M = softmax\left(\frac{QK^T + N}{\sqrt{d_k}}\right) \tag{3.6}$$

Nous n'avons appliqué cette approche que pour les deux dernières couches d'auto-attention de BioBERT. L'intuition derrière l'ajout de la représentation de la matrice biomédicale et NER N pendant le calcul de la matrice d'auto-attention M est, comme mentionné dans notre hypothèse, d'augmenter les valeurs d'attention entre l'adjectif interrogatif de la question et les jetons biomédicaux et NER dans le passage contextuel. La Figure 20 montre les poids d'attention de la tête numéro 9 de la première couche pour une instance (Question : Which disease is Dasatinib used to treat? Passage contextuel : Patients with chronic myeloid leukemia) du dataset BioASQ utilisé avec et sans l'ajout de la représentation matricielle biomédicale et NER N. Comme on peut le voir sur la figure, l'enrichissement de l'attention biomédicale et NER a entraîné des valeurs d'attention plus fortes entre le mot « Which » de la question et les entités biomédicales dans le passage contextuel « myeloid » et « leukemia ». Les figures 45 et 46 en annexe montrent les mêmes informations pour toutes les têtes d'attention dans les deux couches inférieures. L'architecture globale de la méthode est présentée dans la Figure 19. Nous présentons et discutons nos résultats dans le reste du chapitre.

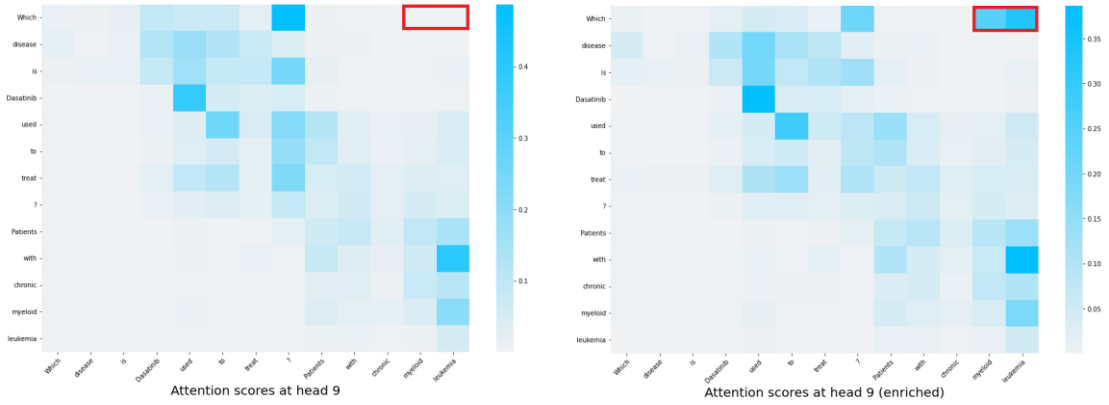

Figure 20 Scores d'attention de la tête N° 9 de la première couche avant et après l'enrichissement de l'attention avec les entités biomédicales et NER

Le tableau 9 ci-dessous énumère les détails techniques d'implémentation. La taille du modèle désigne le nombre de paramètres entraînables par le modèle. La longueur maximale de la séquence désigne la taille maximale de la question + le contexte de passage en jetons. Plus la longueur est grande plus il y a la possibilité de prendre en compte de longs contextes de passages. L'entrée question + contexte de passage qui dépasse 300 tokens va être tronquée. La valeur de 300 est la longueur maximale permise par notre GPU. La taille du lot « batch size » est aussi choisi par rapport à nos ressources matérielles. Le taux d'apprentissage « learning

rate » de 9e-6 est une bonne valeur de départ pour les modèles de type BERT selon plusieurs expérimentations. Le nombre d'époques d'entraînement « train epochs » de 3 est recommandé par les auteurs de BERT [1]

Paramètre	Valeur
Modèle de base	BioBERT
Taille du modèle	Large
Longueur maximale de la séquence	300
Taille du lot « Batch size »	8
Taux d'apprentissage « Learning rate »	9e-6
Époques d'entraînement « Train epochs »	3
GPU	NVIDIA RTX 6000 24gb
Framework Deep learning	Tensorflow 1.14.0

Table 9 Détails techniques d'implémentation de notre approche BQA pour les questions de type factoïde et liste

3.4 Evaluation et discussion

Afin de valider davantage notre approche, nous avons expérimenté avec le dataset BioASQ [4], qui est actuellement le plus grand dataset BQA annoté par des experts. C'est actuellement le benchmark de référence dans le domaine du BQA. Cette année marque la 10éme édition de la compétition de BioASQ. La Figure 21 montre les équipes participantes à cette 10éme édition. Notre équipe « LaRSA » représentée par le logo de l'université Mohammed Premier, est la seule équipe participante d'Afrique et du monde arabe. À chaque édition, les organisateurs de la compétition publient cinq jeux de tests. Sauf pour la 10éme édition où un sixième lot de test est ajouté. Chaque lot contient des questions de type factoïde, liste, oui/non et résumé. Les systèmes participant au défi sont évalués par rapport à chaque lot, et pour chaque type de question supporté. Chaque système doit renvoyer une liste de cinq réponses classées par ordre de confiance décroissant.

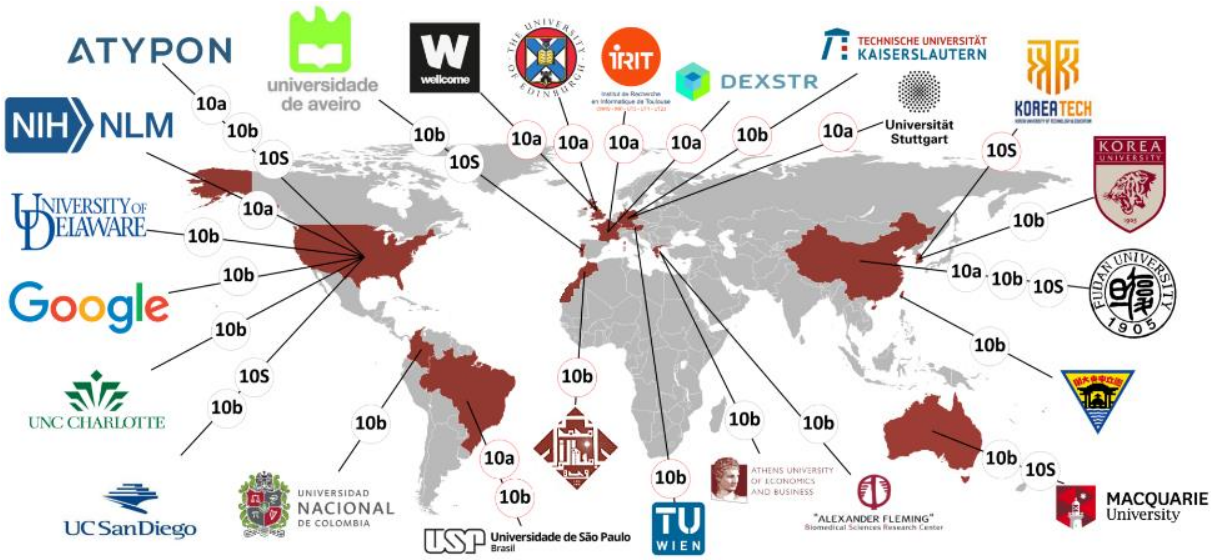

Figure 21 Equipes participantes dans la dixième édition (2022) du challenge BioASQ

Dans le tableau 14, nous comparons les performances de notre approche avec le modèle de base BioBERT-large, et avec le même modèle après réglage fin sur le jeu de données SQuAD [5]. Cette comparaison a été effectuée sur le premier lot de la 9ème édition du dataset BioASQ. Comme on peut le voir dans le tableau, l'apprentissage par transfert avec SQuAD a considérablement amélioré les résultats sur les trois métriques. En appliquant notre approche au modèle BioBERT-large + SQuAD, nous avons obtenu des gains encore plus considérables. En particulier sur la métrique de précision stricte, où une amélioration de 6,9 points a été enregistrée. Des résultats similaires ont été obtenus sur d'autres lots et dans les autres jeux de données 10b, 8b et 7b.

Modèle	Strict Acc.	Lenient Acc.	MRR
BioBERT-large	0.4137	0.5517	0.4626
BioBERT-large + SQuAD	0.4482	0.5862	0.5000
BioBERT-large + SQuAD + notre approche	**0.5172**	**0.5862**	**0.5345**

Table 10 Comparaison entre notre approche et le modèle de base BioBERT sur le premier lot de la 9ème édition du jeu de données BioASQ.

La Figure 22 montre la précision stricte du modèle par époque dans les phases d'entraînement et de test.

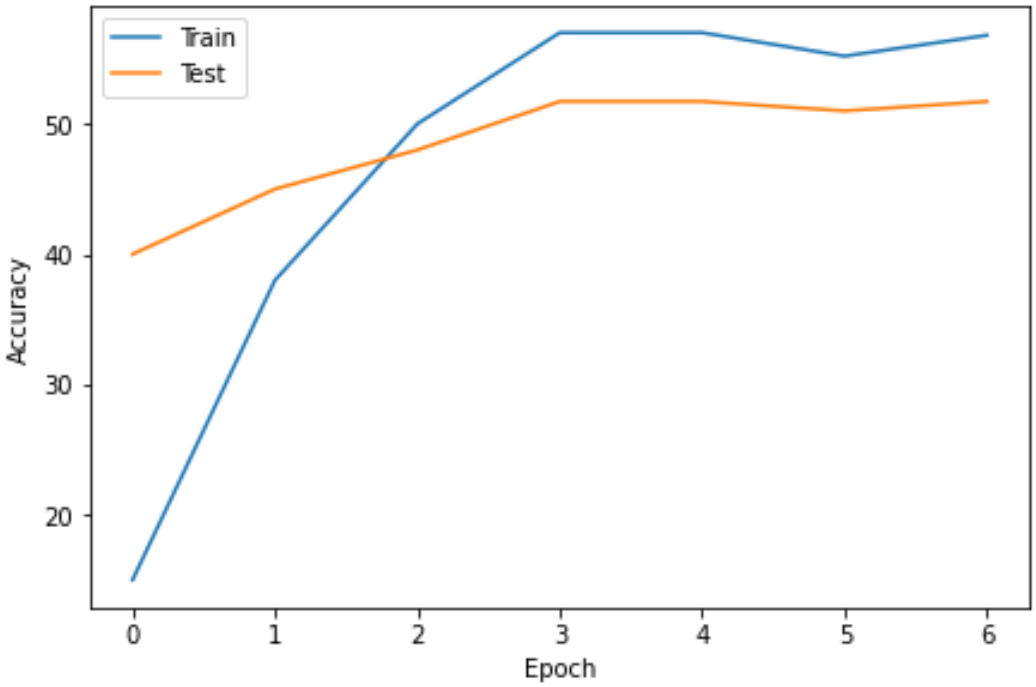

Figure 22 Précision du modèle par époque pour les phases d'entrainement et de test

Nous présentons ci-dessous nos résultats dans BioASQ 10b, auquel nous avons participé avec notre système. Afin de confirmer les performances de notre approche, nous avons également expérimenté avec les jeux de test des trois dernières éditions, 9b, 8b et 7b. Les meilleurs scores sont directement tirés du classement public[5] de BioASQ.

3.4.1 Résultats des questions de type factoïdes

Pour le jeu de données de la 7éme édition, le modèle que nous avons proposé a pu obtenir le résultat de pointe (SOTA) pour toutes les métriques en deux lots. Le modèle a permis des gains de deux à huit points.

[5] http://participants-area.bioasq.org/results/10b/phaseB

Numéro de lot	Strict Acc.		Lenient Acc.		MRR	
	Score	Top	Score	Top	Score	Top
1	0.3799	0.4103	0.5311	0.5385	0.4432	0.4637
2	**0.5809**	0.5200	**0.7240**	0.6400	**0.6257**	0.5667
3	0.4003	0.4483	0.5957	0.6552	0.4784	0.5115
4	0.5504	0.5882	0.8087	0.8235	0.6616	0.6912
5	**0.3067**	0.2857	**0.5868**	0.5143	**0.3882**	0.3638

Table 11 Résultats pour les questions de type factoïde en BioASQ 7b

Pour le jeu de données de la 8éme édition, nous avons pu atteindre le résultat de pointe (SOTA) sur trois lots. Avec des gains allant de un à six points.

Numéro de lot	Strict Acc.		Lenient Acc.		MRR	
	Score	Top	Score	Top	Score	Top
1	0.3594	0.3750	**0.6820**	0.6250	**0.4972**	0.4688
2	0.1357	0.2800	0.4170	0.4800	0.2530	0.3533
3	**0.3371**	0.3214	0.5570	0.5714	**0.4145**	0.3970
4	**0.5745**	0.5588	**0.7335**	0.7353	**0.6558**	0.6284
5	0.5469	0.5625	0.7445	0.7813	0.6248	0.6354

Table 12 Résultats pour les questions de type factoïde en BioASQ 8b

Quant au jeu de données de la 9ème édition, nous avons pu obtenir le résultat de pointe (SOTA) sur deux lots. En plus d'avoir la meilleure performance dans une métrique sur deux autres lots. Avec des gains allant jusqu'à dix points.

Numéro de lot	Strict Acc.		Lenient Acc.		MRR	
	Score	Top	Score	Top	Score	Top
1	**0.5172**	0.4138	**0.5862**	0.5862	**0.5345**	0.4632
2	0.4412	0.5000	0.7647	0.7941	0.5524	0.6127
3	**0.5833**	0.5833	0.6666	0.7222	0.6027	0.6319
4	0.6071	0.6429	**0.8571**	0.8214	0.6916	0.6929
5	**0.5833**	0.5556	0.6666	0.7222	0.5995	0.6019

Table 13 Résultats pour les questions de type factoïde en BioASQ 9b

Pour la 10éme édition, nous n'avons réussi à obtenir la première position que pour la précision indulgente dans le premier lot. La raison de la baisse de performance de notre approche proposée dans cette dernière édition par rapport aux éditions précédentes peut s'expliquer par le fait que cette édition a vu l'utilisation de modèles plus récents comme PubMedBERT [126], BioELECTRA [201], et BioM-ALBERT [202] Ces modèles ont obtenu de meilleurs résultats que BioBERT, qui est notre modèle de base. Ces modèles nouvellement introduits sont également basés sur l'architecture du transformateur [61]. L'approche que nous proposons peut donc leur être appliquée. Ce que nous avons l'intention de faire en tant que travail futur. Ce qui se traduira par des gains considérables dans le dernier jeu de données de la 10éme édition.

Numéro de lot	**Strict Acc.**			**Lenient Acc.**			**MRR**		
	Score	**Top**	**Rank**	**Score**	**Top**	**Rank**	**Score**	**Top**	**Rank**
1	0.2941	0.4118	12/21	**0.5588**		**1/21**	0.4118	0.4608	9/21
2	0.4412	0.5588	15/23	0.5882	0.6765	10/23	0.5098	0.6000	15/23
3	0.5000	0.5313	**3/29**	0.6563	0.6875	**2/29**	0.5677	0.5792	4/29
4	0.4516	0.5806	5/29	0.6129	0.6774	**3/29**	0.5129	0.5995	11/29
5	0.4138	0.4828	**3/27**	0.5517	0.6207	**3/27**	0.4540	0.5098	5/27
6	0.1667	0.3333	**2/20**	0.1667	0.5000	**3/20**	0.1667	0.3333	9/20

Table 14 Résultats pour les questions de type factoïde en BioASQ 10b

Le tableau 15 ci-dessous montre un exemple d'une prédiction correcte de notre modèle pour une question de type factoïde tirée du premier lot de test du dataset BioASQ 10b

Question	Which disease is caused by mutations in the gene PRF1?
Passage	The presence of mutations in PRF1, UNC13D, STX11 and STXBP2 genes in homozygosis or compound heterozygosis results in immune deregulation. Most such cases lead to clinical manifestations of haemophagocytic lymphohistiocytosis (HLH).
Réponse	hemophagocytic lymphohistiocytosis

Table 15 Exemple d'une prédiction correcte de notre modèle pour une question de type factoïde tirée du premier lot de test du dataset BioASQ 10b

La Figure 23 est similaire à la Figure 20, elle montre les scores d'attention pour la tête 9 dans la couche 1 avant et après l'enrichissement de l'attention par les entités biomédicales et NER pour l'instance (Question : Which disease is Dasatinib used to treat? Passage en contexte : Patients with chronic myeloid leukemia) du premier lot de test dataset BioASQ 8b. La première

réponse candidate retournée pour cette question par notre modèle est "chronic myeloid leukemia". Comme on peut le voir sur la figure, l'enrichissement de l'attention a entraîné des valeurs d'attention plus fortes pour les entités biomédicales dans le passage contextuel "myeloid" et "leukemia", ce qui a aidé notre modèle à renvoyer la bonne réponse.

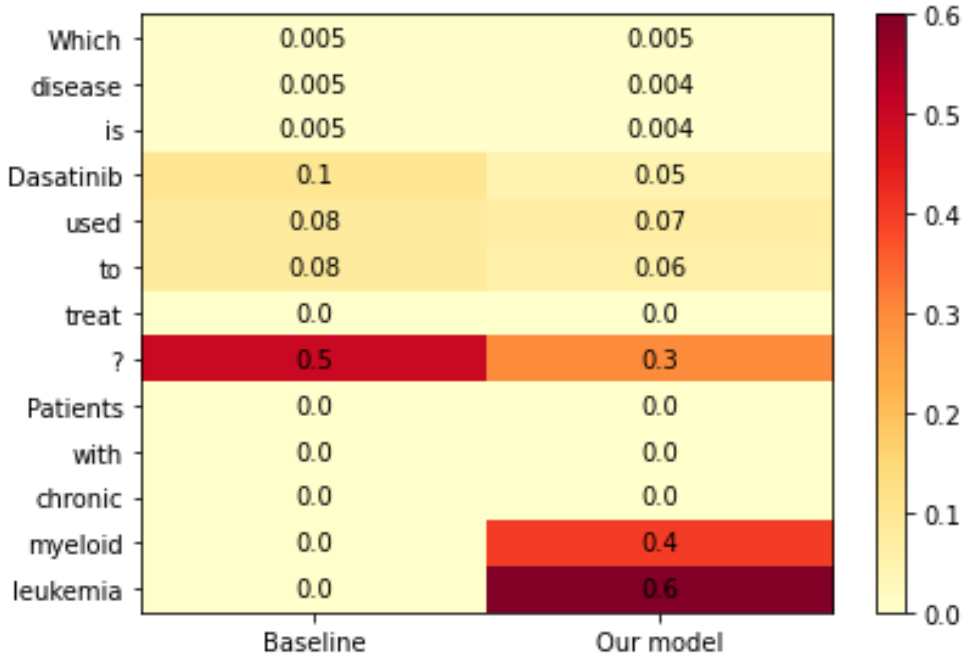

Figure 23 Visualisation des scores d'attention pour la tête 9 dans la couche 1 avant et après l'enrichissement de l'attention par les entités biomédicales et NER

3.4.2 Résultats des questions de type Liste

Pour le jeu de données de la 7éme édition, le modèle que nous avons proposé a pu obtenir le résultat SOTA pour deux métriques en trois lots. Le modèle a permis des gains de un à deux points.

Numéro de lot	Mean Prec.		Recall		F-Measure	
	Score	Top	Score	Top	Score	Top
1	**0.6321**	0.4792	0.4939	0.6026	**0.5279**	0.3276
2	**0.6484**	0.5826	0.4754	0.5143	**0.5123**	0.4732
3	**0.5043**	0.4267	0.3598	0.5338	**0.3713**	0.3298
4	0.4687	0.5024	0.4418	0.6957	0.4216	0.4539
5	0.5071	0.5653	0.3249	0.5873	0.3730	0.4619

Table 16 Résultats pour les questions de type liste en BioASQ 7b

Pour le jeu de données de la 8éme édition, nous n'avons pas pu avoir le résultat SOTA sur aucun lot.

Numéro de lot	Mean Prec.		Recall		F-Measure	
	Score	Top	Score	Top	Score	Top
1	0.4583	0.4875	0.3957	0.5629	0.3962	0.4315
2	0.3392	0.5643	0.3464	0.5929	0.3280	0.4735
3	0.4861	0.7361	0.4163	0.4972	0.3921	0.5020
4	0.4861	0.6520	0.4163	0.5597	0.3921	0.4571
5	0.4201	0.6458	0.3179	0.5645	0.3347	0.5247

Table 17 Résultats pour les questions de type liste en BioASQ 8b

Quant au jeu de données de la 9ème édition, nous avons pu obtenir le résultat SOTA sur une métrique dans trois lots.

Numéro de lot	Mean Prec.		Recall		F-Measure	
	Score	Top	Score	Top	Score	Top
1	**0.5982**	0.5730	0.3507	0.6667	0.4051	0.5339
2	0.5093	0.6829	0.3790	0.6471	0.3803	0.5047
3	0.6517	0.6862	0.4419	0.7221	0.4911	0.5721
4	**0.7292**	0.7167	0.6329	0.8531	0.5841	0.7405
5	**0.5915**	0.5289	0.3448	0.6204	0.4112	0.5306

Table 18 Résultats pour les questions de type liste en BioASQ 9b

Pour la 10éme édition, nous n'avons pas pu avoir des résultats de pointe (SOTA) pour la même raison que pour les questions des type factoïde, à savoir que cette édition a vu l'utilisation de modèles plus récents comme PubMedBERT [126], BioELECTRA [201], et BioM-ALBERT [202]. Ces modèles ont obtenu de meilleurs résultats que notre modèle de base BioBERT.

Numéro de lot	Mean Prec.		Recall		F-Measure	
	Score	Top	Score	Top	Score	Top
1	0.5714	0.7201	0.4821	0.8464	0.4959	0.7469
2	0.6143	0.7042	0.4281	0.7530	0.4490	0.7051
3	0.4923	0.6273	0.4128	0.6742	0.4052	0.5655
4	0.4736	0.6162	0.3104	0.5844	0.3048	0.5386
5	0.4486	0.6799	0.4188	0.6407	0.4191	0.6016
6	0.5044	0.5730	0.3669	0.4690	0.3854	0.4534

Table 19 Résultats pour les questions de type liste en BioASQ 10b

Notre modèle pour les questions de type liste a donné des résultats insatisfaisants par rapport à celui des questions de type factoïde. Pourtant, nous avons utilisé la même technique de mécanisme d'enrichissement de l'attention pour les questions de type factoïde et de type liste. Nous prévoyons d'étudier ce fait maintenant que la compétition BioASQ 10b est terminée.

Le tableau 20 ci-dessous montre un exemple d'une prédiction correcte de notre modèle pour une question de type liste tirée du premier lot de test du dataset BioASQ 10b

Question	Which drugs are included in the CABENUVA pill?
Passage	The clinical phase III trials of FLAIR and ATLAS, showed two-drug injectable cabotegravir (CAB) and rilpivirine (RPV) formulation is potent, safe, and tolerable in HIV-infected patients. The recent approval of cabenuva (CAB+RPV) by Health Canada is a milestone in the development of long-term therapies for HIV infection.
Réponses	cabotegravir rilpivirine

Table 20 Exemple d'une prédiction correcte de notre modèle pour une question de type liste tirée du premier lot de test du dataset BioASQ 10b

3.4.3 Discussion des résultats

En plus des résultats de pointe (SOTA) dans certains lots de tests au niveau de tous les datasets d'évaluation, à savoir BioASQ 10b, 9b, 8b, et 7b. Nous avons aussi réussi à avoir des résultats compétitifs dans presque l'ensemble des autres lots de tests. Néanmoins, nous n'avons pas pu

avoir des résultats de pointe (SOTA) dans tous les lots de tests d'une des datasets d'évaluation. En général, notre modèle remporte les premiers scores dans deux à trois lots de tests, avec des résultats inférieurs pour les autres lots. Cette variation de performances entre les différents lots de tests peut s'expliquer par trois choses, **(1)** chaque lot de test comporte bien évidemment des questions différentes, mais aussi des types de question/réponse différentes (entité biomédical, entité nommée, mode d'administration médicament, procédure médicale, …), vu que notre approche exploite les noms d'entités biomédicales et entités nommées pour enrichir le modèle de base BioBERT donc elle n'offrira pas forcément les mêmes performances pour les autres types de questions/réponses (mode d'administration médicament, procédure médicale,…). **(2)** chaque lot de test peut voir l'introduction d'une nouvelle équipe participante. Les meilleurs scores qui sont affichés dans les tableaux des résultats appartiennent en général à différentes équipes à chaque lot de test. Donc par exemple, un système A qui a réussi à avoir le meilleur score face aux systèmes B, et C dans le premier lot de test. Peut se faire dépasser par un nouveau système entrant D dans le deuxième lot de test. **(3)** les modèles d'apprentissage profond souffrent toujours d'un problème de reproductibilité des résultats [203], le même modèle peut avoir des résultats un peu différents à chaque évaluation.

Un autre constat, que nous avons déjà mentionné dans la partie des résultats, est la dégradation des performances de notre modèle lors de la dernière édition du challenge BioASQ, à savoir, le dataset 10b. La raison de cette baisse de performance par rapport aux éditions précédentes peut s'expliquer par le fait que cette édition a vu l'utilisation de modèles plus récents comme PubMedBERT [126], BioELECTRA [201], et BioM-ALBERT [202]. Ces modèles ont obtenu de meilleurs résultats que BioBERT, qui est notre modèle de base.

Comme travail futur, nous comptons travailler sur ces limites par, **(1)** analyser les fausses prédictions de notre modèle dans différents lots de tests, et déterminer la corrélation exacte entre la prédiction du modèle et les types de question/réponse (entité biomédical, entité nommée, mode d'administration médicament, procédure médicale, …). **(2)** Utiliser notre approche avec un PLM biomédical plus récent.

3.5 Conclusion

Dans ce chapitre, nous avons présenté un nouveau mécanisme d'enrichissement de l'auto-attention pour la tâche de BQA. Nous nous attaquons spécifiquement aux questions de type factoïde et liste. La méthode proposée est basée sur l'hypothèse que les jetons dans le passage contextuel avec des scores d'attention plus élevés ont une plus grande probabilité de faire partie de la réponse prédite. Dans notre approche, nous avons enrichi la couche d'auto-attention de BioBERT avec des informations biomédicales et des informations sur les entités nommées extraites de la question et du passage contextuel. La méthode proposée a obtenu des résultats de pointe (SOTA) sur plusieurs lots des jeux de données 10b, 9b, 8b et 7b du dataset BioASQ. L'hypothèse de la corrélation entre les scores d'attention et la prédiction du modèle a également été confirmée. Notre approche est actuellement utilisée avec BioBERT. Mais nous prévoyons de l'utiliser avec les modèles les plus récents comme PubMedBERT, BioM-ELECTRA et BioM-ALBERT. Car ils sont tous basés sur l'architecture du transformateur qui est supportée par notre approche. Dans le chapitre suivant, nous allons présenter nos modèles BQA pour les questions de type Oui/Non et résumé.

Chapitre IV :
Application de l'apprentissage par transfert pour les questions BQA de type Oui/Non et résumé

4.1 Introduction

Bien que la technologie traditionnelle d'apprentissage machine (ML) ait connu un grand succès et ait été appliquée avec succès dans de nombreuses applications pratiques, elle présente encore certaines limites pour certains scénarios du monde réel. Le scénario idéal de l'apprentissage automatique consiste à disposer d'un grand nombre d'instances d'apprentissage étiquetées, qui présentent la même distribution que les données de test. Cependant, la collecte d'un nombre suffisant de données d'apprentissage est souvent coûteuse, longue, voire irréaliste dans de nombreux scénarios. L'apprentissage semi-supervisé peut résoudre en partie ce problème en assouplissant le besoin de données étiquetées en masse. En général, une approche semi-supervisée ne nécessite qu'un nombre limité de données étiquetées et utilise une grande quantité de données non étiquetées pour améliorer la précision de l'apprentissage. Mais dans de nombreux cas, les instances non étiquetées sont également difficiles à collecter, ce qui rend généralement les modèles traditionnels résultants insatisfaisants.

L'apprentissage par transfert, qui se concentre sur le transfert des connaissances entre les domaines, est une méthodologie d'apprentissage automatique prometteuse pour résoudre le problème ci-dessus. L'idée derrière l'apprentissage par transfert est de tirer profit des caractéristiques et des connaissances acquises par un modèle sur une tâche, et de les appliquer à une nouvelle tâche connexe pour améliorer les performances du modèle sur la nouvelle tâche. Il est actuellement à l'origine d'améliorations considérables dans une variété de tâches NLP.

Dans ce chapitre, nous allons décrire notre méthode BQA pour les questions de types Oui/Non et résumé, qui repose sur l'apprentissage par transfert. D'abord, nous allons introduire les concepts de l'apprentissage par transfert, la génération de texte « Natural Language Generation » (NLG), et le modèle BART [204] que nous avons utilisé pour les questions de type résumé. Ensuite, nous allons citer les méthodes et systèmes BQA de pointes pour les questions de types Oui/Non et résumé. Après ceci, nous allons détailler notre approche pour les deux types de questions cités en ce qui concerne les datasets utilisés pour le transfert d'apprentissage, et les modèles et les architectures adoptées. Nous allons terminer la présentation de notre approche avec les résultats obtenus toujours sur le dataset BioASQ [4]. Enfin, nous allons terminer ce chapitre avec une conclusion.

4.1.1 L'apprentissage par transfert

L'apprentissage par transfert est un domaine de recherche dans l'apprentissage machine « Machine Learning » (ML) qui se concentre sur le stockage des connaissances acquises lors de la résolution d'un problème et leur application à un problème différent mais connexe (voir la Figure 24). Par exemple, les connaissances acquises lors de l'apprentissage de la reconnaissance des voitures peuvent être appliquées à la reconnaissance des camions. L'un des principaux avantages de l'apprentissage par transfert est qu'il permet de réduire considérablement la quantité de données étiquetées et les ressources informatiques nécessaires pour entraîner un modèle pour une nouvelle tâche, car le modèle peut utiliser les connaissances qu'il a déjà acquises lors de la première tâche. Cela peut être particulièrement utile lorsque les données pour la nouvelle tâche sont limitées ou lorsque la nouvelle tâche est similaire à une tâche qui a déjà été bien étudiée et pour laquelle il existe déjà un modèle de haute performance.

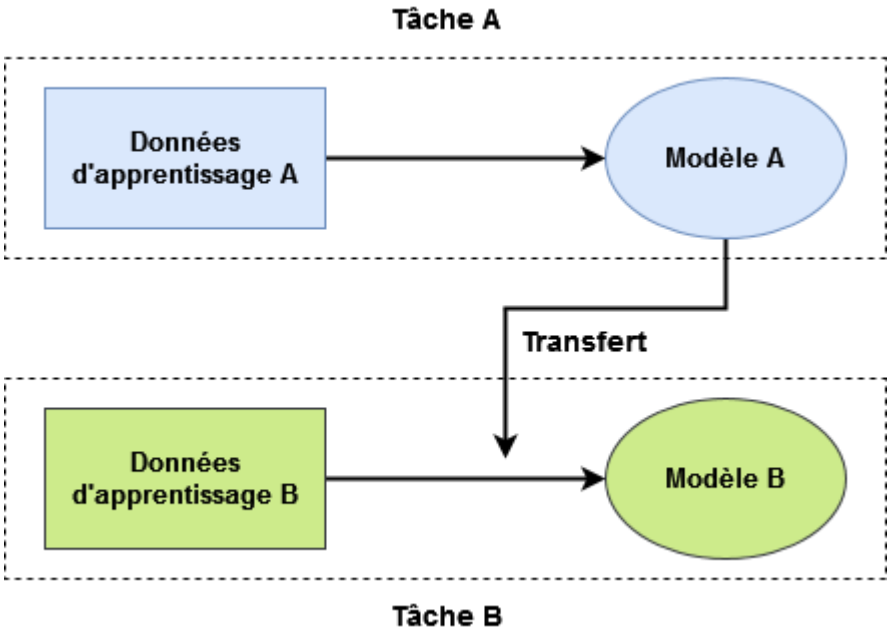

Figure 24 Principe de l'apprentissage par transfert "Transfert learning"

La définition de l'apprentissage par transfert est donnée en termes de domaines et de tâches. Un domaine $\mathcal{D}$ consiste en : un espace de caractéristiques $\mathcal{X}$ et une distribution de probabilité marginale $P(X)$ où $X = \{x_1, \dots, x_n\} \in \mathcal{X}$. Étant donné un domaine spécifique $\mathcal{D} = \{\mathcal{X}, P(X)\}$, une tâche consiste en deux composantes : un espace d'étiquettes $\mathcal{Y}$ et une fonction objective de

prédiction $f : \mathcal{X} \rightarrow \mathcal{Y}$. La fonction f est utilisée pour prédire l'étiquette correspondante $f(x)$ d'une nouvelle instance x. Cette tâche, désignée par $\mathcal{T} = \{\mathcal{Y}, f(x)\}$, est apprise à partir des données d'apprentissage constituées de paires $\{x_i, y_i\}$, où $x_i \in X$ et $y_i \in \mathcal{Y}$. Étant donné un domaine source $\mathcal{D}_S$ et une tâche d'apprentissage $\mathcal{T}_S$, un domaine cible $\mathcal{D}_T$ et une tâche d'apprentissage $\mathcal{T}_T$, où $\mathcal{D}_S \neq \mathcal{D}_T$, ou $\mathcal{T}_S \neq \mathcal{T}_T$, l'apprentissage par transfert vise à aider à améliorer l'apprentissage de la fonction prédictive cible $f_T(\cdot)$ dans $\mathcal{D}_T$ en utilisant les connaissances dans $\mathcal{D}_S$ et $\mathcal{T}_S$.

4.1.2 La génération de texte

La génération en langage naturel, en anglais « Natural Language Generation » (NLG) est un sous-domaine de l'NLP qui s'intéresse à la construction de systèmes informatiques capables de produire des textes compréhensibles en anglais ou dans d'autres langues humaines à partir d'une représentation non-linguistique sous-jacente de l'information. Les systèmes de génération de langage naturel combinent des connaissances sur la langue et le domaine d'application pour produire automatiquement des documents, des rapports, des explications, des messages d'aide et d'autres types de textes.

En général, la tâche NLG vise à trouver une séquence optimale $y_{<T+1} = (y_1, y_2, \dots, y_T)$ qui satisfait à :

$$y_{<T+1} = \underset{y<T+1 \in \mathcal{Y}}{\operatorname{argmax}} \log P_\theta(y_{<T+1}|x) = \underset{y<T+1 \in \mathcal{Y}}{\operatorname{argmax}} \sum_{t=1}^{T} \log P_\theta(y_t|y_{<t}, x) \qquad (4.1)$$

Où T représente le nombre de jetons de la séquence générée, $\mathcal{Y}$ représente un ensemble contenant toutes les séquences possibles, et $P_\theta(y_t|y_{<t}, x)$ est la probabilité conditionnelle du jeton suivant y_t en fonction de ses jetons précédents $y_{<t} = (y_1, y_2, \dots, y_{t-1})$ et de la séquence source x avec des paramètres de modèle θ.

Il existe plusieurs approches pour l'NLG, notamment les systèmes basés sur des règles, qui utilisent des règles et des modèles « templates » prédéfinis pour générer du texte, et les systèmes basés sur l'apprentissage automatique, qui utilisent des données pour apprendre des modèles et des relations dans le langage et générer du texte en fonction de ces modèles. Les méthodes NLG récentes utilisent principalement : Réseau neuronal récurrent « Recurrent Neural Network »

(RNN) [58], transformateur [61], mécanisme d'attention [49], mécanismes de copie et de pointage « Copy and Pointing Mechanisms » [205], réseau adversarial génératif « Generative Adversarial Network » (GAN) [206], réseau de mémoire « Memory Network » [207, 208], réseau neuronal graphique « Graph Neural Network » (GNN) [209] et les PLM [210].

4.1.3 Le modèle de génération de texte BART

BART [204] est un auto-codeur de débruitage « denoising autoencoder » pour le pré-entraînement de modèles séquence-à-séquence. Il est entraîné en corrompant le texte avec une fonction de bruitage arbitraire et en apprenant un modèle pour reconstruire le texte original. Il utilise une architecture neuronale standard de traduction automatique basée sur le Transformer [61] qui, malgré sa simplicité, peut être considérée comme une généralisation de BERT [1] (grâce à l'encodeur bidirectionnel), de GPT [2] (avec le décodeur gauche-droite) et d'autres schémas de pré-entraînement récents. BART est particulièrement efficace lorsqu'il est ajusté pour la génération de texte, mais fonctionne bien également pour les tâches de compréhension. Il obtient de nouveaux résultats de pointe sur une série de tâches de dialogue abstrait, de QA et de résumé, avec des gains allant jusqu'à 3,5 ROUGE. BART fournit également une augmentation de 1,1 BLEU par rapport à un système de rétro-traduction pour la traduction automatique, avec seulement un pré-entraînement en langue cible.

Un modèle BART finement ajusté peut prendre une séquence de texte (par exemple, l'anglais) en entrée et produire une séquence de texte différente en sortie (par exemple, le français). Ce type de modèle est pertinent pour la traduction automatique (traduction d'un texte d'une langue à une autre), la réponse à des questions (production de réponses à une question donnée sur un corpus spécifique), le résumé de texte (résumé ou paraphrase d'un long document textuel), ou la classification de séquences (catégorisation de phrases ou de jetons de texte en entrée).

4.2 État de l'art : approches pour les questions BQA de types Oui/Non et résumé

Approches pour les questions BQA de type Oui/Non : Dans [211], Sarrouti et al. ont utilisé la bibliothèque CoreNLP de Stanford [210] pour la tokénisation et le marquage des parties du discours, en anglais « part-of-speech tagging » de tous les passages pertinents d'une question

de type oui/non. Ils ont ensuite attribué un score de sentiment basé sur SentiWordNet [213] à chaque mot des passages. Enfin, la décision sur les réponses "oui" ou "non" est basée sur le score sentiment-passages obtenu : "oui" pour un score sentiment-passages final positif et "non" pour un score final négatif. Les évaluations expérimentales effectuées sur le dataset BioASQ 3b [4] montrent que la méthode proposée est plus efficace que la méthode de pointe à l'époque, qu'elle surpasse de 15,68 % en moyenne en termes de précision. Les auteurs ont également incorporé cette approche dans le système qu'ils ont proposé, SemBioNLQA [95], pour obtenir des résultats de pointe (SOTA) dans les 3éme et 4éme éditions de BioASQ. Dans [214], Telukuntla et al. ont utilisé l'implication « entailment » pour les questions Oui/Non de la 7éme édition de BioASQ. Étant donné une question, ils ont itéré à travers les phrases candidates et ont essayé de trouver une phrase candidate qui contredit la question (avec un niveau de confiance supérieur à 50 %), si c'est le cas, "Non" est renvoyé comme réponse, sinon "Oui" est renvoyé. L'utilisation de cette approche a permis d'augmenter les performances d'environ 13% (score F1 macro) par rapport au modèle de base de BioASQ. Dans [215], Kommaraju et al. ont expérimenté avec BioBERT [123] et SciBERT [124], ainsi qu'avec l'apprentissage par transfert à partir des datasets SQuAD [5] et PubMedQA [79]. Ils ont également utilisé une technique de traduction par cloze [66] pour augmenter les données d'entraînement où un extracteur d'entités nommées détecte les entités nommées à partir d'un contexte donné. Ils sélectionnent ensuite un nom d'entité biomédicale présent dans le contexte. Pour une instance positive, ils introduisent le nom correct de l'entité biomédicale dans la question ainsi que le contexte non modifié dans le modèle et la réponse sera "Oui". Pour une instance négative, ils remplacent aléatoirement le nom de l'entité biomédicale dans le contexte par une autre entité biomédicale, puis ils associent le même nom d'entité biomédicale erroné en tant que question et l'envoient au modèle ; la réponse à cette question sera un "Non". L'utilisation de cette approche avec l'apprentissage par transfert a amélioré les performances du modèle. Dans [216], Kazaryan et al. ont expérimenté avec deux modèles, ALBERT [62] ajusté à l'aide de SQuAD [5] et BioBERT [123] ajusté à l'aide de PubMedQA [79]. Les deux modèles ont donné des résultats compétitifs sur les 8éme et 9éme éditions de BioASQ. Dans [140], l'équipe de BioBERT a expérimenté avec l'apprentissage par transfert à partir des datasets SQuAD puis MNLI [217] pour les questions Oui/Non de BioASQ. L'apprentissage par transfert en utilisant ces deux datasets et dans cet

ordre a donné des résultats de pointe (SOTA) sur la 7ème édition de BioASQ pour l'équipe de BioBERT. Alrowili et al. dans [218] ont suivi la même stratégie avec leurs modèles BioM-ELECTRA, et BioM-ALBERT et ont obtenu des résultats de pointe (SOTA) sur les 9ème et 10ème éditions de BioASQ pour les questions de type Oui/Non.

Approches pour les questions BQA de type résumé : Dans [219], Chen et al. ont proposé un système de résumé d'informations médicales basé sur des questions et utilisant les connaissances ontologiques de l'UMLS et l'ontologie de la « National Library of Medicine ». L'algorithme de résumé est basé sur les termes, et seuls les termes définis dans UMLS sont reconnus et traités. La procédure de résumé est la suivante : a) révision de la question avec les connaissances ontologiques de l'UMLS ; b) calcul de la distance de chaque phrase du document par rapport à la question finalisée ; c) calcul des distances par paires entre les phrases candidates, puis division des phrases candidates en groupes en fonction d'un seuil et sélection de la phrase la mieux classée de chaque groupe. Lorsqu'il est déterminé quelles phrases seront incluses dans le résumé, trois scores différents sont générés et normalisés avec la longueur de la phrase. Dans [220] Yuan et al. ont proposé un modèle de langage génératif auto-régressif biomédical, BioBART, pré-entraîné sur les corpus biomédicaux. Ils ont adopté BART [204], un PLM génératif qui a obtenu des résultats SOTA sur différentes tâches NLG dans le domaine général, en pré-entraînant continuellement BART sur des résumés PubMed pour réaliser une adaptation au domaine biomédical. Ils ont évalué BioBART sur les tâches NLG biomédicales existantes. Le modèle BioBART dans le domaine surpasse le modèle BART et établit des « baselines » solides pour plusieurs tâches NLG. Dans [216], Kazaryan et al. ont utilisé une architecture traditionnelle de transformateur codeur-décodeur [61], où le codeur est basé sur BioMed-RoBERTa [221], tandis que le décodeur est entrainé à partir de zéro, en suivant BertSUM [222]. Tout d'abord, ils ont pré-entraîné le modèle sur un ensemble de données de résumé basé sur PubMed, où la cible est un intervalle arbitraire du résumé et la source est un morceau de texte, à partir duquel la cible peut être dérivée. Ensuite, ils ont affiné le modèle pour produire des résumés à partir de la question et de la concaténation d'extraits pertinents de l'ensemble de données d'entraînement BioASQ, séparés par un jeton spécial. Le système proposé a donné des résultats compétitifs à la 8éme édition de BioASQ pour les questions de type résumé. Dans [223], Luo et al. ont proposé BioGPT, un modèle de langage génératif spécifique au domaine,

pré-entraîné sur la littérature biomédicale. BioGPT a été évalué sur six tâches de traitement du langage naturel biomédical et a démontré qu'il surpasse les modèles précédents dans la plupart des tâches, y compris la génération de résumés.

4.3 Notre approche pour les questions de type Oui/Non

4.3.1 Processus d'apprentissage par transfert

Comme mentionné dans l'introduction, nous avons utilisé l'apprentissage par transfert pour traiter les questions Oui/Non. Nous avons commencé par transférer les connaissances depuis le dataset BoolQ [224], puis du dataset PubMedQA [79] dans cet ordre. L'ordre de l'apprentissage par transfert séquentiel est important pour combler l'écart entre les différentes tâches. Les performances s'améliorent lorsque la fonction objective de la tâche à réglage fin devient similaire à celle de la tâche en aval [140]. Par conséquent, nous avons commencé par BoolQ [224], qui est un dataset QA du type Oui/Non du domaine général, puis PubMedQA [79], un dataset QA de type Oui/Non biomédical, qui est plus proche de BioASQ, notre dataset cible. La Figure 25 ci-dessous illustre notre processus d'apprentissage par transfert. Le tableau 21 donne des informations clés sur BoolQ et PubMedQA, les deux datasets que nous avons utilisés dans notre processus.

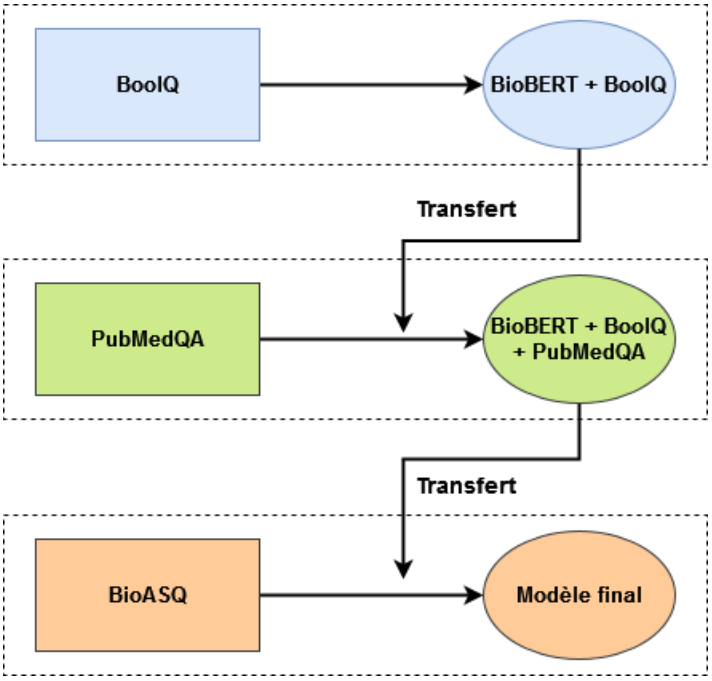

Figure 25 Notre processus d'apprentissage par transfert pour les questions de type Oui/Non

BoolQ : Le jeu de données BoolQ [224] (pour Boolean Questions) contient 16 mille questions de type Oui/Non. Chaque question est associée à un paragraphe de Wikipedia qu'un annotateur indépendant a marqué comme contenant la réponse. La tâche consiste donc à prendre une question et un passage en entrée, et à renvoyer "oui" ou "non" en sortie. Les questions sont recueillies à partir de requêtes anonymes et agrégées sur le moteur de recherche Google. Les questions ne sont conservées que si une page Wikipedia est retournée comme l'un des cinq premiers résultats, auquel cas la question et la page Wikipedia sont données à un annotateur humain pour un traitement ultérieur. Les annotateurs étiquettent les paires question/article selon un processus en trois étapes. Tout d'abord, ils décident si la question est bonne, c'est-à-dire si elle est compréhensible, non ambiguë et si elle demande des informations factuelles. Ce jugement est fait avant que l'annotateur ne voie la page Wikipédia. Ensuite, pour les bonnes questions, les annotateurs trouvent un passage dans le document qui contient suffisamment d'informations pour répondre à la question. Les annotateurs peuvent marquer les questions comme "sans réponse" si l'article de Wikipédia ne contient pas les informations demandées. Enfin, les annotateurs indiquent si la réponse à la question est "oui" ou "non". Annoter des données de cette manière est assez coûteux, car les annotateurs doivent rechercher des documents Wikipédia entiers pour trouver des preuves pertinentes et lire le texte attentivement. Les questions portent souvent sur le divertissement (comme la télévision, les films et la musique), ainsi que sur d'autres sujets populaires comme le sport. Cependant, il y a encore une bonne partie des questions qui demandent des connaissances factuelles plus générales, notamment sur des événements historiques ou le monde naturel.

PubMedQA [79] : il s'agit d'un nouveau dataset BQA collecté à partir des résumés de PubMed. La tâche de PubMedQA est de répondre à des questions de recherche avec oui/non/peut-être (par ex : Les statines préopératoires réduisent-elles la fibrillation auriculaire après un pontage aorto-coronarien ?) en utilisant les résumés correspondants. PubMedQA possède mille instances annotées par des experts, 61,2 mille instances non étiquetées et 211,3 mille instances générées artificiellement. Chaque instance PubMedQA est composée de (1) d'une question qui est soit un titre d'article de recherche existant, soit un titre dérivé, (2) d'un contexte qui est le résumé correspondant sans sa conclusion, (3) d'une réponse longue, qui est la conclusion du résumé et qui, vraisemblablement, répond à la question de recherche, et (4) d'une réponse

oui/non/peut-être qui résume la conclusion. PubMedQA est le premier ensemble de données BQA où le raisonnement sur des textes de recherche biomédicale, en particulier leur contenu quantitatif, est nécessaire pour répondre aux questions.

	BoolQ	**PubMedQA**
Domaine	General	Biomédical
Nombre d'instances	16 mille	211,3 mille
Longueur moyenne des questions (en jetons)	8.9	16.3
Longueur moyenne des passages (en jetons)	108	238
Nécessite du raisonnement	Oui	Oui
Source des données	Requêtes de recherche Google et Wikipedia	Abstracts PubMed
Type de construction	Annoté manuellement	Annoté manuellement/généré automatiquement

Table 21 Statistiques sur les datasets BoolQ et PubMedQA

4.3.1 Architecture

Comme pour les questions de type factoïde et liste, nous avons utilisé le même modèle, à savoir BioBERT [123], aussi pour les questions de type Oui/Non. En suivant toujours la même procédure d'injection de données que pour les questions factoïdes et listes. L'entrée du modèle BioBERT est $X = \{[CLS] \,||\, Q' \,||\, [SEP] \,||\, P' \,||\, [SEP]\}$ où Q' et P' sont la question Q et le passage contextuel P tokenizés par le WordPiece tokenizer de BioBERT. [CLS] est un jeton BERT [1] prédéfini utilisé dans les tâches de classification. Le jeton BERT prédéfini [SEP] est utilisé comme séparateur entre les entrées du modèle. La question et le passage contextuel tokenisés, ainsi que les jetons spéciaux de BERT, sont ensuite concaténés pour former une seule entrée. Le vecteur de représentation caché du $i^{ième}$ token d'entrée est noté $h_i \in \mathbb{R}^H$ où H est la taille cachée.

Pour calculer la probabilité de « oui » P^{yes}, nous avons projeté une matrice de transformation linéaire $M \in R^H$ pour transformer la représentation cachée du jeton [CLS] $C \in R^H$. Dans la classification binaire, la fonction sigmoïde est utilisée pour calculer la probabilité de « oui » comme suit :

$$P^{yes} = \frac{1}{1 + e^{-C \cdot M^T}} \tag{4.2}$$

La perte binaire d'entropie croisée « binary cross entropy loss » est utilisée entre la probabilité de « oui » P^{yes} et sa réponse correcte correspondante a_{yes}. Notre perte totale est calculée comme suit.

$$Loss = -\left(a_{yes} \log P^{yes} + \left(1 - a_{yes}\right) log(1 - P^{yes})\right) \tag{4.3}$$

La Figure 26 ci-dessous montre notre architecture décrite en haut.

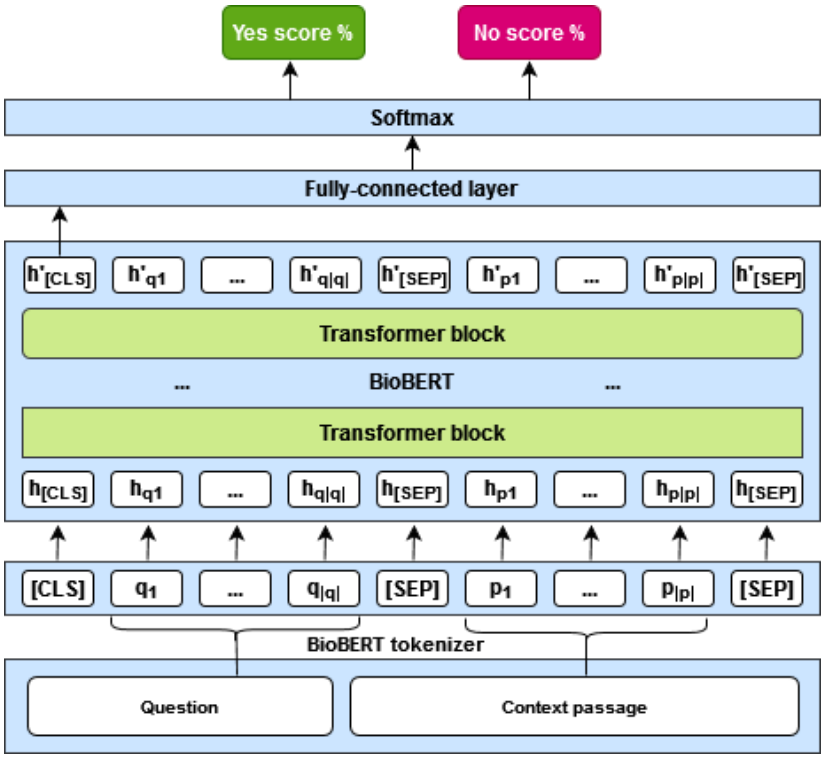

Figure 26 Architecture globale de notre méthode pour les questions de type Oui/Non

Le tableau 22 ci-dessous présente les détails d'implémentation techniques. Nous avons choisi presque les mêmes paramètres de configuration que dans notre modèle pour les questions de type factoïde et liste. Nous avons fixé la longueur maximale de séquence à 300, ce qui signifie qu'on peut traiter jusqu'à 300 mots à la fois. La taille du lot « batch size » est de 16, ce qui signifie que le modèle traite 16 exemples à la fois pendant l'entraînement. Le taux d'apprentissage « learning rate » est de 8e-6, ce qui détermine la vitesse à laquelle le modèle apprend à partir des données d'entraînement. Le modèle a été entraîné pendant 3 époques, ce

qui signifie qu'il a été exposé aux données d'entraînement 3 fois. Le modèle a été entraîné sur une GPU NVIDIA RTX 6000 de 24 Go de mémoire, ce qui a permis d'accélérer le processus d'apprentissage. Le modèle a été développé à l'aide du Framework Pytorch.

Paramètre	Valeur
Modèle de base	BioBERT
Taille du modèle	Large
Longueur maximale de la séquence	300
Taille du lot « Batch size »	16
Taux d'apprentissage « Learning rate »	8e-6
Époques d'entraînement « Train epochs »	3
GPU	NVIDIA RTX 6000 24gb
Framework Deep learning	Pytorch 1.5.1

Table 22 Détails techniques d'implémentation (questions de type Oui/Non)

4.3.2 Evaluation et résultats

Comme pour les questions de type factoïde et liste, nous présentons ci-dessous nos résultats pour les questions de type Oui/Non dans BioASQ, auquel nous avons aussi participé avec notre système.

Dans le tableau 23, nous comparons les performances de notre modèle avec le modèle de base BioBERT-large, et avec le même modèle après réglage fin sur le dataset BoolQ. Cette comparaison a été effectuée sur le premier lot de la 9ème édition du dataset BioASQ. Comme on peut le voir dans le tableau, l'apprentissage par transfert avec BoolQ et PubMedQA a considérablement amélioré les résultats sur les quatre métriques. Des résultats similaires ont été obtenus sur d'autres lots et sur les autres jeux de données 10b, 8b et 7b.

Modèle	Macro F1	F1 Yes	F1 No	Précision
BioBERT-large	0.6376	0.6268	0.6600	0.6134
BioBERT-large + BoolQ	0.6666	0.6648	0.6896	0.6400
BioBERT-large + BoolQ + PubMedQA	**0.7037**	**0.7032**	**0.7142**	**0.6920**

Table 23 Comparaison des performances les modèles BioBERT, BioBERT + BoolQ, et BioBERT + BoolQ + PubMedQA sur le premier lot de la 9ème édition du jeu de données BioASQ

Dans BioASQ 10b notre modèle pour les questions oui/non s'est classé dans les trois premières positions dans au moins une métrique pour les lots un, quatre et six. Dans le deuxième lot, un problème technique dans notre script de génération de résultats nous a empêchés de soumettre les résultats du modèle, nous avons donc fixé la réponse à « Non » pour toutes les questions, par manque de temps. Cela explique la grande différence de scores dans le deuxième lot.

Numéro de lot	Macro F1			F1 Yes		F1 No		Précision		
	Score	Top	Rank	Score	Top	Score	Top	Score	Top	Rank
1	0.9464	1.0000	**2/25**	0.9697	1.0000	0.9231	1.0000	0.9565	1.0000	**2/25**
2	0.3378	1.0000	11/35	0.1538	1.0000	0.5217	1.0000	0.3889	1.0000	10/35
3	0.8252	0.9600	4/38	0.9500	0.9730	0.8000	0.9231	0.8800	0.9480	4/38
4	0.9473	1.0000	**3/39**	0.9714	1.0000	0.9231	1.0000	0.9583	1.0000	**2/39**
5	0.7333	0.9286	10/41	0.8000	0.9375	0.6667	0.9231	0.7500	0.9282	6/41
6	0.6250	1.0000	4/32	0.7500	1.0000	0.5000	1.0000	0.6667	1.0000	**3/32**

Table 24 Résultats de notre approche pour les questions de type Oui/Non en BioASQ 10b

Afin de confirmer les performances de notre approche, nous avons également expérimenté avec les jeux de test des trois dernières éditions, 9b, 8b et 7b. Les meilleurs scores sont directement tirés du classement public de BioASQ. Le tableau 24 présente nos résultats pour les questions de type Oui/Non en BioASQ 9b. Malgré le fait que nous n'avons pas eu des résultats SOTA sur ce jeu de test, nos résultats restent compétitifs.

Numéro de lot	Macro F1		F1 Yes		F1 No		Précision	
	Score	Top	Score	Top	Score	Top	Score	Top
1	0.7037	0.9258	0.7032	0.9286	0.7142	0.9231	0.6920	0.9259
2	0.8181	0.9454	0.7904	0.9677	0.8666	0.9231	0.7142	0.9545
3	0.8333	0.9473	0.7983	0.9714	0.8823	0.9231	0.7142	0.9583
4	0.8000	0.9480	0.7619	0.9730	0.8571	0.9231	0.6666	0.9600
5	0.6315	0.8081	0.6144	0.8889	0.6956	0.7273	0.5333	0.8421

Table 25 Résultats de notre approche pour les questions de type Oui/Non en BioASQ 9b

Pour le jeu de données de la 8éme édition, nous avons pu atteindre le résultat SOTA en deux métriques dans le lot de test trois. Le tableau 25 montre nos résultats sur ce jeu de test de l'édition 8b.

Numéro de lot	Macro F1		F1 Yes		F1 No		Précision	
	Score	Top	Score	Top	Score	Top	Score	Top
1	0.6800	0.8800	0.6323	0.9143	0.7647	0.8235	0.5000	0.8663
2	0.8055	0.9444	0.7304	0.9630	0.8727	0.8889	0.5882	0.9259
3	**0.9032**	0.9032	0.8963	0.9189	**0.9230**	0.8966	0.8695	0.9028
4	0.6153	0.8462	0.6060	0.8571	0.6666	0.8333	0.5454	0.8452
5	0.7941	0.8529	0.7939	0.8649	0.8000	0.8485	0.7878	0.8528

Table 26 Résultats de notre approche pour les questions de type Oui/Non en BioASQ 8b

Quant au jeu de données de la 7éme édition, le modèle que nous avons proposé a pu obtenir le résultat SOTA en « F1 No » dans 4/5 lots de test, en plus du résultat SOTA en « Macro F1 » dans le cinquième lot.

Numéro de lot	Macro F1		F1 Yes		F1 No		Précision	
	Score	Top	Score	Top	Score	Top	Score	Top
1	0.7931	0.8276	0.7175	0.8980	**0.8636**	0.4444	0.5714	0.6712
2	0.8000	0.8333	0.7916	0.8387	**0.8333**	0.8276	0.7500	0.8331
3	0.7391	0.8696	0.6617	0.9231	**0.8235**	0.5714	0.5000	0.7473
4	0.5652	0.8696	0.4866	0.9189	0.6875	0.7273	0.2857	0.8208
5	**0.8286**	0.8286	0.8250	0.8500	**0.8500**	0.8000	0.800	0.8250

Table 27 Résultats de notre approche pour les questions de type Oui/Non en BioASQ 7b

Le tableau 28 ci-dessous montre un exemple d'une prédiction correcte de notre modèle pour une question de type Oui/Non tirée du premier lot de test du dataset BioASQ 10b

Question	Is Sotrovimab effective for COVID-19?
Passage	It seems that monoclonal antibodies (e.g., low dosage bamlanivimab, baricitinib, imatinib, and sotrovimab) are a better choice for treating severe or non-severe COVID-19 patients.
Réponse	Yes

Table 28 Exemple d'une prédiction correcte de notre modèle pour une question de type Oui/Non tirée du premier lot de test du dataset BioASQ 10b

4.4 Notre approche pour les questions de type résumé

4.4.1 Processus d'apprentissage par transfert

Comme pour les questions de type Oui/Non, nous avons utilisé l'apprentissage par transfert pour traiter les questions de type résumé. Nous avons commencé par transférer les

connaissances depuis le dataset CNN/Daily Mail [55] qui est un dataset de type résumé « summarization » du domaine général, puis du dataset Ebmsum [225], un dataset de type résumé du domaine biomédical. La Figure 27 ci-dessous illustre notre processus d'apprentissage par transfert. Le tableau 29 donne des informations clés sur CNN/Daily Mail et Ebmsum, les deux datasets que nous avons utilisés dans notre processus.

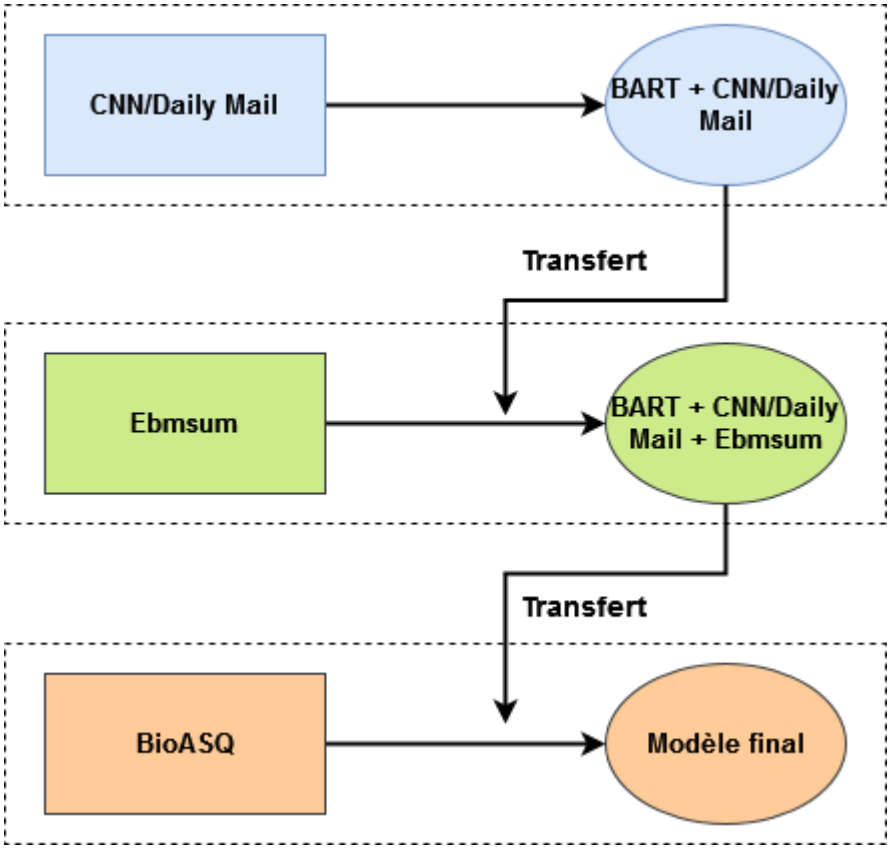

Figure 27 Notre processus d'apprentissage par transfert pour les questions de type résumé

CNN/Daily Mail [55]: est un dataset pour le résumé de texte. Des résumés abstraits générés par des humains ont été générés à partir d'articles d'actualité sur les sites Web de CNN et du Daily Mail sous forme de questions (avec une des entités cachées), et les articles sous forme de passages correspondants à partir desquels le système doit répondre à la question de type remplir le vide « fill-in the-blank ». Les auteurs ont publié les scripts qui explorent, extraient et génèrent les paires de passages et de questions à partir de ces sites web.

Au total, le corpus compte 286 817 paires d'entraînement, 13 368 paires de validation et 11 487 paires de test, telles que définies par leurs scripts. Les documents sources de l'ensemble

d'entraînement contiennent 766 mots et 29,74 phrases en moyenne, tandis que les résumés contiennent 53 mots et 3,72 phrases.

Ebmsum [225]: est un dataset de résumés multi-documents basés sur des requêtes pour le domaine médical, où les résumés se concentrent sur les réponses et les preuves cliniques liées à des questions médicales spécifiques posées par un médecin généraliste ou un médecin de famille. Les données proviennent de la section « Clinical Inquiries », accessible au public, du Journal « Family Practice »[6]. Les données contiennent la question clinique, des résumés de la réponse à plusieurs niveaux de détail, et des informations supplémentaires telles que la qualité de la preuve et les références aux articles publiés pertinents. L'annotation des données a été réalisée par le biais d'une série de méthodes d'annotation comprenant l'extraction automatique des données de la source, l'annotation manuelle et la reformulation du texte.

	CNN/Daily Mail	**Ebmsum**
Domaine	Général	Biomédical
Nombre d'instances	311 672	10 mille
Longueur moyenne des documents (en jetons)	766	536
Longueur moyenne des résumés (en jetons)	53	38
Source des données	Sites web CNN et Daily Mail	Journal « Family Practice »
Type de construction	Automatique (technique cloze)	Manuelle/ Automatique

Table 29 Statistiques sur les datasets CNN/Daily Mail et Ebmsum

4.4.2 Architecture

Comme déjà mentionné, nous avons utilisé le modèle de génération de texte BART [204] pour les questions de type résumé. Nous passons la concaténation de la question et le passage comme une seule entrée au modèle. BART génère par la suite un résumé du texte d'entrée. Ce résumé

[6] http://www.jfponline.com

est considéré comme réponse à la question. La Figure 28 ci-dessous montre notre architecture décrite en haut.

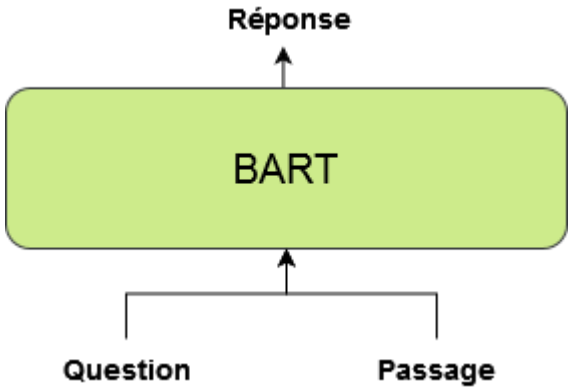

Figure 28 Architecture de notre méthode pour les questions de type résumé

Le tableau 30 ci-dessous présente les détails d'implémentation techniques de notre approche pour les questions de type résumé.

Paramètre	Valeur
Modèle de base	BART
Taille du modèle	Large
Longueur maximale de la séquence	500
Taille du lot « Batch size »	8
Taux d'apprentissage « Learning rate »	6e-5
Époques d'entraînement « Train epochs »	3
GPU	NVIDIA RTX 6000 24gb
Deep learning Framework	Pytorch 1.5.1

Table 30 Détails techniques d'implémentation (questions de type résumé)

4.4.3 Evaluation et résultats

Comme pour les autres types de questions, nous présentons ci-dessous nos résultats pour les questions de type résumé dans BioASQ 10b, auquel nous avons aussi participé.

Nous n'avons commencé à soumettre des résultats pour les questions de type résumé qu'à partir du troisième lot de test. Notre modèle a obtenu le premier score pour la métrique R-

2 (F1) dans le sixième lot. Nous avons également pu obtenir la deuxième place pour deux métriques dans le quatrième lot.

Numéro de lot	R-2 (Rec)		R-2 (F1)		R-SU4 (Rec)		R-SU4 (F1)	
	Score	Top	Score	Top	Score	Top	Score	Top
3	0.4616	0.5851	0.2877	0.3761	0.4663	0.5948	0.2735	0.3689
4	0.5458	0.5752	0.4132	0.4229	0.5482	0.5884	0.4022	0.4165
5	0.4926	0.6071	0.3500	0.4020	0.4975	0.5984	0.3423	0.3916
6	0.1782	0.1927	**0.1528**		0.2043	0.2347	0.1691	0.1705

Table 31 Résultats de notre approche pour les questions de type résumé en BioASQ 10b

Le tableau 32 ci-dessous montre un exemple d'une prédiction correcte de notre modèle pour une question de type résumé tirée du troisième lot de test du dataset BioASQ 10b

Question	What is Etizolam?
Passage	Etizolam is a thienodiazepine derivative, with high affinity for the benzodiazepine site of GABAA receptors. It is often referred to as a new psychoactive substance, a 'designer' benzodiazepine or a 'street benzodiazepine'. Etizolam is a novel psychoactive substance and novel benzodiazepine of the thienotriazolodiazepine class that has recently seen an increasing trend in use worldwide. Etizolam is a benzodiazepine analogue that is approved for use in Japan, Italy and India Etizolam is a drug from the thienotriazoldiazepine class, widely prescribed as anxiolytic due to its apparently secure toxicological profile. The prevalence of benzodiazepine consumption in Japan is one of the highest worldwide. Etizolam is the most abused drug of the benzodiazepine class. Etizolam is an anxiolytic drug with a pharmacologic profile similar to that of the classic benzodiazepines. Neurochemical research suggests that etizolam may have selectivity for the subpopulation of Y-aminobutyric acid type A receptors associated with anxiety (ie, alpha1, beta2, gamma2).
Réponse	Etizolam is an anxiolytic drug of the thienotriazolodiazepine class with a pharmacologic profile similar to that of the classic benzodiazepines. It is available in the US, Europe, Japan, Italy, and India.

Table 32 Exemple d'une prédiction correcte de notre modèle pour une question de type résumé tirée du troisième lot de test du dataset BioASQ 10b

4.5 Discussion

Malgré le fait que nos modèles pour les questions de types Oui/Non et résumé ont remporté des résultats de pointe (SOTA) dans certains lots de tests. Leur performance reste inférieure à notre modèle pour les questions de type factoïde et liste. Les raisons de cette faiblesse en performance sont d'une part semblables à ceux de notre modèle pour les questions de type factoïde te liste.

À savoir, **(1)** la différence dans la nature des questions entre les lots de tests. **(2)** La participation de différentes équipes dans différents lots de tests. **(3)** Le problème de reproductibilité des prédictions faites par les modèles d'apprentissage profond. Mais d'autre part, par le fait que nous n'avons pas pu expérimenter avec d'autres datasets plus grands que ceux que nous avons utilisés, comme MNLI [217] par exemple, à cause de nos limites matérielles, notamment en GPU. Avoir plus de ressources matérielles dans le futur pourra nous aider à tirer profit des datasets à grande échelle lors de l'apprentissage par transfert.

Une autre limitation de notre approche pour les questions de types Oui/Non et résumé, réside dans le fait qu'elle se base entièrement sur l'apprentissage par transfert. En fait, nous avons adopté les modèles BioBERT [123], et BART [204] sans apporter de changements au niveau du prétraitement de textes d'entrée, ni au niveau de l'architecture des modèles. Comme travail futur, nous comptons étendre ces deux modèles avec d'autres couches supplémentaires, ou bien agir directement sur le fonctionnement interne des modèles à l'image de notre approche pour les questions de type factoïde et liste. Ceci bien évidemment en plus de l'apprentissage par transfert.

4.6 Conclusion

Dans ce chapitre, nous avons présenté nos modèles BQA pour les questions de types Oui/Non et résumé, qui se basent sur l'apprentissage par transfert. Les modèles proposés ont obtenu des résultats de pointe (SOTA) sur plusieurs lots de test des jeux de données 10b, 9b, 8b et 7b du dataset BioASQ. Dans le chapitre suivant, nous allons présenter notre contribution à un autre volé BQA. Celui de la construction de datasets d'entraînement et évaluation des modèles.

Chapitre V : FrBMedQA : le premier dataset BQA en langue française

5.1 Introduction

Contrairement au QA, il existe un nombre limité de datasets BQA [226]. De plus, la majorité de ces datasets sont de très petite taille. Le dataset TREC Genomics Track [72], par exemple, ne comporte que 28 questions. BioASQ [4] ne comporte également que 3 mille instances de questions-réponses. Les autres jeux de données sont soit construits automatiquement, soit générés artificiellement. La construction d'un dataset BQA annoté par des humains est coûteuse et prend du temps, car l'annotation doit être effectuée par des experts médicaux. Mais ces ensembles de données sont de meilleure qualité que ceux construits automatiquement ou générés artificiellement, qui souffrent d'un taux de bruit élevé. Néanmoins, l'application de modèles statistiques ou d'apprentissage profond à la tâche de BQA nécessite un grand nombre d'échantillons d'entraînement. Par conséquent, les grands ensembles de données construits automatiquement et générés artificiellement sont aussi utiles que les petits ensembles de données annotés par des humains.

La principale technique utilisée pour construire automatiquement un dataset QA est la technique cloze [66]. Elle traduit la tâche QA en un problème de remplissage des blancs, en cachant un mot ou un ensemble de mots dans un texte, et en demandant de deviner les mots cachés en revenant au texte contextuel dans lequel ils ont été pris. Le premier dataset qui a adopté cette technique dans le domaine de QA était le children’s books dataset [227]. Les datasets « cable news network » (CNN) et Daily Mail [55] ont aussi utilisé la même technique.

Plusieurs datasets BQA ont également utilisé la même technique de cloze. Le premier est BioRead [68], suivi de « biomedical knowledge comprehension » (BMKC) [69] et de « biomedical machine reading comprehension » (BioMRC) [67]. Dans le contexte de BQA, le mot caché doit être un terme biomédical. Ces datasets ont donc utilisé des outils d'annotation biomédicale pour identifier automatiquement les termes biomédicaux.

Alors qu'il existe de nombreux jeux de données BQA en anglais, il n'y a pas de jeu de données BQA en français au moment de la rédaction de ce mémoire. Il s'agit sans doute du défi numéro un pour la BQA en français, car les datasets sont la première condition préalable à l’entraînement et à l'évaluation des systèmes QA/BQA. Comme première étape vers la résolution de ce défi, nous présentons le dataset FrBMedQA. Avec plus de 41 mille instances,

ce jeu de données a été collecté à partir d'articles biomédicaux de Wikipedia en français, puis construit de manière cloze, comme les autres jeux de données BQA mentionnés.

Afin d'évaluer la validité et la difficulté du dataset, et également comme première étape vers un classement public des modèles BQA pour la langue française, nous avons implémenté et expérimenté avec plusieurs modèles de base, un modèle de langage biomédical basé sur les réseaux de neurones, et deux modèles de langage monolingues français. Nous avons également procédé à des évaluations humaines d'un sous-ensemble de l'ensemble de tests. Nous avons mis à disposition le dataset et le code permettant de reproduire nos résultats.

Nous avons organisé le reste de ce chapitre de la manière suivante : dans la section suivante, nous donnons un aperçu des travaux connexes qui ont été réalisés sur les datasets BQA et sur le QA en français. Dans la section trois, nous décrivons en détail les aspects suivants du dataset FrBMedQA, la récupération et l'annotation du corpus, la stratégie cloze que nous avons utilisée, et nous donnons une analyse détaillée des propriétés textuelles et biomédicales du jeu de données. Dans la section quatre, nous décrivons les modèles de base ainsi que les modèles neuronaux que nous avons appliqués au dataset, en donnant également les résultats de nos expériences et en les discutant. Nous terminons le chapitre par une conclusion dans laquelle nous évoquons les futures directions de recherche après la publication du dataset.

5.2 État de l'art : approches de constructions de datasets BQA

Datasets BQA : S'il n'existe pas à ce jour des datasets BQA en français, il en existe de nombreux pour l'anglais. En 2006, « The TREC Genomics Track » [72] a inclus pour la première fois une tâche QA. Un jeu de données a été construit pour cette tâche, en collectant plus de 162 mille articles de recherche biomédicale en texte intégral. Les systèmes participants devaient extraire des passages pertinents des articles de recherche en réponse à une question, ces passages étant ensuite évalués par des juges humains. Bien que cet ensemble de données ait été un pionnier en matière de BQA, il a souffert de deux grandes limitations. Seules 28 questions étaient fournies avec les articles en texte intégral. Plus important encore, aucune instance n'a été fournie pour l'entraînement. Par conséquent, ce jeu de données ne pouvait pas être utilisé par des systèmes d'apprentissage automatique (ML) ou d'apprentissage profond (DL).

Le dataset le plus connu de BQA est BioASQ [4], contrairement à la majorité des jeux de données de BQA, BioASQ est annoté manuellement par des experts médicaux, bien qu'il ne contienne que ~3 mille instances. Ce jeu de données a été publié pour la première fois en 2015 dans le cadre d'un concours plus large de traitement du langage naturel (NLP) biomédical qui a lieu chaque année. Les questions de ce jeu de données sont de quatre types : factoïde, oui/non, liste et résumé. Le seul facteur limitant de ce jeu de données est sa petite taille.

En réponse au problème de la petite taille des jeux de données BQA, BioRead [68] a été introduit en 2018. Il s'agit actuellement du plus grand ensemble de données BQA avec ~16,4 millions d'instances. Il a été construit automatiquement à partir d'articles de recherche biomédicale en texte intégral PubMed en suivant la technique « cloze » [66]. Un nombre prédéfini de phrases est sélectionné à chaque fois comme passage, la phrase suivante servant de question. METAMAP [74] a été utilisé pour annoter les entités biomédicales trouvées dans le passage et la question, une entité de la question est ensuite masquée. La tâche consiste alors à trouver l'entité masquée parmi toutes les entités du passage comme candidats possibles. L'avantage évident de cet ensemble de données est sa taille, mais il souffre de deux limitations : alors que BioASQ, par exemple, prend en charge quatre types de questions, BioRead n'en prend en charge qu'un seul, à savoir un choix unique, en raison de la technique « cloze » adoptée. L'autre limite est le fait que de nombreuses instances de questions de passage ont été prises depuis la section des références, les légendes des figures et des tableaux, les notes de bas de page, etc.

BMKC [69] est un autre dataset BQA de type « cloze » avec ~500 mille instances de passage-question, construit à partir de résumés d'articles de recherche biomédicale PubMed. Le titre des articles a été choisi comme question, dans un autre cadre, ils ont utilisé la dernière phrase du résumé comme question. Le choix de n'utiliser que les résumés et les titres a été fait pour réduire le bruit. Comme BioRead [68], les auteurs de ce jeu de données ont automatiquement annoté les entités biomédicales dans la question et le passage, et ont masqué une entité de la question pour qu'elle soit devinée ensuite à partir des autres entités du passage.

Un autre ensemble de données annotées automatiquement est MedQuAD [78], qui contient plus de 47 mille paires question-réponse extraites et générées à partir de différents sites web médicaux fiables.

PubMedQA [79], construit à partir des résumés de PubMed et comprenant mille annotations d'experts, ~61 mille instances non étiquetées et ~211 mille instances générées artificiellement. Ce jeu de données ne prend en charge que les questions de type oui/non/peut-être.

En 2020, le dataset BioMRC [67] a été introduit comme une version améliorée de BioRead [68]. Il compte 812 mille instances de question-passage et suit la même stratégie « cloze » que son prédécesseur. Pour réduire le bruit, qui est le principal inconvénient de BioRead, les auteurs de BioMRC n'ont utilisé que les abstracts pour générer des instances de question-passage, contrairement à l'utilisation du texte intégral dans BioRead. Ils ont également utilisé les annotations d'entités biomédicales DNORM (disease named entity recognition and normalization) [81], qui sont plus précises que MetaMap utilisé dans BioRead.

QA française : À notre connaissance, il n'existe actuellement que deux dataset QA français, et aucun dataset BQA. FQuAD [228] est le premier jeu de données QA en français qui a été introduit en 2020. Il contient environ 60 mille instances de passage-question annotées manuellement. Comme pour SQuAD [5], les données ont été collectées à partir de Wikipedia en suivant la même stratégie. Pour expérimenter avec ce jeu de données, les auteurs ont appliqué deux familles de modèles, des modèles monolingues français natifs CamemBERT [229] et FlauBERT [230], et des modèles multilingues mBERT [231] et XLM-RoBERTa [232]. Les expériences ont montré que le modèle CamemBERT [229], monolingue et de langue française, était plus performant.

Peu de temps après la publication de FQuAD [228], le jeu de données PIAF [233] a été introduit. Comme FQuAD, il a été collecté à partir de Wikipedia en suivant la même stratégie. Mais contrairement à FQuAD, il ne contient que 3835 paires question-réponse. Pour expérimenter avec ce jeu de données, les auteurs du jeu de données n'ont testé que le modèle le plus performant sur FQuAD, à savoir CamemBERT, en utilisant plusieurs stratégies de réglage fin.

5.3 Construction du dataset FrBMedQA

5.3.1 Recherche et annotation de corpus

La plupart des datasets BQA utilisent PubMed comme source de leur corpus. Dans notre cas, nous n'avons pas pu l'utiliser, car PubMed ne fournit pas d'abstracts pour les articles écrits en français. D'autre part, les équivalents français de PubMed comme Lissa[7] interdisent l'utilisation ou la redistribution des abstracts. Face à ce défi, nous avons décidé d'utiliser Wikipedia comme source de notre jeu de données.

Sur les cinq millions d'articles français de Wikipédia, nous avons d'abord récupéré 243 mille articles biomédicaux, en utilisant deux stratégies de filtrage. Pour la première, nous avons récupéré tous les articles ayant un « InfoBox » biomédical. La seconde consiste à récupérer les articles ayant au moins un terme biomédical dans leur titre, et au moins dix termes biomédicaux dans leur texte. Pour ce faire, nous nous sommes appuyés sur une liste de termes biomédicaux que nous avons construite à partir de différents dictionnaires biomédicaux français. Après le filtrage, nous avons nettoyé le texte des parties bruyantes comme les références et certaines balises html, et l'avons divisé en paragraphes. Nous avons écarté les paragraphes composés de moins de trois phrases, ou contenants moins de 23 jetons. Après avoir collecté le corpus, nous avons utilisé SIFR annotator [234], un outil biomédical français de reconnaissance d'entités nommées (NER), pour annoter les entités biomédicales trouvées dans le corpus. Afin de limiter l'annotation aux termes biomédicaux, nous n'avons considéré que les entités appartenant aux groupes sémantiques de l'UMLS (Unified Medical Language System) présentés dans le tableau 33 ci-dessous.

[7] https://www.lissa.fr

Groupe sémantique	ID	Nombre d'entités	Pourcentage
Produits chimiques et médicaments	CHEM	62820	29.77%
Anatomie	ANAT	54906	26.02%
Physiologie	PHYS	30660	14.53%
Troubles	DISO	30045	14.24%
Phénomènes	PHEN	16007	7.59%
Procédures	PROC	12588	5.96%
Gènes et séquences moléculaires	GENE	3475	1.75%
Appareils	DEVI	476	0.02%

Table 33 Groupes sémantiques UMLS considérés pour l'annotation

La Figure 29, montre les pourcentages des groupes sémantiques UMLS considérés pour l'annotation des entités biomédicales.

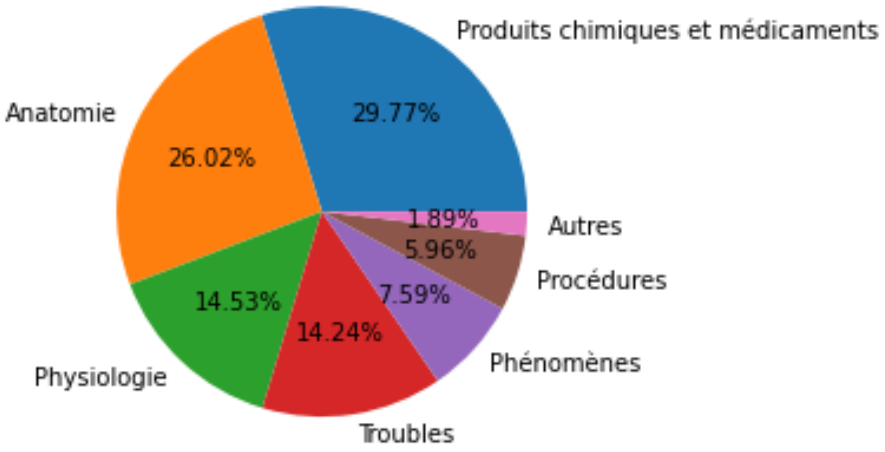

Figure 29 Pourcentages de groupes sémantiques UMLS considérés pour l'annotation

5.3.2 La génération d'instances du dataset

Une instance du dataset FrBMedQA est un tuple contenant, un passage contextuel, une question, des réponses candidates, et la réponse. La tâche consiste alors à sélectionner la bonne réponse à la question dans la liste des réponses candidates qui apparaissent également dans le passage. Pour générer les instances, nous avons utilisé la même stratégie de type cloze que celle utilisée pour construire les jeux de données CNN et Daily Mail [55], avec quelques

modifications mineures. Tout d'abord, nous avons remplacé toutes les entités biomédicales par des pseudo-jetons de la forme @entityID, où ID est un identifiant entier unique pour chaque entité biomédicale, nous commençons avec ID zéro et l'incrémentons avec chaque nouvelle entité biomédicale. D'autres ensembles de données BQA de type cloze, comme BioRead [68] et BioMRC [67], suivent deux stratégies lors du remplacement des entités biomédicales par des pseudo-jetons. La première consiste à redémarrer à partir de l'ID zéro pour chaque nouvelle instance, et la seconde à conserver le même ID pour la même entité biomédicale pour toutes les instances. Nous avons choisi de suivre uniquement la deuxième stratégie après avoir constaté qu'elle donne de meilleurs résultats pour les systèmes à base de neurones sur BioRead et BioMRC. Cela s'explique par le fait que, dans le second cas, les systèmes neuronaux sont capables d'apprendre les propriétés utiles des pseudo-jetons en s'entraînant sur plusieurs instances. L'anonymisation des termes biomédicaux en les remplaçant par des pseudo-jetons empêche les systèmes d'utiliser leurs connaissances de base et les oblige à lire et à comprendre le passage du contexte. Sans l'anonymisation, même un modèle n-gram entraîné préalablement sur le même corpus sera capable de retrouver le @placeholder manquant.

Pour générer le passage et la question, nous parcourons toutes les phrases d'un paragraphe en commençant par la première, puis nous recherchons les pseudo-jetons dans la phrase actuelle, qui apparaissent également dans le reste des phrases. Lorsque nous trouvons une correspondance, nous choisissons la phrase actuelle comme question, et le reste des phrases comme passage. Cette approche est différente des autres jeux de données de type cloze, où la question est soit la première, soit la dernière phrase du paragraphe. Avec notre approche, nous avons pu avoir plus d'instances, car dans de nombreux cas, la première ou la dernière phrase ne partagent pas d'entités biomédicales avec le reste du texte. Cependant, nous veillons à ce qu'une phrase ne soit pas choisie comme question dans d'autres passages. Ensuite, nous remplaçons le pseudo-jeton que nous avons trouvé dans la question par un placeholder de la forme @placeholder, le pseudo-jeton devient alors la réponse, et l'ensemble de tous les pseudo-jetons dans le contexte et la question deviennent les réponses candidates. Si plusieurs pseudo-jetons sont trouvés dans la question, la même opération est répétée pour chaque pseudo-jeton. Pour illustrer davantage le processus de collecte de corpus, d'annotation et de génération d'instances, l'algorithme 1 montre l'algorithme exact que nous avons utilisé.

Algorithm 1: L'algorithme global de collecte de corpus, d'annotation et de génération d'instances.

```
Input:
    data ← Données de Wikipédia en français
    infoBoxes ← liste des InfoBox biomédicales de Wikipédia
    medTerms ← liste des termes biomédicaux français
    semanticGroups ← liste des groupes sémantiques autorisés
Output: Liste d'instances (passage, question, réponses candidates, réponse) tuples
articles ← {}
For each article in data do
  if article.InfoBox in infoBoxes or article.title has one of medTerms then
    Supprimer les jetons bruyants de l'article
    paragraphs ← diviser l'article en paragraphes
    for each paragraph in paragraphs do
      annotations ← appeler le service web de l'annotateur SIFR avec le texte du paragraphe
      for each entity in annotations do
        if entity.semanticGroup in semanticGroups then
          pseudoToken ← générer un pseudo-jeton de remplacement pour l'entité
          remplacer chaque occurrence de l'entité dans le paragraphe par un pseudo-jeton
          sentences ← diviser le paragraphe en phrases
          for each sentence in sentences do
            restOfText ← sentences - sentence
            if sentence has pseudoToken and restOfTex has pseudoToken then
              question ← remplacer le pseudo-jeton dans la phrase par @placeholder
              candidateAnswers ← liste de tous les pseudo-jetons dans le paragraphe
              articles ← append: {restOfText, question, candidateAnswers, pseudoToken}
```

```
27                    end if
28                  end for
29                end if
30              end for
31            end for
          end if
        end for
```

Un exemple d'une instance aléatoire tirée du dataset est présenté dans le tableau 34, montrant le contexte, la question, les entités candidates et la réponse, avant et après l'application de l'étape d'encodage de type cloze.

Passage: Le traitement secondaire va être le traitement définitif. L'enfant est mis sous traitement antibiotique avant l'opération et il le continuera 48 heures après qu'elle sera passée. Ce traitement permet de prévenir toute infection, car c'est la plus grosse complication possible. Elle peut être due à l'entérocolite ou à une contamination par les selles. Cette chirurgie peut être pratiquée avant les 3 premiers mois de vie si nécessaire. Il existe différentes techniques pour ce traitement mais toutes ces opérations ont le même but. Elles permettent de rétablir la continuité du tube digestif après avoir effectué une ablation partielle de la partie du côlon malade. Après cette ablation de la portion pathologique du côlon, le segment de l'iléon est relié au segment du côlon qui reste, avec du fil ou des agrafes. Cette intervention n'entraîne généralement pas de conséquences sur le fonctionnement du tube digestif. Quand l'ensemble du côlon est atteint, c'est l'iléon normalement innervé qui doit être amené au niveau du rectum ou même de l'anus. Une technique alternative consiste en une	**Passage**: Le traitement secondaire va être le traitement définitif. L'enfant est mis sous traitement antibiotique avant l'opération et il le continuera 48 heures après qu'elle sera passée. Ce traitement permet de prévenir toute @entity0, car c'est la plus grosse complication possible. Elle peut être due à l'@entity1 ou à une contamination par les selles. Cette chirurgie peut être pratiquée avant les 3 premiers mois de vie si nécessaire. Il existe différentes techniques pour ce traitement mais toutes ces opérations ont le même but. Elles permettent de rétablir la continuité du @entity2 après avoir effectué une ablation partielle de la partie du @entity3 malade. Après cette ablation de la portion pathologique du @entity3, le segment de l'iléon est relié au segment du @entity3 qui reste, avec du fil ou des agrafes. Cette intervention n'entraîne généralement pas de conséquences sur le fonctionnement du @entity2. Quand l'ensemble du @entity3 est atteint, c'est l'iléon normalement innervé qui doit être amené au niveau du @entity4 ou même de l'anus. Une technique alternative consiste en une résection transanale du colon distal avec des résultats

résection transanale du colon distal avec des résultats comparables voire supérieurs à la technique classique	comparables voire supérieurs à la technique classique
Question: Autrement dit, le but recherché est de supprimer les zones intestinales ne contenant plus de cellules neuro-ganglionnaires, et de relier les intestins qui fonctionnent normalement à la partie terminale du tube digestif, c'est-à-dire le rectum, si celui-ci possède ces cellules, sinon à l'anus	**Question**: Autrement dit, le but recherché est de supprimer les zones intestinales ne contenant plus de @entity5 neuro-ganglionnaires, et de relier les @entity6 qui fonctionnent normalement à la partie terminale du @entity2, c'est-à-dire le @placeholder, si celui-ci possède ces @entity5, sinon à l'anus
Entités candidates: Infection, Entérocolite, Tube digestif, Côlon, Rectum, Cellules, Intestins	**Entités candidates**: @entity0: Infection @entity1: Entérocolite @entity2: Tube digestif @entity3: Côlon @entity4: Rectum @entity5: Cellules @entity6: Intestins
Réponse: Rectum	**Réponse**: @entity4: Rectum

Table 34 Exemple d'instance montrant le contexte, la question, les entités candidates et la réponse, avant et après l'application de l'étape d'encodage de style cloze.

5.4 Analyse du dataset :

Le dataset est divisé en trois ensembles : 80 % pour l'entraînement, 10 % pour la validation et la 10 % restante pour le test (comme illustré dans la Figure 30). Le tableau 35 montre différentes statistiques sur le dataset, comme le nombre d'instances dans chaque ensemble. Il indique également la longueur moyenne, maximale et minimale de la question, du contexte et des réponses candidates.

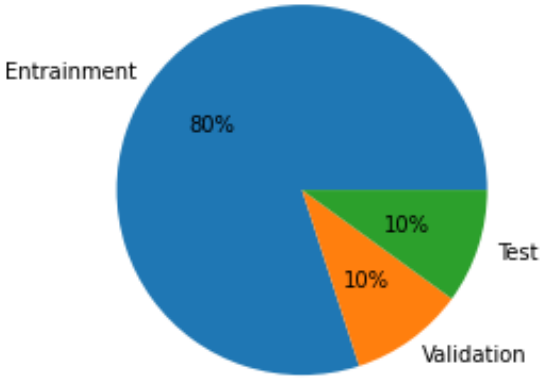

Figure 30 Distribution des instances du datset

	Entrainment	Validation	Test	Total
Nombre d'Instances	32888	4111	4110	41109
Nombre moyen d'entités candidates	4.81	4.89	4.73	4.81
Nombre max d'entités candidates	42	38	28	42
Nombre min d'entités candidates	2	2	2	2
Longueur moyenne du passage	111.02	111.44	109.94	110.95
Longueur max du passage	807	715	715	807
Longueur min du passage	19	19	19	19
Longueur moyenne de la question	38.54	39	38.32	38.56
Longueur max de la question	685	587	542	685
Longueur min de la question	5	5	5	5

Table 35 Statistiques sur le dataset FrBMedQA (longueur en jetons)

Dans la prochaine mise à jour du dataset, nous prévoyons d'augmenter la taille du jeu de données, qui est actuellement de plus de 41 mille instances. Nous prévoyons également de diminuer la différence entre le nombre maximal et minimal de réponses candidates. Étudier la corrélation entre les résultats de performance et la longueur des passages et des questions, afin de pouvoir choisir les limites de longueur maximale et minimale les plus adéquates. La réalisation de ces optimisations sur l'état actuel du jeu de données entraînera une forte diminution de sa taille. Nous prévoyons de rassembler davantage de données biomédicales françaises afin de pouvoir écarter les instances ayant certaines propriétés spécifiques.

La Figure 31 montre la distribution de la longueur du contexte (passage) (a), et de la longueur de la question (b). La majorité des instances sont regroupées autour de la valeur moyenne de 110,95 pour la longueur du contexte, et de 38,56 pour la longueur de la question. Le fait de disposer de plus de données à l'avenir va nous aider à écarter les valeurs aberrantes.

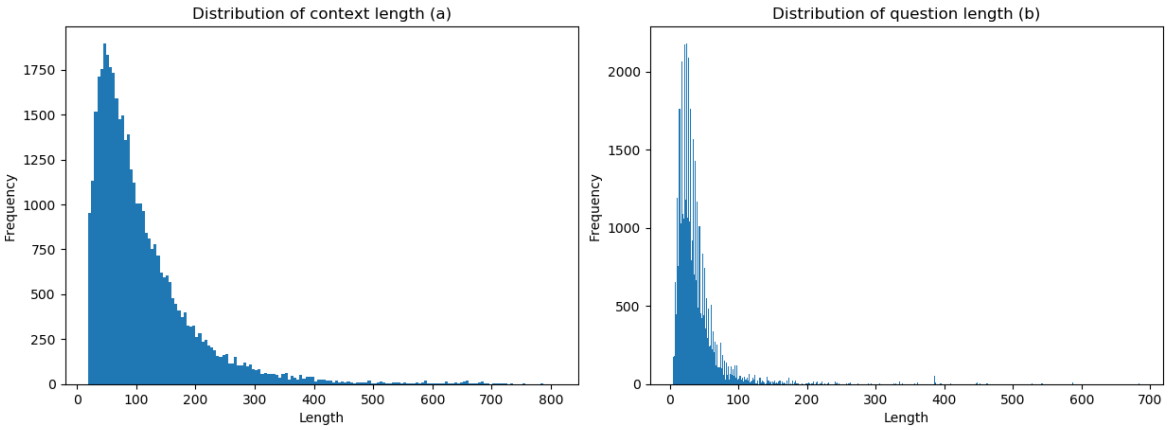

Figure 31 La distribution de la longueur du contexte (a), et la distribution de la longueur de la question (b)

La Figure 32 montre la distribution du nombre d'entités candidates (a), et la distribution des entités biomédicales par groupe sémantique UMLS (b). La majorité des instances ont entre deux et quatre entités candidates. Un système qui choisit au hasard une entité comme réponse à partir de la liste des entités candidates peut en fait obtenir de bons résultats. Pour surmonter cette limitation, nous avons l'intention d'éliminer les instances ayant moins de quatre entités uniques lors de la prochaine mise à jour du dataset. Quant à la distribution des entités biomédicales par groupe sémantique UMLS, nous pensons qu'elle reflète correctement la réalité des corpus biomédicaux.

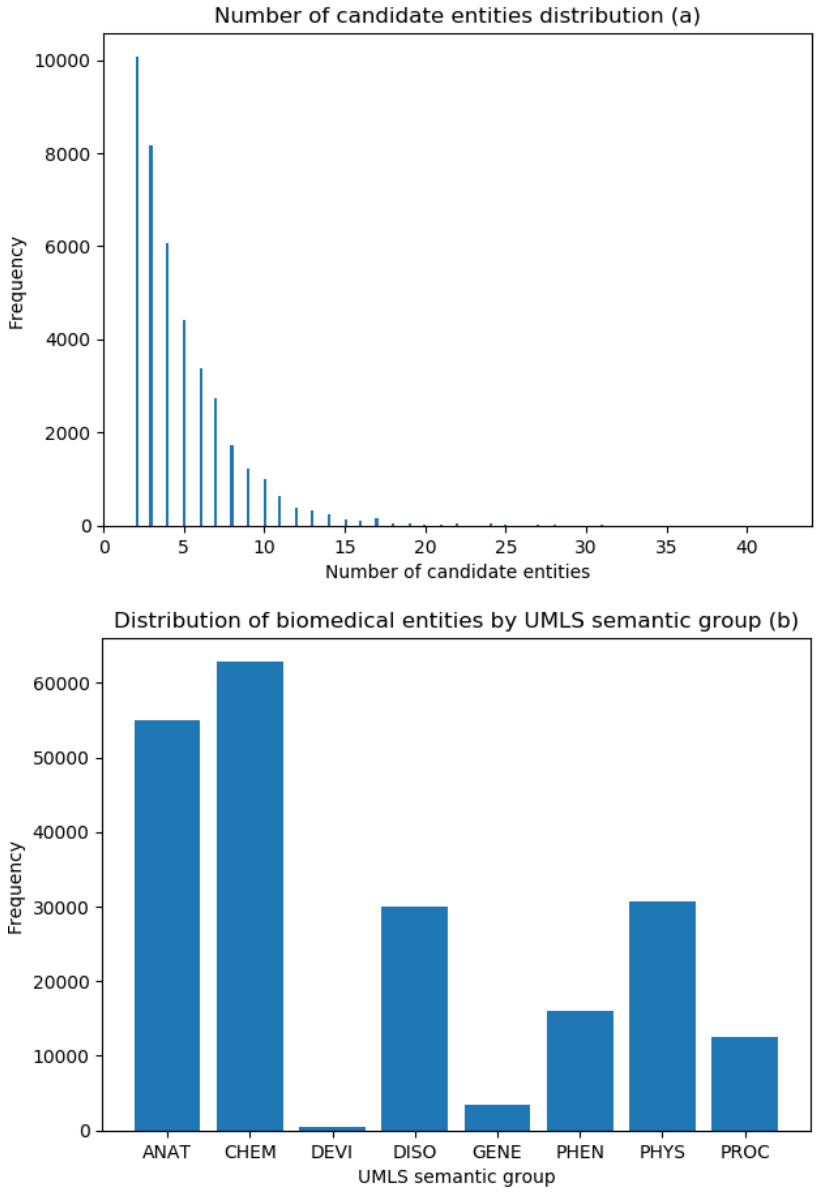

Figure 32 La distribution du nombre d'entités candidates (a), et la distribution des entités biomédicales par groupe sémantique UMLS (b)

5.5 Evaluation et discussion

Afin de tester la validité et la difficulté du dataset, et aussi pour offrir une première étape vers un classement public des modèles BQA français, nous avons implémenté et expérimenté avec trois modèles de base déjà testés sur BioMRC [67], plus un autre modèle de base. Nous avons également implémenté et testé trois modèles neuronaux basés sur BERT [1]. En outre, nous

avons procédé à une évaluation humaine sur un sous-ensemble de l'ensemble de tests. Les modèles que nous avons expérimentés avec sont les suivants.

Baseline 1 : sélection aléatoire d'une entité dans la liste des entités candidates. Le tableau des résultats de l'expérience montre la moyenne de trois essais.

Baseline 2 : renvoie l'entité (@entityID) qui apparaît le plus dans le passage et la question, de raison que cette entité est plus susceptible d'avoir été convertie en @placeholder.

Baseline 3 : renvoie l'entité qui apparaît en premier dans le passage. L'entité apparaissant en premier est fort probable le point central du passage, donc plus susceptible d'avoir été répétée dans la question et convertie en @placeholder.

Baseline 4 : ici, nous commençons par extraire tous les jetons n-gram (n = 2) de la question qui contient le jeton @placeholder. Ensuite, nous parcourons la liste des réponses candidates, en remplaçant le jeton @placeholder par chaque réponse candidate pour tous les n-grammes extraits, et en comptant le nombre d'occurrences du n-gramme résultant. La réponse candidate donnant le plus grand nombre d'occurrences est alors retournée comme réponse.

SciBERT [124]: un modèle de langage basé sur BERT et pré-entraîné sur des corpus scientifiques biomédicaux. Nous avons utilisé la même implémentation que celle utilisée par BioMRC.

CamemBERT [229]: un modèle de langage basé sur BERT et pré-entraîné sur des corpus textuels français. Nous avons utilisé la même implémentation que SciBERT, les deux modèles étant basés sur BERT.

FlauBERT [230]: un autre modèle de langue pré-entraîné basé sur BERT pour le français.

Performance humaine : nous avons sélectionné au hasard 30 instances de l'ensemble de test, après avoir retiré les réponses des instances, nous les avons données à trois participants humains non-experts sans connaissances biomédicales. Ils ont ensuite reçu comme instruction de choisir une entité de la liste des entités candidates comme réponse, après avoir lu le passage et la question, même en cas d'incertitude. La précision moyenne des trois participants est indiquée

dans le tableau des résultats de l'expérience. Le tableau 36, énumère les résultats obtenus avec chaque modèle, en plus de la précision humaine.

Modèle	Précision
Baseline 1	44.69
Baseline 2	52.72
Baseline 3	46.57
Baseline 4	41.85
SciBERT	46.86
CamemBERT	44.95
FlauBERT	44.73
Performance humaine	**61.11**

Table 36 Résultats des expériences

La Figure 33 montre une comparaison entre la précision des deux meilleurs modèles et la performance humaine.

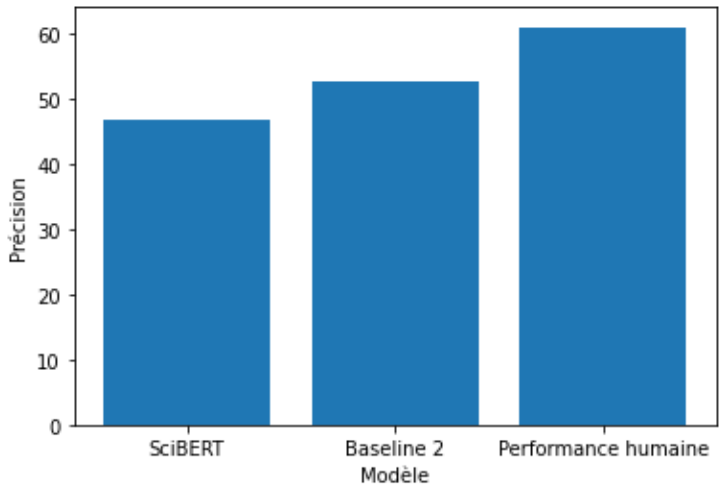

Figure 33 Comparaison entre la précision des deux meilleurs modèles et la performance humaine

Le modèle le moins performant est le modèle Baseline 4. Le deuxième modèle le moins performant est Baseline 1, mais, étant donné que ce modèle sélectionne au hasard une entité parmi les entités candidates comme réponse, nous pouvons dire que ce modèle est étonnamment performant. En revenant à la Figure 32, nous pouvons voir que la majorité des instances n'ont

qu'entre deux et quatre réponses candidates, la majorité en ayant deux. Ceci explique la performance surprenante du modèle Baseline 1. Dans la prochaine mise à jour du dataset, nous prévoyons de recueillir davantage de données, afin de pouvoir écarter les paragraphes contenant moins de quatre entités biomédicales uniques. Les deux modèles de langue française, CamemBERT et FlauBERT, ne présentent pratiquement aucune différence avec le modèle Baseline 1. Cela confirme ce que nous avons dit sur la nécessité de disposer de datasets BQA en français pour pouvoir faire progresser la recherche et les performances dans le domaine de BQA en français, car même les modèles de langage monolingues pré-entraînés en français n'ont pas pu dépasser le modèle de base le plus performant. Le Baseline 3 et SciBERT partagent pratiquement le même score. Le fait que SciBERT ne dépasse pas le modèle de base le plus performant ne nous a pas surpris, car il n'a été entraîné qu'avec des corpus textuels biomédicaux anglais. Ce fait souligne la nécessité de disposer d'un modèle de langue biomédicale français dédié, entraîné uniquement ou conjointement sur des corpus biomédicaux français. Le modèle le plus performant est baseline 2, qui renvoie simplement l'entité la plus fréquente dans le passage et la question. Dans la prochaine mise à jour du dataset, lorsque nous disposerons de plus de données, nous prévoyons d'éliminer les instances où l'entité la plus fréquente dans le passage est également la réponse. Enfin, avec 61,11%, la performance humaine non-experte est en tête du classement avec plus de 8% d'écart du modèle Baseline 2. Ce qui suggère qu'il y a une grande marge d'amélioration pour atteindre et dépasser la performance humaine non experte. Les résultats globaux montrent qu'il y a beaucoup de travail à faire dans le domaine de BQA français, qu'il s'agisse d'avoir plus de jeux de données ou d'avoir des modèles de langage biomédical français.

5.6 Conclusion

Nous avons construit et mis à la disposition du public le premier dataset français de réponses à des questions biomédicales, contenant plus de 41 mille instances question-passage. Le corpus a été collecté à partir des articles biomédicaux français de Wikipedia, les passages et les questions ont été générés de manière « cloze ». Dans un premier temps, nous avons appliqué et expérimenté avec quatre modèles de base, trois modèles neuronaux, un modèle de langue biomédicale et deux modèles de langue française. Aucun modèle basé sur les réseaux de neurones n'a pu dépasser le modèle de base le plus performant, ce qui suggère que pour relever

le défi de BQA en français, nous avons besoin de modèles de langue biomédicaux en français entrainé sur un corpus biomédical français. Avec la publication de ce dataset et du tableau de classement « leaderboard »[8], nous espérons voir apparaître des modèles de langage biomédicaux français ou multilingues dans le futur.

[8] https://apps.ump.ma/FrBMedQA

Chapitre VI :
Couplage d’un modèle BQA avec un moteur de recherche d'information IR

6.1 Introduction

Chaque jour, en moyenne, plus de trois mille articles de recherche biomédicale sont ajoutés au site PubMed. La recherche d'informations dans cette vaste littérature est aujourd'hui plus difficile et plus longue que jamais. En utilisant les moteurs de recherche biomédicaux existants, les chercheurs biomédicaux et les professionnels de la santé passent de plus en plus de temps par requête de recherche. Une étude sur les pratiques de recherche d'information des chercheurs biomédicaux et des professionnels de la santé [3] a révélé qu'il faut en moyenne trois minutes pour évaluer la pertinence d'un document renvoyé par le moteur de recherche.

Avec la propagation de la pandémie de COVID-19, ce problème de temps passé par requête de recherche dans les moteurs de recherche biomédicaux existants est désormais plus pressant. En effet, d'une part, le nombre d'articles de recherche biomédicale publiés quotidiennement a fortement augmenté depuis le début de la pandémie. D'autre part, les chercheurs biomédicaux et les professionnels de la santé sont pressés par le temps pour trouver des traitements et des vaccins efficaces, ce qui augmente de façon considérable la recherche dans la littérature scientifique liée au COVID-19.

Compte tenu de l'importance et de l'urgence de ce défi, la recherche sur la fouille de textes biomédicaux [65], la recherche d'information (RI) et la réponse aux questions biomédicales (QA biomédicale) a suscité une attention croissante de la part de la communauté de recherche NLP. Cette attention a donné naissance à de nombreux datasets et défis IR et QA biomédicales [226], comme le célèbre challenge BioASQ [4], et à un grand nombre de systèmes, de modèles, d'architectures et de techniques spécifiques au IR et BQA [78, 95, 69, 79, 174]. L’un des axes les plus recherchés dans les dernières années dans le domaine de recherche d’information est comment coupler efficacement un moteur IR classique avec un modèle QA neuronal. En fait, un moteur de recherche IR classique se contente seulement de retourner les documents pertinents pour une question donnée. Souvent exprimée en mots-clés. Alors qu’un modèle QA/BQA retourne directement la réponse à une question donnée. Souvent exprimée en langue naturelle. D’où vient l’intérêt de coupler les moteurs IR classiques avec les modèles QA/BQA. Afin de tirer avantage des performances et particularités offertes par les deux approches IR et QA.

Dans ce chapitre, nous présentons la méthode que nous proposons pour coupler un modèle BQA avec un système IR afin de former un moteur de recherche biomédical, basé sur des modèles IR et BQA de pointe. Pour choisir l'architecture de notre moteur de recherche, nous avons étudié les trois principales architectures utilisées dans les moteurs de recherche biomédicaux existants, à savoir l'architecture basée seulement sur un moteur IR classique, l'architecture moteur RI + modèle QA, et les architectures neuronales de bout en bout. Ensuite, nous avons exécuté différents benchmarks pour évaluer les algorithmes IR et les modèles BQA les plus précis et les plus pratiques à utiliser dans le moteur de recherche que nous proposons.

Avec l'arrivé de la pandémie de COVID-19, nous avons décidé d'appliquer notre approche de couplage moteur IR classique avec modèle QA pour construire un moteur de recherche intelligent spécifique aux questions en relation avec la maladie de COVID-19. L'objectif de ce moteur de recherche intelligent baptisé « INKAD COVID-19 IntelliSearch » est d'aider les chercheurs Marocains à suivre l'évolution de la recherche sur le COVID-19, les chercheurs qui seront directement concernés sont les biologistes et les médecins particulièrement les virologues et les infectiologues, mais aussi les chercheurs en sciences sociales. Ce moteur de recherche s'inscrit dans le cadre de notre réponse à un appel à projets en relation avec COVID-19 lancé par le Centre National pour la Recherche Scientifique et Technique (CNRST). La deuxième partie de ce chapitre est consacrée à la présentation de « INKAD COVID-19 IntelliSearch ».

6.1.1 Les moteurs de recherche d'information IR

La recherche d'informations « Information Retrieval » (IR) est la science de la recherche d'informations dans les bases de données relationnelles, les documents, les textes, les fichiers multimédia et le World Wide Web. Mooers [235] a défini la IR comme suit : "La recherche d'informations est le nom du processus ou de la méthode par lequel un utilisateur potentiel d'informations est capable de convertir son besoin d'informations en une liste réelle de citations de documents stockés contenant des informations utiles pour lui." Les systèmes IR sont couramment utilisés dans divers contextes, notamment dans les bibliothèques, les moteurs de recherche et les bases de données. Dans une bibliothèque, les systèmes IR peuvent être utilisés pour aider les utilisateurs à trouver des livres ou d'autres documents dans la collection de la

bibliothèque. Sur le web, les moteurs de recherche utilisent des techniques IR pour indexer et rechercher parmi des milliards de pages web afin de fournir aux utilisateurs des résultats de recherche pertinents. Dans un contexte de base de données, les systèmes IR peuvent être utilisés pour effectuer des recherches dans de grandes collections de données afin de trouver des enregistrements ou des informations spécifiques.

La Figure 34 montre l'architecture globale des systèmes IR. Dans cette figure, l'utilisateur qui a besoin d'informations émet une requête (requête utilisateur) au système de recherche par le biais du module d'opérations de requête. Le module de recherche utilise l'index des documents pour récupérer les documents qui contiennent certains termes de la requête (ces documents sont susceptibles d'être pertinents pour la requête), calcule des scores de pertinence pour eux, puis classe les documents récupérés en fonction des scores. Les documents classés sont ensuite présentés à l'utilisateur. La collection de documents est également appelée base de données textuelle, qui est indexée par l'indexeur pour une recherche efficace. L'objectif de tout système IR est de produire une liste de documents pertinents pour le besoin d'information de l'utilisateur ou la requête fournie par celui-ci. L'utilisateur a généralement besoin de documents pertinents même si les termes exacts qu'il a utilisés dans sa requête ne sont pas présents dans ces documents.

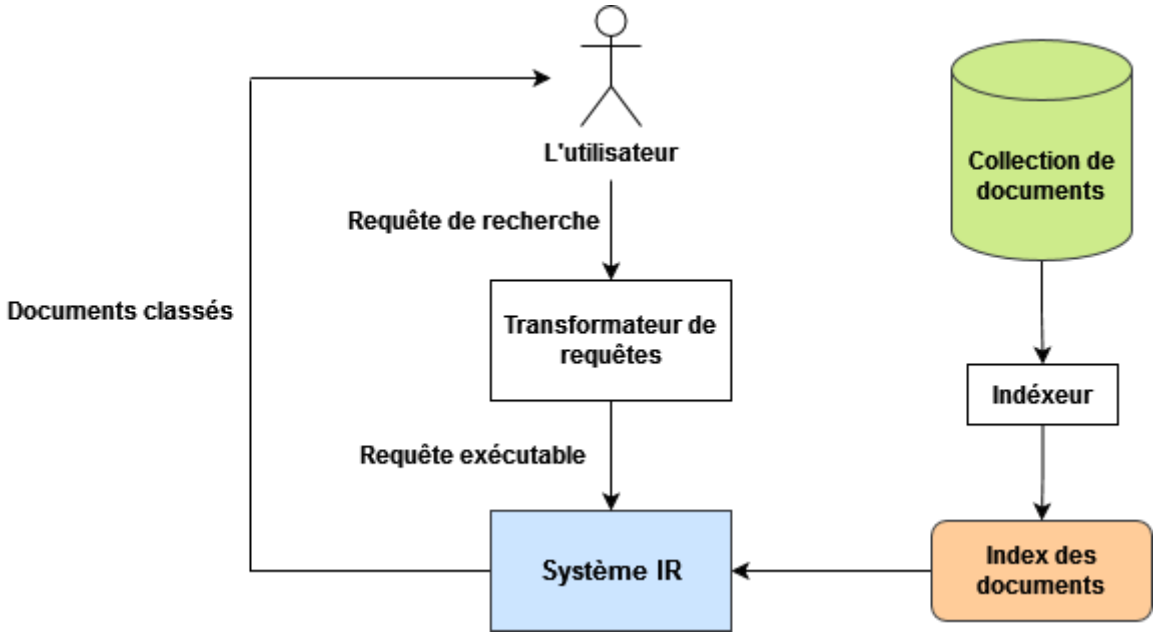

Figure 34 L'architecture globale des systèmes IR

De nombreux facteurs peuvent affecter l'efficacité d'un système IR, notamment la qualité de l'algorithme de recherche, l'exactitude et la pertinence des résultats de recherche, ainsi que la capacité de l'utilisateur à formuler une requête efficace. Par conséquent, le domaine de la IR est en constante évolution, les chercheurs et les praticiens s'efforçant de développer des techniques nouvelles et plus efficaces pour la recherche et l'extraction d'informations.

6.1.2 Les métriques d'évaluation en recherche d'information

Nous décrivons ici les métriques d'évaluation que nous avons utilisées dans nos expériences. Les définitions des métriques suivantes sont valables dans le contexte IR et QA. Ces métriques ont parfois d'autres définitions dans d'autres contextes.

6.1.2.1 Métriques IR

Précision : La précision est le nombre de documents pertinents retrouvés par une recherche divisé par le nombre total de documents retrouvés par cette recherche.

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \qquad (6.1)$$

Plutôt que de considérer tous les documents, la précision peut également être évaluée pour un nombre donné de documents récupérés, où le modèle est évalué en considérant uniquement les documents les plus importants. Dans ce cas, la métrique de précision est appelée précision à k ou P@K.

Recall : Le rappel est le nombre de documents pertinents récupérés par une recherche divisé par le nombre total de documents pertinents existants.

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|} \qquad (6.2)$$

Mean Average Precision : Pour un seul besoin d'information, la précision moyenne est la moyenne de la valeur de précision obtenue pour l'ensemble des documents top-k existant après la récupération de chaque document pertinent, cette valeur est ensuite moyennée sur les besoins d'information. L'équation de la précision moyenne (MAP) est la suivante.

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk}) \quad (6.3)$$

Avec, l'ensemble des documents pertinents pour un besoin d'information $q_j \in Q$ est $\{d_1, \dots, d_{mj}\}$ et R_{jk} est l'ensemble des documents de recherche classés à partir du premier résultat jusqu'au document d_k

Mean Reciprocal Rank : Le rang réciproque moyen est défini comme la moyenne des rangs inverses pour toutes les requêtes Q

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (6.4)$$

Où $rank_i$ est la position du premier document pertinent pour la requête i

6.1.2.2 Métriques QA

Précision Top-1 : la précision lorsque seule la première réponse retournée est prise en compte comme base d'évaluation.

Précision Top-5 : la précision lorsque les cinq premières réponses retournées sont prises en compte comme base d'évaluation.

Top-1 EM : le pourcentage de réponses correctes, lorsque seule la première réponse retournée est prise en compte comme base d'évaluation, et qu'il s'agit exactement de la réponse souhaitée (Exact match).

Top-5 EM : le pourcentage de réponses correctes, lorsque les cinq premières réponses retournées sont prises en compte comme base pour l'évaluation, et qu'elles sont exactement les réponses souhaitées (Exact match)

Top-1 F1 : la précision F1 lorsque seule la première réponse retournée est prise en compte comme base d'évaluation.

Top-5 F1 : la précision F1 lorsque les cinq premières réponses retournées sont prises en compte comme base d'évaluation.

6.2 État de l'art : approches de couplage d'un modèle QA/BQA avec un moteur de recherche d'information IR

Dans l'article [236], les auteurs proposent un système qui combine un module IR avec un module de compréhension de lecture basé sur SciBERT [124], affiné sur les jeux de données SQuAD 2.0 [16] et QuAC [8]. Les auteurs de l'article [237] ont travaillé sur un système qu'ils ont appelé COVIDASK qui utilise l'exploration de textes biomédicaux, la IR et du QA pour répondre aux questions en temps réel. "AskMe" [238] est un autre moteur de recherche et d'extraction.

Dans leur article [239], les auteurs présentent "CAiRE-COVID", un système QA et de résumé multi-documents en temps réel, composé d'un module de recherche de documents IR, d'un module QA basé sur le modèle HLTC-MRQA [240] et le modèle BioBERT [123], et d'un module de résumé abstrait basé sur BART [204] affiné sur CNN/DailyMail [55]. "CO-Search [241] est une autre proposition de moteur de recherche sémantique de type retriever-ranker. Dans le module d'extraction, les auteurs ont utilisé un codeur SiameseBERT qui est composé linéairement avec un vectoriseur TF-IDF, et fusionné réciproquement avec un vectoriseur BM25. Dans l'article [242], les auteurs décrivent la construction d'un moteur de recherche neuronal qu'ils ont appelé "Neural Covidex" et qui exploite les dernières architectures neuronales de classement « Ranking ». Une autre contribution est un moteur de recherche conversationnel pour COVID-19 [241], construit en adaptant le modèle Poly-encoder [244] pour la recherche d'informations à partir des « Frequently asked questions » (FAQ)

6.3 Notre approche de couplage d'un modèle BQA avec un moteur de recherche d'information IR

En général, il existe trois types d'architectures qui sont utilisées dans les moteurs de recherche biomédicaux. Dans le premier, seul un système IR est utilisé pour indexer les documents et répondre aux requêtes de recherche avec une liste de documents pertinents. Dans le second, la liste de documents pertinents qui sont retournés par le système IR est transmise à un modèle QA, qui reclasse ensuite les documents, extrait les passages pertinents, et enfin, il extrait et renvoie la réponse de la requête de recherche à partir des passages pertinents. Dans la troisième

architecture, un seul modèle neuronal est entraîné de bout en bout pour récupérer les documents pertinents et extraire les passages et les réponses.

Dans le tableau 37, nous comparons les trois types d'architectures en termes de précision et de vitesse. En raison de la nature pratique de notre moteur de recherche, nous choisissons d'adopter la deuxième architecture qui offre un bon équilibre entre la précision et la vitesse. L'architecture globale de notre approche est illustrée dans la Figure 38. Nous avons choisi d'utiliser l'algorithme BM25 [245] avec le système IR ElasticSearch[9], et le modèle BioBERT [123] comme modèle QA. Nous avons fait ces choix après avoir comparé deux algorithmes IR et plusieurs modèles QA. Les résultats de ces benchmarks sont présentés dans le reste de cette section.

Architecture	Précision	Rapidité
IR	Moins précis	Le plus rapide
IR + QA	Précis	Rapide
Bout en bout	Précis	Lent

Table 37 Comparaison des architectures des moteurs de recherche biomédicaux

Deux grandes familles d'algorithmes et de modèles sont utilisées dans les systèmes IR. Il existe des algorithmes de sac de mots (BOW) qui calculent le nombre de cooccurrences de mots entre la requête et les documents et utilisent ce nombre pour évaluer les documents pertinents. TF-IDF et BM25 sont les deux principaux algorithmes de cette famille. D'autre part, il existe des modèles neuronaux comme Dense Passage Retrieval (DPR) [246]. L'algorithme BM25 est considéré comme l'algorithme de pointe de sa famille. Le modèle DPR est le modèle de pointe de la famille des modèles basés sur les réseaux neuronaux. Par conséquent, nous avons choisi d'évaluer la précision et la rapidité de ces deux modèles sur notre système IR. Nous avons également évalué plusieurs modèles pour le système QA. Dans les deux évaluations (IR et QA), nous avons utilisé le jeu de test covid-qa [247]. Les résultats de ces évaluations sont présentés

[9] https://www.elastic.co

dans les tableaux 38 et 39 dans la suite de cette section. Les métriques d'évaluation utilisées dans ces expériences sont définies dans l'introduction.

Algorithme/modèle	Recall	Mean Avg Precision	Mean Reciprocal Rank	Temps moyen/question
BM25	0.907	0.907	0.907	0.005
Dense Passage Retrieval (DPR)	0.944	0.944	0.944	0.020

Table 38 Résultats de l'évaluation comparative des deux modèles BM25 et DPR

Comme on peut le voir dans le tableau, les performances du modèle DRP sont un peu meilleures que celles de l'algorithme BM25 ; en revanche, BM25 est 4 fois plus rapide que DRP. Étant donné le gain moyen en performance avec DRP, nous avons choisi d'utiliser BM25 comme algorithme de recherche dans notre système IR.

Modèle	Top-1 Acc.	Top-5 Acc.	Top-1 EM	Top-5 EM	Top-1 F1	Top-k F1	Temps moyen/requête
RoBERTa Base [69]	0.327	0.714	0.265	0.592	0.295	0.666	1.439
RoBERTa Base + CORD	0.163	0.755	0.122	0.510	0.143	0.634	1.447
MiniLM [246]	0.163	0.653	0.122	0.551	0.156	0.596	**0.717**
BERT Base [1]	0.143	0.612	0.143	0.531	0.143	0.568	1.335
ELECTRA Base [247]	0.245	0.714	0.204	0.571	0.225	0.667	1.351
XLM-RoBERTa Base [248]	0.306	0.694	0.306	0.612	0.306	0.657	1.830
BioBERT [123]	**0.413**	**0.867**	**0.456**	**0.713**	**0.404**	**0.718**	1.387

Table 39 Résultats de l'évaluation comparative des modèles QA

La Figure 35 montre une comparaison entre la précision top-1 des différents modèles QA utilisés dans nos expérimentations.

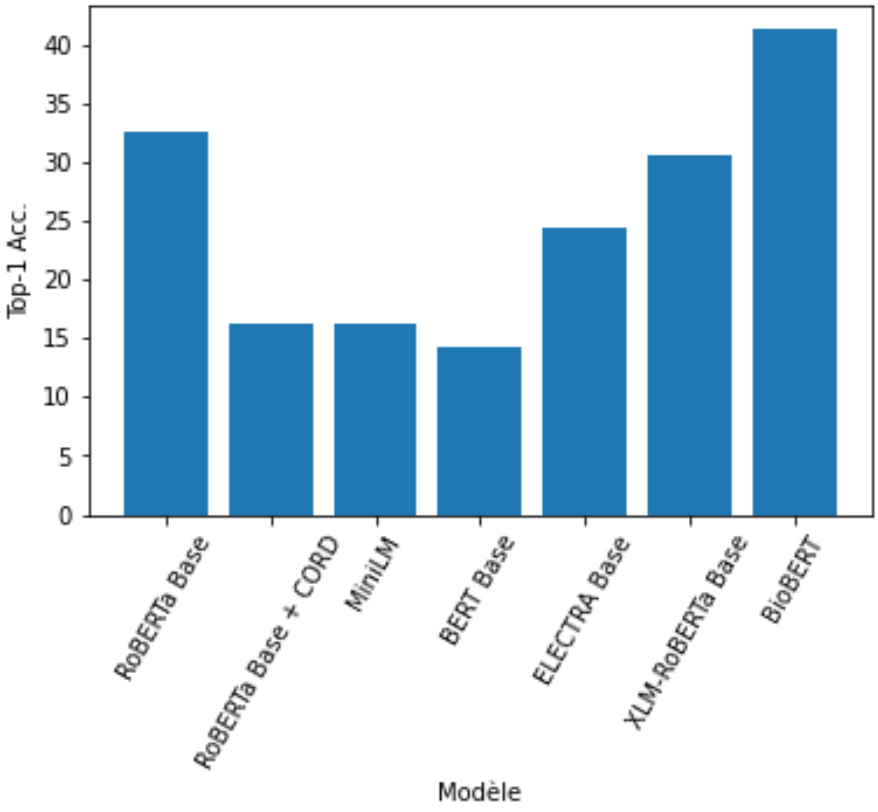

Figure 35 Comparaison entre la précision top-1 des différents modèles QA

Pour tous les modèles QA présentés dans le tableau 39, nous avons utilisé les implémentations du hub huggingface[10], tous les modèles ont également été entraînés sur le dataset SQuAD 2.0 [16] avec les hyperparamètres par défaut.

Constats :

- Le modèle "MiniLM" est le plus rapide (comme le montre la Figure 36), mais faible en précision.

[10] https://huggingface.co/models

- Le modèle BioBERT est le plus performant de tous les modèles (en calculant la somme de toutes les métriques, comme le montre la Figure 37)

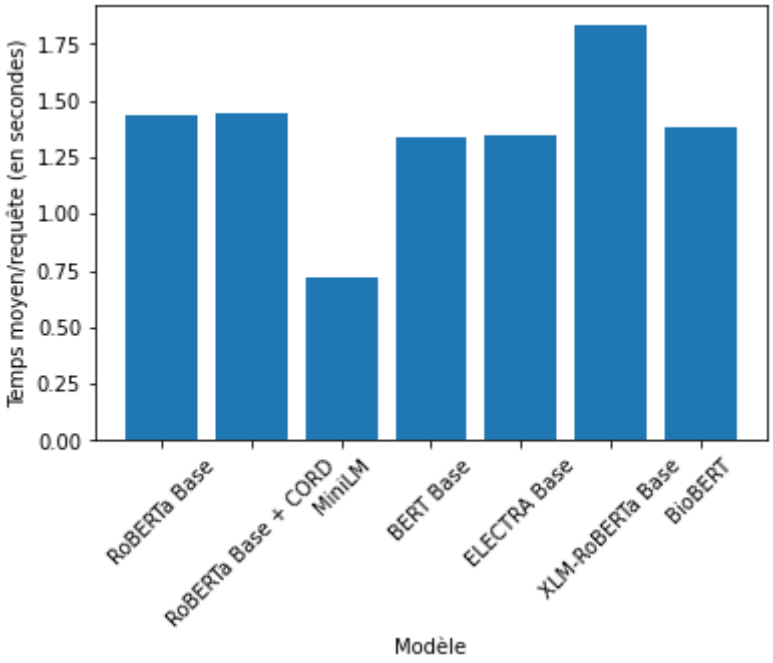

Figure 36 Comparaison entre le temps moyen par requête (en secondes) des différents modèles QA

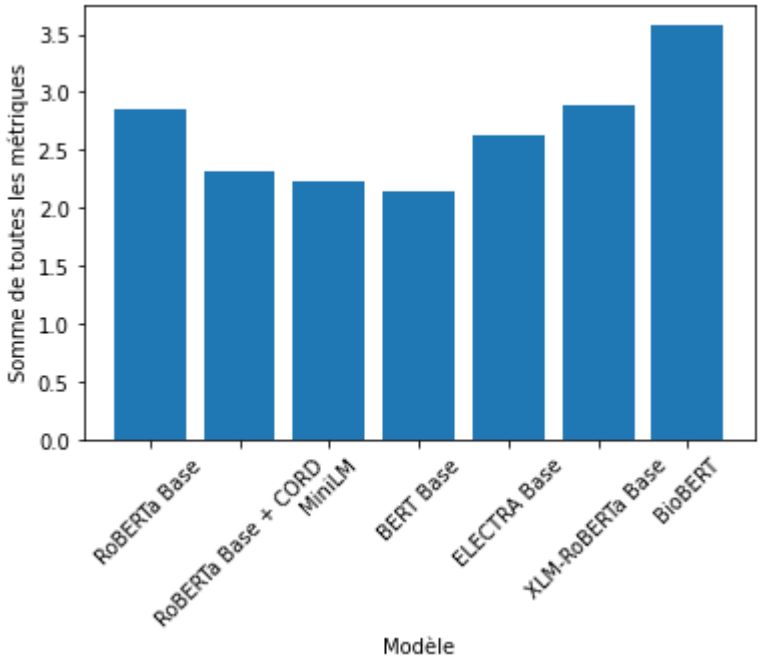

Figure 37 Comparaison entre la performance des différents modèles QA par somme de toutes les métriques

A l'issue de ces résultats, nous avons choisi d'utiliser le modèle BioBERT comme modèle du système QA dans notre démarche de couplage. Nous avons aussi utilisé un « reranker » basé sur l'encodeur croisé « ms-marco-MiniLM »[11] entraîné sur la tâche « MS Marco Passage Ranking »[12]. Nous avons implémenté notre moteur de recherche en utilisant le framework « haystack »[13]. Nous avons fixé le nombre de documents à renvoyer par le « retriever », le « reranker », et le lecteur à 200, 50 et 10 respectivement. La Figure 38 illustre l'architecture globale de notre approche de couplage système IR et modèle BQA.

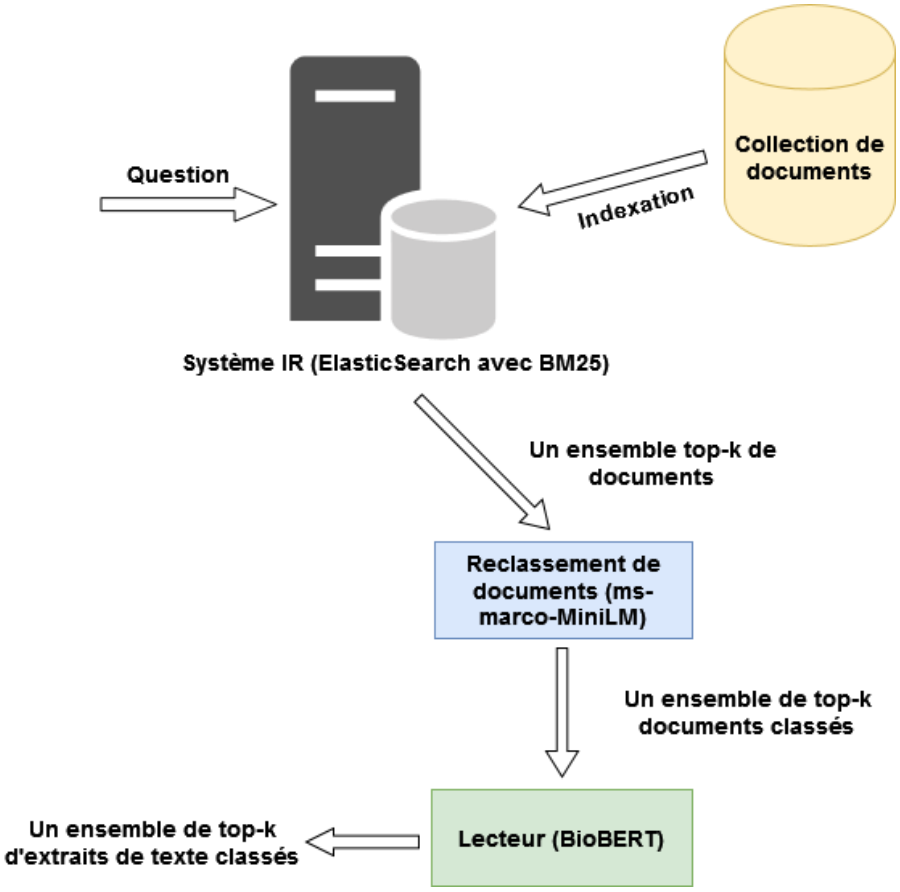

Figure 38 L'architecture globale de notre approche pour le couplage d'un système IR avec un modèle BQA

[11] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2
[12] https://microsoft.github.io/MSMARCO-Passage-Ranking
[13] https://haystack.deepset.ai

Le tableau 40 présente les détails d'implémentation techniques de notre approche de couplage système IR et modèle BQA. Le système de recherche IR utilisé est ElasticSearch, qui est un moteur de recherche open source capable de traiter de grandes quantités de données à haute vitesse. L'algorithme de recherche utilisé est BM25. Le modèle QA utilisé est BioBERT, avec le modèle de reclassement des documents Ms-marco-MiniLM. La longueur maximale de la séquence est de 250, ce qui signifie que le modèle peut traiter des séquences de 250 jetons au maximum. La taille du lot « batch size » est de 8, ce qui signifie que le modèle va traiter 8 exemples à chaque itération d'entraînement. Le taux d'apprentissage « learning rate » est de 9e-6, ce qui signifie que le modèle va ajuster les poids de ses couches en fonction de la perte avec un taux de 9e-6. Le nombre d'itérations d'entraînement « train epochs » est de 3, ce qui signifie que le modèle va être entraîné sur les données d'entraînement 3 fois. Le GPU utilisé est NVIDIA Tesla K80 24gb, qui est une carte graphique dédiée au calcul haute performance. Le framework de deep learning utilisé est HayStack, qui est un framework open source pour le développement de modèles de traitement du langage naturel. Le nombre de résultats à retourner par ElasticSearch est de 200, ce qui signifie que le système IR d'ElasticSearch à l'aide de l'algorithme BM25, va retourner les 200 documents les plus pertinents par rapport à la requête de recherche. Le nombre de résultats à retourner après le reclassement est de 50, ce qui signifie que le modèle "Ms-marco-MiniLM" va reclasser les documents trouvés par le système IR d'ElasticSearch pour retourner les 50 meilleurs documents trouvés. Le nombre de résultats à retourner par le modèle BQA est de 10, ce qui signifie que le modèle BQA BioBERT-base va lire et comprendre les 50 meilleurs documents retournés par le modèle de reclassement et retourner les 10 réponses les plus pertinentes à la question.

Paramètre	**Valeur**
Système IR	ElasticSearch
Algorithme IR	BM25
Modèle BQA	BioBERT
Modèle de reclassement (reranker)	Ms-marco-MiniLM
Taille du modèle	Base
Longueur maximale de la séquence	250

Paramètre	Valeur
Taille du lot « Batch size »	8
Taux d'apprentissage « Learning rate »	9e-6
Nombre d'itérations d'entraînement « Train epochs »	3
GPU	NVIDIA Tesla K80 24gb
Framework deep learning	HayStack
Top-k retriever	200
Top-k reranker	50
Top-k reader	10

Table 40 Détails techniques d'implémentation de notre approche de couplage d'un moteur IR avec un modèle BQA

6.4 Application sur le contexte de la pandémie de COVID-19

Depuis le début de la pandémie de COVID-19, le nombre de projets de recherche et d'initiatives dans ce domaine a considérablement augmenté. L'initiative la plus importante a été la publication du COVID-19 Open Research Dataset (CORD-19) [108] par la Maison Blanche et certaines des plus grandes entreprises technologiques et instituts de recherche. CORD-19 est une collection de plus de 400 mille articles de recherche biomédicale liés au COVID-19, au SRAS-CoV-2 et aux autres maladies de la famille coronavirus. Cette collection a été mise gratuitement à la disposition des chercheurs NLP afin qu'ils puissent appliquer les avancées les plus récentes dans les domaines de la IR et BQA. L'objectif était d'appliquer ces techniques pour trouver des réponses pertinentes aux questions les plus urgentes sur le COVID-19, comme les modes de transmission, la période d'incubation, les traitements possibles et les informations relatives aux vaccins.

La publication de CORD-19 a entraîné une vague de recherche et d'innovation dans les systèmes IR et BQA. Plusieurs groupes de recherche ont construit et publié des moteurs de recherche biomédicaux complets sur la base de CORD-19.

Dans cette deuxième partie de ce chapitre, nous présentons notre proposition de moteur de recherche biomédical multilingue appelé INKAD COVID-19 IntelliSearch, basé sur CORD-19 et utilisant notre approche de couplage système IR et modèle BQA. Ce moteur de recherche s'inscrit dans le cadre de notre réponse à un appel à projets en relation avec COVID-19 lancé

par le Centre National pour la Recherche Scientifique et Technique (CNRST). Avant de commencer à travailler sur le moteur de recherche, nous avons d'abord interrogé les chercheurs biomédicaux sur leur expérience de l'utilisation des moteurs de recherche biomédicaux existants dans des tâches de recherche liées à COVID-19. Nous avons pris en compte ce retour d'expérience lors de notre travail sur INKAD COVID-19 IntelliSearch.

6.4.1 Etude sur les pratiques de recherche scientifiques liées au COVID-19

Afin de connaître les défis et les besoins réels des chercheurs et des professionnels de la santé en termes de recherche d'information COVID-19. Nous avons créé et distribué une enquête afin de recueillir les retours de la communauté des chercheurs sur leur expérience avec les moteurs de recherche biomédicaux existants. Nous avons envoyé l'enquête aux chercheurs universitaires du Maroc travaillant dans les domaines de la médecine, de la biologie et de la virologie. Nous avons reçu des réponses de 23 chercheurs universitaires de différentes universités marocaines. Vous trouverez ci-dessous les questions incluses dans l'enquête avec les réponses associées.

Quels moteurs de recherche utilisez-vous actuellement pour effectuer vos recherches liées à COVID-19 ?

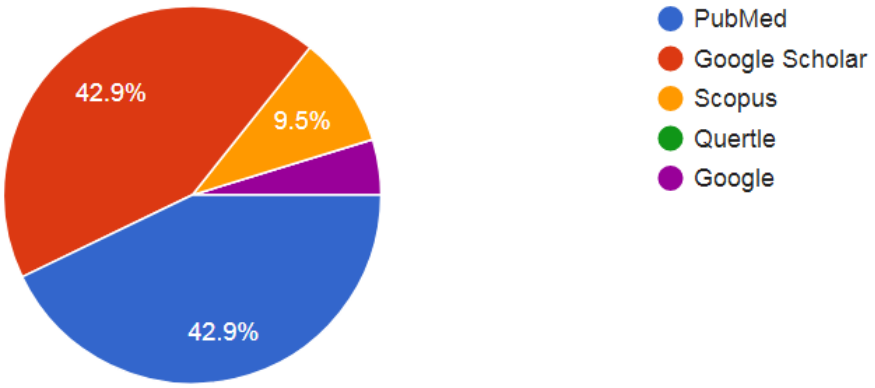

Figure 39 Réponses à la question : Quels moteurs de recherche utilisez-vous actuellement pour effectuer vos recherches liées à COVID-19 ?

Quelles langues utilisez-vous pour interroger ces moteurs de recherche ?

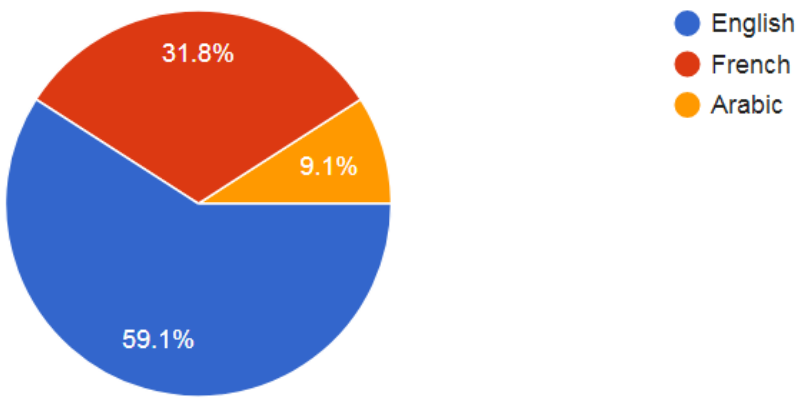

Figure 40 Réponses à la question : Quelles langues utilisez-vous pour interroger ces moteurs de recherche ?

Quels sont les problèmes que vous rencontrez dans vos recherches liées à COVID-19 ?

Synthèse des réponses les plus courantes :

- Manque de données
- La plupart du temps, des réponses non-pertinentes sont renvoyées pour des requêtes formulées en français.
- Informations imprécises et non crédibles

Quelles sont les fonctionnalités que vous aimeriez avoir dans un moteur de recherche biomédical (notamment en relation avec COVID-19) ?

Synthèse des réponses les plus courantes :

- Des réponses plus précises
- Renvoyer des réponses directes dans un délai court
- Simplicité et facilité d'utilisation
- Plus d'articles de recherche en français

Combien de temps passez-vous sur chaque article retourné par le moteur de recherche avant de trouver exactement l'information que vous recherchez ?

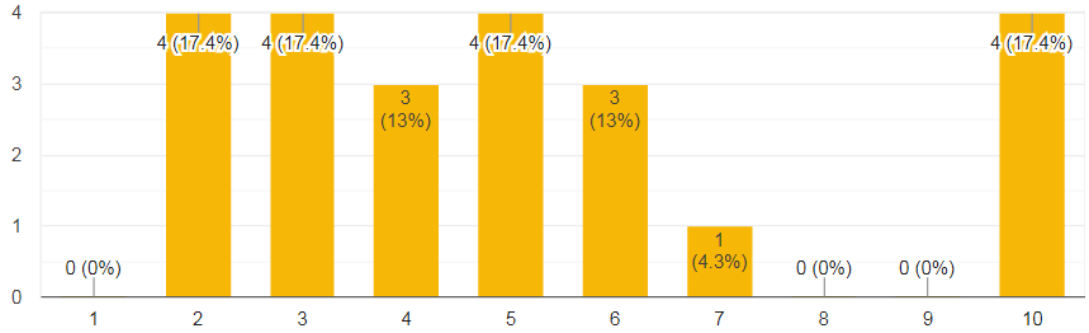

Figure 41 Réponses à la question : Combien de temps passez-vous sur chaque article retourné par le moteur de recherche avant de trouver exactement l'information que vous recherchez ? (En minutes)

Ce que l'on peut conclure des réponses issues de l'enquête est que les chercheurs et les professionnels de la santé marocains participant à cette étude, utilisent une variété de moteurs de recherche, et formulent leurs requêtes de recherche principalement en anglais, et en français. Ils passent, en moyenne, cinq minutes sur chaque document retourné par un moteur de recherche avant de trouver exactement l'information recherchée. Ils souhaitent un meilleur support du français dans les moteurs de recherche biomédicale, et des réponses plus précises et directes à leurs requêtes de recherche.

6.4.2 INKAD COVID-19 IntelliSearch

Nous avons utilisé le COVID-19 Open Research Dataset (CORD-19) [108] comme source principale d'articles scientifiques pour notre moteur de recherche, en plus des articles COVID-19 relatifs au Maroc. Nous avons indexé les articles CORD-19 dans le système IR ElasticSearch. Nous avons adopté la même architecture de couplage système IR et modèle BQA décrite dans la première partie de ce chapitre. La Figure 42 illustre l'architecture globale de notre moteur de recherche biomédical INKAD COVID-19 IntelliSearch. Le moteur de recherche est accessible en ligne sur https://apps.ump.ma/inkadsearch

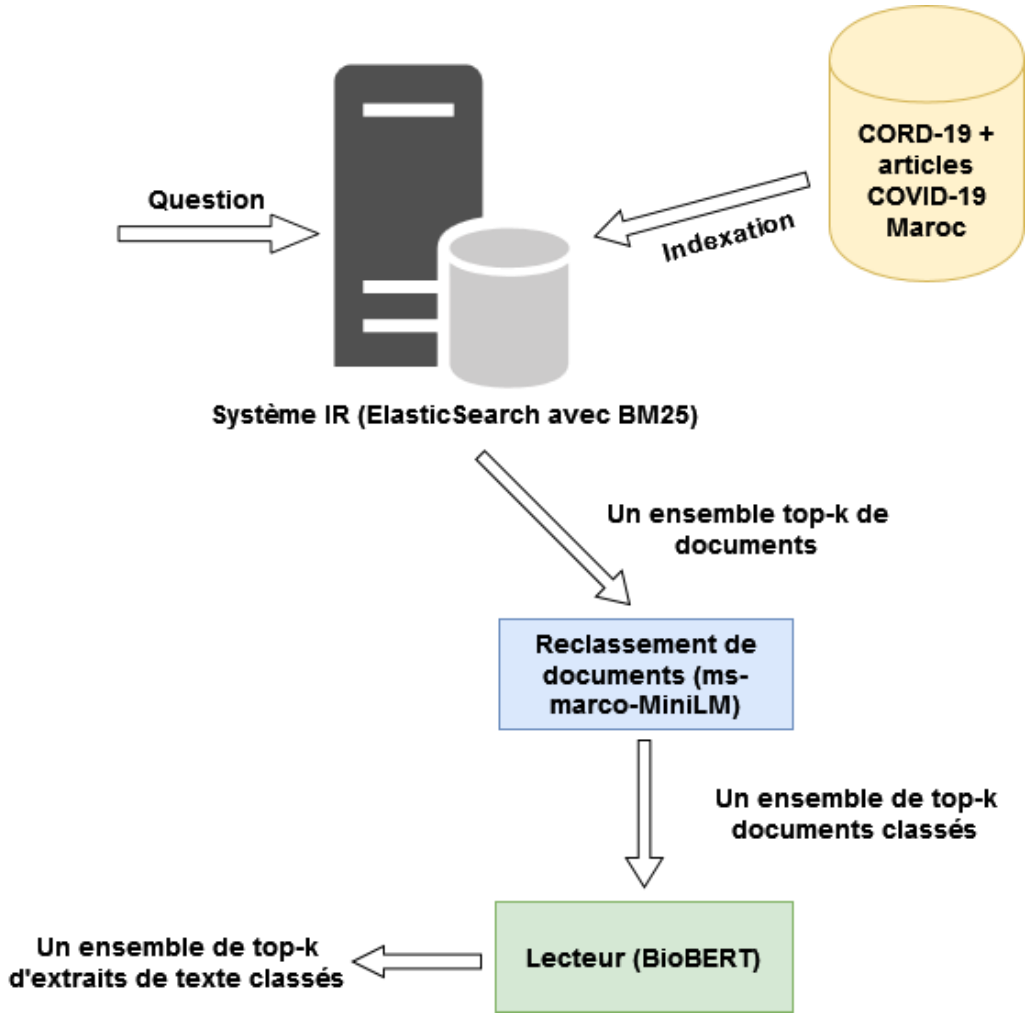

Figure 42 L'architecture globale de notre moteur de recherche biomédical INKAD COVID-19 IntelliSearch

Ci-dessous des captures d'écran de la page d'accueil du moteur de recherche INKAD COVID-19 IntelliSearch, ainsi que de la page des résultats de recherche.

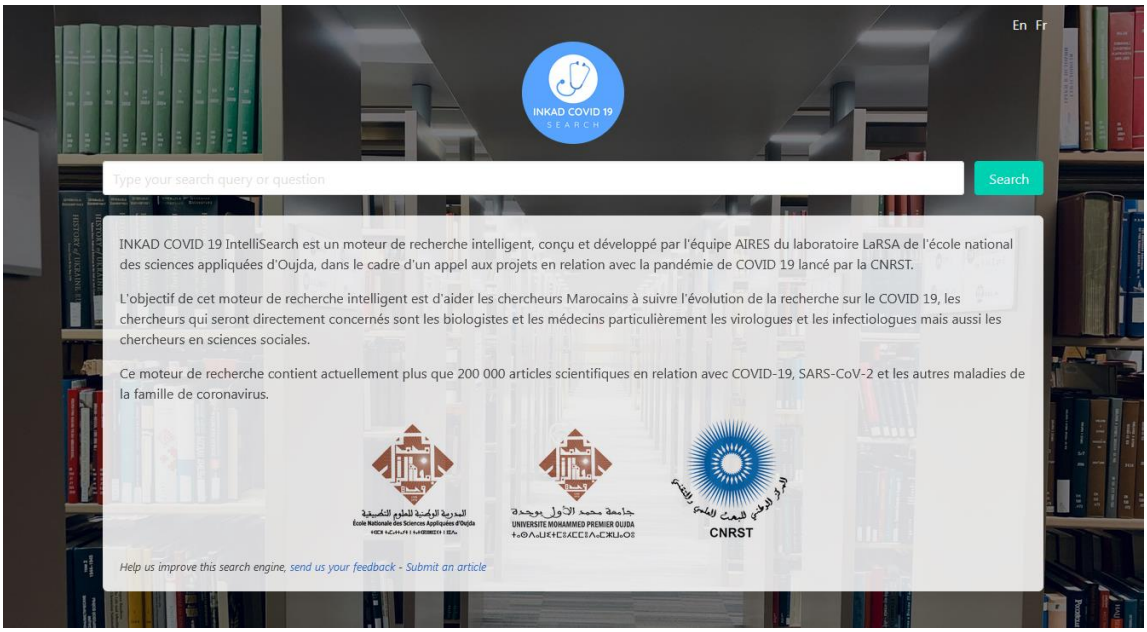

Figure 43 Page d'accueil du moteur de recherche INKAD COVID-19 IntelliSearch

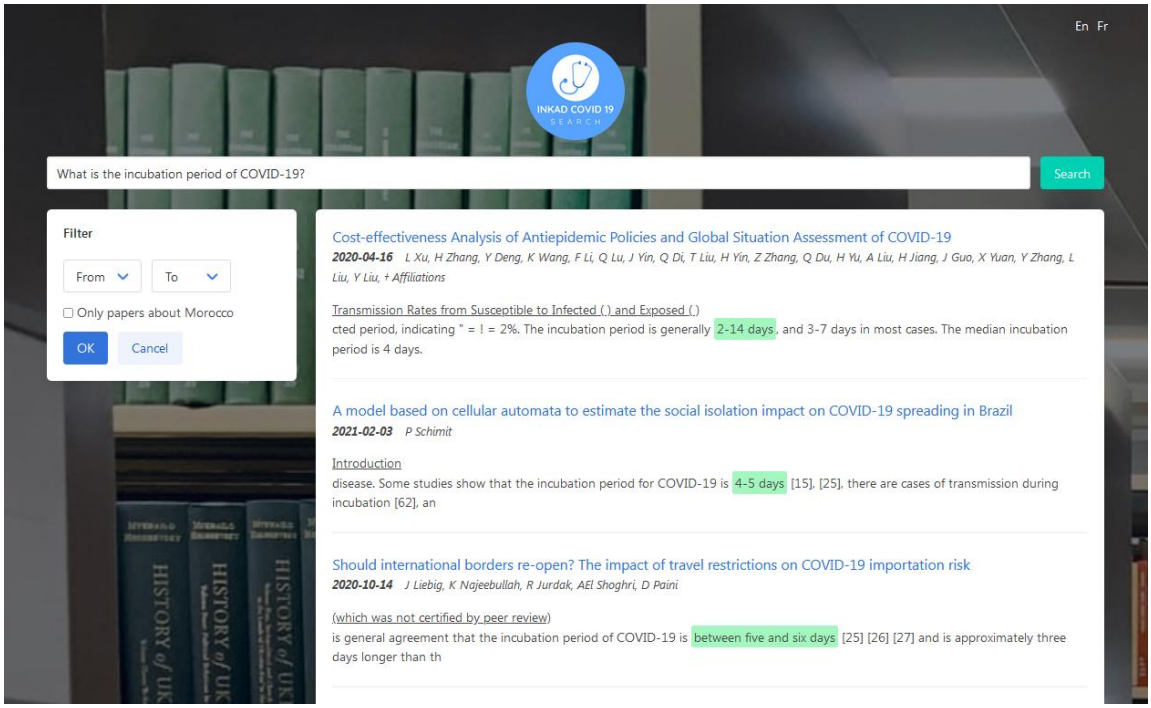

Figure 44 Page des résultats de recherche du moteur de recherche INKAD COVID-19 IntelliSearch

Comme vous pouvez le constater depuis la capture d'écran ci-dessus, en réponse à une question, le moteur de recherche INKAD COVID-19 IntelliSearch, ne se contente pas seulement de retourner la liste d'articles scientifiques contenants probablement la réponse à la question, mais

retourne aussi l’extrait de texte contenant la réponse, ainsi que la réponse elle-même, surlignée en vert.

6.5 Conclusion

Dans la première partie de ce chapitre, nous avons présenté l'architecture que nous proposons pour coupler un système IR avec un modèle BQA. Les expériences que nous avons menées ont démontré l'efficacité de l'architecture adoptée ainsi que des modèles IR et BQA que nous avons utilisés. Dans la deuxième partie du chapitre, nous avons présenté INKAD COVID-19 IntelliSearch, un moteur de recherche destiné à aider les chercheurs et les professionnels de la santé à trouver des informations précises et pertinentes sur le COVID-19 en peu de temps. Le moteur de recherche biomédical proposé est basé sur l'architecture décrite dans la première partie du chapitre. Il s'agit d'un travail en cours, nous cherchons d'autres moyens d'augmenter les performances des composants IR et QA, et d'améliorer d'autres parties du moteur de recherche. Le moteur de recherche est accessible en ligne[14] et sera mis à jour au fur et à mesure.

[14] https://apps.ump.ma/inkadsearch

Conclusion générale et perspectives

Notre problématique de recherche en cette thèse est l'amélioration des performances des systèmes QA dans le domaine biomédical. La recherche en médecine et biologie se développe rapidement, ce qui se traduit par une augmentation considérable des articles de recherche biomédicale. Trouver des informations pertinentes dans cette littérature en pleine expansion devient de plus en plus difficile pour les chercheurs et les professionnels de la santé, ce qui accroît également le fossé entre la recherche et la pratique professionnelle.

Les systèmes IR classiques tels que PubMed, bien que très utiles, renvoient toujours beaucoup plus de résultats de recherche que ce qui est idéalement souhaitable [3]. Il faut donc plus de temps pour évaluer la pertinence des documents renvoyés, puis extraire l'information requise et la synthétiser sous une forme qui peut facilement informer la prise de décision en matière de soins de santé. D'autre part, les systèmes QA ont le potentiel de surmonter les lacunes des systèmes IR classiques et de transformer positivement l'expérience de recherche. En effet, plutôt que de renvoyer des documents entiers, les systèmes QA peuvent extraire et même synthétiser des réponses précises à des questions formulées naturellement. Malheureusement, lorsqu'ils sont appliqués directement à la littérature biomédicale, ces modèles QA donnent souvent des résultats insatisfaisants.

Nous nous sommes focalisés sur trois axes de recherches qui s'inscrivent dans notre problématique. **(1)** Nouvelles approches pour la BQA, **(2)** La construction de nouveaux ensembles de données (datasets) BQA, et **(3)** Couplage d'un système IR avec un modèle BQA. Afin de répondre à ces axes de recherche, nos contributions sont :

Axe 1 – Nouvelles approches pour la BQA :

- L'introduction d'une nouvelle méthode BQA pour les questions de type factoïde et liste. Cette méthode a donné des résultats de pointe (SOTA) sur plusieurs lots des datasets 10b, 9b, 8b et 7b du challenge BioASQ [16].

- L’exploitation de l’apprentissage par transfert pour les questions de types Oui/Non et résumé.

Axe 2 – La construction de nouveaux ensembles de données (datasets) BQA :

- La construction du premier dataset BQA français
- Le lancement du premier tableau de classement public des modèles BQA français

Axe 3 – Couplage d’un système IR avec un modèle BQA :

- La proposition d’une démarche de couplage d’un moteur IR avec un modèle BQA
- La création d’un moteur de recherche biomédical spécial au COVID-19, basé sur la démarche proposée. Ce moteur de recherche s’inscrit dans le cadre de notre réponse à un appel à projets en relation avec COVID-19 lancé par le Centre National pour la Recherche Scientifique et Technique (CNRST)

Bien que nous ayons pu avoir des résultats de pointe (SOTA) avec nos différentes approches proposées, ce travail présente encore certaines limites. Tout d’abord, concernant notre nouvelle approche pour les questions BQA de type factoïde et liste. Nos gains en performances ne sont pas constants sur tous les lots de tests des datasets d’évaluation. En plus, nos résultats pour les questions de type liste sont inférieurs à nos résultats pour les questions de type facoide. Malgré le fait que nous avons utilisé la même approche. Pour les questions de types Oui/Non et résumé, la performance de notre modèle proposé est inférieure à notre modèle pour les questions de type factoïde et liste. Aussi, le fait de se baser uniquement sur l’apprentissage par transfert s’est avéré insuffisant pour avoir de bonnes performances dans ces types de questions. En ce qui concerne notre dataset BQA français. Sa seule limitation est le fait qu’il est généré automatiquement. Donc forcément d’une qualité inférieure à un dataset annoté manuellement. Tandis que pour notre démarche de couplage système IR avec un modèle BQA. Elle nécessite encore plus d’expérimentation pour trouver la configuration idéale entre les différents composants.

Comme travail futur, nous comptons travailler sur ces limites par **(1)** analyser les fausses prédictions de nos modèles dans différents lots de tests, et déterminer la corrélation exacte entre

la prédiction du modèle et les types de question/réponse (entité biomédical, entité nommée, mode d'administration médicament, procédure médicale, …). **(2)** Utiliser nos approches avec un PLM biomédical plus récent. **(3)** Augmenter nos ressources matérielles pour tirer profit des datasets à grand échelle lors de l'apprentissage par transfert. **(4)** Construire un deuxième dataset BQA français, cette fois-ci annoté manuellement. Et **(5)** continuer à expérimenter pour trouver la combinaison idéal des différents composants lors du couplage d'un système IR avec un modèle BQA.

Annexe : visualisation des scores d'attention

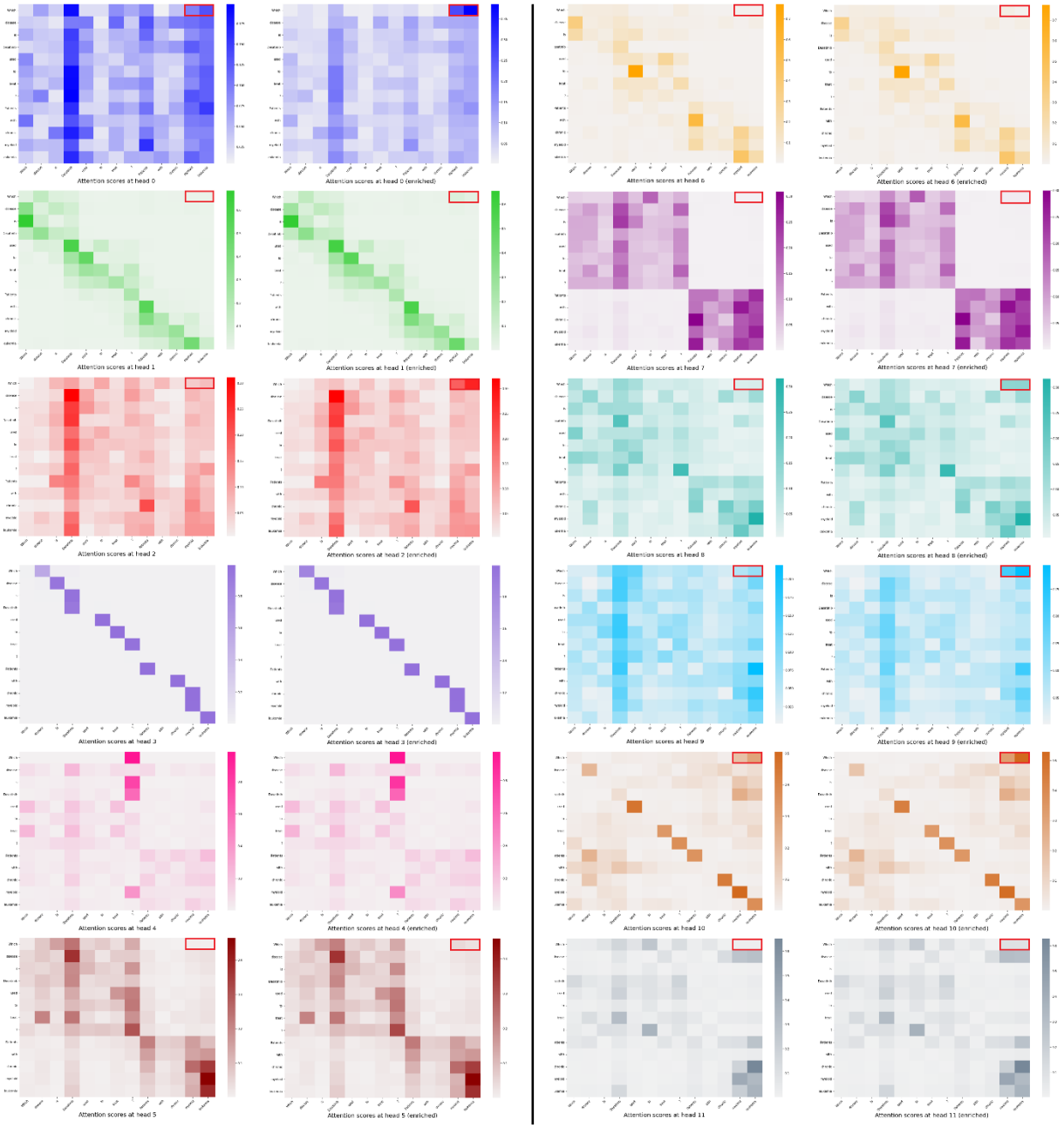

Figure 45 Scores d'attention dans l'ensemble des 12 têtes de la couche 1 avant et après l'enrichissement de l'attention biomédicale et NER

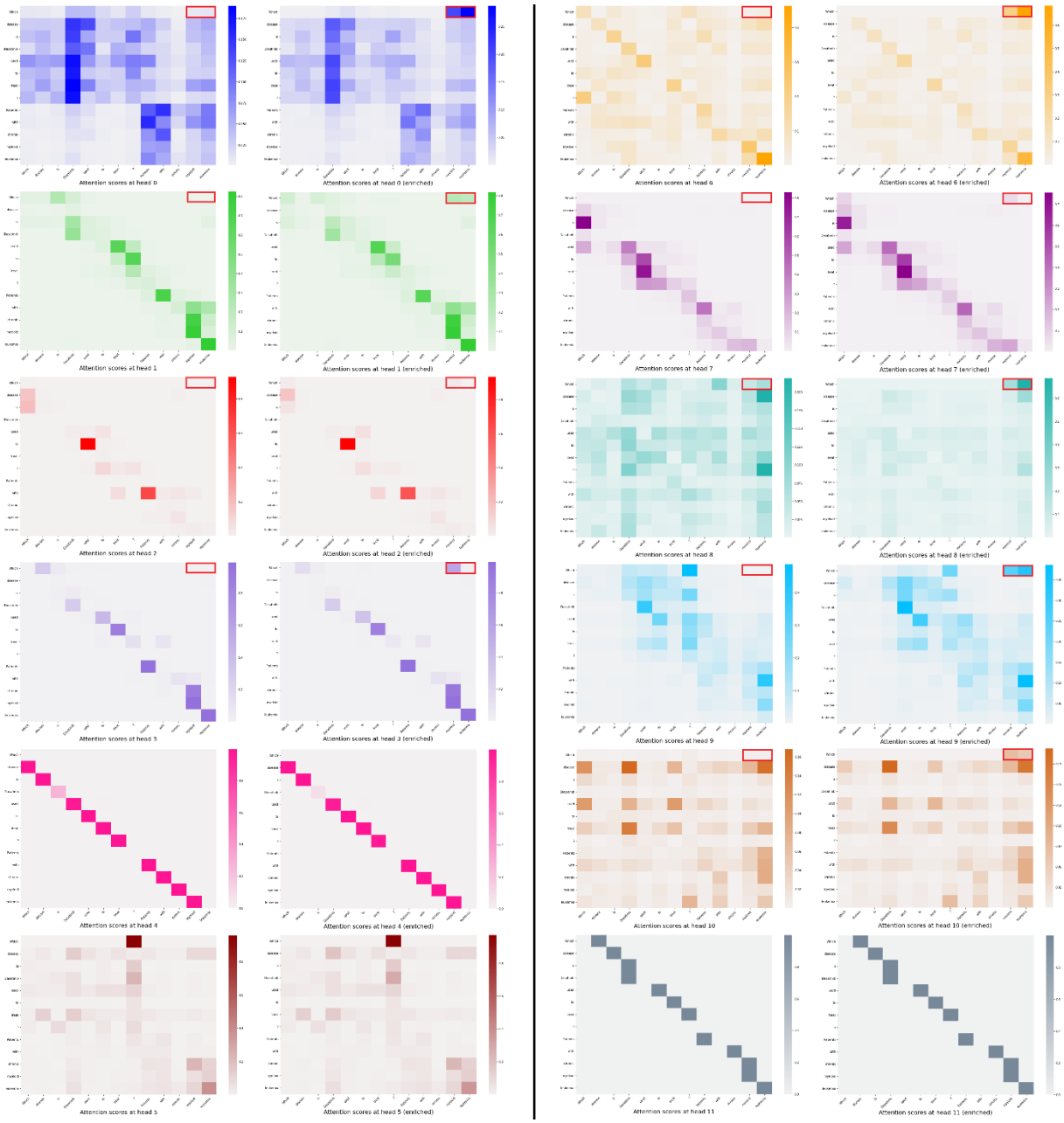

Figure 46 Scores d'attention dans l'ensemble des 12 têtes de la couche 2 avant et après l'enrichissement de l'attention biomédicale et NER

Glossaire

Intelligence artificielle : est l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine

Dataset : est une collection de données utilisée surtout pour l'entraînement, validation et l'évaluation des modèles d'apprentissage automatique

Réseau neuronal artificiel : est un système dont la conception est à l'origine schématiquement inspirée du fonctionnement des neurones biologiques, et qui par la suite s'est rapproché des méthodes statistiques

Baseline : est un modèle de base sur lequel se basent d'autres modèles

Transformers : est un modèle d'apprentissage profond qui adopte le mécanisme de l'attention, en pesant l'influence de différentes parties des données d'entrée

Encodeur-décodeur : est une architecture d'apprentissage profond pour le traitement automatique de langage naturel qui se base sur deux composants le 1er pour le codage du texte et le deuxième pour le décodage du vecteur représentatif du texte codé

Mécanisme d'attention : est la capacité d'apprendre à se concentrer sur des parties spécifiques d'une donnée complexe, par exemple un mot dans une phrase ou une partie d'une image

Apprentissage par transfert : (transfer learning en anglais) est l'un des champs de recherche de l'apprentissage automatique qui vise à transférer des connaissances d'une ou plusieurs tâches sources vers une ou plusieurs tâches cibles

Métrique : est une mesure (quantitative) pour la mesure et l'évaluation d'un système

Époque : indique le nombre de passes de l'ensemble de données d'entraînement complet que l'algorithme d'apprentissage automatique a effectué

Taille du lot (batch size) : est un hyperparametre pour les réseaux de neurones artificiels, qui définit le nombre d'échantillons par époque qui seront propagés à travers le réseau avant chaque modification de ses poids

Train Set : est l'ensemble de données utilisées pour l'entraînement d'un modèle d'apprentissage automatique

Dev Set : est l'ensemble de données utilisées pour la validation d'un modèle d'apprentissage automatique

Test Set : est l'ensemble de données utilisées pour l'évaluation et le test d'un modèle d'apprentissage automatique

PubMed : la plus grande base de données d'articles de recherche biomédicale

Apprentissage profond : c'est un sous-ensemble de l'apprentissage automatique, qui consiste essentiellement en un réseau neuronal comportant trois couches ou plus.

Chatbot : C'est un logiciel qui simule des conversations de type humain avec des utilisateurs par le biais de messages textuels sur le chat.

Ontologie biomédicale : décrit les concepts des terminologies médicales et les relations entre eux, permettant ainsi le partage des connaissances médicales.

Reconnaissance d'entités nommées (NER) : C'est une sous-tâche de l'extraction d'informations qui vise à localiser les entités nommées mentionnées dans un texte non structuré dans des catégories prédéfinies telles que des noms de personnes, des organisations, des lieux...

Public leaderboard : Affiche le classement des équipes participantes à une compétition en fonction de leur score.

Apprentissage supervisé : est une sous-catégorie de l'apprentissage automatique et de l'intelligence artificielle. Elle se définit par son utilisation d'ensembles de données étiquetées pour former des algorithmes qui classent les données ou prédisent les résultats avec précision.

Apprentissage de séquence à séquence : il consiste à former des modèles pour convertir des séquences d'un domaine (par exemple, des phrases en anglais) en séquences d'un autre domaine (par exemple, les mêmes phrases traduites en français).

Pointer Net : Un réseau de pointeurs apprend la probabilité conditionnelle d'une séquence de sortie dont les éléments sont des jetons discrets correspondant à des positions dans une séquence d'entrée.

Réseau récurrent : Un réseau de neurones récurrents est un réseau de neurones artificiels présentant des connexions récurrentes.

Convolution : La convolution est une méthode mathématique permettant de combiner deux signaux pour en former un troisième.

Liste d'or : en anglais "Golden list", désigne la référence des prédictions correctes, avec lesquelles on peut évaluer un modèle machine learning

F-mesure : un indicateur de synthèse permettant d'évaluer les algorithmes de classification de données textuelles

n-gramme : Un n-gramme est un ensemble de *n* éléments successifs dans un document texte pouvant comprendre des mots, des nombres, des symboles et de la ponctuation

Technique cloze : une stratégie d'enseignement qui utilise des passages de texte avec des mots manquants. Elle est exploitée en machine learning dans la phase d'apprentissage.

UMLS : ensemble de fichiers et de logiciels qui rassemblent de nombreux vocabulaires et normes sanitaires et biomédicaux afin de permettre l'interopérabilité entre les systèmes informatiques.

Knowledge Base (KB) : une technologie utilisée pour stocker des informations complexes, structurées et non structurées, utilisée par un système informatique.

RDF : est un modèle standard d'échange de données sur le Web.

Réseau convolutif de graphes (GCN) : architecture de réseaux neuronaux pour l'apprentissage automatique sur les graphes.

Mécanisme de co-attention : permet l'apprentissage des attentions par paire.

Régression logistique : estime les paramètres d'un modèle logistique (les coefficients de la combinaison linéaire).

Word embedding : est une représentation apprise pour le texte où les mots qui ont la même signification ont une représentation similaire.

Méta-apprentissage : un sous-domaine de l'apprentissage automatique dans lequel les algorithmes d'apprentissage automatique sont appliqués aux métadonnées relatives aux expériences d'apprentissage automatique.

Pré-entraînement : consiste à former d'abord un modèle sur une tâche ou un ensemble de données. Ensuite, on utilise les paramètres ou le modèle de cette formation pour former un autre modèle sur une autre tâche ou un autre ensemble de données.

Réseau à action directe « Feed-Forward Network » : un réseau neuronal artificiel dans lequel les connexions entre les nœuds ne forment pas un cycle.

Réglage fin « Fine-tuning » : est le processus par lequel les paramètres d'un modèle doivent être ajustés très précisément afin de correspondre à certaines observations.

Perplexité : est une mesure de la capacité d'un modèle de probabilité à prédire un échantillon. Dans le contexte de l'NLP, la perplexité est une façon d'évaluer les modèles de langage.

Taille du modèle : le nombre de paramètres du modèle pouvant être entraînés

Longueur maximale de la séquence : la longueur maximale en caractères de la séquence d'entrée à un modèle

Taux d'apprentissage « Learning rate » : est un paramètre dans un algorithme d'optimisation qui détermine la taille du pas à chaque itération tout en se déplaçant vers un minimum d'une fonction de perte.

GPU : est un circuit électronique spécialisé conçu pour manipuler et modifier la mémoire afin d'accélérer la création d'images.

Instances d'apprentissage étiquetées : les données étiquetées sont des données accompagnées d'une étiquette, comme un nom, un type ou un nombre.

Apprentissage semi-supervisé : L'apprentissage semi-supervisé utilise des données étiquetées et non étiquetées pour former un modèle.

Réseau adversarial génératif « Generative Adversarial Network » (GAN) : ce sont des architectures algorithmiques qui utilisent deux réseaux neuronaux

Réseau neuronal graphique « Graph Neural Network » (GNN) : une classe de méthodes d'apprentissage profond conçues pour effectuer des inférences sur des données décrites par des graphes.

Auto-codeur de débruitage « denoising autoencoder » : un modèle qui résoudre un problème en corrompant volontairement les données en mettant aléatoirement à zéro certaines des valeurs d'entrée.

Tokénisation : est un moyen de séparer un morceau de texte en unités plus petites appelées jetons.

Marquage des parties du discours « part-of-speech tagging » : classer les mots d'un texte en correspondance avec une partie particulière du discours, en fonction de la définition du mot et de son contexte.

Token : un morceau résultant de la phase de tokénisation

TF-IDF : une statistique numérique destinée à refléter l'importance d'un mot dans un document d'une collection ou d'un corpus.

BM25 : est une fonction de classement utilisée par les moteurs de recherche pour estimer la pertinence des documents par rapport à une requête de recherche donnée.

ElasticSearch : est un moteur de recherche basé sur la bibliothèque Lucene

Algorithmes de sac de mots (BOW) : est une représentation du texte qui décrit l'occurrence des mots dans un document.

Publications et Communications

Articles publiés/soumis dans des journaux indexés

Kaddari Z., Bouchentouf T., FrBMedQA: the first French biomedical question answering dataset, IAES International Journal of Artificial Intelligence, 2022, doi: http://doi.org/10.11591/ijai.v11.i4.pp1588-1595

Kaddari Z., Bouchentouf T., A novel self-attention enriching mechanism for biomedical question answering, Expert Systems with Applications, 2023, doi: https://doi.org/10.1016/j.eswa.2023.120210

Kaddari Z., Mellah Y., Haja Z., Berrich J., Bouchentouf T., Transfer learning for Yes/No biomedical question answering, International Journal of Advanced Computer Science and Applications (IJACSA), 2023 (Soumis)

Communications publiées dans des conférences indexées

Kaddari Z., Bouchentouf T., LaRSA at BioASQ 10b: classical and novel approaches for biomedical document retrieval and question answering, CEUR Workshop Proceedings, 2022, 3180, pp. 274–280

Kaddari Z., Mellah Y., Berrich J., Belkasmi M.G., Bouchentouf T., OctaNLP: A Benchmark for Evaluating Multitask Generalization of Transformer-Based Pre-trained Language Models, Lecture Notes in Electrical Engineering, 2022, 745, pp. 223–232

Kaddari Z., Berrich J., Rahmoun N., Belouali S., Bouchentouf T., INKAD COVID-19 IntelliSearch: A multilingual search engine for answering questions about COVID-19 in real-time from the scientific literature, 5th International Conference on Intelligent Computing in Data Sciences, ICDS 2021, 2021

Mellah Y., Kaddari Z., Berrich J., Bouchentouf T., Belkasmi M.G, Text to Code Conversion Using Deep Learning for NLP, 2020 International Symposium on Advanced Electrical and Communication Technologies, ISAECT 2020, 2020

Kaddari Z., Mellah Y., Berrich J., Bouchentouf T., Belkasmi M.G., Biomedical Question Answering: A Survey of Methods and Datasets, 4th International Conference on Intelligent Computing in Data Sciences, ICDS 2020, 2020

Kaddari Z., Mellah Y., Berrich J., Bouchentouf T., Belkasmi M.G., Applying the T5 language model and duration units normalization to address temporal common sense understanding on the MCTACO dataset, 2020 International Conference on Intelligent Systems and Computer Vision, ISCV 2020, 2020

Chapitres livres indexées

Kaddari Z., Mellah Y., Berrich J., Belkasmi M.G., Bouchentouf T., Natural language processing: Challenges and future directions, Lecture Notes in Networks and Systems, 2021, 144, pp. 236–246

Bibliographie

[1] J. Devlin, M. Chang, K. Lee et K. Toutanova, «BERT: pre-training of deep bidirectional transformers for language understanding,» *CoRR,* vol. 1810.04805, 2018.

[2] Brown, B. Tom, Benjamin and Ryder et Nick and Subbiah, «Language Models are Few-Shot Learners,» *CoRR,* 2020.

[3] T. ussell-Rose et J. Chamberlain, «Expert search strategies: The information retrieval practices of healthcare information professionals,» *JMIR Med Inform,* vol. 5, n° %14, p. 33, 2017.

[4] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara et S. Petridis, «An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition,» *BMC Bioinformatics,* vol. 16, 2015.

[5] P. Rajpurkar, J. Zhang, K. Lopyrev et P. Liang, «Squad: 100, 000+ questions for machine comprehension of text,» *CoRR,* vol. 1606.05250, 2016.

[6] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov et C. D. Manning, «Hotpotqa: A dataset for diverse, explainable multi-hop question answering,» *CoRR,* vol. 1809.09600, 2018.

[7] S. Reddy, D. Chen et C. D. Manning, «Coqa: A conversational question answering challenge,» *CoRR,* vol. 1808.07042, 2018.

[8] E. Choi, H. He, M. Iyyer, M. Yatskar, W. Yih, Y. Choi, P. Liang et L. Zettlemoyer, «Quac : Question answering in context,» *CoRR,* vol. 1808.07036, 2018.

[9] T. Kocisky, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis et E. Grefenstette, «The NarrativeQA reading comprehension challenge,» *Transactions of the Association for Computational Linguistics,* vol. 6, pp. 317-328, 2018.

[10] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee, K. N. Toutanova, L. Jones et M. W. Chang, «Natural questions: a benchmark for question answering research,» chez *Transactions of the Association of Computational Linguistics*, 2019.

[11] R. Kadlec, M. Schmid, O. Bajgar et J. Kleindienst, «Text understanding with the attention sum reader network,» *CoRR,* vol. 1603.01547, 2016.

[12] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu et G. Hu, «Attention-over-attention neural networks for reading comprehension,» *CoRR,* vol. 1607.04423, 2016.

[13] M. J. Seo, A. Kembhavi, A. Farhadi et H. Hajishirzi, «Bidirectional attention flow for machine comprehension,» *CoRR,* vol. 1611.01603, 2016.

[14] Z. Zhang, J. Yang et H. Zhao, «Retrospective Reader for Machine Reading Comprehension,» *CoRR,* vol. 2001.09694, n° %12020.

[15] A. Shrestha et A. Mahmood, «Review of deep learning algorithms and architectures,» *IEEE Access,* vol. 7, 2019.

[16] P. Rajpurkar, R. Jia et P. Liang, «Know What You Don't Know: Unanswerable Questions for SQuAD,» chez *The 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[17] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary et R. Majumder, «MS MARCO: A Human Generated MAchine Reading COmprehension Dataset,» *CoRR,* vol. 1611.09268, 2016.

[18] T. Kwiatkowski, J. Palomaki, O. Redfield et M. Collins, «Natural questions: A benchmark for question answering research,» chez *Assoc. Comput. Linguist.*, 2019.

[19] Y. Yang, W. T. Yih et C. Meek, «Wikiqa: A challenge dataset for open-domain question answering,» chez *The 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[20] Z. Yang, P. Qi, S. Zhang, Y. Bengio et W. Cohen, «HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering,» chez *The 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[21] T. Kocisky, J. Schwarz, P. Blunsom, C. Dyer et k. Hermann, «The narrativeqa reading comprehension challenge,» chez *Trans. Assoc. Comput. Linguist*, 2018.

[22] J. Welbl, P. Stenetorp et S. Riedel, «Constructing datasets for multi-hop reading comprehension across documents,» chez *Trans. Assoc. Comput. Linguist*, 2018.

[23] D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay et D. Roth, «Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences,» chez *The 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

[24] A. Talmor, J. Herzig, N. Lourie et J. Berant, «CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge,» chez *The 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

[25] S. Zhang, X. Liu, J. Liu, J. Gao et K. Duh, «Record: Bridging the gap between human and machine commonsense reading comprehension,» *CoRR,* 2018.

[26] T. Mihaylov, P. Clark, T. Khot et A. Sabharwal, «Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering,» *CoRR,* 2018.

[27] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh et M. Gardner, «DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs,» *CoRR,* 2019.

[28] Lai, Guokun and Xie, Qizhe and Liu et Hanxiao, «RACE: Large-scale ReAding Comprehension Dataset From Examinations,» *CoRR,* 2017.

[29] Xie, Qizhe and Lai et Eduard, «Large-scale Cloze Test Dataset Created by Teachers,» 2017.

[30] Peas, Anselmo, Eduard et Rodrigo, QA4MRE 2011-2013: Overview of Question Answering for Machine Reading Evaluation, Springer Berlin Heidelberg, 2013.

[31] E. Voorhees et D. Tice, «Building a Question Answering Test Collection,» chez *SIGIR*, 2000.

[32] Richardson, M. Burges, Renshaw et Erin, «MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text,» chez *The 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.

[33] Suster, S. Daelemans et Walter, «CliCR: a Dataset of Clinical Case Reports for Machine Reading Comprehension,» chez *The 2018 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

[34] J. Welbl, F. Nelson et Matt Gardner, «Crowdsourcing Multiple Choice Science Questions,» *CoRR,* 2017.

[35] Grail, Quentin and Perez et Julien, «ReviewQA: a relational aspect-based opinion reading dataset,» *CoRR,* 2018.

[36] Tushar Khot, S. Ashish et P. Clark, «SciTaiL: A Textual Entailment Dataset from Science Question Answering,» *AAAI,* 2018.

[37] Dasigi, Pradeep and Lo, Kyle and Beltagy et Iz and Cohan, «A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers,» *CoRR,* 2021.

[38] K. Aniruddha, S. Minjoon, S. Dustin et H. Hannaneh, «Are You Smarter Than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension,» chez *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[39] Yagcioglu, E. Semih, E. Aykut et Nazli, «RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes,» chez *The 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[40] Inoue, Naoto and Furuta, Ryosuke and Yamasaki et Kiyoharu, «Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation,» *CoRR,* 2018.

[41] Tapaswi, Makarand and Zhu, Yukun and Stiefelhagen et Sanja, «MovieQA: Understanding Stories in Movies through Question-Answering,» *CoRR,* 2015.

[42] Dhingra, Bhuwan and Mazaitis et Kathryn and Cohen, «Quasar: Datasets for Question Answering by Search and Reading,» *CoRR,* 2017.

[43] Dunn, Matthew and Sagun, Levent and Higgins, Mike and Guney et Kyunghyun, «SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine,» *CoRR,* 2017.

[44] Sun, Kai and Yu, Dian and Chen et Claire, «DREAM: A Challenge Dataset and Models for Dialogue-Based Reading Comprehension,» *CoRR,* 2019.

[45] Saeidi, Marzieh and Bartolo, Max and Lewis, Guillaume and Riedel et Sebastian, «Interpretation of Natural Language Rules in Conversational Machine Reading,» *CoRR,* 2018.

[46] A. Marco et S. P. Fernando, «A literature review on question answering techniques, paradigms and systems,» *Journal of King Saud University - Computer and Information Sciences,* vol. 32, n° %16, pp. 635-646, 2020.

[47] F. A. Gers, J. Schmidhuber et F. Cummins, «Learning to forget: continual prediction with LSTM,» chez *Ninth International Conference on Artificial Neural Networks ICANN 99*, 1999.

[48] Huang, Zhiheng and Xu, Wei and Yu et Kai, «Bidirectional LSTM-CRF Models for Sequence Tagging,» *CoRR,* 2015.

[49] Bahdanau, Dzmitry and Cho, Kyunghyun and Bengio et Yoshua, «Neural Machine Translation by Jointly Learning to Align and Translate,» *CoRR,* 2014.

[50] Wang, Shuohang and Jiang et Jing, «Machine Comprehension Using Match-LSTM and Answer Pointer,» *CoRR,* 2016.

[51] Wang, Shuohang and Jiang et Jing, «Learning Natural Language Inference with LSTM,» *CoRR,* 2015.

[52] Vinyals, Oriol and Fortunato, Meire and Jaitly et Navdeep, «Pointer Networks,» *CoRR,* 2015.

[53] Xiong, Caiming and Zhong, Victor and Socher et Richard, «Dynamic Coattention Networks For Question Answering,» *CoRR,* 2016.

[54] Seo, Minjoon and Kembhavi, Aniruddha and Farhadi et Hannaneh, «Bidirectional Attention Flow for Machine Comprehension,» *CoRR,* 2016.

[55] K. Hermann, T. Kočiský, E. Grefenstette et P. Blunsom, «Teaching Machines to Read and Comprehend,» *CoRR,* 2015.

[56] W. Wang, N. Yang, F. Wei et M. Zhou, «Gated Self-Matching Networks for Reading Comprehension and Question Answering,» chez *The 55th Annual Meeting of the Association for Computational Linguistics*.

[57] T. Rocktäschel, E. Grefenstette, K. Hermann et P. Blunsom, «Reasoning about Entailment with Neural Attention,» *CoRR,* 2015.

[58] J. J. Hopfield, «Neural networks and physical systems with emergent collective computational abilities,» chez *Proceedings of the National Academy of Sciences*, 1982.

[59] A. Yu et D. Wei, «QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension,» *CoRR,* 2018.

[60] M. Peters, M. Neumann et M. Iyyer, «Deep contextualized word representations,» *CoRR,* 2018.

[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser et I. Polosukhin, «Attention Is All You Need,» *CoRR,* vol. abs/1706.03762, 2017.

[62] Z. Lan, M. Chen, S. Goodman et R. Soricut, «ALBERT: A Lite BERT for Self-supervised Learning of Language Representations,» *CoRR,* 2019.

[63] Z. Yang, Z. Dai et V. Quoc, «XLNet: Generalized Autoregressive Pretraining for Language Understanding,» *CoRR,* 2019.

[64] Y. Liu, M. Ott, N. Goyal et V. Stoyanov, «RoBERTa: A Robustly Optimized BERT Pretraining Approach,» *CoRR,* 2019.

[65] A. W. Fleuren, «Application of text mining in the biomedical domain,» *Methods,* vol. 74, pp. 97-106, 2015.

[66] W. L. Taylor, «cloze procedure: A new tool for measuring readability,» *Journalism Quarterly,* vol. 30, n° %14, pp. 415-433, 1953.

[67] P. Stavropoulos, D. Pappas, I. Androutsopoulos et R. McDonald, «BIOMRC: A Dataset for Biomedical Machine Reading Comprehension,» *CoRR,* vol. 2005.06376, 2020.

[68] D. Pappas, I. Androutsopoulos et H. Papageorgiou, «BioRead: A new dataset for biomedical reading comprehension,» chez *LREC*, 2018.

[69] S. Kim, D. Park, Y. Choi, K. Lee, B. Kim, M. Jeon et J. Kim, «A pilot study of biomedical text comprehension using an attention-based deep neural reader: Design and experimental analysis,» *JMIR Med Inform,* vol. 6, n° %11, 2018.

[70] L. Chin-Yew, «ROUGE: A package for automatic evaluation of summaries,» chez *Association for Computational Linguistics*, 2004.

[71] K. Papineni, S. Roukos et T. Ward, «Bleu: a Method for Automatic Evaluation of Machine Translation,» chez *The 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[72] A. Hersh et W. Cohen, «Trec 2006 genomics track overview,» chez *The Fifteenth Text Retrieval Conference (TREC 206)*, 2006.

[73] «Trec 2007 genomics track overview,» chez *The Sixteenth Text Retrieval Conference (TREC 2007)*, 2007.

[74] A. Aronson, «Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program,» chez *Proceedings. AMIA Symposium*, 2001.

[75] S. Lee et D. Kim, «Best: Next-generation biomedical entity search tool for knowledge discovery from biomedical literature,» chez *PLOS ONE*, 2016.

[76] X. Zhang, J. Wu, Z. He et Y. Su, «Medical Exam Question Answering with Large-scale Reading Comprehension,» *CoRR,* 2018.

[77] A. Pampari, P. Raghavan et J. Liang, «emrQA: A Large Corpus for Question Answering on Electronic Medical Records,» chez *The 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[78] D. Ben Abacha, «A question-entailment approach to question answering,» *BMC Bioinformatics,* 2019.

[79] Q. Jin et B. Dhingra, «PubMedQA: A Dataset for Biomedical Research Question Answering,» *CoRR,* 2019.

[80] C. Wei et R. Harris, «Accelerating literature curation with text-mining tools: a case study of using pubtator to curate genes in pubmed abstracts,» *Database: the journal of biological databases and curation,* 2012.

[81] L. Leaman et R. Islamaj Dogan, «DNORM with pairwise learning to rank,» *Bioinformatics,* 2013.

[82] Y. Niu, G. Hirst et M. Gregory, «Answering clinical questions with role identification,» chez *The ACL 2003 Workshop on Natural Language Processing in Biomedicine*, 2003.

[83] D. Demner-Fushma et J. Lin, «Knowledge extraction for clinical question answering: Preliminary results,» chez *AAAI Workshop - Technical Report*, 2005.

[84] D. Demner-Fushman et J. Lin, «Answer extraction, semantic clustering, and extractive summarization for clinical question answering,» chez *The 21st International Conference on*

Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, 2006.

[85] D. Demner-Fushman et J. Lin, «Answering clinical questions with knowledge-based and statistical techniques,» *Computational Linguistics,* 2007.

[86] X. Huang, J. Lin et D. Demner-Fushman, «Evaluation of pico as a knowledge representation for clinical questions,» chez *AMIA annual symposium proceedings*, 2006.

[87] L. Minsuk, J. Cimino, Hai Ran et C. Sable, «Beyond information retrieval—medical question answering,» chez *AMIA annual symposium proceedings*, 2006.

[88] Z. Shi, M. Gabor et Y. Wang, «Question answering summarization of multiple biomedical documents,» chez *Conference of the Canadian Society for Computational Studies of Intelligence*, 2007.

[89] G. Melli, Y. Wang et S. Anoop, «Description of squash,the sfu question answering summary handler for the duc-2005 summarization task,» 2005.

[90] R. Terol et M. Patricio, «A knowledge based method for the medical question answering problem,» *Computers in biology and medicine,* 2007.

[91] S. Cruchet et A. Gaudinat, «Supervised approach to recognize question type in a qa system for health,» *Studies in Health Technology and Informatics,* 2008.

[92] R. Lin, Justin Liang-Te Chiu et Hong-Jei Dai, «Biological question answering with syntactic and semantic feature matching and an improved mean reciprocal ranking measurement,» chez *IEEE International Conference on Information Reuse and Integration*, 2008.

[93] J. Gobeill, E. Patsche et D. Theodoro, «Question answering for biology and medicine,» chez *9th Internationa Conference on Information Technology and Applications in Biomedicine*, 2009.

[94] Y. Cao et L. Feifan, «Askhermes: An online question answering system for complex clinical questions,» *Journal of biomedical informatics,* 2011.

[95] M. Sarrouti et S. Ouatik El Alaoui, «SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions,» *Artificial Intelligence in Medicine,* vol. 102, p. 101767, 2020.

[96] M. Ashburner, C. Ball et J. Blake, «Gene ontology: tool for the unification of biology,» *Nature genetics,* 2000.

[97] Gene Ontology Consortium, «The gene ontology resource: 20 years and still going strong,» *Nucleic acids research,* 2019.

[98] M. Stearns, C. Price et K. Spackman, «Snomed clinical terms: overview of the development process and project status,» chez *The AMIA Symposium*, 2001.

[99] F. Rinaldi, J. Dowdall et G. Schneider, «Answering questions in the genomics domain,» chez *The Conference on Question Answering in Restricted Domains*, 2004.

[100] D. Molla, R. Schwitter et M. Hess, «Extrans, an answer extraction system,» *In TAL,* 2000.

[101] A. Ben Abacha et P. Zweigenbaum, «Medical question answering: translating medical questions into sparql queries,» chez *The 2nd ACM SIGHIT international health informatics*, 2012.

[102] A. Ben Abacha et P. Zweigenbaum, «Means: A medical question-answering system combining nlp techniques and semantic web technologies,» *Information processing & management,* 2015.

[103] J.-D. Kim et B. Cohen, «Natural language query processing for sparql generation: A prototype system for snomed ct,» chez *Biolink*, 2013.

[104] D. Li, B. Hu et Q. Chen, «Towards medical machine reading comprehension with structural knowledge and plain text,» chez *The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[105] T. Kipf et M. Welling, «Semisupervised classification with graph convolutional networks,» *CoRR,* 2016.

[106] B. Hao, H. Zhu et I. Paschalidis, «Enhancing clinical BERT embedding using a biomedical knowledge base,» chez *The 28th International Conference on Computational Linguistics*, 2020.

[107] Z. Yuan, Z. Zhao et S. Yu, «Coder: Knowledge infused cross-lingual medical term embedding for term normalization,» *CoRR,* 2020.

[108] L. L. Wang, K. Lo, Y. Chandrasekhar et R. Reas, «Cord-19: The covid-19 open research dataset,» *CoRR,* 2020.

[109] Y. Liu, «The university of alberta participation in the bioasq challenge: The wishart system,» chez *Semantic Indexing Question Answering*, 2013.

[110] Z.-X. Jin, B.-W. Zhang, F. Fang et L.-L. Zhang, «A multi-strategy query processing approach for biomedical question answering: USTB PRIR at BioASQ 2017 task 5B,» chez *Association for Computational Linguistics*, 2017.

[111] L. Bonnefoy, R. Deveaud et P. Bellot, «Do social information help book search?,» 2012.

[112] N. Zhiltsov, A. Kotov et F. Nikolaev, «Fielded sequential dependence model for adhoc entity retrieval in the web of data,» chez *The 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015.

[113] S. Clinchant et E. Gaussier, «Bridging language modeling and divergence from randomness models: A log-logistic model for ir,» chez *The Theory of Information Retrieval*, 2009.

[114] G. Brokos, P. Liosis, R. McDonald et D. Pappas, «AUEB at BioASQ 6: Document and snippet retrieval,» chez *The 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, 2018.

[115] D. Pappas, R. McDonald et G.-I. Brokos, «Aueb at bioasq 7: document and snippet retrieval,» chez *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2019.

[116] D. Pappas, P. Stavropoulos et I. Androutsopoulos, «Aueb-nlp at bioasq 8: Biomedical document and snippet retrieval,» 2020.

[117] K. Hui, A. Yates et K. Berberich, «PACRR: A position-aware neural IR model for relevance matching,» chez *The 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[118] J. Guo, Y. Fan et Q. Ai, «A deep relevance matching model for ad-hoc retrieval,» chez *The 25th ACM International on Conference on Information and Knowledge Management*, 2016.

[119] W. Yin, H. Schutze et B. Xiang, «ABCNN: Attention-based convolutional neural network for modeling sentence pairs,» chez *Transactions of the Association for Computational Linguistics*, 2016.

[120] S. Chakraborty, E. Bisong, S. Bhatt et T. Wagner, «BioMedBERT: A pre-trained biomedical language model for QA and IR,» chez *The 28th International Conference on Computational Linguistics*, 2020.

[121] D. Weissenborn, G. Wiese et L. Seiffe, «Making neural QA as simple as possible but not simpler,» chez *The 21st Conference on Computational Natural Language Learning*, 2017.

[122] Q. Jin, B. Dhingra et W. Cohen, «Probing biomedical embeddings from language models,» chez *The 3rd Workshop on Evaluating Vector Space Representations for NLP*, 2019.

[123] L. Jinhyuk, Y. Wonjin, K. Sungdong, K. Donghyeon, K. Sunkyu, S. Chan Ho et K. Jaewoo, «BioBERT: a pre-trained biomedical language representation model for biomedical text mining,» *Bioinformatics,* vol. 36, p. 1234–1240, 2020.

[124] B. Iz, K. Lo et A. Cohan, «SciBERT: A Pretrained Language Model for Scientific Text,» chez *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[125] E. Alsentzer, J. Murphy et Boag, «Publicly available clinical BERT embeddings,» chez *The 2nd Clinical Natural Language Processing Workshop*, 2019.

[126] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao et H. Poon, «Domain-Specific Language Model Pretraining for Biomedical Natural,» *CoRR,* vol. abs/2007.15779, 2020.

[127] I. Partalas, E. Gaussier et A.-C. Ngonga, «Results of the first bioasq workshop,» 2013.

[128] G. Balikas, I. Partalas et A. Krithara, «Results of the bioasq tasks of the question answering lab at clef 2014,» chez *clef 2014*, 2014.

[129] D. Weissenborn, G. Tsatsaronis et M. Schroeder, «Answering factoid questions in the biomedical domain,» 2013.

[130] D. Weissenborn, G. Tsatsaronis et Y. Zhang, «The fudan participation in the 2015 bioasq challenge: Large-scale biomedical semantic indexing and question answering,» chez *CEUR Workshop Proceedings*, 2015.

[131] S. Choi, «Snumedinfo at clef qa track bioasq 2015,» 2015.

[132] F. Schulze, R. Schuler, T. Draeger et D. Dummer, «Hpi question answering system in bioasq 2016,» chez *The Fourth BioASQ workshop*, 2016.

[133] G. Erkan et D. Radev, «Lexrank: Graph-based lexical centrality as salience in text summarization,» *Journal of artificial intelligence research,* 2004.

[134] A. Bhandwaldar et W. Zadrozny, «UNCC QA: Biomedical question answering system,» chez *The 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, 2018.

[135] G. Wiese, D. Weissenborn et M. Neves, «Neural domain adaptation for biomedical question answering,» chez *The 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 2017.

[136] G. Wiese, D. Weissenborn, B. Dhingra et D. Danish, «Simple and effective semi-supervised question answering,» chez *The 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

[137] B. Dhingra, H. Liu et Z. Yang, «Gated-attention readers for text comprehension,» chez *The 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

[138] Y. Du, B. Pei et X. Zhao, «Hierarchical multi-layer transfer learning model for biomedical question answering,» chez *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018.

[139] Y. Du, W. Guo et Y. Zhao, «Hierarchical question-aware context learning with augmented data for biomedical question answering,» chez *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019.

[140] J. Kang, «Transferability of natural language inference to biomedical question answering,» *CoRR,* 2020.

[141] W. Yoon, J. Lee et D. Kim, «Pre-trained language model for biomedical question answering,» chez *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2019.

[142] S. Harabagiu et A. Hickl, «Methods for using textual entailment in open-domain question answering,» chez *The 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006.

[143] A. Ben Abacha et D. Demner-Fushman, «Recognizing question entailment for medical question answering,» chez *AMIA Annual Symposium Proceedings*, 2016.

[144] P. Nakov, A. Moschitti et W. Magdy, «SemEval-2016 task 3: Community question answering,» chez *The 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016.

[145] a. Luo, G.-Q. Zhang, S. Wentz et L. Cui, «Simq: real-time retrieval of similar consumer health questions,» *Journal of medical Internet research,* 2015.

[146] D. Campbell et S. Johnson, «A transformational-based learner for dependency grammars in discharge summaries,» chez *The ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*, 2002.

[147] D. Wang et E. Nyberg, «Cmu oaqa at trec 2017 liveqa: A neural dual entailment approach for question paraphrase identification,» chez *TREC*, 2017.

[148] V. Nair et G. Hinton, «Rectified linear units improve restricted boltzmann machines,» chez *The 27th International Conference on International Conference on Machine Learning*, 2010.

[149] W. Zhu et X. Zhou, «Panlp at mediqa 2019: Pre-trained language models, transfer learning and knowledge distillation,» chez *BioNLP@ACL*, 2019.

[150] X. Liu, P. He et W. Chen, «Multi-task deep neural networks for natural language understanding,» *CoRR,* 2019.

[151] K. Simonyan et A. Zisserman, «Very deep convolutional networks for large-scale image recognition,» *CoRR,* 2014.

[152] K. He, X. Zhang et S. Ren, «Deep residual learning for image recognition,» chez *The IEEE conference on computer vision and pattern recognition*, 2016.

[153] X. Yan, L. Li, C. Xie et J. Xiao, «Zhejiang university at imageclef 2019 visual question answering in the medical domain,» chez *CLEF*, 2019.

[154] F. Ren et Y. Zhou, «Cgmvqa: A new classification and generative model for medical visual question answering,» *IEEE Access,* 2020.

[155] M. Lin, Q. Chen et S. Yan, «Network in network,» *CoRR,* 2013.

[156] B. Nguyen, T.-T. Do, B. Nguyen, T. Do et E. Tjiputra, «Overcoming data limitation in medical visual question answering,» chez *In International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019.

[157] C. Finn, P. Abbeel et S. Levine, «Model-agnostic meta-learning for fast adaptation of deep networks,» *CoRR,* 2017.

[158] J. Masci et U. Meier, «Stacked convolutional auto-encoders for hierarchical feature extraction,» chez *International conference on artificial neural networks*, 2011.

[159] Z. Yang, X. He, J. Gao et L. Deng, «Stacked attention networks for image question answering,» chez *The IEEE conference on computer vision and pattern recognition*, 2016.

[160] A. Fukui, D. Park et D. Yang, «Multimodal compact bilinear pooling for visual question answering and visual grounding,» *CoRR,* 2016.

[161] Z. Yu, J. Yu et J. Fan, «Multi-modal factorized bilinear pooling with co-attention learning for visual question answering,» chez *The IEEE international conference on computer vision*, 2017.

[162] Z. Yu, J. Yu, C. Xiang et J. Fan, «Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering,» *IEEE transactions on neural networks and learning systems,* 2018.

[163] L. Li, M. Yatskar et D. Yin, «Visualbert: A simple and performant baseline for vision and language,» 2019.

[164] H. Tan et M. Bansal, «LXMERT: Learning cross-modality encoder representations from transformers,» chez *The 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[165] F. Ren et Y. Zhou, «Cgmvqa: A new classification and generative model for medical visual question answering,» *IEEE Access,* 2020.

[166] Y.-C. Chen, L. Li, L. Yu et Y. Cheng, «Uniter: Learning universal image-text representations,» *CoRR,* 2019.

[167] C. Yiming, Z. Wei-Nan, C. Wanxiang, L. Ting, C. Zhigang et W. Shijin, «Multilingual multi-aspect explainability analyses on machine reading comprehension models,» *iScience,* vol. 25, 2022.

[168] M. Pande, A. Budhraja, P. Nema, P. Kumar et M. M. Khapra, «The heads hypothesis: A unifying statistical approach towards understanding,» *CoRR,* vol. abs/2101.09115, 2021.

[169] v. Aken, B. Winter, A. Löser et F. A. Gers, «How Does BERT Answer Questions? A Layer-Wise Analysis of Transformer Representations,» chez *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019.

[170] J. Lei, R. Kiros et H. Geoffrey, «Layer Normalization,» *CoRR,* 2016.

[171] E. Alsentzer, J. Murphy et W. Boag, «Publicly Available Clinical {BERT} Embeddings,» *CoRR,* 2019.

[172] H. Kexin, J. Altosaar et R. Ranganath, «ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission,» *CoRR,* 2019.

[173] A. Adhikari, R. Achyudh et R. Tang, «DocBERT: BERT for Document Classification,» *CoRR,* 2019.

[174] J.-S. Lee et J. Hsiang, «PatentBERT: Patent Classification with Fine-Tuning a pre-trained BERT Model,» *CoRR,* 2019.

[175] Y. Zheng, L. Yijia, T. Chuanqi, H. Songfang et H. Fei, «Improving Biomedical Pretrained Language Models with Knowledge,» chez *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021.

[176] M. George, W. Yuanxin, K. Hussam, C. Helen et W. Alexander, «UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus,» chez *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[177] Y. He, Z. Zhu, Y. Zhang, Q. Chen et J. Caverlee, «Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognitio,» chez *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[178] M. E. Peters, M. Neumann, R. Logan, R. Schwartz, V. Joshi, S. Singh et N. A. Smith, «Knowledge Enhanced Contextual Word Representations,» chez *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[179] A. Gajbhiye, N. Al Moubayed et S. Bradley, «ExBERT: An External Knowledge Enhanced {BERT} for Natural Language,» *CoRR,* vol. abs/2108.01589.

[180] R. Speer, J. Chin et C. Havasi, «ConceptNet 5.5: An Open Multilingual Graph of General Knowledge,» *CoRR,* vol. abs/1612.03975, 2016.

[181] T. Khot, A. Sabharwal et P. Clark, «SciTaiL: A Textual Entailment Dataset from Science Question Answering,» chez *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[182] S. R. Bowman, G. Angeli, C. Potts et C. D. Manning, «A large annotated corpus for learning natural language inference,» chez *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

[183] L. Zhongli, Z. Qingyu, L. Chao, X. Ke et C. Yunbo, «Improving BERT with Syntax-aware Local Attention,» chez *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.

[184] T. Xia, Y. Wang, Y. Tian et Y. Chang, «Using Prior Knowledge to Guide BERT's Attention in Semantic Textual Matching Tasks,» chez *Proceedings of the Web Conference*, 2021.

[185] M. de Jong, Y. Zemlyanskiy, N. FitzGerald, F. Sha et W. W. Cohen, «Mention Memory: incorporating textual knowledge into Transformers through entity mention attention,» chez *International Conference on Learning Representations*, 2022.

[186] X. Gezheng, R. Wenge, W. Yanmeng, O. Yuanxin et X. Zhang, «External features enriched model for biomedical question answering,» *BMC Bioinformatics,* 2021.

[187] K. Peng, C. Yin, W. Rong, C. Lin, D. Zhou et Z. Xiong, «Named Entity Aware Transfer Learning for Biomedical Factoid Question Answering,» *IEEE/ACM Transactions on Computational Biology and Bioinformatics.*

[188] J. Mandar, L. Kenton, L. Yi et T. Kristina, «Contextualized Representations Using Textual Encyclopedic Knowledge,» *CoRR,* vol. abs/2004.12006, 2020.

[189] A. Koufakou, E. W. Pamungkas, V. Basile et V. Patti, «HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language,» chez *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 2020.

[190] B. Wang, Z. Zhang, K. Xu, G.-Y. Hao, Y. Zhang, L. Shang, L. Li, X. Chen, X. Jiang et Q. Liu, «DyLex: Incorporating Dynamic Lexicons into BERT for Sequence Labeling,» chez *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[191] M. Neumann, D. King, I. Beltagy et W. Ammar, «ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing,» chez *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019.

[192] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz et L. I. Furlong, «DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants,» *Nucleic Acids Research,* vol. 45, 2017.

[193] R. Apweiler, A. Bairoch, C. Wu, W. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. Martin, D. Natale, C. O'Donovan, N. Redaschi et L. Yeh, «UniProt: the Universal Protein knowledgebase,» *Nucleic Acids Research,* 2004.

[194] J. Amberger, C. Bocchini, F. Schiettecatte, A. Scott et A. Hamosh, «OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders,» *Nucleic Acids Research,* vol. 43, 2015.

[195] S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder et L. J. Jensen, «DISEASES: Text mining and data integration of disease–gene associations,» *Methods,* vol. 74, pp. 83-89, 2015.

[196] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, J. Wiegers, T. C. Wiegers et C. J. Mattingly, «Comparative Toxicogenomics Database (CTD): update 2021,» *Nucleic Acids Research,* vol. 49, 2021.

[197] G. Brown, V. Hem, K. Katz, M. Ovetsky, C. Wallin, O. Ermolaeva, I. Tolstoy, T. Tatusova, K. Pruitt, D. Maglott et T. Murphy, «Gene: a gene-centered information resource at NCBI,» *Nucleic Acids Research,* 2015.

[198] D. Wishart, C. Knox, A. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam et M. Hassanali, «DrugBank: a knowledgebase for drugs, drug actions and drug targets,» *Nucleic acids research,* 2008.

[199] S. Avram, C. G. Bologa, J. Holmes, G. Bocci, T. B. Wilson, D.-T. Nguyen, R. Curpan, L. Halip, A. Bora, J. J. Yang, J. Knockel, S. Sirimulla, O. Ursu et T. I. Oprea, «DrugCentral 2021 supports drug discovery and repositioning,» *Nucleic Acids Research,* vol. 49, 2021.

[200] A. Chang, L. Jeske, S. Ulbrich, J. Hofmann, J. Koblitz, I. Schomburg, M. Neumann-Schaal, D. Jahn et D. Schomburg, «BRENDA, the ELIXIR core data resource in 2021: new developments and updates,» *Nucleic Acids Research,* 2021.

[201] K. Kamal raj, K. Bhuvana et S. Malaikannan, «BioELECTRA:Pretrained Biomedical text Encoder using Discriminators,» chez *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021.

[202] S. Alrowili et V. Shanker, «BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA,» chez *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021.

[203] S. ALAHMARI, D. GOLDGOF et L. HALL, «Challenges for the Repeatability of Deep Learning Models,» *IEEE Access,* 2020.

[204] M. Lewis, Y. Liu et N. Goyal, «BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,» chez *The 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[205] J. Gu, Z. Lu, H. Li et V. Li, «Incorporating Copying Mechanism in Sequence-to-Sequence Learning,» *CoRR,* 2016.

[206] I. Goodfellow, J. Pouget-Abadie et M. Mirza, «Generative Adversarial Networks,» *CoRR,* 2014.

[207] J. Weston, S. Chopra et A. Bordes, «Memory Networks,» *CoRR,* 2014.

[208] Sukhbaatar, S. Sainbayar, W. Arthur et J. Fergus, «End-To-End Memory Networks,» *CoRR,* 2015.

[209] J. Zhou, G. Cui et S. Hu, «Graph Neural Networks: A Review of Methods and Applications,» *CoRR,* 2018.

[210] Q. XiPeng, S. TianXiang, X. YiGe et S. YunFan, «Pre-trained models for natural language processing: A survey,» *CoRR,* 2020.

[211] M. Sarrouti et S. El Alaoui, «A Yes/No Answer Generator Based on Sentiment-Word Scores in Biomedical Question Answering,» *International Journal of Healthcare Information Systems and Informatics (IJHISI),* 2017.

[212] C. Manning, S. Mihai, J. Bauer et J. Finkel, «The Stanford CoreNLP Natural Language Processing Toolkit,» chez *Association for Computational Linguistics*, 2014.

[213] S. Baccianella, A. Esuli et F. Sebastiani, «SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining,» chez *LREC*, 2010.

[214] S. Telukuntla, A. Kapri et W. Zadrozny, «UNCC Biomedical Semantic Question Answering Systems. BioASQ: Task-7B, Phase-B,» *Communications in Computer and Information Science,* 2020.

[215] V. Kommaraju, K. Gunasekaran et T. Bansal, «Unsupervised Pre-training for Biomedical Question Answering,» *CoRR,* 2020.

[216] A. Kazaryan, U. Sazanovich et V. Belyaev, «Transformer-Based Open Domain Biomedical Question Answering at BioASQ8 Challenge,» chez *ceur*, 2020.

[217] W. Adina, N. Nikita et S. Bowman, «A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference,» chez *Association for Computational Linguistics*, 2018.

[218] S. Alrowili et Vijay-Shanker, «Exploring Biomedical Question Answering with BioM-Transformers At BioASQ10B challenge: Findings and Techniques,» chez *ceur*, 2022.

[219] P. Chen et R. Verma, «A Query-Based Medical Information Summarization System Using Ontology Knowledge,» chez *The 19th IEEE Symposium on Computer-Based Medical Systems*, 2006.

[220] Y. Hongyi, Y. Zheng, G. Ruyi et Z. Jiaxing, «BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model,» chez *Association for Computational Linguistics*, 2022.

[221] S. Gururangan, A. Marasović, S. Swayamdipta et K. Lo, «Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks,» chez *Association for Computational Linguistics*, 2020.

[222] Y. Liu et M. Lapata, «Text Summarization with Pretrained Encoders,» *CoRR,* 2019.

[223] R. Luo, L. Sun et Y. Xia, «BioGPT: generative pre-trained transformer for biomedical text generation and mining,» *Briefings in Bioinformatics,* 2022.

[224] C. Clark, K. Lee, M.-W. Chang et T. Kwiatkowski, «BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions,» chez *Association for Computational Linguistics*, 2019.

[225] D. Mollá, M. Santiago-Martínez et A. Sarker, «A corpus for research in text processing for evidence based medicine,» *Language Resources and Evaluation,* 2016.

[226] Z. Kaddari, Y. Mellah, J. Berrich, T. Bouchentouf et M. Belkasmi, «Biomedical Question Answering: A Survey of Methods and Datasets,» chez *Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)*, 2020.

[227] F. Hill, A. Bordes, S. Chopra et J. Weston, «THE GOLDILOCKS PRINCIPLE: READING CHILDREN’S BOOKS WITH EXPLICIT MEMORY REPRESENTATIONS,» chez *ICLR*, 2016.

[228] M. Hoffschmidt, W. Belblidia, T. Brendlé et Q. Heinrich, «FQuAD: French Question Answering Dataset,» *CoRR,* 2020.

[229] M. Louis, M. Benjamin et O. Suárez, «CamemBERT: a Tasty French Language Model,» chez *The 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[230] H. Le, L. Vial, J. Frej et V. Segonne, «FlauBERT: Unsupervised Language Model Pre-training for French,» *CoRR,* 2020.

[231] T. Wolf, L. Debut, V. Sanh et J. Chaumond, «HuggingFace's Transformers: State-of-the-art Natural Language Processing,» *CoRR,* 2020.

[232] A. Conneau, K. Khandelwal, N. Goyal et V. Chaudhary, «Unsupervised Cross-lingual Representation Learning at Scale,» *CoRR,* 2020.

[233] R. Keraron, G. Lancrenon, M. Bras et F. Allary, «Project PIAF: Building a Native French Question-Answering Dataset,» chez *LREC 2020*, 2020.

[234] A. Tchechmedjiev, A. Abdaoui et V. Emonet, «SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes,» *BMC Bioinformatics,* 2018.

[235] C. Mooers, «Information retrieval viewed as temporal signaling,» chez *The International Congress of Mathematicians*, 1952.

[236] A. Otegi, A. Campos et G. Azkune, «Automatic evaluation vs. user preference in neural textual QuestionAnswering,» chez *Association for Computational Linguistics*, 2020.

[237] J. Lee, S. Yi et M. Jeong, «Answering questions on COVID-19 in real-time,» chez *Association for Computational Linguistics*, 2020.

[238] K. Suderman, N. Ide, V. Marc, B. Cochran et J. Pustejovsky, «AskMe: A LAPPS Grid-based NLP query and retrieval system for covid-19 literature,» chez *EMNLP*, 2020.

[239] D. Su, Y. Xu et T. Yu, «Caire-covid: A question answering and query-focused multi-document summarization system for covid-19 scholarly information management,» 2020.

[240] D. Su, Y. Xu et I. Winata, «Generalizing question answering system with pre-trained language model fine-tuning,» chez *Association for Computational Linguistics*, 2019.

[241] A. Esteva, A. Kale, R. Paulus et K. Hashimoto, «Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization,» 2020.

[242] E. Zhang, N. Gupta, R. Nogueira et K. Cho, «Rapidly deploying a neural search engine for the COVID-19 Open Research Dataset,» chez *ACL*, 2020.

[243] S. Lee et J. Sedoc, «Using the poly-encoder for a COVID-19 question answering system,» chez *EMNLP*, 2020.

[244] S. Humeau, K. Shuster, M. Lachaux et J. Weston, «Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring,» 2020.

[245] S. Robertson, «The Probabilistic Relevance Framework: BM25 and Beyond,» 2009.

[246] Oguz, S. Min, P. Lewis et S. Edunov, «Dense passage retrieval for open-domain question answering,» 2020.

[247] R. Tang, R. Nogueira, E. Zhang et N. Gupta, «Rapidly bootstrapping a question answering dataset for covid-19,» 2020.

[248] W. Wang, F. Wei et L. Dong, «Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers,» 2020.

[249] K. Clark, M. Luong, V. Le et C. Manning, «Electra: Pre-training text encoders as discriminators rather than generators,» 2020.

[250] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary et G. Wenzek, «Unsupervised cross-lingual representation learning at scale,» 2020.