Recherche et extraction d'information dans des cas cliniques. Présentation de la campagne d'évaluation DEFT 2019

Natalia Grabar[1, 2] Cyril Grouin[2] Thierry Hamon[2, 3] Vincent Claveau[4]
(1) STL, CNRS, Université de Lille, Domaine du Pont-de-bois, 59653 Villeneuve-d'Ascq cedex, France
(2) LIMSI, CNRS, Université Paris-Saclay, Campus universitaire d'Orsay, 91405 Orsay cedex, France
(3) Université Paris 13, 99 avenue Jean-Baptiste Clément, 93430 Villetaneuse, France
(4) IRISA, CNRS, Campus universitaire de Beaulieu, 35042 Rennes cedex, France
natalia.grabar@univ-lille.fr, {cyril.grouin,thierry.hamon}@limsi.fr,
vincent.claveau@irisa.fr

RÉSUMÉ
Cet article présente la campagne d'évaluation DEFT 2019 sur l'analyse de textes cliniques rédigés en français. Le corpus se compose de cas cliniques publiés et discutés dans des articles scientifiques, et indexés par des mots-clés. Nous proposons trois tâches indépendantes : l'indexation des cas cliniques et discussions, évaluée prioritairement par la MAP (mean average precision), l'appariement entre cas cliniques et discussions, évalué au moyen d'une précision, et l'extraction d'information parmi quatre catégories (âge, genre, origine de la consultation, issue), évaluée en termes de rappel, précision et F-mesure. Nous présentons les résultats obtenus par les participants sur chaque tâche.

ABSTRACT
Information Retrieval and Information Extraction from Clinical Cases. Presentation of the DEFT 2019 Challenge

This paper presents the DEFT 2019 challenge on the analysis of clinical texts in French. These texts are Clinical Cases, published and discussed within scientific papers, and indexed by keywords. We propose three independent tasks : the indexing of clinical cases and discussions, primarily evaluated using the mean average precision (MAP), the pairing between clinical cases and discussions, evaluated using precision, and the information extraction among four categories (age, gender, origin of consultation, outcome), evaluated in terms of recall, precision and F-measure. We present the results obtained by the participants on each task.

MOTS-CLÉS : Cas clinique, fouille de texte, extraction d'information, recherche d'information, évaluation.

KEYWORDS: Clinical cases, text-mining, information extraction, information retrieval, evaluation.

1 Introduction

L'édition 2019 du défi fouille de textes (DEFT 2019, https://deft.limsi.fr/2019/) porte sur l'analyse de cas cliniques rédigés en français. Cette édition se compose de trois tâches autour de la recherche d'information et de l'extraction d'information. Bien que ces tâches aient déjà fait l'objet de campagnes d'évaluation dans le passé (l'identification de mots-clés dans DEFT 2012 et DEFT 2016, l'appariement entre une recette et ses ingrédients lors de DEFT 2013), c'est la première fois qu'une

campagne d'évaluation porte sur des textes cliniques en français. Les cas décrivent les situations cliniques de patients, réels ou fictifs. Les cas cliniques sont publiés dans plusieurs sources de données (scientifique, didactique, associatif, juridique) sous forme anonymisée. L'objectif consiste à présenter des situations cliniques typiques (cadre didactique) ou bien des situations rares (cadre scientifique).

Déroulement de la campagne Les annonces informant de cette campagne ont été faites entre décembre 2018 et avril 2019 sur plusieurs listes de diffusions du traitement automatique des langues, de l'ingénierie des connaissance, et du domaine biomédical, en français (AIM, ARIA, EGC, Info-IC, LN, MadICS) et en anglais (BioNLP, Corpora). L'accès aux données d'entraînement a été possible dès le 18 février, tandis que la phase de test s'est déroulée du 9 au 15 mai, sur une période de trois jours définie par chacun des participants. Afin de participer, chaque équipe a signé un accord d'utilisation des données fixant les conditions d'accès et de précautions à prendre concernant les données.

Huit équipes se sont inscrites et ont participé jusqu'au bout. Nous comptons cinq équipes académiques (LGI2P/Mines Alès, Nîmes ; LIMICS/INRA, Paris ; LIPN/STIH, Paris ; TALN-LS2N, Nantes ; Université Assane Seck de Ziguinchor, Sénégal), deux équipes industrielles (EDF Lab, Palaiseau ; Qwant, Paris) et une équipe mixte (Synapse/IRIT, Toulouse).

2 Corpus

2.1 Origine des données

Le corpus mis à disposition pour DEFT 2019 fait partie d'un corpus de cas cliniques plus grand, avec des annotations et informations associées plus riches (Grabar *et al.*, 2018). Pour cette édition, nous nous sommes concentrés sur les cas cliniques pour lesquels existent une indexation au moyen de mots-clés et une discussion. Les cas proposés sont liés à différentes spécialités médicales (cardiologie, urologie, oncologie, obstétrique, pneumologie, gastro-entérologie) et concernent plusieurs pays francophones (France, Belgique, Suisse, Canada, pays africains et tropicaux).

2.2 Données de référence

Les données de référence de la compétition sont consensuelles et obtenues à partir de deux annotations effectuées de manière indépendante (Grabar *et al.*, 2019). Le tableau 1 donne les accords inter-annotateurs évalués au moyen de la F-mesure sur les annotations de la tâche d'extraction d'information, d'abord entre les deux annotateurs, puis entre chaque annotateur et le résultat du consensus.

Catégorie	*Annotateur 1/Annotateur 2*	*Annotateur 1/consensus*	*Annotateur 2/consensus*
âge	0,9844	0,9887	0,9944
genre	0,8044	0,9903	0,8143
issue	0,4654	0,6204	0,8152
origine	0,8734	0,8886	0,9755

TABLE 1 – Accords inter-annotateurs (F-mesure) calculés avec BRATeval (comparaison des portions pour *âge* et *origine*, des valeurs normalisées pour *genre* et *issue*)

3 Présentation des tâches

3.1 Tâche 1 : indexation des cas cliniques

Dans cette première tâche, nous fournissons les cas cliniques avec les discussions correspondantes et le nombre de mots-clés attendus pour chaque couple cas clinique/discussion. Nous donnons également la liste de l'ensemble des mots-clés du corpus, classés par ordre alphabétique, tels qu'ils ont été choisis par les auteurs des articles dont sont issus les cas cliniques (voir tableau 2). Un même mot-clé peut servir à indexer plusieurs cas cliniques, il n'apparaîtra cependant qu'une seule fois dans la liste fournie. Puisque la liste des mots-clés fournie porte sur l'intégralité du corpus, certains mots-clés ne sont utilisés que dans le corpus d'entraînement (par exemple *agénésie*), d'autres uniquement dans le corpus de test (tel que *agénésie déférentielle*), d'autres encore ne sont pas utilisés dans le corpus de la tâche 1 mais servent à indexer les documents utilisés dans le corpus de la tâche 2 (ces documents pourront servir pour une nouvelle tâche d'indexation lors d'une prochaine édition).

Mot-clé	*Cas clinique et discussion indexés*	*Sous-corpus*
agénésie	1136550700.txt 2300836250.txt	Entraînement
agénésie déférentielle	1139700160.txt 2354143280.txt	Test
agénésie rénale agénésie rénale unilatérale	Inutilisés dans la tâche d'indexation en 2019	

TABLE 2 – Extrait de la liste des mots-clés du corpus avec indexation de cas cliniques et discussions

L'objectif de cette tâche est d'identifier, parmi la liste des mots-clés du corpus, les mots-clés servant à indexer chaque couple cas clinique/discussion. Les participants ont la possibilité de fournir davantage de mots-clés que le nombre attendu, en les classant par ordre de pertinence décroissant. La principale mesure d'évaluation de la tâche est la Mean Average Precision (MAP), la mesure secondaire étant une R-précision, c'est-à-dire la précision au rang N, Prec@N, avec N le nombre de mots-clés attendus.

3.2 Tâche 2 : similarité sémantique entre cas cliniques et discussions

Dans cette deuxième tâche, nous fournissons un ensemble de cas cliniques, et un ensemble de discussions qui correspondent aux cas cliniques du premier ensemble. Parce que les articles scientifiques intègrent parfois plusieurs descriptions de cas cliniques, il est possible qu'une même discussion porte sur plusieurs cas cliniques. Dans ce cas, une même discussion correspond à différents fichiers. L'objectif de cette tâche consiste à apparier les cas cliniques avec les discussions. L'évaluation des résultats est de type booléen agrégé sous la forme d'une précision et d'un rappel classique (ces deux mesures sont égales si le système renvoie une réponse pour chaque cas clinique). Lors de l'évaluation, les fichiers de discussion sont dédoublonés : il suffit qu'un des fichiers de la liste de discussions doublons soit trouvé.

3.3 Tâche 3 : extraction d'information

Cette dernière tâche s'intéresse aux informations démographiques et cliniques générales présentes dans le corpus. Nous nous intéressons à quatre types d'information.

L'âge de la personne dont le cas est décrit, au moment du dernier élément clinique rapporté, normalisé sous la forme d'un entier ("0" pour un nourrisson de moins d'un an, "1" pour un enfant de moins de deux ans, y compris un an et demi, "20" pour un patient d'une vingtaine d'années, etc.).

Le genre de la personne dont le cas est décrit, parmi deux valeurs normalisées : *féminin, masculin* (il n'existe aucun cas de dysgénésie ou d'hermaphrodisme dans le corpus).

L'origine ou motif de la consultation ou de l'hospitalisation, pour le dernier événement clinique ayant motivé la consultation. Cette catégorie intègre généralement les pathologies, signes et symptômes (*"une tuméfaction lombaire droite, fébrile avec frissons"* ou *"un contexte d'asthénie et d'altération de l'état général"*), plus rarement les circonstances d'un accident (*"une chute de 12 mètres, par défénestration, avec réception ventrale"*, *"un AVP moto"* ou *"pense avoir été violée"*). Le suivi clinique se trouve dans la continuité d'événements précédents. Il ne constitue pas un motif de consultation.

L'issue parmi cinq valeurs possibles :

- *guérison*, le problème clinique décrit dans le cas a été traité et la personne est guérie : *"Le recul était de deux ans sans récidive locale ni incident notable"*, *"Les fuites urinaires ont disparu dans les suites opératoires"*
- *amélioration*, l'état clinique est amélioré sans qu'on ne puisse conclure à une guérison : *"évolution favorable de l'état de la patiente"*, *"Les suites ont été simples"*
- *stable*, soit l'état clinique reste stationnaire, soit il est impossible de choisir entre amélioration et détérioration : *"Les images ne se sont pas modifiées à 20 mois de recul"*, *"la patiente présente toujours une constipation opiniâtre terminale, équilibrée sous traitement médical. Sur le plan sexuel, aucune amélioration notable n'a été notée dans les suites de la neuromodulation"*
- *détérioration*, l'état clinique se dégrade : *"Un mois plus tard, le patient a été hospitalisé pour toxoplasmose cérébrale et pneumocytose pulmonaire, actuellement en cours de traitement"*, *"Une EER de contrôle à 3 ans a été réalisée et montrait la persistance de cette masse kystique mais avec des parois et des végétations endoluminales plus épaisses, denses et homogènes"*
- *décès*, lorsque le décès concerne directement le cas clinique décrit : *"Le patient est décédé au 6ème mois après l'intervention"*, *"Elle est décédée quinze ans après la première intervention par récidive tumorale importante et envahissement des viscères adjacents"*

Dans le cas de documents se rapportant à plusieurs patients, les âges et genres de chacun des patients devront être identifiés (par exemple, dans le cas d'un greffon issu d'un même donneur qui aura été greffé à deux patients successifs, l'âge et le genre des deux personnes greffées devront être identifiés). Il n'est pas nécessaire de relier l'âge avec le genre. Pour le cas où seraient mentionnés plusieurs âges se rapportant à une même personne (l'âge actuel et un âge dans les antécédents), seul l'âge au moment du cas clinique décrit doit être rapporté. Quelques rares documents ne permettent cependant pas d'instancier l'ensemble des quatre catégories. Dans cette situation, la valeur est NUL.

La figure 1 présente un extrait de cas clinique dont les portions annotées renvoient à ces quatre catégories. Sur la base de ces annotations, nous construisons la référence en normalisant la valeur de trois catégories ("masculin" pour la catégorie *genre*, "60" pour *âge*, et "décès" pour *issue*) ou en conservant la portion annotée pour la catégorie *origine*.

Les valeurs d'âge, genre et issue, sont évaluées de manière stricte (même valeur entre hypothèse et référence). Il n'est pas demandé de rapporter la portion textuelle ayant permis de fournir ces valeurs. L'origine de la consultation est évaluée en tenant compte du taux de recouvrement de la portion textuelle fournie par rapport à la portion textuelle de référence.

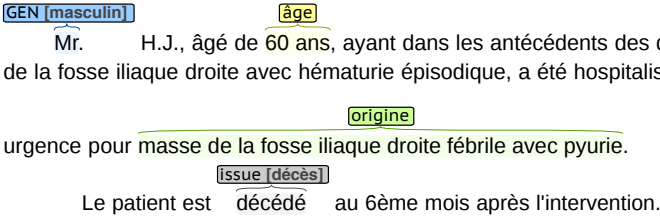

FIGURE 1 – Extrait d'un cas clinique annoté en âge, genre, origine et issue, avec valeurs normalisées

4 Résultats

Nous présentons dans cette section les résultats des soumissions (runs) des équipes participantes. Pour chaque tâche, nous décrivons les mesures d'évaluation employées et proposons une étude de la significativité statistique des différences constatées. Pour chacune des tâches, nous proposons des systèmes *baseline* avec la philosophie suivante : ces systèmes ne doivent pas recourir à des données externes mais s'appuyer sur des méthodes simples ou éprouvées du domaine. Les résultats obtenus permettent d'évaluer la difficulté de la tâche et de mettre en valeur les gains obtenus par les participants.

4.1 Tâche 1 : indexation des cas cliniques

Participants Le tableau 3 présente les résultats obtenus par les participants pour chacune des soumissions (runs) de la tâche d'indexation, classés par ordre alphabétique des noms d'équipe, évalués en termes de MAP et de R-précision (R-Prec). Les meilleurs résultats obtenus par chaque équipe sur la mesure principale (MAP) sont en gras.

Équipe	EDF Lab			LGI2P			LIPN		
Soumission	1	2	3	1	2	3	1	2	3
MAP	**0,362**	0,273	—	0,401	0,397	**0,478**	0,126	**0,220**	**0,220**
R-Prec	0,324	0,236	—	0,459	0,451	0,451	0,122	0,240	0,240

Équipe	LS2N			Synapse			UASZ		
Soumission	1	2	3	1	2	3	1	2	3
MAP	**0,405**	0,232	0,404	0,365	**0,446**	0,365	0,276	**0,396**	0,317
R-Prec	0,467	0,283	0,460	0,439	0,439	0,439	0,343	0,455	0,378

TABLE 3 – Résultats (MAP et R-Prec) sur la tâche 1. Les meilleurs résultats par équipe sont en gras

Baseline Nous avons produit deux systèmes *baseline* s'appuyant sur des principes issus de la Recherche d'Information pour pondérer les termes-clés candidats (voir Grabar *et al.* (2019) pour une description complète). Sur le test, la première baseline obtient une MAP de 0,177 et une R-Précision de 0,236 ; la deuxième baseline obtient une MAP de 0,434 et une R-Précision de 0,428.

Significativité statistique Pour mesurer la pertinence des écarts constatés entre les meilleurs runs des participants, nous avons calculé leur significativité en utilisant un t-test pairé sur les MAP avec une p-valeur fixée à 0,05. Ainsi, sous ces conditions, le run 3 de LGI2P est jugé significativement meilleur que le run 2 de Synapse (p=0,0428). Ce dernier n'est pas significativement meilleur que la baseline 2 mais est jugé significativement meilleur que le run 1 de LS2N. Les différences constatées entre les runs 1 et 2 de LGI2P, run 2 de UASZ, runs 1 et 3 de LS2N ne sont pas jugées significatives.

4.2 Tâche 2 : similarité sémantique entre cas cliniques et discussions

Participants Le tableau 4 présente les résultats obtenus par les participants sur la tâche d'appariement, classés par ordre alphabétique des noms d'équipe, évalués en termes de précision.

Equipe	EDF Lab			LGI2P			LIPN		
Soumission	1	2	3	1	2	3	1	2	3
Précision	0,888	**0,953**	0,935	**0,907**	**0,907**	0,902	**0,617**	0,107	0,126
Equipe	Qwant			Synapse			UASZ		
Soumission	1	2	3	1	2	3	1	2	3
Précision	**0,841**	0,762	0,832	**0,617**	0,561	0,631	0,874	**0,883**	0,832

TABLE 4 – Résultats (précision) sur la tâche 2. Les meilleurs résultats par équipe sont en gras

Baseline Nous avons produit un système *baseline* s'appuyant là encore sur des principes issus de la Recherche d'Information pour calculer la similarité entre cas et discussion (voir Grabar *et al.* (2019) pour une description complète). Sur le corpus de test, cette baseline obtient une Précision de 0,953.

Significativité statistique Comme précédemment, nous reportons la significativité statistique des écarts constatés entre les meilleurs runs, au sens du t-test pairé (p=0,05). Les différences entre la baseline et les runs 2 et 3 d'EDF Lab ne sont pas jugée significatives. En revanche, la différence entre la baseline et les runs 1 et 2 de LGI2P est significative. Les différences entre les paires de runs suivants ne sont pas non plus jugées significatives : EDF Lab (run 3) vs. LGI2P (run 1), LGI2P (run 1) vs. LGI2P (run 2), LGI2P (run 2) vs. LGI2P (run 3), LGI2P (run 3) vs. EDF Lab (run 1).

4.3 Tâche 3 : extraction d'information

Participants Le tableau 5 présente les résultats obtenus par les participants, ainsi que les deux systèmes de baseline (par règles et par apprentissage, noté ML), évalués en termes de précision, rappel et F-mesure sur les catégories *Age*, *Genre* et *Issue*, et une évaluation au moyen des macro et micro mesures, ainsi que taux de bonne prédiction de mots de la référence (*Accuracy*) sur la catégorie *Origine*. Les meilleurs résultats sont présentés en gras. Lorsqu'une équipe a soumis plusieurs fichiers de résultats, nous observons que les variations de résultats portent uniquement sur la catégorie *Issue*.

Baseline Nous avons produit deux systèmes de *baseline* (voir Grabar *et al.* (2019) pour une description complète). La première baseline repose sur un ensemble limité de règles propres à chaque

Équipe		EDF Lab			LAI		Qwant	Baselines	
Soumission		1	2	3	1	2	1	Règles	ML
Age	P	0,939	0,939	0,939	0,980	0,980	0,975	0,813	0,961
	R	0,467	0,467	0,467	0,919	0,919	0,902	0,807	0,912
	F	0,624	0,624	0,624	**0,948**	**0,948**	0,937	0,810	0,936
Genre	P	0,967	0,967	0,967	0,981	0,981	0,942	0,934	0,960
	R	0,472	0,472	0,472	0,974	0,974	0,947	0,928	0,954
	F	0,634	0,634	0,634	**0,978**	**0,978**	0,944	0,931	0,957
Issue	P	0,329	0,362	0,352	0,486	0,498	0,520	0,502	0,532
	R	0,164	0,180	0,176	0,405	0,492	0,492	0,485	0,525
	F	0,219	0,241	0,234	0,442	0,495	**0,505**	0,493	0,528
Origine (macro)	P	0,534	0,534	0,534	0,582	0,582	0,785	0,465	0,514
	R	0,323	0,323	0,323	0,722	0,722	0,579	0,009	0,565
	F	0,403	0,403	0,403	0,645	0,645	**0,666**	0,018	0,538
Origine (micro)	P	0,302	0,302	0,302	0,628	0,628	0,658	0,037	0,771
	R	0,349	0,349	0,349	0,735	0,735	0,640	0,020	0,556
	F	0,324	0,324	0,324	**0,677**	**0,677**	0,649	0,026	0,646
	Acc	0,278	0,278	0,278	0,600	0,600	0,589	0,017	0,497
Macro-F globale		0,470	0,475	0,474	0,753	**0,766**	0,763	0,563	0,739

TABLE 5 – Résultats (P=précision, R=rappel, F=F-mesure, Acc=Accuracy i.e. taux de mots de la référence bien prédits) par catégorie sur la tâche 3. Les meilleurs résultats pour chacune des sous-tâches sont en gras. La dernière ligne du tableau indique la moyenne des F-mesures (macro) sur l'ensemble des catégories

catégorie. Ce système a des performances élevées pour le genre (F=0,931) et l'âge (F=0,810). En revanche les performances des deux autres catégories sont plus basses : issue (F=493) et origine (F_{micro}=0,026, F_{macro}=0,018). La seconde baseline exploite des approches par apprentissage artificiel (catégorisation supervisée par régression logistique pour le genre et l'issue, et comme problème d'étiquetage par CRF pour l'âge et l'admission). Ce dernier système montre des performances élevées pour l'âge (F=0,936) et le genre (F=0,957). Les deux autres catégories ont des performances un peu plus modestes mais qui restent élevées : issue (F=0,528) et origine (F_{micro}=0,646, F_{macro}=0,538).

Significativité statistique Nous étudions comme précédemment si les différences constatées sont statistiquement significatives. Pour cette tâche, nous avons subdivisé aléatoirement le jeu de test en 20 portions sur lesquels nous avons évalué les performances des algorithmes. Pour l'âge, le genre et l'issue, nous avons pris en compte la F-mesure tandis que pour l'origine, nous avons considéré l'overlap comme mesures principales. Nous disposons donc de 20 F-mesures pour l'âge sur chacun des runs, et ainsi de suite. Sur la base de ces mesures, nous effectuons les t-tests pairés (p=0,05). Nous ne reportons les différences qu'entre les meilleurs runs de chaque équipe. Les différences entre LAI (run 2) et EDF Lab (run 2) sont statistiquement significatives pour chacune des mesures. C'est également le cas entre Qwant et EDF Lab. En revanche, seule la différence sur le genre est significative entre Qwant et LAI (run 2). Les autres différences observées ne sont donc pas significatives.

5 Méthodes des participants

De manière générale, nous observons que la majorité des participants a appliqué des étapes de pré-traitements classiques (homogénéisation de la casse, tokénisation, lemmatisation, racinisation, étiquetage en parties du discours, suppression des mots outils, etc.), quelle que soit la tâche considérée. Certains participants (Maudet *et al.*, 2019) ont également fait le choix de supprimer l'ambiguïté des acronymes en les remplaçant par une forme expansée.

En fonction des approches utilisées, Maudet *et al.* (2019); Sileo *et al.* (2019) ont parfois eu recours à des données externes pour compléter les corpus fournis. Ces données se composent de pages Wikipédia du domaine médical, des fiches médicaments de l'EMEA (agence européenne du médicament), ou des résumés d'articles scientifiques Cochrane.

Les approches à base de réseaux de neurones ont assez peu été employées dans cette campagne. Le peu de données annotées, correspondant à un cadre réel dans lequel ces données sont rares et coûteuses, peut expliquer en partie ce choix de s'appuyer sur des approches non neuronales. Notons cependant que le LIPN (Buscaldi *et al.*, 2019) et Synapse (Sileo *et al.*, 2019) dans la tâche d'appariements et Qwant (Maudet *et al.*, 2019) dans la tâche d'extraction d'informations se sont appuyés sur des architectures classiques, avec pour Qwant, des résultats intéressants.

5.1 Tâche 1 : indexation des cas cliniques

Après l'application de pré-traitements classiques sur les cas cliniques, la majorité des participants a opté pour la vectorisation de documents au moyen d'une représentation fondée sur le TF*IDF (Bouhandi *et al.*, 2019; Dramé *et al.*, 2019; Mensonides *et al.*, 2019). Des plongements lexicaux *(word embeddings)* ont été utilisés par Suignard *et al.* (2019), parfois en comparant les résultats d'algorithmes (Sileo *et al.*, 2019). L'information mutuelle a été employée par Buscaldi *et al.* (2019).

Les principales différences portent sur les classifieurs utilisés pour calculer la similarité entre les mots-clés et les cas cliniques. Certains ont choisi des approches différentes selon que le mot-clé est syntaxiquement simple ou complexe (Suignard *et al.*, 2019). Parmi les approches employées, nous relevons le classifieur Naïve Bayes par Dramé *et al.* (2019) ou un gradient boosting par Bouhandi *et al.* (2019). Un angle de vue original abordé par Bouhandi *et al.* (2019) a consisté, non pas à identifier les termes qui sont potentiellement des mots-clés, mais plutôt les termes qui ne sont pas des mots-clés.

5.2 Tâche 2 : similarité sémantique entre cas cliniques et discussions

Sileo *et al.* (2019) ont utilisé des approches à base de réseaux de neurones convolutionnels et une activation ReLu. Les autres équipes ont travaillé sur des représentations vectorielles, notamment fondées sur word2vec pour Suignard *et al.* (2019) ou provenant d'un choix après comparaison de plusieurs modèles pour Dramé *et al.* (2019). L'étude des espaces sémantiques et d'algorithmes d'indexation sémantique latente ont été employés par Dramé *et al.* (2019) et Mensonides *et al.* (2019), ainsi que des modèles de langue par Maudet *et al.* (2019). L'algorithme hongrois, pour optimiser l'attribution discussion-cas, a été utilisé par Buscaldi *et al.* (2019); Suignard *et al.* (2019) et dans notre baseline. Le coefficient de Dice et le score de perplexité sont respectivement utilisés par Mensonides *et al.* (2019) et Maudet *et al.* (2019) pour calculer la similarité.

5.3 Tâche 3 : extraction d'information

Les approches utilisées varient en fonction de la catégorie traitée. L'identification du genre se fonde généralement sur l'utilisation de lexiques de termes spécifiques aux genres féminin ou masculin tandis que des règles ont été développées pour l'âge (Hilbey *et al.*, 2019; Suignard *et al.*, 2019) et l'origine (Hilbey *et al.*, 2019). L'identification de l'issue a été envisagée comme une tâche de classification multi-classes (Maudet *et al.*, 2019), ou fondée sur une représentation vectorielle pour un clustering (Suignard *et al.*, 2019) ou sur une analyse des fréquences de n-grammes (Hilbey *et al.*, 2019).

Un second type d'approche utilisée par Maudet *et al.* (2019) et dans notre baseline envisage la tâche comme un problème d'étiquetage. Notre baseline utilise un CRF standard, tandis que Maudet *et al.* (2019) ont opté pour un réseaux de neurones avec un enchaînement d'une couche convolutive (CNN), d'un réseau de neurones récurrent (Bi-LSTM), d'un CRF, avec des activations ReLu.

6 Conclusion

La compétition DEFT 2019 a proposé aux participants de travailler sur des données médicales originales et récentes, en français, proches des données cliniques grâce au corpus de cas cliniques constitué et annoté manuellement. Les cas cliniques proviennent de publications scientifiques librement disponibles et accessibles. Les tâches de la compétition sont inspirées par les données qui accompagnent les cas cliniques dans les publications sources. Il s'agit d'une part de mots-clés et d'autres part de discussions. Cela a permis de fournir les données pour les tâches 1 et 2. Par ailleurs, des annotations manuelles ont été effectuées par deux annotateurs et ont ensuite été soumises à un consensus. Ces annotations portent sur l'âge et le genre du patient, la raison de la consultation et l'issue de la consultation. Cette annotation a permis de fournir les données pour la tâche 3.

Huit équipes ont soumis des résultats, représentatifs de différents types de méthodes, en fonction des tâches et des catégories : à base de règles, par apprentissage, ou issues de la recherche d'information. Les méthodes des participants présentent des performances assez homogènes pour la plupart des tâches. Les résultats sont légèrement supérieurs aux baselines simples que nous proposons. Cela tend à montrer la difficulté des tâches, notamment du fait du peu de données annotées, qui rend l'usage de techniques neuronales plus difficile.

Ce corpus de cas cliniques, utilisé pour la première fois dans ce contexte, peut fournir des données de référence pour d'autres campagnes d'évaluation. Nous espérons que sa disponibilité encouragera les travaux de TAL sur les données médicales de type clinique, notamment pour le français.

Remerciements

Ce travail a bénéficié d'une aide de l'État attribuée au labex COMIN LABS et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01. Ce travail s'inscrit également dans le projet *CLEAR (Communication, Literacy, Education, Accessibility, Readability)* financé par l'ANR sous la référence ANR-17-CE19-0016-01. Nous remercions également l'ensemble des participants pour l'intérêt porté à cette nouvelle édition du défi fouille de texte (DEFT) et pour la diversité des méthodes employées.

Références

BOUHANDI M., BOUDIN F., GALLINA Y. & HAZEM A. (2019). DeFT 2019 : Auto-encodeurs, gradient boosting et combinaisons de modèles pour l'identification automatique de mots clés. participation de l'équipe TALN du LS2N. In *Actes de DEFT*, Toulouse, France.

BUSCALDI D., GHOUL D., LE ROUX J. & LEJEUNE G. (2019). Indexation et appariements de documents cliniques pour le deft 2019. In *Actes de DEFT*, Toulouse, France.

DRAMÉ K., DIOP I., FATY L. & NDOYE B. (2019). Indexation et appariement de documents cliniques avec le modèle vectoriel. In *Actes de DEFT*, Toulouse, France.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). Cas : French corpus with clinical cases. In *Proc of LOUHI*, p. 122–128, Brussels, Belgium.

GRABAR N., GROUIN C., HAMON T. & CLAVEAU V. (2019). Corpus annoté de cas cliniques en français. In *Actes de TALN*, Toulouse, France.

HILBEY J., DELÉGER L. & TANNIER X. (2019). Participation de l'équipe LAI à DEFT 2019. In *Actes de DEFT*, Toulouse, France.

MAUDET E., CATTAN O., DE SEYSSEL M. & SERVAN C. (2019). Qwant Research @DEFT 2019 : appariement de documents et extraction d'informations à partir de cas cliniques. In *Actes de DEFT*, Toulouse, France.

MENSONIDES J.-C., JEAN P.-A., TCHECHMEDJIEV A. & HARISPE S. (2019). DÉfi fouille de textes 2019 : indexation par extraction et appariement textuel. In *Actes de DEFT*, Toulouse, France.

SILEO D., VAN DE CRUYS T., MUELLER P. & PRADEL C. (2019). Apprentissage non-supervisé pour l'appariement et l'étiquetage de cas cliniques en français - DEFT2019. In *Actes de DEFT*, Toulouse, France.

SUIGNARD P., BOTHUA M. & BENAMAR A. (2019). Participation d'EDF R&D à DEFT 2019 : des vecteurs et des règles. In *Actes de DEFT*, Toulouse, France.