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

SPQR@Deft2023: Résolution automatique de QCM médicaux à partir de corpus de domaine et de mesures de similarité

Julien Bezançon[1,2] Toufik Boubehziz[1] Corina Chutaux[1] Oumaima Zine[1]
Laurie Acensio[1] Caroline Koudoro-Parfait[1,4] Andrea Briglia[1,3] Gaël Lejeune[1,2]
(1) STIH, 28 rue Serpente, 75006 Paris, France
(2) CERES, 28 rue Serpente, 75006 Paris, France
(3) UMR 1253 iBrain, 10 Boulevard Tonnellé, 37000 Tours, France
(4) Obtic, SCAI, 4 Pl. Jussieu, 75005 Paris, France
prenom.nom@sorbonne-universite.fr

RÉSUMÉ

Nous présentons le travail de SPQR (Sorbonne Question-Réponses) au DÉfi Fouille de Textes 2023 sur la réponse automatique à des questionnaires à choix multiples dans le domaine de la pharmacologie. Nous proposons une approche fondée sur la constitution de corpus de spécialité et la recherche de phrases similaires entre ces corpus et les différentes réponses possibles à une question. Nous calculons une similarité cosinus sur des vecteurs en n-grammes de caractères pour déterminer les bonnes réponses. Cette approche a obtenu un score maximal en Hamming de 0,249 sur les données de test (0,305 sur le dev) et de 0,0997 en Exact Match Ratio (0,16 sur le dev).

ABSTRACT

SPQR@Deft2023 : Answering automatically to MCQ in the medical domain with similarity measures and domain-specific corpora

We exhibit the approach of the SPQR team in the 2023 French Text Mining Challenge (DEFT). This challenge focused on automatically answering Multiple Choice Questions (MCQ) in the pharmacology domain. We proposed an approach that takes advantage of domain-specific corpora in order to find similarities between possible answers and sentences in the corpora. We compute a cosine similarity on character n-gram vectors to compare them. The best scores we obtained were 0,294 for the Hamming score on the test set (0,305 on the dev set) and 0,997 for the Exact Match ratio (0,16 on the dev set).

MOTS-CLÉS : QCM, FrenchMedMCQA, pharmacologie, similarité, n-grammes de caractères, systèmes de question-réponse.

KEYWORDS: MCQ, FrenchMedMCQA, pharmacology, similarity, character n-grams, question-answering systems.

1 Introduction

Cette nouvelle édition du Défi Fouille de Textes (DEFT) porte sur le corpus FrenchMedMCQA (Labrak *et al.*, 2022a) composé de 3 105 questions fermées, issues des annales d'examens de pharmacie en français. Chaque question possède un identifiant, cinq options de réponses et les corrections. Deux tâches sont proposées dans cette édition du DEFT(Labrak *et al.*, 2022b). La tâche principale proposée consiste en l'identification de réponses correctes parmi cinq réponses possibles

proposées. La tâche annexe propose d'identifier le nombre de réponses (entre 1 et 5) potentiellement correctes pour une question. Nous avons participé aux deux tâches. Le corpus FrenchMedMCQA est découpé en 3 sous-parties. Le corpus d'entraînement (70 % du corpus total), le corpus de développement (10 %) et le corpus de test (20 %). Il était attendu que les performances des systèmes soient évaluées avec la métrique *Exact Match Ratio*, ou taux de réponses parfaitement juste, et le *Hamming Score*, taux de bonnes réponses parmi l'ensemble des réponses données par le système.

Nous faisons l'hypothèse que l'identification des bonnes réponses à une question de manière automatique s'apparente à une recherche de similarité entre les réponses possibles et des données textuelles de référence sur le sujet. Plus précisément, qu'il s'agit de chercher comment nous pouvons montrer que les bonnes réponses à une question sont similaires à des phrases trouvées dans un texte de référence. Pour démarrer nos expériences, nous avons exploré différentes techniques pour constituer des corpus de référence, que nous présentons dans l'article. Nous nous sommes appuyés sur la détection de technolectes biomédicaux dans les questions et les réponses du corpus d'entraînement pour les lier soit avec des définitions issues du manuel Merck (Beers *et al.*, 2008) soit pour interroger l'API OpenAI [1] (Brown *et al.*, 2020). Enfin, nous avons constitué un corpus de manière intrinsèque en transformant les questions et les réponses du jeu d'entraînement en énoncés définitoires. Cet article débute par un état de l'art sur l'évaluation des systèmes de questions-réponses à choix multiple (Section 2) puis nous proposons une description du jeu de données FrenchMed MCQA dans la Section 3. La construction des corpus qui servent de base à nos méthodes est présentée dans la Section 4, nous décrivons les différentes méthodes développées dans la Section 5 puis nous présentons nos résultats et des éléments de discussion dans la Section 6.

2 Méthodes pour les systèmes de question-réponse

Les systèmes de questions-réponses à choix multiple (ou *Multiple Choice Question Answering*) visent à sélectionner la ou les bonne(s) réponse(s) parmi un ensemble d'options données à une question. Cette tâche est particulièrement utilisée dans le domaine de l'éducation (Touissi *et al.*, 2022) (Soares *et al.*, 2021). Ce défi du traitement du langage naturel (TAL) est un problème complexe du fait qu'il nécessite *a priori* une compréhension fine du langage utilisé pour analyser la question d'une part, et, d'autre part, d'appliquer des techniques de résolution de problèmes différents pour sélectionner la ou les bonne(s) réponse(s) parmi plusieurs options. Ainsi, les systèmes de questions-réponses peuvent fonctionner en analysant la typologie des questions (factuelle, liste, booléenne, définition) selon la catégorisation proposée par (Falco, 2012) et en fonction des réponses qu'ils attendent. Concernant la détection des bonnes réponses, les modèles et techniques sont divers : des méthodes de raisonnement telles que le raisonnement multi-sauts (Clark *et al.*, 2018) ou encore le raisonnement logique (Liu & Lee, 2020; Yu *et al.*, 2020; Baggetto *et al.*, 2022). On trouve aussi dans la littérature des travaux s'appuyant sur des sources sémantiques incluant des connaissances généralistes (ou de sens commun) (Talmor *et al.*, 2018; Mihaylov *et al.*, 2018) et des connaissances spécialisées (ou scientifiques) (Clark *et al.*, 2018; Huang *et al.*, 2019). Il est aussi possible d'adopter des méthodes de déduction par l'erreur, sur le principe de l'élimination des réponses, ces méthodes impliquent la sélection de toutes les mauvaises options afin de faciliter la détermination de la bonne réponse. Cette stratégie utilisée par (Kim & Fung, 2020) vise à entraîner le modèle en imitant la stratégie où le répondant exclut intuitivement les options improbables.

1. https ://openai.com/blog/openai-api

Sans être exhaustif, cet état de l'art montre que les modèles existants se spécifient selon la tâche d'application et ne sont pas directement applicables dans le cadre de notre étude. Le contexte de l'étude induit une complexité linguistique qui se caractérise principalement par la compréhension de la question/réponse qui est relativement courte mais fortement spécialisée. La section suivante de cet article propose une description du jeu de données de questions-réponses du DEFT.

3 Description et analyse du jeu de données

Le corpus FrenchMed MCQA est un ensemble de questions à choix multiple extraites d'examens de pharmacologie français. Les questions et les réponses ont été créées manuellement par des experts médicaux et destinés à des étudiants de niveau universitaire.

3.1 Considérations générales sur le corpus

Le jeu de données fourni pour le DEFT 2023 se compose de 2 174 questions pour les données d'entraînement et de 312 questions pour les données de développement. Le tableau 1 présente la distribution des données du corpus d'entraînement et de développement. Nous observons par ailleurs que la majorité des questions comportent 1 à 3 bonnes réponses tandis que les questions ayant l'intégralité des bonnes réponses proposées sont minoritaires. Il est à préciser que le nombre de bonnes réponses n'est pas explicitement indiqué dans la formulation dans la question.

	Apprentissage	**Développement**
Nombre total de mots	130 290	20 428
Nombre total de questions	2 172	313
Nombre total de réponses correctes	5 159	597
Moyenne de réponses correctes par question	2,37	1,9

TABLE 1 – Description de données du Deft 2023

Chaque question est associée à un identifiant, cinq réponses possibles étiquetées de A à E, la ou les réponse(s) correcte(s) et le nombre de réponses correctes. Le tableau 2 représente un exemple des questions et de réponses du corpus d'entraînement.

ID	**Question**	**Réponses**	**Bonnes Réponses**
...	Parmi les substances suivantes, une seule ne traverse pas la barrière placentaire. Laquelle ?	a. Dicoumarine b. Glucose c. Héparine d. Tétracycline e. Amplicilline	c

TABLE 2 – Exemple de question extraite du corpus d'entraînement

3.2 Classification des questions

Les questions du corpus comportent trois parties P_i pour $i = 1, 2, 3$. La séparation entre ces parties est généralement faite avec des virgules. Cela nous donne la forme générale suivante :

$$[P_1][P_2][P_3] \tag{1}$$

— $[P_1]$ est l'entête de la question qui commence souvent par : Parmi les propositions suivantes, cocher la (les) proposition(s) exacte(s), quelle(s) affirmation(s) est(sont) exacte(s), etc.
— $[P_2]$ est donne le type des questions attendues (exacte, inexacte, vraie, fausse) et éventuellement des précisions sur le nombre des réponses possibles.
— $[P_3]$ comporte le contexte (l'information) pharmacologique.

Nous pouvons toutefois avoir d'autres variants à la forme 1 :
— $[P_1 + P_2][P_3]$; $[P_1][P_2 + P_3]$; $[P_3]$
Parfois, une partie $[P_i]$, pour $i = 1 : 3$, est divisée avec des ponctuations : "." "," ":" "?". Nous pouvons alors classer les questions du corpus d'apprentissage en 5 variantes majeures :

$$\begin{cases} [\mathbf{A} + \cdots][\mathbf{NEG/TF/1N/IMP}][\mathbf{MED}] : \\ [\mathbf{A} + \cdots + \mathbf{MED}][\mathbf{NEG/TF/1N/IMP}][\mathbf{MED}] : \\ [\mathbf{A} + \cdots + \mathbf{NEG/TF/1N/IMP}][\mathbf{MED}] : \\ [\mathbf{A} + \cdots + \mathbf{MED}][\mathbf{NEG/TF/1N/IMP}] : \\ [\mathbf{MED}] : \end{cases} \tag{2}$$

Ces variantes sont composées des éléments récurrents suivants :
— **A** : termes utilisés dans la formulation des questions (parmi, donner, indiquer, cocher, on observe, quelle(s), laquelle(s), sélectionner)
— **TF** : termes utilisés dans l'affirmation ou la négation ((in)exacte(s), juste(s), fausse(s), vraie(s))
— **1N** : indication précise sur le nombre des réponses possibles
— **IMP** : indication implicite sur le nombre des réponses possibles
— **MED** : information médicale
— **NEG** : négation
En termes de structure, on peut trouver :

1. Des questions qui peuvent être de type interrogatif :
   Exemple : Parmi les principes actifs suivants, lequel (lesquels) est-il contre-indiqué ou fortement déconseillé d'associer à l'aspirine ?
2. Des questions peuvent être de type assertif :
   Exemple : Parmi les propositions suivantes, indiquer celle qui est exacte. Le crack est une forme : . . .")
3. Des cas où le nombre exact de bonnes réponses n'est pas explicite dans la formulation de la question. Néanmoins, certaines formulations indiquent qu'une réponse unique est attendue :
   Exemple : Parmi les propositions suivantes, une seule est fausse. Indiquez laquelle ? La rifampicine : (. . .)

On peut aussi classer les questions d'une autre manière en cherchant à répertorier les questions en fonction de la ponctuation utilisée en fin d'énoncé (Tableau 3). Ces deux manières de décrire les énoncés des corpus d'entraînement nous ont permis d'aboutir à la construction du Corpus *feedback* qui, aux côtés de trois autres corpus détaillés et explicités dans la partie 4, servira à la constitution du corpus de référence final.

	Terminaison de la question	Nombre de questions
" ?"	(point d'interrogation)	1147
" :"	(deux points)	766
" "	(pas de ponctuation finale)	118
...	(ellipse)	78
"."	(point)	52
","	(virgule)	10

TABLE 3 – Nombre de questions par type de terminaison

4 Construction des corpus de référence

Afin de procéder à la résolution automatique des questions, nous constituons cinq corpus de référence. Ces corpus ont été constitués à partir de diverses ressources (jeu de données d'entraînement du défi, livres numérisés, sites web, ...). L'enjeu est de tester plusieurs ressources médicales afin de déterminer lesquelles sont les plus adaptées à la résolution de la tâche.

Corpus FEEDBACK Nous sommes partis du postulat que les questions et les réponses s'entrecroisaient. Nous avons donc procédé à un test de similarité sur le jeu d'entraînement qui a permis une première validation de l'hypothèse. Ainsi, nous avons réfléchi à la construction d'un corpus à partir du jeu d'entraînement fourni et nous avons formulé, à partir des phrases interrogatives et des réponses correctes, des propositions assertives. Nous avons pu répartir les questions en deux catégories principales : questions de forme affirmatives et questions de forme véritablement interrogative. Le jeu d'entraînement contient 1 147 questions interrogatives et 1 024 affirmatives. Nous avons déterminé trois transformations pour la construction du corpus FEEDBACK en nous appuyant sur cette typologie :

1. Concaténation des questions et des réponses correctes (la réponse complète la question) :
   Parmi les propositions suivantes, indiquer celle qui est exacte. Le crack est une forme :
   Réponse : *De cocaïne*
   Phrase : Le crack est une forme *de cocaïne*.
2. Transformation des interrogatives en subordonnées relatives :
   Dans une des conditions suivantes, l'antigène est tolérogène, laquelle ?
   Réponse : *Administration par voie intra-veineuse*
   Phrase : Lors de *l'administration par voie intraveineuse* l'antigène est tolérogène.
3. La paraphrase :
   Quels sont les toxiques dépresseurs du système nerveux central ?
   Réponse : 1. *Le cannabis*, 2. *Les opiacés*
   Phrase : *Le cannabis et les opiacés* sont les toxiques dépresseurs du système nerveux central.

Cette constitution itérative d'un corpus de référence a été rendue possible par les stratégies de classification des questions que nous avons élaborées. Dans ce contexte, il a été possible de concaténer automatiquement, en appliquant des conditions, les phrases se terminant par " :", ".", "dernier mot", "..." et ",", ce qui a représenté 1024 questions sur un total de 2171. Une classification des phrases d'un point de vue syntaxique nous a permis de les rapprocher au niveau de la forme et de traiter la transformation des interrogatives en subordonnées relatives. En ce qui concerne la paraphrase,

la transformation a été plus difficile à réaliser. Nous nous sommes appuyés sur une classification syntaxique et lexicale, effectuée en amont, pour établir un patron et procéder à la modification.

Corpus FEEDBACK BIS Bien que l'objectif de ce corpus soit similaire à celui du corpus FEEDBACK, qui consiste à créer un corpus à partir du jeu de données d'entraînement, l'approche adoptée présente quelques nuances. Pour créer ce corpus, on va traiter deux types de questions : celles qui comportent les termes 'exacte(s)' ou 'inexacte(s)'. Elles représentent plus de 46% des questions du jeu d'entraînement. À partir de ces questions et de leurs réponses, nous avons formulé des propositions assertives. Pour construire ce corpus, nous avons identifié deux transformations essentielles :

1. Transformation des questions en assertions avec des expressions régulières pour les deux classes (cf. Section 3.2). Les phrases ainsi obtenues sont purgées de la ponctuation ainsi que les terms spécifiques aux questions ("quelles", "lesquelles", "indiquer", "celles", etc.)
2. Concaténation de chaque assertion obtenue à l'étape 1 avec les réponses correctes (pour la classe "exacte") ou les réponses incorrectes (pour la classe "inexacte"), par exemple :

   Parmi les affirmations suivantes, indiquer la (les) affirmation(s) inexacte(s). La spectrofluorimétrie est une technique :
   a : Très sensible,
   b : Applicable à toutes les molécules organiques,
   c : Utilisable pour des dosages quantitatifs,
   d : Ne nécessitant pas de gamme d'étalonnage,
   e : Effectuée sur des solutions congelées à basse température.
   Réponses correctes : b, d, e
   Les assertions ajoutées au corpus :
   a : La spectrofluorimétrie est une technique très sensible,
   c : La spectrofluorimétrie est une technique utilisable pour des dosages quantitatifs.

Contrairement à FEEDBACK, ici chaque question se voit attribuer une assertion distincte pour chaque réponse correcte. Ainsi, nous évitons de regrouper toutes les réponses sous une seule phrase assertive, comme illustré dans l'exemple ci-dessus. Ce corpus comporte 2 383 phrases.

Corpus CHATGPT Afin d'augmenter la taille du corpus FEEDBACK et ainsi élargir la couverture à des termes médicaux absents des données d'entraînement, nous avons choisi de recourir à l'API OpenAI (Brown *et al.*, 2020). Chat GPT a été utilisé pour interroger le modèle GPT-3.5-turbo, une version améliorée du célèbre modèle de langue profond GPT-3, afin de générer des complétions de phrases à partir des assertions du corpus FEEDBACK. Chaque assertion est envoyée au modèle sous forme de requête CURL via l'API qui renvoie un ou plusieurs paragraphes qui viennent compléter l'assertion. Le corpus ainsi généré est présenté sous forme de paires "assertions" + "complétion". Cette approche nous a permis d'élargir la couverture du vocabulaire pharmaceutique lié aux thématiques abordées dans le jeu de données, tout en économisant le temps de recherche de textes pharmaceutiques aléatoires et les coûts liés à la création d'un corpus de phrases médicales. L'exemple suivant présente un extrait du corpus résultant sous forme d'une paire Assertion/complétion :

> "Le crack est une forme de cocaïne" : "le crack est une forme de cocaïne qui est transformée en une substance solide et cristalline, généralement fumée plutôt que sniffée. Le crack est considéré comme une drogue très addictive et dangereuse en raison de son effet rapide et intense sur le système nerveux central."

Corpus MERCK Le corpus Merck se base sur le manuel Merck (Beers *et al.*, 2008). Il s'agit d'un manuel de médecine disponible en ligne et dont nous avons récupéré le contenu. Nous l'avons ensuite découpé en phrases afin de constituer un corpus de référence complémentaire de 66 845 phrases.

Corpus ALL IN ONE Ce corpus est obtenu par concaténation de tous les corpus décrits précédemment. La création de ce corpus va permettre de nourrir notre système de similarité avec des termes techniques propres à la médecine et à la pharmacologie, des technolectes médicaux ainsi que du contexte.

5 Méthodes de résolution automatique

Nous avons procédé à la détection des technolectes médicaux à l'aide d'une similarité *cosinus* entre les phrases de nos corpus de référence et les questions/réponses du défi.

5.1 Méthode de détection des termes médicaux

La détection des technolectes médicaux est une étape essentielle dans la résolution automatique des questions du FrenchMedMCQA. Elle permet d'extraire les termes médicaux des questions et de la base de données de référence. Pour réaliser cette tâche, nous avons développé une approche qui consiste à stocker les mots des questions du corpus d'entraînement considérés comme des technolectes médicaux en utilisant une liste préalablement ajustée de mots courants de la langue française. Après avoir dépassé un certain seuil d'apprentissage, la liste ajustée est remplacée par la liste stockée afin de récupérer un maximum de technolectes récurrents dans le corpus d'entraînement. Cela assure le bon fonctionnement de la méthode de résolution des questions présentée dans la section suivante. Afin d'associer chaque question d'un questionnaire donné (jeu de données cible) avec la (les) bonne(s) réponse(s) correspondante(s), nous proposons une approche qui repose sur l'utilisation d'un corpus de référence (voir Section 4) et le jeu de données cible lui-même. La méthode comprend trois phases :

1. Extraction des technolectes médicaux de chaque phrase du corpus de référence et de chaque paire question/réponse
2. Recherche des réponses correctes à l'aide de mesures de similarité
3. Évaluation des résultats obtenus

Extraction des technolectes L'extraction des termes médicaux s'opère à la fois sur le corpus de référence et le jeu de données ciblé. Une ressource additionnelle de mots courants en français (décrite dans la Section 5.1) a été créée pour mieux détecter les technolectes. Pour le corpus de référence, nous retirons de chaque phrase tokénisée tous ces mots courants, qui ne sont pas *a priori* des technolectes médicaux. Chaque phrase du corpus de référence devient ainsi une séquence de termes du domaine. Pour le jeu de données ciblé, nous commençons par concaténer les questions avec chacune [Q] de leurs réponses possibles $[R_j]$ pour $j = 1 : 5$.

Nous donnons ici un exemple de traitement d'une question à résoudre :

$$\begin{cases} [Q] \ : \textbf{Parmi les éléments suivants, quel est celui qui n'entre pas dans l'uréogenèse ?} \\ [R_1] \ : \textit{CO2} \\ [R_2] \ : \textit{NH3} \\ [R_3] \ : \textit{Valine} \\ [R_4] \ : \textit{ATP} \\ [R_5] \ : \textit{Ornithine} \end{cases} \tag{3}$$

À partir de cette question et des réponses possibles, on obtient :

— [Q] + [R_1], [Q] + [R_2], [Q] + [R_3] ...

Une fois ce travail effectué, nous tokénisons chaque paire question/réponse concaténée et nous filtrons avec la même méthode les mots courants du français.

Mesures de similarité La désignation des réponses correctes pour une question du jeu de données cible est faite à l'aide d'une mesure de similarité de manière à vérifier l'existence d'une phrase similaire du corpus de référence. Si pour une paire question [Q] / réponse ([R_k] où k est la $k^{\text{ème}}$ réponse possible pour $k \in \{1, 2, 3, 4, 5\}$) une phrase similaire existe dans le corpus de référence, nous ajoutons la réponse de cette paire dans la liste des réponses potentiellement correctes à la question. Le test de similarité est réalisé en faisant varier différents paramètres :

— Type de vectorisation (mots, n-grammes de caractères libres ou à l'intérieur des mots ou mots)
— Taille des n-grammes (bi-grammes, tri-grammes, ...)
— Seuil minimal de similarité pour sélectionner la réponse résultat (0.6, 0.7, 0.8, ...)
— Mesure de similarité (Similarités Cosinus Bray-Curtis, Indices de Dice et de Jaccard)

Nous commençons par vectoriser l'ensemble du corpus de référence à l'aide de la librairie SKLEARN avec une vectorisation en fréquence absolue (COUNTVECTORIZER) et avec une pondération Tf-IDf (TFIDFVECTORIZER). Pour chaque paire question/réponse, nous la vectorisons également selon les mêmes paramètres et l'ajoutons à la matrice de vecteurs du corpus de référence. Nous calculons ensuite la similarité *cosinus*, les autres mesures testées ayant des résultats notablement moindres, au contraire de ce qui a été observé dans (Buscaldi *et al.*, 2020). Afin de vérifier que pour une paire question/réponse, il existe une phrase similaire dans le corpus de référence, nous mettons en place différentes méthodes :

— Recherche par seuil (BYSEUIL) : Toutes les réponses dont la similarité est supérieur à un seuil déterminé sont sélectionnées.
— Recherche par le maximum (BYMAX) : Nous sélectionnons simplement la réponse disposant de la plus grande similarité.
— Fusion des deux méthodes précédentes (BYFUSION) : Nous gardons la réponse avec la similarité auxquelles nous ajoutons toutes les réponses dont la similarité dépasse le seuil.

Pour chaque question du jeu de données cible, nous obtenons une liste de bonnes réponses potentielles selon l'une des trois méthodes décrites. Notons que la méthode BYSEUIL peut ne pas sélectionner de réponse dans le cas où aucune réponse n'a une similarité supérieure au seuil choisi. Cette méthode s'étant avérée nettement moins performante, nous n'en présenterons pas les résultats ici. Nous avons testé toutes les combinaisons des paramètres décrits. En plus d'éliminer la méthode BYSEUIL, ceci

Hamming	Paramètres	EMR	Corpus
0.548	BYFUSION_2-3_char_cosine_0.8	0.377	All In One
0.541	BYFUSION_3-3_char_cosine_0.8	**0.388**	All In One
0.532	BYFUSION_3-3_char_cosine_0.7	0.331	All In One
0.532	BYFUSION_2-3_char_cosine_0.9	0.373	All In One
0.518	BYFUSION_2-3_char_cosine_0.7	0.252	All In One
0.505	BYFUSION_3-3_char_cosine_0.6	0.197	All In One
0.486	BYFUSION_2-3_char_cosine_0.6	0.094	All In One
0.475	BYFUSION_3-3_char_cosine_0.8	0.352	Feedback bis
0.468	BYFUSION_2-3_char_cosine_0.8	0.332	Feedback bis
0.462	BYFUSION_3-3_char_cosine_0.9	0.258	All In One

TABLE 4 – 10 meilleurs résultats en Hamming sur le jeu d'entraînement (vectorisation en bi-grammes et tri-grammes de caractères).

nous a amené à écarter la pondération Tf-Idf (moins efficace que la simple valeur absolue) ainsi que les représentations en mots. La section suivante présente en plus en détails les résultats obtenus.

6 Résultats et discussion

La chaîne de traitement décrite dans la section 5 nous a permis de chercher les meilleures combinaisons de paramètres pour les deux mesures du défi, le *Hamming Score* et le *Exact Match Ratio*. Tout d'abord, nous avons pu voir que vectoriser en bi-grammes/tri-grammes de caractères était, avec notre méthode, systématiquement plus efficace que de vectoriser en mots (2 à 5 points de pourcentages selon les cas). Le Tableau 4 présente les dix meilleurs résultats obtenus sur le jeu d'entraînement fourni par les organisateurs. Nous avons atteint un score de Hamming maximal de 54 % pour un taux d'exactitude de 37 %. Ces résultats semblaient très prometteurs comparés à ceux obtenus par (Labrak *et al.*, 2022a). En ce qui concerne la tâche annexe, nous obtenons une précision de 49 % pour un F1_macro de 43 %. avec les paramètres BYFUSION_2-3_CHAR_COSINE_0.8 et le corpus All In One.

Les meilleurs résultats viennent de la méthode BYFUSION. Ce résultat n'est pas surprenant puisque les résultats obtenus avec la méthode BYMAX amènent une seule réponse par question (favorisant la précision) et que la méthode BYSEUIL amenait elle un meilleur rappel. La fusion de ces deux méthodes semble donc être l'approche la plus adaptée à la recherche de bons résultats. Le corpus permettant les meilleurs résultats est sans surprise le corpus *All In One*, contenant l'ensemble des corpus assemblés. Le Tableau 5 présente cette fois les meilleurs résultats obtenus avec le jeu de données de développement.

Nous constatons immédiatement des scores inférieurs aussi bien avec Hamming qu'avec l'EMR. Cela est certainement dû à un sur-apprentissage venant de nos corpus issus du jeu de données d'entraînement. De plus, nous remarquons que le corpus le plus productifs n'est plus le corpus *All In One*. Il s'agit des corpus Feedback bis et Feedback. Les ressources extérieures créées sont donc moins productives que les ressources assemblées à partir du jeu de données d'entraînement. Pour la tâche annexe, nous obtenons une précision de 25 % pour un F1_macro de 19 %. avec les paramètres BYFUSION_2-3_CHAR_COSINE_0.6 et le corpus All In One.

Enfin, pour le jeu de données test, nous avons obtenu un Hamming de 24,93 % et un EMR de 8,52 % lors de l'évaluation officielle. Pour la tâche annexe, nous avons obtenu une précision de 23 % et un

Hamming	Paramètres	EMR	Corpus
0.303	BYFUSION_2-3_char_cosine_0.6	0.051	All In One
0.293	BYFUSION_2-3_char_cosine_0.6	0.112	Feedback bis
0.279	BYFUSION_2-3_char_cosine_0.6	0.041	Merck
0.277	BYFUSION_3-3_char_cosine_0.6	0.147	Feedback bis
0.266	BYFUSION_2-3_char_cosine_0.7	0.150	Feedback bis
0.265	BYFUSION_3-3_char_cosine_0.6	0.099	All In One
0.261	BYFUSION_3-3_char_cosine_0.6	0.150	Feedback
0.258	BYFUSION_3-3_char_cosine_0.7	0.157	Feedback
0.254	BYFUSION_2-3_char_cosine_0.6	0.102	Feedback
0.253	BYMAX_3-3_char_cosine	**0.160**	Feedback

TABLE 5 – 10 meilleurs résultats en Hamming sur le jeu de développement (vectorisation en bi-grammes et tri-grammes de caractères).

Hamming	Paramètres	EMR	Corpus
0.329	BYFUSION_2-3_char_cosine_0.6	0.056	All In One
0.301	BYFUSION_2-3_char_cosine_0.6	0.078	Merck
0.270	BYFUSION_2-3_char_cosine_0.6	0.074	Feedback bis
0.269	BYFUSION_2-3_char_cosine_0.6	0.099	Feedback
0.249	BYFUSION_3-3_char_cosine_0.6	0.085	All In One
0.243	BYFUSION_2-3_char_cosine_0.7	0.090	All In One
0.239	BYFUSION_3-3_char_cosine_0.6	0.099	Feedback bis
0.237	BYFUSION_2-3_char_cosine_0.7	0.117	Feedback
0.231	BYMAX_2-3_char_cosine	**0.120**	Feedback
0.231	BYFUSION_2-3_char_cosine_0.9	**0.120**	Feedback

TABLE 6 – 10 meilleurs résultats en Hamming sur le jeu de test (vectorisation en bi-grammes et tri-grammes de caractères), en gris les résultats envoyés pour le défi

F1_macro de 43 %. Les meilleurs résultats obtenus sur le jeu de données de test sont présentés dans le Tableau 6.

Nous remarquons que nous aurions pu obtenir de meilleurs résultats que ceux obtenus lors du défi avec un choix plus judicieux du jeu de paramètre. Tout comme pour le jeu de développement, nous observons que les corpus les plus productifs sont les corpus assemblés à partir du jeu d'entraînement.

En résumé, nous avons présenté un système de question réponse fondé sur la recherche de similarités de phrases. Cette méthode ne se fonde pas sur un apprentissage et ne nécessite qu'un, ou plusieurs, corpus de référence du domaine. Notre méthode cherche la réponse la plus vraisemblable en recherchant dans des corpus spécialisés la ou les phrases les plus similaires aux réponses possibles. Nos corpus de référence ont été constitué de différentes manières : un corpus d'énoncés assertifs reconstruits à partir des données d'entraînement (corpus FEEDBACK), un corpus d'énoncés issus de l'interrogation de ChatGPT (corpus CHATGPT) ainsi que deux corpus issus de données disponibles en lignes (corpus MERCK et corpus CHATGPT). Si cette méthode simple n'obtient pas des résultats aussi élevés que des méthodes supervisées fondées notamment sur du *deep learning*, elle est assez interprétable puisque l'on peut sans rétro-ingénierie complexe retrouver les segments, phrases ou paragraphes par exemple, qui ont présidé au choix de telle ou telle réponse.

Références

BAGGETTO P., RAMOS S., GARCIA J. & NAVARRO J. R. (2022). Study on text comprehension and mcqa in spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), A Coruna, Spain. CEUR Workshop Proceedings. the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, volume 1, p. 4171–4186.

BEERS M. H., PORTER R. S., JONES T. V., KAPLAN J. L. & BERKWITS M. (2008). *Le Manuel Merck de diagnostic et thérapeutique*. Edition d'Après, 4e édition.

BROWN T. B., MANN B., RYDER N., SUBBIAH M., KAPLAN J., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. In *Proceedings of the 37th International Conference on Machine Learning*.

BUSCALDI D., FELHI G., GHOUL D., LE ROUX J., LEJEUNE G. & ZHANG X. (2020). Calcul de similarité entre phrases : quelles mesures et quels descripteurs ?(sentence similarity : a study on similarity metrics with words and character strings). In *Traitement Automatique des Langues Naturelles (TALN, 27e édition). Atelier DÉfi Fouille de Textes*, p. 14–25.

CLARK P., COWHEY I., ETZIONI O., KHOT T., SABHARWAL A., SCHOENICK C. & TAFJORD O. (2018). Think you have solved question answering ? try arc, the ai2 reasoning challenge. DOI : 10.48550/ARXIV.1803.05457.

FALCO M.-H. (2012). Typologie des questions à réponses multiples pour un système de question-réponse (typology of multiple answer questions for a question-answering system)[in french]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 3 : RECITAL*, p. 191–204.

HUANG L., BRAS R. L., BHAGAVATULA C. & CHOI Y. (2019). Cosmos qa : Machine reading comprehension with contextual commonsense reasoning. DOI : 10.48550/ARXIV.1909.00277.

KIM H. & FUNG P. (2020). Learning to classify the wrong answers for multiple choice question answering (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**(10), 13843–13844. DOI : 10.1609/aaai.v34i10.7194.

LABRAK Y., BAZOGE A., DUFOUR R., DAILLE B., GOURRAUD P., MORIN E. & ROUVIER M. (2022a). Frenchmedmcqa : A french multiple-choice question answering dataset for medical domain. p. 41–46.

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

LIU C.-L. & LEE H.-Y. (2020). Unsupervised multiple choices question answering : Start learning from basic knowledge. *arXiv preprint arXiv :2010.11003*.

MIHAYLOV T., CLARK P., KHOT T. & SABHARWAL A. (2018). Can a suit of armor conduct electricity ? a new dataset for open book question answering. DOI : 10.48550/ARXIV.1809.02789.

SOARES T. G., AZHARI A., ROKHMAN N. & WONARKO E. (2021). Education question answering systems : a survey. In *Proceedings of The International MultiConference of Engineers and Computer Scientists*.

TALMOR A., HERZIG J., LOURIE N. & BERANT J. (2018). Commonsenseqa : A question answering challenge targeting commonsense knowledge. DOI : 10.48550/ARXIV.1811.00937.

TOUISSI Y., HJIEJ G., HAJJIOUI A., IBRAHIMI A. & FOURTASSI M. (2022). Does developing multiple-choice questions improve medical students' learning ? a systematic review. *Medical Education Online*, **27**(1), 2005505.

YU W., JIANG Z., DONG Y. & FENG J. (2020). Reclor : A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv :2002.04326*.