Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne DEFT 2021

Cyril Grouin[1] Natalia Grabar[2] Gabriel Illouz[1]
(1) Université Paris Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400 Orsay, France
(2) Univ. Lille, CNRS, UMR 8163 - STL - Savoirs Textes Langage, 59000 Lille, France
prenom.nom@lisn.upsaclay.fr, prenom.nom@univ-lille.fr

RÉSUMÉ

Le défi fouille de textes (DEFT) est une campagne d'évaluation annuelle francophone. Nous présentons les corpus et baselines élaborées pour trois tâches : (i) identifier le profil clinique de patients décrits dans des cas cliniques, (ii) évaluer automatiquement les réponses d'étudiants sur des questionnaires en ligne (Moodle) à partir de la correction de l'enseignant, et (iii) poursuivre une évaluation de réponses d'étudiants à partir de réponses déjà évaluées par l'enseignant. Les résultats varient de 0,394 à 0,814 de F-mesure sur la première tâche (7 équipes), de 0,448 à 0,682 de précision sur la deuxième (3 équipes), et de 0,133 à 0,510 de précision sur la dernière (3 équipes).

ABSTRACT

Clinical cases classification and automatic evaluation of student answers : Presentation of the DEFT 2021 Challenge

DEFT is an annual French-speaking text mining challenge. We present the corpora, tasks, and baselines we produced : (i) identify the clinical profile of patients described in clinical cases, (ii) automatically assess student answers from online survey (Moodle) based on the teacher's correction, and (iii) continue to evaluate student answers based on answers already assessed by the teacher. Results ranged from 0.394 to 0.814 F-score on the first task (7 participants), 0.448 to 0.682 accuracy on the second one (3 participants), and 0.133 to 0.510 accuracy on the last one (3 participants).

MOTS-CLÉS : Extraction d'information, cas cliniques, réponses courtes d'étudiants.

KEYWORDS: Information extraction ; Clinical cases ; Short Answer Grading ; Student answers.

1 Introduction

Le défi fouille de textes (DEFT) est une campagne d'évaluation annuelle francophone qui permet à plusieurs équipes de confronter leurs méthodes sur une ou plusieurs tâches régulièrement renouvelées. Pour cette nouvelle édition, nous proposons deux thématiques principales. La première concerne le domaine clinique au travers d'une tâche d'extraction d'information depuis des cas cliniques. La deuxième concerne l'enseignement et le traitement de copies d'étudiants avec deux tâches sur l'évaluation automatique des réponses d'étudiants dans des questionnaires en ligne de type Moodle [1].

Les tâches d'extraction d'information constituent une première étape d'accès aux informations pré-

1. https://moodle.org

sentes dans des documents, pour des objectifs plus généraux (recherche de cas similaires, résumé automatique, etc.). A l'image des campagnes de repérage d'entités nommées, nous proposons de repérer les maladies, signes et symptômes des patients décrits dans des cas cliniques, dans la perspective de dresser le profil clinique des patients. Cette tâche fait suite aux précédentes éditions sur l'identification d'informations démographiques et cliniques (Grabar *et al.*, 2019), et l'extraction d'information fine autour des patients, de la pratique clinique, des traitements et du temps (Cardon *et al.*, 2020). L'évaluation automatique de réponses d'étudiants constitue une tâche originale qui n'a jamais été abordée dans les campagnes DEFT. Elle voit son utilité dans l'assistance automatique lors de la correction des copies et dans la comparaison qualitative entre une copie et une référence.

Organisation de la compétition. Les participants ont pu s'inscrire à la compétition et accéder aux données d'entraînement à partir du 12 février 2021. Ils ont eu accès aux scripts d'évaluation officiels le 26 avril. La phase de test s'est déroulée entre le 17 et le 23 mai, sur une période de trois jours choisie par chaque équipe de participants. L'atelier de clôture s'est déroulé le 28 juin 2021. Quatorze équipes se sont inscrites, toutes françaises, parmi lesquelles quatre entreprises (y compris trois spécifiques au domaine médical), une équipe hospitalo-universitaire, trois équipes formées d'étudiants uniquement (niveau master ou doctorat), et six équipes académiques. Trois équipes ont abandonné la compétition avant la phase de test et une équipe a abandonné la compétition pendant la phase de test.

Dans le contexte du règlement général européen sur la protection des données [2] (RGPD), nous relevons l'intérêt des acteurs du domaine clinique pour accéder à des données de type clinique. Nous observons que la majorité des participants s'intéresse aux cas cliniques (cinq équipes académiques : DOING, Orléans ; ISME, Grenoble ; LIRMM, Montpellier ; QUEER, Paris ; Team Stel, Paris, et deux entreprises : BL.Santé, Toulouse ; Everteam Lab, Lyon). Trois équipes se sont consacrées à l'évaluation des réponses d'étudiants, avec un industriel (EDF Lab, Palaiseau) et deux équipes d'étudiants en master (Nantalco, Univ. Nanterre et INaLCO) ou en thèse (Proofreaders, Univ. Nantes).

2 Corpus

2.1 Domaine clinique

Concernant le domaine clinique, nous reprenons le corpus de cas cliniques (Grabar *et al.*, 2018; Grouin *et al.*, 2019) que nous avons proposé aux participants les années passées. Le corpus d'entraînement se compose des données d'entraînement et de test de DEFT 2020 (soit 167 cas cliniques), tandis que le corpus de test rassemble 108 nouveaux cas. Les annotations des campagnes de 2019 et 2020 sont mises à disposition comme informations complémentaires mais leur utilisation reste facultative.

Afin de préparer les données de référence de 2021, nous avons annoté toutes les maladies, signes ou symptômes, correspondant à un descripteur français du MeSH (Medical Subject Headings) (NLM, 2001), en prenant comme label l'un des vingt-six axes de l'arborescence du chapitre C uniquement [3]. Cependant, vingt-trois axes sont présents et annotés dans le corpus (voir tableau 1).

Le corpus se compose d'un total de 275 cas cliniques annotés sous BRAT (Stenetorp *et al.*, 2012)

2. https://www.cnil.fr/fr/reglement-europeen-protection-donnees
3. Le chapitre C renvoie aux maladies, voir http://mesh.inserm.fr/FrenchMesh/view/index.jsp

C01	Infections bactériennes et mycoses	C13	Maladies de l'appareil urogénital féminin et complications de la grossesse
C02	Maladies virales	C14	Maladies cardiovasculaires
C03	Maladies parasitaires	C15	Hémopathies et maladies lymphatiques
C04	Tumeurs	C16	Malformations et maladies congénitales, héréditaires et néonatales
C05	Maladies ostéomusculaires	C17	Maladies de la peau et du tissu conjonctif
C06	Maladies de l'appareil digestif	C18	Maladies métaboliques et nutritionnelles
C07	Maladies du système stomatognathique	C19	Maladies endocriniennes
C08	Maladies de l'appareil respiratoire	C20	Maladies du système immunitaire
C09	Maladies oto-rhino-laryngologiques	C23	États, signes et symptômes pathologiques
C10	Maladies du système nerveux	C25	Troubles dus à des produits chimiques
C11	Maladies de l'œil	C26	Plaies et blessures
C12	Maladies urogénitales de l'homme		

TABLE 1 – Les axes du MESH et les types de maladies, signes et symptômes annotés en 2021

autour de trois dimensions : les informations démographiques et cliniques [4] (DEFT 2019), les informations cliniques fines [5] (DEFT 2020), et les types de maladies (DEFT 2021). Dans nos annotations et dans la référence, les axes sont désignés au moyen d'un mot-clé représentatif (« infections » pour l'axe C01, « homme » en C12, « femme » en C13, « hemopathies » en C15, « peau » en C17, etc.).

Nous présentons dans le tableau 2 un extrait de cas clinique (fichier filepdf-144-2-cas) avec l'ensemble des axes du MeSH à identifier pour ce fichier. Les pathologies permettant d'identifier les différents axes sont mises en italiques dans le texte (par exemple, les problèmes immunitaires, hémopathiques, et de tumeur de ce patient sont identifiables par la mention d'un myélome dans le texte).

Texte	Axes associés aux pathologies identifiées		
Mr. E., âgé de 70 ans, était suivi depuis cinq ans pour un *myélome*, une *polyarthrite rhumatoïde*, et une *amylose rectale et rénale*. Alors qu'il n'avait pas de troubles mictionnels anciens, il a présenté un *épisode de rétention aiguë d'urine*.	immunitaire	hemopathies	tumeur
	peau	osteomusculaires	
	nutritionnelles	digestif	
	homme		

TABLE 2 – Extrait de cas clinique avec tous les axes du MeSH à identifier pour ce fichier (les axes sont inscrits en vis à vis de la pathologie présente dans le texte)

Le tableau 3 présente le nombre d'annotations par classe dans le corpus total de 275 cas cliniques.

4. Soit 4 types d'information : âge ; genre ; origine / motif de la consultation ou de l'hospitalisation ; issue du traitement

5. Soit 12 classes parmi quatre domaines : (i) anatomie ; (ii) examen, pathologie, signe ou symptôme ; (iii) substance, dose, durée, fréquence, mode d'administration, traitement (chirurgical ou médical) ; et (iv) date, moment

DEFT 2019	âge	271	genre	276	origine	178	issue	180
DEFT 2020	anatomie	4780	examen	3355	pathologie	764	sosy	5240
	substance	2009	dose	562	durée	375	fréquence	383
	mode	484	traitement	1311	date	250	moment	970
DEFT 2021	blessures	13	cardiovasc.	103	chimiques	89	digestif	91
	endocrinien	30	état/sosy	546	femme	134	génétique	33
	hémopathies	75	homme	199	immunitaire	34	infections	53
	nerveux	109	nutrition	43	œil	21	ORL	11
	ostéomuscul.	44	parasitaire	11	peau	84	respiratoire	70
	stomato.	12	tumeur	276	virales	14		
inutilisées	fonction	21	organisme	20	poids	5	taille	3
	température	9	valeur	1743				

TABLE 3 – Nombre d'annotations par classe dans le corpus DEFT 2021 (275 cas cliniques)

2.2 Réponses d'étudiants

Des évaluations pour les réponses courtes d'étudiants existent en langue anglaise (Mohler & Mihalcea, 2009). Le lecteur intéressé trouvera un état de l'art dans Burrows *et al.* (2015). Des campagnes d'évaluation ont été menées sur des données en anglais, telles que les compétitions Kaggle [6] ou certaines éditions de SemEval (Dzikovska *et al.*, 2013). Les dispositifs d'aide à la correction ont aussi donné lieu à des expériences. Le but, contrairement à mettre une note en ayant la correction, consiste à trouver comment corriger par regroupement de réponses (Basu *et al.*, 2013; Horbach *et al.*, 2014).

Nous avons constitué un corpus d'une centaine d'énoncés produits par des étudiants en informatique pendant les contrôles de programmation web et de bases de données. Les identités des étudiants ont été anonymisées. Ces énoncés sont composés de questions ouvertes et fermées. Nous distinguons les questions qui impliquent l'écriture du code informatique (« *Modifiez le code XML ci-dessous pour le rendre valide* ») de celles qui appellent une réponse en langue naturelle (« *À quoi sert l'attribut alt de la balise <img> ?* »). Nous avons réparti ces deux types de questions entre les corpus d'entraînement et de test. Les énoncés sont accompagnés de la correction de l'enseignant et des réponses produites par une cinquantaine d'étudiants en moyenne par question. Les énoncés ont été collectés sur deux années d'enseignement. Comme indiqué, ces énoncés correspondent aux réponses formulées sur des questionnaires en ligne. Ils intègrent des balises XML de mise en forme pour l'affichage.

Le tableau 4 présente un extrait du fichier de questions du corpus d'entraînement [7], avec une question en langue naturelle (numéro 1001) et une question de code (numéro 2045). Pour chaque question, une ou plusieurs corrections de l'enseignant sont associées. Un commentaire de l'enseignant peut également être présent pour pénaliser certaines réponses (sur la question 2045, l'enseignant attribue 0,5 point si la réponse fournie est partiellement exacte).

Le tableau 5 présente un ensemble de réponses faites par les étudiants, sans correction orthographique, aux questions du tableau 4. En cas d'absence de réponse, la mention « NO_ANS » est indiquée.

Toutes les notes attribuées aux étudiants ont été normalisées sur un point avec une décimale conservée.

6. https://www.kaggle.com/spscientist/students-performance-in-exams et https://www.kaggle.com/c/asap-sas

7. Le corpus distribué aux participants contient cinq colonnes : l'identifiant de la question, la note maximale d'origine, le numéro de question (similaire à l'identifiant), et la correction de l'enseignant.

Id	Question	Correction ou commentaire de l'enseignant
1001	Qu'est-ce que le World Wide Web ?	système hypertexte fonctionnant sur internet = une des applications d'internet, comme courrier électronique, messagerie instantanée...
2045	Pourquoi le code HTML suivant ne respecte-t-il pas les principes d'accessibilité de WCAG ? <code> Site de la RATP </code>	car la légende de l'image ne lui est pas associée (avec un figcaption par exemple) .5 pour ceux qui ont dit que le texte alternatif n'était pas suffisamment précis

TABLE 4 – Fichier de questions avec correction et commentaire de l'enseignant

Id	Note	Etudiant	Réponse de l'étudiant
1001	0.5	student101	Ce sont les pages web accessible par tout navigateur.\n
1001	0	student108	Un réseau mondial \n
1001	1	student3	C'est le systeme hypertexte qui sert à consulter des documents et des pages hébergés sur le réseau internet\n
1001	0	student95	NO_ANS
2045	0	student101	Les mal-voyant ne peuvent pas y accéder.
2045	1	student109	L'image n'a pas de légende. Les malvoyants ne pourront pas savoir qu'il y a une image.\nL'utilisation de la balise <figcaption>logo RATP</figcaption> aurait permit de respecter le principe d'accessibilité.\nComme alt peut etre lu par les lecteur d'ecran, changer "RATP" en "Logo RATP" pour plus de comprehension.
2045	0.2	student42	Il n'y a pas une description qui décrit l'image
2045	0.8	student70	Le texte par défaut de l'image ne décrit pas l'image précisément.

TABLE 5 – Fichier de réponses des étudiants avec note associée

Afin de comprendre la valeur de certaines notes (0.2 ou 0.8 sur la question 2045), la note maximale d'origine est indiquée dans le fichier de questions (généralement comprise entre 1 et 2,5).

3 Description des tâches

3.1 Tâche 1 : identification du profil clinique du patient

La tâche vise à identifier le profil clinique du patient décrit dans chaque cas. Cela revient à normaliser les pathologies décrites, en identifiant l'axe correspondant du MeSH pour le chapitre C (voir tableau 1). Si une pathologie renvoie à plusieurs axes, tous devront être identifiés. Le choix entre « maladies de l'appareil urogénital féminin et complications de la grossesse » (C13) et « maladies urogénitales de l'homme » (C12) dépend du genre de la personne dont le cas est décrit. Nous fournissons les annotations des éditions DEFT 2019 et 2020 (voir tableau 3) comme aide facultative.

3.2 Tâches autour des réponses d'étudiants

Nous proposons deux tâches autour de l'évaluation de réponses d'étudiants en considérant un enseignant qui souhaiterait améliorer la qualité de son évaluation tout en économisant le temps passé à cette activité. Nous considérons deux situations : (i) celle où l'enseignant dispose déjà des corrections et souhaite développer un système d'évaluation automatique, et (ii) celle où il n'existe pas encore de correction, mais où les premières réponses évaluées permettent déjà de se faire une idée des réponses et des notes à leur associer. La première situation vise à fournir une base, l'enseignant vérifiant la pertinence de l'évaluation automatique. La deuxième évite à l'enseignant de perdre du temps à évaluer les réponses proches, en associant à ces réponses la note des réponses proches déjà corrigées.

3.2.1 Tâche 2 : évaluation automatique de copies d'après une référence existante

À partir d'une liste de questions avec corrections et commentaires de l'enseignant (tableau 4) et d'une liste de réponses d'étudiants à ces questions (tableau 5), l'objectif consiste à évaluer et à noter (sur un point) les réponses des étudiants, en se fondant sur la correction de l'enseignant.

3.2.2 Tâche 3 : poursuite de l'évaluation de réponses à partir de premières évaluations

À partir d'une liste de questions sans aucune correction de l'enseignant (la dernière colonne du tableau 4 est systématiquement indiquée « NO_CORR ») et d'une liste de réponses d'étudiants à ces questions avec de premières notes fournies, l'objectif consiste à évaluer et à noter (sur un point) les réponses des étudiants qui n'ont pas encore été corrigées, à partir des réponses déjà évaluées (la deuxième colonne du tableau 5 comprend, pour une minorité de réponses, la note de l'enseignant, et pour une majorité de réponses la mention « A_CORRIGER »). Le tableau 6 fournit le nombre de questions et de réponses proposées dans chaque tâche dans les corpus d'entraînement et de test.

Corpus	Tâche 2		Tâche 3		Total
	train	test	train	test	
Questions	50	21	11	6	87
Réponses	3820	1644	769	387	6620

TABLE 6 – Nombre de questions et de réponses par corpus pour chaque tâche

Pour le corpus de test de la troisième tâche, nous fournissons 5 % des réponses déjà corrigées pour trois questions (5005, 5011, 5012), et 10 % des réponses déjà corrigées pour les trois autres (5001, 5009, 2012). Les réponses déjà corrigées ont été aléatoirement choisies. Pour que les participants puissent se mettre dans les conditions du test, nous avons fourni le corpus d'entraînement en deux versions : une version avec toutes les notes, et une version avec 5 ou 10 % des réponses déjà corrigées.

4 Baselines

Tâche 1. Cette baseline repose sur les annotations du corpus d'entraînement dont nous avons extrait une centaine de concepts représentatifs, jusqu'à 21 concepts par axe. Certains concepts sont communs

à plusieurs axes. Les concepts peuvent renvoyer à des pathologies (*Pott's Puffy tumor, asthme, brugada, dermatite, pneumonie*), des symptômes (*agitation, fatigue, rash, toux*), des adjectifs ou noms de parties anatomiques (*hépatique, poumon, pulmonaire, rectal, rein, rénal*), ou aux informations de genre (*femme, fille, patiente, vagin, homme, pénis*). Nous présentons quelques axes et leurs concepts :

- infections (C01) : Pott's Puffy tumor, tuberculose
- ostéomusculaire (C05) : fatigue, ostéolyse, ostéosarcome, Pott's Puffy tumor
- œil (C11) : mydriase, myosis, Pott's Puffy tumor
- immunitaire (C20) : allergie, urticaire, VIH
- chimique (C25) : anticholinergique, cirrhose, datura, intoxication

Pour chaque concept identifié dans un document, nous conservons l'axe ou les axes associés. Sur le corpus d'entraînement, nous obtenons une F-mesure globale de 0,568 (rappel de 0,468 et précision de 0,723) et pour le test, une F-mesure globale de 0,546 (rappel de 0,416 et précision de 0,796).

Tâche 2. Cette baseline consiste à compter le nombre de mots en commun entre (1) les mots de la réponse de l'étudiant et (2) les mots de la question et de la réponse de l'enseignant. Cette comparaison repose sur des mots mis en minuscules, sans chiffre ni ponctuation, d'au-moins quatre caractères. Ce décompte est divisé par le nombre de mots conservés dans la question et la réponse de l'enseignant pour produire un score, multiplié par deux (pour compenser le nombre réduit de mots et les fautes d'orthographe). Ce score est normalisé : valeur finale de 1 si le score est supérieur ou égal à 0,5 ; 0,5 si supérieur ou égal à 0,4 ; les autres valeurs sont conservées. Par exemple :

- Question 2032 normalisée : quel est intérêt utiliser du code ajax
- Réponse et commentaire enseignant normalisés : permet échange de données avec le serveur sans mise à jour complète de la page ok pour permet de màj une partie de la page sans avoir à la recharger complètement.
- Réponse normalisée (student7) : utiliser du code ajax permet de mettre à jour certaines parties une page web sans recharger toute la page.
- Soit 5 mots en communs (*jour, page, permet, recharger, sans*) sur 17 mots conservés dans la question et réponse enseignant. Score de $(5/17) \times 2 = 0,59$ normalisé à 1 (score final).

Sur le corpus d'entraînement, nous obtenons une précision de 0,484 (1 847 évaluations correctes pour 1 973 incorrectes) et sur le test, une précision de 0,477 (785 évaluations correctes et 859 incorrectes).

Tâche 3. Nous avons produit deux baselines pour la tâche 3. La première s'inspire de celle utilisée pour la tâche 2, en comptant le nombre de mots communs entre les réponses déjà corrigées et celles restant à évaluer, pour attribuer la note de la réponse corrigée la plus proche. Sur le corpus d'entraînement, nous obtenons une précision de 0,439 (corrélation de 0,60) et une précision de 0,397 sur le test (corrélation de 0,41). La deuxième baseline est une implémentation des k plus proches voisins. Sur le corpus d'entraînement, nous obtenons une précision de 0,561 (corrélation de 0,62) et une précision de 0,462 sur le test (corrélation de 0,58).

5 Résultats

Les performances des systèmes ont été évaluées en termes de rappel, précision et F-mesure :

$$\text{Rappel} = \frac{\text{prédictions correctes}}{\text{prédictions attendues}} \qquad \text{Précision} = \frac{\text{prédictions correctes}}{\text{prédictions réalisées}} \qquad \text{F-mesure} = \frac{2 \times \text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

Tâche 1. Le tableau 7 présente les résultats (rappel, précision, F-mesure) des participants sur la tâche 1, classés par F-mesure décroissante, ainsi que le rang dans le classement final. Sept équipes ont participé à cette tâche, pour une F-mesure moyenne de 0,636 et une médiane de 0,700.

Equipe	Run	R	P	F	Rang
DOING (Hiot *et al.*, 2021)	2	0,750	**0,888**	**0,814**	1
	1	0,713	0,873	0,785	–
QUEER (Dupont *et al.*, 2021)	1	0,734	0,838	0,782	2
Team Stel (Gérardin *et al.*, 2021)	1	0,874	0,696	0,775	3
	2	**0,895**	0,677	0,771	–
	3	0,872	0,689	0,770	–
QUEER	2	0,684	0,843	0,755	–
	3	0,677	0,819	0,741	–
DOING	3	0,769	0,686	0,725	–
Advanse LIRMM	2	0,637	0,783	0,703	4
	1	0,627	0,784	0,697	–
BL.Santé (Billami *et al.*, 2021)	3	0,730	0,558	0,633	5
	1	0,677	0,570	0,619	–
	2	0,471	0,786	0,589	–
Everteam Lab (Bailly *et al.*, 2021)	1	0,683	0,370	0,480	6
ISME (Mannion *et al.*, 2021)	2	0,423	0,496	0,457	7
	3	0,398	0,439	0,417	–
	1	0,390	0,444	0,416	–
Everteam Lab	3	0,637	0,298	0,406	–
	2	0,651	0,283	0,394	–

TABLE 7 – Résultats et classement sur la tâche 1. Les meilleurs résultats sont en gras

La figure 1 présente les valeurs moyennes de rappel, précision, et F-mesure, calculées sur l'ensemble des soumissions, pour chacun des axes du chapitre C du MeSH, classées par F-mesure croissante. Les axes les mieux identifiés par les participants, en termes de F-mesure moyenne, sont : signes ou symptômes (F=0,955) ; maladies parasitaires (F=0,821) ; maladies urogénitales de l'homme (F=0,765) ; tumeurs (F=0,756). Cette réussite s'explique par le nombre élevé d'exemples en corpus (l'axe signes ou symptômes est le plus représenté) et leur relative régularité (le mot *tumeur* ou les parties anatomiques masculines). A l'opposé, les axes les plus compliqués sont : maladies virales (F=0,319) ; malformations et maladies congénitales, héréditaires et néonatales (F=0,351) ; plaies et blessures (F=0,372) ; et maladies de la peau et du tissu conjonctif (F=0,393). Ces axes sont peu représentés dans les corpus, renvoient à des entités plus complexes à identifier et nécessitent de réelles connaissances médicales (telles que les maladies congénitales).

Les méthodes employées par les participants sont des méthodes de classification multi-labels supervisées, éventuellement complétées par des plongements lexicaux tels que CamemBERT par Bailly *et al.* (2021) et Gérardin *et al.* (2021), ou en réentraînant des plongements avec Word2Vec comme réalisé par Billami *et al.* (2021). Une autre approche, suivie par Dupont *et al.* (2021), repose sur l'utilisation du MeSH, complété de termes du corpus, pour indexer le contenu des documents. Les meilleurs résultats, obtenus par Hiot *et al.* (2021), reposent sur des transducteurs à états finis et des listes de mots-clés issus du MeSH et de MedDRA. Les auteurs soulignent la simplicité de cette méthode et

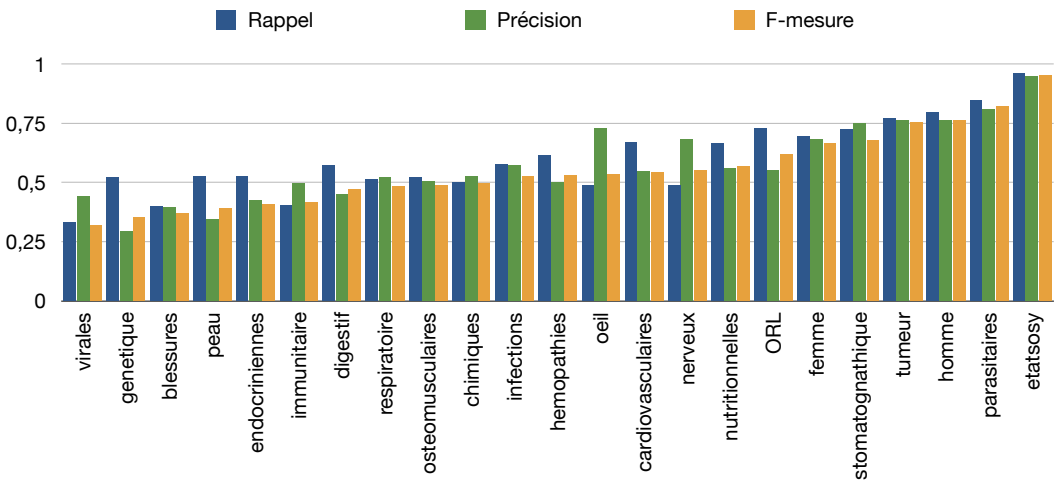

FIGURE 1 – Valeurs moyennes de rappel (bleu), précision (vert), et F-mesure (jaune), pour chaque axe, sur l'ensemble des soumissions des participants, par F-mesure croissante

son coût environnemental faible puisque ne nécessitant aucun entraînement.

Les pré-traitements de tokénisation, lemmatisation, et suppression des mots outils ont été appliqués par l'ensemble des participants pour réduire le nombre de descripteurs. Dupont *et al.* (2021) ont également travaillé sur la détection des phrases négatives et hypothétiques tandis que Hiot *et al.* (2021) ont identifié les négations et les informations de genre.

Tâche 2. Le tableau 8 présente les résultats (précision pour l'ensemble et moyenne par question des corrélations de Pearson) et le classement sur la tâche 2 (RP=rang précision, classement officiel, RC=rang corrélation), par précision décroissante. Trois équipes ont participé, pour une précision moyenne de 0,607 et une médiane de 0,627. Une partie du fichier de la deuxième soumission de l'équipe EDF Lab étant compromise, une évaluation du fichier corrigé a été réalisée hors-délais, sans remettre en cause le classement.

Les meilleurs résultats ont été obtenus avec des algorithmes classiques de classification, tel que Random Forest sous WEKA utilisé par Suignard *et al.* (2021), ou en calculant une similarité entre vecteurs de la question et de la réponse tel que réalisé par Wang *et al.* (2021) au moyen d'un SVM et des informations contextuelles, ou par Dupont *et al.* (2021) qui ont comparé plusieurs coefficients de similarité. Nous observons que les approches utilisant CamemBERT et SentenceBERT (autres soumissions EDF Lab et Nantalco) ont obtenu de moins bons résultats.

Tâche 3. Pour l'évaluation des résultats de la troisième tâche, nous avons également utilisé la corrélation de Pearson (formule 1, avec Cov(X,Y) la covariance des variables X et Y, et σX et σY les écarts-types de ces variables) en complément de la précision.

$$r = \frac{Cov(X,Y)}{\sigma X \sigma Y} \quad (1)$$

Equipe	Run	Précision	RP	Corrélation	RC
EDF Lab (Suignard *et al.*, 2021)	1	0,682	1	0,57 (1 N/A)	1
	*2**	*0,589**	–	*	–
	3	0,638	–	0,57	–
QUEER (Dupont *et al.*, 2021)	1	0,448	–	0,46	–
	2	0,624	–	0,52	2
	3	0,630	3	0,47 (2 N/A)	–
Nantalco (Wang *et al.*, 2021)	1	0,639	2	0,52 (5 N/A)	3
	2	0,580	–	0,38 (2 N/A)	–
	3	0,627	–	0,40 (2 N/A)	–

TABLE 8 – Résultats et classement des équipes participantes à la tâche 2 (RP=rang précision, classement officiel, RC=rang corrélation ; *évaluation hors-délais, N/A : nombre de question où on ne peut calculer la corrélation)

Le tableau 9 présente les résultats (précision pour l'ensemble et moyenne par question des corrélations de Pearson) et le classement sur la tâche 3 (RP=rang précision, classement officiel, RC=rang corrélation), par précision décroissante. Avec trois équipes, la moyenne est de 0,264 et la médiane de 0,241.

Equipe	Run	Précision	RP	Corrélation	RC
EDF Lab (Suignard *et al.*, 2021)	1	0,510	1	0,45	–
	2	0,382	–	0,40	–
	3	0,292	–	0,47	2
QUEER (Dupont *et al.*, 2021)	1	0,278	2	0,00	–
	2	0,241	–	-0,01	–
	3	0,212	–	-0,04	–
Proofreaders (Poulain & Connes, 2021)	1	0,170	3	0,54	1
	2	0,159	–	0,49	–
	3	0,133	–	0,51	–

TABLE 9 – Résultats et classement des équipes participantes à la tâche 3 (RP=rang précision, classement officiel, RC=rang corrélation)

Alors que les résultats des participants à la deuxième tâche sont parfois très proches entre soumissions de deux participants, nous observons que les résultats obtenus sur la troisième tâche conservent les soumissions groupées par participant, témoignant à la fois de la difficulté de la tâche, et des différences méthodologiques, qui ne permettent pas à un participant de dépasser un autre participant.

Sur cette tâche, les meilleurs résultats ont été obtenus avec une simple similarité fondée sur des trigrammes de caractères, utilisée par Suignard *et al.* (2021), qui obtient de bien meilleurs résultats que SentenceBERT (autres soumissions EDF Lab), ou que les réseaux de neurones LSTM utilisés par Dupont *et al.* (2021). Enfin, Poulain & Connes (2021) ont travaillé sur une approche d'extraction de traits lexicaux en combinant les corpus du défi avec un sous-corpus de textes pédagogiques issus de Wikilivres. Nous observons que ces participants ont également comparé les mesures de similarité disponibles pour comparer des vecteurs.

Les questions appellent des réponses soit sous la forme d'un résultat attendu (une réponse qui peut

être discrétisée), soit en langue formelle (du code informatique), soit en langue naturelle. Pour les réponses sous la forme d'un résultat attendu, certains étudiants produisent des réponses en langue naturelle pour justifier leur réponse. Les résultats pour le meilleur run et en prenant pour chaque question le système qui répond le mieux sont donnés dans le tableau 10. On note pour la tâche 2, que les précisions suivent l'ordre attendu de facilité à noter les réponses (Résultat, Formelle, puis Naturelle). Pour la tâche 3, malgré le peu de questions (6), la même tendance est observée (il n'existe aucune réponse de type résultat attendu sur la tâche 3).

Type de réponse	Tâche 2		Tâche 3	
	MRun(=EDF1)	MSPQ	MRun	MSPQ
Naturelle	0,66	0,71	0,16(EDF-2)	0,18
Formelle	0,78	0,81	0,76(EDF-1)	0,76
Résultat	0,80	0,82	—	—

TABLE 10 – Précisions par type de questions (MRun : meilleur run ; MSPQ : Meilleur système par question)

Nous remercions les participants à la tâche 3 qui était assez expérimentale, et pour lequelle nous espérons à long terme l'intégration d'outils d'évaluation dans Moodle, permettant un gain de temps et de qualité lors des évaluations. A l'instar des campagnes SemEval, nous envisageons de limiter le nombre de notes (deux ou trois niveaux). Une autre direction d'évaluation serait d'introduire une évaluation par rubrique (évaluation des réponses des étudiants selon plusieurs dimensions : syntaxe, sémantique, etc.) comme abordée par Mizumoto *et al.* (2019).

6 Conclusion

L'édition 2021 du défi fouille de texte (DEFT) a traité le domaine clinique dans la continuité de DEFT 2019 et DEFT 2020 d'une part, et pour la première fois dans DEFT le domaine des réponses d'étudiants à des questionnaires en ligne de type réponses courtes (dans Moodle) d'autre part.

La première tâche a rassemblé sept équipes et visait à établir le profil clinique des patients décrits dans un corpus de 275 cas cliniques, en normalisant les pathologies, signes ou symptômes par rapport à l'un des 23 axes du chapitre C du MeSH. Les F-mesures obtenues par les participants sur le corpus de test varient de 0,394 à 0,814, avec une moyenne de 0,636 et une médiane de 0,700. Notre baseline (identification de concepts représentatifs de chaque axe) obtient une F-mesure de 0,546. L'utilisation de transducteurs à états finis avec des listes de mots-clés s'est révélée la plus efficace.

La deuxième tâche a rassemblé trois équipes et portait sur l'évaluation automatique de réponses d'étudiants à des questionnaires en ligne, en prenant pour référence la correction de l'enseignant. Les précisions obtenues par les participants varient de 0,448 à 0,682, avec une moyenne de 0,607 et une médiane de 0,627. Notre baseline (nombre de mots en commun entre réponse et question/correction) obtient une précision de 0,477. La classification avec Random Forest a permis l'obtention des meilleurs résultats.

Enfin, la dernière tâche a également rassemblé trois équipes et concernait la poursuite de l'évaluation de réponses d'étudiants, en prenant pour référence les réponses déjà évaluées par l'enseignant (entre 5 et 10 % des réponses à chaque question étaient déjà évaluées). Les précisions obtenues varient de

0,133 à 0,510, avec une moyenne de 0,264 et une médiane de 0,241, et les corrélations de Pearson varient de 0,02 à 0,65. Nos baselines obtiennent une précision de 0,397 (recherche des réponses évaluées les plus similaires) et de 0,561 (k plus proches voisins). Comme pour la précédente tâche, la méthode la plus simple a obtenu les meilleurs résultats, grâce à une similarité de trigrammes de caractères.

Cette nouvelle édition du défi fouille de texte se termine avec une variété de méthodes testées sur chacune des tâches proposées, et un constat que les méthodes les plus simples continuent, pour l'instant, de surpasser les approches à base de plongements lexicaux.

Références

BAILLY A., BLANC C. & GUILLOTIN T. (2021). Classification multi-label de cas cliniques avec CamemBERT. In *Actes de DEFT*, Lille, France.

BASU S., JACOBS C. & VANDERWENDE L. (2013). Powergrading : a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, **1**, 391–402.

BILLAMI M. B., NICOLAIEFF L., GOSSET C. & BORTOLASO C. (2021). Participation de Berger-Levrault (BL.Research) à DEFT 2021 : de l'apprentissage des seuils de validation à la classification multi-labels de documents. In *Actes de DEFT*, Lille, France.

BURROWS S., GUREVYCH I. & STEIN B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, **25**(1), 60–117.

CARDON R., GRABAR N., GROUIN C. & HAMON T. (2020). Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In *Actes DEFT*, p. 1–13, Nancy, France : ATALA. HAL : hal-02784737.

DUPONT Y., GONZÁLEZ-GALLARDO C.-E., LEJEUNE G., MILLOUR A. & TANGUY J.-B. (2021). QUEER@DEFT2021 : Identification du profil clinique de patients et notations automatique de copies d'étudiants. In *Actes de DEFT*, Lille, France.

DZIKOVSKA M. O., NIELSEN R. D., BREW C., LEACOCK C., GIAMPICCOLO D., BENTIVOGLI L., CLARK P., DAGAN I. & DANG H. T. (2013). Semeval-2013 task 7 : The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, p. 263–274, Atlanta, Georgia.

GÉRARDIN C., VAILLANT P., WAJSBÜRT P., GILAVERT C., BELLAMINE A., KEMPF E. & TANNIER X. (2021). Classification multilabel de concepts médicaux pour l'identification du profil clinique du patient. In *Actes de DEFT*, Lille, France.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French corpus with clinical cases. In *Proc of LOUHI*, p. 122–128, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5614.

GRABAR N., GROUIN C., HAMON T. & CLAVEAU V. (2019). Recherche et extraction d'information dans des cas cliniques. présentation de la campagne d'évaluation DEFT 2019. In *Actes DEFT*, p. 1–10, Toulouse, France : ATALA. HAL : hal-02280852.

GROUIN C., GRABAR N., HAMON T. & CLAVEAU V. (2019). Clinical case reports for NLP. In *Proc of BioNLP*, p. 273–282, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-5029.

HIOT N., MINARD A.-L. & BADIN F. (2021). DOING@DEFT : utilisation de lexiques pour une classification efficace de cas cliniques. In *Actes de DEFT*, Lille, France.

HORBACH A., PALMER A. & WOLSKA M. (2014). Finding a tradeoff between accuracy and rater's workload in grading clustered short answers. In *LREC*, p. 588–595 : Citeseer.

MANNION A., CHEVALIER T., SCHWAB D. & GOEURIOT L. (2021). Identification de profil clinique du patient : Une approche de classification de séquences utilisant des modèles de langage français contextualisés. In *Actes de DEFT*, Lille, France.

MIZUMOTO T., OUCHI H., ISOBE Y., REISERT P., NAGATA R., SEKINE S. & INUI K. (2019). Analytic score prediction and justification identification in automated short answer scoring. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 316–325.

MOHLER M. & MIHALCEA R. (2009). Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, p. 567–575.

NLM (2001). *Medical Subject Headings*. National Library of Medicine, Bethesda, Maryland. https://www.nlm.nih.gov/mesh/meshhome.html.

POULAIN T. & CONNES V. (2021). DEFT 2021 : Évaluation automatique de réponses courtes, une approche basée sur la sélection de traits lexicaux et augmentation de données. In *Actes de DEFT*, Lille, France.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). brat : a web-based tool for NLP-Assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 102–107, Avignon, France : Association for Computational Linguistics.

SUIGNARD P., BENAMAR A., MESSOUS N., CHRISTOPHE C., JUBAULT M. & BOTHUA M. (2021). Participation d'EDF R&D à DEFT 2021. In *Actes de DEFT*, Lille, France.

WANG X., LIU X. & YUE Y. (2021). Mesure de similarité textuelle pour une évaluation automatique de copies d'étudiants. In *Actes de DEFT*, Lille, France.