Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Participation de Berger-Levrault (BL.Research) à DEFT 2021 : de l'apprentissage des seuils de validation à la classification multi-labels de documents

Mokhtar Boumedyen Billami, Lina Nicolaieff, Camille Gosset, Christophe Bortolaso
Berger-Levrault, 64 Rue Jean Rostand, 31670 Labège, France
{mb.billami, lina.nicolaieff, camille.gosset,
christophe.bortolaso}@berger-levrault.com

RESUME

Cet article présente notre participation à l'édition 2021 du DÉfi Fouille de Textes (DEFT) et plus précisément à la première tâche liée à l'identification du profil clinique du patient. Cette tâche consiste à sélectionner, pour un document décrivant l'état d'un patient, les différents types de maladies rencontrées correspondant aux entrées génériques des chapitres du MeSH (*Medical Subject Headings*). Dans notre travail, nous nous sommes intéressés aux questions suivantes : (1) Comment améliorer les représentations vectorielles de documents, voire de classes ? (2) Comment apprendre des seuils de validation de classes ? Et (3) Une approche combinant apprentissage supervisé et similarité sémantique peut-elle apporter une meilleure performance à un système de classification multi-labels ?

ABSTRACT

Berger-Levrault (BL.Research) submission to DEFT 2021: from learning validation thresholds to multi-label document classification.

This article presents the participation of Berger-Levrault team to the DEFT's 2021 challenge, and more precisely to the first task related to the identification of the patient's clinical profile. This task consists in selecting, for a document describing a patient's health status, the different types of diseases encountered corresponding to the generic entries of the MeSH (Medical Subject Headings) chapters. In our work, we were interested in the following questions: (1) How to improve vector representations of documents, or even classes? (2) How to learn class validation thresholds? And (3) Can an approach combining supervised learning and semantic similarity bring a better performance to a multi-label classification system?

MOTS-CLES : Apprentissage supervisé, Représentation sémantique de classes, Similarité sémantique, Réentraînement de plongements lexicaux, MeSH.

KEYWORDS: Supervised learning, Semantic representation of classes, Semantic similarity, Fine-Tuning, MeSH.

1 Introduction

L'édition 2021 du DÉfi Fouille de Textes (DEFT) (Grouin et al., 2021) est consacrée à trois tâches différentes, à savoir : (1) l'identification du profil clinique du patient ; (2) l'évaluation automatique

de copies d'étudiants d'après une référence existante ; et (3) la poursuite automatique de la correction d'après de premières corrections. Berger-Levrault s'est fortement intéressée à participer à la première tâche dont l'enjeu scientifique est la classification de cas cliniques. La direction BL.Research a tenu à proposer des systèmes de classification multi-labels traitant des données provenant du domaine de la santé. Nous avons appelé notre équipe BL.Santé pour faire référence aux données cliniques.

La première tâche de DEFT 2021 s'inscrit dans la continuité des deux éditions précédentes (Cardon et al., 2020 ; Grabar et al., 2019), à savoir : le traitement des cas cliniques rédigés en français (descriptions de situations cliniques rares utilisées à des fins pédagogiques, scientifiques ou thérapeutiques). L'édition DEFT 2019 s'est concentrée sur la recherche et l'extraction d'informations (*âge*, *genre*, *origine* et *issue*) à partir de documents. L'édition 2020, quant à elle, s'est poursuivie en partie sur la tâche d'extraction d'information avec de nouveaux types autour des patients (*anatomies*), de la pratique clinique (*examen*, *pathologie*, *signe* ou *symptôme*), des traitements médicamenteux et chirurgicaux (*substance*, *dose*, *durée*, *fréquence*, *mode d'administration*, *traitement* (*chirurgical* ou *médical*) et *valeur*) et autour du temps (*date* et *moment*). Ces différentes informations ont été proposées comme annotations dans les corpus de données de DEFT 2021.

Les données DEFT 2021 proviennent d'un ensemble plus vaste composé de cas cliniques, porteur d'annotations. Les cas cliniques sont anonymes et couvrent différentes spécialités médicales (*cardiologie, urologie, oncologie, obstétrique, pulmonaire, gastro-entérologie, etc.*). Ils décrivent des cas qui se sont produits dans différents pays francophones (France, Belgique, Suisse, Canada, pays africains, pays tropicaux, etc.). L'objectif principal que nous nous fixons pour la première tâche de DEFT 2021 est le suivant : pour un cas clinique donné, nous nous intéresserons à identifier le profil clinique du patient concerné par le type de maladie de toutes les pathologies présentes dans le cas.

Après avoir présenté en section 2 les corpus d'apprentissage et de test de DEFT 2021, nous décrivons notre méthodologie de classification multi-labels en section 3. À ce stade, nous présentons différents systèmes que nous avons proposés lors de la campagne. Par la suite, dans la section 4, nous décrivons les résultats d'évaluation avant de conclure en section 5.

2 Corpus de données

Le corpus se compose de cas cliniques décrits dans un format textuel. Il regroupe des documents pour l'apprentissage (cf. sous-section 2.1) et d'autres documents pour le test (cf. sous-section 2.2). Chaque corpus est accompagné d'annotations provenant des deux éditions précédentes de DEFT, avec 25 types au total pour l'apprentissage (*poids*, *taille*, *changement*, *état*, *prise*, *AnnotatorNotes*, *date*, *âge*, *origine*, *durée*, *fréquence*, *issue*, *norme*, *mode*, *genre*, *dose*, *pathologie*, *moment*, *assertion*, *traitement*, *valeur*, *substance*, *examen*, *anatomie* et *sosy*) et 28 types pour le test dont 3 nouveaux (*température*, *organisme* et *fonction*). Par la suite, nous présentons dans la sous-section 2.3 une analyse comparative entre les deux corpus. Pour le contenu textuel des corpus, le lecteur peut consulter le travail mené par Grabar et al. (2018) pour plus de détails.

2.1 Corpus d'apprentissage

Ce corpus regroupe 167 documents. Pour l'ensemble des 25 types d'annotation, nous avons 15 802 instances. Par exemple, *rachianesthésie* est une instance pour le type *traitement*, *épileptique* est une instance pour *pathologie*, voire *légère somnolence* pour *sosy*. Le corpus d'apprentissage a la particularité d'avoir de nouvelles annotations par rapport aux éditions précédentes. En effet, nous

disposons pour DEFT 2021 de nouvelles instances associées aux chapitres du MeSH. Nous appellerons ces instances dans ce qui suit par "termes-clés". Par exemple, *pyurie itératives* est un terme-clé du chapitre *infections*, *plombémie élevée* pour *chimiques*, *adénocarcinome à cellules claires* pour *tumeur*, voire *lombalgie gauche* pour *etatsosy*. Il est à noter que ce corpus d'apprentissage propose des instances d'annotation pour 23 chapitres du MeSH. La FIGURE 1 présente le nombre d'instances/occurrences et la taille du vocabulaire (termes-clés uniques) associés à chaque chapitre du MeSH. Le corpus d'apprentissage dispose de 2 115 instances de chapitres au total pour avoir 17 917 annotations tout type confondu.

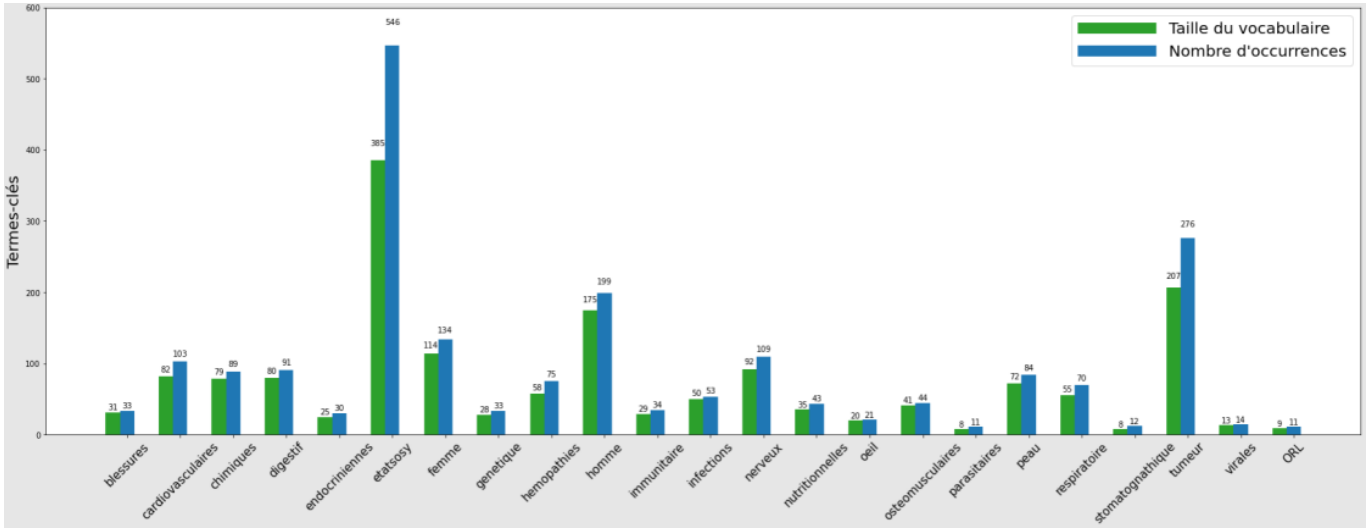

FIGURE 1: Distribution du nombre de termes-clés par chapitre du MeSH

Par ailleurs, le nombre d'annotations en chapitres dans ce corpus est 773. Sur l'ensemble des 167 documents, nous avons en moyenne 4,63 (≈ 5) classes (chapitres) par cas clinique. La classe la plus fréquente dans les documents est *etatsosy* avec 141 cas sur 167 contre la classe la moins fréquente *stomatognathique* avec 3 cas sur 167.

2.2 Corpus de test

Ce corpus regroupe 108 documents. Contrairement au corpus d'apprentissage, nous ne disposons pas d'instances de chapitres. Pour les annotations fournies dans ce corpus, nous avons 9 856 instances pour 28 types. Le type le plus fréquent est *sosy* avec 2 127 annotations contre le type le moins fréquent *état* avec 2 instances. En prenant en considération les chapitres de référence, nous constatons que sur l'ensemble des 108 documents, nous avons en moyenne 5,01 (≈ 5) chapitres par cas clinique. Cela revient à la même moyenne en comparaison avec le corpus d'apprentissage.

2.3 Analyse comparative des deux corpus

Dans cette sous-section, nous présentons une analyse comparative de la distribution des documents par chapitre entre apprentissage et test. En prenant en considération les annotations de référence des deux corpus, nous illustrons dans la figure FIGURE 2 le nombre d'exemples fournis pour chaque chapitre. Nous constatons qu'à l'exception des chapitres *ostéomusculaires* et *stomatognathique* avec un nombre d'exemples équitables entre apprentissage et test, la plupart du temps, nous avons plus d'exemples en apprentissage.

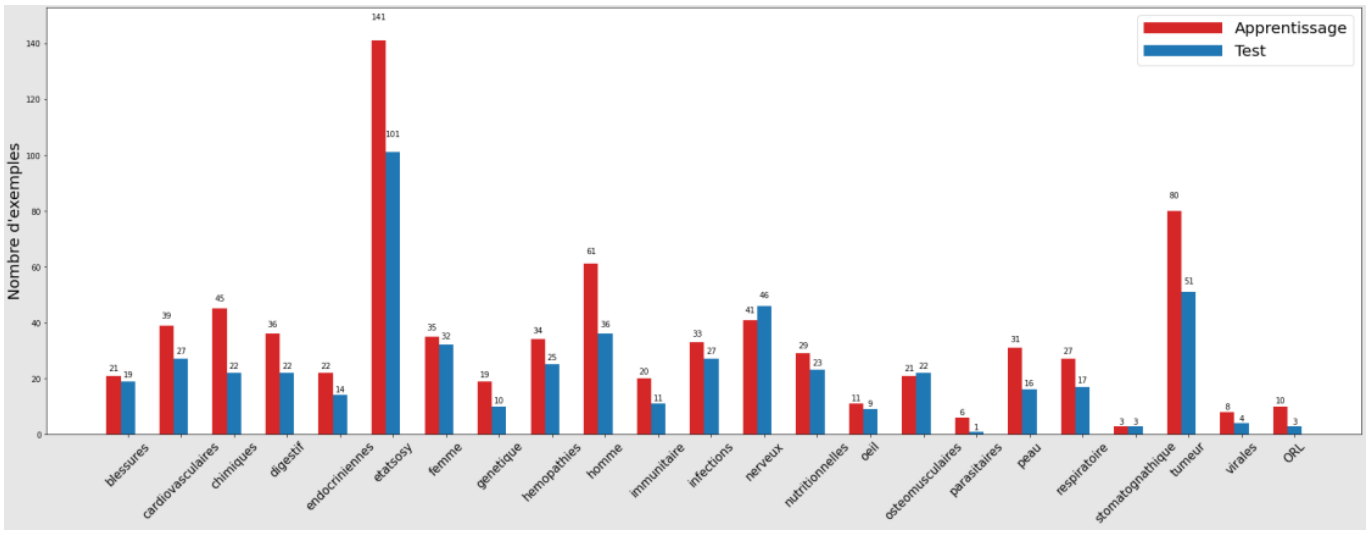

FIGURE 2: Distribution du nombre d'exemples (documents) par chapitre entre apprentissage et test

Le chapitre le mieux représenté en apprentissage est *etatsosy* mais cela revient aussi du fait qu'il est couvert par 84,4 % du corpus. Le chapitre *tumeur*, quant à lui, est couvert par 47,9 % du corpus d'apprentissage contre 47,22 % pour le corpus du test. Les trois chapitres *stomatognathique*, *parasitaires* et *virales* sont les moins représentés dans les deux corpus, avec moins de 10 exemples pour chacun.

3 Méthodologie

Dans cette section, nous présentons deux approches différentes pour répondre à la tâche de classification automatique multi-labels. La première approche (cf. sous-section 3.1) consiste à utiliser des plongements lexicaux provenant du domaine général et les réentraîner sur un corpus de spécialité, c'est-à-dire, le corpus d'apprentissage de DEFT 2021. Cette même approche consiste à créer des représentations sémantiques de documents, et de classes, et permet d'apprendre des seuils de validation pour valider les rapprochements sémantiques entre les documents et les classes. La deuxième approche (cf. sous-section 3.2), quant à elle, consiste à utiliser une succession de classificateurs binaires basés sur des représentations vectorielles de sacs de mots (*Bag-of-Words*). Nous proposons ensuite une combinaison des deux approches pour augmenter la couverture des classes non prédites par l'une des deux approches (cf. sous-section 3.3).

Par la suite, nous présentons dans cette même section une méthode d'extraction de terme-clés pour des documents provenant du corpus de test et faisant référence à des instances pour les chapitres du MeSH (cf. sous-section 3.4). Cette méthode repose principalement sur l'utilisation d'expressions régulières apprises à partir du corpus d'apprentissage.

3.1 Réentraînement de plongements lexicaux standards et apprentissage des seuils de validation

Cette première approche consiste à utiliser principalement des plongements lexicaux (*Word Embeddings*) et permet d'apprendre des seuils de validation des chapitres pour des cas cliniques. La figureFIGURE 3 présente l'architecture globale de notre approche. Dans l'ensemble du processus d'apprentissage, 8 étapes sont essentielles.

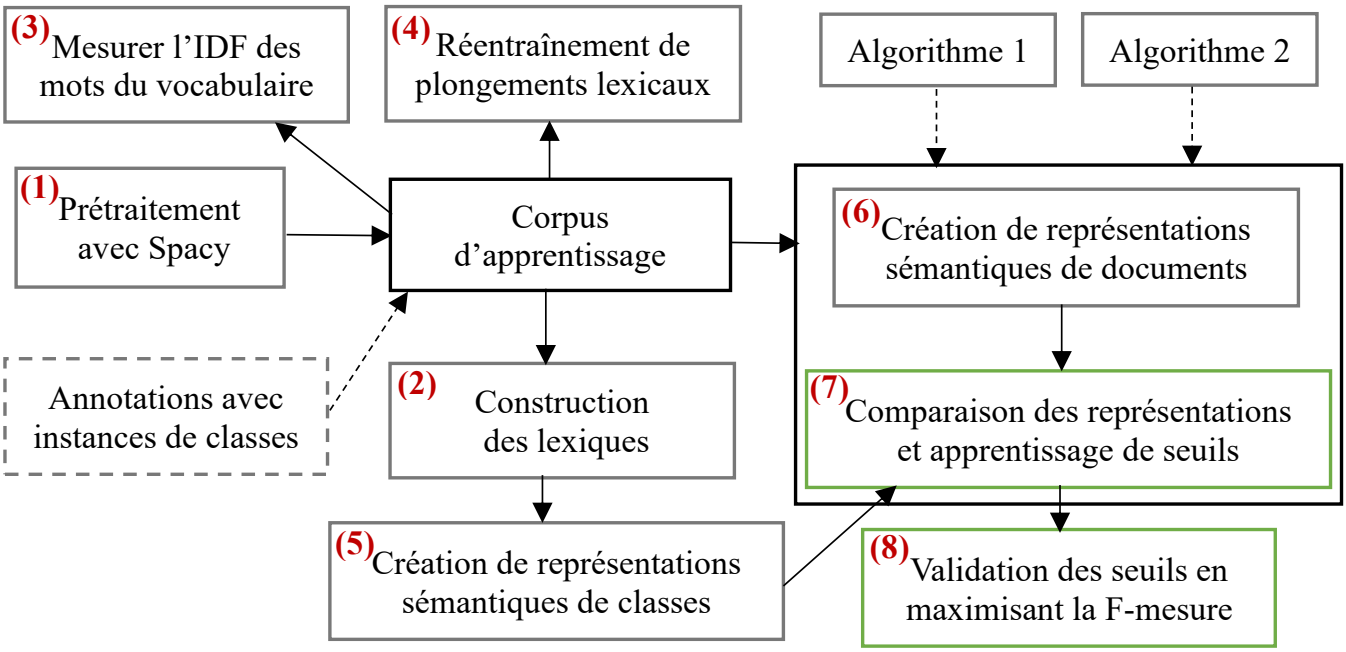

FIGURE 3: Architecture globale du premier système – apprentissage des seuils

Le principe de cette approche consiste à projeter dans un même espace vectoriel les documents et les classes. Pour cela, des vecteurs de documents et des vecteurs de classes sont créés. Nous détaillons ci-après chaque étape du processus.

- **Prétraitement des données :** Avec l'utilisation d'une chaîne de traitement du langage naturel comme Spacy[1] (Honnibal et al., 2020), nous prenons en considération seulement les mots portant du sens (noms, adjectifs, adverbes et verbes) dans leur format lemmatisé. Ce prétraitement est effectué sur le contenu textuel des documents et est pris en considération lors de l'utilisation des annotations.
- **Construction des lexiques :** nous prenons en considération uniquement les annotations associées aux instances de chapitres. Ces annotations, considérées comme termes-clés, alimentent des lexiques. Concrètement, nous construisons un lexique pour chaque classe (chapitre).
- **Mesurer la fréquence inverse du document pour chaque mot :** la fréquence IDF, *Inverse Document Frequency,* (Jones, 1972) est calculée sur le corpus d'apprentissage. Elle sert principalement à apporter des pondérations dans la création des représentations sémantiques de documents et classes (cf. étapes 5 et 6).
- **Réentraînement de plongements lexicaux :** à cette étape, nous prenons en considération un modèle Word2Vec (Mikolov et al., 2013) déjà entraîné (Fauconnier, 2015)[2] sur un corpus de grande taille, à savoir : FRWAC (Baroni et al., 2009), *The WaCky wide web* pour le français. Ce modèle est de type CBOW (*Continuous Bag of Words*) avec 500 dimensions pour la taille des vecteurs *embeddings*. Nous effectuons ensuite un réentraînement de type *fine-tuning* sur le corpus d'apprentissage à l'aide de la bibliothèque Gensim[3] tout en gardant un vocabulaire plus riche. À cette étape, le corpus est fourni dans un format prétraité avec Spacy. Par exemple, « ... *Le matin du jour trois, étant donné la persistance des nausées, des*

1 https://spacy.io/
2 http://fauconnier.github.io/#data
3 https://radimrehurek.com/gensim/models/word2vec.html

vomissements et de l'hypersalivation, l'équipe traitante augmente l'ondansétron à … » est considéré de la façon suivante « … *matin jour donner persistance nausée vomissement hypersalivation équipe traitant augmenter ondansétron …* ». Le nombre d'*epochs* a été fixé à 500. Cela n'a pas posé de problèmes pour les temps de calcul puisque la taille du corpus d'apprentissage reste relativement petite.

- **Création de représentations sémantiques de classes (chapitres) :** à partir du lexique construit pour chaque classe, nous créons des vecteurs centroïdes (moyens) pondérés avec l'IDF. Les vecteurs de mots proviennent du modèle fine-tuné.
- **Création de représentations sémantiques de documents :** dans le même principe que l'étape précédente, des vecteurs moyens pondérés sont créés en prenant en considération seulement les mots pleins des documents. Il est à noter qu'en phase de test, et dans le cas où de nouveaux mots apparaissent, l'IDF est considéré à une valeur égale à 1. Cela est dans le but de prendre en considération les vecteurs de mots du corpus de test non reconnus dans le corpus d'apprentissage. Dans cette étape, nous créons deux types de représentation sémantique pour les documents : (1) soit plusieurs vecteurs, chacun est associé à un paragraphe donné du document à traiter (cf. algorithme 1), (2) soit un vecteur représentant le texte intégral (cf. algorithme 2). Nous avons utilisé deux algorithmes dans la même approche afin d'associer le meilleur des deux pour chaque classe (chapitre).
- **Comparaison des représentations sémantiques, apprentissage et validation des seuils :** selon l'algorithme utilisé, nous comparons les vecteurs de classes avec les vecteurs de documents. Formellement, l'équation 1 ci-après fait référence au premier algorithme et l'équation 2 fait référence au deuxième algorithme.

$$Sim_1(C_i, Doc_j) = \underset{p \in paras\,(Doc_j)}{argmax} \; (1 - DistanceCosinus\,(Vec_{C_i}, Vec_p)) \quad (1)$$

$$Sim_2(C_i, Doc_j) = 1 - DistanceCosinus\,(Vec_{C_i}, Vec_{Doc_j}) \quad (2)$$

Avec Sim_1 et Sim_2 comme fonctions de similarité sémantique entre un chapitre C_i et un document Doc_j ; $paras$ représente l'ensemble des paragraphes d'un document Doc_j, Vec_{C_i} le vecteur du chapitre C_i, Vec_p le vecteur du paragraphe p et Vec_{Doc_j} le vecteur du document Doc_j.

Étant donné les annotations de référence du corpus d'apprentissage, nous avons appris les seuils de validation pour chaque classe permettant d'obtenir les meilleurs scores de F-mesure (mesure détaillée en section 4.1). Pour cela, nous avons varié les seuils de $0{,}1$ à 1 par pas de $0{,}01$.

Le Table 1 présente les seuils appris pour chaque algorithme et les meilleurs scores de F-mesure obtenus. Pour le premier algorithme, nous constatons que le seuil de validation minimal revient à 0,58 (pour *etatsosy*), à l'exception de la classe *tumeur* où le seuil est 0,46. Pour le deuxième algorithme, les seuils sont plus élevés puisque le minimum revient à 0,61 pour la classe *etatsosy*. De plus, dans la plupart du temps et pour chaque chapitre, les seuils pour le deuxième algorithme sont plus élevés à l'exception des classes *ostéomusculaires*, *stomatognathique* et *ORL*. Par ailleurs, l'utilisation du texte intégral permet d'avoir de meilleures scores pour la F-mesure à l'exception des classes *génétique, respiratoire, virales* et *ORL*. Ainsi, nous avons pris le choix d'utiliser l'algorithme 1 pour ces 4 classes et l'algorithme 2 pour toutes les autres classes.

Chapitre	Paragraphe le plus proche		Texte intégral	
	Best Seuil	**F-mesure**	**Best Seuil**	**F-mesure**
blessures	0,74	0,345	0,75	0,478
cardiovasculaires	0,65	0,505	0,70	0,553
chimiques	0,64	0,727	0,68	0,737
digestif	0,69	0,416	0,77	0,506
endocriniennes	0,74	0,444	0,76	0,489
etatsosy	0,58	**0,916**	0,61	**0,936**
femme	0,64	0,418	**0,80**	0,467
génétique	**0,79**	0,364	**0,80**	0,323
hémopathies	0,67	0,488	0,72	0,558
homme	0,65	0,618	**0,80**	0,662
immunitaire	0,65	0,545	0,67	0,558
infections	0,66	0,449	0,70	0,467
nerveux	0,59	0,466	0,68	0,545
nutritionnelles	0,68	0,514	0,73	0,549
œil	0,73	0,333	0,78	0,429
ostéomusculaires	0,77	0,32	0,72	0,333
parasitaires	0,70	0,222	0,74	0,364
peau	0,73	0,481	0,76	0,515
respiratoire	0,65	0,458	0,69	0,412
stomatognathique	0,69	0,667	0,67	0,667
tumeur	0,46	0,721	0,71	0,836
virales	0,69	0,5	0,70	0,471
ORL	0,70	0,333	0,64	0,235

TABLE 1 : Apprentissage du seuil de validation pour chaque chapitre du MeSH et présentation de la meilleure F-mesure obtenue

Toutefois, deux classes peuvent porter une certaine ambigüité, à savoir : (1) *Maladies urogénitales de l'homme (homme)* et *Maladies de l'appareil urogénital féminin et complications de la grossesse (femme)*. Afin de différencier ces deux classes et dans le cas où les deux classes sont prédites, nous utilisons une stratégie d'identification du genre. Concrètement, nous avons pris les annotations associées au type *genre* du corpus d'apprentissage. Le chapitre *homme* est validé pour un cas clinique donné seulement et seulement si aucun genre féminin n'est identifié. Cela est le cas aussi pour le chapitre *femme* à sa validation, c'est-à-dire, aucun genre masculin n'est identifié.

3.2 Apprentissage supervisé par utilisation de représentations vectorielles à base de sacs de mots (*Bag-of-Words*)

Cette deuxième approche consiste à utiliser des modèles d'apprentissage supervisé pour satisfaire le besoin de la classification multi-labels. Nous proposons d'utiliser plusieurs classificateurs binaires, chacun pour prédire un chapitre donné du MeSH. Ainsi, nous avons 23 classificateurs. Plusieurs

modèles de classification ont été testés, à savoir : la régression logistique, la classification naïve bayésienne, un classificateur de forêts aléatoires (*Random Forest Classifier*, RFC), voire un classificateur linéaire par vecteurs de support faisant partie de la famille des machines à vecteurs de support (SVM). Pour les représentations vectorielles de documents, nous nous sommes intéressés aux vecteurs TF-IDF (*Term Frequency-Inverse Document Frequency*) (Jones, 1972) et sacs de mots (*Bag-of-Words*). L'avantage de ces types de représentation, en comparaison aux représentations par plongements lexicaux, est qu'ils nous permettent de nous affranchir du réentraînement d'un modèle de langage, sur le vocabulaire spécifique médical.

En utilisant seulement le corpus d'apprentissage, et afin de sélectionner le meilleur modèle, nous avons mis en place une stratégie de validation basée sur 5 jeux de tests construits en gardant une même répartition des classes présentées dans ce corpus. L'équilibrage des classes dans les 5 jeux de données a été réalisé à l'aide la librairie *Scikit-multilearn*[4] spécialisée dans la classification multi-labels et basée sur Scikit-learn (Buitinck et al., 2013). Le nombre de classes prédites correspond au nombre de classes différentes en sortie du système. Le meilleur modèle aura ainsi un juste équilibre entre la F-mesure et une bonne répartition des classes dans les prédictions. Le tableau 2 présente les résultats obtenus pour la validation du modèle le plus pertinent à sélectionner.

Modèle	**F-mesure**		**Nombre de classes prédites**	
	TF-IDF	**Sac de mots**	**TF-IDF**	**Sac de mots**
Régression Logistique	**0,78**	0,61	4	**20**
Naïve Bayes	0,74	0,64	5	16
RFC	0,72	**0,66**	7	9
SVM	0,57	0,57	**15**	19

TABLE 2 : Résultats des performances de chacun des modèles testés en fonction de la représentation vectorielle choisie

L'approche ayant retenu notre attention par l'obtention des meilleurs validations combine des représentations de sac de mots avec un modèle de régression logistique. Cette méthode offre une meilleure diversité dans les classes prédites ainsi qu'une F-mesure considérable. La figure 4 illustre ainsi l'architecture globale de ce deuxième système.

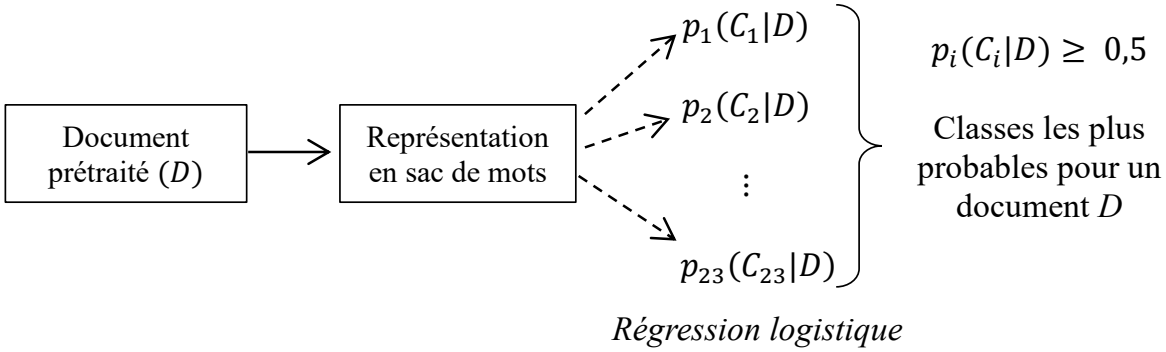

FIGURE 4: Architecture globale du deuxième système – apprentissage supervisé

Tout d'abord, nous tenons à mentionner qu'un prétraitement a été effectué sur tout le contenu textuel des documents de façon à ne garder que les mots portant du sens. Cela dit, les caractères spéciaux et

[4] http://scikit.ml/

les mots-vides (*stopwords*) sont supprimés. La fonction de perte de la régression logistique est optimisée à l'aide du solveur *liblinear*[5] et la diversité des classes prédites est améliorée grâce à la régularisation $L1$. Formellement, l'équation 3 permet de mesurer la probabilité d'associer la classe C_i au document D :

$$p_i(C_i|D) = \frac{e^{\beta_K X}}{1 + e^{\beta_K X}} \quad (3)$$

Avec k le nombre de mots retenus pour construire le dictionnaire servant à l'élaboration des vecteurs de sacs de mots (*Bag-of-Words*), X la matrice contenant l'ensemble des vecteurs représentant les documents et β les coefficients de régression estimés par le modèle. Le modèle permet d'estimer une probabilité de la classe C_i pour le document D. Par exemple, si le système offre une probabilité de 0,6 alors le document à 60 % de chance d'appartenir à cette classe.

3.3 Approche hybride : et si nous tenions compte des deux premiers systèmes

Pour un troisième système, nous nous sommes intéressés à prendre le cumul des deux premiers systèmes. Concrètement, nous nous intéressons ici à mieux couvrir les classes non prédites par l'un des deux systèmes. En effet, cela permettra d'augmenter la mesure du rappel (cf. sous-section 4.1). Toutefois, le risque d'une diminution de la mesure de précision (cf. sous-section 4.1) reste présent. Les résultats obtenus sur le corpus de test sont discutés dans la sous-section 4.2.

3.4 Extraction de termes-clés et association « Terme-Clé – Chapitre du MeSH »

Afin de satisfaire le besoin d'associer un chapitre à un terme-clé se trouvant dans un cas clinique, nous proposons une méthode d'extraction de termes-clés et d'association d'un terme-clé à un chapitre donné. Cette méthode s'inspire de l'utilisation des lexiques que nous avons créés et détaillés dans les sections précédentes.

Le principe de notre méthode est d'identifier automatiquement des termes-clés similaires à ceux des lexiques. Pour cela, nous utilisons les expressions régulières. L'utilisation de ces expressions offre plusieurs avantages. Tout d'abord, cela permet d'englober toutes les formes fléchies d'un terme-clé, c'est-à-dire, avoir la possibilité de récupérer un terme-clé sous différentes formes. En effet, la suite de termes-clés peut-être sous forme singulière ou plurielle (par exemple, *kyste pyélogénique* ou *kystes pyélogéniques*). Aussi, on peut récupérer plusieurs temps de conjugaison pour un verbe donné. Cela permet de retrouver dans un texte une suite de termes-clés au présent comme au passé (par exemple, *perdu conscience* ou *perdre conscience*).

Par ailleurs, nous tenons à récupérer des motifs de phrases contenant des termes-clés. Par exemple, dans l'expression *10 fois la dose*, l'intérêt est de pouvoir retrouver des expressions du genre *N fois la dose* où N est un entier > 0. En effet, dans le corpus de test, si la dose est prescrite, elle ne sera pas forcément d'une parfaite égalité à 10 (comme dans le corpus d'apprentissage). Pour le traitement de ces cas, nous récupérons la forme lemmatisée à l'aide de Spacy (Honnibal et al., 2020) pour chacun des termes-clés. Le traitement des chiffres est effectué en utilisant l'expression régulière ([0-9](, ?))+. Cette expression signale qu'il est possible de récupérer une suite de chiffres séparables à tout moment par une virgule. Le TABLE 3 présente un exemple pour l'application de ces expressions régulières.

[5] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

Exemple tiré du corpus d'apprentissage	*10*	*fois*	*la*	*dose*
Expression régulière associée (concept)	([0-9](, ?))+	foi*	l*	dos*
Exemple concordant	5	fois	les	doses

TABLE 3 : Exemple de traitement de termes-clés avec des expressions régulières

Enfin, pour le traitement des majuscules, deux cas sont possibles : (1) nous pouvons avoir un terme complet en lettres capitales. Pour ce cas, l'expression régulière prend en compte la forme à la fois en capitale et en minuscule ; (2) nous pouvons avoir une majuscule seulement sur la première lettre des mots composant le terme-clé. Pour ce cas, l'expression régulière prend en compte la forme normalisée.

4 Résultats et discussion

Dans cette section, nous présentons tout d'abord les mesures d'évaluation recommandées par la campagne (cf. sous-section 4.1) avant de présenter les résultats obtenus (cf. sous-section 4.2).

4.1 Mesures d'évaluation

Souvent, le rappel (R), la précision (P) et la F-mesure (F) sont utilisés pour évaluer les performances des systèmes de classification automatique de textes. Le rappel permet de répondre à la question : Quelle proportion de résultats positifs réels est identifiée correctement ? La précision, quant à elle, répond à la question : Quelle proportion d'identifications positives est effectivement correcte ? La F-mesure est une moyenne harmonique du rappel et de la précision. Elle permet de mesurer la capacité d'un système à donner toutes les solutions pertinentes et à refuser les autres. Formellement, la description mathématique de ces mesures pour une étiquette de classe donnée C_i (cf. un chapitre C du MeSH), avec $C_i \in [1, 23]$, est présentée dans les équations suivantes :

$$P_{C_i} = \frac{TP_{C_i}}{TP_{C_i} + FP_{C_i}} \quad (4)$$

$$R_{C_i} = \frac{TP_{C_i}}{TP_{C_i} + FN_{C_i}} \quad (5)$$

$$F_{C_i} = \frac{2 \times P_{C_i} \times R_{C_i}}{P_{C_i} + R_{C_i}} \quad (6)$$

Dans les équations, 3 variables sont à déterminer, à savoir TP_{C_i}, FP_{C_i} et FN_{C_i} :

- TP_{C_i} représente le nombre de documents correctement classés pour la C_i-ième classe (vrais positifs)
- FP_{C_i} représente le nombre de documents qui sont incorrectement classés en C_i-ième classe (faux positifs)
- FN_{C_i} est le nombre de documents qui appartiennent à la C_i-ième classe, mais qui sont incorrectement classés (faux négatifs)

Pour le calcul des mesures sur toutes les classes (cf. mesures globales), la moyenne sur la somme est obtenue.

4.2 Résultats d'évaluation

Nous avons testé nos trois systèmes sur le corpus de test ayant 108 cas cliniques. Les résultats obtenus sont présentés dans le tableau 4. Nous nous comparons dans ce tableau avec le meilleur système ayant participé à la campagne DEFT 2021 (Grouin et al., 2021). Dans le tableau, le Run1 fait référence à notre premier système utilisant les plongements lexicaux et les seuils validés par apprentissage. Le Run2 fait référence au deuxième système à base d'apprentissage supervisé et utilisant la régression logistique. Le Run3 fait référence au troisième système utilisant la combinaison des deux premiers Runs.

Système	Rappel	Précision	F-mesure
Run1	0,677	0,570	0,619
Run2	0,471	**0,786**	0,589
Run3	**0,730**	0,558	**0,633**
Meilleur système	**0,750**	**0,885**	**0,812**

TABLE 4 : Résultats obtenus pour la classification multi-labels en chapitres du MeSH

Les résultats montrent que nous obtenons un bon rappel avec le Run3. Toutefois, le Run1 est le système ayant permis d'accroître cette bonne couverture. Nous constatons aussi que le Run2 possède un faible rappel. Cependant, il propose une meilleure précision que le Run1. Comme pressenti, la combinaison des deux (cf. Run3) permet seulement d'augmenter le rappel sans faire autant pour la précision. Néanmoins, ce Run3 permet d'avoir notre meilleur score pour la F-mesure. Par ailleurs, en comparaison avec le meilleur système de la campagne, le rappel du Run3 est proche avec un écart de 2 %. Pour la précision, l'écart avec le Run2 est près de 10 %. La figure 5 présente les résultats de F-mesure pour chaque chapitre du MeSH.

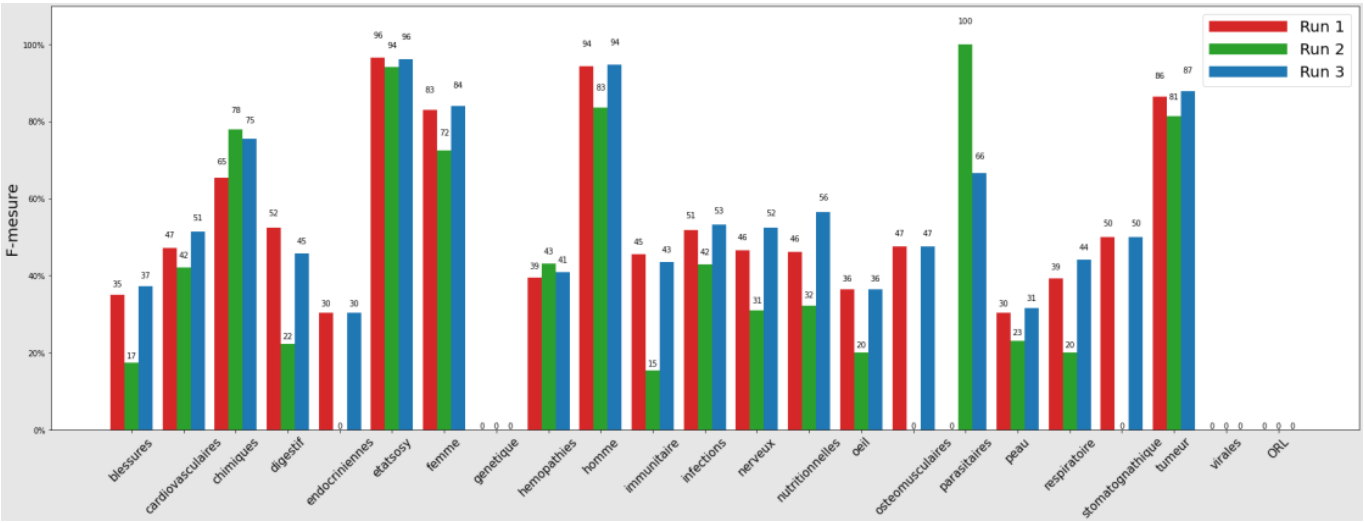

FIGURE 5: Résultats en F-mesure des trois systèmes (Runs) proposés pour chaque chapitre du MeSH

Sur ces résultats plus détaillés, nous constatons une certaine cohérence entre les validations effectuées sur le corpus d'apprentissage et les résultats obtenus en test. En effet, nous avons pour la prédiction du chapitre *etatsosy* une F-mesure de 96 % pour le Run1 et 94 % pour le Run2 (101 documents de référence sur 108). Les chapitres *femme* et *homme* sont aussi bien traités avec 83 % (*femme*, Run1) et

94 % (*homme*, Run1). Toutefois, le Run2 offre une bonne F-mesure pour certains chapitres comme *chimiques*, *hémopathies* voire *parasitaires*. Par ailleurs, aucun de nos systèmes n'a pu faire une identification des chapitres *génétique* (10 cas), *virales* (4 cas) ou *ORL* (3 cas). L'enrichissement des lexiques pour ces chapitres est une bonne piste afin d'améliorer les performances de nos trois Runs.

5 Conclusion

Notre participation à la campagne DEFT 2021 nous a permis de proposer deux approches totalement différentes pour satisfaire le besoin de la classification automatique multi-labels pour une application dans le cadre du domaine médical. De l'apprentissage des seuils et la comparaison de représentations sémantiques à la proposition d'un modèle d'apprentissage supervisé, nous avons testé ces techniques sur des données médicales. Nous constatons que ces méthodes peuvent être améliorées si nous enrichissons les lexiques construits, avec une utilisation de bases de connaissances (par exemple, des ontologies du domaine médical). En effet, si l'écart entre les représentations sémantiques de classes est grand, cela ne peut que diminuer le nombre de faux positifs et faux négatifs pour ainsi augmenter les valeurs de la précision et du rappel.

Remerciements

Nous tenons à remercier toutes les personnes ayant contribué à la réalisation de ce travail dans sa globalité. Nos remerciements vont tout particulièrement aux organisateurs de DEFT 2021 pour la disponibilité, la qualité et tout l'effort de l'annotation du corpus de travail.

Références

BARONI M., BERNARDINI S., FERRARESI A. & ZANCHETTA E. (2009). The WaCky wide web : a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, p. 209–226.

BUITINCK L., LOUPPE G., BLONDEL M., PEDREGOSA F., MUELLER A., GRISEL O., NICULAE V., PRETTENHOFER P., GRAMFORT A., GROBLER J., LAYTON R., VANDERPLAS J., JOLY A., HOLT B. & VAROQUAUX G. (2013). API design for machine learning software: experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, p. 108–122.

CARDON R., GRABAR N., GROUIN C. & HAMON T. (2020). Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, Nancy, France. p. 1–13. HAL : hal-02784737v3.

FAUCONNIER J.-P. (2015). French Word Embeddings. URL : http://fauconnier.github.io.

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. *DEFT 2019 - Défi fouille de texte*, Toulouse, France. p. 1–10. HAL : hal-02280852.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS: French Corpus with Clinical Cases. *LOUHI 2018 - The Ninth International Workshop on Health Text Mining and Information Analysis*, Bruxelles, France. p. 1–7. HAL : hal-01937096.

GROUIN C., GRABAR N. & ILLOUZ G. (2021). Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne DEFT 2021. *Actes de DEFT*. Lille.

HONNIBAL M., MONTANI I., VAN LANDEGHEM S. & BOYD A. (2020). spaCy: Industrial-strength Natural Language Processing in Python, *Zenodo*, DOI :10.5281/zenodo.1212303.

JONES K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, vol. 28, p. 11–21.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations*, p. 1–12.