DEFT 2020 - Extraction d'information fine dans les données cliniques : terminologies spécialisées et graphes de connaissance

Thomas Lemaitre, Camille Gosset, Mathieu Lafourcade, Namrata Patel, Guilhem Mayoral

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

DEFT 2020 - Extraction d'information fine dans les données cliniques : terminologies spécialisées et graphes de connaissance

Thomas Lemaitre[1], Camille Gosset[1], Mathieu Lafourcade[1], Namrata Patel[2,3], Guilhem Mayoral[3]
(1) LIRMM, Université Montpellier, Montpellier, France
(2) AMIS, Université Paul-Valéry Montpellier 3, Montpellier, France
(3) Onaos, Montpellier, France
(1) prenom.nom@lirmm.fr
(2) prenom.nom@univ-montp3.fr
(3) prenom.nom@onaos.com

RÉSUMÉ

Nous présentons dans cet article notre approche à base de règles conçue pour répondre à la tâche 3 de la campagne d'évaluation DEFT 2020. Selon le type d'information à extraire, nous construisons (1) une terminologie spécialisée à partir de ressources médicales et (2) un graphe orienté basé sur les informations extraites de la base de connaissances généraliste et de grande taille - JeuxDeMots.

ABSTRACT

Fine-grained Information Extraction in Clinical Data : Dedicated Terminologies and Knowledge Graphs

This paper presents our rule-based approach for fine-grained information extraction in clinical data, submitted in reponse to Task 3 at the DEFT 2020 evaluation campaign. We design (1) a dedicated medical terminology from existing medical references and (2) a knowledge graph based on the semantically rich knowlege base - JeuxDeMots.

MOTS-CLÉS : données cliniques, extraction d'information fine, graphes de connaissance.

KEYWORDS: clinical data, fine-grained information extraction, knowledge graphs.

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 2020 (Cardon *et al.*, 2020). Cette tâche porte sur l'extraction d'informations dans un corpus de cas cliniques (Grabar *et al.*, 2018). Notre système est conçu à base de règles et s'appuie sur deux sources de données différentes :

— une base de connaissances JeuxDeMots qui est généraliste et de grande taille (Lafourcade, 2015)
— Une terminologie médicale dédiée construite pour la détection de chacune des catégories d'informations à extraire.

Un ensemble de règles est ensuite appliqué pour détecter les différentes catégories d'information recherchées. Ce travail a été réalisé en collaboration avec un expert du domaine médical, qui a été principalement impliqué dans la constitution de notre terminologie spécialisée ainsi que la

Atelier DÉfi Fouille de Textes, pages 55–65.
HAL : hal-02784742.

définition de règles d'extraction d'information ayant un fondement médical. Notre choix d'approche symbolique nous a permis également de faire une analyse fine des résultats produits au fur et à mesure de l'évolution du système, afin d'en améliorer les lexiques et les règles.

2 Méthodologie

Afin de répondre à la tâche d'extraction d'informations fines dans des dossiers cliniques, nous adoptons des méthodes à base de règles s'appuyant sur deux sources de données différentes, selon le type d'information à extraire.

Pour les catégories spécialisées autour du patient et des pratiques cliniques, nous avons construit une terminologie dédiée à partir de ressources existantes, organisée et complétée selon les besoins de la tâche. Un ensemble de règles est ensuite appliqué pour détecter les différentes catégories d'information recherchées. Nous appliquons les pré-traitements classiques liées aux approches symboliques à base de règles (tokénisation, lemmatisation, étiquetage en parties du discours, etc.) afin de générer nos résultats.

Pour les catégories autour du temps et des traitements médicaux, nous explorons une nouvelle approche exploitant la base de connaissances JeuxDeMots, basée sur le parcours d'un graphe orienté.

2.1 Extraction d'informations autour des patients et des pratiques cliniques

L'extraction automatique des informations de cette catégorie de données, en particulier la classe *signe ou symptôme* qui selon le DEFT définit à la fois des symptômes cliniques mais aussi les observations ou résultats d'examens complémentaires, est essentielle à l'établissement d'un diagnostic médical précis. L'association de ces éléments permet en effet, de définir pour un patient donné, un état pathologique particulier.

Même selon une approche automatisée, il est indispensable pour un médecin utilisateur d'un système informatique d'accéder aux explications et de pouvoir critiquer la catégorisation des entités trouvées automatiquement dans le dossier clinique de son patient. Pour cette raison nous avons, en collaboration avec un médecin, choisi une approche symbolique d'extraction d'information, permettant une transparence complète sur les résultats du système.

L'algorithme d'extraction d'information a été défini selon la démarche suivante :

- constitution d'une terminologie spécialisée cohérente aux annotations de référence
- définition de règles linguistiques permettant de compléter la détection de termes issus de la terminologie
- définition de règles médicales pour les cas complexes, tels que *signe ou symptôme* ou *pathologie*
- amélioration récursive de la terminologie et des règles par analyse des faux positifs (FP) et faux négatifs (FN) générés à chaque itération.

2.1.1 Constitution d'une terminologie spécialisée

Ressource principale Les données issues des ressources "Logical Observation Identifiers Names and Codes" (LOINC) et "Unified Medical Language System" (UMLS) ont servi de support à la constitution d'une base de données spécialisée en terminologie médicale. Chacune des entrées UMLS étant classées par un « semantic type », nous avons sélectionné celles qui correspondent aux catégories définies pour la tâche 3 du DEFT 2020, notamment les informations liées aux *sites anatomiques*, *pathologies* et *signes ou symptômes*. Ainsi, les « semantic type »issus de l'UMLS ont été rapprochées des catégories d'entités recherchées, en accord avec le guide d'annotation du DEFT 2020. Nous détaillons dans ce qui suit notre sélection de classes UMLS :

Anatomie :
- ANAT|Anatomy|T017|Anatomical Structure
- ANAT|Anatomy|T029|Body Location or Region
- ANAT|Anatomy|T023|Body Part, Organ, or Organ Component
- ANAT|Anatomy|T030|Body Space or Junction ...

Pathologie :
- DISO|Disorders|T049|Cell or Molecular Dysfunction
- DISO|Disorders|T047|Disease or Syndrome
- DISO|Disorders|T191|Neoplastic Process

Signes ou symptômes :
- DISO|Disorders|T184|Sign or Symptom
- DISO|Disorders|T033|Finding
- DISO|Disorders|T046|Pathologic Function

Examen :
- PROC|Procedures|T060|Diagnostic Procedure
- PROC|Procedures|T059|Laboratory Procedure
- PHEN|Phenomena|T034|Laboratory or Test Result

Traitement :
- PROC|Procedures|T061|Therapeutic or Preventive Procedure

Substance :
- CHEM|Chemicals and Drugs|T131|Hazardous or Poisonous Substance
- CHEM|Chemicals and Drugs|T125|Hormone
- CHEM|Chemicals and Drugs|T121|Pharmacologic Substance
- CHEM|Chemicals and Drugs|T127|Vitamin
- CHEM|Chemicals and Drugs|T195|Antibiotic
- CHEM|Chemicals and Drugs|T200|Clinical Drug
- LIVB|Living Beings|T002|Plant

Ressources supplémentaires Nos terminologies spécialisées ont été complétées selon les besoins de la tâche en fonction de termes manquants des ressources principales, celles-ci étant riches en anglais, mais pas entièrement traduites en français.

2.1.2 Écriture de règles linguistiques et médicales avec la terminologie spécialisée

Grâce aux étapes de prétraitement des approches classiques d'extraction d'information (tokenisation, etc.), nous avons élaboré des règles linguistiques et médicales visant à raffiner la détection des entités issues des terminologies.

Les entités plus complexes telles que *signes ou symptômes* ou *pathologie* ont nécessité la définition de règles complexes, exploitant à la fois (1) des informations linguistiques (étiquettes des parties du discours, dépendances morphologiques, etc.) et (2) des informations médicales combinant souvent plusieurs entités médicales, elles-mêmes extraites par le système.

Par exemple, le terme « sténose » issu du lexique *signe ou symptome* a permis au système de détecter l'entité complète suivante :

> « *sténose complète de l'uretère gauche au niveau de la jonction iliopelvienne étendue environ sur 5 cm* »

par application d'une règle qui détecte le groupe nominal associé à « sténose », contenant des entités *anatomie* (uretère gauche, jonction iliopelvienne) et *valeur* (5 cm).

2.2 Extraction d'informations autour du temps et des traitements médicaux

Les entrées du système sont des textes tirés de dossiers patient. Le but est de détecter des entités nommées. Une entité nommée peut représenter un « moment » ou une « durée », par exemple. Nous souhaitons utiliser une bonne structure de données capable de regrouper un maximum d'informations sur les éléments du texte. Notre choix de structure de données s'est orienté vers une structure de graphe. Cette structure de graphe est couplée à une base de connaissances JeuxDeMots qui est généraliste et de grande taille (Lafourcade *et al.*, 2015). L'intérêt de notre structure est de pouvoir regrouper toutes les informations récupérées de JeuxDeMots pour chaque élément du texte. Cela permet ensuite d'obtenir une trace concrète d'inférences faites par notre système. JeuxDeMots est l'outil d'un programme de recherche en Traitement Automatique du Langage Naturel conçu par Mathieu Lafourcade au sein de l'équipe TEXTE du LIRMM (Lafourcade, 2007). JeuxdeMots est un jeu sérieux appartenant à la famille des jeux avec buts, et permet d'enrichir un réseau lexico-sémantique représentant les connaissances avec des relations orientés, typées et pondérées.

2.2.1 Représentation des connaissances

Comme pour le réseau de JeuxDeMots, notre structure de graphe utilise des relations orientés, typées et pondérées. Pour chaque mot de notre texte, nous créons un nœud dans lequel nous pouvons récupérer la chaîne de caractères associée, l'index du début de la chaîne de caractères dans le texte et l'index de fin . L'ordre des mots est représenté par la relation `r_succ` entre un nœud et son nœud successeur.

Une fois notre graphe de base obtenu, nous utilisons JeuxDeMots pour effectuer une lemmatisation et l'étiquetage des mots (*Postagging*). La lemmatisation consiste à déterminer la racine d'un mot. Par exemple, le mot « mangera » a pour forme lemmatisée « manger ». L'étiquetage des mots, quant à lui, consiste à assigner à un mot une ou plusieurs classes grammaticales possibles. Par exemple, le mot « attaque », hors contexte, a les classes grammaticales de nom commun et de verbe. Pour réaliser ces

deux opérations nous utilisons respectivement les relations `r_lemma` et `r_pos` qui relient un nœud donné avec ses formes lemmatisées et ses étiquettes.

Ensuite, il est primordial de ne pas perdre le sens d'un mot en le décortiquant. Lorsque le texte est mis sous forme de graphe, chaque mot est décortiqué un à un. Cependant, le mot peut être dit composé. Un mot composé est une succession de plusieurs mots qui donne un sens précis. Par exemple, le mot composé « mettre les pieds dans le plat » obtient un sens propre à cette expression différent des mots successifs un à un. Afin de retrouver ces mots composés, nous utilisons un dictionnaire de mots composés bien fourni extrait depuis JeuxDeMots. Nous parcourons notre graphe en prenant une suite de mots et vérifions l'existence d'un tel mot composé dans le dictionnaire.

Finalement, nous utilisons un ensemble de règles générales. Ces règles permettent de réaliser une analyse syntaxique du texte. L'analyse syntaxique permet de s'intéresser à la structure du texte en établissant des relations entre les mots basés sur des motifs. Grâce à cet ensemble de règles, le système va créer de nouveaux nœuds. En s'appuyant sur l'étiquetage des mots, il va former de nouveaux groupes. Par exemple, on obtiendra un groupe nominal à partir d'un nom et un adjectif. Ces deux mots sont naturellement successeurs. Cette phase permet également de retrouver des mots composés constitués de plusieurs mots consécutifs qui ne se réfèrent pas forcément au dictionnaire précédemment utilisé.

Lorsqu'une règle est appliquée, celle-ci créée le nœud du groupe de mot. Puis, elle relie ce nœud nouvellement créé au reste du graphe, afind de créer un chemin alternatif dans le graphe. Grâce à la relation `r_isa` de JeuxDeMots, il est possible de vérifier si certains mots représentent des mois ou des moments. Cela permet de regrouper une suite de mots et d'améliorer la détection des mots qui doivent être annotés avec l'utilisation d'un second ensemble de règles spécialisées pour le DEFT.

2.2.2 Utilisation des règles

Précédemment, nous avons expliqué que nous utilisions un système de règles afin de regrouper des groupes de mots. Nous avons, donc, rédigé un ensemble de règles. Ce système de règles ainsi que les règles elles-mêmes ont été rempli par nos soins. Le but de cette analyse syntaxique est n'utiliser que peu de règles afin de ne pas exploser en complexité computationnelle.

Chaque règle est constituée de deux parties : une partie de conditions et une partie d'application. La partie de conditions peut être composée du symbole $\&$ et/ou du symbole $\|$. Cette partie contient un ensemble de conditions non limité. Lorsque des conditions sont entourées par le symbole $\&$ alors cet ensemble se doit de respecter toutes les conditions. Lorsque des conditions sont entourées par le symbole $\|$, alors seulement une condition se doit d'être respectée. Le système réalise une évaluation paresseuse. Si la partie de conditions est validée alors la partie d'application pourra être lancée sur les nœuds sélectionnés. La partie d'application peut contenir plusieurs actions. Chacune des actions pourra être appliquée sur les nœuds sélectionnés préalablement. Une condition ou une application est représentée par un triplet.

Deux type de triplets sont possibles pour la partie condition. Le premier type est *triplet link* qui permet de vérifier l'existence d'une relation entre deux nœuds variables. Un nœud variable peut être remplacé par un nœud constante. Par exemple, on cherchera à sélectionner les liens de type $\$x\ r\_pos\ Nom$:. Le deuxième type est *triplet equals* et permet d'indiquer si une chaîne de caractère d'un nœud est bien égale à un autre nœud ou à une valeur donnée.

Lorsque l'on lance la phase de conditions, on récupère un ensemble de nœud. Dans le cas où cet ensemble n'est pas vide, la partie application pourra lancer les triplets sur ces nœuds sélectionnés. Deux types de triplet d'application sont possibles. Le premier consiste à créer un nouveau lien et se nomme *triplet link*. Il s'utilise de la même manière que celui de la partie condition. Le second consiste à créer un nouveau nœud représentant la composition de plusieurs nœuds ou bien du label de la constante indiquée (*triplet makeNode*).

Basé sur ce même système règles, l'utilisateur peut poser des questions sur le graphe ou sur des informations liées au graphe et présentes dans JeuxDeMots. Notre système de règles est extensible et réutilisable. Nous l'avons donc facilement adapté afin de répondre aux questions. Une requête est de la forme $condition \rightarrow Return : \ elements\ a\ retourner$. Le but est de savoir quels nœuds correspondent à la condition de la condition pour ensuite retourner les nœuds sélectionnés.

2.2.3 Génération des résultats

Finalement, lorsque l'ensemble des règles est appliqué nous parcourons une dernière fois l'ensemble des nœuds du graphe dans le but de récupérer les nœuds représentants les entités détectées. Pour chaque nœud nous recherchons des relations indiquant des annotations possibles. Chaque annotation possède un nœud dans le graphe et les nœuds détectés comme une entité nommée possèdent une relation `-r_annoted->` vers les annotations. Par exemple : "durant 3 mois" `r_annoted` « Durée ». Ainsi, pour chaque nœud d'entités détectées, nous rajoutons dans le fichier de sortie, la chaîne de caractères associée avec son annotation, et ses index de début et de fin.

3 Résultats

Nous présentons ci-dessous les résultats issus des deux approches décrites. Notre choix d'approche symbolique nous permettant de réaliser une analyse fine des résultats obtenus, nous distinguons deux manières de calculer les mesures d'évaluation : (1) les mesures strictes (officielles) et (2) des mesures plus souples qui prennent en compte les annotations partiellement correctes.

La mise en perspective du différentiel de performance entre ces deux mesures permet d'isoler les entités détectées de manière « aberrante » (FP) et entièrement non-détectées par le système (FN). Cette distinction nous permet de discuter finement les divergences obtenues entre notre système et les annotations de référence.

3.1 Autour du patient et des pratiques cliniques

Mesures officielles	TP	FP	FN	Précision	Rappel	F1
anatomie	631	297	489	0,6800	0,5634	0,6162
examen	417	217	400	0,6577	0,5104	0,5748
substance	171	107	142	0,6151	0,5463	0,5787
traitement	125	170	179	0,4237	0,4112	0,4174
signe ou symptôme	356	540	923	0,3973	0,2783	0,3274
pathologie	69	277	97	0,1994	0,4157	0,2695

Mesures souples	TP	FP	FN	Précision	Rappel	F1
anatomie	818	110	317	0,8815	0,7207	0,7930
examen	563	71	265	0,888	0,68	0,7702
substance	199	79	114	0,7158	0,6358	0,6734
traitement	210	85	101	0,7119	0,6752	0,6931
signe ou symptôme	753	143	530	0,8404	0,5869	0,6911
pathologie	116	230	50	0,3353	0,6988	0,4531

Notre analyse des résultats du système face aux annotations de référence permet d'identifier les typologies principales d'erreurs suivantes :

— mauvaises étiquettes des étapes de prétraitement
— terminologie spécialisée incomplète/divergente
— règles linguistiques/médicales non exhaustives
— règles médicales en accord avec le guide d'annotation mais en désaccord avec l'annotation de référence

Nous discutons dans ce qui suit, les détails de cette analyse par catégorie d'entité, accompagnée des exemples d'erreurs les plus notables.

Anatomie :

FN (35% partiellement corrects) : Presque 30% des entités entièrement non détectées sont représentés par des termes relatifs à des *liquides biologiques* (e.g. : "sang", "urines") et par des termes qui ne sont pas en rapport avec une partie anatomique (e.g. : "anatomopathologique", "psychomoteurs"), non présents dans notre terminologie spécialisée *anatomie*.

FP (63% partiellement corrects) : La plupart des FP restants sont des erreurs dues aux étiquettes de prétraitement. Il en existe un faible nombre (13) qui sont des oublis d'annotation de la référence.

Examen :

FN (34% partiellement corrects) : Presque 25% des entités entièrement non détectés par notre système concernent des termes génériques et mettent en évidence *la limite d'utiliser des terminologies hyperspécialisées pour répondre à la tâche*. Ainsi, notre terminologie issue de l'UMLS, intégrait des « procédures médicales diagnostiques », mais n'intégrait pas *des termes d'ordre plus général employés seuls*, tels que : "analyse", "bilan", "consultation", "examen", "exploration", "interrogatoire", "investigation", "surveillance".

FP (67% partiellement corrects) : Nous avons intégré dans notre terminologie d'examens certains noms de substances qui sont normalement mesurées dans des examens biologiques : ces termes détectés seuls ont produit des FP.

Substance :

FN (20% partiellement corrects) : Près de 30% des entités non détectés par notre système sont dues à une construction incomplète de notre terminologie spécialisée. Il s'agit de termes génériques ne renvoyant pas à un nom de substance médicamenteuse précis, mais, soit à une *classe médicamenteuse* ("anti-androgène", "antiarythmique", "traitement anti-bacillaire"), soit à des *types de traitement* ("traitement de rattrapage", "traitement local d'appoint", "traitement

prophylactique", "transfusions", "antibiothérapie", "analgésie"). Notre terminologie de référence était majoritairement constituée de noms de molécules et ne pouvait permettre à notre système de détecter ce type de termes.

FP (26% partiellement corrects) : Les FP générés par notre système ne présentent pas de cas aberrants, il s'agit d'erreurs dues au problèmes de prétraitement.

Traitement :

FN (44% partiellement corrects) : La difficulté principale avec cette catégorie a été la détection de cas complexes telles que :

> « fond du néo-vagin avait été créé en suturant par un surjet résorbable le moignon colique droit ».

FP (50% partiellement corrects) : Dans cette catégorie, les FN et les FP sont expliqués par la même difficulté.

Signe ou symptôme :

FN (43% partiellement corrects) : Pour cette catégorie on note un différentiel important de performance entre les mesures strictes et souples. Ceci s'explique par la complexité inhérente de l'information représentée. Le guide d'annotation précise qu'il s'agit *« des résultats d'observations cliniques ou à un examen avec son résultat »*. Ceci peut constituer des groupes nominaux complexes qui doivent faire appel à des règles linguistiques complexes, par exemple :

> Gold : « masse de 1,5 x 1,0 cm au niveau du pôle inférieur du rein gauche, se rehaussant après injection du produit de contraste »
> Système : « masse de 1,5 x 1,0 cm au niveau du pôle inférieur du rein gauche »

FP (73% partiellement corrects) : Le nombre aussi élevé d'annotations partiellement correctes s'explique par une discordance relevée entre nos résultats et ceux de référence en rapport avec les termes qualifiant *l'absence ou la présence* d'un *signe ou symptôme*, par exemple :

> Gold : « zone hypoéchogène centro-rénale gauche »
> Système : « présence d'une zone hypoéchogène centro-rénale gauche »

Pathologie :

FN (48% partiellement corrects) : Le nombre élevé de FN pour cette catégorie s'explique par *le manque d'exhaustivité des règles expertes* qui ne peuvent couvrir l'ensemble des cas de figure rencontrés en médecine. Ainsi que par des *divergences entre règles expertes et annotation de référence*, cette dernière annote ces termes comme *pathologie* tels que : "adénopathies", "adénopathies périphériques", "cystalgies", "oedème réactionnel important de tout le fourreau du pénis", "mydriase bilatérale aréactive", alors qu'il s'agit de *signe ou symptôme*.

FP (17% partiellement corrects) : Parmi toutes les catégories d'entités traitées, c'est la seule pour laquelle le rappel est nettement meilleur que la précision. Ceci s'explique par le grand nombre de FP générés par le système. On peut l'expliquer par une divergence entre la classe d'entités UMLS *disease or syndrome* que nous avions rapproché de *pathologie* et qui comporte en fait d'authentiques *signes ou symptômes* tels que : "hématurie", "thrombopénie", "anémie".

3.2 Autour du temps et des traitements médicaux

Mesures officielles	TP	FP	FN	Précision	Rappel	F1
dose	10	11	42	0,4762	0,1923	0,2740
mode	16	7	73	0,6957	0,1798	0,2857
moment	85	243	80	0,2591	0,5152	0,3448
valeur	155	115	277	0,5741	0,3588	0,4416

Mesures souples	TP	FP	FN	Précision	Rappel	F1
dose	12	09	40	0,5714	0,2308	0,3288
mode	20	3	69	0,8696	0,2247	0,3571
moment	156	172	28	0,4756	0,8478	0,6094
valeur	225	45	216	0,8333	0,5102	0,6329

En regardant les résultats obtenus dans ces catégories, nous pouvons observer une précision plus importante par rapport au rappel . Ceci est dû à notre approche qui priorise la précision au rappel, pour la majorité des catégories traités. Ainsi dans les catégories avec un faible rappel, les règles déjà présentent sont efficaces mais il manque des règles pour récupérer les éléments manquants. Dans le cas de la catégorie moment nous pouvons remarquer que la précision est plus faible que le rappel, ce qui montre que certaines sont trop générales et nécessitent d'être affinées.

3.3 Observations générales d'ordre médical

Concernant la détection des classes d'entités liées au patient et aux pratiques cliniques, les *pathologies* et *signe ou symptôme* sont beaucoup plus dépendantes des règles médicales que les autres classes, pour lesquelles la bonne construction des terminologies et des règles linguistiques apportent déjà de bonnes performances de détection. Or, médicalement, la distinction *pathologie* versus *signe ou symptôme* est difficile pour plusieurs raisons :

Consensus entre experts : Intrinsèquement certaines observations cliniques ou paramédicales assimilables donc, à des *signes ou symptômes*, sont aussi d'authentiques *pathologies*. Ainsi, dans les annotations de référence, clairement identifier ce qui relève de la pathologie, de ce qui relève d'observations d'examens cliniques ou paracliniques peut ne pas être consensuel, même entre experts médecins. On peut citer le cas des anomalies morphologiques évocatrices d'une affection congénitale telles que *« rein en fer à cheval »* ou *« hypospadia »*, toutes deux classées en *signes ou symptômes* par la référence du DEFT mais qui pourraient également être reclassées en *pathologie* (congénitale).

La question des « syndromes » : Selon la définition du Larousse, un syndrome est *« un ensemble de plusieurs symptômes ou signes en rapport avec un état pathologique donné et permettant, par leur groupement, d'orienter le diagnostic »* et un syndrome biologique est un *« ensemble des modifications biochimiques, physiques, sérologiques, bactériologiques caractérisant un état pathologique donné »*. Il semblerait donc que nosologiquement les syndromes puissent être assimilés à des *signes ou symptômes*, or certains syndromes sont utilisés pour qualifier d'authentiques *états pathologiques* : *« syndrome néphrotique »* par exemple.

Ambiguïté d'une règle d'annotation : Il faut relever l'ambiguïté de la règle d'annotation du DEFT suivante, concernant les tumeurs : *« Les tumeurs malignes sont annotées* pathologie *tandis que les tumeurs bénignes seront annotées* signe ou symptôme *»*. Le degré de malignité

semblait donc constituer la frontière entre *pathologie* et *signe ou symptôme*. Or une tumeur même bénigne peut constituer une pathologie, citons « l'adénome prostatique » qui est une *pathologie* bien qu'une tumeur bénigne. Le fondement médical de cette distinction reste à préciser.

Importance du contexte pour désambiguïser des termes : la désambiguïsation fait aussi appel au contexte de la phrase dans lequel apparaît le terme. Ainsi dans la phrase suivante : *« à l'examen clinique, le patient présente une hypertension artérielle »*, « *hypertension artérielle »* est à classer en *signe ou symptôme*. A l'inverse, dans l'exemple suivant issu du texte 140-2 des données de test : *« Mme R.S, âgée de 60 ans, suivie depuis 10 ans pour hypertension artérielle, a été admise pour une douleur lombaire gauche évoluant depuis 1 mois »*, « *hypertension artérielle »* pourtant annotée en *signe ou symptôme* par la référence, est sans aucun doute une authentique *pathologie*.

4 Conclusion

Dans cet article nous avons présenté les méthodes que nous avons conçues pour répondre à la tâche 3 de la campagne d'évaluation DEFT 2020. Pour les catégories liées aux patients et aux pratiques cliniques, nous avons adopté une approche classique d'extraction d'informations à base de règles et de lexiques spécialisés. Pour les catégories autour du temps et des traitements médicaux, nous avons développé une approche basée sur les graphes de connaissances, exploitant le réseau sémantique JeuxDeMots.

Nous avons constaté que la performance de notre système a été souvent bonne en précision, grâce aux règles linguistiques et médicales complexes, mais ne pouvait couvrir de manière exhaustive l'ensemble des cas de figure rencontrés en médecine. Cela nécessiterait pour les experts médicaux une charge de travail trop importante et ceci nous interroge sur la possibilité de maintenir une approche uniquement basée sur les règles dans le domaine médical.

En perspectives, nous allons explorer des approches hybrides qui permettraient d'exploiter à la fois (1) les performances des approches par apprentissage, qui sont efficaces pour le traitement de l'exhaustivité, et (2) les connaissances métier apportées par les experts médicaux.

Plus précisément, nous avons identifié lors de la détection d'entités complexes, des cas où nos règles médicales ont été prises en défaut par la non détection d'entités simples (*anatomie, valeur, examen*), servant de point d'ancrage dans le texte à l'application de ces règles. Afin de viser une meilleure couverture des cas de figure rencontrés en médecine, nous pourrions nous appuyer sur des approches par apprentissage automatique pour la détection de ces entités simples, tout en conservant nos règles d'expertise liées aux entités plus complexes telles que *signes ou symptômes* ou *pathologie*.

Références

CARDON R., GRABAR N., GROUIN C. & HAMON T. (2020). Présentation de la campagne d'évaluation deft 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In *Actes de DEFT*.

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

122–128, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5614.

LAFOURCADE M. (2007). Making people play for lexical acquisition.

LAFOURCADE M. (2015). Jeux de mots, pour quoi faire ?

LAFOURCADE M., LE BRUN N. & JOUBERT A. (2015). Jeux et intelligence collective - résolution de problèmes et acquisition de données sur le web. collection science cognitive et management des connaissances.