Fouille de motifs et CRF pour la reconnaissance de symptômes dans les textes biomédicaux

Pierre Holat, Nadi Tomeh, Thierry Charnois, Delphine Battistelli, Marie-Christine Jaulent, Jean-Philippe Metivier

▶ **To cite this version:**

Pierre Holat, Nadi Tomeh, Thierry Charnois, Delphine Battistelli, Marie-Christine Jaulent, et al.. Fouille de motifs et CRF pour la reconnaissance de symptômes dans les textes biomédicaux. 23e conférence sur le Traitement Automatique des Langues Naturelles (TALN'16), Jul 2016, Paris, France. pp.194-206. halshs-01727081

HAL Id: halshs-01727081
https://shs.hal.science/halshs-01727081v1

Submitted on 12 Mar 2018

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.

Fouille de motifs et CRF pour la reconnaissance de symptômes dans les textes biomédicaux

Pierre Holat[1] Nadi Tomeh[1] Thierry Charnois[1]
Delphine Battistelli [2] Marie-Christine Jaulent[3] Jean-Philippe Métivier[4]
(1) LIPN, Université Paris 13, Sorbonne Paris Cité, Paris, France
(2) MoDyCo, Université Paris Ouest Nanterre La Défense, Paris, France
(3) Inserm, Paris, France
(4) GREYC, Université de Caen Basse-Normandie, Caen, France

RÉSUMÉ

Dans cet article, nous nous intéressons à l'extraction d'entités médicales de type symptôme dans les textes biomédicaux. Cette tâche est peu explorée dans la littérature et il n'existe pas à notre connaissance de corpus annoté pour entraîner un modèle d'apprentissage. Nous proposons deux approches faiblement supervisées pour extraire ces entités. Une première est fondée sur la fouille de motifs et introduit une nouvelle contrainte de similarité sémantique. La seconde formule la tache comme une tache d'étiquetage de séquences en utilisant les CRF (champs conditionnels aléatoires). Nous décrivons les expérimentations menées qui montrent que les deux approches sont complémentaires en termes d'évaluation quantitative (rappel et précision). Nous montrons en outre que leur combinaison améliore sensiblement les résultats.

ABSTRACT

Pattern mining and CRF for symptoms recognition in biomedical texts.

In this paper, we tackle the issue of symptoms recognition in biomedical texts. There is not much attention to this problem in the literature and it does not exist to our knowledge an annotated dataset to train a model. We propose two weakly-supervised approaches to extract these entities. The first is based on pattern mining and introduces a new constraint based on semantic similarity. The second represents the task as sequence labeling using CRF (Conditional Random Fields). We describe our experiments which show that the two approaches are complementary in terms of quantification (recall and precision). We further show that their combination significantly improves the results.

MOTS-CLÉS : Extraction d'information, Fouille de motifs, CRF, Reconnaissance de symptômes, Texte biomédicaux.

KEYWORDS: Information extraction, Pattern mining, CRF, Symptoms recognition, Biomedical texts.

1 Introduction

Le travail décrit dans cet article s'inscrit dans le contexte du projet Hybride [1] dont l'un des objectifs est la capitalisation des connaissances sur les maladies rares. Une maladie rare est une maladie qui

1. hybride.loria.fr

affecte moins d'une personne sur deux-mille. Il y a entre six-mille et huit-mille maladies rares et trente millions de personnes concernées en Europe. L'un des axes du projet est de constituer ou mettre à jour automatiquement des fiches de synthèse résumant les connaissances actuelles sur chaque maladie rare (prévalence, symptôme, étiologie, modes de transmission...). Ce travail de collecte est actuellement réalisé par des experts humains qui surveillent manuellement la littérature médicale sur le sujet. Un enjeu majeur est donc d'apporter une aide à ce processus par la découverte de connaissances en rapport avec les maladies rares. Nous nous concentrons sur la tâche de *reconnaissance de symptômes* dans les résumés de d'articles scientifiques médicaux.

Nous utilisons le terme *symptôme* pour désigner la manifestation d'une maladie, comme ressentie et décrite par un patient (signe fonctionnel, par exemple "*mal de tête*"), ou comme observé par un professionnel de la santé (signe clinique, dans cet exemple "*céphalée*") sans distinction. La structure linguistique des symptômes est généralement plus complexe que les autres entités nommées biomédicales (Cohen, 2010) pour diverses raisons (Martin *et al.*, 2014). Les symptômes peuvent s'exprimer sous des formes très diverses, du simple nom à une phrase entière, et comportent un certain nombre d'ambigüités syntaxiques et sémantiques. Dans l'exemple suivant, les symptômes identifiés par un expert sont en gras :

— With disease progression patients additionally develop **weakness** and **wasting of the limb and bulbar muscles**.
— Diagnosis is based on clinical presentation, and **glycemia and lactacidemia levels after a meal (hyperglycemia and hypolactacidemia), and after three to four hour fasting (hypoglycemia and hyperlactacidemia)**.

De plus, peu de travaux se sont concentrés sur la reconnaissance de symptômes et de ce fait les ressources existantes sont limitées ou incomplètes : à notre connaissance il n'existe pas de corpus avec la totalité des symptômes annotés qui permettrait d'entrainer un apprentissage supervisé.

Pour résoudre ces problèmes, nous proposons une approche faiblement supervisée pour la reconnaissance de symptômes qui combine trois ressources (Section 2.1) : un corpus de résumés d'articles médicaux non annotées et deux dictionnaires, un pour les maladies rares et un autre pour les symptômes.

Nous projetons les dictionnaires sur les résumés pour créer un jeu de données annotées permettant d'entrainer nos modèles. Puisque les dictionnaires ne sont pas exhaustifs, l'annotation n'est que partielle. Grâce à ce jeu de données annotées, nous formalisons le problème de la reconnaissance de symptômes de deux manières complémentaires : (a) comme une tâche de classification de séquence supervisée pour laquelle nous utilisons des champs conditionnels aléatoires (CRF) (Lafferty, 2001) (Section 2.3) ; et (b) comme une tâche de fouille de motifs séquentiels sous contraintes (Béchet *et al.*, 2015) (Section 2.4). Nous combinons ces deux approches dans une architecture modulaire (Section 2).

Notre contribution est triple. Premièrement, nous montrons expérimentalement (Section 3) que la classification de séquences et la fouille de motifs sont deux formalisations adéquates de la tâche, elles sont plus performantes qu'une simple projection de dictionnaire sur le texte. De plus nous montrons que leur combinaison est bénéfique puisque les CRF permettent une représentation plus riche des mots alors que les motifs séquentiels favorisent la modélisation du contexte. Deuxièmement, nous introduisons une nouvelle contrainte en fouille de motifs séquentiels basée sur la similarité distributionnelle entre les mots (Section 2.4.1). Troisièmement, nous avons créé un corpus d'évaluation avec les symptômes manuellement annotés (gold standard) par des experts humains (Section 2.1).

2 Approche pour la reconnaissance de symptômes

Dans cette section nous décrivons notre approche modulaire itérative pour la reconnaissance de symptômes, dans l'esprit de (Métivier *et al.*, 2015). La Figure 2 présente l'architecture globale de notre système. Les données d'entrée de notre approche contiennent une collection de résumés d'articles et deux dictionnaires, un pour les maladies et un pour les symptômes. Premièrement, les dictionnaires sont projetés sur les résumés pour obtenir une annotation partielle ; ces données annotées sont utilisées pour entrainer un classifieur de séquence CRF et comme entrée à un algorithme de fouille de motifs séquentiels ; le modèle CRF appris et les motifs extraits sont utilisés, séparément ou en combinaison, pour extraire les symptômes des données d'évaluation ; ces symptômes sont alors comparés aux symptômes gold standard manuellement extraits pour une évaluation utilisant le F-score. Il est à noter que les modèles appris peuvent être appliqués sur les données d'apprentissage (en utilisant une validation croisée) pour découvrir de nouveaux symptômes à ajouter aux dictionnaires, et que tout ce processus peut être réitéré.

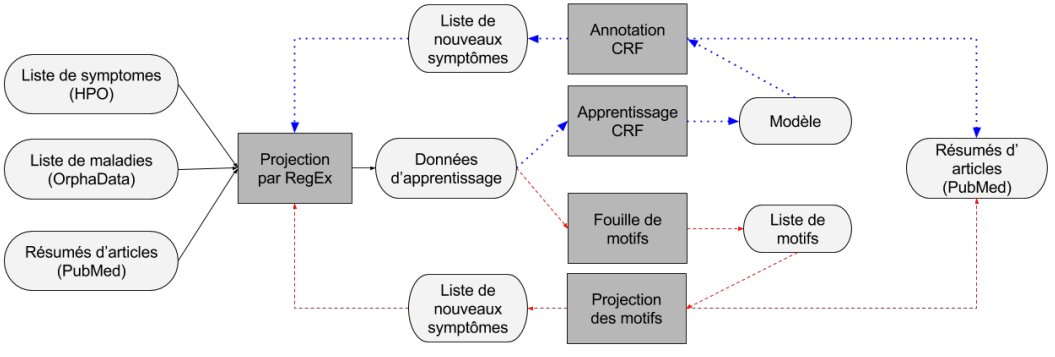

FIGURE 1 – Fonctionnement global du système

2.1 Jeux de données et évaluation

Les données d'entrée de notre système sont composées de trois ressources indépendantes disponibles en ligne :

— Le premier jeu de données est un corpus de 10 000 résumés d'articles extraits de la littérature biomédicale disponible sur PubMed [2]. Pour le constituer, nous avons extrait 100 résumés d'articles biomédicaux pour chacune des 100 maladies rares sélectionnées à l'avance par un expert.

— Le second jeu de données est un dictionnaire de 17 469 anomalies phénotypiques distinctes, 34 257 avec les variations, fournies par HPO (Human Phenotype Ontology [3]). Le phénotype est l'ensemble des caractères observables d'une personne (morphologiques, biochimiques, physiologiques). Il résulte de l'interaction du génotype avec son milieu (l'environnement dans

2. www.ncbi.nlm.nih.gov/pubmed
3. human-phenotype-ontology.github.io

lequel il se développe). Comme de nombreuses maladies rares sont génétiques, nous avons suivi les conclusions de (Martin *et al.*, 2014) et donc considéré les anomalies phénotypiques comme des symptômes.

— Le troisième jeu de données est un dictionnaire de 16 576 noms de maladies rares distincts, 29 803 avec les variations, fournis par OrphaData [4], une ressource de haute qualité sur les maladies rares et les médicaments orphelins.

Le corpus de test contient 50 résumés d'articles, avec une moyenne de 184 mots par résumé, dont chaque symptôme a été annoté manuellement. Nos experts ont relevé 407 symptômes et leurs positions dans ce corpus, soit une moyenne de 8.1 symptômes par résumé.

La performance de notre système est évaluée grâce aux mesures standards de précision, rappel et F-score. [5]

2.2 Annotation par projection

Une fois les données d'entrée préparées, la première étape de notre système est de projeter les dictionnaires de symptômes et de maladies sur les résumés d'articles biomédicaux. Cette projection ne produit qu'une annotation partielle puisque HPO et OrphaData ne sont pas exhaustifs ; ils ne contiennent pas tous les symptômes et toutes les maladies, ni les différentes formes linguistiques qu'ils peuvent prendre.

Les résumés d'articles et les dictionnaires ont été pré-traités avec l'outil TreeTagger (un outil d'étiquetage morphosyntaxique (Schmid, 1994, 1995)) pour ajouter des informations comme le lemme et la catégorie morphosyntaxique de chaque mot. Nous avons ensuite implémenté quelques vérifications supplémentaires sur la liste d'annotations créée par l'annotation automatique. Une première vérification est d'utiliser le *Noun Phrase Chunking* (Ramshaw & Marcus, 1995), dont le but est de découper une phrase en morceaux considérés comme des syntagmes nominaux. Nous l'utilisons pour étendre une annotation de symptôme, si un symptôme et certains mots contigus sont considérés comme étant une phrase nominale alors le symptôme est étendu à cette phrase. En effet les termes venant d'HPO sont très génériques (e.g "weakness") alors que les symptômes dans les textes médicaux sont souvent entourés d'adjectifs ou de compléments d'objets (e.g. "severe weakness of the tongue"). Une seconde vérification va s'occuper de rechercher tous les acronymes du corpus. La dernière va éclater les énumérations (suite séparées par *"and"*, *"or"* ou *","*).

Ce corpus partiellement annoté par projection servira de données d'apprentissage pour les modèles suivants.

2.3 Reconnaissance de symptômes par annotation de séquence

Nous formalisons le problème de la reconnaissance de symptômes comme une tâche d'annotation de séquence, avec un format BIO. Un résumé d'article peut être vu comme une séquence de mots, chacun d'eux est annoté avec une des trois annotations possibles : B (Beginning, le début d'un symptôme), I (Inside, l'intérieur d'un symptôme) et O (Outside, en dehors d'un symptôme). Ainsi un symptôme

4. www.orphadata.org

5. En utilisant le script fourni par `www.cnts.ua.ac.be/conll2000/chunking/output.html` qui utilise le même format de données (BIO) que nos données.

est un segment de mots correspondant à une annotation B potentiellement suivie par des annotations consécutives I, comme montré en figure 2.

Nous utilisons les CRF (Conditional Random Field ou champs conditionnels aléatoires) qui sont des modèles statistiques très utilisés en apprentissage automatique et en traitement du langage. Les CRF ont été introduit par (Lafferty, 2001) : les lecteurs curieux pourront également se tourner vers cette introduction (Sutton & McCallum, 2011). L'avantage principal des CRF est leur nature conditionnelle qui permet des représentations riches des mots d'une séquence. Il est possible d'incorporer de multiples sources d'informations sous la forme de fonctions caractéristiques (feature functions) sans avoir à explicitement modéliser leurs interactions. En effet, il serait excessivement couteux de faire de la fouille de motifs sur de telles représentations.

Nous utilisons le même jeu de descripteurs que la tâche de reconnaissance d'entités nommées de (Finkel *et al.*, 2005) implémenté dans l'extracteur d'entités nommés de Standford (Standford NER) [6]. Cela inclut le mot courant, le précédent et le suivant, tous les mots d'une fenêtre de taille donnée (n-grammes), des descripteurs orthographique caractérisant la forme des mots, ainsi que préfixes et suffixes.

we	PP	O
find	VBD	O
that	IN	O
clinically	RB	O
silent	JJ	B-symptom
tumour	NNS	I-symptom
often	RB	O
demonstrate	VBP	O
subclinical	JJ	B-symptom
hormonal	JJ	I-symptom
activity	NN	I-symptom
.	SENT	O

FIGURE 2 – Exemple de séquence BIO

2.4 Reconnaissance de symptômes par fouille de motifs séquentiels

Alors que l'approche d'annotation de séquences présentée précédemment peut facilement utiliser des représentations riches des mots, elle ne peut pas capturer la richesse du contexte. Par exemple, considérer toutes les sous-séquences possibles comme contexte pour le mot courant engendrerait des calculs insolubles à cause du nombre de sous-séquences exponentiellement croissant sur la taille de la séquence.

Pour résoudre ce problème, nous proposons d'utiliser la fouille de motifs séquentiels en se focalisant sur le contexte plus que sur les mots eux-mêmes. La fouille de motifs séquentiels permet de prendre en compte la séquentialité du langage, l'ordre dans lequel les mots apparaissent étant primordial pour comprendre le sens d'une phrase. La fouille de séquences est une tâche très étudiée, mais complexe,

6. nlp.stanford.edu/software/CRF-NER.shtml

en fouille de donnée. Depuis son introduction (Agrawal & Srikant, 1995), beaucoup de chercheurs ont développé des approches pour fouiller les séquences, le problème est donc bien formalisé :

Soit $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$ un ensemble fini de littéraux appelés $items$. Un $itemset$ est un ensemble non vide d'$items$. Une séquence S sur $\mathcal{I}$ est une liste ordonnée $\langle it_1, \ldots, it_k \rangle$ non vide, où les it_j sont des $itemsets$ de $\mathcal{I}$ et $j = 1 \ldots k$. Une séquence $s_a = \{a_1, a_2, \ldots, a_n\}$ est incluse dans une autre séquence $s_b = \{b_1, b_2, \ldots, b_n\}$ s'il existe des entiers $1 \leq i_1 \leq i_2 \leq \ldots \leq i_n \leq m$ tels que $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, \ldots a_n \subseteq b_{i_n}$. Si la séquence s_a est incluse dans s_b, alors s_a est une sous-séquence de s_b et s_b est une super-séquence de s_a, noté $s_a \preceq s_b$. Le $Support$ d'un motif séquentiel dans une base de séquence est le nombre de séquences dans lesquelles est inclus le motif.

Nos données d'apprentissage contiennent le lemme et la catégorie morphosyntaxique de chaque mot. Donc dans ce contexte, soit $\mathcal{I}$ l'ensemble fini de littéraux composés de tous les lemmes et catégories morpho-syntaxiques ainsi que l'$item$ spécial $\#symptom\#$. Un $itemset$ est un ensemble non vide contenant le lemme et la catégorie morphosyntaxique d'un mot. L'$item$ $\#symptom\#$ servira de remplacement pour chaque symptôme dans les données partiellement annotées d'entrainement, comme le montre cet exemple repris de la Figure 2 :

$< \{we, PP\}\{find, VBD\}\{that, IN\}\{clinically, RB\}\{$**#symptom#**$\}$
$\{often, RB\}\{demonstrate, VBP\}\{$**#symptom#**$\}\{., SENT\} >$

L'extraction de motifs séquentiels sur de telles séquences nous permet d'extraire des patrons linguistiques couvrant des symptômes symboliques. Cependant l'extraction de motifs séquentiels pose encore aujourd'hui un problème quand à la quantité des motifs extraits. Selon les paramètres utilisés les résultats peuvent être trop nombreux pour pouvoir être traités par un expert ou à l'opposé trop génériques pour être intéressants. Les contraintes introduites par (Srikant & Agrawal, 1996) sont un paradigme puissant pour cibler les motifs pertinents (Pei *et al.*, 2007). Nous avons donc repris les contraintes les plus utilisées dans la littérature. La contrainte de *fréquence minimale* (nombre d'occurrences minimum d'un motif dans les données d'apprentissage, contrainte également appelée *support minimal*), la contrainte de *gap* (qui définit l'écart minimal et maximal – en nombre d'itemsets de la séquence – entre deux itemsets consécutifs d'un motif séquentiel) et une contrainte d'appartenance spécifique à notre tâche (un motif doit contenir l'item #symptom#).

2.4.1 Contrainte de similarité sémantique

En plus de l'utilisation des contraintes de fréquence minimale de gap maximal et d'appartenance, nous présentons dans cet article une nouvelle contrainte basée sur les propriétés distributionnelles des mots, estimées à partir d'un corpus non annoté de très grande taille.

Nous avons remarqué pendant nos expérimentations initiales que nous avions un haut niveau de redondance dans les motifs extraits, comme des successions de conjonction ou de prépositions par exemple. Nous avons conçu une contrainte avec une limite sur la similarité entre deux items adjacents d'un motif. Par conséquent, cette contrainte est conçue pour élaguer les motifs redondants, peu informatifs.

Pour être capable de quantifier le niveau de redondance d'un motif, nous avons utilisé le principe de distributionalité (Harris, 1954) : des mots qui apparaissent dans le même contexte ont tendance à avoir le même sens. Cette hypothèse est la base des modèles comme *Word2Vec* (Mikolov *et al.*, 2013a,b), qui apprennent une représentation vectorielle continue de faible dimension des mots à partir

d'une grande quantité de texte. Le modèle *Word2Vec* appris est chargé par l'algorithme de fouille de motifs, chaque $item$ aura donc un vecteur associé permettant de mesurer la distance cosinus entre deux $items$ consécutifs. Si la distance cosinus entre l'$item$ courant et le précédent est supérieur à un paramétrage donnée par l'utilisateur, le motif est élagué puisque considéré comme redondant.

Pour entraîner le modèle Word2Vec, nous avons extrait 7 031 643 résumés d'articles biomédicaux, soit 8,7Go de textes. Le modèle final de 1,10Go contient au total 1 373 138 mots issus d'un contexte biomédical.

3 Expérimentations et Résultats

3.1 Résultats individuels des différents modules

Dans cette section, nous comparons les résultats de reconnaissance de symptômes, présentés dans la Table 1, pour chacun des modules de reconnaissance séparément des autres.

Module	Paramètres	Précision	Rappel	F-Score
Dictionnaire		**57,58**	14,00	22,53
CRF	mot	**56,31**	14,25	22,75
CRF	ngram, ngramLength=6	**56,14**	15,72	24,57
Fouille	freq=0.05%, gap=0, dist=0.4	23,12	**38,57**	**28,91**

TABLE 1 – Détails des meilleurs résultats pour chaque système

3.1.1 Résultats de la projection de dictionnaire.

Nous avons d'abord créé une baseline par projection des 34 257 symptômes que nous avons recueillis dans notre dictionnaire sur les données de test. De tous nos résultats, cette baseline a la meilleure précision mais le pire rappel.

3.1.2 Résultats du CRF

Le module CRF a appris avec succès une certaine connaissance du contexte, car il y a une augmentation de 12% [7] dans le rappel et 9% en F-mesure pour une petite baisse de 2,5% de précision. Cette tendance est la même quel que soit le type de caractéristiques utilisé, sac de mots ou de n-grammes.

3.1.3 Résultats de la fouille de motifs séquentiels

Puisque nous avons remplacé chaque symptôme par l'$item$ $\#symptom\#$ dans les données d'entrainement, nous ne représentons que le contexte. Un motif comme $< \{such, JJ\}\{as, IN\}\{\#symptom\#\} >$ peut être appliqué sur les données de test et découvrir des symptômes qui n'étaient pas dans les données d'apprentissage. Par conséquent, l'approche par

7. Dans cette section, la comparaison des scores est donnée en valeur relative.

fouille de motifs séquentiels montre une augmentation de 175% en rappel, mais une perte de 60% en précision pour une amélioration finale du F-score de 28%. Il était à prévoir que la précision chuterait, mais dans notre contexte d'aide au traitement de données pour un expert humain, le rappel est la mesure à maximiser : manquer un potentiellement nouveau symptôme est plus gênant que de produire un grand nombre de faux-positifs. La Figure 3 liste quelques-uns des motifs extraits. La Table 2 montre les différences entre les annotations de chaque approche et les annotations de nos experts humains.

$< \{development, NN\}\{IN\}\{\#symptom\#\} >$: 45
$< \{treatment, NN\}\{IN\}\{\#symptom\#\} >$: 62
$< \{development, NN\}\{of, IN\}\{\#symptom\#\} >$: 43
$< \{patient, NNS\}\{with, IN\}\{\#symptom\#\} >$: 295
$< \{diagnosis, NN\}\{IN\}\{\#symptom\#\} >$: 98
$< \{patient, NNS\}\{IN\}\{\#symptom\#\} >$: 306
$< \{case, NN\}\{of, IN\}\{\#symptom\#\} >$: 48
$< \{such, JJ\}\{as, IN\}\{\#symptom\#\} >$: 91
$< \{IN\}\{patient, NNS\}\{IN\}\{\#symptom\#\} >$: 163
$< \{NNS\}\{such, JJ\}\{as, IN\}\{\#symptom\#\} >$: 46
$< \{in, IN)\}\{patient, NNS\}\{IN\}\{\#symptom\#\} >$: 89
$< \{in, IN\}\{patient, NNS\}\{with, IN\}\{\#symptom\#\} >$: 88
$< \{IN\}\{patient, NNS\}\{with, IN\}\{\#symptom\#\} >$: 161
$< \{NN\}\{IN\}\{patient, NNS\}\{with, IN\}\{\#symptom\#\} >$: 69

FIGURE 3 – Meilleurs motifs extraits, format : "<motif> : nombre d'occurrences"

Phrase	diffuse	palmoplantar	keratoderma	and	precocious
Annotation Expert	B-S.	I-S.	I-S.	O	B-S.
Annotation Fouille	B-S.	I-S.	I-S.	I-S.	I-S.
Annotation CRF	B-S.	I-S.	I-S.	O	O

Phrase	primary	immunodef.	disorders	with	residual	cell-mediated	immunity
Expert	B-S.	I-S.	I-S.	O	O	O	O
Fouille	B-S.	I-S.	I-S.	I-S.	I-S.	I-S.	I-S.
CRF	O	B-S.	O	O	O	O	O

TABLE 2 – Comparaison des annotations de chaque module avec ceux de l'expert.

3.2 Impact de la quantité de donnée d'entrainement

La Figure 4 montre la variation en précision, rappel et F-score de nos modules avec l'augmentation de la taille des données d'entrainement. Chaque approche est sensible à la quantité de données disponible pour l'entrainement.

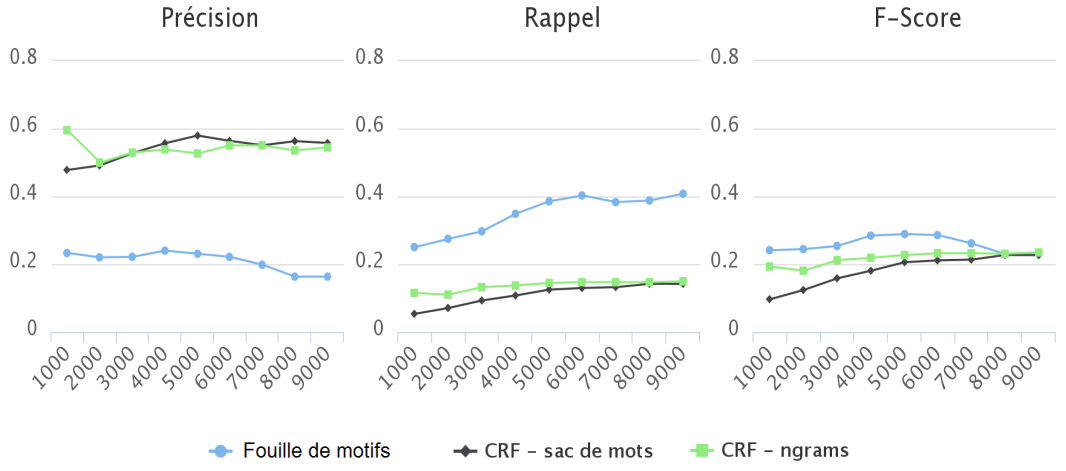

FIGURE 4 – Impact de la quantité de donnée d'apprentissage (en nombre de résumés) sur les scores de classification

3.3 Impact de la contrainte de similarité sémantique

Le but de la contrainte de similarité sémantique est de réduire le nombre de motifs extraits pour faciliter la lecture des modèles sans compromettre la mesure de performance. Comme avec toute contrainte en fouille de motifs le choix des paramètres doit être fait avec quelques précautions. La Figure 5 montre la réduction significative du nombre de motifs extraits. Elle montre également qu'un paramétrage trop bas de la contrainte réduit la performance. Avec une similarité sémantique en dessous de 0,4 la contrainte a tendance à produire des motifs qui ne sont pas aussi bons. En terme de distance cosinus, une valeur de 0,2 est si basse que des mots sémantiquement différents seraient considérés comme similaire. Avec un paramétrage intelligent de la contrainte de similarité sémantique il est donc possible de réduire le nombre de motifs extraits sans impact négatif sur les performances de reconnaissance des symptômes.

3.4 Analyse de l'apprentissage itératif

Pour cette expérience, contrairement aux résultats présentés précédemment, le processus d'annotation complet est itéré un certain nombre de fois pour essayer de découvrir de nouveau symptômes. Pour la fouille de motifs séquentiels, il n'y a pas d'amélioration parce qu'après la première itération l'item $\#symptom\#$ se propage dans les séquences, les nouveaux symptômes sont donc bruités. Nous ne trouvons pratiquement jamais de nouveaux symptômes après la première itération. Une approche intéressante serait d'essayer cette méthode itérative sur des jeux de données différents à chaque itération. Le comportement du CRF est différent, il découvre de nouvelles entités à chaque itération, mais malheureusement c'est principalement du bruit puisque la performance globale a tendance à diminuer.

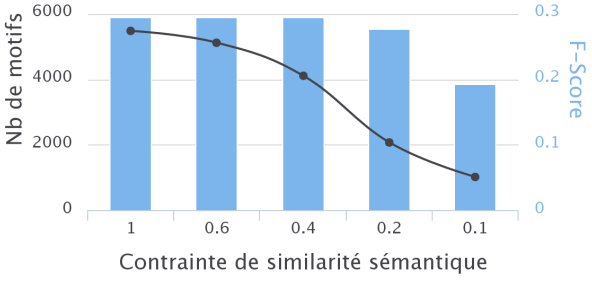

FIGURE 5 – Impact de la contrainte de similarité sémantique (abscisse) sur le nombre de motifs extraits (ligne noire) et le score de classification (hauteur des colonnes) en utilisant ces motifs comme modèle . Fréquence minimum de 0.125% et Gap maximum de 3.

3.5 Analyse de la combinaison de modèle

La Figure 4 montre clairement la différence de ratio précision/rappel entre l'approche numérique et l'approche symbolique des modèles même si ils sont très proches en termes de F-score. Parce que le modèle CRF maximise la précision et que le modèle de fouille maximise le rappel, nous avons essayé de combiner les résultats du meilleur modèle de CRF et du meilleur modèle de fouille de motifs séquentiels. Pour ce faire nous avons comparé chaque observation une à une : si les deux modèles sont d'accord sur l'annotation on ne change rien. En cas de désaccord si l'un des deux modèles a l'annotation *B-symptom* et que l'autre est *I-symptom* ou *O*, nous choisissons *B-symptom*. Nous maximisons le label *B-symptom* parce que nous avons constaté qu'il y a beaucoup d'énumérations dans les résumés et parce que la couverture des motifs à tendance à tout englober. La précision du CRF permet justement de segmenter cette couverture en plusieurs petits morceaux plus précis. Cette combinaison donne un score supérieur au score individuel des modules (Table 3). Ce même principe de combinaison a été réalisé entre chacun de nos modèles et une projection du dictionnaire de symptôme utilisé pour la génération du corpus d'apprentissage. On remarque que la projection améliore sensiblement les résultats individuels de chacun de nos modules. Étonnamment la tâche de segmentation effectuée par le module CRF est faite tout aussi efficacement avec la projection du dictionnaire de symptômes. On constate même que lorsque l'on combine un modèle de fouille avec la projection du dictionnaire, l'ajout du modèle CRF n'a pratiquement plus aucun impact. Étant donné que nous générons le corpus d'apprentissage en projetant le dictionnaire de symptômes sur des résumés d'articles, le CRF est en fait une généralisation du dictionnaire. Nous pensons que c'est la raison pour laquelle la combinaison du meilleur modèle de fouille et du meilleur modèle CRF est à 100% identique à la combinaison de ces mêmes modèles avec la projection du dictionnaire (Table 3). Ceci montre que l'ajout du dictionnaire à la combinaison des deux modèles n'apporte pas d'amélioration en phase de détection.

Système	Précision	Rappel	F-Score
Dictionnaire	**57.58**	14.00	22.53
CRF - ngram	**56.14**	15.72	24.57
Fouille sémantique	23.12	**38.57**	28.91
Combinaison	23.46	**39.31**	**29.38**
CRF - ngram et dictionnaire	**56.90**	16.22	25.24
Fouille sémantique et dictionnaire	23.35	**39.07**	29.23
Combinaison	23.46	**39.31**	**29.38**

TABLE 3 – Détails du meilleur résultat pour chaque système et de la combinaison de ces deux meilleurs systèmes en un seul corpus annoté.

4 État de l'art

À notre connaissance, il n'existe pas de jeu de données annotées qui pourrait être utilisé pour apprendre un modèle supervisé pour la reconnaissance de symptômes. La plupart des études sont basées sur des corpus de rapports ou des dossiers cliniques sans annotation de symptômes et par conséquent ces données ne peuvent pas être utilisées dans notre contexte de veille. Les corpus de ce genre comprennent le Mayo Clinic Corpus (Savova *et al.*, 2010) et le 2010i2b2/VA Challenge corpus (Uzuner *et al.*, 2011). D'autres jeux de données biomédicales existants annotent uniquement les maladies ; cela inclue le NCBI disease corpus (Doğan *et al.*, 2014) qui consiste en 793 résumés d'articles issus de PubMed avec 6 892 maladies annotées et 790 concepts uniques de maladies en liaison avec le Medical Subject Headings (MeSH) [8], et le Arizona Disease Corpus (AZDC) (Leaman *et al.*, 2009) qui contient 2 784 phrases issus de résumés d'articles de MEDLINE avec les maladies annotées en liaison avec le Unified Medical Language System (UMLS) [9].

La reconnaissance de symptômes (Martin *et al.*, 2014) est une tâche relativement nouvelle, souvent incluses dans des catégories plus générales comme *concepts cliniques* (Wagholikar *et al.*, 2013), *problèmes médicaux* (Uzuner *et al.*, 2011) ou *information phénotypique* dans (South *et al.*, 2009). Même dans ces catégories, peu d'études ont considéré le contexte linguistique dans lequel le symptôme apparait, elles sont plus concentrées sur l'analyse linguistique. (Savova *et al.*, 2010) remarque que la plupart des symptômes sont données en relation avec une localisation anatomique ; (Kokkinakis, 2006) relatent, après annotation de leur corpus avec MeSH, que 75% des signes cliniques et symptômes co-occurrent avec jusqu'à cinq autres signes cliniques et symptômes dans une même phrase. Aucun des travaux cités ci-dessus ne se sont penchés sur la reconnaissance entièrement automatisé de symptôme.

5 Conclusion

Nous avons présenté une approche permettant d'utiliser différents modèles d'apprentissages pour une tâche de reconnaissance de symptômes dans des textes biomédicaux. Nous avons dans cette étude utilisé un modèle CRF pour l'approche numérique qui a maximisé la précision, et un modèle de fouille de motifs pour l'approche symbolique qui a maximisé le rappel. Nous avons introduit une

8. www.nlm.nih.gov/mesh/meshhome.html
9. www.nlm.nih.gov/research/umls/

nouvelle contrainte sémantique pour l'extraction des motifs qui en réduit le nombre sans dégrader les performances de la tâche. Les deux approches (fouille et CRF) bien qu'étant de nature différente ont montré des résultats comparables en termes de F-score mais une répartition précision/rappel inversée. Au vu de cette observation nous avons combiné ces deux approches et amélioré ainsi les scores de classification. Dans un cadre peu supervisé où on ne dispose pas de corpus d'apprentissage annoté, cette combinaison semble donc prometteuse.

Il serait intéressant de poursuivre les expérimentations en utilisant différents corpus d'apprentissages entre chaque itération de l'apprentissage pour améliorer la détection de nouveaux symptômes non présents dans le dictionnaire. Une autre perspective est de tester si cette combinaison de modèles d'apprentissages numériques et symboliques est aussi intéressante sur d'autres tâches d'extraction d'information telle que l'extraction de relations dans le cadre faiblement ou non supervisée de l'*OpenIE (Open Information Extraction)*.

Remerciements

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'Avenir portant la référence ANR-10-LABX-0083 et du projet Hybride ANR-11-BS02-002.

Références

AGRAWAL R. & SRIKANT R. (1995). Mining Sequential Patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, p. 3–14, Washington, DC, USA : IEEE Computer Society.

BÉCHET N., CELLIER P., CHARNOIS T. & CRÉMILLEUX B. (2015). Sequence mining under multiple constraints. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing, Salamanca, Spain, April 13-17, 2015*, p. 908–914.

COHEN K. B. (2010). BioNLP : biomedical text mining. In *Handbook of Natural Language Processing, Second Edition*.

DOĞAN R. I., LEAMAN R. & LU Z. (2014). Ncbi disease corpus : a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, **47**, 1–10.

FINKEL J. R., GRENAGER T. & MANNING C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, p. 363–370 : Association for Computational Linguistics.

HARRIS Z. S. (1954). Distributional structure. *Word*, **10**(2-3), 146–162.

KOKKINAKIS D. (2006). Developing resources for swedish bio-medical text mining. In *Proceedings of the 2nd International Symposium on Semantic Mining in Biomedicine (SMBM)*.

LAFFERTY J. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. p. 282–289 : Morgan Kaufmann.

LEAMAN R., MILLER C. & GONZALEZ G. (2009). Enabling recognition of diseases in biomedical text with machine learning : corpus and benchmark. *Proceedings of the 2009 Symposium on Languages in Biology and Medicine*, **82**(9).

MARTIN L., BATTISTELLI D. & CHARNOIS T. (2014). Symptom extraction issue. In *Proceedings of BioNLP 2014*, p. 107–111, Baltimore, Maryland : Association for Computational Linguistics.

MÉTIVIER J.-P., SERRANO L., CHARNOIS T., CUISSART B. & WIDLÖCHER A. (2015). Automatic symptom extraction from texts to enhance knowledge discovery on rare diseases. In *Artificial Intelligence in Medicine*, p. 249–254. Springer International Publishing.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

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

PEI J., HAN J. & WANG W. (2007). Constraint-based sequential pattern mining : the pattern-growth methods. *Journal of Intelligent Information Systems*, **28**(2), 133–160.

RAMSHAW L. A. & MARCUS M. P. (1995). Text chunking using transformation-based learning. *arXiv preprint cmp-lg/9505040*.

SAVOVA G. K., MASANZ J. J., OGREN P. V., ZHENG J., SOHN S., KIPPER-SCHULER K. C. & CHUTE C. G. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes) : architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, **17**(5), 507–513.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, p. 44–49 : Citeseer.

SCHMID H. (1995). Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop* : Citeseer.

SOUTH B. R., SHEN S., JONES M., GARVIN J., SAMORE M. H., CHAPMAN W. W. & GUNDLAPALLI A. V. (2009). Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. *BMC bioinformatics*, **10**(9), 1.

SRIKANT R. & AGRAWAL R. (1996). Mining sequential patterns : Generalizations and performance improvements. In *Proceedings of the 5th International Conference on Extending Database Technology : Advances in Database Technology*, EDBT '96, p. 3–17, London, UK, UK : Springer-Verlag.

SUTTON C. & MCCALLUM A. (2011). An introduction to conditional random fields. *Machine Learning*, **4**(4), 267–373.

UZUNER Ö., SOUTH B. R., SHEN S. & DUVALL S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, **18**(5), 552–556.

WAGHOLIKAR K. B., TORII M., JONNALAGADDA S. R. & LIU H. (2013). Pooling annotated corpora for clinical concept extraction. *Journal of Biomedical Semantics*, **4**(1), 1–10.