

DOING@DEFT : utilisation de lexiques pour une classification efficace de cas cliniques

Nicolas Hiot, Anne-Lyse Minard, Flora Badin

▶ **To cite this version:**

Nicolas Hiot, Anne-Lyse Minard, Flora Badin. DOING@DEFT : utilisation de lexiques pour une classification efficace de cas cliniques. Traitement Automatique des Langues Naturelles, 2021, Lille, France. pp.41-53. hal-03265924

HAL Id: hal-03265924
https://hal.science/hal-03265924

Submitted on 23 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

DOING@DEFT : utilisation de lexiques pour une classification efficace de cas cliniques

Nicolas Hiot [1] Anne-Lyse Minard[2] Flora Badin[2]
(1) Université d'Orléans, LIFO, Orléans, France
(2) Université d'Orléans, LLL-CNRS, Orléans, France
nicolas.hiot@etu.univ-orleans.fr, anne-lyse.minard@univ-orleans.fr,
flora.badin@univ-orleans.fr

RÉSUMÉ

Nous présentons dans cet article notre participation à la tâche 1 de la campagne d'évaluation francophone DEFT 2021, sur l'identification du profil clinique du patient. Nous proposons une méthode évolutive et efficace en temps et en ressources pour la classification de documents médicaux pouvant être facilement adaptée à d'autres domaines de recherche. Notre système a obtenu les meilleures performances sur cette tâche avec une F-mesure de 0,814.

ABSTRACT

In this paper, we present our participation to the DEFT 2021 task 1. The task focuses on the identification of patient clinical profile. We propose a method that is upgradable and efficient in time and resources for medical document classification. The method can be easily adapted to other domains. Our system has obtained the best results on this task, with an F-measure of 0,814.

MOTS-CLÉS : cas clinique, transducteur fini, lexique, classification.

KEYWORDS: clinical case, final state transducer, lexicon, classification.

1 Introduction

Nous présentons dans cet article le système que nous avons développé pour l'identification du profil clinique du patient, et qui nous a permis de participer à la tâche 1 de DEFT 2021 [1] (Grouin *et al.*, 2021). La tâche consistait à identifier les types de maladies d'un patient décrit dans un cas clinique et correspondant aux entrées génériques du chapitre C du MeSH [2], un thésaurus bilingue anglais-français du domaine médical. Chaque document décrivant le cas clinique d'un patient, l'extraction des classes du MeSH peut être vue comme un problème de classification (pour chaque classe du MeSH on cherche à savoir si la classe est représentée dans le cas ou non).

L'idée était de concevoir un système capable de répondre au problème efficacement en temps et en ressources. Nous avons mis au point une méthode symbolique, basée sur des transducteurs finis et des lexiques. Le système développé peut être utilisé en temps réel, mis à jour facilement et à un coût environnemental faible. En effet, aucun apprentissage n'est effectué et les pré-traitements consistent

1. https://deft.lisn.upsaclay.fr/2021/
2. http://mesh.inserm.fr/FrenchMesh/index.htm

Atelier DÉfi Fouille de Textes (DEFT), pages 41–53.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

uniquement à tokeniser le texte et à raciniser les mots. Nous avons également fait le choix de ne pas utiliser les annotations manuelles disponibles dans le corpus de test pour pouvoir utiliser notre système sur n'importe quel texte brut. Nous avons uniquement utilisé les annotations manuelles en genre pour les cas pour lesquels notre méthode n'était pas en mesure de détecter le genre (2 sur 57). Nous montrons dans la section 3.4 que nous aurions pu nous en passer sans que les performances de notre système ne soient impactées. Notre système a obtenu les meilleurs résultats de la campagne DEFT 2021 pour la tâche 1.

Les organisateurs de DEFT nous ont fourni un corpus d'entraînement composé de 167 cas cliniques rédigés en français, pour un total de 69 256 mots. Les cas cliniques sont anonymes et couvrent différentes spécialités médicales (cardiologie, urologie, oncologie, obstétrique, pulmonaire, gastro-entérologie, etc.). Le corpus a été annoté manuellement sous BRAT en pathologies, signes ou symptômes, parties anatomiques, examen, substances, traitement, dose, mode, moment, fréquence, durée, valeur, etc. (e.g. *« échographie» de type «examen»*). Un enrichissement est effectué sur l'absence ou la présence de certains concepts, un changement, un état, une prise, etc. (e.g. *« diminution» de type «changement» et « possible» de type « assertion»*)

Ces cas sont des descriptions de situations cliniques rares utilisées à des fins pédagogiques, scientifiques ou thérapeutiques. Les classer automatiquement permettrait entre autres d'indexer les cas automatiquement et pourrait aider à l'identification de pathologies.

Ce travail a été réalisé dans le cadre du groupe de travail régional DOING [3], qui s'intéresse à la transformation des données en information, puis en connaissance, en favorisant la collaboration de chercheurs en TAL, en bases de données et en IA.

Après avoir fait un rapide état de l'art du domaine (section 2), nous présentons notre système en section 3, puis les résultats obtenus en section 4. Nous terminerons avec une partie discussion autour des choix d'implémentation du système et une analyse des erreurs (section 5).

2 État de l'art

La tâche d'identification de types de maladie s'assimile à une tâche de classification de documents, mais repose en partie sur l'extraction d'information dans le texte, plus particulièrement sur la reconnaissance d'entités. Cette dernière est une tâche très importante en TAL pour le domaine médical, elle est souvent associée à une tâche de linkage d'entités ou normalisation d'entités. Pour l'anglais nous pouvons par exemple citer les travaux sur la détection des maladies et des troubles et leur normalisation via les CUI (Concept Unique Identifier) de l'UMLS (The Unified Medical Language System) dans le cadre de différentes campagnes d'évaluation : ShARe/CLEF eHealth 2013 Evaluation Lab Task 1 (Pradhan *et al.*, 2013), SemEval 2015 tâche 4 (Elhadad *et al.*, 2015), etc. Le système qui a obtenu les meilleures performances à SemEval 2015 tâche 4 (Pathak *et al.*, 2015) utilise une approche supervisée basée sur des CRF (Conditional Random Fields) et des SVM (Support Vector Machine) pour l'identification des entités. Pour la normalisation des entités, ils recherchent d'abord l'entité identifiée dans l'UMLS, puis ils construisent automatiquement des variantes de ces entités et les recherchent, et enfin s'ils n'ont toujours pas trouvé le CUI ils calculent la similarité entre des

3. DOING (https://www.univ-orleans.fr/lifo/evenements/doing/) est un groupe de travail proposé en 2018 dans le cadre du réseau régional DIAMS (https://www.univ-orleans.fr/lifo/evenements/RTR-DIAMS/). En 2020, DOING était également un atelier du GdR MADICS et est devenu une action du GdR en 2021 https://www.madics.fr/actions/doing/.

chaînes proches dans l'UMLS. En français l'annotation d'entités dans des cas cliniques a fait l'objet de la tâche 3 de l'édition de DEFT 2020 (Cardon *et al.*, 2020) qui portait sur l'extraction des examens, des traitements, des signes ou symptômes, etc. (Minard *et al.*, 2020) ont proposé une méthode basée sur une cascade de CRF pour identifier ces informations. Malgré des performances dans l'ensemble plutôt bonnes, le système ne détecte correctement que la moitié des entités de type *signe ou symptôme* et *pathologie* (meilleure F-mesure respectivement de 0,55 et 0,44). Afin d'identifier les types de maladie, nous aurions pu nous baser sur ces deux types d'information, mais les performances ne semblent pas assez bonnes pour y arriver. Nous avons donc décidé de considérer la tâche comme une tâche de classification basée sur un lexique.

Dans la littérature, la classification de documents est souvent traitée comme un problème de classification traditionnelle. Étant donné des individus (documents) représentés sous forme vectorielle, l'objectif est de prédire la classe (un label) de chaque individu à partir d'un modèle qui a été entraîné sur un corpus d'exemples. Lorsque l'on traite des documents textuels, la représentation vectorielle ne parait pas évidente. La méthode souvent retenue est celle du sac de mots, l'idée est de sélectionner un ensemble de mots de taille n qui représenteront les dimensions de notre espace vectoriel. Ainsi, chaque document est représenté par un vecteur de booléen (représentant l'apparition du mot ou non dans le texte) ou un vecteur d'entier représentant le nombre d'occurrences des termes dans le texte. Afin de travailler sur un espace à dimensions réduites, une sélection des mots les plus révélateurs de chaque classe est nécessaire. L'idée est de trier chaque terme selon un certain critère et de sélectionner les m éléments ayant le score le plus élevé. Une mesure simple est le TF-IDF (pour term frequency-inverse document frequency) (Jones, 1972). Intuitivement, cette mesure est un ratio entre la fréquence d'apparition du terme dans une classe donnée et sa fréquence dans l'ensemble du corpus qui repose sur la loi de Zipf (un terme a plus de chance d'être révélateur d'une classe s'il y est souvent présent ; au contraire, si un terme est trop fréquent dans le corpus, il n'est pas assez discriminant). Dans (Weng *et al.*, 2017), les auteurs cherchent à extraire le domaine médical auquel appartiennent un ensemble de documents. Ils comparent une approche utilisant des réseaux de neurones et une approche utilisant des sacs de mots pondérés par TF-IDF avec un classifieur SVM. Ils montrent des résultats similaires pour les deux approches avec un gain en explicabilité pour l'approche sac de mots. D'autres mesures de pondération existent comme l'index de Jaccard. (Mihalcea & Tarau, 2004) proposent une autre méthode de pondération des termes basée sur l'algorithme de PageRank (Brin & Page, 1998) dans un graphe représentant les interactions entre les mots.

Nous avons fait le choix d'utiliser un transducteur fini et un lexique de "mots clés" plutôt que les méthodes décrites précédemment pour avoir une méthode utilisable « en temps réel » et utilisant peu de ressources. Défini dans la section 3.2, les transducteurs finis sont une forme d'automate fini qui reconnaissent un langage mais qui sont aussi capables de produire une sortie. Ils sont formellement définis comme des machines de Turing à deux rubans. A notre connaissance, (Gross, 1987) est le premier à introduire l'utilisation des transducteurs finis dans le traitement automatique de la langue naturelle. Les transducteurs finis peuvent être utilisés pour de l'analyse syntaxique (Briscoe & Carroll, 2002) mais aussi pour l'extraction d'entités (Gaio & Moncla, 2017). (Mihov & Maurel, 2000) ont introduit un algorithme permettant de construire un transducteur minimal à partir d'une liste triée de mots reconnus par le langage avec leur sortie. Cet algorithme est celui implémenté dans pour la construction des FST dans Apache Lucene, un moteur d'indexation de texte notamment utilisé par Apache SolR.

3 Système

Le système utilisé, présenté dans la figure 1, utilise un lexique permettant l'extraction des termes révélateurs de chaque classe (section 3.1). L'extraction est réalisée par un transducteur fini dont le principe général et la construction sont définis dans la section 3.2. Une phase de pré-traitement du texte est nécessaire avant l'extraction des termes. Elle permet de supprimer les mots vides et de normaliser les mots en récupérant leur racine. Un pré-traitement visant à transformer les mots en n-grammes a aussi été envisagé, mais a été jugé trop coûteux en ressources pour un gain de qualité trop faible.

Nous avons également mis en place une phase de post-traitement des annotations retournées par le transducteur afin de nettoyer les résultats obtenus. Elle se charge de gérer les négations (section 3.3) et de traiter l'ambiguïté engendrée par certains termes sur le genre (section 3.4).

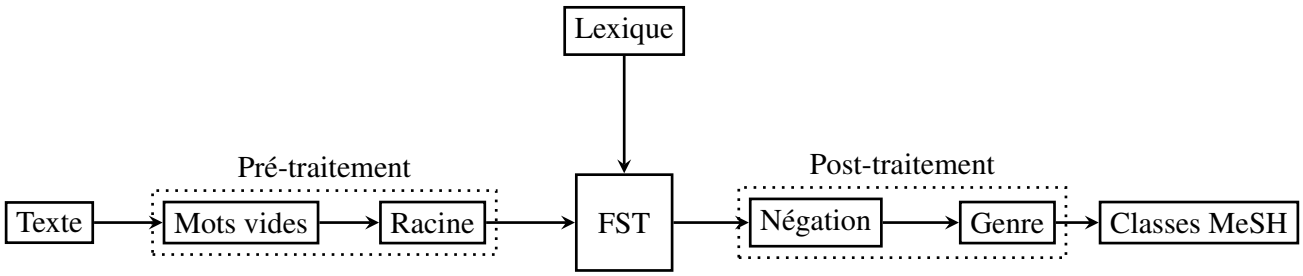


FIGURE 1 – Schéma du système d'identification du profil clinique du patient

3.1 Lexique

Définition Dans cet article nous définissons un lexique comme une collection de valeurs correspondant aux entrées du lexique pour lesquelles sont associées un ensemble de lexèmes la représentant. Dans notre cas les entrées du lexique sont les types de maladie représentés par 23 classes du chapitre C du MeSH (voir section 3.1.1).

Les lexiques sont construits à partir de thésaurus, de terminologies ou de bases de connaissances qui représentent des ressources riches qui s'enrichissent avec le temps (souvent semi-automatiquement à partir de corpus annotés) et qui sont très souvent surveillées par une autorité qui vérifie les informations et se charge du nettoyage des entrées. Le domaine médical ne fait pas exception et fait peut-être partie des plus représentés, notamment avec des institutions comme la NLM (U.S. National Library of Medicine) qui a regroupé, dans le méta-thésaurus UMLS, un grand nombre de ressources [4] pouvant être très utiles pour le traitement du langage naturel. Parmi les ressources de l'UMLS on peut notamment citer MeSH, MedDRA©, SNOMED et RxNORM. Notre méthode s'appuie sur les deux premiers décrits dans la suite de la section.

3.1.1 MeSH

Le MeSH (Medical Subject Headings) est un thésaurus du domaine biomédical, à l'origine en anglais, qui est géré par la NLM. Il permet entre autres d'indexer et d'interroger des bases de données comme

4. https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html

MEDLINE/PubMed. Une traduction du MeSH en français a été faite par l'INSERM, elle est mise à jour chaque année. La version bilingue anglais-français peut être interrogée depuis une interface en ligne (http://mesh.inserm.fr/FrenchMesh/), et il est également possible de télécharger le thésaurus au format XML. Le MeSH est organisé en 16 catégories thématiques, mais seule la catégorie C, maladie, a été utilisée pour cette tâche. Elle contient 26 classes, qui correspondent aux types de maladie à identifier dans les cas cliniques. Dans la table 1 nous donnons l'identifiant et le nom de ces classes (trois classes ne sont pas mentionnées dans le tableau car nous n'avions pas à les utiliser). Chaque classe est structurée en une arborescence de descripteurs, eux-mêmes constitués de concepts auxquels sont associés des termes. Nous avons extrait pour chaque classe du chapitre C tous les termes associés à des concepts. Au total 40 052 termes ont été extraits. Il est à noter qu'un même terme peut être associé à plusieurs classes. Par exemple "diabète gestationnel" est associé aux classes C13, C18 et C19.

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

TABLE 1 – Classes du chapitre C du MeSH

3.1.2 MedDRA©

Afin d'augmenter la couverture de notre lexique, nous avons cherché d'autres terminologies pour le français. Nous avons choisi d'utiliser MedDRA© puisque certaines de ses classes semblaient correspondre à celles du MeSH.

MedDRA© [5] (Dictionnaire Médical des Affaires Réglementaires) (Brown *et al.*, 1999) est un dictionnaire terminologique médical utilisé par les autorités réglementaires et l'industrie biopharmaceutique. MedDRA© est disponible en plusieurs langues, dont le français. Il contient aussi bien des termes référant à des symptômes, des examens ou encore des traitements, structurés en 5 niveaux. Le niveau le plus haut étant une classification par discipline médicale. Il existe 26 classes, par exemple *Affections vasculaires*, *Affections du rein et des voies urinaires*, *Affections du système immunitaire*.

Nous avons utilisé les termes de 13 classes de MedDRA©, par exemple *affections congénitales, familiales et génétiques*, *affections gastro-intestinales*, *affections de la peau et du tissu sous-cutané*.

5. La marque MedDRA© est enregistrée par l'IFPMA au nom du CIH. MedDRA© est développé par le Conseil International d'Harmonisation des exigences techniques pour l'enregistrement des médicaments à usage humain (CIH).

Nous avons relié ces classes aux classes du MeSH, dans les exemples précédents respectivement *congenitales* (C16), *digestif* (C06), *peau* (C17). Le lexique contient ainsi 58 071 termes en plus.

3.1.3 Corpus d'entraînement

Nous avons aussi utilisé le corpus d'entraînement pour supprimer ou ajouter des termes dans le lexique. Les résultats de notre système sur le corpus d'entraînement apportent une indication sur les valeurs repérées et les classes associées à ces valeurs. Il est repéré :

— les faux positifs (101 associations terme/classe) : termes qui ont toujours amené l'identification d'une mauvaise classe
hépatite B / virales ; mastite / femme ; plaie opératoire / blessures ; syphilis / infections ; morsure / chimiques

— les faux négatifs (1110 associations) : termes associés à une classe que notre système ne repère pas
inflexion épidermoïde / peau ; myélome / hemopathies ; tabagisme / chimiques ; cachectique / nutritionnelles

La liste des termes à supprimer a été nettoyée manuellement afin d'ignorer les termes les plus précis comme *accidents cérébrovasculaires*. En effet, nous souhaitons récupérer dans cette liste les termes ambigus ou trop génériques, par exemple *ampoule* pour la classe *peau*. Au total 54 termes ont été supprimés du lexique et 199 termes ont été ajoutés.

3.2 Transducteur fini (FST)

Afin d'extraire les lexèmes d'un lexique en « temps réel », tout en minimisant les ressources utilisées, nous proposons l'utilisation de transducteurs finis.

Définition En théorie des langages, un transducteur fini $T = (\Sigma^{in}, \Sigma^{out}, Q, I, F, \delta)$ est un automate fini qui reconnaît un langage $L = \{w_1, \ldots, w_n\}$ sur un alphabet Σ^{in} où les transitions possèdent deux labels ($l^{in} \in \Sigma^{in}$ et $l^{out} \in \Sigma^{out}$). Le premier caractérise la transition et le second constitue la sortie de l'automate. Q est l'ensemble des états, I les états initiaux, F les états finaux, δ l'ensemble des transitions et ϵ est le mot vide. La sortie de l'automate (quand un mot w est reconnu, c.-à-d. $w \in L$) peut être une somme des labels de sortie, leur concaténation ou, comme ici, une unique valeur (la classe). Un transducteur n'est pas obligatoirement déterministe et peut donc, pour un même mot, retourner plusieurs valeurs de sortie.

Les transducteurs sont des structures plus optimisées en mémoire que d'autres structures comme les tables triées, mais au détriment d'un accès plus coûteux en ressources processeur. Ils sont par conséquent, très utiles pour traiter des langages de grande taille qui ne pourraient pas normalement tenir en mémoire tout en offrant un accès suffisamment rapide. Nous utilisons l'implémentation fournie dans Apache SolR/Lucene [6] par le projet OpenSextant [7]. Elle repose sur l'algorithme de (Mihov & Maurel, 2000) qui permet d'obtenir le transducteur minimal efficacement.

Les transducteurs gardent aussi l'avantage d'être facilement mis à jour. Il est possible d'ajouter ou de supprimer de nouveaux mots dans le langage sans avoir à reconstruire l'automate entièrement.

6. https://solr.apache.org
7. https://github.com/OpenSextant/SolrTextTagger

Construction Comme présenté dans la Section 3.1, nos lexiques sont définis comme une application surjective $\forall v_i \in V \; \exists X_i \subset X, \; Lex : v_i \rightarrow X_i$ où V est l'ensemble des valeurs (ici classes du MeSH) du lexique et X est l'ensemble des lexèmes présents dans le lexique. X_i est défini comme l'ensemble des exemples de la valeur v_i, c.-à-d. pour MeSH, l'ensemble des lexèmes qui représente une classe du MeSH.

Afin de construire notre transducteur pour le lexique $Lex_{MeSH} : V_{MeSH} \rightarrow X_{MeSH}$, nous définissons les alphabets $\Sigma^{in}_{MeSH} = \{t_i \mid t_i \in \textit{token}(x_j) \; \forall x_j \in X_{MeSH}\} \cup \{\epsilon\}$ où *token* est une fonction qui retourne l'ensemble des tokens t_i d'un lexème x_j et $\Sigma^{out}_{MeSH} = V_{MeSH} \cup \{\epsilon\}$. Notre langage L_{MeSH} est alors naturellement défini comme l'ensemble des lexèmes X_{MeSH} du lexique, c.-à-d. chaque lexème est un mot du langage.

La fonction *token* permet l'extraction des tokens utilisés pour la construction du transducteur. Cette fonction est aussi appliquée aux textes en entrée afin de les faire correspondre à l'alphabet Σ^{in}_{MeSH}. Elle a pour rôle :

- Le découpage des lexèmes (mot ou suite de mots) en tokens. Le découpage est réalisé sur les caractères d'espacement, les ponctuations, les traits d'union et les chiffres accolés à du texte (ex : 50mg devient {50, mg}) ;
- Le passage en minuscule de l'ensemble des tokens ;
- Le filtrage des tokens correspondant à des mots vides (basé sur une liste) ;
- La transformation de tous les tokens non ASCII par leur équivalent (suppression des accents) ;
- Le remplacement de chaque token par sa racine en utilisant l'algorithme Snowball (Porter, 2001).

Détection des valeurs Notre système transforme le texte en entrée en une liste de tokens avec l'aide de la fonction *token*. La liste de tokens est ensuite passée dans le transducteur afin d'extraire l'ensemble des valeurs du lexique. Si plusieurs classes sont trouvées pour un même mot de L_{MeSH}, les multiples classes sont gardées. Cependant, si deux mots différents se recoupent (ex : *aggravation transitoire des symptômes* et *symptômes cardiovasculaires*) seulement le plus grand est gardé (ici, *aggravation transitoire des symptômes*). Les tokens restants (*cardiovasculaires*) sont remis en jeux (au cas où ils pourraient former un autre mot de L_{MeSH}). Cette approche permet de sélectionner les plus grands lexèmes qui sont plus discriminant de par leur taille.

Exemple 1. Prenons comme exemple trois lexèmes du MeSH avec des classes arbitraires dans la table 2. A partir des exemples pour chaque classe, nous pouvons construire le transducteur Figure 2.

Classe	Lexème	Tokens
C1	Exacerbation transitoire des symptômes	{exacerb, transitoir, symptom}
C1	Aggravation transitoire des symptômes	{aggrav, transitoir, symptom}
C2	Aggravation passagère des symptômes	{aggrav, passager, symptom}

TABLE 2 – Liste de lexèmes avec leur classe associée et la liste des tokens obtenus avec la fonction *token*

3.3 Détection de la négation et de l'incertitude

Dans les cas cliniques, il est possible que l'auteur notifie l'absence ou l'incertitude de certains symptômes. Bien que utile pour le corps médical, ces informations ne font pas partie du profil du

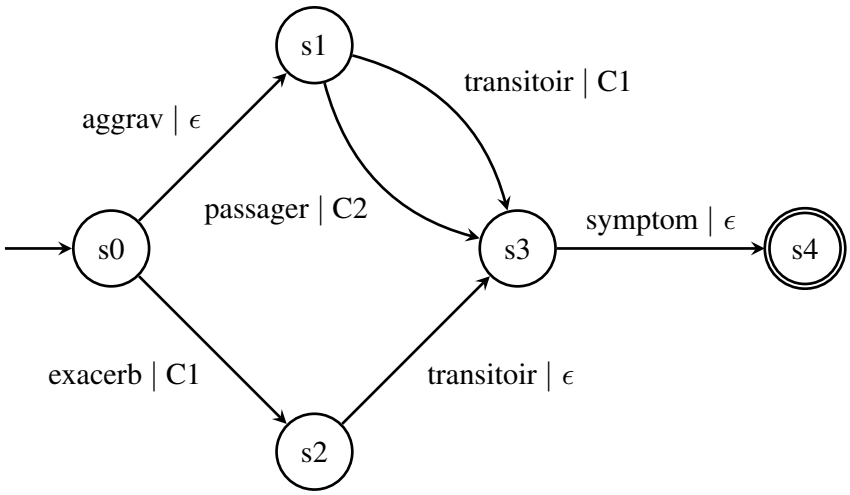


FIGURE 2 – Exemple d'un transducteur fini construit à partir de la table 2

patient. Dans la phrase « Le patient n'avait eu ni **traumatisme**, ni **piqûre d'insecte** » le MeSH nous permet de reconnaître les termes *traumatisme* et *piqûre d'insecte*. Cependant, ils doivent être ignorés pour la classification.

Pour ce faire, nous avons construit un nouveau lexique contenant l'ensemble des marqueurs de négation et d'incertitude extrait automatiquement à partir du corpus CAS (Grabar *et al.*, 2019). Une mesure de distance est ensuite appliquée entre chaque lexème extrait et les marqueurs de négation. Si la distance est inférieure à un certain seuil, l'exemple de la classe est rejeté. Nous utilisons ici la distance en offset de caractères (distance entre la fin du marqueur et le début du lexème). Nous avons testé des seuils de 2, 5, 10 et 20, et nous avons remarqué très peu de différences dans les résultats entre les seuils 2, 5 et 10 (entre 0,736 et 0,737 de F-mesure sur le corpus d'entraînement). Nous avons donc choisi de fixer le seuil à 10, pour améliorer la précision. (Garcelon *et al.*, 2014) propose une méthode plus évoluée pour réduire la porté de la négation, mais dans notre cas, ce n'est pas nécessaire car on recherche uniquement à classer des documents, et si un cas clinique possède une classe C, il y a une grande probabilité que d'autres exemples révélateurs de cette classe se trouvent ailleurs dans le document. Au contraire, un terme hypothétique a plus de chance de n'apparaître que peu de fois et donc d'être supprimé par cette approche. Toutefois, une simple mesure d'occurrences n'est pas suffisante, car certaines classes peuvent n'être reconnues qu'à l'aide d'un seul exemple isolé dans le document.

3.4 Recherche du genre du patient

Dans les cas cliniques du corpus d'entraînement, la classe du genre (homme, femme) est très présente. Cette classe est repérée pour les maladies liées à l'appareil urogénital et aux complications de grossesse. Notre système détecte l'information dans 98 cas sur 167 au total. Pour 62 d'entre eux, un double genre est affecté. Au vu de cette proportion, nous proposons un post-traitement déterminant le genre du patient pour chaque cas.

A l'aide du corpus d'entraînement nous établissons un lexique lié au genre comme « mlle », « madame », « âgée », « hospitalisée », « patiente », « monsieur », « homme », « masculin », etc. Ces mots sont repérés dans le premier paragraphe, le genre étant souvent indiqué en début de cas. Si nous n'avons pas repéré le genre à cette étape alors nous ajoutons au lexique une liste de termes plus

spécifiques à l'anatomie (« testicule », « utérus », « ovaire », etc.) et nous étendons la recherche aux autres paragraphes.

Dans le corpus d'entraînement, pour deux cas nous n'arrivons pas à déterminer le genre, pour l'un il s'agit de plusieurs personnes pour l'autre, aucun indice notable ne permet de le savoir. S'il s'agit de cas cliniques pour lesquels l'information du genre est nécessaire nous gardons l'annotation du transducteur.

Au niveau du corpus test, 57 cas sont liés à ces maladies selon notre système.Pour deux cas notre post-traitement ne nous permet pas de choisir entre la classe homme et la classe femme. L'annotation manuelle effectuée sur les données test sont utilisées pour lever l'ambiguïté. La F-mesure passe de 0.784 à 0.785, il semble donc plus intéressant de se passer de l'annotation manuelle. Dans une prochaine version du système, nous garderons donc les deux genres possibles lorsque nous ne sommes pas en mesure de désambiguïser.

3.5 Suppression de termes non spécifiques au domaine médical

Dans le MeSH nous avons remarqué qu'il y avait certains termes très génériques (e.g. « maladie ») et/ou ambigus (e.g. « pris »). Pour supprimer ces termes qui risquent de nous apporter beaucoup de faux positifs, nous nous sommes basés sur la fréquence des mots dans un corpus non spécifique au domaine médical. Nous avons choisi d'utiliser le corpus Wikipédia FR 2008 [8] pour lequel un fichier avec les fréquences de chaque token est disponible. Grâce à ce corpus, nous avons supprimé les lexèmes du lexique composés d'un seul mot qui sont très fréquents et donc possiblement non spécifiques au domaine médical ou ambigus. Nous avons évalué différents seuils sur le corpus d'entraînement, nous obtenons la meilleure précision avec un seuil à 100, le meilleur rappel avec un seuil à 5000 et la meilleure F-mesure avec un seuil à 1000. Nous avons donc choisi de supprimer les lexèmes présents plus de 1000 fois dans le corpus Wikipedia.

Cette étape constitue un post-traitement, mais elle pourrait également être utilisée lors de la préparation du lexique pour le nettoyer. Cela permettrait à la fois d'avoir un lexique plus petit et d'éviter une étape de post-traitement supplémentaire. Nous intégrerons cette modification dans une nouvelle version du système.

Dans le corpus d'entraînement, nous avons ainsi supprimé 324 termes détectés, qui représentent 33 termes uniques du lexique (e.g. « syndrome », « fièvre », « malade »).

4 Résultats

Dans cette section nous présentons les résultats officiels obtenus à DEFT ainsi que les résultats d'expérimentations supplémentaires sur le corpus de test permettant de mettre en évidence l'apport de chaque post-traitement et l'impact des modifications sur le lexique.

Le corpus de test se compose de 108 cas cliniques rédigés en français, représenté par 41 478 mots. Le corpus comporte des annotations manuelles plus conséquentes que le corpus d'entraînement : genre, âge, poids, taille, température etc.

8. http://redac.univ-tlse2.fr/corpus/wikipedia.html

Nous avons soumis trois runs à DEFT, qui différent selon les ressources utilisées.

— Run1 : MeSH
— Run2 : MeSH + annotation du corpus d'entraînement
— Run3 : MeSH + MedDRA©

Dans la partie gauche de la table 3, nous présentons les résultats officiels fournis par les organisateurs de DEFT. Dans la partie droite nous indiquons le nombre de cas pour lesquels la précision est de 1, c.-à-d. que toutes les classes de ce fichier ont été trouvées, et ceux pour lesquels la précision est de 0, c.-à-d. qu'aucune classe n'a été trouvée pour ce fichier. Nous obtenons les meilleurs résultats avec le run 2, c.-à-d. en nettoyant et complétant le lexique avec les annotations provenant du corpus d'entraînement. Cette configuration nous permet également d'avoir la meilleure précision (0,885). Pour le run 3, nous avons utilisé MedDRA©, ce qui a permis d'augmenter le lexique de façon considérable (+58 071 termes). Cette augmentation permet d'augmenter un peu le rappel par rapport aux run 1 et 2, mais fait chuter la précision à 0,679. Le run1 pour lequel nous n'avons utilisé que le MeSH, nous permet d'obtenir des bons résultats, proche du run 2 et supérieurs à la médiane de la tâche.

Les classes pour lesquelles nous avons obtenues les moins bons résultats sont la classe *blessures* (F-mesure entre 0.49 et 0.55), *chimiques* (F-mesure entre 0.36 et 0.54) et *virales* (1 classe sur 4 a été identifiée dans le corpus de test).

	Évaluation globale			Nombre de cas	
	Précision	Rappel	F1	P=1	P=0
Run1	0,873	0,713	0,785	25	6
Run2	**0,888**	0,750	**0,814**	29	4
Run3	0,686	**0,769**	0,725	42	5

TABLE 3 – Résultats officiels obtenus sur la tâche 1.

Dans la table 4 nous présentons des résultats d'expérimentations supplémentaires sur le corpus de test permettant de mettre en évidence l'impact des post-traitements et de la modification du lexique sur les performances du système.

Configuration	Précision	Rappel	F1
MeSH	0,725	0,739	0,732
MeSH + négation	0,739	0,738	0,738
MeSH + genre	0,739	0,798	0,768
MeSH + négation + genre	0,816	0,738	0,775
MeSH + négation + genre + fréquence	0,873	0,713	0,785
MeSH + négation + genre + fréquence + annotation train	0,888	0,750	0,814

TABLE 4 – Résultats d'expérimentations supplémentaires sur l'impact des post-traitements.

Nous observons que le post-traitement lié à l'identification du genre du patient permet une amélioration importante des performances (+ 0,036 pour la F-mesure). Un gain important est également observé avec l'utilisation des annotations du train pour nettoyer et augmenter le lexique.

5 Conclusion

Dans cet article nous avons proposé une méthode rapide, simple, et efficace pour la classification de documents médicaux. Notre approche montre que les lexiques constituent une ressource riche pour cette tâche, facilement mise à jour, et permettent d'obtenir une bonne qualité de classification.

Une analyse d'erreur rapide nous a permis de mettre en évidence que la plupart des erreurs étaient dues à l'absence de certains termes dans le lexique (e.g. *kystes biliaires hépatiques*), ou au fait que certains termes présents ne sont pas reliés à la classe identifiée manuellement (e.g. *fistule* et *fistule cutanée* sont contenus dans la classe *etatsosy* et *peau* pour le deuxième, mais pas dans la classe *tumeur*, contrairement à ce qui a été annoté manuellement). Nous observons aussi quelques cas qui nécessitent un raisonnement, par exemple *violente chute* qui induit des blessures.

Comme discuté dans les parties précédentes, certaines améliorations du système peuvent être envisagées. Notamment pour le traitement de la négation, où il est envisageable de mettre en place une méthode sémantique cherchant à mieux identifier la porté des marqueurs de négation. Il peut être aussi intéressant de tenter de détecter des négations plus complexes comme dans la phrase « Elle n'a présenté des nausées que durant la nuit et aucun vomissement ».

Pour le traitement du genre, il n'est pas nécessaire d'utiliser les annotations manuelles, si nous avons trop de non détection nous pouvons envisager une augmentation du lexique avec des listes de prénoms ou des participes passés (« s'est présenté »).

La suppression de certains lexèmes dans le lexique basée sur leur fréquence dans le corpus Wikipedia FR 2008 est réalisée en post-traitement. Nous pensons qu'il serait préférable de se servir de ces fréquences pour nettoyer le lexique en amont. L'avantage est d'avoir un lexique plus petit et de ne pas chercher des variants inutilement. Nous avons testé cette nouvelle configuration et nous obtenons pour le run 1 une F-mesure de 0,788 (au lieu de 0,785).

A l'heure des réseaux de neurones énergivores, coûteux en maintenance et en entraînement, notre méthode s'inscrit dans une démarche éco-responsable. Selon (Strubell *et al.*, 2019), entraîner un modèle de réseaux de neurones profond « à l'état de l'art » pour faire de la traduction correspondrait à l'impact de la durée de vie de 5 voitures. Les durées d'entraînement pour ces modèles peuvent aussi aller de quelques jours à plusieurs semaines. Les résultats obtenus montrent que les approches symboliques restent efficaces pour ce genre de tâche. L'utilisation de structures de données adaptées permet de minimiser les ressources nécessaires, ce qui diminue le coût mais aussi le temps d'exécution. Avec notre approche, nous avons pu atteindre un temps $\approx$ 1min pour la classification de 108 documents. Ce système peut aussi être facilement déployé sur des systèmes embarqués.

Références

BRIN S. & PAGE L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, **30**(1-7), 107–117.

BRISCOE T. & CARROLL J. A. (2002). Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain* : European Language Resources Association.

BROWN E. G., WOOD L. & WOOD S. (1999). The Medical Dictionary for Regulatory Activities (MedDRA). *Drug Safety*, **20**(2), 109–117. DOI : 10/czv6mb.

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

ELHADAD N., PRADHAN S., GORMAN S. L., MANANDHAR S., CHAPMAN W. W. & SAVOVA G. K. (2015). Semeval-2015 task 14 : Analysis of clinical text. In D. M. CER, D. JURGENS, P. NAKOV & T. ZESCH, Éds., *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, p. 303–310 : The Association for Computer Linguistics. DOI : 10.18653/v1/s15-2051.

GAIO M. & MONCLA L. (2017). Extended Named Entity Recognition Using Finite-State Transducers : An Application To Place Names. In *The Ninth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2017)*, Nice, France. HAL : hal-01492994.

GARCELON N., SALOMON R. & BURGUN A. (2014). Enrichissement sémantique associé à la détection de la négation et des antécédents familiaux dans un entrepôt de données hospitalier. In *JFIM*, p. 83–93.

GRABAR N., GROUIN C., HAMON T. & CLAVEAU V. (2019). Corpus annoté de cas cliniques en français. In *TALN 2019 - 26e Conference on Traitement Automatique des Langues Naturelles*, p. 1–14, Toulouse, France. HAL : hal-02391878.

GROSS M. (1987). The use of finite automata in the lexical representation of natural language. In *LITP Spring School on Theoretical Computer Science*, p. 34–50 : Springer.

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

JONES K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

MIHALCEA R. & TARAU P. (2004). Textrank : Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, p. 404–411.

MIHOV S. & MAUREL D. (2000). Direct construction of minimal acyclic subsequential transducers. In *Implementation and Application of Automata, 5th International Conference, CIAA 2000, London, Ontario, Canada, July 24-25, 2000, Revised Papers*, volume 2088 de *Lecture Notes in Computer Science*, p. 217–229 : Springer.

MINARD A., ROQUES A., HIOT N., ALVES M. H. F. & SAVARY A. (2020). Doing@deft : cascade de CRF pour l'annotation d'entités cliniques imbriquées (doing@deft : cascade of CRF for the annotation of nested clinical entities). In R. CARDON, N. GRABAR, C. GROUIN & T. HAMON, Éds., *Actes de la 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, June 8-19, 2020*, p. 66–78 : ATALA et AFCP.

PATHAK P., PATEL P., PANCHAL V., SONI S., DANI K., PATEL A. & CHOUDHARY N. (2015). ezdi : A supervised NLP system for clinical narrative analysis. In D. M. CER, D. JURGENS, P. NAKOV & T. ZESCH, Éds., *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, p. 412–416 : The Association for Computer Linguistics. DOI : 10.18653/v1/s15-2071.

PORTER M. F. (2001). Snowball : A language for stemming algorithms. Published online. Accessed 11.03.2008, 15.00h.

PRADHAN S., ELHADAD N., SOUTH B. R., MARTÍNEZ D., CHRISTENSEN L. M., VOGEL A., SUOMINEN H., CHAPMAN W. W. & SAVOVA G. K. (2013). Task 1 : Share/clef ehealth evaluation lab 2013. In P. FORNER, R. NAVIGLI, D. TUFIS & N. FERRO, Éds., *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*, volume 1179 de *CEUR Workshop Proceedings* : CEUR-WS.org.

STRUBELL E., GANESH A. & MCCALLUM A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv :1906.02243*.

WENG W.-H., WAGHOLIKAR K. B., MCCRAY A. T., SZOLOVITS P. & CHUEH H. C. (2017). Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC medical informatics and decision making*, **17**(1), 1–13.