AfIA

Association française
pour l'Intelligence Artificielle

DEFT

Défi Fouille de Textes
(atelier TALN-RECITAL)

PFIA 2019

Table des matières

Éditorial

La reproductibilité des résultats et la robustesse des outils constituent des enjeux critiques en traitement automatique des langues, en particulier parce que le TAL commence à fournir des méthodes et outils mûrs qui sont de plus en plus utilisés dans d'autres domaines, comme le domaine médical. L'objectif des compétitions en TAL est de fournir des corpus et les données de référence qui permettent aux chercheurs de développer des outils et de les tester ensuite. Un tel contexte permet également d'avoir une première comparaison entre les méthodes et approches utilisées par les participants du défi, dans des conditions expérimentales parfaitement identiques.

L'édition 2019 du défi fouille de textes (DEFT 2019, `https://deft.limsi.fr/2019/`) a porté sur l'analyse de cas cliniques rédigés en français. Trois tâches ont été proposées autour de la recherche d'information et de l'extraction d'information, en s'inspirant de tâches réelles et utiles pour le domaine médical. La particularité de cette édition concerne ainsi le domaine traité (médical) et les documents utilisés (des cas cliniques). C'est la première fois qu'une compétition a lieu sur des textes cliniques en français. Les cas cliniques décrivent les situations cliniques de patients, réels ou fictifs. Ils sont publiés dans différentes sources de données (scientifique, didactique, associatif, juridique, etc.), de manière anonymisée. L'utilité des cas consiste à présenter des situations cliniques typiques ou rares, notamment à des fins pédagogiques. Le corpus de cas cliniques utilisé lors de la campagne DEFT 2019 se compose de cas librement accessibles en ligne.

Lors du déroulement de la campagne, 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. 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).

Ces actes rassemblent la présentation des objectifs de la campagne (corpus, tâches, évaluation...), les résultats obtenus sur les différentes tâches et la description des systèmes participants.

Les organisateurs remercient le comité de programme pour avoir apporté leur soutien et leur expertise à la campagne d'évaluation DEFT 2019.

Natalia Grabar, Cyril Grouin, Thierry Hamon et Vincent Claveau

Comités

Organisateurs de DEFT

— Vincent CLAVEAU (IRISA, CNRS)
— Natalia GRABAR (STL, CNRS, Université de Lille)
— Cyril GROUIN (LIMSI, CNRS, Université Paris-Saclay)
— Thierry HAMON (LIMSI, CNRS, Université Paris-Saclay ; Université Paris XIII)

Comité de programme de DEFT

— Patrice BELLOT (LSIS, Aix-Marseille Université)
— Leonardo CAMPILLOS LLANOS (LIMSI, CNRS, Université Paris-Saclay ; Madrid)
— Vincent CLAVEAU (IRISA, CNRS)
— Natalia GRABAR (STL, CNRS, Université de Lille)
— Cyril GROUIN (LIMSI, CNRS, Université Paris-Saclay)
— Vincent GUIGUE (LIP6, Sorbonne Université)
— Thierry HAMON (LIMSI, CNRS, Université Paris-Saclay ; Université Paris XIII)
— Véronique MORICEAU (LIMSI, Université Paris-Sud, Université Paris-Saclay ; IRIT)
— Fleur MOUGIN (Bordeaux Population Health, Université de Bordeaux)
— Mathieu ROCHE (TETIS, CIRAD)
— Patrick RUCH (HEG Geneva, BiTeM)
— Frantz THIESSARD (Bordeaux Population Health, Université de Bordeaux, Inserm ; CHU de Bordeaux, SIM pôle santé publique, unité médicale Informatique et archivistique médicales)

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 agénésie déférentielle	1136550700.txt 2300836250.txt 1139700160.txt 2354143280.txt	Entraînement 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.

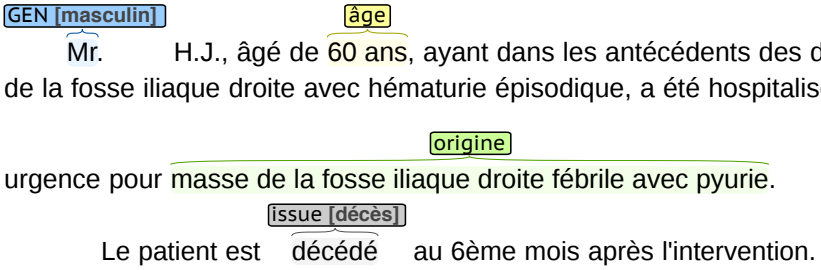


FIGURE 1 – Extrait d'un cas clinique annoté en âge, genre, origine et issue, avec valeurs normalisées

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.

Participation d'EDF R&D à DEFT 2019 : des vecteurs et des règles !

Philippe SUIGNARD[1] Meryl BOTHUA[1] Alexandra BENAMAR[1]
(1) EDF R&D, 7 Boulevard Gaspard Monge, 91120 Palaiseau
prenom.nom@edf.fr

RÉSUMÉ

Ce papier décrit la participation d'EDF R&D à la campagne d'évaluation DEFT 2019. Notre équipe a participé aux trois tâchés proposées : Indexation de cas cliniques (Tâche T1) ; Détection de similarité entre des cas cliniques et des discussions (Tâche T2) ; Extraction d'information dans des cas cliniques (Tâche 3). Nous avons utilisé des méthodes symboliques et/ou numériques en fonction de ces tâches. Aucune donnée supplémentaire, autre que les données d'apprentissage, n'a été utilisée. Notre équipe obtient des résultats satisfaisants sur l'ensemble des taches et se classe première sur la tache 2. Les méthodes proposées sont facilement transposables à d'autres tâches d'indexation et de détection de similarité qui peuvent intéresser plusieurs entités du groupe EDF.

ABSTRACT

EDF R&D submission to DEFT 2019

This paper describes the participation of EDF R&D at DEFT 2019 evaluation campaign. Our team participated in the tree proposed tasks : Clinical Cases Indexation (Task T1) ; Semantic Similarity Detection between Cases and Discussions (Task T2) ; Information Extraction in Clinical Cases (Task T3). We used both symbolic and numerical methods. No additional data other than the training data was used. Our team achieves satisfactory results on all tasks and ranks first on task 2. The proposed methods are easily transferable to other Indexation or Semantic Similarity Detection tasks and may interest several entities of the EDF group.

MOTS-CLÉS : données cliniques, indexation, détection de similarité sémantique, Word2Vec, détection de multimots, extraction d'information, clustering.

KEYWORDS: clinical data, indexation, semantic similarity detection, information extraction, Word2Vec, multiwords detection, clustering.

1 Introduction

Plusieurs éléments nous ont motivés à participer à l'édition 2019 du défi DEFT (Grabar *et al.*, 2019) :

— S'évaluer sur des données différentes comme les données médicales (Grabar *et al.*, 2018)
— Participer à DEFT était l'occasion de travailler sur plusieurs méthodes de détection de similarité dont les résultats contribueront directement à EDF Commerce et à d'autres entités du groupe EDF.

2 Tâche 1 : indexation, calcul de mots-clés

La tâche 1 est une tâche d'indexation, qui consiste à trouver les mots-clés associés à un document, ici le couple cas/discussion dans le domaine médical. La liste des mots clés est fournie, elle est composée de mots simples, de mots composés et de multi mots.

2.1 Méthode 1 : *embeddings* + similarités

La méthode proposée est basée sur les embeddings de mots ou vecteurs-mots. Après avoir entrainé un modèle Word2Vec sur le corpus d'apprentissage, la méthode consiste à calculer une représentation vectorielle des mots-clés potentiels ainsi que des documents puis à retenir les N mots-clés les plus similaires au document. Liste des étapes :

1. **Calcul des embeddings sur le corpus** : La méthode commence par entraîner un modèle Word2Vec ((Mikolov *et al.*, 2013)) sur les données. Les différents paramètres de W2V ont été choisis de manière à optimiser les résultats de la tâche 2.

2. **Calcul des embeddings pour les mots clés** : Quand le mot-clé est constitué d'un seul mot comme « vessie » ou de plusieurs mots, mais dont un seul fait partie du modèle W2V comme « fonction cognitive », l'embedding du mot-clé sera égal à l'embedding du mot. Quand le mot-clé est constitué de plusieurs mots, l'embedding du mot-clé est égal à la moyenne des embeddings de chaque mot. Deux variantes ont été considérées, soit en pondérant les mots clés par l'inverse de la fréquence (IDF) (Sparck Jones, 1972), soit sans pondération.

3. **Calcul des embeddings pour les documents** : Pour calculer l'embedding d'un document, on commence par supprimer les mots appartenant à une « stop liste » constitué d'environ 1000 mots. Puis on moyenne les différents vecteurs-mots en les pondérant par l'inverse de la fréquence (IDF).

4. **Constitution d'une liste de mots-clés potentiels** : Sur le corpus d'apprentissage, on observe que certains documents ont pour mots-clés, des mots qui ne sont pas présents dans le document. Mais comme ces situations sont loin d'être majoritaires, pour éliminer le risque de bruit, on constitue une liste de mots-clés potentiels en gardant uniquement les mots clés constitués des mots présents dans le document lui-même.

5. **Similarité entre les mots-clés potentiels et les documents** : Pour chaque mot-clé potentiel, on calcule sa similarité avec le document. Puis on conserve les N mots clés ayant la similarité la plus élevée avec ce document. En comparant les mots-clés ainsi obtenus sur le corpus d'apprentissage à ceux attendus, on s'aperçoit d'une sur-représentation des mot-clés composés de 2 mots, 3 mots, etc. par rapport aux mots clés composés d'un seul mot. On va donc pondérer le calcul de similarité en fonction du nombre de mots du mot-clé :

$$sim = coef * \cos(Vec(Mot - cle), Vec(Document)) \quad (1)$$

avec $coef = 1$ si $n = 1$, $coef = 0,9$ si $n = 2$, $coef = 0,8$ si $n = 3$ et $coef = 0,7$ si $n >= 4$, et n le nombre de mots du mot-clé.

2.2 Méthode 2 : détection de multi-mots et exploitation des étiquettes morpho-syntaxiques

Les mots-clés à retrouver dans ce corpus sont des mots uniques (mots simples, par exemple « tumeur » et composés, par exemple « Wolff-Parkinson ») et des expressions multi-mots (comme, par exemple, « tumeur rénale ») . Après une analyse fréquentielle sur les mots-clés à retrouver dans le corpus, nous remarquons qu'il y a à peu près autant de mots simples que d'expressions multi-mots (Cf. Table 1).

	mots simples	multi-mots	total
sans doublons	358	329	687
total	684	445	1.129

TABLE 1 – Occurrence des mots-clés à retrouver dans le corpus d'entraînement

2.2.1 Détection des mots simples

La fréquence des mots simples dans le corpus est un indicateur pouvant être déterminant dans le filtrage des mots obtenus après indexation (Cf. Table 2). Les mots-clés obtenus (simples et composés) apparaissent légèrement plus dans la discussion que dans le cas associé, mais cela ne semble pas significatif. Il y a 46 mots-clés qui n'apparaissent pas dans le corpus, ce qui complexifie la tâche d'indexation : nous ne pourrons donc pas nous baser uniquement sur le contenu du texte. Des pré-traitements ont été réalisés pour le comptage des mots-clés : minusculisation et suppression des accents.

Mots simples			Total
Occurence cas	Occurence discussion	Déduction annotateur	-
247	277	46	358

TABLE 2 – Occurrence des mots à trouver par type de mots (simples ou multi-mots) dans le corpus

Afin de comprendre au mieux le choix d'attribution des mots-clés que nous devons retrouver dans le corpus, nous avons utilisé TreeTagger [1], qui permet d'étiqueter les mots sur le plan morpho-syntaxique (Cf. Table 3). Nous avons obtenu les résultats suivants : les mots-clés simples sont principalement des noms (« NOM ») et des adjectifs (« ADJ »). Cette information nous permettra d'effectuer un filtre sur notre corpus, et de supprimer 248.161 mots, soit 65% de nos données textuelles.

	Catégorie morpho-syntaxique			Total
	NOM	ADJ	Autres	-
mots-clés	293	19	46	358
corpus	91.152	45.412	248.161	384.725

TABLE 3 – Catégories morpho-syntaxique des mots simples à retrouver dans le corpus (obtenues avec TreeTagger)

1. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Afin de réduire le bruit, nous avons ensuite calculé une pondération des mots obtenus en comparant la fréquence d'un mot-clé w_i dans un document d_j, à sa fréquence dans tout l'ensemble des documents d. La fonction de pondération f est définie par :

$$f(w_i) = \frac{freq(w_i, d_j)}{freq(w_i, d)} \tag{2}$$

On note $freq(w_i, d_j)$ la fréquence du mot dans le document et $freq(w_i, d)$ sa fréquence dans le corpus (équation 3).

$$freq(w_i|d_j) = \frac{count(w_i, d_j)}{count(w, d_j)}; freq(w_i|d) = \frac{count(w_i, d)}{count(w, d)} \tag{3}$$

L'objectif est alors de trier les mots-clés obtenus afin de maximiser f, ce qui revient à classer les mots selon leur importance dans le document par rapport à leur importance dans tout le corpus. Après quelques pré-traitements (minusculisation et suppression des accents), utilisation de la morpho-syntaxe (« NOM » et « ADJ ») et de la liste de référence fournie, nous obtenons les résultats suivants :

	Référence			
	Présents		Absents	Total
	Présents dans corpus	Déduction annotateur		
Trouvés	534	0	4.402	4.936
Non trouvés	104	46	-	150
Total	638	46	4.402	5.086

TABLE 4 – Matrice de confusion des mots simples retrouvés dans le corpus après filtrage morpho-syntaxique (« NOM » et « ADJ ») et filtrage sur la liste de mots-clés.

Il est important de noter que les 46 déductions d'annotateurs ne peuvent pas être retrouvés avec cette méthode, nous soustrayons donc que le nombre de mots-clés simples à obtenir au nombre de déductions annotateurs.

Afin de réduire le bruit, nous avons fixé un seuil de $w_i > 0,05$, ce qui a permis de réduire la liste de mots-clés trouvés (Cf. Table 5).

	Référence			
	Présents		Absents	Total
	Présents dans corpus	Déduction annotateur		
Trouvés	520	0	3.836	4.356
Non trouvés	118	46	-	164
Total	638	46	3.836	4.520

TABLE 5 – Matrice de confusion des mots simples retrouvés dans le corpus après filtrage morpho-syntaxique (« NOM » et « ADJ »), filtrage sur la liste de mots-clés et fixation d'un seuil.

Pour mesurer la qualité des résultats obtenus, nous avons utilisé trois indicateurs : précision, rappel et f-mesure (Cf. Table 6). Le seuil utilisé semble avoir légèrement amélioré les résultats obtenus en précision et en f-mesure, ce qui s'explique par la réduction du bruit généré. Cependant, on remarque

que le rappel diminue, ce qui signifie que nous avons supprimé de la liste des mots-clés qui étaient pertinents pour le cas clinique étudié.

	Précision	Rappel	F-mesure
Table 4	0,1082	0,7807	0,1900
Table 5	0,1194	0,7602	0,2063

TABLE 6 – Mesures de qualité calculées sur les matrices de confusion dans les tableaux 4 et 5

2.2.2 Détection des multi-mots

Comme pour la détection de mots simples, il existe des expressions multi-mots qui n'apparaissent pas dans le corpus : les déductions annotateurs. Une expression est une détection annotateur si elle n'apparaît pas telle quelle dans le corpus (après avoir réalisé un pré-traitement sur les données, identique à celui effectué précédemment). Il y a donc 101 expressions multi-mots qui sont des déductions annotateurs, parmi les 329 expressions à trouver, soit près d'un tiers des expressions (Cf. Table 7). Cela nous indique que la détection des expressions strictes dans les données ne sera pas suffisante pour obtenir de bonnes performances d'extraction de mots-clés.

Mots simples			Total
Occurence cas	Occurence discussion	Déduction annotateur	-
247	277	101	329

TABLE 7 – Occurrence des expressions multi-mots à trouver dans le corpus

Dans un premier temps, nous avons extrait les expressions multi-mots qui apparaissent dans les cas cliniques et discussions associés à chaque patient ; cette méthode basique est celle que l'on a nommée *hard matching*. Les résultats obtenus montrent que 20% des multi-mots obtenus sont pertinents pour le cas clinique associé. Parmi les multi-mots de référence, il y en a moins d'un tiers que l'on ne retrouve pas dans le corpus. Cela s'explique par le fait que 97% d'entre-eux n'apparaissent pas tels quels dans le corpus. Pour cette méthode, nous obtenons un faible rappel et une faible précision (cf. Table 8).

Afin d'améliorer le système existant et de s'intéresser aux expressions multi-mots n'apparaissant pas dans le corpus, nous avons testé une méthode que nous avons appelé *fuzzy matching*. Cette méthode consiste à regarder, pour chaque expression multi-mots, si tous les mots qui la composent sont présents dans le corpus. Par exemple, si le multi-mot est « cancer du sein » et que « cancer » et « sein » sont dans le corpus, l'expression est retenue pour le cas clinique. Cette méthode permet de retrouver 78% des expressions multi-mots à trouver, soit 25% de plus qu'avec la méthode *hard matching*. Néanmoins, nous récupérons 3.405 expressions en trop avec cette méthode.

Nous avons ensuite cherché à attribuer un rang de pertinence des mots clés trouvés. Pour ce faire, nous avons traité séparément les expressions obtenues avec un *hard matching* et celles obtenues en *fuzzy matching*. Les expressions apparaissant telles quelles dans le texte seront en tête de liste d'expressions multi-mots. Ensuite, pour ordonner les autres mots-clés, nous leur avons donné la

	Précision	Rappel	F-mesure
Hard Matching	0,2210	0,5393	0,2536
Fuzzy Matching	0,0927	0,7820	0,1658

TABLE 8 – Mesures de qualité calculées avec la méthode *hard matching* et la méthode *fuzzy matching*

somme des occurences de chaque mot w_i qui les composent dans un document d_j. Une liste de mots vides a été construite afin de ne pas comptabiliser les mots de type préposition :

$$weight(w_{1..n}|d_j) = \frac{\sum_{i=1}^{n} count(w_i, d_j)}{n} \quad (4)$$

2.2.3 Ranking

Nous avons cherché à détecter séparément les mots-clés simples et les expressions multi-mots et nous les avons ordonnés différemment. L'objectif est maintenant de mélanger les deux sorties afin d'ordonner les mots simples et les expressions multi-mots ensemble. Pour cela, une règle évidente s'est imposée à nous : les expressions multi-mots qui apparaissent dans le texte tels quels (avec la méthode *hard matching*) sont placés en début de sortie. Il nous reste ensuite à ordonner les mots simples et les expressions multi-mots qui apparaissent uniquement avec la méthode *fuzzy matching*. Pour cela, nous avons utilisé une règle 50/50, qui consiste à récupérer un mot-clé simple et une expression multi-mot, jusqu'à ce qu'il n'y ait plus de mots-clés dans notre liste.

2.3 Résultats

Les meilleurs scores ont été obtenus avec la méthode 1, soit l'utilisation d'*embeddings* suivi d'un calcul de similarité. Malheureusement, nous sommes en dessous de la moyenne des résultats obtenus à cette compétition (qui est de 38,5%).

Méthode	MAP	R-Precision
Run 1 : *embeddings* + similarité	0.3617	0.3243
Run 2 : multi-mots et morpho-syntaxe	0.2732	0.2362

TABLE 9 – Scores obtenus pour la tâche 1

3 Tâche 2 : similarité entre documents

La tâche 2 est une tâche d'appariement entre documents. Les documents à apparier sont d'un côté des « cas » et de l'autre des « discussions ». Les méthodes proposées ici sont basée sur les embeddings de mots ou vecteurs-mots. Après avoir entrainé les embeddings sur le corpus d'apprentissage, la méthode consiste à calculer une représentation vectorielle des cas d'un côté et des discussions de l'autre, puis à calculer des similarités deux à deux entre ces cas et ces discussions, puis à choisir la configuration qui maximise les paires de similarités cas/discussion à l'aide de l'algorithme hongrois (ref). On commence par utiliser Word2Vec (Mikolov *et al.*, 2013) pour transformer les mots en

vecteurs. Plusieurs paramètres ont été testés : la version Skip-Gram et CBOW (qui consiste à prévoir les contextes d'un mot à partir du mot lui-même ou qui consiste à prévoir le mot à partir de son contexte), le voisinage de mot à gauche et à droite (de 2 à 5), la taille de la couche cachée (de 25 à 300) et le nombre d'itérations (500 ou 1000). Les meilleurs résultats sont obtenus avec : Skip-Gram, un voisinage de 5 mots, une couche cachée de taille 300 et 1000 itérations. La fréquence minimale pour les mots est fixée à 3. Plusieurs méthodes de passage des embeddings de mots aux embeddings de documents ont été testées correspondant aux différents run : SWEM-aver, DoCov et DoCov + P-Mean. Une fois calculées les similarités deux à deux entre les cas et les discussions, la configuration optimale est choisie en faisant appel à « l'algorithme hongrois ou méthode hongroise, aussi appelé algorithme de Kuhn-Munkres, algorithme d'optimisation combinatoire, qui résout le problème d'affectation en temps polynomial. C'est donc un algorithme qui permet de trouver un couplage parfait de poids maximum dans un graphe biparti dont les arêtes sont valuées. » [2].

3.1 Run 1 : SWEM-average

La première méthode pour passer des embeddings de mots aux embeddings de documents est la plus simple, elle consiste à moyenner tous les vecteurs-mots du document. Elle est appelée SWEM-aver dans (Shen *et al.*, 2018) pour « Simple Word Embedding Model – average » :

$$z = \frac{1}{L}\sum_{i=1}^{L} v_i \tag{5}$$

avec z, la représentation vectorielle du document, L le nombre de mots du document, v_i le vecteur mot du i^{eme} mot du document. Les meilleurs résultats ont été obtenus avec la variante consistant à pondérer les mots par leur IDF.

3.2 Run 2 : DoCov

La méthode DoCoV (Torki, 2018) pour « Document Co Variance » va calculer une représentation vectorielle du document à partir du triangle supérieur de la matrice de covariance. Ce vecteur aura pour taille $d * \frac{d+1}{2}$ si d est la taille des embeddings (cf. Figure 1).

$$O = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ x_{21} & \cdots & x_{2d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{pmatrix} \quad \sigma_{X,Y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N}$$

$$C = \begin{pmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_d} \\ \sigma_{X_1X_2} & \sigma^2_{X_2} & \cdots & \sigma_{X_2X_d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_1X_d} & \sigma_{X_2X_d} & \cdots & \sigma^2_{X_d} \end{pmatrix} \quad v = \mathbf{vect}(C) = \begin{cases} \sqrt{2}C_{p,q} & \text{if } p < q \\ C_{p,q} & \text{if } p = q \end{cases}$$

FIGURE 1 – DoCov

2. https://fr.wikipedia.org/wiki/Algorithme_hongrois

De même que précédemment, nous avons modifié ces formules pour introduire l'IDF. La taille des embeddings résultant est assez impressionnante. Avec d=300, on obtient un vecteur de taille 45150 !

3.3 Run 3 : DoCov + Pmean

La méthode des « Power Mean » présentée dans (Rücklé *et al.*, 2018) consiste à concaténer plusieurs types de moyennes :

$$\left(\frac{x_1^p + ... + x_n^p}{n}\right)^{1/p} ; p \in \mathbb{R} \cup \{\pm\infty\} \tag{6}$$

Calcul réalisé pour chaque composante des vecteurs, avec x_i la i^{eme} composantes des n vecteurs composant le document. Avec $p = 1$, on retrouve la mesure SWEM-aver précédente. De la même manière que précédemment, nous avons introduit le coefficient IDF. Les valeurs des composantes x_i doivent obligatoirement être positives. Comme ce n'est pas le cas avec nos données, nous avons choisi des valeurs paires pour p (sauf pour $p = 1$). Il suffit ensuite de concaténer les vecteurs obtenus avec p=1 et p=2 pour obtenir un vecteur de taille $2 * d$. L'idée intuitive est de « capter » plus d'informations par ces concaténations. Nous avons fait des tests avec $p = 1, 2$ puis $p = 1, 2, 4$ puis $p = 1, 2, 4, 6$ puis $p = 1, 2, 4, 6, 8$. Comme DoCoV obtenait de bons résultats, on vient ajouter aux embeddings obtenus avec DoCoV ceux obtenus avec PMean avec p=1,2.

3.4 Résultats

Le meilleur score est obtenu avec la méthode DoCoV, qui est le meilleur résultat obtenu à cette compétition.

Méthode	Score
Run 1 : SWEM-Aver	88,79 %
Run 2 : DoCov	95,32 %
Run 3 : DoCov + Pmean (1,2)	93,45 %

TABLE 10 – Scores obtenus pour la tâche 2

4 Tâche 3 : extraction d'information

Dans cette tâche d'extraction d'information, il est nécessaire de repérer, dans les cas cliniques, les informations démographiques et cliniques.

4.1 Détection du genre des patients

A partir des cas cliniques, l'objectif est de déterminer si la personne est de genre « féminin », « masculin » ou s'il s'agit de plusieurs personnes, auquel cas « féminin|masculin ». Il y a également une catégorie d'individus non classés.

4.1.1 Méthode

La distribution des genres des patients est présentée dans la Table 11. Bien qu'il y ait plus de « masculin » que de « féminin », les proportions ne sont pas significativement différentes pour les utiliser dans l'étude. On remarque que « féminin » désigne à la fois un cas clinique sur une personne de genre « féminin », mais aussi sur plusieurs personnes de ce genre, tout comme « masculin ».

féminin	masculin	féminin|masculin	NA
117	168	4	1

TABLE 11 – Répartition des genres dans le corpus

Nous avons également remarqué un cas difficile à détecter dans le corpus : on parle d'un patient de 57 ans, et on précise bien, à de nombreuses reprises, qu'il s'agit d'un patient de sexe masculin, et le genre associé au patient est « féminin ».

Pour extraire les informations sur le genre, nous avons utilisé deux lexiques :

1. un lexique qui contient des mots relatifs à une désignation d'une personne de genre « féminin » comme « fillette », « femme », « patiente », « sexe féminin » etc. Nous avons également ajouté les situations qui sont plus fréquentes chez une personne de genre « féminin » : « violée », « ménopausée », « enceinte » etc. Enfin, nous avons ajouté les verbes conjugués qui reviennent très souvent dans le corpus : « âgée de » et « née le » ainsi que les parties du corps présentent le plus souvent chez ces personnes : « vagin », « vulve », « ovaire » etc.

2. un lexique qui contient des mots des mots relatifs à une désignation d'une personne de genre « masculin » comme « garçon », « monsieur », « patient », « sexe masculin » etc. On a également ajouté des situations qui sont plus fréquentes chez une personne de genre « masculin » : « infertilité », « circoncit », « éjaculer », « flaccide » etc. Enfin, nous avons ajouté des verbes conjugués qui reviennent très souvent dans le corpus : « âgé de » et « né le » ainsi que des parties du corps présentent le plus souvent chez ces personnes : « pénis », « testicule », « prostate » etc.

Ces lexiques ont été utilisés pour nous permettre de comptabiliser les mots significatifs pouvant apparaître dans un cas clinique. Si on ne trouve que des mots appartenant au lexique « féminin », on associera ce genre au cas clinique, de même pour le genre masculin. En revanche, si on voit des mots appartenant aux deux lexiques, le mot majoritaire « gagne », et la classe « féminin|masculin » est obtenue uniquement s'il y a autant de mots des deux lexiques dans le cas clinique.

4.2 Détection de l'âge des patients

A partir des cas cliniques, l'objectif est de déterminer l'âge du ou des patients. Il y a également une catégorie d'individus pour lesquels il n'y a pas d'âge (non classés). Pour cette tâche, l'objectif est d'utiliser un ensemble de règles d'extraction pour pouvoir déterminer l'âge du ou des patients en utilisant le cas clinique.

4.3 Méthode

Afin d'extraire l'âge du ou des patients, nous avons choisi d'utiliser une méthode à base de règles d'extraction. Pour cela, il a fallu prendre en compte différents cas :

— l'âge de l'individu est indiqué en années : nous avons seulement à récupérer l'âge associé à l'individu.
— l'âge de l'individu est indiqué en mois : une conversion en années s'impose, en arrondissant le résultat obtenu à l'âge inférieur auquel il correspond. Par exemple, pour un enfant de 3 mois, nous arrondirons l'âge à 0 ans.
— l'âge de l'individu, en mois ou en années, est écrit en lettres : nous utilisons un dictionnaire qui nous permet d'associer un chiffre ou un nombre écrit en toutes lettres à l'écriture chiffrée correspondante.
— un adjectif, comme « quinquagénaire », est donné pour indiquer l'âge du patient : un dictionnaire est également utilisé pour parer à cette éventualité.
— plusieurs patients sont présents dans le cas clinique : extraction de déclencheurs comme « âgées de », « âges respectifs » etc.

4.4 Détection de l'issue

A partir des cas cliniques, l'objectif est de déterminer l'état du patient parmi : amélioration, stable, détérioration ou décès. Il y a également une catégorie d'individus non classés.

4.4.1 Méthode

Dans cette partie, l'objectif est d'essayer de retrouver les classes en regroupant les documents avec des techniques de *clustering*. Cela nous permet d'évaluer la similarité des documents entre plusieurs matrices. Nous avons utilisé une méthode de *clustering* de type *spherical k-means*, utilisant la distance cosinus.

La classification sera réalisée à partir d'une matrice **document-vecteur** obtenue après entraînement d'un **modèle doc2vec** PV-DBOW (Le & Mikolov, 2014) sur le corpus : prédit le mot cible à partir de plusieurs mots du contexte. Exemple : prédiction de « lit » à partir de la séquence « le chat s'asseoit sur le ».

Nous avons utilisé un lexique de mots nous permettant, avant le calcul de doc2vec, de repérer les cas cliniques ayant une issue « décès », ces cas étant assez faciles à détecter à l'aide de règles d'extraction.

4.5 Résultats

Malheureusement, les résultats obtenus sont en dessous de ceux obtenus durant la compétition.

Age		Genre		Issue	
Precison	Rappel	Precison	Rappel	Precison	Rappel
0.93925	0.46744	0.96667	0.47209	0.36150	0.18033

TABLE 12 – Scores obtenus pour la tâche 3

5 Conclusion

Participer à la campagne DEFT 2019, nous a permis de tester plusieurs méthodes basées sur des règles linguistiques et des plongements de mots. Aucune donnée supplémentaire, autre que les données d'apprentissage, n'a été utilisée. Les résultats obtenus sont satisfaisants. Les méthodes que nous avons mises en œuvre sont facilement transposables à d'autres tâches et peuvent intéresser plusieurs entités du groupe EDF.

Références

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). Cas : French corpus with clinical cases. In *LOUHI 2018 : The Ninth International Workshop on Health Text Mining and Information Analysis*.

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 de DEFT*.

LE Q. & MIKOLOV T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, p. 1188–1196.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, p. 3111–3119.

RÜCKLÉ A., EGER S., PEYRARD M. & GUREVYCH I. (2018). Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv :1803.01400*.

SCHAKEL A. M. & WILSON B. J. (2015). Measuring word significance using distributed representations of words. *arXiv preprint arXiv :1508.02297*.

SHEN D., WANG G., WANG W., MIN M. R., SU Q., ZHANG Y., LI C., HENAO R. & CARIN L. (2018). Baseline needs more love : On simple word-embedding-based models and associated pooling mechanisms. *arXiv preprint arXiv :1805.09843*.

SPARCK JONES K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, **28**(1), 11–21.

TORKI M. (2018). A document descriptor using covariance of word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 527–532.

WILSON B. J. & SCHAKEL A. M. (2015). Controlled experiments for word embeddings. *arXiv preprint arXiv :1510.02675*.

Participation de l'équipe LAI à DEFT 2019

Jacques Hilbey[1] Louise Deléger[2] Xavier Tannier[3]
(1) Inserm, LIMICS
(2) MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
(3) Sorbonne Université, Inserm, LIMICS, Paris, France
jacques.hilbey@inserm.fr, louise.deleger@inra.fr,
xavier.tannier@sorbonne-universite.fr

RÉSUMÉ

Nous présentons dans cet article les méthodes conçues et les résultats obtenus lors de notre participation à la tâche 3 de la campagne d'évaluation DEFT 2019. Nous avons utilisé des approches simples à base de règles ou d'apprentissage automatique, et si nos résultats sont très bons sur les informations simples à extraire comme l'âge et le sexe du patient, ils restent mitigés sur les tâches plus difficiles.

ABSTRACT

Participation of team LAI in the DEFT 2019 challenge

We present in this article the methods developed and the results obtained during our participation in task 3 of the DEFT 2019 evaluation campaign. We used simple rule-based or machine-learning approaches ; our results are very good on the information that is simple to extract (age, gender), they remain mixed on the more difficult tasks.

1 Introduction

Nous présentons dans cet article les méthodes que nous avons conçues pour répondre à la tâche 3 de la campagne d'évaluation DEFT 2019. Cette tâche porte sur l'extraction d'informations dans un corpus de cas cliniques (Grabar *et al.*, 2018). Quatre types d'information doivent être identifiés :

— l'âge du patient ;
— le genre du patient ;
— l'origine : le motif de la consultation ou de l'hospitalisation ;
— l'issue, à déterminer parmi cinq valeurs possibles : guérison, amélioration, stable, détérioration ou décès.

Le jeu d'entraînement fourni par les organisateurs est composé de 290 documents, tandis que le jeu de test est composé de 427 cas. Plus de détails sur le défi sont présents dans Grabar *et al.* (2019).

2 Méthode

Nous avons mis en place des méthodes à base de règles et/ou d'apprentissage automatique, selon le type d'information à extraire.

2.1 Âge et genre

En étudiant le corpus d'entraînement, nous avons constaté que les informations d'âge et de sexe étaient souvent exprimées sous des formes similaires avec relativement peu de variations dans les documents. Ceci a orienté notre choix de méthode vers une approche à base de règles. Il nous a en effet semblé que la majorité de ces informations pourrait être capturée à l'aide d'un ensemble relativement restreint de règles.

Nous avons implémenté nos règles à l'aide de l'outil PyRATA (Hernandez & Hazem, 2018). PyRATA est une librairie python qui permet d'écrire des règles de type expressions régulières s'appliquant à des structures de données plus complexes qu'une chaîne de caractères, en particulier sur des suites de tokens possédant différents attributs (lemmes, étiquettes morpho-syntaxiques, etc.).

Pour l'âge, nous avons conçu un ensemble de motifs composés soit de déclencheurs (*âgé de, patient de, etc.*) suivi d'un nombre (en chiffre ou en lettre) et d'une unité de temps, soit de noms dénotant des classes d'âge (*quadragénaire, quinquagénaire, etc.*). Les âges exprimés en mois ont été convertis en années (en pratique, 0 ou 1).

Pour le genre, nous avons établi des listes d'expressions évocatrices du genre masculin (*homme, patient, Monsieur, testicule, un enfant, prénom masculin, etc.*) et du genre féminin (*femme, patiente, Madame, utérus, une enfant, prénom féminin, etc.*).

Dans les rares cas où plusieurs personnes sont concernées par le cas, nous nous sommes assurés de la cohérence entre le nombre d'âges et de genres retournés, en considérant le plus petit nombre.

Avant d'appliquer nos règles, nous pré-traitons le corpus (segmentation en mots, analyse morphosyntaxique) à l'aide de la librairie python spaCy [1].

2.2 Origine

Comme pour l'âge et le sexe, nous avons adopté une approche à base de règles, implémentées à l'aide de l'outil pyRATA, pour capturer l'information relative à l'origine de l'admission.

Celle-ci se trouve le plus souvent, sans surprise, dans la première ou la deuxième phrase du cas. Un premier motif est apparu : une portion de phrase ayant pour introducteur la préposition *pour* et pour terminateur le point final de la phrase. Nous avons procédé ensuite par ajout d'autres introducteurs possibles liés soit au patient (*présenter, souffrir, subir, se plaindre, etc.*) soit au processus médical (*prise en charge, diagnostic, tableau, bilan, etc.*).

Dans un deuxième temps, nous avons étendu les terminateurs possibles à d'autres ponctuations et préféré éventuellement, au simple *pour*, des expressions plus spécifiques (*hospitaliser / admettre / consulter pour*).

2.3 Issue

L'issue caractérise l'évolution de l'état clinique du patient entre son admission et la fin de la consultation ou de l'hospitalisation. Nous avons suivi deux voies différentes pour capter cette évolution.

1. `https://spacy.io/`

Une première méthode (*run 1*) a consisté à utiliser des tests du χ^2 afin de déterminer quels N-grammes (avec N de 1 à 3), dans les cas cliniques présentant une même issue, lui étaient le plus fortement associés, puis à sélectionner manuellement pour chaque issue une dizaine de ces suites de mots en évitant autant que possible ceux qui paraissaient trop caractéristiques du jeu d'entraînement. En recherchant dans chaque cas clinique ces N-grammes, il a été possible d'établir pour chacun un score par issue et d'associer au cas l'issue ayant le score maximal. En cas de score nul, l'issue attribuée était 'NUL' et en cas d'égalité entre plusieurs issues, la plus fréquente de celles-ci dans le jeu d'entraînement.

La deuxième méthode (*run 2*) a consisté à décrire la deuxième moitié de chaque cas clinique (où les informations relatives à l'issue étaient le plus susceptibles de se trouver) dans un espace vectoriel de type *one-hot vectors* (sac de mots), avec une pondération TF-IDF, puis à appliquer aux vecteurs ainsi constitués des classifieurs multi-classe courants (régression logistique, bayésien naïf multinomial, séparateur à vaste marge linéaire, forêt aléatoire, arbre de décision, k plus proches voisins) dont nous avons fait varier les paramètres pour choisir finalement la meilleure configuration sur le jeu d'entraînement, par validation croisée.

Le cas de plusieurs issues, rarissime dans le corpus d'entraînement, a été ignoré.

	Précision	Rappel
Âge	0.986	0.953
Genre	0.993	0.983
Issue (*run 1*)	0.693	0.619
Issue (*run 2*)	Exactitude (*Accuracy*)	
	0.524	

TABLE 1 – Performances pour l'âge, le genre et l'issue sur les données d'entraînement (pour le *run 2*, performance en validation croisée utilisée pour le choix du meilleur classifieur).

	Précision	Rappel	F-Mesure	Meilleure F-mesure
Âge	0.980	0.919	0.948	0.948
Genre	0.981	0.974	0.978	0.978
Issue (*run 1*)	0.486	0.405	0.442	0.505
Issue (*run 2*)	0.498	0.492	0.495	

TABLE 2 – Performances pour l'âge, le genre et l'issue sur les données de test (résultats officiels fournis par les organisateurs du challenge), comparées aux meilleurs résultats parmi les participants.

3 Résultats et discussion

Les règles de reconnaissance de l'âge et du genre du patient donnent de très bons résultats sur le corpus d'entraînement (tableau 1). Les performances sur le corpus de test sont également bonnes mais légèrement en baisse (tableau 2), en particulier pour l'âge qui perd environ 3 points en rappel. Nos règles semblent donc dans l'ensemble assez robustes, mais manquent certains cas.

Dans les mêmes tableaux, il est intéressant mais pas surprenant de constater que l'approche à base de règles pour l'issue connaît une forte baisse de performance entre le jeu d'entraînement et le jeu de

Origine	Macro-	Précision	0.582
		Rappel	0.722
		F-mesure	0.645
		Meilleure F-mesure	0.666
	Micro-	Précision	0.628
		Rappel	0.735
		F-mesure	0.677
		Meilleure F-mesure	0.677
		overlap-accuracy	0.600

TABLE 3 – Performances pour l'origine (admission) sur les données de test (résultats officiels fournis par les organisateurs du challenge), comparées aux meilleurs résultats parmi les participants.

test. L'approche par apprentissage est en revanche plus robuste (le classifieur finalement retenu a été un séparateur à vaste marge (SVM) linéaire avec les paramètres $C = 10^{-6}$ et $tol = 10^{-6}$) et conduit à notre meilleur résultat sur le test. Dans les deux cas, la difficulté de la tâche cumulée au faible nombre de documents d'entraînement ne permet pas d'obtenir des résultats satisfaisants avec ces approches simples.

Les matrices de confusion pour l'entraînement (Figure 1) et pour le test (Figure 2) permettent de mettre en évidence la difficulté pour nos deux approches de distinguer efficacement les classes *amélioration* et *guérison*, ce qui représente la plus grande partie des erreurs des systèmes. Il serait intéressant de savoir si les chiffres d'accords inter-annotateurs du corpus illustrent les mêmes difficultés.

Enfin, la table 3 présente les résultats obtenus pour l'extraction du motif de l'admission (origine) avec les métriques officielles du défi.

4 Conclusion

Si nos résultats sont, selon les cas, les meilleurs ou tout proches des meilleurs du challenge, nous ne sommes malheureusement pas parvenus à produire des approches innovantes et performantes sur cette tâche du défi DEFT 2019. Nos résultats proviennent d'approches classiques et conduisent à des résultats sans surprise : très bons sur les informations simples à extraire comme l'âge et le sexe, plus mitigés sur les informations pouvant s'exprimer de façon très variable dans les textes.

Nous avons également réalisé des expérimentations sur la tâche 1 (extraction de mots-clés), avec une extraction de termes puis une adaptation de la Word Mover Distance (Kusner *et al.*, 2015) pondérée par les tf.idf des termes, pour estimer les termes les plus représentatifs de chaque paire cas/discussion, mais les résultats se sont avérés peu convaincants.

Remerciements

Nous remercions les organisateurs pour la création du corpus ainsi que pour l'organisation du défi, ainsi qu'Ivan Lerner pour ses expériences préliminaires.

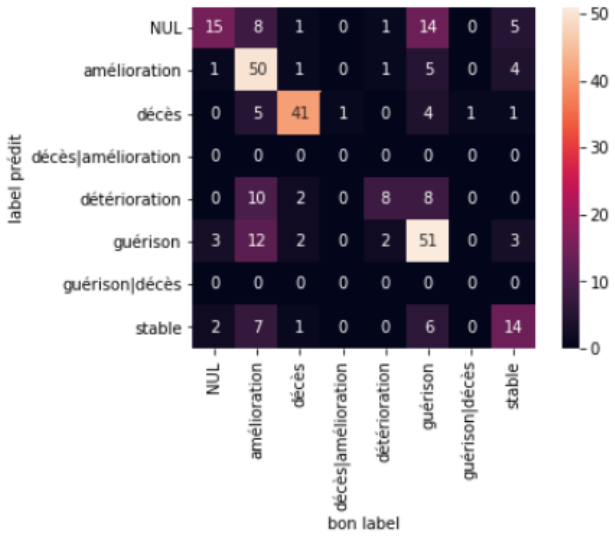
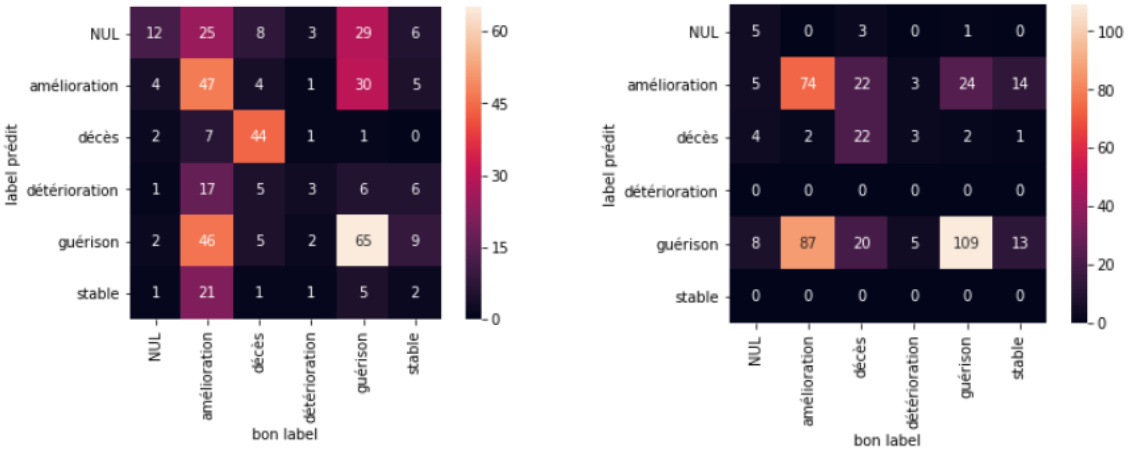


FIGURE 1 – Matrice de confusion pour l'*issue (run 1)* sur le jeu d'entraînement.



run 1 *run 2*

FIGURE 2 – Matrice de confusion pour l'*issue* sur le jeu de test (run 1 à base de règles et run 2 par apprentissage).

Références

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French corpus with clinical cases. In *LOUHI 2018 : The Ninth International Workshop on Health Text Mining and Information Analysis*.

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 de DEFT*.

HERNANDEZ N. & HAZEM A. (2018). PyRATA, Python Rule-based feAture sTructure Analysis. In N. C. C. CHAIR), K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Eds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France : European Language Resources Association (ELRA).

KUSNER M. J., SUN Y., KOLKIN N. I. & WEINBERGER K. Q. (2015). From Word Embeddings To Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, Lille, France.

DÉfi Fouille de Textes 2019: indexation par extraction et appariement textuel

Jean-Christophe Mensonides[1] Pierre-Antoine Jean[1]
Andon Tchechmedjiev[1] Sébastien Harispe[1]
(1) LGI2P, IMT Mines Ales, Univ Montpellier, Ales, France
{prénom}.{nom}@mines-ales.fr

RÉSUMÉ

Cet article présente la contribution de l'équipe du Laboratoire de Génie Informatique et d'Ingénierie de Production (LGI2P) d'IMT Mines Alès au DÉfi Fouille de Textes (DEFT) 2019. Il détaille en particulier deux approches proposées pour les tâches liées à (1) l'indexation et à (2) la similarité de documents. Ces méthodes reposent sur des techniques robustes et éprouvées du domaine de la Recherche d'Information et du Traitement Automatique du Langage Naturel, qui ont été adaptées à la nature spécifique du corpus (biomédical/clinique) et couplées à des mécanismes développés pour répondre aux spécificités des tâches traitées. Pour la tâche 1, nous proposons une méthode d'indexation par extraction appliquée sur une version normalisée du corpus (MAP de 0,48 à l'évaluation) ; les spécificités de la phase de normalisation seront en particulier détaillées. Pour la tâche 2, au-delà de la présentation de l'approche proposée basée sur l'évaluation de similarités sur des représentations de documents (score de 0,91 à l'évaluation), nous proposons une étude comparative de l'impact des choix de la distance et de la manière de représenter les textes sur la performance de l'approche.

ABSTRACT

DEFT 2019 : extraction-based document indexing and textual document similarity matching

This paper presents the contribution of the LGI2P (Laboratoire de Génie Informatique et d'Ingénierie de Production) team from IMT Mines Alès to the DEFT 2019 challenge (DÉfi Fouille de Textes). We detail two approches we devised for the tasks pertaining to (1) the indexing and to (2) the similarity of documents. Said approaches rely on proven and robust techniques from Information Retrieval and Natural Language Processing that have been adapted to the specificities of the corpus (biomedical text) and of the formulation of the tasks. For task 1, we propose an indexing-by-extraction approach applied on the corpus after a normalisation procedure (MAP=0.48) that we will detail further. For task 2, we proposed a similarity-based approach computed on vector representation of the documents (score=0.910) and study the impact of the choice of the similarity metric and of the document representation method on task performance.

MOTS-CLÉS : Indexation de documents, similarité sémantique, recherche d'information, corpus biomédical.

KEYWORDS: Document indexing, semantic similarity, information retrieval, biomedical corpus.

1 Tâche 1 : Indexation des cas cliniques

La tâche d'extraction de mots-clés consiste à distinguer les mots de l'unité linguistique analysée (*e.g.* document, corpus) qui sont caractéristiques de l'unité au regard d'un objectif prédéfini. Elle est généralement exploitée dans l'indexation (Marchand *et al.*, 2016) et le résumé de documents textuels (Gupta & Lehal, 2010). Les mots-clés peuvent dans certains cas être conceptualisés comme un sous-ensemble de méta-données de nature sémantique associées à ces documents. La stratégie utilisée pour l'obtention des mots-clés permet de distinguer deux types d'approches au sein de la tâche d'indexation de documents : les approches par extraction et les approches par assignation (Chartier *et al.*, 2016). La principale différence entre ces deux types d'approches repose sur la provenance des mots-clés. Les approches par extraction s'emploient à annoter un document avec des mots issus du document, tandis que les approches par assignation visent à aligner un document avec une liste de mots-clés issus d'un vocabulaire contrôlé sans nécessairement contraindre les mots-clés sélectionnés à apparaître dans le document.

Le premier type d'approche, par extraction, repose principalement sur une évaluation de la *pertinence* des termes d'un document pour sa caractérisation. La notion de pertinence peut être abordée de différentes manières, *e.g.* statistique, par apprentissage supervisé ou non-supervisé. L'étude statistique permet notamment d'évaluer les termes au travers de leur usage au sein des documents et du corpus. Une des approches les plus courantes repose sur le calcul du coefficient de pondération TF-IDF (Jones, 1972). Ce coefficient cherche à retranscrire l'importance relative d'un terme par rapport à un document et à l'ensemble du corpus. Sur la base de ce modèle, l'importance d'un terme sera d'autant plus grande si le terme apparaît fréquemment dans le document et peu fréquemment dans le corpus. D'autres exemples de coefficients de pondération employés dans le domaine de la classification de textes pour la pondération de termes conditionnée à une classe donnée peuvent également être cités *e.g.* le RDF (*Relevant Document Frequency*), l'IG (*Information Gain*), le MI (*Mutual Information*) ou bien encore le *Chi Square* (Nanas *et al.*, 2003; Hamdan, 2015).

La pertinence d'un terme pour caractériser un document peut également être évaluée à partir de méthodes d'apprentissage supervisé et non-supervisé. Parmi les méthodes d'apprentissage supervisé, KEA - *Keyphrase Extraction Algorithm* (Witten *et al.*, 2005) - exploite un modèle bayésien en utilisant le TF-IDF et le ratio de la première occurrence du mot-clé dans un texte comme fonctions caractéristiques. Une autre approche, proposée par Zhang (2008), se base sur un CRF (*Conditional Random Field*) qui exploite des fonctions caractéristiques liées à la position (*e.g.* présence/absence d'un mot au sein du résumé ou dans le corpus du texte), lexicales (*e.g.* fonction grammaticale du mot) et statistiques (*e.g.* TF-IDF) afin d'identifier et de généraliser les mots-clés au sein des textes. Concernant les modèles d'apprentissage non supervisé, le LSI (*Latent Semantic Indexing*) est un exemple d'approches fréquemment utilisées (Deerwester *et al.*, 1990). Elle se base sur une décomposition matricielle de type SVD (*Singular Value Decomposition*) appliquée à une matrice termes-documents décrivant les occurrences des termes dans les documents. Cette approche permet alors d'indexer les documents par un ensemble de *concepts* (compositions linéaires des représentations des termes dans la matrice initiale).

Le second type d'approche, par abstraction, se focalise plus particulièrement sur le rapprochement sémantique des termes. Ce type de procédé peut s'appuyer sur des mesures de similarité sémantique basées sur une structuration de la connaissance fournie *a priori* et/ou sur l'analyse de corpus de textes. Les mesures de similarité sémantique basées sur une structuration de la connaissance (*e.g.* ontologie, vocabulaire structuré) permettent d'évaluer la similarité de sens entre concepts/termes

selon l'information modélisée par la structuration de la connaissance (Harispe *et al.*, 2014). Par exemple, l'approche USI - *User-oriented Semantic Indexer* (Fiorini *et al.*, 2015) - exploite cette stratégie d'abstraction afin d'indexer des articles biomédicaux.

Les approches basées sur l'analyse de corpus de textes reposent très souvent explicitement ou implicitement sur l'hypothèse distributionnelle qui considère que le sens d'un terme est donné par son voisinage dans les textes, *i.e.* son usage. On distingue aujourd'hui les analyses traditionnelles des fréquences de co-occurrences, des approches par plongements sémantiques de mots récemment popularisées par l'approche WORD2VEC (Mikolov *et al.*, 2013). Des travaux liés au résumé textuel par abstraction démontrent l'efficacité d'employer ce type d'approche couplée à un réseau neuronal récurrent (Nallapati *et al.*, 2016).

Enfin, nous pouvons évoquer des approches hybrides dont la spécificité reposent sur la manière de représenter les textes sous la forme de graphes afin d'en extraire des mots-clés pertinents par extraction ou par abstraction. Ces approches s'appuient sur des algorithmes de parcours de graphes pondérés (Wang *et al.*, 2014; Mahata *et al.*, 2018) ou non pondérés (Litvak & Last, 2008) ou bien de recherche de motifs *e.g.*, recherche des cliques maximales (Kim *et al.*, 2014). Ces graphes peuvent être dirigés dans le cas par exemple où le graphe est construit en tenant compte de la succession des termes ou non dirigés si le graphe s'appuie sur une matrice de co-occurrences.

La stratégie d'indexation définie par notre équipe pour cette première tâche a été déterminée suite à l'étude statistique du corpus présentée en section 1.1. La section 1.2 présente la méthodologie mise en place et les différents résultats obtenus.

1.1 Description du corpus et métriques d'évaluation

1.1.1 Description et statistiques descriptives du corpus

Les tâches de cette quinzième édition du défi DEFT reposent sur un corpus d'entraînement constitué de 290 couples de textes de cas clinique/discussion (Grabar *et al.*, 2019). A chaque couple est associé un ensemble de mots-clés qui ont été définis manuellement au terme d'un consensus entre deux annotateurs. L'ensemble des mots-clés utilisés pour annoter les couples cas clinique/discussion forment un vocabulaire contrôlé. Le tableau 1 résume différentes statistiques liées aux données fournies dans le cadre du défi DEFT 2019.

# couples cas clinique/discussion dans $\mathcal{C}$	290
# moyen de mots dans les cas cliniques de $\mathcal{C}$	332
# moyen de mots dans les discussions de $\mathcal{C}$	764
# de mots-clés dans le vocabulaire contrôlé	1311
# de mots-clés utilisés dans $\mathcal{C}$	1123 (0.85%)
# de mots-clés avec une correspondance exacte dans les cas cliniques de $\mathcal{C}$	441
# de mots-clés avec une correspondance exacte dans les discussions de $\mathcal{C}$	658
# de mots-clés abstraits au sein d'un couple de $\mathcal{C}$	390

TABLE 1 – Statistiques sur les couples cas clinique/discussion et sur les mots-clés associés. Avec $\mathcal{C}$ le jeu d'entraînement et # un symbole signifiant nombre.

La figure 1 présente également des statistiques intéressantes liées au corpus et au vocabulaire contrôlé.

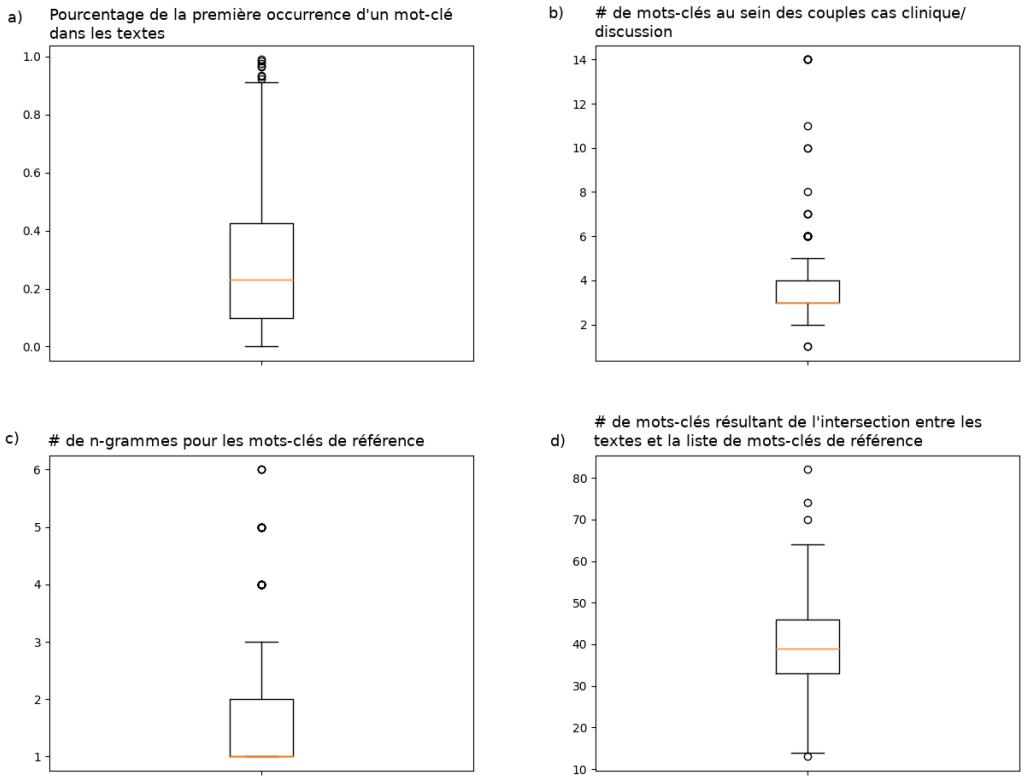


FIGURE 1 – Statistiques, présentées sous forme de *boxplot*, du corpus de textes et de mots-clés.

Ces dernières sont venues nourrir nos réflexions sur les méthodologies à privilégier pour la tâche 1. Nous discutons les différentes informations présentées dans cette figure :

a) la proportion de première occurrence d'un mot-clé au sein d'un cas clinique et d'une discussion. La plupart des premiers mots-clés qui apparaissent dans les textes apparaissent au début du texte.

b) le nombre de mots-clés à attribuer à chaque couple cas clinique/discussion. La plupart des couples sont annotés par 3 ou 4 mots-clés.

c) le nombre de grammes (mots) par mot-clé. Les couples sont essentiellement annotés par des unigrammes ou bi-grammes.

d) le nombre de mots-clés potentiels lorsque l'on considère l'intersection entre le vocabulaire contrôlé et les couples de textes. En médiane, nous observons qu'un couple contient des occurrences de 40 mots-clefs.

Nous avons aussi analysé le taux de recouvrement entre les mots-clés qui apparaissent dans les couples et les mots-clés attendus pour ces couples. Ce taux est calculé à partir de l'équation 1 dans laquelle $\mathcal{C}$ représente l'ensemble des couples cas clinique/discussion, M l'ensemble des mots-clés issus du vocabulaire contrôlé, T_c l'ensemble des mots du couple cas clinique/discussion c et K_c l'ensemble des mots-clés associés au couple cas clinique/discussion c.

$$R = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{|(M \cap T_c) \cap K_c|}{|K_c|} \quad (1)$$

Le taux de recouvrement global R des mots-clés attendus est de 0,72 *i.e.* 72% des mots-clefs attendus dans le jeu d'entraînement apparaissent dans les couples correspondants. Ainsi, dans le meilleur des cas, la MAP (cf. sous-section 1.1.2) pouvant être obtenue à l'aide d'une approche par extraction sera de 0,72 pour le jeu d'entraînement - un score pouvant être considéré comme respectable lorsque la tâche d'annotation est complexe. En tenant compte de l'observation d) proposée ci-dessus - en médiane, un couple contient des occurrences de 40 mots-clefs -, nous notons qu'il est possible de réduire la complexité de la tâche d'annotation en la redéfinissant comme une tâche d'extraction (au prix d'une réduction des performances théoriques maximales pouvant être atteintes, acceptable si la tâche d'annotation s'avère complexe). En la redéfinissant comme telle, il s'agit de définir une approche qui distinguera les mots-clefs pertinents parmi ceux observés dans les couples. Cette approche par extraction a l'avantage de réduire l'espace de recherche de manière significative : lors de l'annotation d'un couple, l'espace de recherche composé initialement de 1311 mots-clés est réduit à 40 mots-clés en médiane par couple. Les nombreux tests préliminaires effectués sur cette tâche sans réduction de l'espace de recherche, *i.e.* sans se restreindre à une approche par extraction, avec des stratégies multiples (statistique et par apprentissage), nous ont permis d'apprécier la potentielle difficulté de la tâche, et nous ont alors amené à concentrer nos efforts sur des approches par extraction. Ceux-ci seront détaillés par la suite.

1.1.2 Métrique d'évaluation

La métrique d'évaluation de cette première tâche est la MAP (*Mean Average Precision*). C'est une métrique populaire dans le domaine de la recherche d'information (Yue *et al.*, 2007). Chaque entrée est représentée par un vecteur binaire p, qui est l'espace des mots-clés de référence où la pertinence d'un mot-clé est symbolisée par un 1, et par un vecteur $\hat{p}$ dans lequel les mot-clés sont classés du plus pertinents au moins pertinent. Par conséquent, le classement des mots-clés récupérés a une incidence sur le score final. L'équation 2 formalise la modalité de calcul de la MAP.

$$MAP(p, \hat{p}) = \frac{1}{rel} \sum_{j:p_j=1} Prec@j \tag{2}$$

où $rel = |i : p_i = 1|$ est le nombre de mots-clés attendus et $Prec@j$ est le nombre de mots-clés pertinents dans les j premiers mots-clés issus du vecteur $\hat{p}$. À noter que lors des années précédentes du défi DEFT (2012 et 2016) la F1-mesure avait été utilisée.

1.2 Méthodologies et résultats

La méthodologie proposée repose sur 3 principales phases : i) un pré-traitement sur l'ensemble des mots-clés M issus du vocabulaire contrôlé et sur les textes des couples cas clinique/discussion, ii) le calcul d'un score de pondération TF-IDF des mots-clés de M pour chaque couple cas clinique/discussion et iii) le classement des mots-clés suite à une étape de post-traitement tenant compte des coefficients calculés lors de la précédente phase. Ces 3 étapes sont détaillées par la suite.

1.2.1 Pré-traitement des données

De nombreux termes au sein de la liste des mots-clés sont lexicalement proches et sémantiquement identiques, tels que « urètre » et « urèthre ». Ces termes ne sont cependant pas considérés comme similaires lors de la phase d'évaluation dû à leur différence orthographique, bien qu'ils représentent de manière implicite une même entité conceptuelle. En vue de réduire tant que possible la variabilité des observations lors de la phase de calcul des coefficients de pondération TF-IDF, et cela sans induire une perte dommageable d'information, nous avons souhaité adopter une approche permettant de considérer de tels termes comme identiques.

Un premier traitement est appliqué à l'ensemble des unigrammes extraits des éléments de M (mots-clés du vocabulaire contrôlé). La ponctuation, les chiffres ainsi que les *stopwords* ont été supprimés. Chaque unigramme est lemmatisé et racinisé pour obtenir un ensemble U d'unigrammes normalisés.

Dans l'objectif de regrouper les unigrammes proches sémantiquement (*e.g* « urètre » et « urèthre »), un second traitement est appliqué. Un nouvel ensemble U' d'unigrammes est construit à l'aide d'une mesure de similarité sémantique appliquée sur les éléments de U. Pour chaque couple d'unigrammes $(u_i, u_j) \in U^2$ avec $i \neq j$, les radicaux r_{u_i} et r_{u_j} sont extraits en soustrayant un préfixe et un suffixe communs à u_i et u_j, s'ils existent dans des listes prédifinies [1]. Une distance cosinus $cos(Emb(r_{u_i}), Emb(r_{u_j}))$ est ensuite calculée, avec $Emb(.)$ la fonction vecteur de plongement sémantique. Si $cos(Emb(r_{u_i}), Emb(r_{u_j})) > \lambda_0$ avec $\lambda_0 \in \mathbb{R}^+$, u_i et u_j sont considérés comme similaires, et dans ce cas seul l'unigramme le plus court est conservé. L'intérêt des préfixes et des suffixes est de limiter la similarité sémantique d'unigrammes spécifiques au domaine étudié. En effet, si deux unigrammes (u_i, u_j) ont pour suffixe « sarcome », tels que « liposarcome » et « carcinosarcome », $cos(Emb(u_i), Emb(u_j))$ sera proche de 1 bien qu'une distinction entre les deux soit nécessaire, alors que $cos(Emb(\text{"}lipo\text{"}), Emb(\text{"}carcino\text{"}))$ sera plus faible. Les vecteurs caractérisants le plongement sémantique de chaque unigramme, obtenus avec la méthode de Bojanowski *et al.* (2017) [2], correspondent à une moyenne pondérée de n-grammes de caractères, et sont donc particulièrement adaptés au traitement de racines d'unigrammes. Enfin, une abstraction M' de la liste des mots-clés de référence M est construite en substituant les unigrammes des éléments de M par ceux de U'.

L'ensemble des mots composants les cas cliniques et les discussions bénéficient d'un traitement similaire à celui appliqué aux mots-clés. Chaque couple est représenté par la concaténation du texte normalisé du cas clinique et de la discussion. Seuls les unigrammes présents dans U' sont conservés. Certains unigrammes ne sont caractéristiques que d'un seul mot-clé, tels que « *escherichia* » qui ne peut former que le mot-clé « *escherichia coli* ». Afin de pouvoir attribuer un coefficient de pondération TF-IDF non nul à ces mots-clés lorsqu'ils ne sont que partiellement observés, nous substituons ces unigrammes par la succession des unigrammes formant le seul mot-clé qu'ils peuvent constituer. La même procédure est appliquée aux bi-grammes ne pouvant former qu'un seul n-gramme strictement supérieur à 2.

1. Liste des préfixes utilisés : *acetyl, acetal, ana, anti, angio, antibiot, ante, ben, meth, eth, prop, but, pent, hex, hept, di, tri, tetra, carboxy, sulf, alca, hyper, hypo, cardio, psych, poly, pneumo, myco, meso, lymph, intra, hydro, immun, homo, endo, dys, chondro, met, micro, osteo, retro, hemangio*. Liste des suffixes utilisés : *tom, plast, scop, graph, oid, sarcom, log, om*.

2. Des vecteurs pré-entraînés ont été utilisés, disponibles sur `https://github.com/facebookresearch/fastText/blob/master/docs/pretrained-vectors.md`

1.2.2 Estimation du coefficient de pondération pour chaque mot-clé

L'objectif de cette phase est, pour chaque couple $c \in \mathcal{C}$, d'attribuer un coefficient de pondération pour chaque mot-clé de M'. A cette fin, un coefficient de pondération TF-IDF est calculé en tenant compte des n-grammes de rang 1 à 5 (cf. équation 3). Seuls les n-grammes présents dans M' sont conservés.

$$tfidf(t, c) = idf(t) \times (1 + log\ \ tf(t, c)) \tag{3}$$

où $tfidf(t, c)$ est le coefficient de pondération TF-IDF du n-gramme t pour le couple c, $tf(t, c)$ la fréquence du n-gramme t au sein du couple c et $idf(t)$ la fréquence inverse de document de t calculé selon l'équation 4.

$$idf(t) = 1 + log\frac{1 + |\mathcal{C}|}{1 + df(t)} \tag{4}$$

où $df(t)$ représente le nombre de couples $c \in \mathcal{C}$ contenant le n-gramme t.

Enfin, étant donné que certains n-grammes sont introduits par l'intermédiaire d'une substitution d'unigrammes (*e.g* « *escherichia coli* » substitue « *escherichia* »), leur fréquence est pondérée afin de marquer leur partielle observation (cf. équation 5).

$$tf(t, c) = entier(tf(t, c) \times \lambda_1) \tag{5}$$

où $entier(.)$ représente la fonction partie entière et $\lambda_1 \in [0, 1]$.

1.2.3 Post-traitement et classement des mots-clés

Dans l'objectif d'améliorer la MAP, un post-traitement est appliqué sur les coefficients de pondération TF-IDF obtenus précédemment. Dans un premier temps, les mots-clés de K, avec K mots-clés indexant les couples cas clinique/discussion dans le jeu d'entraînement, sont favorisés. Cela se traduit par l'équation 6.

$$tfidf(t, c) = tfidf(t, c) \times (1 + freq(t) \times \lambda_2) \tag{6}$$

où $freq(t)$ est la fréquence d'occurrence du terme t dans K et $\lambda_2 \in \mathbb{R}^+$.

Dans un second temps en observant le jeu d'entraînement, la tendance semble être que les n-grammes porteurs de l'information la plus spécifique sont privilégiés. Par exemple, si « tumeur » et « tumeur du rein » semblent pertinents pour indexer un couple, « tumeur du rein » est généralement favorisé. A cette fin, 3 stratégies sont appliquées de manière séquentielle sur les mots-clés à classer :

— $\forall(t_i, t_j) \in M'^2$ avec $i \neq j$, le coefficient de pondération TF-IDF w_i de t_i est incrémenté de $w_j \times \lambda_3$ pour chaque unigramme dans $t_i \cap t_j$, avec $\lambda_3 \in \mathbb{R}^+$.

— $\forall(t_i, t_j) \in M'^2$ avec $i \neq j$, si t_i est un n-gramme de rang supérieur à t_j, et $t_i \cap t_j \neq \{\varnothing\}$ et $w_i - w_j < max(w_i, w_j) \times \lambda_4$, avec $\lambda_4 \in \mathbb{R}^+$, alors $w_i := max(w_i, w_j)$ et w_j n'est plus candidat à l'indexation du couple cas clinique/discussion à traiter.

— $\forall(t_i, t_j) \in M'^2$ avec $i \neq j$, si $t_i \cap t_j \neq \{\varnothing\}$ et $w_i - w_j > max(w_i, w_j) \times \lambda_5$, avec $\lambda_5 \in \mathbb{R}^+$, alors w_j n'est plus candidat à l'indexation du couple cas clinique/discussion à traiter.

Enfin, pour chaque couple cas clinique/discussion c, la version non abstraite dans M des k_c mots-clés abstraits ayant le meilleur coefficient de pondération suite aux précédents traitements est utilisée comme indexe, où k_c représente le nombre de mots-clés attendus pour l'indexation du couple c. Cependant le processus de transformation d'un élément de M en un élément de M' n'est pas une application injective. Par exemple, « cancer de la prostate » $\in M$ et « cancer de prostate » $\in M$ correspondent au même élément normalisé « *canc prost* » $\in M'$. La stratégie de correspondance vers les versions non abstraites des mots-clés revient à utiliser l'élément de M correspondant à l'élément de M' sélectionné le plus représenté dans K. Lorsqu'il est impossible de départager les candidats, un choix par ordre alphabétique est effectué.

La MAP obtenue sur le jeu d'entraînement est de 0,42 avec $\lambda_0 = 0,6$, $\lambda_1 = 0,33$, $\lambda_2 = 110$, $\lambda_3 = 0,15$, $\lambda_4 = 0,25$ et $\lambda_5 = 0,45$. La MAP obtenue sur le jeu d'évaluation est de 0,40 avec ces mêmes paramètres. Sans limitation sur le nombre de mots-clés k_c à renvoyer pour l'indexation de chaque document c, et en ne considérant que les mots-clés dont le coefficient de pondération est non nul, la MAP obtenue est de 0,48.

2 Tâche 2 : Similarité sémantique entre les cas cliniques et les discussions

La tâche d'appariement textuel est la seconde tâche du défi DEFT 2019. Une tâche similaire avait été proposée lors du défi DEFT 2011 sur un corpus constitué de revues en Sciences Humaines et Sociales (Grouin *et al.*, 2011). Initialement, les organisateurs imaginaient une tâche associée au résumé automatique de textes mais de part les questions sous-jacentes liées à la complexité à évaluer une telle tâche (qu'est ce qui constitue un résumé de référence ? Comment évaluer la pertinence d'un résumé ?), ils l'ont transformée en une tâche d'appariement entre résumé et contenu d'article scientifique. Ils partent de l'hypothèse qu'un module de résumé textuel doit être en capacité d'évaluer le degré d'association entre le contenu d'un article et son résumé. Lors de cette précédente édition, deux phases de test avaient été réalisées. Ces phases se différenciaient sur le contenu des articles ; la première conservait la globalité de l'article et la seconde en supprimait l'introduction et la conclusion. Cette suppression part du principe qu'un même auteur aura tendance a paraphraser le résumé au travers de ces deux sections. Les résultats obtenus lors de cette tâche avaient été particulièrement bon : quatre équipes ont obtenu le score maximal lors de la première phase et deux d'entre elles réitèrent ce même score lors de la seconde phase.

2.1 Contexte de la tâche

2.1.1 Métrique d'évaluation

La métrique d'évaluation est la même que la précédente édition *i.e.* une évaluation binaire des résultats dans laquelle chaque prédiction exacte équivaut à 1 sinon à 0. Mis en formule, pour chacun des N cas cliniques r_i, le score $s(a_p(r_i), a_r(r_i))$ donné à chaque prédiction vaut 0 ou 1 selon que la discussion prédite $a_p(r_i)$ est, ou pas, la discussion de référence $a_r(r_i)$ (cf. équation 7).

$$s(a_p(r_i), a_r(r_i)) = \begin{cases} 1 & \text{si } a_p(r_i) = a_r(r_i) \\ 0 & \text{sinon.} \end{cases} \quad (7)$$

Le score global correspond à la moyenne des scores obtenus sur l'ensemble des prédictions (cf. équation 8).

$$S(p) = \frac{1}{N} \sum_{i=1}^{N} s(a_p(r_i), a_r(r_i)) = \frac{1}{N} \sum_{i=1}^{N} |r_i; a_p(r_i) = a_r(r_i)| \quad (8)$$

2.1.2 Méthodes proposées lors de DEFT 2011

Plusieurs méthodes proposées lors de DEFT 2011 ont obtenu le score maximal au moins pour la phase 1. C'est le cas par exemple de Hoareau *et al.* (2011) ayant obtenu 100% et 99,5% respectivement à la phase 1 et 2. Les auteurs proposent de modéliser chaque document au sein d'espaces sémantiques construits au travers de projections aléatoires par le biais d'une méthodologie intitulée *Random Indexing*. Ils calculent ensuite une matrice de distances en exploitant la distance euclidienne pondérée entre chaque document de l'espace sémantique. Cette matrice, modélisant un graphe à N noeuds et N^2 arcs, permet aux auteurs de construire un graphe biparti dans lequel un article est associé au résumé le plus proche. En cas d'ambiguïté, *i.e.* plusieurs articles pour un même résumé, une procédure itérative basée sur la minimisation des distances entre les différents articles concernés est mise en place. Un autre exemple d'utilisation des espaces sémantiques a été proposé dans le cadre d'une approche supervisée (Bestgen, 2011). Cette approche a obtenu 100% lors des deux phases de test. Les espaces sémantiques sont obtenus ici au travers d'une approche de réduction matricielle basée sur une SVD (*Singular Value Decomposition*) : la méthode LSA (*Latent Semantic Analysis*). L'algorithme d'apprentissage utilisé est un SVM (*Support Vector Machine*) qui tient compte des espaces sémantiques comme caractéristiques pour l'apprentissage. Le SVM exploite une approche multi-classes ainsi qu'une stratégie d'appariement par le meilleur d'abord en tenant compte des valeurs de décision de la procédure SVM comme un score de comptabilité avec chacune des catégories et donc avec chaque article. Enfin, un dernier exemple obtenant 100% et 90,9% sur la phase 1 et 2 propose une approche orientée autour des constructions lexicales (Lejeune *et al.*, 2011). Les auteurs emploient la méthode *rstr-max* permettant de détecter les chaînes de caractères répétées maximales entre les articles et les résumés. Cette méthode permet de définir une notion d'affinité caractérisée par la taille en caractères de la plus grande chaîne de caractères et le nombre total de chaînes de caractères en commun pour chaque couple potentiel. Par le biais de ces affinités, le résumé adéquat pour un article sera celui qui partagera le plus grand nombre d'affinités avec un article.

2.1.3 Spécificités du corpus de DEFT 2019

Les expérimentations que nous avons menées ont permis de détecter des propriétés potentiellement indésirables du jeu de données. En effet, ce dernier référence uniquement une seule discussion pour chacun des cas cliniques. Cependant, plusieurs discussions possèdent des copies identiques avec des identifiants différents. Par conséquent, le fait qu'un cas clinique ne référence pas toutes les copies de la discussion qui lui est rattachée génère des erreurs arbitraires lors de l'évaluation. Dans le cas présent, la mise en place d'une stratégie d'appariement (minimisation des distances, stratégie par le

meilleur d'abord, maximisation des affinités) ne permet pas de différencier l'identifiant exact parmi les différentes copies d'une même discussion par rapport à un cas clinique donné. Par conséquent, les méthodes décrites en sous-section 2.2 ne tiennent pas compte d'une stratégie d'appariement mais uniquement d'un tirage avec remise parmi l'ensemble des résumés possibles pour chaque cas clinique.

2.2 Méthodologies et résultats

Chaque approche expérimentée s'appuie sur des cas cliniques et des discussions lemmatisés par l'intermédiaire de l'outil *TreeTagger*. L'approche ayant servi de *baseline* repose sur une analyse des similarités lexicales partagées entre les textes par l'intermédiaire d'une adaptation de la similarité de Lin (Lin, 1998). Comme montré dans l'équation 9, la similarité entre un cas clinique s_1 et une discussion s_2 est calculée comme le logarithme de la probabilité de la somme pondérée des unigrammes en communs entre les deux textes, divisée par le logarithme de la probabilité de la somme pondérée de tous les mots dans les deux textes - formulation du coefficient de DICE basée sur des métriques proposées par la théorie de l'information).

$$sim(s_1, s_2) = \frac{2 \times \sum_{w \in s_1 \cap s_2} \log P(w)}{\sum_{w \in s_1} \log P(w) + \sum_{w \in s_2} \log P(w)} \tag{9}$$

Les probabilités $P(w)$ des unigrammes w sont estimées sur les données d'entraînement (Agirre *et al.*, 2016). Cette méthodologie obtient un score de 0,638.

Les deux prochaines approches ont été soumises lors de la phase de test. Elles diffèrent principalement sur la manière de représenter les cas cliniques et les discussions. La première approche exploite une méthode des *k* plus proches voisins avec une distance euclidienne (cf. tableau 2). Les cas cliniques et les discussions sont modélisés au travers d'une représentation vectorielle des valeurs de TF-IDF associées à leurs n-grammes [3].

Distances	Score
Euclidienne	**0,748**
Manhattan	0,141
Chebyshev	0,352
Hamming	0,010
Canberra	0,045
Braycurtis	0,741

TABLE 2 – Résultats issus de la méthode des *k* plus proches voisins en tenant compte de différentes distances.

Tandis que la seconde approche mise en place se base sur les espaces sémantiques des différents documents calculés à partir de la méthode LSA (cf. tableau 3) et d'une représentation vectorielle des textes qui tient compte des valeurs de TF-IDF associées aux n-grammes [4]. Ces espaces permettent ensuite de calculer une matrice de distances l^2 normalisée entre les cas cliniques et les discussions à

3. Les n-grammes considérés vont de l'unigramme au pentagramme.

4. La représentation vectorielle des textes au travers d'une valeur binaire de présence ou d'absence d'un terme a également été testée mais elle obtient un score moins performant.

partir d'une distance euclidienne. L'appariement est ensuite réalisée en minimisant la distance entre un cas clinique et les discussions.

Dimensions	Score
200	0,707
300	0,745
400	0,728
500	0,734
1000	**0,755**

TABLE 3 – Résultats en fonction du nombre de dimensions sélectionnées par l'intermédiaire de la méthode LSA.

Lors de la phase de test, l'approche obtenant le meilleur résultat est la méthode des *k* plus proches voisins utilisant la mesure euclidienne avec un score de 0,86. Toutefois, les organisateurs de la tâche ont recalculé les scores en tenant compte des doublons du corpus de test. L'actualisation du corpus a permis d'améliorer les performances de cette même approche avec un score de 0,91.

3 Conclusion

Ces travaux présente la contribution du LGI2P pour la tâche 1 et 2 du défi DEFT 2019. Ces tâches ont porté respectivement sur l'indexation et la similarité entre documents. La particularité de ce défi portait sur la nature biomédicale et clinique du jeu de données qui était constitué d'un ensemble de couples cas clinique/discussion. Les méthodologies développées pour ces deux tâches s'appuient sur des techniques provenant du domaine de la recherche d'information notamment au travers de l'utilisation du coefficient de pondération TF-IDF. Lors des phrases d'évaluation la meilleure approche pour la tâche 1 a obtenu un score 0,48, tandis que pour la tâche 2 la meilleure approche a obtenu un score de 0,91.

Références

AGIRRE E., BANEA C., CER D., DIAB M., GONZALEZ-AGIRRE A., MIHALCEA R., RIGAU G. & WIEBE J. (2016). Semeval-2016 task 1 : Semantic textual similarity, monolingual and cross-lingual evaluation. p. 497–511 : International Workshop on Semantic Evaluation.

BESTGEN Y. (2011). Lsvma : au plus deux composants pour apparier des résumés à des articles. p. 105–114 : Actes du septième DÉfi Fouille de Textes.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. volume 5, p. 135–146.

CHARTIER J.-F., FOREST D. & LACOMBE O. (2016). Alignement de deux espaces sémantiques à des fins d'indexation automatique. p. 13–19 : Actes du douxième DÉfi Fouille de Textes.

DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K. & HARSHMAN R. (1990). Indexing by latent semantic analysis. In *Journal of the American Society for Information Science*, p. 391–407.

FIORINI N., RANWEZ S., MONTMAIN J. & RANWEZ V. (2015). Usi : a fast and accurate approach for conceptual document annotation. In *BMC Bioinformatics*, p. 1471–2105.

GRABAR N., GROUIN C., HAMON T. & CLAVEAU V. (2019). : Actes du quinzième DÉfi Fouille de Textes.

GROUIN C., FOREST D., PAROUBEK P. & ZWEIGENBAUM P. (2011). Présentation et résultats du défi fouille de texte deft2011. p. 3–14 : Actes du septième DÉfi Fouille de Textes.

GUPTA V. & LEHAL G. S. (2010). A survey of text summarization extractive techniques. In *Journal of emerging technologies in web intelligence*, p. 258–268.

HAMDAN H. (2015). Sentiment analysis in social media. In *P.h.D thesis at Aix-Marseille*, p. 165.

HARISPE S., SÁNCHEZ D., RANWEZ S., JANAQI S. & MONTMAIN J. (2014). A framework for unifying ontology-based semantic similarity measures : A study in the biomedical domain. In *Journal of biomedical informatics*, p. 38–53.

HOAREAU Y. V., AHAT M., PETERMANN C. & BUI M. (2011). Couplage d'espaces sémantiques et de graphes pour le deft 2011 : une approche automatique non supervisée. p. 115 : Actes du septième DÉfi Fouille de Textes.

JONES K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. In *Journal of documentation*, p. 11–21.

KIM T.-Y., KIM J., LEE J. & LEE J.-H. (2014). A tweet summarization method based on a keyword graph. In *Conference on Ubiquitous Information Management and Communication*, p.9̃6 : ACM.

LEJEUNE G., BRIXTEL R. & GIGUET E. (2011). Deft 2011 : appariements de résumés et d'articles scientifiques fondés sur des distributions de chaînes de caractères. p. 53–64 : Actes du septième DÉfi Fouille de Textes.

LIN D. (1998). An information-theoretic definition of similarity. volume 98, p. 296–304 : Proceedings of the Fifteenth International Conference on Machine Learning.

LITVAK M. & LAST M. (2008). Graph-based keyword extraction for single-document summarization. In *Multi-source Multilingual Information Extraction and Summarization*, p. 17–24 : Association for Computational Linguistics.

MAHATA D., KURIAKOSE J., SHAH R. R. & ZIMMERMANN R. (2018). Key2vec : Automatic ranked keyphrase extraction from scientific articles using phrase embeddings. In *Association for Computational Linguistics*.

MARCHAND M., FOUQUIER G., MARCHAND E. & PITEL G. (2016). Représentation vectorielle de documents pour l'indexation de notices bibliographiques. p. 34–40 : Actes du douxième DÉfi Fouille de Textes.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, p. 3111–3119.

NALLAPATI R., ZHOU B., GULCEHRE C., DOS SANTOS C. & XIANG B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Conference on Computational Natural Language Learning*.

NANAS N., UREN V. & ROECK A. D. (2003). A comparative study of term weighting methods for information filtering. In *Knowledge Media Institue*.

WANG R., LIU W. & MCDONALD C. (2014). Corpus-independent generic keyphrase extraction using word embedding vectors. In *Software Engineering Research Conference*, volume 39.

WITTEN I. H., PAYNTER G. W., FRANK E., GUTWIN C. & NEVILL-MANNING C. G. (2005). Kea : Practical automated keyphrase extraction. In *Design and Usability of Digital Libraries : Case Studies in the Asia Pacific*, p. 129–152 : IGI Global.

YUE Y., FINLEY T., RADLINSKI F. & JOACHIMS T. (2007). A support vector method for optimizing average precision. In *SIGIR*, p. 271–278 : ACM.

ZHANG C. (2008). Automatic keyword extraction from documents using conditional random fields. In *Journal of Computational Information Systems*, p. 1169–1180.

Indexation et appariements de documents cliniques pour le Deft 2019

Davide Buscaldi[1] Dhaou Ghoul[2] Joseph Le Roux[1] Gaël Lejeune[1]
(1) LIPN UMR 7030, Université Paris XIII, 99 avenue Jean-Baptiste Clément, 93430 Villetaneuse, France
(2) STIH EA 4509, Sorbone Université, 1 rue Victor Cousin, 75005 Paris, France
(1) prenom.nom@lipn.univ-paris13.fr, (2)
prenom.nom@sorbonne-universite.fr

RÉSUMÉ

Dans cet article, nous présentons nos méthodes pour les tâches d'indexation et d'appariements du Défi Fouile de Textes (Deft) 2019. Pour la tacĥe d'indexation nous avons testé deux méthodes, une fondée sur l'appariemetn préalable des documents du jeu de tset avec les documents du jeu d'entraînement et une autre méthode fondée sur l'annotation terminologique. Ces méthodes ont malheureusement offert des résultats assez faible. Pour la tâche d'appariement, nous avons dévellopé une méthode sans apprentissage fondée sur des similarités de chaînes de caractères ainsi qu'une méthode exploitant des réseaux siamois. Là encore les résultats ont été plutôt décevant même si la méthode non supervisée atteint un score plutôt honorable pour une méthode non-supervisée : 62% .

ABSTRACT

Indexing and pairing texts of the medical domain

In this paper, we propose different methods for the indexation and pairing tasks of the 2019 edition of the french text mining challenge (Deft). For indexing texts we proposed two different methods, one exploit a pairing method developed for task 2 in order to use keywords found in similar documents.. The second method takes advantage of terminological annotation obtained via the MeSH (Medical Subject Headings) resource. For the pairing task, we also proposed two methods. The first one is unsupervised and relies on similarities at character-level, in the second one we exploited siamese networks. In both tasks our results were quite disappointing except for task 2 where the unsupervised method achieved an honourable 62% score.

MOTS-CLÉS : Appariement, Indexation, Réseaux Siamois, MESH, Modèles en Caractères.

KEYWORDS: Pariring, Indexation, Siamese Networks, MESH, Character-level Models.

1 Introduction

L'édition 2019 du Défi Fouille de Textes (Grabar *et al.*, 2019) proposait trois tâches différentes, exploitant toutes un corpus de textes du domaine médical, en l'espèce des cas cliniques et de discussions à propos de ces cas cliniques. Le développement d'outils de TAL pour ce type particulier de données textuelles est une perspective de recherche très important puisqu'il s'agit d'exploiter des données très riches et très fournies en terminologie afin d'extraire des connaissances et de faciliter

le stockage et l'échange d'information entre les praticiens. Ce sujet de recherche est très actif et intéresse différentes branches du TAL parmi lesquelles la Recherche et l'Extraction d'Information.

Nous avons travaillé sur la tâche d'indexation (Tâche 1) et sur la tâche d'appariement (Tâche 2) et nous avons laissé de côté la partie extraction d'Information (tâche 3). Pour la tâche d'indexation (Section 2), nous avons conçu une méthode endogène fondée sur la réutilisation des résultats d'un système d'appariements de textes développé pour la tâche 2. Nous présentons également une méthode exogène qui exploite la ressource MESH (*Medical Subject Headings*). Pour la tâche d'appariement de cas cliniques et de discussions (Section 3), nous avons proposé deux approches : une approche non-supervisée fondée sur des similarités de chaînes de caractères (Section 3.1) et une approche d'apprentissage profond exploitant des réseaux siamois (Section 3.2).

2 Méthodes pour la tâche d'indexation (Tâche 1)

2.1 Indexation fondée sur des appariements de documents (run1)

Notre hypothèse initiale consistait à factoriser le travail effectué sur la tâche 2. Il s'agissait d'utiliser l'appariement préalable des textes pour associer à un texte nouveau (texte du jeu de test) les mots-clés figurant dans un texte similaire et déjà indexé (présent donc dans le jeu d'entraînement). Notre hypothèse qu'un cas clinique doit être indexé peu ou prou de la même manière que la discussion à laquelle il se rapporte (et inversement pour les discussions). Indexer un cas clinique C (resp. une discussion D) du jeu de test revient à l'apparier avec une discussion D' (resp. un cas clinique C') du jeu d'entraînement et lui assigner les mots-clés correspondants. Étant donné que la tâche consistait à indexer les paires (cas clinique – discussion), nous avons exploité les mots-clés des deux appariements obtenus.

Pour chaque paire à indexer (C_i, D_i), on calcule une paire appariée (D_j, C_j). On obtient donc deux jeux de mots-clés candidats : KW_{Di} et KW_{Cj}. Dans un premier temps on conserve l'intersection : $KW_{Di} \cup KW_{Cj}$, si la taille de l'intersection est inférieure au nombre de mots-clés attendus alors on utilise l'union : $KW_{Di} \cap KW_{Cj}$. Les mots-clés ainsi obtenus sont rangés par ordre d'importance en fonction de leur longueur en caractères.

Si l'idée semblait séduisante, et permettait de factoriser le travail effectué sur la tâche 2, il s'est avéré que son efficacité était significativement inférieure à une simple *baseline* vérifiant la présence des mots-clés de la référence et les rangeant dans l'ordre inverse de leur longueur (run 2 de la tâche 1).

2.2 Indexation fondée sur l'annotation terminologique (run4)

2.2.1 Annotation hybride

Cette méthode, déjà présentée pour DEFT2016 (Buscaldi & Zargayouna, 2016) combine une annotation fondée sur le volet terminologique avec une annotation supervisée pour maximiser le rappel.

L'annotation terminologique utilise un moteur de recherche d'information. D'abord, la ressource sémantique (MeSH français) est indexée en utilisant Whoosh [1]. L'index relie chaque identifiant de

1. https ://whoosh.readthedocs.io/en/latest/

concept à ses représentations lexicales ou étiquettes (principales ou autres) qui constituent le volet terminologique de la ressource. Les fragments textuels qui constituent les étiquettes peuvent ainsi être cherchés par le moteur de recherche.

En phase d'annotation, le texte du document à annoter d est analysé pour extraire les fragments à soumettre au moteur de recherche en tant que requêtes. Annoter un document revient donc à interroger la ressource indexée en utilisant des séquences de mots extraites à partir du texte (dans notre cas, on avance trigramme de mots par trigramme de mots). Le moteur de recherche renvoie une liste ordonnée de résultats contenant les identifiants des concepts avec leurs poids. Le meilleur résultat est ajouté à l'ensemble d'annotations du documents.

Cet algorithme d'annotation est utile quand les étiquettes sont présentes dans le texte même partiellement. Le bruit peut provenir des étiquettes partageant la même racine ainsi que des concepts retournés en premier mais avec un faible score.

2.2.2 Annotation fondée sur l'Information Mutuelle

Quand les concepts n'ont pas de représentations lexicales dans le texte, les algorithmes d'annotation fondés sur la terminologie échouent. Cet algorithme se fonde sur la co-occurrence entre le concept et les termes du documents. L'idée est que si, dans un corpus de textes annotés, un ou plusieurs termes co-occurrent souvent avec le même concept c, alors si on retrouve le même terme ou la même combinaison de termes dans un document sans annotation, on peut l'annoter avec le concept c.

Nous avons d'abord procédé à quelques filtrages. Nous avons éliminé tous les cas pour lesquels les co-occurrences sont dues au pur hasard. Pour cela, dans la phase d'entraînement nous avons pris en compte seulement les concepts qui ont servi pour annoter au moins 3 documents dans le corpus de référence. Au final, nous avons calculé l'information mutuelle uniquement si le terme et le concept co-occurrent dans au moins 2 documents. Par rapport à l'édition 2016, on a réduit les deux paramètres (originellement fixés respectivement 5 et 3) à cause du nombre inférieur de documents dans le corpus d'entrainement.

L'information mutuelle ponctuelle est définie comme :

$$IP(t,c) = \log \frac{p(t,c)}{p(t)p(c)}$$

Où t est un terme (substantif ou adjectif dans notre cas) et c un concept ; $p(t,c)$ est la probabilité conjointe entre t et c, calculée comme $freq(t,c)/N$, $p(t) = freq(t)/N$ et $p(c) = freq(c)/N$, où N est le nombre de documents dans le corpus d'entraînement.

2.2.3 Construction du modèle

Nous construisons une matrice M avec $|C|$ lignes et $|T|$ colonnes, où C est l'ensemble des étiquettes dans le corpus d'entraînement avec fréquence > 5, et T est l'ensemble des mots du dictionnaire (donc tous les substantifs et adjectifs observés dans le corpus) avec fréquence > 3. Chaque élément $M_{i,j}$ est calculé comme suit :

$$M_{i,j} = \begin{cases} IP(t_i, c_j) & if \;\; IP(t_i, c_j) > 0 \\ 0 & else \end{cases}$$

Sur cette matrice, nous appliquons l'analyse sémantique latente, en décomposant M en valeurs singulières (algorithme SVD) $M = U\Sigma V^T$, et en approximant M avec $\hat{M} = U_k\Sigma_k V_k^T$, avec les meilleurs k valeurs singulières sélectionnés ($k = 100$).

Nous espérons améliorer la couverture en appliquant LSA sur la matrice. En effet, LSA permet de trouver des termes fortement associés avec d'autres termes qui sont très caractéristiques pour certaines étiquettes, même si dans le corpus d'entraînement l'étiquette n'apparaît pas avec ce mot. Par exemple, si « Paris » et « français » co-occurrent souvent, mais dans la collection on ne trouve que « France » associé au terme « français », on peut déduire que « France » est une annotation plausible si dans le texte on trouve « Paris », même si on ne les a jamais vus ensemble dans le corpus l'entraînement.

Finalement, nous appliquons un filtre sur la matrice $\hat{M}$ de la façon suivante :

$$\hat{M}^{\sigma}_{i,j} = \begin{cases} \hat{M}_{i,j} & if \;\; \hat{M}_{i,j} \geq \Big(0.5 * \sigma'(\hat{M}_{*,j}) + \mu'(\hat{M}_{*,j})\Big) \\ 0 & else \end{cases}$$

Où $\hat{M}_{*,j}$ est le $j-$ème vecteur colonne de la matrice $\hat{M}$, $\sigma'(\mathbf{x})$ est l'écart type des éléments non nuls du vecteur $\mathbf{x}$ et $\mu'(\mathbf{x})$ est la moyenne des éléments non nuls du vecteur $\mathbf{x}$. La matrice $\hat{M}^{\sigma}$ ainsi obtenue est utilisée comme *modèle* pour l'annotation des documents.

2.2.4 Annotation

Nous associons au document d à annoter un vecteur binaire $\mathbf{b}$ de taille $|T|$ (taille du vocabulaire) où chaque élément $\mathbf{b}_i$ est à 1 si le terme correspondant est dans le texte du document d, 0 autrement. Nous calculons pour chaque étiquette l_i le score $s(l_i) = \mathbf{b} \cdot \hat{M}^{\sigma}_{i,*}$. Le document est annoté finalement avec les 20 étiquettes avec les meilleurs scores > 0. S'il y a moins de 20 étiquettes avec des scores positifs, alors une annotation est générée avec toutes les étiquettes ayant un score positif. Le choix de 20 a été défini empiriquement après les tests sur le corpus de développement.

3 Méthodes pour la tâche d'appariement (Tâche 2)

3.1 T2 : Appariements fondés sur des distributions de chaînes de caractères (run1)

Afin d'apparier les cas cliniques et les discussions nous avons eu l'idée d'exploiter le phénomène de recopie, de réutilisation de séquences langagières. de manière à rendre la méthode aussi endogène et générique que possible, nous nous sommes efforcés de nous affranchir d'une approche purement lexicale. Nous avons fait l'hypothèse que les discussions pouvaient être vues comme des prolongements, des développements des cas cliniques. Ceci nous a amené à nous intéresser à une proposition faite pour le Défi Fouille de Textes en 2011 dans une autre tâche d'appariement visant cette fois à associer des résumés et des articles scientifiques. Dans cet article(Lejeune *et al.*, 2011), nous proposions de considérer que l'association entre un résumé et un article était liée à des correspondances uniques dans le corpus nommées "affinités". Une affinité y était définie comme une sous-chaîne de caractère saillante dans l'article (répétée à des positions clés) et partagée par un seul résumé du corpus.

Une chaîne de caractère était considérée comme saillante dans l'article si elle était présente dans l'introduction ou la conclusion et dans le corps de l'article.

D'un point de vue fonctionnel, l'expérience consistait à considérer que les articles étaient des célibataires et que les résumés étaient des prétendants. Chaque célibataire était présenté tour à tour à tous les prétendants et il s'agissait de calculer les sous-chaînes de caractères communes à un célibataire et à un seul prétendant. Autrement dit, il s'agit de sous-chaînes qui sont hapax dans le sous-corpus des prétendants. On cherche donc toutes les affinités entre le célibataire et chacun des prétendants (critère $Card - Aff$) et on s'intéresse aussi à l'affinité la plus longue (critère $Aff - Max$).

On apparie un prétendant P_i à un célibataire C s'il respecte les deux conditions :

— Maximiser $Card - Aff$: P_i est le prétendant avec lequel C partage le plus d'affinités. Si plusieurs prétendants partagent le même nombre d'affinités alors nous considérons que l'appariement ne peut être validé.

— Maximiser $Aff - Max$: P_i et C partagent la plus longue affinité détectée. De la même manière, l'appariement n'est effectué que si un seul prétendant partage C une affinité de taille N.

En d'autres termes, pour chaque célibataire on apparie le prétendant qui possède à la fois le plus grand nombre d'affinités ainsi que l'affinité la plus longue. Les célibataires sont présentés tout à tour mais il apparaît expérimentalement que l'ordre d'apparition des célibataires a un impact marginal voire nul sur la qualité des appariements.

Nous avons utilisé le code disponible en ligne [2] pour reproduire cette méthode. Par rapport à l'exploitation qui en a été faite en 2011 nous avons opéré deux modifications. La première est que nous n'avons pas utilisé le critère de saillance. En effet, les cas cliniques comme les discussions sont moins richement structurés que des articles scientifiques de sorte que la pertinence de la notion de saillance était moins évidente. D'autre part, ce sont des documents beaucoup plus courts et plus denses que des articles scientifiques de sorte que les phénomènes de répétition sont moins prégnants. Les textes sont très condensés et n'appartiennent pas au genre très normé qui est celui de l'article scientifique. La seconde est que dans le Deft 2011 chaque article correspondait à un seulr ésumé et vice-versa. Une fois un appariement effectué entre un célibataire et un prétendant P_i, ce prétendant était écarté pour la suite du processus. Autrement dit, il s'agissait d'un tirage sans remise. Ici, cette stratégie n'était pas possible pour deux raisons. La première c'est que la même discussion peut correspondre à plusieurs cas cliniques. Ecarter une discussion déjà appariée n'était donc pas pertinent. La deuxième raison est plus prosaïque, la tâche du Deft 2019 était singulièrement plus difficile. Là où sur le Deft 2011 la première phase d'appariement permettait d'obtenir 80% de rappel avec une précision proche de 100%, sur le défi de cette année le rappel était de l'ordre de 20% avec une précision de 80%. Les premiers appariements étant moins nombreux et moins sûrs, le tirage avec remise était donc plus adapté.

3.2 T2 : Réseau siamois et algorithme hongrois (run2 et run3)

Les réseaux siamois (Bromley *et al.*, 1993) sont des architectures neuronales spécialisées pour l'appariement de structures similaires [3]. Ils sont composés de deux sous-réseaux identiques qui permettent de transformer deux vecteurs d'entrée, représentant les documents à évaluer, vers un espace de caractéristiques commun. Une dernière couche prend en entrée ces deux vecteurs de caractéristiques et calcule une énergie, censée représenter la proximité entre les deux structures.

2. `https://github.com/rundimeco/deft2011`
3. Ils ont été développés pour la reconnaissance automatique de signatures de chèques.

Plus concrètement, notre seconde méthode fonctionne comme suit.

1. Les documents sont filtrés via SpaCY pour ne garder que les noms communs [4]. À chaque document, on attribue comme vecteur représentatif la moyenne des vecteurs associés à ses mots après filtrage. Les méthodes par réseaux récurrents n'ont rien apporté.
2. Pour quantifier la proximité entre deux documentsd d_1 et d_2, on donne leur vecteur représentatif comme entrée au sous-réseau du réseau siamois, implémentée comme un perceptron à une couche cachée d'activation par rectificateur linéaire (ReLU). On obtient alors en sortie deux vecteurs v_1 et v_2
3. À la prédiction, on utilise comme énergie simplement la distance euclidienne entre v_1 et v_2.
4. Pour prédire globalement les appariements d'un ensemble de documents, on réduit le problème à trouver le couplage parfait de poids minimal dans un graphe bi-partite complet $G = (V_1 \cup V_2, V_1 \times V_2)$ où les éléments de V_1 et V_2 représentent respectivement les cas cliniques et les discussions et où les poids des arcs (v_1, v_2) sont données par la distance euclidienne. On utilise pour cela l'algorithme de (Munkres, 1957), aussi appelé algorithme *hongrois*, de complexité en temps $O(n^3)$ où n est le nombre de sommets.

Nous suivons la méthode de (Chopra *et al.*, 2005) avec à l'apprentissage une énergie dite *contrastive* qui permet de diminuer l'énergie pour les paires de structures similaires et de l'augmenter pour les paires de structures dissimilaires. Pour tous les triplets (d_1, d_2, l) constitués de documents d_1, d_2 (cas cliniques ou discussions) et d'un label $l \in \{0, 1\}$ indiquant si les documents sont similaires ou non. Deux documents sont similaires (label 0) si : ils sont identiques, ils sont une paire cas clinique/discussion de références (ou contiennent le même texte), sont deux cas cliniques associés à la même discussion. Dans tous les autres cas, le label est 1. On peut ainsi définir la perte constrastive : $L(d_1, d_2, l) = (1 - l)\|v(d_1) - v(d_2)\|_2^2 + l\frac{1}{2}\max(0, m - \|v(d_1) - v(d_2)\|_2)^2$. Cette fonction est parfaitement minimisée, et donc égale à zéro pour toute paire, si la distance euclidienne des paires similaires est nulle et si la distance entre des documents non similaires est au moins m.

4 Résultats et discussion

4.1 Tâche 1 : Indexation

Run	MAP	R-Precision
Run1 (Appariements)	0.126	0.122
Run2 (Baseline)	0.220	0.240
Run4 (MeSH)	0.044	0.034

TABLE 1 – Résultats officiels sur la tâche 1

Nous présentons dans le tableau 1 les résultats que nous avons obtenu sur la tâche d'indexation. Les deux hypothèses que nous avions formulé à savoir l'utilisation de l'appariement de documents (Run1) et l'exploitation du MESH (Run4) se sont avérées inadaptées. En effet, ces méthodes se situent très nettement en retrait de la *baseline* que nous avions développé.

4. Nous avons essayé sans filtrage, ou avec d'autres parties du discours, mais avec de moins bons résultats.

4.2 Tâche 2 : Appariements

Pour la tâche 2, nous avons également obtenu des résultats plutôt décevants (Table 2). Nos méthodes fondées sur les réseaux siamois ont assez vite plafonné, que ce soit pour la variante *average* qui utilise un modèle moyenné sur les différentes itérations et le jeu de développement ou pour la variante *single* qui correspond simplement au modèle obtenant la meilleur évaluation sur le jeu de développement. Notre système fondé sur la similarité de chaînes de caractères s'est avéré plus satisfaisant, avec l'avantage de ne pas subir le phénomène de sur-apprentissage.

Run	Précision
Run1 (Similarité en caractères)	0.617
Run2 (Réseau Siamois *average*)	0.107
Run3 (Réseau Siamois *single*)	0.126

TABLE 2 – Résultats officiels sur la tâche 2

4.3 Discussion

Il est toujours ardu de constater que l'on s'est appuyé sur des hypothèses inadaptées, en particulier lorsque la différence de résultat est aussi grande. Sur la tâche 1, nous sommes bons derniers avec 16 points de pourcentage de moins que la moyenne et 18 de moins que la médiane. C'est peut être sur cette tâche que nous aurions pu améliorer nos résultats. En effet, sur la tâche 2 notre méthode sans apprentissage est encore plus loin de la moyenne (19 points) et de la médiane (25 points). Toutefois, il est difficile d'imaginer comment nous pourrions l'améliorer sans en dénaturer l'esprit.

Références

BROMLEY J., BENTZ J. W., BOTTOU L., GUYON I., LECUN Y., MOORE C., SÄACKINGER E. & SHAH R. (1993). Signature verification using a siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, **7**(04), 669–688.

BUSCALDI D. & ZARGAYOUNA H. (2016). LIPN@DEFT2016 : Annotation de documents en utilisant l'Information Mutuelle. In *DÉfi Fouille de Texte 2016 – DEFT2016*, Paris, France.

CHOPRA S., HADSELL R., LECUN Y. *et al.* (2005). Learning a similarity metric discriminatively, with application to face verification. In *CVPR (1)*, p. 539–546.

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 de DEFT*, Toulouse, France.

LEJEUNE G., BRIXTEL R., GIGUET E. & LUCAS N. (2011). Deft 2011 : appariements de résumés et d'articles scientfiques fondés sur des distributions de chaînes de caractères. In *Proceedings of DEfi Fouille de Texte (DEFT'11)*, p. 53–64.

MUNKRES J. (1957). Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, **5**(1), 32–38.

DeFT 2019 : Auto-encodeurs, *Gradient Boosting* et combinaisons de modèles pour l'identification automatique de mots-clés. Participation de l'équipe TALN du LS2N

Mérième Bouhandi Florian Boudin Ygor Gallina
LS2N, Université de Nantes
prénom.nom@univ-nantes.fr

RÉSUMÉ

Nous présentons dans cet article la participation de l'équipe TALN du LS2N à la tâche d'indexation de cas cliniques (tâche 1). Nous proposons deux systèmes permettant d'identifier, dans la liste de mots-clés fournie, les mots-clés correspondant à un couple cas clinique/discussion, ainsi qu'un classifieur entraîné sur la combinaison des sorties des deux systèmes. Nous présenterons dans le détail les descripteurs utilisés pour représenter les mots-clés ainsi que leur impact sur nos systèmes de classification.

ABSTRACT

Autoencoders, gradient boosting and ensemble systems for automatic keyphrase assignment : The LS2N team participation's in the 2019 edition of DeFT

In this article, we present the participation of the TALN team at the LS2N in the clinical case indexing task (task 1). We propose two systems to identify for each clinical case/discussion pair its corresponding keywords in a given thesaurus, as well as a classifier trained on the the two systems outputs combination. We will present in detail the features used to represent the keywords and their impact on the given task.

MOTS-CLÉS : Identification automatique de mots-clés, autoencoders, gradient boosting, TAL.

KEYWORDS: Automatic keyword assignment, autoencoders, gradient boosting, NLP.

1 Introduction

Dans cet article, nous présentons nos travaux réalisés dans le cadre de l'édition 2019 du Défi Fouille de Texte (DeFT) (Grabar *et al.*, 2019). Portant sur l'analyse de cas cliniques rédigés en français, cette édition se compose de trois tâches autour de la recherche et de l'extraction d'information. Nous avons choisi de participer à la tâche d'indexation des cas cliniques (tâche 1) qui consiste à retrouver les mots-clés les plus pertinents, pour une paire de cas clinique/discussion donnée, dans une liste de mots-clés fournie.

Dans un premier temps (§2), nous détaillons l'ensemble des descripteurs utilisés pour représenter les mots-clés, ainsi que les différents modèles utilisés pour les identifier automatiquement, puis quelques expériences exploratoires. Dans un second temps (§3), nous présentons les résultats obtenus. Finalement (§4), nous concluons et discutons des perspectives de travaux futurs.

2 Approches

Le corpus mis à disposition cette année dans le cadre de la compétition DeFT est composé de cas cliniques associés à une discussion et des mots-clés (Grabar *et al.*, 2018). Les documents proposés sont liés à différentes spécialités médicales (cardiologie, urologie, oncologie, obstétrique, pneumologie, gastro-entérologie, etc.) et ont été tirés de publications francophones (France, Belgique, Suisse, Canada, pays africains, etc.).

Un cas clinique est tiré du dossier médical d'un patient et correspond à la description des symptômes, des diagnostics et propositions thérapeutiques d'un médecin. Les discussions médicales couvrent un sujet particulier de façon plus complète et sont généralement plus longues (cf. Figure 1.

Cas
Un échantillon de sérum a été prélevé (puis congelé) 30 minutes après son admission pour une demande de recherche d'amphétamines et d'acide gamma hydroxybutyrique, et une mèche de cheveux ...

Discussion
Par ailleurs, à la suite de prises croissantes de GHB sur une période de 28 jours (30, 45, 45 et 60 mg/kg de poids corporel) chez un volontaire, l'analyse des cheveux prélevés a présenté les résultats suivants : les segments témoins présentaient des concentrations moyennes de 0,62 ng/mg et ...

Mots-clés de réference :
analyse de cheveux ; acide gamma hydroxybutyrique ; intoxication sanguine.

FIGURE 1 – Exemple de couple cas clinique / discussion et mots-clés associés.

Les mots-clés associés sont variés. On y trouve des termes simples (*"cancer"*, *"prostate"*) ou des termes complexes, aussi bien morphologiquement complexes (*"urétéroscopie"*) que polylexicaux (*"syndrome de la fente médiane"*).

Les mots-clés ne sont pas toujours composés que de caractères alphanumériques. On y retrouve, par exemple, des mots contenants des caractères spéciaux (*"fighter®"*) ou plusieurs termes séparés par des virgules *"atropine, scopolamine, hyoscyamine, hallucinogène"*. Certains mots-clés sont rpésents sous différentes variantes morphologiques d'un mots-clés peuvent être utilisés (*"analyse de cheveux"*, *"analyse des cheveux"*). De plus, un mot-clé n'est pas toujours présent dans les deux documents cas/discussion et il n'est parfois présent dans aucun des deux.

2.1 Pré-traitements effectués

Certains mots étant coupés en deux, *"lésion"* devenant *"lési on"* ou *"chimiothérapie"* devenant *"chi miothérapie"* par exemple, nous avons appliqué une normalisation du corpus en utilisant le vocabulaire présent dans le corpus ainsi que les mots de la liste de mots-clés fournie : pour deux mots qui se suivent dans le corpus, si la concaténation des deux est présente dans le vocabulaire issu du corpus ou la liste de mots-clés fournie, alors ils sont fusionnés. De plus, nous avons observé que ces césures apparaissent généralement lorsque le premier mot (*"lési"* dans l'exemple) se termine par une des

lettres de la suite *"ïïǐkltv"*. Ce problème est probablement dû à une mauvaise gestion des césures dans les documents originaux.

Aussi, nous avons fait le choix de ne pas pré-traiter le texte, en lui appliquant des normalisations telles la suppression des mots vides ou une lemmatisation, pour éviter de perdre de l'information.

2.2 Descripteurs

Nous avons essayé de caractériser les particularités linguistiques et stylistiques qui caractérisent les mots-clés au moyen des descripteurs suivants :

— **Descripteurs fréquentiels**

1. Fréquence du terme (TF) : fréquence d'un terme dans la paire de documents
2. Fréquence inverse de document (IDF) : logarithme de l'inverse de la proportion de documents du corpus qui contiennent le terme
3. $TF - IDF$

— **Descripteurs positionnels**

1. Position de la première occurrence (Occ_{first}) : position (en nombre de caractères) de la première occurrence du mot dans la paire de documents. Si le mot est absent de la paire de documents, $Occ_{first} = -1$
2. Mesure d'étalement ($Spread$) : Nombre de caractères entre la première et la dernière occurrence dans la paire de documents.

— **Descripteurs statistiques**

1. Présence dans les deux documents (P_{both}) : vrai si le mot est présent dans les deux documents de la paire cas/discussion, faux sinon
2. Mesure du lien du mot et son contexte (W_{rel}) : mesure de la singularité du mot dans le corpus. Plus un mot candidat co-occure avec des termes différents, plus ce mot candidat est susceptible d'être peu important dans le document. On le calcule comme suit :

$$W_{rel} = (0.5 + ((WL \cdot \frac{TF(w)}{MaxTF}) + PL)) + (0.5 + ((WR \cdot \frac{TF(w)}{MaxTF}) + PR))$$

   avec $TF(w)$ la fréquence du terme dans la paire de document, $MaxTF$ la fréquence du terme le plus fréquent, WL [ou WR] le rapport entre le nombre de mots différents qui co-occurrent avec le mot candidat à gauche (ou à droite) et le nombre de mots total qui co-occurrent avec celui-ci et PL (ou PR) mesure le rapport entre le nombre de mots différents qui co-occurrent avec le terme candidat à gauche (ou à droite) et le $MaxTF$.
3. Occurrence de sous-parties ($Occ_{subparts}$) : somme de la fréquence de chaque mot composant le terme
4. Occurrence de variantes ($Occ_{variants}$) : compte des variantes du terme dans la paire de documents. Pour *"analyse des cheveux"*, nous prenons aussi en compte *"analyses des cheveux"* et *"analyse de cheveux"*, par exemple.
5. Longueur de la paire de documents normalisée (D_{len}) : somme de la longueur de la paire de document

Descripteur	**Références**
Fréquence du terme (TF)	(Jones, 2004)
Fréquence inverse de document (IDF)	(Jones, 2004)
$TF - IDF$	(Jones, 2004)
Position de la première occurrence (Occ_{first})	(Aquino *et al.*, 2014)
Mesure d'étalement ($Spread$)	(Hasan & Ng, 2014)
Présence dans les deux documents (P_{both})	-
Mesure du lien du mot et son contexte (W_{rel})	(Campos *et al.*, 2018)
Occurrence de sous-parties ($Occ_{subparts}$)	-
Occurrence de variantes ($Occ_{variants}$)	(Claveau & Raymond, 2012)
Longueur du document normalisée (D_{len})	-
Z-Score du mot (W_z)	(Aquino *et al.*, 2014)

TABLE 1 – Récapitulatif des descripteurs utilisés pour caractériser les mots-clés.

6. Z-Score du mot (W_z) : fréquence normalisée du terme en utilisant sa fréquence moyenne dans le corpus et son écart-type

Ainsi, pour chaque paire de cas/discussion, nous représentons chaque mot de la liste de mots-clés fournie en utilisant ces descripteurs. Ils sont normalisés en utilisant le *StandardScaler* de la librarie python *scikit-learn*, qui transforme une valeur x en une valeur $z = \frac{x-u}{s}$, avec u la moyenne des x et s son écart-type.

Ces descripteurs sont plus ou moins importants pour caractériser les mots-clés (Figure 2). Après plusieurs essais, nous avons empiriquement déterminé que seuls les éléments qui corrèlent à plus de 0.125 avec notre classe cible sont conservés.

Ces descripteurs (Table 2) seront utilisés en entrée de tous nos systèmes sauf celui présenté en §2.3.

Descripteur	**Corrélation avec la cible**
$TF - IDF$	0,44
Fréquence du terme (TF)	0,36
Fréquence inverse de document (IDF)	0,32
Présence dans les deux documents (P_{both})	0,28
Mesure d'étalement ($Spread$)	0,21
Mesure du lien du mot et son contexte (W_{rel})	0,21
Position de la première occurrence (Occ_{first})	0,13

TABLE 2 – Récapitulatif des descripteurs utilisés pour caractériser les mots-clés.

2.3 Vecteurs multi-hot & TF-IDF

Ce système est basé sur la comparaison entre la paire cas/clinique et chaque mot-clés fournis. Les paires cas/discussions sont représentés par leur vecteur Tf-Idf et les mots-clés par un vecteur multi-hot dans le même espace vectoriel. La mesure cosinus est alors utilisée pour calculer la similarité ,et ainsi la pertinence, du mot-clé par rapport à la paire de document.

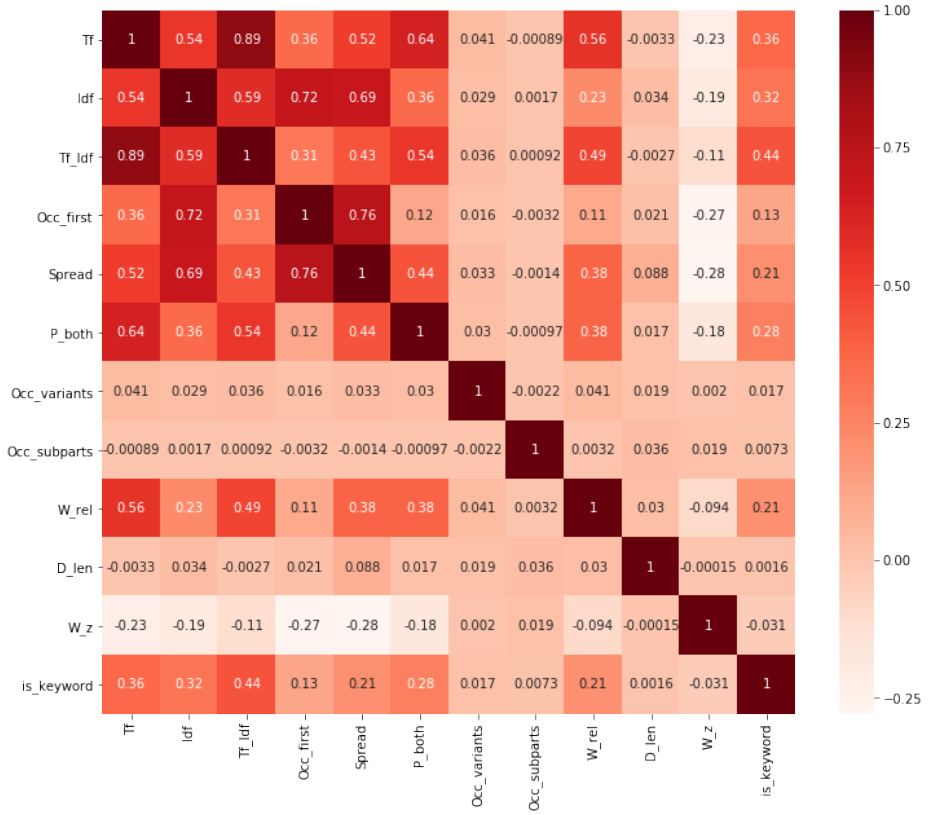


FIGURE 2 – Corrélation des différents descripteurs et de la cible (*est/n'est pas un mot-clé*).

Nous montrons dans la Table 3 comment sont construits ces vecteurs pour les mots-clés et pour les paires cas/discussions.

	analyse des cheveux	*acide gamma hydroxybutyrique*	*intoxication sanguine*	*cas + discussion*
	Compte			TFIDF
acide	0	1	0	0,0147
analyse	1	0	0	0,0147
cheveux	1	0	0	0,0294
gamma	0	1	0	0,0147
hydroxybutyrique	0	1	0	0,0147
intoxication	0	0	1	0
sanguine	0	0	1	0

TABLE 3 – Représentations vectorielles des mots de la liste de mots-clés fournie ainsi que des paires cas/discussions.

2.4 Gradient Boosting

Le *boosting* est une méthode d'apprentissage ensembliste qui consiste apprendre itérativement plusieurs classifieurs dont les poids des individus sont corrigés au fur et à mesure pour mieux prédire les valeurs difficiles. Les classifieurs sont alors pondérés selon leurs performances et agrégés itérativement.

Nous utilisons le modèle *XGBoost* (*eXtreme Gradient Boosting*) qui est une implémentation très populaire, notamment lors des compétitions *Kaggle*, du modèle *Gradient Boosting*. La principale différence entre *boosting* classique (*AdaBoost*) et le *Gradient Boosting* se trouve au niveau de la fonction de coût : ce dernier utilise des gradients dans sa fonction de coût alors que le premier se contente d'appliquer des poids plus importants aux individus mal classifiés.

Nous prenons en entrée les descripteurs présentés dans la Table 1, labellisés. Le classifieur nous renvoie en sortie la probabilité que le mot soit effectivement mot-clé de la paire de documents donnée. Les paramètres utilisés sont présentés dans la Table 4.

Paramètres	**Valeurs**
alpha	10
colsample_bytree	0.3
early_stopping_rounds	10
learning_rate	0.1
max_depth	5
metrics	rmse
nfold	3
num_boost_round	50
objective	reg :linear
seed	123

TABLE 4 – Paramètres utilisés pour l'entraînement du classifieur XGBoost.

2.5 Auto-encodeur

La deuxième stratégie mise en place est l'utilisation d'un *auto-associative neural network* ou auto-encodeur. Plutôt que de classifier, l'objectif de l'auto-encodeur est de reconstruire en sortie l'ensemble de données d'entrée.

Ainsi, étant donné un ensemble d'entrée X et un ensemble de sortie X', on peut mesurer l'erreur de reconstruction commise par l'auto-encodeur en calculant la somme des différences au carré :

$$e(X) = \sum_{i=1}^{n} (X_i - X'_i)^2$$

L'intuition derrière l'utilisation de cette méthode pour l'identification automatique des mots-clés est la suivante : nous n'avons pas suffisamment de données pour entraîner un système à associer à chaque document l'ensemble de mots-clés correspondants. De plus, les classes (*est/n'est pas un mot-clé*) sont fortement non balancées. En effet, pour chaque paire de documents, sur les 1311 mots

de la liste de mots-clés fournie, seuls 4 mots sont en moyenne associés à chaque document et sur 38k combinaisons liste mots-clés/documents, seules 765 correspondent à des mots-clés. Il est donc plus facile d'apprendre à reconnaître un mot qui n'est pas un mot-clé, plutôt qu'un terme l'étant. En traitant cette tâche comme une tâche de détection d'évènements rares, l'erreur de reconstruction sur les mots-clés sera particulièrement élevée par rapport à celle de mots qui ne le sont pas.

L'erreur de reconstruction est alors utilisée comme mesure pour classer les mots-clés par ordre de pertinence : plus cette erreur est élevée, plus important est le mot-clé. Nous récupérons alors les N premiers pour chaque paire de documents, avec N le nombre de mots-clés attendu.

Les paramètres utilisés sont récapitulés dans la Table 5.

Paramètres	**Valeurs**
batch_size	128
epoch	100
loss	mean squared error
optimizer	adam
learning_rate	0.001
encoder (input) $\rightarrow$ *activation*	tanh
encoder (hidden) $\rightarrow$ *activation*	relu
decoder (hidden) $\rightarrow$ *activation*	tanh
decoder (output) $\rightarrow$ *activation*	relu
EarlyStopping $\rightarrow$ *monitor*	val_loss
EarlyStopping $\rightarrow$ *patience*	10

TABLE 5 – Paramètres utilisés pour l'entraînement de l'auto-encodeur (utilisation de la librairie *Keras*).

2.6 Combinaison des systèmes

Dans l'optique de tirer partie des résultats obtenus avec les deux modèles, nous proposons deux approches pour les combiner.

2.6.1 Combinaison des scores

La première approche se fonde sur le calcul d'un score moyen pour un mot m de vecteur x en combinant la sortie $p(c|x)$ de XGBoost et l'erreur de reconstruction $e(x)$ de l'auto-encodeur au moyen d'une moyenne harmonique, pour éviter de sur-estimer le score si l'un des deux est significativement plus élevé que l'autre :

$$\overline{H}(x,c) = \frac{2 \cdot p(c|x) \cdot e(x)}{p(c|x) + e(x)}$$

2.6.2 Stacking

La seconde approche – *stacking* – est une technique d'apprentissage ensembliste permettant de combiner plusieurs modèles de classification entraînés chacun sur l'ensemble des données d'apprentissage. Les sorties de ces modèles sont alors fournies en données d'entrée d'un méta-classificateur pour prédire un résultat final. Nous utilisons pour cela XGBoost qui prend en entrée la sortie $p(c|x)$ du précédent XGBoost sur l'intégralité des données ainsi l'erreur de reconstruction $e(x)$ de l'auto-encodeur.

3 Résultats

Notre travail a donné lieu à de nombreuses expérimentations, notamment plusieurs combinaisons et variantes de nos systèmes. Nous présentons ici les systèmes les plus performants. Les mesures d'évaluation utilisées sont la MAP et la P@N (précision rang N, avec N le nombre de mots-clés attendus).

3.1 Résultats tâche 1

Nous reportons Table 6 les résultats obtenus pour nos différents systèmes sur le corpus d'apprentissage et Table 7 ceux obtenus pour les 3 runs sur le corpus de test. Pour la classification avec XGBoost, nous avons choisi une validation croisée en 3 *folds*. Pour les autres, nous avons découpés le corpus d'apprentissage en 80% pour le *train* et 20% pour le *test*.

Nous remarquons que la majorité des systèmes peinent à atteindre les 50% de MAP et de précision. Le modèle basé sur XGBoost donne les meilleurs résultats seul et gagne 3 points en étant associé avec l'auto-encodeur. Seul, ce dernier est assez médiocre, ce qui est lié à la petite taille du corpus d'apprentissage (Figure 3).

Modèle	**MAP**	**P@N**
TF-IDF	16,1	20,5
Auto-Encodeur (AE)	22,6	29,2
XGBoost (XGB)	**43,4**	**45,1**
TF-IDF + AE (moyenne harmonique)	**30,9**	**35,3**
TF-IDF + AE (XGB sur sorties)	27,2	**35,1**
AE + XGB (moyenne harmonique)	**45,3**	**48,7**
AE + XGB (XGB sur sorties)	44,8	**48,4**

TABLE 6 – Résultats sur le corpus d'apprentissage.

4 Conclusion

Dans cet article, nous avons présentés nos travaux réalisés dans le cadre de l'édition 2019 du Défi Fouille de Texte (DeFT) pour la tâche 1, qui consistait à retrouver les mots-clés les plus pertinents,

Modèle	Run	MAP	P@N
Auto-Encodeur (AE)	1	23,2	28,3
XGBoost (XGB)	2	**40,4**	46,0
AE + XGB (moyenne harmonique)	3	**40,4**	**46,7**

TABLE 7 – Résultats sur le corpus de test.

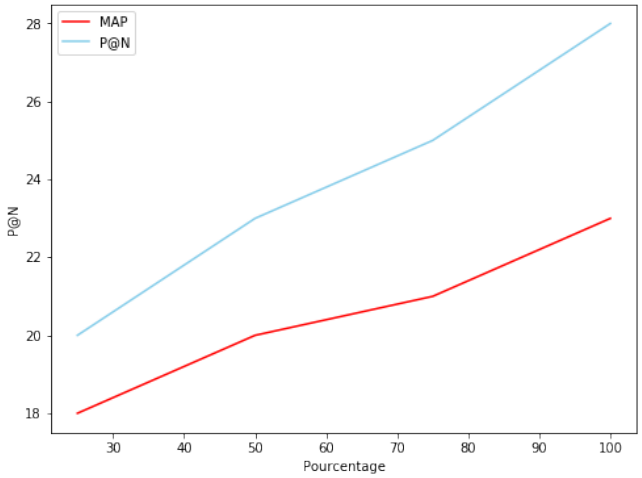


FIGURE 3 – MAP et Précision rang N pour différentes tailles de corpus d'apprentissage pour l'auto-encodeur)

pour une paire de cas clinique/discussion donnée, dans une liste de mots-clés fournie.

Nous avons obtenu des résultats moyens en utilisant des méthodes classiques telles les méthodes à base de *boosting* et d'arbres de décision, ce qui nous laisse une nette marge de progression. Les méthodes neuronales ont elles démontrés de moins bons résultats, en partie dus à la taille du corpus qui ne permettaient pas un apprentissage optimal. Nous restons cependant, avec notre meilleur système, dans la moyenne des résultats, puisque sur 6 participants, nous nous situons au dessus de la moyenne (38,5%) et de la médiane (40,1%) avec notre système *AE-XGBoost* (40,4%).

Finalement, nous avons montré qu'une combinaison de systèmes n'apportait finalement qu'une très légère amélioration.

Références

AQUINO G., HASPERUÉ W. & LANZARINI L. (2014). Keyword extraction using auto-associative neural networks.

CAMPOS R., MANGARAVITE V., PASQUALI A., JORGE A. M., NUNES C. & JATOWT A. (2018). A text feature based automatic keyword extraction method for single documents. In G. PASI, B. PIWOWARSKI, L. AZZOPARDI & A. HANBURY, Eds., *Advances in Information Retrieval*, p. 684–691, Cham : Springer International Publishing.

CLAVEAU V. & RAYMOND C. (2012). Participation de l'IRISA à DeFT2012 : recherche d'information et apprentissage pour la génération de mots-clés. In *JEP-TALN-RECITAL 2012, Workshop*

DEFT 2012 : DÉfi Fouille de Textes (DEFT 2012 Workshop : Text Mining Challenge), p. 49–60, Grenoble, France, France : ATALA/AFCP.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French Corpus with Clinical Cases. In *LOUHI 2018 - The Ninth International Workshop on Health Text Mining and Information Analysis*, Ninth International Workshop on Health Text Mining and Information Analysis (LOUHI) Proceedings of the Workshop, p. 1–7, Bruxelles, France.

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 de TALN 2019 (Traitement automatique des langues naturelles)*, ateliers DEFT 2018.

HASAN K. S. & NG V. (2014). Automatic keyphrase extraction : A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1262–1273, Baltimore, Maryland : Association for Computational Linguistics.

JONES K. S. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, **60**(5), 493–502.

Qwant Research @DEFT 2019 : appariement de documents et extraction d'informations à partir de cas cliniques

Estelle Maudet, Oralie Cattan, Maureen de Seyssel, Christophe Servan
QWANT RESEARCH, 7 Rue Spontini, 75116 Paris, France
initiale.nom@qwant.com

RÉSUMÉ

Dans ce papier, nous présentons la participation de Qwant Research aux tâches 2 et 3 de l'édition 2019 du défi fouille de textes (DEFT) portant sur l'analyse de documents cliniques rédigés en français. La tâche 2 est une tâche de similarité sémantique qui demande d'apparier cas cliniques et discussions médicales. Pour résoudre cette tâche, nous proposons une approche reposant sur des modèles de langue et évaluons l'impact de différents pré-traitements et de différentes techniques d'appariement sur les résultats. Pour la tâche 3, nous avons développé un système d'extraction d'information qui produit des résultats encourageants en termes de précision. Nous avons expérimenté deux approches différentes, l'une se fondant exclusivement sur l'utilisation de réseaux de neurones pour traiter la tâche, l'autre reposant sur l'exploitation des informations linguistiques issues d'une analyse syntaxique.

ABSTRACT

Document matching and information retrieval using clinical cases.

This paper reports on Qwant Research contribution to tasks 2 and 3 of the DEFT 2019's challenge, focusing on French clinical cases analysis. Task 2 is a task on semantic similarity between clinical cases and discussions. For this task, we propose an approach based on language models and evaluate the impact on the results of different preprocessings and matching techniques. For task 3, we have developed an information extraction system yielding very encouraging results accuracy-wise. We have experimented two different approaches, one based on the exclusive use of neural networks, the other based on a linguistic analysis.

MOTS-CLÉS : Similarité sémantique, extraction d'information, modèle de langues, modèle de vraisemblance de la requête, réseaux de neurones, analyse syntaxique.

KEYWORDS: Semantic similarity, information extraction, language model, query likelihood model, neural network, syntactic analysis.

1 Introduction

L'analyse et l'extraction d'informations pertinentes au sein d'un corpus médical est une tâche qui peut se montrer particulièrement difficile en raison de l'extrême spécificité du domaine. L'édition 2019 du défi fouille de texte (DEFT) porte sur cette problématique (Grabar *et al.*, 2019), et met à disposition un corpus de cas cliniques français, eux-mêmes issus du corpus CAS (Grabar *et al.*, 2018).

Nos motivations pour participer cette année au DEFT sont multiples. L'accent mis sur l'aspect médical de l'édition 2019 est particulièrement stimulant, du fait de l'impact médical d'éventuelles avancées dans le domaine. De plus, la spécificité du domaine, ses problématiques d'accès et la taille

restreinte des ressources associées en font un défi particulièrement intéressant. Il s'agit de notre première participation au défi. C'est pour nous une opportunité de nous confronter à d'autres équipes réfléchissant à des problématiques similaires. Enfin, il nous tient à coeur de pouvoir contribuer à la recherche dans le domaine du traitement automatique des langues en France.

Nous avons participé à deux des trois tâches proposées dans le cadre de la campagne cette année. Nous détaillons dans la section 2 notre contribution à la tâche 2 fondé sur l'appariement des cas cliniques et des discussions. La section 3 décrit les méthodes employées dans le cadre de la tâche 3 dont le but est d'extraire des informations des cas cliniques.

2 Tâche 2 : Mise en correspondance des cas cliniques et discussions par vraisemblance de la requête

L'objectif de la tâche 2 est de faire un appariement entre un cas clinique et la discussion correspondante. Le corpus d'entraînement contient 290 discussions et 290 cas cliniques. Tandis que chaque cas clinique est unique, plusieurs discussions peuvent être identiques. Le corpus de test contient 214 discussions et cas cliniques, présentant les mêmes caractéristiques que le corpus d'entraînement.

2.1 Approches

L'approche utilisée dans le cadre de la tâche 2 est la même que celle proposée par Ponte & Croft (1998) pour le calcul de la similarité basé sur des modèles de langue.

L'idée principale est de générer un modèle de langue par discussion et de mesurer leur proximité avec chacun des cas cliniques, le plus proche étant celui qui sera apparié. La mesure utilisée est la perplexité, calculée suivant l'équation 1 :

$$PPL = \hat{P}(w_1, ..., w_m)^{-\frac{1}{m}} \quad (1)$$

L'approche étant basée sur la forme de surface des mots, nous avons appliqué différents pré-traitements et étudié l'impact de ces derniers sur les résultats. Nous avons également étudié l'effet de différentes méthodes d'appariement de données.

2.2 Pré-traitement des données

Pour tenir compte des particularités linguistiques du domaine, nous avons appliqué un certain nombre de pré-traitements que nous détaillons dans cette partie.

Le texte est dans un premier temps converti en minuscules et tokenisé à l'aide de notre outil interne Qnlp-toolkit [1]. Différents autres traitements ont été appliqués en fonction des expérimentations : racinisation, suppression de mots vides et désabrègement.

Racinisation Aussi appelée dé-suffixation, la racinisation permet de regrouper l'ensemble des déclinaisons autour d'une même racine. Elle a été réalisée suivant l'algorithme Snowball. (Porter, 2001).

1. https ://github.com/QwantResearch/qnlp-toolkit

Suppression des mots vides Les mots très courants (souvent appelés « mots vides »), tels que « le », « et » ou « de », sont généralement ignorés dans les recherches. Ils ne contiennent habituellement pas autant d'information que les autres mots recherchés, qui eux permettent l'appariement des cas et discussions. Une situation dans laquelle conserver les mots vides pourrait être important est la correspondance de l'information stylistique des cas et discussions (si les paires de cas et discussions avaient systématiquement été rédigées par la même personne).

Désabrègement Compte tenu de l'abondance des sigles, acronymes, symboles et autres abrègements rencontrés dans les cas cliniques et dans les discussions, il nous a semblé intéressant de procéder à leur désabrègement automatique. Dans le domaine médical, l'abrègement est un procédé de construction lexicale très utilisé pour des raisons mnémotechniques et d'économie du langage qui permet de réduire les longues compositions de mots (souvent savants). Il arrive fréquemment que ces abrègements soient définis et repris tout au long du texte. Afin de lever toute ambiguïté lexicale en écartant au maximum les cas d'abrègements polysémiques (e.g. IVG peut à la fois être utilisé pour signifier « insuffisance ventriculaire gauche » ou « interruption volontaire de grossesse »), nous avons constitué un lexique des formes étendues, élaboré à partir des sigles, acronymes et de leurs définitions rencontrés en contextes à partir de l'ensemble du corpus. La mise en correspondance de la forme développée et de son abrègement a été réalisée selon plusieurs règles de recherche et nous avons complété notre lexique initial par des ressources terminologiques recensant les abrègements reconnus par la communauté. Ainsi, à partir du corpus comprenant les cas cliniques et les discussions s'y rapportant, 227 abrègements ont été relevés, ce qui a engendré un nombre de substitutions égal à 9016.

2.3 Apprentissage des modèles de langue

Les différents modèles de langue (LM) ont été appris à l'aide de l'outil *SRILM* (Stolcke, 2002), selon trois ordres différents : unigrammes, bigrammes et trigrammes. Afin de permettre une comparaison entre les différentes perplexités (voir Section 2.4), tous les modèles ont été créés utilisant un vocabulaire commun, correspondant à l'ensemble des mots existants dans le corpus (cas cliniques et discussions).

2.4 Appariements

Les appariements peuvent se faire de différentes manières. Nous avons choisi deux approches : apparier les cas cliniques avec les discussions et apparier les discussions aux cas cliniques (notées respectivement *c2d* et *d2c*). Dans la première méthode, les modèles de langue sont entraînés sur les cas cliniques, et nous estimons la perplexité de chaque modèle sur les discussions. Dans la seconde méthode, nous faisons l'inverse : nous entraînons les modèles de langue sur les discussions et nous estimons les scores de perplexité sur chacun des cas cliniques.

Nous avons également testé deux techniques permettant le choix des meilleures paires en fonction du score de perplexité. La première, non-exclusive (NE), consiste à choisir indépendemment et pour chaque LM le texte avec la perplexité la plus basse. Cela signifie que le même texte peut être apparié à plusieurs LMs. Dans la seconde technique, exclusive (E), un cas ne peut être apparié qu'une seule fois avec une discussion (et vice-versa). Pour ce faire, nous choisissons de façon itérative la paire de cas et discussion générant la perplexité la plus faible sur toutes les paires possibles, avant de

supprimer ces cas et discussions de la liste de choix futurs. Cela exige que les scores de perplexités soient comparables, ce qui est le cas ici, le même vocabulaire ayant été utilisé pour générer tous les modèles de langue.

2.5 Expériences & Discussions

Nous avons testé différentes combinaisons des techniques de pré-traitement et d'appariement présentées ci-dessus. Toutes les expériences ont été effectuées sur l'entièreté du corpus d'apprentissage, soit 290 paires de discussions et cas cliniques. Les scores obtenus sur le corpus d'évaluation pour les soumissions finales sont également présentés.

2.5.1 Effet du pré-traitement de texte

Les résultats présentés dans le tableau 1 mettent en exergue l'effet de différentes techniques de pré-traitement sur les corpus de cas et discussions. Dans ce tableau, nous avons comparé uniquement les différentes approches en fonction de la méthode d'appariement *d2c* décrite dans la section 2.4. De plus, tous les LMs sont d'ordre 2. La racinisation est notée *rac*, la suppression des mots-vides *mv* et le désabrègement *des*.

Le système initial, qui ne comporte aucun pré-traitement, atteint un score de 61,38 en précision et rappel. Les pré-traitements classiques de racinisation et de suppression de mots-vides améliorent logiquement les scores de près de 11 points. le processus de désabrègement automatique, seul, offre des scores de précision et de rappel de 80,69, soit près de 19 points d'amélioration. Lorsqu'on combine les deux pré-traitements, malheureusement, les améliorations ne se cumulent pas. Au contraire, on observe une légère contre-performance de 0,7 points par rapport au meilleur système.

Pré-traitement	Apprentissage
Initial	61,38
rac+mv	72,41
des	**80,69**
rac+mv+des	80,00

TABLE 1 – Scores de précision sur les données d'apprentissage de la tâche 2, en fonction du pré-traitement testé. Tous les LMs sont d'ordre 2, et l'appariement s'est fait de façon exclusive, en direction d2c. *rac : racinisation ; mv : mots vides supprimés ; des : désabrègement.*

2.5.2 Effet de l'ordre du modèle de langue

Utilisant des modèles de langue, nous nous sommes intéressés à l'impact de l'ordre de ces derniers. Le tableau 2 présente les résultats obtenus. Nous avons utilisé une configuration *d2c*, avec le pré-traitement de désabrègement. On peut constater que les modèles de langue d'ordre 2 et 3 obtiennent de meilleurs résultats que les modèles d'ordre 1 (amélioration de près de 3 points à l'ordre 2).

Ordre	Précision
1-gram	77,24
2-gram	**80,68**
3-gram	79,31

TABLE 2 – Scores de précision pour les données de la tâche 2, en fonction de l'ordre du LM. L'appariement s'est fait de façon exclusive, en direction *d2c*. Le corpus d'apprentissage a été pré-traité uniquement avec les abréviations normalisées (*des*).

2.5.3 Impact des techniques d'appariement

Nous avons également testé les différentes techniques d'appariement introduites dans la Section 2.4. La Table 3 souligne l'amélioration apportée par la technique d'exclusivité, avec un score de précision et de rappel systématiquement plus haut que pour les mêmes systèmes n'utilisant pas cette technique. En effet, sans ce procédé, il est probable que si un cas est relativement général dans les termes qui le composent, il soit apparié à une grande majorité de discussions (ou vice-versa). Puisque la tâche 2 nécessite qu'un texte ne soit apparié qu'une seule fois, nous avons choisi d'utiliser cette technique dans nos soumissions.

L'importance de la direction utilisée pour l'appariement (*d2c* ou *c2d* - voir Section 2.4), est aussi mise en exergue dans les résultats présentés Table 3. Il semble ainsi que mesurer la proximité de modèles de langue estimés sur les discussions par rapport à chacun des cas cliniques (*d2c*) produise des résultats plus probants que dans le cas inverse. Les résultats peuvent s'expliquer par la plus grande longueur des discussions par rapport aux cas (environ 393 mots en moyenne pour les cas versus 919 mots pour les discussions). Les discussions permettant ainsi d'estimer des modèles de langue plus variés. Une expérience intéressante pour le futur serait de combiner les deux approches, et sélectionner les meilleurs scores sur toutes les paires possibles, avec les deux directions *c2d* et *d2c*.

Direction	NE/E	Précision
c2d	NE	40,34
c2d	E	71,03
d2c	NE	48,96
d2c	E	**80,69**

TABLE 3 – Scores de précision pour les données de la tâche 2, en fonction de la méthode d'appariement. L'appariement s'est fait de façon exclusive (*E*) ou non-exclusive (*NE*), dans les deux directions (*d2c* et *c2d*). Le corpus d'apprentissage a été pré-traité uniquement avec les abréviations normalisées (*des*).

2.5.4 Résultats soumis

La Table 4 récapitule les résultats obtenus sur les trois contributions que nous avons soumis pour la tâche 2 de DEFT 2019. En accord avec les conclusions tirées Section 2.5.3, nous utilisons l'approche d2c et la technique d'exclusion pour les trois soumissions.

La première version testée (*run-1*) correspond aux meilleurs résultats obtenus lors de tous nos essais sur le corpus d'apprentissage. Le texte a été racinisé, désabrégé et les mots vides supprimés. Les modèles de langue sont d'ordre 1, et ont été créés sur les discussions (direction *d2c*). Pour la

deuxième soumission (*run-2*), nous avons choisi l'approche qui serait théoriquement la meilleure en se basant exclusivement sur les conclusions tirées lors des expériences sur le corpus d'apprentissage. Nous avons ainsi utilisé un modèle de langue d'ordre 2 et le seul pré-traitement effectué est le désabrègement. Enfin, la troisième version (*run-3*) est plus expérimentale. Nous avons décidé de jouer sur l'ordre du modèle de langue (choisissant un modèle d'ordre 3) et d'utiliser outre cette variable les mêmes caractéristiques que pour la version 1 (désabrègement, racinisation, suppression des mots vides, direction *d2c* et technique d'exclusion).

Version	Ordre	Pré-traitement	Précision	
			Apprentissage	Évaluation
run-1	1-gram	rac+mv+des	**81,72**	**84,11**
run-2	2-gram	des	80,69	76,17
run-3	3-gram	rac+mv+des	80,34	83,18

TABLE 4 – Scores de précision pour les résultats soumis pour la tâche 2. L'appariement s'est systématiquement fait de façon exclusive, en direction *d2c*.

Les approches 1 et 3 produisent les résultats attendus sur les données de test, proches de ceux obtenus sur le corpus d'entraînement. La seconde version (run-2) cependant, produit sur le corpus de test des résultats en deçà de ceux obtenus sur le corpus d'apprentissage. Puisque nous savons que l'impact de l'ordre des modèles de langue est restreint, il est probable que ces résultats viennent du pré-traitement choisi. Cela peut souligner la possible importance de la racinisation et de la suppression de mots-vides lors de l'utilisation de corpus de taille limitée.

Il est également intéressant d'observer que l'approche donnant lieu aux meilleurs résultats utilise des modèles de langue d'ordre 1. Ce type de modèle de langue (ou modèle « sac de mots »), qui ne prend en compte que la fréquence des mots, ignorant leur ordre, peut donc suffire pour ce type d'exercice. En effet, la taille extrêmement restreinte des cas et discussions sur lesquelles les modèles de langue ont été estimés ne permet aucun gain d'information en utilisant des modèles d'ordre plus élevé.

3 Tâche 3 : Extraction d'information sur des cas cliniques

La tâche 3 est une tâche d'extraction d'information de type démographique et médicale. Ses objectifs concernent l'identification de cinq types d'informations correspondant au moment du dernier élément clinique rapporté dans le cas : l'âge et le genre de la personne concernée, l'origine (motif de la consultation ou de l'hospitalisation) et l'issue parmi cinq valeurs possibles (guérison, amélioration, stable, détérioration, décès). Pour tous ces cas, il est possible qu'une ou plusieurs des informations soient manquantes. Dans cette situation, la valeur est 'NUL'.

3.1 Approches

Nous présentons ci-dessous deux approches utilisées pour tenter de résoudre cette tâche : une approche neuronale et une approche hybride intégrant des connaissances linguistiques. L'évaluation de leur pertinence sera ensuite proposée.

3.1.1 Pré-traitements des données

La tokénisation et la suppression de la casse ont été réalisées à l'aide de notre outil interne Qnlp-toolkit [2].

La lemmatisation permet d'obtenir la forme canonique des mots. Elle trouve son intérêt dans le cadre de ce travail car elle permet de débarrasser les mots des marques d'inflexion telles que celles de genre (masculin, féminin), de pluriel ou de conjugaison. La lemmatisation d'un verbe est la forme à l'infinitif de ce verbe, celle d'un nom, adjectif ou déterminant, sa forme au masculin singulier. Les mots, ou plus précisément les chaînes de caractères, peuvent ainsi être comparés à un niveau plus fin. Elle est effectuée en utilisant des règles de correspondance à partir des données de WordNet (Fellbaum, 1998).

La lemmatisation a été appliquée uniquement dans le cadre de l'approche hybride. En effet, les modèles neuronaux se basent sur des représentations vectorielles des mots et nécessitent pas de pré-traitement si ce n'est la tokénisation.

3.1.2 Approche neuronale

Nous avons utilisé un modèle neuronal supervisé pour l'étiquetage des empans correspondant aux informations à extraire des cas cliniques en adoptant un schéma d'étiquetage BIO (Ramshaw & Marcus, 1999). L'approche choisie a été proposée dans Ma & Hovy (2016). Elle a l'avantage de ne pas nécessiter un volume de données important pour l'apprentissage et obtient des performances au niveau de l'état de l'art pour la reconnaissances des entités nommées ou l'étiquetage en parties du discours.

Les résultats sont rendus possibles grâce à l'utilisation de représentations pré-entraînées de mots et de caractères ainsi qu'à la combinaison d'un réseau de neurones récurrent (*Bi-directional Long-Short Term Memory*, Bi-LSTM), un réseau de neurones convolutionnel (CNN) et un champ markovien conditionnel (CRF) tels que présentés dans la figure 1.

Pour créer le corpus d'apprentissage, nous avons manuellement étiqueté les 290 cas cliniques selon 4 types d'étiquettes : Âge, Genre, Origine de l'admission et Issue.

Dans le but d'améliorer les performances du modèle et pour palier au faible nombre de données, nous avons généré automatiquement des introductions de cas typiques. En effet, les premières lignes comportent très souvent trois des quatre informations à extraire : âge, genre et raison. Cette génération fut opérée grâce à une grammaire hors-contexte ainsi qu'un certain nombre de ressources linguistiques, telles que des listes de symptômes et de maladies. Les cas suivants ont par exemple pu être générés :

- *mlle a a été mis sous observations suite à hématome temporo-pariétal post-traumatique* ;
- *mme u , 10 ans , a été traité pour une cancer épidermoïde* ;
- *un jeune homme agé de 20 ans avec comme antécédent un perforation digestive instrumentale , a été mis sous observations suite à une surinfection kt jugulaire*.

Un total de 2000 cas ont été générés automatiquement et intégrés aux données d'apprentissage.

En parallèle, nous avons appris sur des corpus plus larges des représentations de mots afin de les utiliser dans le modèle. Nous avons extrait l'ensemble des pages du Wikipédia français appartenant

2. https ://github.com/QwantResearch/qnlp-toolkit

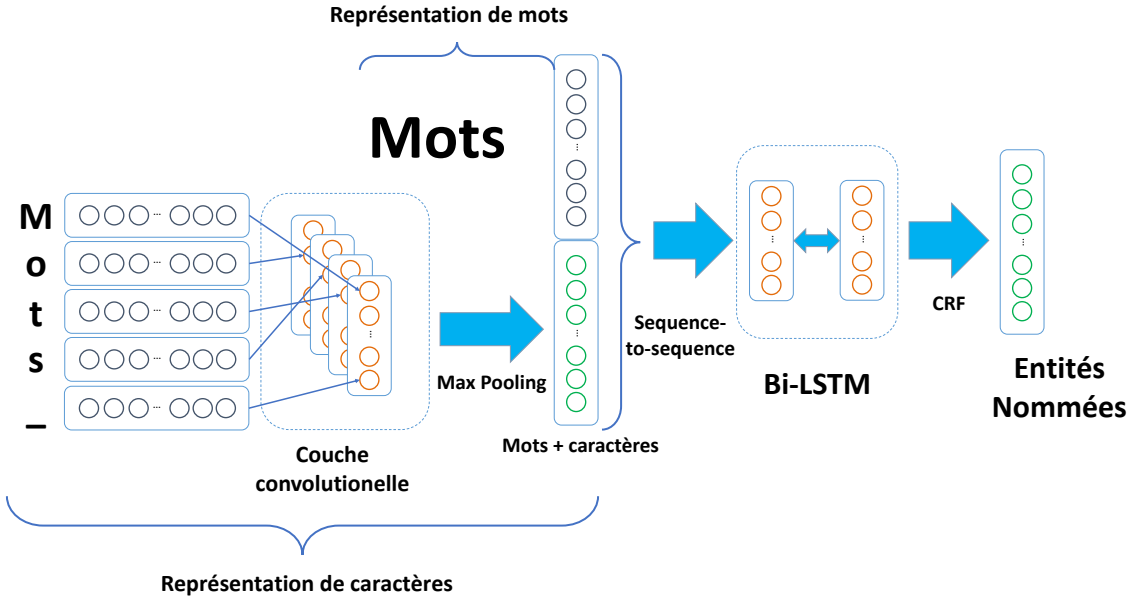


FIGURE 1 – Modèle neuronal utilisé pour l'extraction d'entités nommées fondé sur l'approche proposée par Ma & Hovy (2016).

au portail de la Médecine [3] ainsi que le corpus EMEA contenant des documents PDF de l'agence européenne de Médecine [4]. Sur ce corpus agrégé, nous avons appris des représentations de mots en utilisant l'outil FastText [5] proposé par Bojanowski *et al.* (2017).

Après l'étiquetage du texte, nous inférons les informations demandées, notamment concernant l'âge, le genre, ainsi que l'issue. En effet, une fois l'obtention d'un empan annoté comme "Âge", nous pouvons déduire l'âge en années. A partir d'un certain nombre de règles heuristiques, nous inférons la valeur pour un empan tel que *"18 mois"*. Le résultat obtenu est alors 1 an. De la même manière, un certain nombre d'heuristiques ont été utilisées pour inférer le genre du patient à partir de l'empan correspondant. Concernant l'origine de l'admission, aucune modification n'a été appliquée à l'empan retourné par le modèle.

L'identification de l'issue est quant à elle vue comme un problème de classification multi-classes où l'on considère ses cinq valeurs possibles (guérison, amélioration, stable, détérioration, décès) plus une sixième, utilisée pour nous permettre de considérer les cas où la valeur est NUL. Cette classification repose sur un modèle neuronal proposé par Joulin *et al.* (2017) et est réalisée avec fastText.

3.1.3 Approche hybride

Une méthode alternative à la précédente a été explorée. Elle se fonde sur des connaissances linguistiques pour extraire les unités lexicales correspondant aux âges, genres et origines recherchés, à partir de l'analyse syntaxique en dépendance des cas cliniques et l'utilisation de patrons lexico-syntaxiques prédéfinis.

3. Le portail médecine regroupe les articles appartenant au domaine médical https://fr.wikipedia.org/wiki/Portail:M%C3%A9decine.
4. Le corpus EMEA est accessible à l'adresse http://opus.nlpl.eu/EMEA.php.
5. FastText : https://fasttext.cc

La définition des patrons se déroule selon plusieurs étape. Les textes sont dans un premier temps segmentés en phrases puis filtrés selon la pertinence des informations qui y sont présentes. Les phrases sélectionnées sont ensuite analysées syntaxiquement et les patrons sont construits à partir de leurs analyses.

Certains patrons sont plus productifs que d'autres. On observe par exemple qu'il existe 148 cas où l'âge est exprimé au moyen de l'expression "être âgé de" (Figure 2). Dans 268 cas, il est extrait à partir du chemin de dépendance du modifieur nominal 'an'. Ce dernier peut guider la recherche du

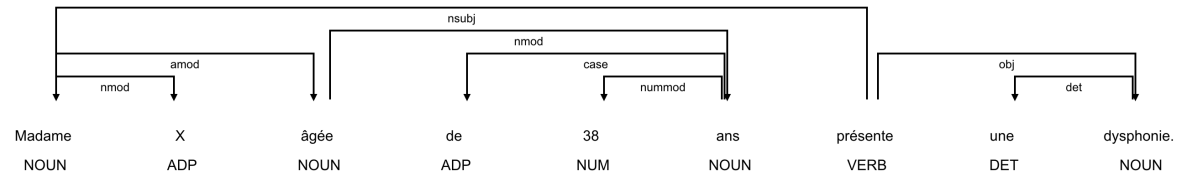


FIGURE 2 – Exemple de phrase sélectionnée après analyse syntaxique.

genre, à partir du syntagme nominal sujet de l'arbre. L'identification du genre est alors réalisée en se référant aux valeurs des traits morpho-syntaxiques de ses constituants.

Dans quelques cas, l'âge n'est pas explicité dans le texte et doit être inféré à partir de dates, par exemple "[...] né le 25 avril 1972 , a consulté en 1996 [...]" (valeur attendue pour âge : 24). Il faut également pouvoir tenir compte des expressions nominales telles que «nouveau né», «quinquagénaire», etc. qui renseignent l'âge.

Pour l'origine, les représentations arborescentes extraites correspondant aux compléments verbaux (syntagmes nominaux et syntagmes prépositionnels) en position postverbale entretenant une relation directe avec les verbes identifiés nous a permis de repérer et d'extraire un ensemble de lexies verbales apparaissant de façon récurrente et se révélant toutes se référer à la tâche de prise en charge (présenter, hospitaliser, admettre, consulter, adresser, etc.).

En ce qui concerne l'identification des issues, la version lemmatisée d'une liste de lexies spécifiques au champ sémantique lié au décès a permis d'identifier les cas de décès, ensuite un classifieur est entraîné pour identifier les autres cas.

3.2 Expériences & Discussions

3.2.1 Corpus d'apprentissage

Dans cette section, nous mesurons la capacité de nos méthodes à extraire les informations recherchées.

Étiquetage automatique Nous présentons tout d'abord les résultats de l'étiquetage automatique du texte. Nous avons retranché dix pour-cent de cas pour l'évaluation. La table 5 présente les scores de F-measure pour chacun des types d'étiquette. Trois cas sont présentés, tout d'abord un apprentissage du modèle sur les données d'apprentissage uniquement, puis l'utilisation de représentations de mots pré-entraînées sur des corpus de données médicales (PRE). Enfin, nous évaluons l'ajout de données générées à l'aide d'une grammaire hors-contexte (GEN).

Modèle / F-measure	Age	Genre	Issue	Origine	All
Étiqueteur	84.21	50.00	51.72	28.57	53.81
Étiqueteur + PRE	87.72	59.26	47.27	36.73	58.60
Étiqueteur + GEN + PRE	90.00	53.33	58.62	43.64	61.80

TABLE 5 – Scores de F-measure sur l'étiquetage automatique de la tâche 3 pour un modèle appris sur 90% du corpus d'entraînement et testé sur le reste. PRE désigne l'utilisation de représentations de mots pré-entraînées sur des corpus de données médicales. GEN correspond à l'extension des données d'apprentissage par des données générées automatiquement à l'aide d'une grammaire hors-contexte.

La combinaison la plus efficace est celle qui regroupe l'utilisation des représentations de mots pré-entraînées et les données générées automatiquement. Le nombre de données du corpus d'apprentissage étant relativement restreint, tout ajout de connaissances extérieures permet d'augmenter sensiblement les performances du modèle.

Nous décidons de conserver le modèle avec utilisation de représentations de mots pré-entraînées ainsi que les données générées automatiquement.

Classification de l'issue Les résultats de la classification de l'issue sont présentés dans la table 6. L'évaluation s'est faite par validation croisée avec 10 plis.

Le premier modèle évalué est appris sur l'ensemble du cas clinique. Étant donné que seule une partie du cas clinique se réfère directement à l'issue finale, nous avons étudié un modèle basé uniquement sur la fin du cas clinique, en se limitant aux dernières phrases. De plus, nous considérons aussi des modèles appris uniquement sur les empans annotés comme "Issue". Nous évaluons, dans un premier temps, un modèle appris et testé sur les empans annotés manuellement pour rendre compte de la validité de l'approche. Dans un second temps, nous testons ce même modèle sur les empans annotés automatiquement.

Portion du cas clinique	Précision	Rappel	F-measure
Cas clinique entier	0.4482	0.4459	0.4471
Deux dernières phrases du cas clinique	0.4448	0.4448	0.4448
Empan de l'issue obtenu par étiquetage manuel	0.6215	0.6215	0.6215
Empan de l'issue obtenu par étiquetage automatique	0.4222	0.4222	0.4222

TABLE 6 – Scores de précision, rappel et F-measure sur l'issue en validation croisée. On compare les résultats sur différentes portions de texte du cas clinique. D'un côté, on considère l'entièreté du cas clinique ainsi que les deux dernières phrases du cas clinique. De l'autre, on considère uniquement les empans étiquetés comme "Issue". Ces empans sont obtenus manuellement dans un cas et automatiquement dans l'autre.

On observe que le meilleur score est celui obtenu en prédisant l'issue uniquement sur la partie du texte s'y référant. Malheureusement, lorsque l'empan est obtenu par étiquetage automatique et non pas manuel, la qualité est grandement dégradée. Cela est dû à la faible qualité de l'étiquetage automatique de l'issue présenté dans le paragraphe précédent.

Pour la phase de test, on décide de conserver le cas où l'on considère le document en entier ainsi que celui basé sur les empans sélectionnés seulement.

3.2.2 Corpus de test

Après avoir choisi les systèmes les plus prometteurs à partir des résultats obtenus sur le corpus d'entraînement, nous avons pu évaluer nos approches bout-à-bout sur le corpus de test.

Âge et genre Les résultats concernant l'âge et le genre sont présentés dans la table 7. L'approche neuronale d'étiquetage de mots et l'approche par arbre syntaxique sont comparées. On observe que l'approche par étiquetage neuronal fonctionne mieux que l'approche lexico-syntaxique. En effet, malgré le faible nombre de données à l'origine (290 données d'apprentissage), l'extension des cas cliniques par génération automatique ainsi que le recours aux représentations de mots pré-entraînées permettent d'obtenir une bonne généralisation, et donc des résultats satisfaisants pour l'étiquetage automatique.

Approche	Age			Genre		
	Précision	Rappel	F1	Précision	Rappel	F1
Étiqueteur + PRE + GEN (*run-1*)	**0.9748**	**0.9023**	**0.9371**	0.9421	0.9465	0.9442
Analyse lexico-syntaxique	0.9719	0.8860	0.9269	**0.9555**	**0.9488**	**0.9521**

TABLE 7 – Scores de précision, rappel et F-measure (F1) pour l'extraction de l'âge et du genre sur le corpus de test. Étiqueteur + PRE + GEN correspond au système envoyé au soumission en première tentative (*run-1*). L'analyse lexico-syntaxique utilise une approche par arbre syntaxique pour extraire les informations.

Approche	macro			micro			*micro overlap accuracy*
	Pr	Rp	F1	Pr	Rp	F1	
Étiqueteur + PRE + GEN (*run-1*)	0.785	0.579	0.666	0.658	0.640	0.649	0.589

TABLE 8 – Scores de micro et macro précision, rappel, F-measure et *micro overlap accuracy* pour l'extraction de l'origine de l'admission sur le corpus de test à partir d'un système étiqueteur neuronal avec représentations de mots pré-entraînées et génération de données d'apprentissage.

Origine de l'admission Les résultats relatifs à l'admission sont présentés dans la table 8. Les scores issus de l'étiqueteur neuronal sont prometteurs et les différences entre précision et rappel (macro et micro) laissent supposer que le modèle retourne un empan de texte trop précis. Nous n'avons pas pu obtenir de résultats concluants par analyse lexico-syntaxique car, contrairement à l'âge et au genre, les cas cliniques ne suivent pas un schéma suffisamment récurrent pour obtenir une bonne extraction.

Issue Les résultats de l'issue sont présentés dans la table 9. Dans les deux premiers cas, nous avons utilisé uniquement un modèle de classification pour prédire l'ensemble des classes. Nous avons tout

d'abord évalué un modèle appris à partir de l'entièreté du cas clinique. Nous avons ensuite comparé ses résultats avec ceux d'un autre classifieur entraîné à prédire sur les données étiquetées automatiquement. Enfin, nous avons identifié les cas de décès en utilisant des connaissances linguistiques puis appris un modèle pour les issues restantes sur les quatre dernières phrases des cas cliniques. Après annotation manuelle du corpus d'entraînement pour étiquetage, nous avons en effet fait plusieurs observations relatives à l'issue. D'une part, les cas de décès se prêtent mieux à une approche lexicale avec un vocabulaire très spécifique. Et d'autre part, dans la majorité des cas, l'empan de texte renvoyant à l'issue apparaît vers la fin du cas clinique. Cette dernière approche est celle qui retourne les meilleurs résultats avec un score de 0.60 et 0.58 en précision et rappel respectivement.

Approche	Issue		
	Précision	Rappel	F1
Cas clinique entier	0.5285	0.5199	0.5241
Empan de l'issue obtenue par étiquetage (*run-1*)	0.5198	0.4918	0.5054
Traitement linguistique de *décès* et quatre dernières phrases	**0.5985**	**0.5831**	**0.5906**

TABLE 9 – Scores de précision, rappel et F-mesure pour l'issue sur le jeu de test. Le première approche considère l'ensemble du cas clinique. Une seconde approche utilise uniquement l'empan "Issue" obtenu par étiquetage automatique. Enfin, une dernière approche utilise un traitement linguistique pour la détection de l'issue "décès", tandis que les quatre dernières phrases sont considérés pour prédire les autres issues.

4 Conclusion

Dans cet article, nous avons présenté notre participation aux tâches 2 et 3 proposées dans le cadre du DEFT 2019, correspondant respectivement aux tâches de recherche de similarité sémantique et d'extraction d'information, appliquées au domaine médical.

La méthode utilisée pour la tâche 2 repose sur l'utilisation de modèles de langue pour estimer la similarité entre documents en ne nous restreignant qu'aux données fournies. Elle a ensuite été étendue à des essais sur l'impact des pré-traitements utilisés en faisant varier les conditions d'appariement.

En ce qui concerne la tâche 3, nous avons réalisé un système entièrement basé sur des réseaux neuronaux qui combine étiquetage de séquences textuelles et classification pour extraire les informations recherchées des cas cliniques. Si le faible nombre de documents disponibles pour l'apprentissage constituait une contrainte forte pour cette méthode, nous avons choisi d'augmenter le corpus en générant de nouveaux cas cliniques, améliorant ainsi nos résultats.

Finalement, les résultats obtenus sur les deux tâches semblent encourageants.

Références

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146.

FELLBAUM C. (1998). *WordNet : An electronic lexical database*. MIT Press.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). Cas : French corpus with clinical cases. In *LOUHI 2018 : The Ninth International Workshop on Health Text Mining and Information Analysis*.

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 de DEFT*.

JOULIN A., GRAVE E., BOJANOWSKI P. & MIKOLOV T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 427–431, Valencia, Spain : Association for Computational Linguistics.

MA X. & HOVY E. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, volume 1, p. 1064–1074.

PONTE J. M. & CROFT W. B. (1998). A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, p. 275–281.

PORTER M. F. (2001). Snowball : A language for stemming algorithms.

RAMSHAW L. A. & MARCUS M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, p. 157–176. Springer.

STOLCKE A. (2002). Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Aprentissage non-supervisé pour l'appariement et l'étiquetage de cas cliniques en français - DEFT2019

Damien Sileo[1, 2] Tim Van de Cruys[2, *] Philippe Muller[2] Camille Pradel[1]
(1) Synapse Développement, 5 Rue du Moulin Bayard, 31000 Toulouse
(2) IRIT, Université Paul Sabatier 118 Route de Narbonne 31062 Toulouse
damien.sileo@synapse-fr.com, camille.pradel@synapse-fr.com,
philippe.muller@irit.fr, tim.van-de-cruys@irit.fr

RÉSUMÉ

Nous présentons le système utilisé par l'équipe Synapse/IRIT dans la compétition DEFT2019 portant sur deux tâches liées à des cas cliniques rédigés en français : l'une d'appariement entre des cas cliniques et des discussions, l'autre d'extraction de mots-clefs. Une des particularité est l'emploi d'apprentissage non-supervisé sur les deux tâches, sur un corpus construit spécifiquement pour le domaine médical en français

ABSTRACT

Unsupervised learning for matching and labelling of french clincal cases - DEFT2019

We present the system used by the Synapse / IRIT team in the DEFT2019 competition covering two tasks on clinical cases written in French : the matching between clinical cases and discussions, and the extraction of key words. A particularity of our submissions is the use of unsupervised learning on both tasks, thanks to a french corpus of medical texts we gathered.

MOTS-CLÉS : TALN bio-médical, DEFT2019, apprentissage non-supervisé.

KEYWORDS: biomedical NLP, DEFT2019, unsupervised learning.

1 Introduction

Les textes du domaine médical sont une source d'information précieuse dont l'analyse automatique peut aider la recherche et le traitement des patients. Cependant, leur nature non structurée fait de leur analyse automatique est un défi, amplifié par la technicité et la spécificité du langage employé. Cette difficulté est exacerbée dans le cas du français, pour lequel les travaux et ressources sont plus rares.

La campagne d'évaluation DEFT2019 (Grabar *et al.*, 2019) est la première à porter sur des textes cliniques français. Les données sont constituées de cas cliniques, accompagnée de discussions correspondantes. Sur ces données, la campagne d'évaluation propose en outre les tâches suivantes :

— L'étiquetage de couples cas cliniques/discussions, par plusieurs expressions clefs choisies dans un ensemble pré-défini de 1311 expressions. (tâche 1)

— L'appariement de cas cliniques avec des discussions originellement correspondantes (tâche 2)

Une autre tâche portant sur l'extraction d'informations (e.g. âge, sexe des patients) n'est pas traitée ici. Dans cet article, on utilise des techniques d'apprentissage non-supervisé pour participer aux deux taches (particulièrement pour la tâche 2).

Pour ce faire, on construit un ensemble de corpus basé sur des ressources en français qui servira à

Jeux de données	Nombre de documents
EMEA	26289
Aranea-Med	11093
Wac-Med	2514
Cochrane	7676
Wiki-Med	4933
Deft	3974
Quaero	3479

TABLE 1 – Nombre de documents dans les jeux de données utilisés

pré-entrainer un réseau de neurones basé sur une concaténation d'embeddings différents, et d'un encodage des textes à base de ces représentations vectorielles de mots par des réseaux convolutifs.

2 Constitution d'un corpus de textes médicaux

Pour servir à l'apprentissage non-supervisé, on constitue un corpus médical à partir des sources suivantes :

— L'ensemble des textes de DEFT2019, y compris les données de test
— Les articles de Wikipédia appartenant au portail de la médecine, que nous nommerons Wiki-Med
— EMEA (Tiedemann, 2012) qui contient des textes de l'european medical agency
— Quaero (Névéol *et al.*, 2014), qui contient des titres Medline et des documents EMEA annotés, pour la reconnaissance d'entités nommées et la normalisation (ici, on ignore ces annotations)
— Des résumés d'articles Cochrane (Grabar & Cardon, 2018)

De plus, on augmente ce corpus en utilisant une technique simple d'adaptation de domaine. On entraine un classifieur FastText (Joulin *et al.*, 2016) (paramètres par défauts, 2 itérations) afin d'apprendre à prédire si des textes viennent du corpus Wiki-Med, ou de sources web (Aranea(Panchenko *et al.*, 2017), FrWac(Ferraresi *et al.*, 2008)) échantillonnées de sorte à contenir deux fois plus de textes que Wiki-Med.

À partir de ce classifieur, on extrait des corpus web les textes qui sont prédits comme appartenant à Wiki-Med. Ces textes sont présents en plus grand nombre que ceux de Wiki-Med. Si le classifieur s'est mépris sur leur origine, c'est, on l'espère, qu'ils sont lexicalement proches de ceux de Wiki-Med, autrement dit qu'ils concernent la médecine, et qu'ils seront utiles pour l'apprentissage de représentations de mots ou d'encodeurs de textes. L'aggrégation de ces corpus, ayant subie une déduplication des textes strictement identiques sera nommé Fr-Med dans la suite de l'article. Les nombres de documents selon ces différentes sources sont donnés dans la table 2.

3 Prétraitement du texte

Les textes passent par la fonction *fix_text* de la librairie ftfy (Speer, 2019) afin de remédier à certains problèmes d'encodage. Ils sont ensuite passés en minuscules. Les nombres entre crochets (citations) sont éliminés, de même que les chiffres entre parenthèses. Les virgules et les mots vides de la liste

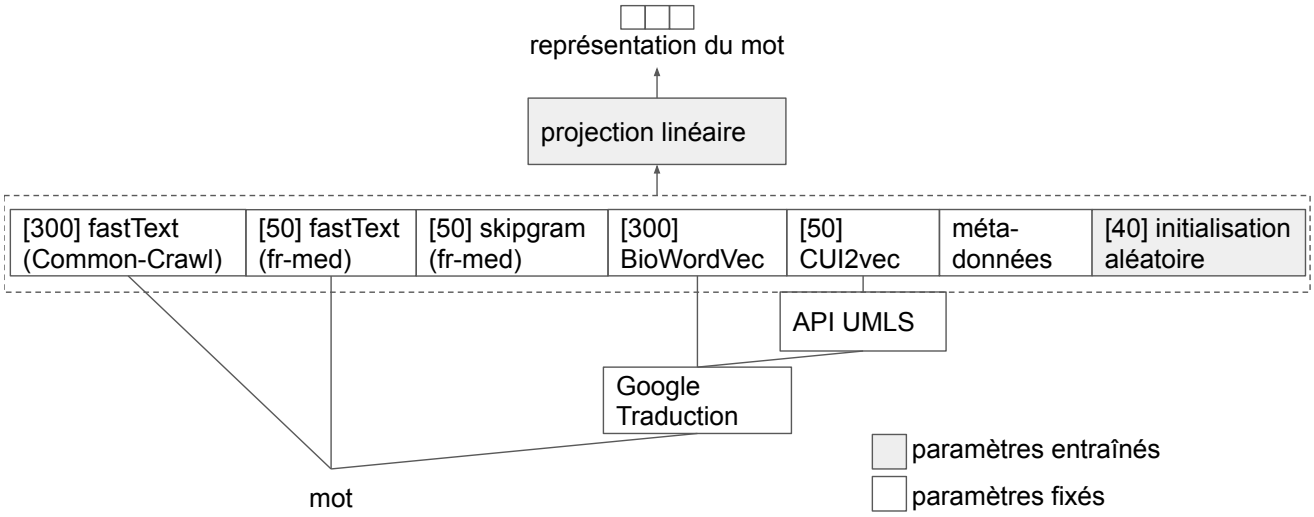


FIGURE 1 – Représentation vectorielle d'un mot. Les chiffres entre crochets désignent la dimension

stopwords-iso [1] sont éliminés.

4 Embeddings

Les embeddings utilisés sont les suivants, aussi représentés dans la figure 1 :

— FastText (Bojanowski *et al.*, 2017) pré-entraînés sur CommonCrawl distribuées sur `fasttext.cc` [2]

— Des embeddings FastText appris sur le corpus *Fr-Med* décrit précédemment, avec une taille de 50, 12 epochs et les paramètres par défaut sinon

— Des embeddings SkipGram appris sur le corpus *Fr-Med* et accédés par PyMagnitude (Patel *et al.*, 2018)

— BioWord2Vec (Chen *et al.*, 2018) [3] utilisés à la suite d'une traduction en anglais utilisant google API [4]

— CUI2Vec (Beam *et al.*, 2018) qui sont des embeddings de Concept Unique Identifier (CUI) UMLS. Le lien entre les mots et les CUI est obtenu en utilisant la fonction de recherche l'API publique UMLS. Leur dimension a été réduite à 50 en utilisant (Raunak, 2017).

— Des méta-données : l'appartenance aux dictionnaires de chaque embeddings (5 booléens), la fréquence d'occurence dans les documents de DEFT (répartie en 12 quantifications binaires)

— Une partie initalisée aléatoirement et apprise lors de l'optimisation des taches finales, de dimension 40

5 Tâche 2 - Appariement des cas cliniques et des discussions

5.1 Modélisation du problème

On traite le problème comme de la classification à partir des paires de phrases, la classe prédite étant l'existence d'un lien entre un cas et une discussion. Les données d'entrainement fournissent déjà

1. https ://github.com/stopwords-iso/stopwords-iso
2. `https://fasttext.cc/docs/en/crawl-vectors.html`
3. `https://ftp.ncbi.nlm.nih.gov/pub/lu/Suppl/BioSentVec/BioWordVec_PubMed_MIMICIII_d200.bin`
4. Les API ont été consommées en mai 2019

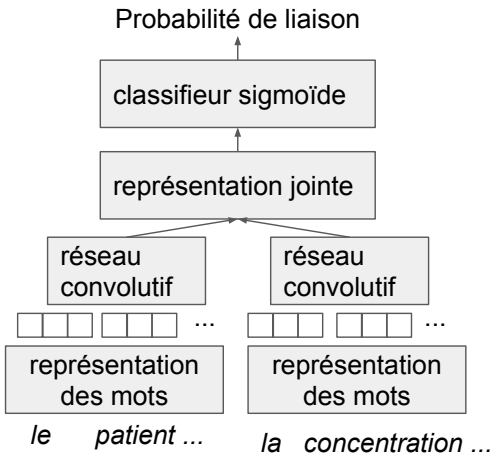


FIGURE 2 – Architecture de la prédiction de liaison entre cas et discussion

des cas et discussions liés. Pour générer des exemples non-liés, on applique un produit cartésien entre les cas et les discussions (en éliminant les cas et dicussions liées). Ensuite, les paires liées sont suréchantillonées par un facteur 10 afin de diminuer le déséquilibre des classes, sans pour autant perdre des données en sous-échantillonnant les exemples non-liés. 15% des données (brutes, c'est à dire avant le produit cartésien) sont réservées à la validation.

La figure 2 montre l'architecture du système utilisé pour la tache 2. Un réseau convolutif compose les représentations de mots décrites précedemment afin de représenter les cas et discussions par un vecteur de taille fixe. Ces vecteurs sont eux mêmes composés en une représentation jointe qui sert à classifier la présence de lien entre cas et discussion. La représentation jointe est $\text{ReLu}(W[u, v, u \odot v, |u - v + t|)]$ (Sileo *et al.*, 2019) où u et v sont les sorties des réseaux convolutifs t est un paramètre de la même dimension que u et v. Les paramètres de ce réseau sont optimisés de sorte à minimiser l'entropie croisée. Les probabilités de laison obtenues permettent de classer pour chaque cas l'ensemble des discussions selon leur pobabilité, et la plus probable est prédite dans les soumissions. La métrique d'évaluation est la proportion de cas pour lequels la bonne discussion a été trouvée, qu'on nomme précision.

5.2 Représentation des séquences de mots

Le réseau convolutif est constitué de $N = 768$ filtres de taille 1, et $N = 768$ filtres de taille 3, concaténés et suivis par une activation ReLu et d'un max-pooling. La taille des séquences de mots d'entrée est limitée à 600.

5.3 Hyperparamètres

On utilise l'optimiseur Adam (Kingma & Ba, 2014) avec le learning rate 0.002, déterminé par validation croisée.

5.4 Pré-entrainement pour l'appariement

On pré-entraine le réseau convolutif et la représentation de mots en utilisant une tâche d'apprentissage non-supervisée inspirée de (Devlin *et al.*, 2018) et (Logeswaran & Lee, 2018) qui consiste découper chaque document du corpus Fr-Med, en deux parties, puis à prédire si deux parties appartiennent au même document. On ne garde que les documents d'au moins 40 mots. On pourrait simplement

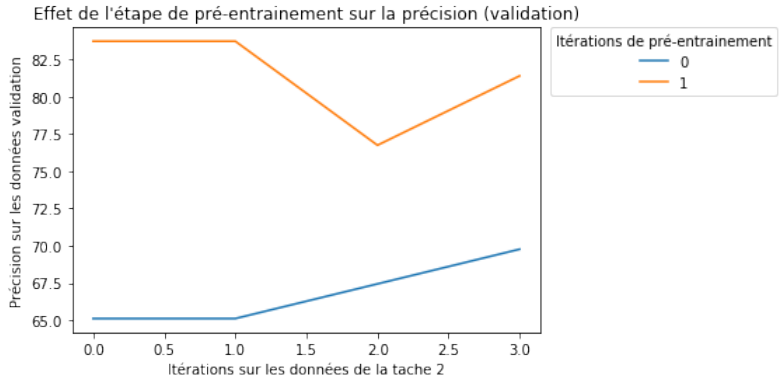


FIGURE 3 – Courbe d'apprentissage pour la tache 2

découper en deux parties égales les documents, mais dès lors le réseau pourrait apprendre à apparier les séquences de même taille, ce qui n'est pas intéressant pour l'apprentissage de représentations. On choisit donc cette division : l'endroit qui divise les deux parties des documents est choisi selon une distrbution de probabilité uniforme telle que la taille de chacun des segments soit supérieure à 20.

Les exemples négatifs sont générés en appariant aléatoirement des segments n'appartenant pas au même document. Les segments appartenant au même document étant a priori thématiquement proches, du moins plus proches en moyenne que des segments issus d'autres documents pris au hasard, cette tâche permet de tirer parti de Fr-Med pour entrainer les encodeurs de textes.

Le jeu de données résultat contient $1.4M$ exemples dont 10% de segments appartenant au même document.

5.5 Influence du pré-entrainement

La figure 3 montre l'influence de cette étape pré-entrainement. Sans l'itération de pré-apprentissage, la précision reste limitée même après plusieurs itérations sur les données de la tâche 2. La tache de pré-entrainement semble donc assez liée à la tache 2 pour être utile.

5.6 Sélection et aggrégation de modèles

On entraine plusieurs modèles, sur des ensemble d'entrainement et de validation distincts, et on selectionne ceux qui ont les meilleures performances sur les données de validation. Les probabilités de liaisons des différents modèles sont aggrégées par une moyenne.

5.7 Résultats

Les résultats des différents runs figurent dans la table 5.7. Le run 2 est issu d'un seul modèle, et les runs 1 et 3 utilisent 4 et 6 modèles.

6 Tâche 1 - Étiquetage des cas cliniques

6.1 Apprentissage non-supervisé pour la détection de mots-clefs

On génère un corpus d'apprentissage non-supervisé pour l'étiquetage de textes avec les données web Aranea(Panchenko *et al.*, 2017). Pour ce faire, on parcourt les pages, et lorsqu'un mot-clef de la liste prédéfinie par la tâche apparait, on le considère comme un label à prédire. Les labels à

	précision (test entier)	précision (test dédupliqué)
run 1	57.94%	61.68%
run 2	54.21%	56.07%
run 3	57.00%	63.08%

TABLE 2 – Résultats sur les données de test de la tache 2 (précision), avec les exemples de tests dans leur totalité, et sur un ensemble exempt de doublons

prédire sont masqués dans 90% des cas pour que le modèle n'apprenne pas une simple détection de la présence stricte des mots clefs. Le corpus résultant contient $16.8k$ exemples contenant au moins des mots-clefs. Le réseau de convolution défini dans la section 5.2, mais avec une seule convolution de taille 1 avec 256 filtres, suivi d'un classifieur log-linéaire, est entrainé de sorte à prédire les mots clefs, Le modèle est pré-entrainé à la prédiction de ces mots-clefs sur ce corpus sur 8 itérations. Le modèle est ensuite optimisé pour la même tâche sur les données de DEFT2019 (1874 exemples) (train/test, cas/discussions) dédupliquées, avec 3 itérations. Notons bien qu'il s'agit de prédiction de mots clefs présents dans le textes et non pas de mots clefs étant ceux qui sont des labels à prédire, fixés par l'annotation mise en oeuvre pour la tâche.

6.2 Modèle de classification

Soit n le nombre de labels (ici 1311) Pour chaque label, on calcule K traits par exemple. Les traits utilisés sont les suivants :

- présence du mot-clef dans un texte $x_i = 1$ si le label i est dans le texte présent, $x_i = 0$ sinon.
- présence du mot-clef sans qu'un autre mot-clef plus petit soit présent dans le texte
- présence prédite par le classifieur de mots clefs décrit dans la sous-section précédente, élevée aux puissances 1,2 et 4.
- score du mot clef selon la librairie fuzzywuzzy [5] qui détecte des mots clefs ou des éditions de leur chaine de caractères en se basant sur des distances de levenstein. La fonction *p*artial_score est utilisée. Les scores sont seulement comptabilisés s'ils dépassent le seuil de 0.85.

Un paramètre θ pondère ces K traits, pour fournir un score pour chaque mot clef dans le cas ou dans la discussion. Les paramètres θ et α sont intialisés de sorte à ce que chaque composante vaille 1. Les scores du cas et de la discussion sont agrégés avec une moyenne pondérée : $\text{score}(\text{exemple}) = \text{score}(\text{discussion}) + a.\text{score}(\text{cas})$ Les paramètres de ce réseau sont optimisés de sorte à minimiser l'entropie croisée.

6.3 Hyperparamètres

La valeur déterminée $a = 0.3$, ce qui signifie que la discussion semble plus utile que le cas pour déterminer les mots clefs. On utilise l'optimiseur Adam (Kingma & Ba, 2014) avec le learning rate 0.01. Une régularisation d'activité L1 est appliquée à la sortie du réseau, avec un coefficient $\lambda = 10^{-5}$ (ce qui revient à ajouter $\lambda|y|_{L1}$ à la fonction de coût). Ces hyperparamètres ont été choisis par validation croisée.

6.4 Résultats

Les résultats des différents runs figurent dans la table 6.4 On entraine une multitude de modèles et les selectionne selon leur précision sur des ensembles de validations différents à chaque fois (la

5. https://github.com/seatgeek/fuzzywuzzy

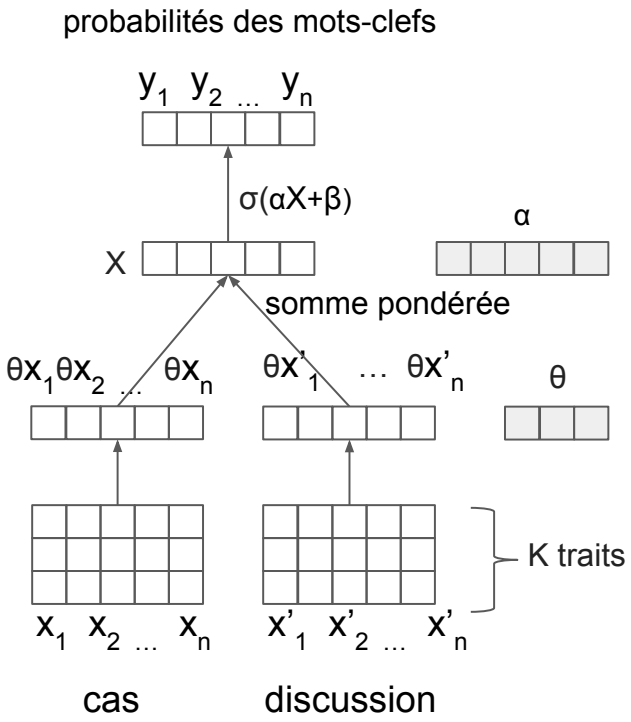


FIGURE 4 – Architecture tache 1. θ et α sont des paramètres libres. σ dénote la fonction sigmoïde. Les vecteurs en gris sont des paramètres appris

	MAP	P@n
run 1	36.47%	43.87%
run 2	44.64%	43.90%
run 3	36.53%	43.87%

TABLE 3 – Résultats sur les données de test de la tache 1

.

partition étant aléatoire, avec 85% de données pour l'entrainement) Le run 1 est un ensemble de 13 modèles ayant une précision supérieure à 29% en validation. Le run 2 est un ensemble de 8 modèles ayant une précision supérieure à 26% sur les données de validation. Le run 3 est un ensemble de 12 modèles ayant une précision supérieure à 30% en validation. Visiblement, la selection de modèles par la précision en validation n'est pas optimale.

7 Conclusion

On a décrit deux systèmes, faisant appel à de l'apprentissage non-supervisé, pour l'étiquetage de cas cliniques et l'appariement avec des discussions. Il serait intéressant d'utiliser des techniques des traduction de manière plus poussée afin de pouvoir bénéficier des techniques récentes d'apprentissage non-supervisé (Devlin *et al.*, 2018) et leurs déclinaisons pour le domaine biomédical(Lee *et al.*, 2019).

Pour la tâche 2, il serait intéressant d'évaluer un modèle non pas simplement pré-entrainé mais multi-tâches entrainé à la fois à la tache 2 et à la tache de pré-entrainement, pour atténuer l'oubli de cette dernière (Kirkpatrick *et al.*, 2016).

Les résultats de la tache 1 pourraient être améliorés en faisant appel à des techniques de recherche d'information plus poussée (e.g. utilisation d'ElasticSearch) pour la création d'autres traits.

Enfin, les deux tâches étant liées à de l'ordonnancement, des fonctions de coûts plus appropriées pourraient être considérées.

Références

BEAM A. L., KOMPA B., FRIED I., PALMER N. P., SHI X., CAI T. & KOHANE I. S. (2018). Clinical concept embeddings learned from massive sources of medical data. *CoRR*, **abs/1804.01486**.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association of Computational Linguistics*, **5**, 135–146.

CHEN Q., PENG Y. & LU Z. (2018). Biosentvec : creating sentence embeddings for biomedical texts.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

FERRARESI A., ZANCHETTA E., BERNARDINI S. & BARONI M. (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english.

GRABAR N. & CARDON R. (2018). CLEAR-Simple Corpus for Medical French. In *ATA*, Tilburg, Netherlands.

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. *Actes de DEFT*.

JOULIN A., GRAVE E., BOJANOWSKI P. & MIKOLOV T. (2016). Bag of Tricks for Efficient Text Classification.

KINGMA D. & BA J. (2014). Adam : A Method for Stochastic Optimization. *International Conference on Learning Representations*, p. 1–13.

KIRKPATRICK J., PASCANU R., RABINOWITZ N., VENESS J., DESJARDINS G., RUSU A. A., MILAN K., QUAN J., RAMALHO T., GRABSKA-BARWINSKA A., HASSABIS D., CLOPATH C., KUMARAN D. & HADSELL R. (2016). Overcoming catastrophic forgetting in neural networks. cite arxiv :1612.00796.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2019). Biobert : a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv :1901.08746*.

LOGESWARAN L. & LEE H. (2018). An efficient framework for learning sentence representations. p. 1–16.

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The QUAERO French medical corpus : A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, p. 24–30.

PANCHENKO A., RUPPERT E., FARALLI S., PONZETTO S. P. & BIEMANN C. (2017). Building a Web-Scale Dependency-Parsed Corpus from Common Crawl. p. 1816–1823.

PATEL A., SANDS A., CALLISON-BURCH C. & APIDIANAKI M. (2018). Magnitude : A fast, efficient universal vector embedding utility package. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 120–126.

RAUNAK V. (2017). Simple and effective dimensionality reduction for word embeddings.

SILEO D., DE CRUYS T. V., PRADEL C. & MULLER P. (2019). Composition of sentence embeddings : Lessons from statistical relational learning. *CoRR*, **abs/1904.02464**.

SPEER R. (2019). ftfy. Zenodo. Version 5.5.

TIEDEMANN J. (2012). Parallel data, tools and interfaces in opus. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, M. U. DOGAN, B. MAEGAARD, J. MARIANI, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey : European Language Resources Association (ELRA).

Indexation et appariement de documents cliniques avec le modèle vectoriel

Khadim Dramé[1, 2] Ibrahima Diop[1, 2] Lamine Faty[1, 2] Birame Ndoye[1]
(1) Université Assane Seck de Ziguinchor, Diabir, Ziguinchor, Sénégal
(2) Laboratoire d'Informatique et d'Ingénierie pour l'Innovation, Ziguinchor, Sénégal
khadim.drame@univ-zig.sn, ibrahima.diop@univ-zig.sn,
lamine.faty@univ-zig.sn, b.ndoye5360@zig.univ.sn

RÉSUMÉ

Dans ce papier, nous présentons les méthodes que nous avons développées pour participer aux tâches 1 et 2 de l'édition 2019 du défi fouille de textes (DEFT 2019). Pour la première tâche, qui s'intéresse à l'indexation de cas cliniques, une méthode utilisant la pondération TF-IDF (term frequency – inverse document frequency) a été proposée. Quant à la seconde tâche, la méthode proposée repose sur le modèle vectoriel pour apparier des discussions aux cas cliniques correspondants ; pour cela, le cosinus est utilisé comme mesure de similarité. L'indexation sémantique latente (latent semantic indexing – LSI) est également expérimentée pour étendre cette méthode. Pour chaque méthode, différentes configurations ont été testées et évaluées sur les données de test du DEFT 2019.

ABSTRACT

Indexing and matching clinical documents using the vector space model.

In this paper, we present the methods that we developed to participate in tasks 1 and 2 of the 2019 edition of the french text mining challenge (DEFT 2019). For the first task, which focuses on the indexing of clinical cases, a method using TF-IDF weighting (term frequency - inverse document frequency) has been proposed. For the second one, the proposed method is based on the vector space model to match discussions with corresponding clinical cases; for this, the cosine is used as similarity measure. The latent semantic indexing (LSI) is also used to extend this method. For each method, different configurations were tested and evaluated on the test data of DEFT 2019.

MOTS-CLÉS : indexation, modèle vectoriel, TF-IDF, indexation sémantique latente, similarité sémantique, cas cliniques.

KEYWORDS: indexing, vector space model, TF-IDF, latent semantic indexing, semantic similarity, clinical cases.

1 Introduction

Le défi fouille de textes (DEFT) est une campagne d'évaluation visant à promouvoir le développement de méthodes et d'applications dans le domaine du traitement automatique de langues naturelles (TALN). Dans son édition de 2019, il s'intéresse à l'analyse de cas cliniques ; il comporte trois tâches traitant essentiellement l'indexation, la recherche et l'extraction d'informations à partir de textes biomédicaux (Grabar *et al.*, 2019).

La tâche 1 consiste à identifier, à partir d'une liste de mots clés, ceux qui sont pertinents pour représenter un couple cas clinique/discussion donné. Cette question d'indexation où chaque document est associé à un ou plusieurs mots clés, peut être considérée comme un problème de classification multi-label. Dans la littérature, ce problème a suscité un grand engouement et différentes approches sont proposées. Certaines approches prônent la décomposition du problème en sous-problèmes de classification binaire (Read *et al.*, 2011), et d'autres l'adaptation des méthodes existantes et notamment l'algorithme des k plus proches voisins (Huang *et al.*, 2011; Dramé *et al.*, 2016). La méthode que nous proposons s'inscrit dans la deuxième approche et utilise la méthode de pondération TF-IDF pour déterminer les mots clés pertinents pour indexer un document.

La tâche 2, quant à elle, s'intéresse à l'appariement des cas cliniques et des discussions. L'idée est de déterminer, pour chaque cas clinique, la discussion correspondante à partir d'un ensemble de discussions. Pour traiter ce type de problème, la similarité (sémantique) est communément utilisée. Dans la littérature, deux approches différentes sont développées : une exploitant des ressources externes (Schuhmacher & Ponzetto, 2014) et une autre basée sur la représentation vectorielle. La deuxième approche est largement explorée; différents modèles sont utilisés : le modèle vectoriel (Vector Space Model - VSM)(Salton *et al.*, 1975), l'indexation sémantique latente (Latent Semantic Indexing - LSI) (Deerwester *et al.*, 1990), l'allocation de Dirichlet latente (Latent Dirichlet Allocation - LDA) (Blei *et al.*, 2003) et récemment les plongements lexicaux (word embeddings) (Mikolov *et al.*, 2013; Le & Mikolov, 2014) . Nous proposons une méthode d'appariement inspirée de cette approche et fondée principalement sur la représentation vectorielle des documents. Le VSM et la LSI sont expérimentés avec la mesure de cosinus.

Ce papier décrit ces méthodes, développées pour participer à ces deux tâches du DEFT 2019. Le reste du papier est structuré comme suit : nos méthodes d'indexation et d'appariement sont présentées respectivement dans les sections 2 et 3 ; les résultats obtenus sont décrits et discutés respectivement dans les sections 4 et 5.

2 Indexation de cas cliniques

Dans cette section, nous décrivons notre méthode d'indexation et ses différentes extensions, visant à améliorer ses résultats.

L'approche proposée est basée sur la représentation vectorielle des documents. Dans la phase de prétraitement, chaque document est d'abord segmenté en phrases et les phrases en tokens. Ensuite, l'ensemble des n-grams (séquences de 1 à n tokens, n fixé empiriquement à 6) sont extraits et appariés aux mots clés ; pour l'appariement des n-grams aux mots clés, ces derniers sont préalablement normalisés (suppression de mots vides, racinisation). A l'issue de cette étape, on a une liste de mots clés extraits avec leurs poids respectifs dans le document. Pour identifier les mots clés les plus pertinents pour représenter un document (couple cas clinique/discussion), nous avons expérimenté les mesures comme la fréquence du mot clé dans le document (TF), la TF-IDF, la première occurrence du mot et la combinaison des ces deux mesures dans un modèle supervisé. Les trois configurations de cette méthode sont décrites ci-dessous :

- *uasz-indexer1* : les fréquences des mots clés (TF) dans le document sont utilisées pour les classer ;
- *uasz-indexer2* : les scores TF-IDF des mots clés sont utilisées pour les classer ;

- *uasz-indexer3* : les mesures utilisées dans les configurations précédentes (TF et TF-IDF) sont combinées avec la première occurrence du mot clé dans une méthode supervisée. Ainsi, pour chaque mots clé extrait, sa pertinence pour le document cible est prédite par un modèle entrainé (en utilisant un algorithme d'apprentissage automatique) sur le corpus d'entrainement. Ensuite, les mots clés sont classés en fonction de leur pertinence et les top K les plus pertinents pour le document sont retournés, K étant fourni. Nous avons choisi le classifieur *Naive Bayes* qui, dans les tests que nous avons réalisés sur le corpus d'entrainement, a donné de meilleurs résultats.

Nous avons aussi expérimenté la méthode proposée dans (Dramé *et al.*, 2016), basée sur l'algorithme des k plus proches voisins, mais les résultats obtenus sont mitigés.

3 Appariement de cas cliniques et discussions

Dans cette section, nous présentons notre méthode d'appariement et son extension avec l'indexation sémantique latente (LSI).

La méthode développée repose sur le modèle vectoriel et utilise la mesure cosinus pour calculer la similarité entre cas cliniques et discussions. Dans la première étape, les documents (cas cliniques et discussions) sont d'abord segmentés en phrases et les phrases en tokens. Les mots vides sont ensuite élagués. Enfin, l'ensemble des mots (concepts dans le cas du modèle LSI) du corpus constitue les dimensions du modèle. Chaque document est ainsi représenté dans cet espace (de grande dimension) par un vecteur de mots (ou concepts) dont les composants sont les poids de ces derniers. La similarité entre deux documents, représentés dans ce modèle, est ainsi assimilée au cosinus de l'angle formé par les vecteurs correspondants.

Plusieurs modèles de représentation vectorielle sont explorés (VSM, LSI, LDA, doc2vec embeddings) mais le modèle vectoriel (VSM) et l'indexation sémantique latente (LSI) ont donné de meilleurs résultats sur nos tests. Nous avons ainsi soumis les trois configurations suivantes :

- *uasz-app1* : elle utilise le modèle vectoriel sur le corpus de test ;
- *uasz-app2* : elle utilise le modèle vectoriel sur l'ensemble du corpus (corpus d'entrainement + corpus de test) ;
- *uasz-app3* : elle utilise l'indexation sémantique latente sur le corpus de test.

4 Evaluation

Dans cette section, nous allons d'abord présenter les jeux de données et les métriques utilisées pour évaluer les systèmes participants au DEFT 2019. Ensuite, les résultats de nos méthodes seront analysés et discutés.

4.1 Jeux de données

Le DEFT 2019 a porté sur l'analyse de cas cliniques rédigés en français. Il est constitué de trois tâches : la première s'intéresse à l'indexation de cas cliniques (tâche 1), la deuxième à l'appariement de cas cliniques aux discussions (tâche 2) et la troisième se focalise sur l'extraction d'informations (tâche 3) (Grabar *et al.*, 2019).

Pour chaque tâche, les organisateurs ont fourni des corpus d'entrainement et de test (Grabar *et al.*, 2018). Pour la première tâche, un corpus d'entrainement constitué de 290 couples de cas cliniques/discussions a été fourni avec, pour chaque couple, les mots clés associés dans l'ordre décroissant de leur pertinence. Le corpus de test est quant à lui constitué de 213 couples de cas cliniques/discussions avec le nombre de mots clés attendu. En ce qui concerne la tâche 2, un corpus d'entrainement constitué de 290 cas cliniques et 290 discussions est fourni avec un appariement de chaque cas à la discussion correspondante. Le corpus de test comporte 214 cas cliniques ainsi que le même nombre de discussions.

4.2 Mesures d'évaluation

La précision moyenne (Mean Average Precision – MAP) et la précision au rang N (P@N) sont utilisées pour mesurer les performances des systèmes participants à la tâche 1. Pour la tâche 2, les mesures classiques (précision et rappel) sont utilisées pour évaluer les appariements cas cliniques/discussions.

4.3 Résultats

Les résultats de nos différents systèmes participants à la tâche 1 sont présentés dans TABLE 1. Nous remarquons que le système *uasz-indexer2*, utilisant la pondération TF-IDF, a obtenu des résultats largement meilleurs selon les deux mesures d'évaluation utilisées (MAP et P@N). Notons également que le système *uasz-indexer3*, qui utilise une méthode supervisée, est plus performante que *uasz-indexer1*, qui lui se sert de la fréquence des mots clés pour les classer. Nous avons également expérimenté une approche supervisée combinant les attributs TF.IDF, TF et la première occurrence du mot clé dans le document avec différents classifieurs (Naive Bayes, Neural Network et Random Forest) mais les résultats obtenus sont moins intéressants. Enfin, l'approche développée dans (Dramé *et al*., 2016), utilisant la méthode des k plus proches voisins, a été explorée mais elle n'a pas permis d'améliorer les résultats.

Systèmes	MAP	P@N
uasz-indexer1	0,2761	0,3433
uasz-indexer2	**0,3957**	**0,4547**
uasz-indexer3	0,3174	0,3783

TABLE 1 : Résultats de nos systèmes d'indexation à la tâche 1 du DEFT 2019

Comparé aux systèmes participants à la tâche 1, *uasz-indexer2*, notre meilleur système, a obtenu des résultats moyens (MAP de 0,395 contre 0,478 pour le système le plus performant) dépassant légèrement la moyenne (0,385) ; il a obtenu une précision moyenne comparable à la médiane (0,401) sans utiliser aucune ressource externe supplémentaire.

Les résultats des systèmes développés pour l'appariement des cas cliniques et des discussions sont présentés dans TABLE 2. Nous constatons que les système *uasz-app1* et *uasz-app2*, basés tous sur le modèle vectoriel, ont obtenu des résultats meilleurs que ceux de *uasz-app3*, qui implémente l'indexation sémantique latente. Les trois systèmes ont obtenu dans l'ensemble des résultats satisfaisants.

Quand la fréquence documentaire (IDF) est calculée sur tout le corpus (données d'entrainement et de test) (*uasz-app2*) plutôt que sur le corpus de test seulement (*uasz-app1*), les résultats sont légèrement améliorés avec le modèle vectoriel.

Systèmes	Précision	Rappel
uasz-app1	0,8738	0,8738
uasz-app2	**0,8832**	**0,8832**
uasz-app3	0,8318	0,8318

TABLE 2 : Résultats de nos systèmes d'appariement à la tâche 2 du DEFT 2019

D'autres approches telles que l'allocation de Dirichlet latente (Blei *et al.*, 2003) et les plongements lexicaux (Le & Mikolov, 2014) sont aussi explorées mais elles ont donné des résultats mitigés. En plus, elles nécessitent plus de ressources et un temps d'exécution plus important.

Comparés aux résultats globaux de la tâche 2, nos systèmes ont obtenu des performances satisfaisantes ; tous les trois ont dépassé la moyenne (0,803) sans l'utilisation d'aucune ressource externe supplémentaire. En plus, *uasz-app2*, bien qu'étant simple, a donné des résultats prometteurs dépassant la précision médiane (0,862). Toutefois, comparé au système le plus performant (0,953), nos résultats restent à améliorer.

5 Discussion

L'évaluation de notre méthode d'indexation sur les données de test du DEFT 2019 montre que cette dernière, bien qu'étant simple, reste performante. L'utilisation de la pondération TF.IDF s'est montrée pertinente. Le score TF.IDF d'un mot clé reste ainsi un bon indicateur pour mesurer son poids dans un document. La fréquence aussi reste un indicateur intéressant. Toutefois, la combinaison de ces deux mesures et la première occurrence dans une méthode supervisée a, à notre surprise, donné des résultats mitigés. Cette faible performance peut s'expliquer par la taille moins conséquente de notre corpus d'entrainement (290 documents).

Les résultats de la TABLE 1 montre que cette méthode permet de prédire correctement 40% des mots clés permettant d'indexer des couples de cas clinique/discussion. Une analyse détaillée des résultats a permis de constater que notre méthode peine à retrouver certains mots clés notamment ceux qui ne sont pas explicitement mentionnés dans les documents. Par exemple, le mot clé *dépression respiratoire*, utilisé dans l'index du couple *1202936314.txt/21688289148.txt* n'apparaît pas explicitement dans ce cas clinique ; parfois un synonyme est utilisé (par exemple, le mot clé *malformation congénitale* est utilisé pour indexer le couple *132378112.txt/2122683392.txt* tandis dans le document il est mentionné *sténose congénitale*). D'autres figurent dans le document mais sont disjoints (par exemple, *antagoniste récepteur nk1* dans le couple *12587038525.txt/22358514900.txt*). Dans certains cas, notre méthode retourne des mots clés plus généraux ; par exemple, elle prédit le mot clé *intoxication* pour le couple *11022315250.txt/21172085750.txt* alors que celui attendu est *intoxication aiguë*.

Pour l'appariement des cas cliniques et des discussions, le modèle vectoriel s'est montré performant ; plus de 88% des couples sont correctement appariés (cf. TABLE 2). En analysant les appariements incorrects, nous avons constaté que la plupart ont des scores de similarité très faibles (68% ont un score de similarité inférieur à 0,20).

6 Conclusion

Dans ce papier, nous avons présenté les méthodes que notre équipe a développées pour participer aux tâches 1 et 2 du DEFT 2019. Pour l'indexation des cas cliniques, une méthode basée sur la pondération TF.IDF a été proposée tandis que pour l'appariement des cas cliniques aux discussions, nous avons utilisé le modèle vectoriel. Comparée aux résultats globaux, notre méthode d'indexation a obtenu des performances encourageantes et nécessite d'être améliorée. Nous envisageons d'exploiter des ressources sémantiques pour surmonter des limites soulignées dans la section 5. Notre méthode d'appariement, quant à elle, a donné des résultats prometteurs que nous envisageons également d'améliorer en explorant l'utilisation des plongements lexicaux sur de gros corpus.

Remerciements

Nous remercions les organisateurs du DEFT 2019.

Références

BLEI D., NG A. Y. & JORDAN M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res*, p. 993–1022.

DEERWESTER S., DUMAIS S., FURNAS G., LANDAUER T. & HARSHMAN R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 41 (6), p. 391–407.

DRAME K., MOUGIN F. & DIALLO G. (2016). Large Scale Biomedical Texts Classification: A KNN and an ESA-Based Approaches. *J. Biomedical Semantics* 7: 40.

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). Recherche et Extraction d'information Dans Des Cas Cliniques. Présentation de la Campagne d'évaluation DEFT 2019. In *Actes de DEFT*, Toulouse, France.

HUANG M., NEVEOL A. & LU Z. (2011). Recommending MeSH Terms for Annotating Biomedical Articles. *Journal of the American Medical Informatics Association* 18 (5), p. 660–67.

LE Q. & MIKOLOV T. (2014). Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning* - Volume 32, p. II–1188–II–1196.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations*.

READ J., PFAHRINGER B., HOLMES G. & FRANK E. (2011). Classifier Chains for Multi-Label Classification. *Machine Learning* 85 (3): 333.

SALTON G., WONG A. & YANG C. S. (1975). A Vector Space Model for Automatic Indexing. In *ACM 18* (11), p. 613–620.

SCHUHMACHER M. & PONZETTO S. P. (2014). Knowledge-Based Graph Document Modeling. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, p. 543–552, NY, USA.