Institut National des Langues et des Civilisations Orientales (INaLCO)

Département *Textes, Informatique, Multilinguisme*

Étude des forums de santé pour la détection d'événements secondaires

---

MÉMOIRE

MASTER TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Ingénierie Multilingue

par

Dalia MEGAHED

Directeurs de recherche :

Cyril Grouin
Pierre Zweigenbaum

présenté et soutenu en
octobre 2014

Remerciements

Ce travail de mémoire a été mené dans le cadre du Master-2 *Traitement Automatique des Langues (TAL)* à l'*Institut National des Langues et des Civilisations Orientales (INaLCO)*, plus précisément lors du stage de fin d'études effectué au *Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (LIMSI-CNRS)*.

Je tiens ainsi à remercier en premier lieu mon directeur de mémoire et mon encadrant de stage, Cyril Grouin, pour toute l'aide qu'il m'a apporté, ainsi que pour ses conseils et son suivi régulier pour que ce mémoire voie le jour.

J'exprime mes profonds remerciements à mon co-directeur de mémoire et co-encadrant, Pierre Zweigenbaum, dont les conseils instructifs ont été très éclairants.

Mes remerciements vont également au groupe *Information, Langues, Écrite et Signée (ILES)* du *LIMSI* pour m'avoir accueilli au sein du laboratoire et permis d'effectuer mon stage dans les meilleures conditions.

Je remercie à l'occasion tous mes professeurs de l'équipe *Textes, Informatique, Multilinguisme (ER-TIM)* de l'*INalCO*. Ce travail est aussi le fruit de leurs enseignements.

RÉSUMÉ

De nombreux travaux de l'état de l'art biomédical ont porté sur la détection d'événements secondaires à partir des rapports médicaux ou des réseaux sociaux à des fins de pharmacovigilance. L'objectif de ce mémoire consiste à construire, en s'appuyant sur certains de ces travaux, un système à base d'apprentissage statistique pour l'extraction des événements secondaires à partir des messages déposés par les patients sur les forums de santé. La méthode implémentée à cet égard repose sur deux grandes étapes, la première est consacrée à l'annotation d'une sous-partie du corpus choisie aléatoirement pour constituer une référence. Cette annotation est en effet effectuée selon un guide élaboré et une grammaire d'annotation définie. Elle comporte deux phases, une sans pré-annotation et une autre avec pré-annotation automatique, et est suivie d'une phase d'adjudication puis d'évaluation. La deuxième étape de la méthode mise en œuvre est dédiée à la création d'un modèle *CRF* et au choix de ces caractéristiques. Ces dernières sont choisies selon leurs pertinences par rapport au corpus étudié et conformément aux types d'entités qui se rapportent aux catégories traitées. Des expériences sont ensuite menées en vue d'évaluer plusieurs hypothèses de travail et tester la validité de la méthode adoptée. Les résultats de ces expériences varient selon la taille du corpus, la qualité des annotations de la référence ainsi que le sujet du forum traité. Les meilleurs résultats sont obtenus par un modèle global appris sur les deux forums étudiés (51.6 de F-mesure pour le corpus *antidépresseurs-anxiolytiques* et 65.52 pour le corpus *migraine*).

Mots-clés : extraction d'information, champs aléatoires conditionnels (CRF), forums de santé, pharmacovigilance, événements secondaires.

TABLE DES MATIÈRES

Liste des figures

Liste des tableaux

INTRODUCTION

Problèmatique de recherche

L'activité pharmacologique et thérapeutique de la gamme des médicaments présente sur le marché a révolutionné la pratique médicale. Toutefois, certains médicaments peuvent provoquer des événements secondaires, c'est-à-dire des réactions positives ou négatives inattendues liées à la prise d'un traitement médical. [Ginn et al., 2014] parlent d'une étude récente [Kongkaew et al., 2008] qui montre que 5.3% des médicaments prescrits sont associés à des événements secondaires. Parmi ces réactions, certains sont difficiles à prévoir par les résultats d'expérimentation toxicologique ou les essais cliniques effectués pendant une durée relativement courte. Cette difficulté émerge notamment en raison des différences qui existent entre les populations exposées aux expérimentations (sexe, âge ou état morbide) ou des changements de spécifications des médicaments postérieurement aux essais.

Par conséquent, la surveillance des réactions inattendues des médicaments, au cours des essais cliniques et de l'utilisation post-commerciale du médicament par les patients, semble cruciale. Cette surveillance fait aujourd'hui partie intégrante du processus de développement du médicament. Elle se concrétise par la pharmacovigilance, activité dont le but principal est la prédiction, aussi rapidement que possible, des événements secondaires provoqués par certains médicaments. Selon l'*Organisation Mondiale de la Santé (OMS)*, *« on entend par pharmacovigilance la notification, l'enregistrement et l'évaluation systématique des réactions adverses des médicaments délivrés avec ou sans ordonnance »* [OMS, 1969].

Pharmacovigilance

Dans un contexte où les événements secondaires sont considérés parmi les causes de décès principaux dans de nombreux pays [Leaman et al., 2010] (quatrième cause de dèces aux États-Unis [Chee et al., 2011]), la sécurité des médicaments constitue une source de préoccupation majeure pour les patients. Depuis les années 1960, l'*OMS* se charge de la détection des événements secondaires provoqués par certains médicaments à l'aide de systèmes de pharmacovigilance appelés également des *systèmes d'alerte précoce*. Le but de ces systèmes est de recueillir des informations concernant les événements secondaires, notamment les événements rares et inconnus, le plus tôt possible après la commercialisation des médicaments.

D'une manière générale, les systèmes de pharmacovigilance reposent sur les notifications des médecins et des pharmaciens. Certaines compagnies pharmaceutiques sont aussi concernées par une telle surveillance. De même, un nombre croissant de pays permettent aux patients de rapporter les événements secondaires potentiels directement aux systèmes de pharmacovigilance. Les essais cliniques et les rapports post-commercialisation rendus accessible par des organismes de santé ou les centres de pharmacovigilance sont ainsi parmi les sources principales d'infor-

mation concernant les événements secondaires. D'autres moyens sont utilisés récemment à des fins de pharmacovigilance comme les requêtes de recherche sur le Web, [White et al., 2013] et [Yates et al., 2013], et les réseaux sociaux. Ces derniers signalent les sujets *tendances*, soit les sujets de discussion les plus populaires, ce qui peut être utile pour les systèmes de pharmacovigilance quand il s'agit de discussions portant sur des sujets médicaux, [de Quincey and Kostkova, 2010] et [Parker et al., 2013].

Objectifs

L'enjeu de ce travail est d'extraire les événements secondaires dans les messages des patients à des fins de pharmacovigilance. Notre problématique de travail repose sur l'extraction automatique des événements secondaire tels qu'ils sont rapportés par les patients sous traitement dans des messages postés sur des forums de santé. L'analyse des résultats obtenus nous permet de dégager les caractéristiques propres aux forums de santé, plus spécifiquement de vérifier si certains événements secondaires rapportés par les patients sont connus et plus fréquents que d'autres.

L'objectif principal de ce travail consiste à implémenter une méthode qui repose sur un système à base d'apprentissage statistique capable de détecter automatiquement les événements secondaires, tout en s'inspirant de certains travaux déjà existants dans l'état de l'art. À cette fin, la méthode adoptée repose sur deux grandes étapes, la première concerne l'annotation manuelle du corpus, en travaillant sur des documents avec et sans pré-annotation automatique, et la seconde vise la mise au point d'un système par apprentissage statistique fondé sur les annotations manuelles précédemment produites. Un tel outil pourrait ainsi être utile à des analyses de données biomédicales grâce au traitement d'un gros volume de données avec un minimum d'intervention humaine.

D'un point de vue linguistique, l'objectif de cette étude s'appuie sur l'observation des particularités du discours non-spécialisé des patients et la comparaison de ces dernières avec celles utilisées par les professionnels de santé. Ces spécificités s'avèrent présentes dans le corpus traité et sont étudiés par l'analyse des résultats des différentes expériences réalisées, ainsi que par certaines informations quantitatives relatives aux données qui constituent le corpus.

Le présent mémoire s'articule autour de trois parties. La première (chapitre 1, page 15) dresse un état de l'art répondant à notre problématique de recherche. La deuxième (chapitre 2, page 25) présente le corpus étudié et la méthodologie adoptée pour la création du modèle de détection d'événements secondaires et la méthode implémentée pour l'apprentissage supervisé. Nous présentons de même les outils utilisés et certains scripts développés pour répondre à nos besoins. Enfin, la troisième (chapitre 3, page 45) expose les résultats d'évaluation du système accompagnés d'exemples de tests effectués. Une discussion s'ensuit pour mettre l'accent sur les problèmes rencontrés.

Projet Vigi4MED

Ce travail s'inscrit dans le cadre des recherches du projet *Vigi4MED* porté par le *centre hôspitalier uinversitaires (CHU)* de *Saint-Étienne* en collaboration avec d'autres partenaires, tels que le *centre de pharmacovigilance (CRPV)* de *Saint-Étienne*, le *CRPV* de l'*Hôpital Européen de Georges Pompidou (HEGP)*, le *Labora-*

toire d'informatique pour la mécanique et les sciences de l'ingénieur (LIMIS-CNRS), l'École des Mines de Saint-Étienne, l'*U1138 Eq22* (*Institut national de la santé et de la recherche médicale (INSERM)* et HEPGP) ainsi que l'*U1142* (INSERM).

L'objectif de ce projet financé par l'*Agence nationale de sécurité du médicament et des produits de santé (ANSM)* est double. Il s'agit d'identifier dans un premier temps les nouveaux événements secondaires rapportés par les patients sur les forums de santé. Dans un deuxième temps, *Vigi4MED* vise à analyser les risques de l'utilisation des médicaments hors autorisation de mise sur le marché (AMM). Ce projet se focalise sur certains cas d'étude, à savoir les pilules contraceptives, les antiacnéiques notamment *Diane 35* et les anticalvites tels que *Propecia*. Deux volets sont ainsi pris en compte dans ce projet de deux ans et demi, un rétrospectif dans lequel s'inscrit ce mémoire de recherche et un autre prospectif consacré pour les médicaments sous surveillance.

CHAPITRE 1

ÉTAT DE L'ART

Sommaire

1.1 Fouille de textes biomédicaux

1.1.1 Contexte et objectifs

La recherche dans le domaine biomédical est en pleine expansion. De nombreux travaux décrivent des progrès réalisés dans le domaine biomédical quant à la recherche d'information, les méthodes d'évaluation et la construction de ressources. Les progrès scientifiques touchent également la reconnaissance d'entités biomédicales, la classification de textes, l'extraction de terminologies ou de relations et la génération d'hypothèses. D'après [Zweigenbaum et al., 2007], certains problèmes tels que la manipulation des abréviations peuvent être considérés comme résolus grâce à ce progrès, tandis que d'autres comme l'identification des gènes à partir des textes médicaux sont susceptibles d'être résolus prochainement. En revanche, nombreux problèmes relatifs à la fouille de textes biomédicaux persistent. Ces problèmes semblent présenter un défi stimulant et offrent des possibilités d'améliorations et de recherches intéressantes.

Dans le cadre du progrès réalisé, le nombre de publications concernant le domaine biomédical témoigne d'une croissance accélérée. Selon [Cohen and Hersh, 2005], le nombre d'articles indexés dans la base de données bibliographiques *Medical Literature Analysis and Retrieval System Online (MEDLINE)* s'accroit avec un taux annuel de 5 000 000 nouvelles citations. Dans ce contexte évolutif, les outils de fouille de textes et d'extraction de connaissances peuvent jouer un rôle dans l'exploitation de cette quantité importante d'information, notamment celles figurant dans les ressources biomédicales non-structurées qui renferment une quantité significative de données difficiles à extraire manuellement.

Plusieurs groupes de recherche tentent ainsi de mettre en œuvre des systèmes de recherche et d'extraction d'information destinés à de multiples usages. Le but de ces systèmes consiste à permettre aux experts, ainsi qu'aux utilisateurs, de se tenir

informer des publications pertinentes relatives à leur discipline ou à des disciplines connexes. Rendre ces systèmes utiles aux personnes qui s'intéressent à ce domaine constitue en effet le défi majeur de la fouille de textes biomédicaux dans les années à venir. [Cohen and Hersh, 2005] affirment que cette orientation vers l'utilisateur nécessite une meilleure compréhension des caractéristiques de la littérature biomédicale et la mise en application de méthodes pertinentes capables de vérifier l'utilité des systèmes élaborés. Ces derniers s'appuient sur des modules de reconnaissance d'entités qui correspondent à des concepts biomédicaux, ainsi que sur des modules d'identification de relations entre ces entités détectées. [Cohen and Hersh, 2005] et [Kang et al., 2014] considèrent la reconnaissance d'entités et l'établissement de relations entre ces dernières comme les composantes essentielles d'une tâche d'extraction d'information.

1.1.2 Reconnaissance d'entités et identification de relations

Les entités nommées sont des entités du texte qui appartiennent à des catégories prédéfinies introduites lors de la campagne *Message Understanding Conferences (MUC6)* en 1995. Ces catégories concernent par exemple des noms de personnes, lieux, produits et organisations. Dans notre cas, il s'agit plutôt d'entités biomédicales qui reflètent des concepts tels que les traitements, les événements secondaires et les pathologies ou des noms de gènes et de protéines. Le but de la reconnaissance de ces entités est d'identifier au sein d'un ensemble de texte toutes les occurrences d'une entité qui sont relatives à une catégorie spécifique. Il s'agit de l'identification des frontières d'une chaîne voulue et d'établir la correspondance entre cette chaîne et une catégorie prédéfinie.

Toutefois, ces frontières sont parfois difficile à être identifiées en raison du contexte et de la structure de phrases. Il en va de même pour le choix de la catégorie pertinente qui traduit la nature d'une entité. Certains termes ou expressions peuvent refléter différents concepts selon le contexte dans lequel ils figurent. À titre d'exemple, le terme *ferritine* peut faire référence à la protéine du corps qui règle le stockage de fer ou au test de sang qui évalue les réserves en fer de l'organisme. De même, certains concepts peuvent se rapporter à plusieurs termes selon le registre utilisé, produisant ainsi une certaine ambigüité. [Dai et al., 2010] témoignent d'une variété concernant les nouveaux noms de gènes et d'une absence de normalisation des noms utilisés (les auteurs utilisent des abréviations ou des variantes en fonction du milieu personnel). Ils indiquent également que les noms de gènes ressemblent dans un grand nombre de cas à des noms de cellules, tissus ou organes. Par exemple, *C1R* est une lignée cellulaire, mais aussi un gène *SwissProt P00736*. Les systèmes de fouille de textes doivent être ainsi en mesure de distinguer ces différents cas de figure.

De nombreux travaux de l'état de l'art ont abordé la question de la détection automatique des entités biomédicales et d'extraction de relations entre ces dernières. Certains ont mis l'accent sur les gènes, [McDonald et al., 2004] et [Demaine et al., 2006] ; et leurs relations avec les protéines, [Fundel et al., 2007] et [Bunescu et al., 2006] ; ou les pathologies, [Chun et al., 2006] et [Bundschus et al., 2008]. D'autres se sont intéressés aux protéines, [Gaizauskas et al., 2003] et [Fukuda et al., 1998] ; leurs interactions, [Bunescu et al., 2006] et [Albert et al., 2003] ; voire même leurs contrastes [jae Kim et al., 2006]. En outre, certaines études combinent les gènes et les protéines, telles que celles de [McDonald and Pereira, 2005], [Chang et al., 2004] et

[Hanisch et al., 2005] ; et d'autres ne limitent pas le problème à une catégorie précise comme [Zhou et al., 2004] et [Karamanis et al., 2007] qui traitent de l'extraction d'entités biomédicales en général. Enfin, il existe de nombreux travaux qui se focalisent sur l'extraction des pathologies, des événements secondaires et des traitements, ainsi que des associations entre eux. Ces travaux sont étroitement liés à notre sujet de recherche et sont mentionnés dans la section 1.2.

1.2 Repérage des événements secondaires et des pathologies

Plusieurs systèmes sont développés à des fins de pharmacovigilance pour repérer, soit à partir de la littérature médicale et les rapports médicaux, soit à travers les forums de santé et les réseaux sociaux, les événements secondaires et d'autres entités comme les pathologies, les symptômes et la posologie. Les méthodes sur lesquelles reposent ces systèmes varient en effet suivant la tâche à accomplir et les spécificités du corpus traité. Outre la co-occurrence des entités qui est considérée comme l'approche de base (*baseline*) dans de nombreux travaux, les approches se divisent généralement en deux grandes catégories, les approches symboliques qui sont des méthodes à base de règles, combinées parfois avec des projections de lexiques et de dictionnaires ou des techniques de traitement automatique des langues (TAL), et les approches statistiques qui se basent sur un apprentissage automatique prenant en compte certaines caractéristiques intrinsèques aux entités, ainsi que des caractéristiques ontologiques ou textuelles. Des approches mixtes (ou hybrides) alliant des techniques qualitatives et d'autres quantitatives sont également utilisées.

1.2.1 Littérature médicale et rapports structurés

Diverses études portent sur l'extraction des événements secondaires au travers des documents structurés ainsi que des documents semi-structurés. En effet, certains ont travaillé sur des dossiers médicaux électroniques, d'autres ont eu recours aux bases de données médicales, telles qu'*EU ADR*, *Mini-Sentinel* ou *DrugBank*, et d'autres encore ont utilisé les connaissances extraites à partir de sources d'informations médicales comme l'*Unified Medical Language System (UMLS)* ou la *FDA Adverse Event Reporting System (FAERS)*.

1.2.1.1 Approche de base (*baseline*)

Parmi les travaux de l'état de l'art médical dans le cadre de la pharmacovigilance, quelques uns ont eu recours à l'approche de base pour repérer les événements secondaires provoqués par certains médicaments et les informations associées. Cette approche ne nécessite pas une analyse linguistique, mais repose sur la fréquence des entités et leur co-occurrence au sein d'un même contexte. Deux entités seront probablement liées si elles figurent au sein d'une fenêtre déterminée. La longueur de cette dernière peut varier (entre une phrase, un paragraphe ou un texte) selon l'objectif de l'étude. Bien qu'elle favorise le rappel au sujet de la précision, cette simple approche est souvent considérée comme une *baseline* pour évaluer les nouvelles méthodes implémentées.

[Chena et al., 2008] ont combiné cette *baseline* avec d'autres méthodes pour améliorer la performance de leur système et aboutir à des résultats plus satisfaisants. Ils

ont appliqué des tests statistiques et des techniques de TAL sur un ensemble de 818 282 articles extraits de *MEDLINE* et 48 360 rapports médicaux de *New-York Presbyterian Hospital* pour identifier des associations entre les traitements et les pathologies. Ils ont utilisé ces tests statistiques pour calculer et évaluer la force d'association entre chaque pathologie et les huit médicaments étudiés. Leur système arrive à identifier des pairs de traitement-pathologie fortement associés, tels que la relation entre la *maladie de Parkinson* et *Carbidopa* (χ^2 varie entre 0.59 et 1 selon l'ensemble du corpus traité), ainsi que d'autres qui le sont moins comme le lien entre la *pneumonie* et *Ceftazidime* (χ^2 entre 0.15 et 0.19). Selon les auteurs, de tels chiffres donnent une idée sur la nature d'un traitement : *Carbidopa* est un traitement spécifique lié à une maladie particulière, tandis que *Ceftazidime* est un médicament plus générique inscrit aux patients pour traiter divers problèmes (*Ceftazidime* est en effet un antibiotique pour traiter les infections et est utilisé parfois pour remédier à certaines conditions concomitantes).

1.2.1.2 Approche symbolique

Par ailleurs, de nombreux travaux se sont intéressés à la détection des événements secondaires dans le cadre de la pharmacovigilance grâce aux méthodes symboliques. Au niveau symbolique, des règles (patrons générés par les experts du domaine selon leur point de vue linguistique souvent sous la forme d'expressions régulières) sont définies manuellement en s'appuyant sur certaines caractéristiques du contexte telles que les préfixes, les suffixes ou les étiquettes morpho-syntaxiques. De plus, certains systèmes à base de connaissances allient ces règles à des méthodes de TAL et d'autres projettent des lexiques, constitués à l'aide de ressources déjà existantes ou créées manuellement, pour pallier les lacunes des règles définies.

À titre d'exemple, [Kang et al., 2014] ont développé un système à base de connaissances reposant sur les informations qui figurent dans l'*UMLS* pour identifier des relations entre les événements et les traitements à partir des résumés de *MEDLINE*. Leur but est de démontrer la possibilité d'adopter une approche symbolique qui ne nécessite pas un corpus d'apprentissage de taille significative pour réaliser une telle tâche. Ce système est ainsi constitué de deux modules, le premier est consacré à l'extraction de concepts qui se rapportent aux deux catégories étudiées et le second permet d'établir les relations entre les entités extraites au sein d'une même phrase. Les auteurs témoignent ainsi d'une augmentation de 34.4% de la F-mesure par rapport aux résultats fournis par une simple méthode de co-occurrence. Malgré cette augmentation, la F-mesure reste faible (50.5) et le système élaboré ne permet pas de distinguer entre les relations traitement-événement secondaire et celles traitement-pathologie.

Bien que les systèmes qui reposent sur des méthodes symboliques offrent des résultats relativement satisfaisants, certains s'avèrent dépendants des ressources linguistiques ou d'une analyse morpho-syntaxique. [Zweigenbaum et al., 2007] estiment que l'intérêt majeur des méthodes à base de règles se concrétise par l'augmentation de la précision, souvent au prix d'un rappel significativement plus faible. Il semble en effet coûteux pour un être humain de définir des règles robustes pour traiter les différentes structures possibles. De plus, [Kang et al., 2014] parlent d'un problème qui se rapporte à l'ambigüité des termes et la présence d'un nombre élevé de variantes pour chaque entité. Selon eux, cette difficulté peut se traduire par l'accumulation des règles créées. C'est ainsi que certains se sont tournés vers d'autres approches

différentes que celles à base de règles. Il s'agit par exemple des approches à base d'apprentissage statistique ou des méthodes hybrides.

1.2.1.3 Approche statistique

Contrairement aux méthodes symboliques, les méthodes statistiques ne s'appuient pas sur la combinaison d'heuristiques, mais reposent sur un apprentissage automatique basé sur des calculs statistiques. Différents algorithmes d'apprentissage automatique tels que les *Support Vector Machine (SVM)* [Kazama et al., 2002], les *modèles de Markov cachés* [Zhao, 2004], les *modèles de Markov à entropie maximale* [Finkel et al., 2004] sont conçus à cet égard. De même, certains travaux tentent d'améliorer la performance des systèmes élaborés, en combinant les résultats de plusieurs algorithmes. [Smith et al., 2008] et [Si et al., 2005] témoignent d'une meilleure exactitude en associant plusieurs algorithmes dans leurs travaux. Dans la plupart du temps, la performance des systèmes à base d'apprentissage statistique est satisfaisante, mais elle est dépendante de la taille du corpus sur lequel le modèle créé a été appris. Les systèmes d'apprentissage automatique nécessitent souvent des jeux de données de taille significative.

Un premier exemple des travaux reposant sur l'apprentissage statistique est celui de [Harpaz et al., 2010]. Ces derniers ont étudié le recours aux approches statistiques pour identifier les événements secondaires provoqués par l'interaction de certains médicaments. Ils ont implémenté une chaîne de traitements fondée sur l'algorithme d'*Apriori* [Agrawal et al., 1993] afin de traiter les notifications spontanées de la *U.S. Food and Drug Administration (FDA)*. Les pairs traitement-événement secondaire extraits sont ensuite filtrés selon certains critères définis. Parmi ces critères figure la force d'association entre les deux entités qui doit atteindre un seuil minimal de support. Ils ont réussi à identifier 3 402 associations traitement-événement secondaire à partir de 163 944 rapports médicaux.

En outre, certains ont traité le problème de détection des événements secondaires comme un problème de classification, notamment en ayant recours aux *SVM* et aux *champs aléatoires conditionnels (CRF)* qui utilisent différents types de caractéristiques (contextuelles, ontologiques, textuelles, etc.) pour pouvoir réaliser une catégorisation. Par exemple, [Wang et al., 2011] ont comparé la performance de plusieurs algorithmes tels que les *SVM*, *Naïve Bayes* et la *régression logistique*, en faisant varier les caractéristiques utilisées dans chaque test. Ils ont défini 21 caractéristiques ontologiques et 14 caractéristiques textuelles pour pouvoir repérer des événements secondaires dans les articles, portant sur la neutropénie, de *PubMed*, le moteur de recherche hébergé par la *Bibliothèque américaine de médecine des instituts américains de santé* qui donne accès à la base de données bibliographiques *MEDLINE*. Leur système enregistre ainsi une exactitude de 0.5 en utilisant uniquement les caractéristiques ontologiques, 0.75 pour les caractéristiques textuelles et 0.6 en combinant les deux types de caractéristiques ensemble. Ils ont en effet prouvé qu'il n'existe pas une grande différence en ce qui concerne la performance des calssifieurs, mais ce sont plutôt les caractéristiques qui jouent un rôle important dans la tâche de classification.

[Roberts et al., 2008] s'appuient, à leur tour, sur un système d'apprentissage automatique supervisé appris sur les caractéristiques extraites des rapports oncologiques en vue de détecter les relations médicales qui existent dans les rapports des patients. Ils entrainent ainsi un classifieur *SVM* pour étudier la variation de performance du système élaboré selon les différentes distances entre les entités à extraire (si les en-

tités se trouvent au sein de la même phrase ou s'étalent sur plusieurs phrases différentes) et la taille du corpus d'apprentissage. [Rink et al., 2011] développent un système de détection de relations entre les traitements, les problèmes (parmi lesquels figurent les événements secondaires) et les tests médicaux pour la tâche d'extraction de relations du challenge *i2b2/VA* de 2010. Le système repose sur un classifieur *SVM* prenant en compte des caractéristiques lexicales, syntaxiques, sémantiques et contextuelles. Enfin, [Gurulingappa et al., 2012] adaptent un système d'apprentissage automatique conçu pour l'extraction de relations relatives aux événements secondaires qui atteint 87.0 de F-mesure.

Tandis que les hyperplans *SVM* qui séparent l'espace dans lequel sont représentées les données en deux sous-espaces fournissent de bonnes résultats, certains travaux ont tendance à détecter automatiquement les événements secondaires des médicaments par le biais des *CRF*. Plus précisément, il s'agit de la création d'un modèle probabiliste discriminant d'apprentissage supervisé fondé sur les distributions conditionnelles. Le principe de base des *CRF* repose en effet sur l'observation de comportements similaires dans des contextes proches. Par exemple, [Bundschus et al., 2008] ont travaillé sur l'identification de relations pahtologie-traitement figurant dans les résumés de *PubMed* (ainsi que sur la classifiaction de relations entre gènes et pathologies qui figurent dans la base de données *Human GeneRIF*). Ils étendent ainsi le cadre de l'utilisation des *CRF* pour considérer des annotations de relations sémantiques. Ils concluent qu'une telle méthode est en mesure de déduire des relations biomédicales avec une exactitude assez satisfaisante (96.9%).

1.2.1.4 Approches hybrides

Pour palier les biais théoriques des approches symboliques et combler les lacunes des méthodes statistiques, une des solutions consiste à allier la performance des systèmes statistiques avec la connaissance des méthodes symboliques. Certains recherches ont eu recours aux méthodes hybrides en combinant l'apprentissage statistique avec d'autres approches. À titre d'exemple, [Wang et al., 2009] ont démontré la validité de l'utilisation des outils de TAL avec les méthodes statistiques à des fins de pharmacovigilance. Malgré l'obtention d'une faible précision (31%), ils ont réussi à identifier 75% des associations traitement-pathologie connues, en reposant sur la fréquence de co-occurrence de ces deux entités et en vérifiant la significativité statistique des résultats par le test du χ^2. Selon eux, cette précision peu satisfaisante peut être interprétée par les nouveaux événements secondaires correctement détectés.

De même, [Minard et al., 2011] traitent plusieurs types de relations entre les entités identifiées lors du challenge *i2b2/VA* [Uzuner et al., 2011] en 2010 pour extraire automatiquement les concepts médicaux (par les *CRF*) et les relations entre ces derniers (par les *SVM*). Il s'agit en effet de relations qui figurent entre les problèmes (parmi lesquels se trouvent les événements secondaires) et les traitements tels que les traitements qui améliorent un problème, les traitements qui dégradent un problème, les traitements qui causent un probème (cas des événements secondaires), les traitements qui peuvent être adminstrés pour un probème et ceux qui ne peuvent pas l'être. Les auteurs confirment que les méthodes hybrides fournissent de milleurs résultats, les méthodes statistiques sont très dépendantes des données d'apprentissage annotées (c'est-à-dire, les classes les mieux représentées fournissent de meilleurs résultats) et les méthodes symboliques ne semblent pas suffisantes face aux nouveaux types de données.

[Li et al., 2013] parlent d'un système hybride appelé *AutoMCExtractor* destiné à recuillir des relations médicales y compris les liens entre traitement-événement secondaire et traitement-pathologie qui figurent dans les étiquettes des médicaments publiées par la *FDA*. Ces étiquettes ont été prétraitées pour en extraire le contenu textuel et identifier huit sections connexes (ex. : événement secondaire, précaution, etc.). Leur méthode repose en effet sur les *CRF* entrainés sur des caractéristiques lingusitiques, sémantiques et de surface, ainsi que sur des règles définies manuellement. Les résultats de leur système sont satisfaisants : il atteint une précision de 90%, un rappel de 81% et une F-mesure de 85. Ils estiment ainsi que l'extraction des relations médicales des étiquettes médicales peut être effectuée en utilisant des méthodes de TAL.

1.2.2 Forums de santé et réseaux sociaux

Par ailleurs, plusieurs études ont mis en évidence l'apport significatif des notifications des patients. Selon [Leaman et al., 2010] et [Yang et al., 2013], les notifications des patients sont de qualité similaire à celles des professionnels de santé. De plus, la motivation d'extraire les associations traitement-événement secondaire à partir des messages des patients semble être une conséquence de l'évolution récente de la population qui tend à discuter plus fréquemment des expériences personnelles sur les forums et les blogs plutôt que de les rapporter à des praticiens (effet blouse blanche), [Benton et al., 2011] et [Yeleswarapu et al., 2014]. Conformément à une étude menée par *Pew Internet & American Life Project* en 2003 [Fox and Fallows, 2003], environ 80% des adultes utilisant Internet, soit 93 millions d'Américains, ont consulté des pages Web portant sur 16 questions majeures de santé. Cela représente une hausse par rapport au 54% rapporté dans leur étude de 2000. De même, une enquête de *Solucient Survey of Healthcare* montre que 45% des utilisateurs ont eu recours à Internet pour obtenir des informations médicales contre 16% seulement qui ont consulté leurs médecins [Zielstorff, 2003]. Le suivi des messages publiés par les patients sur les sites médicaux peut ainsi alerter les compagnies pharmaceutiques et les organismes de réglementation à la fréquence et l'incidence des réactions sérieuses inattendues provoquées par les médicaments.

Malgré la disponibilité d'informations intéressantes sur les réseaux sociaux, peu de travaux ont abordé la détection des événements secondaires à partir de ce type de sources. [Sondhi et al., 2010] ont proposé de nouvelles caractéristiques spécifiques aux données issues des forums pour repérer des phrases connexes au sein de leur corpus tiré d'un forum médical. Ils ont classifié ces phrases, à l'aide des *SVM* et des *CRF*, en deux catégories : celles décrivant un problème médical (événement secondaire ou pathologie) et d'autres représentant un traitement. Selon leurs résultats, ils estiment que les nouvelles caractéristiques (position des phrases au sein du message et position du message au sein des fils de discussion) avec les caractéristiques sémantiques (groupes sémantiques de l'*UMLS*) augmente l'efficacité d'un système. Il s'agit en effet des caractéristiques spécifiques au domaine médical et relatives à la nature des forums qui s'avèrent pertinentes par rapport aux caractéristiques génériques comme le nombre de mots dans une phrase.

[Leaman et al., 2010] projettent un lexique constitué à partir de quatre sources différentes (metathesaurus de l'*UMLS*, base de données des événements secondaires *SIDER*, celle de *MedEffet Canada* et une liste d'expressions dialectales constituée manuellement) sur le contenu de 3 600 messages déposés sur le forum *DailyStrength*.

Leur système enregistre une F-mesure de 73.9, mais ne gère ni les erreurs de frappe, ni les expressions idiomatiques utilisées par les patients pour rapporter les événements secondaires. [Nikfarjam and Gonzalez, 2011] utilisent le même corpus pour mettre en place un système qui détecte les événements secondaires en se basant sur des patrons définis. Leur système obtient une F-mesure de 67.96. [Yang et al., 2013] ont élaboré un système de classification des messages des patients suivant deux exemples, positifs et négatifs. Les messages positifs sont ceux qui portent sur des événements secondaires et ceux négatifs sont les messages qui n'abordent pas des sujets relatifs aux événements secondaires. Selon eux, une telle classification démontre l'utilité des réseaux sociaux pour la surveillance post-commerciale des médicaments.

Enfin, certains travaux se sont focalisés sur un traitement particulier. À titre d'exemple, [Yates and Goharian, 2013] décrivent *ADTrace*, un système s'appuyant sur *Medsyn* et sept patrons définis manuellement pour extraire des trois sources *AskaPatient*, *Drugs* et *DrugRatingZ* les événements secondaires relatifs au cancer du sein. Il en va de même pour [Yates et al., 2013] qui proposent *DepADR*, une nouvelle méthode d'extraction d'événements secondaires reposant sur les relations de dépendances linguistiques et les *CRF*. Pourtant, l'évaluation des résultats montre que le système élaboré favorise la précision (61%) sur le rappel (32%).

Tous ces précédents exemples utilisent en effet les mêmes approches que celles utilisées pour traiter les textes de la littérature médicale et les documents structurés (symboliques, statistiques ou hybrides). Toutefois, ils se doivent de tenir compte des spécificités linguistiques des documents étudiés, notamment quand il s'agit de la correspondance entre le vocabulaire médical spécialisé des praticiens et celui du grand public qui caractérise les forums de santé et les réseaux sociaux.

Vocabulaire spécialisé *vs* vocabulaire grand public

Les patients utilisent souvent un langage familier qui diffère du jargon utilisé dans la pratique médicale. De nombreux travaux ont mis l'accent sur cette différence de registre qui peut entraver la communication entre les patients et les praticiens. Parmi ces travaux figurent par exemple ceux de [Tsea and Soergela, 2003], [Plovnick and Zeng, 2003], [Zeng and Tse, 2006], [Keselman et al., 2007], [Keselman et al., 2008], [Zeng-Treitler et al., 2008] ainsi que [Kandula et al., 2010]. Ces derniers parlent d'une différence entre les termes employés par un patient et ceux employés par un professionnel de la santé. Ils estiment que les deux s'expriment différemment pour décrire un même concept.

Dans le but de combler cette lacune de communication, certains auteurs tentent d'établir une correspondance entre le vocabulaire médical spécialisé et le vocabulaire non-spécialisé du grand public. Cette correspondance s'effectue au travers plusieurs moyens, mais repose principalement sur la simplification d'un texte, en identifiant les termes difficiles (à l'aide de la fréquence des termes ou de leurs contextes), en utilisant des synonymes plus fréquents ou en expliquant les termes.

Par exemple, [Zeng-Treitler et al., 2007], [Elhadad, 2006] et [Cimino et al., 1997] ont eu recours aux hyperliens et aux *info buttons* pour fournir aux utilisateurs les défintions des termes complexes. [Rocha et al., 1993] et [Zeng and Cimino, 1996] ont opté, à leur tour, pour la traduction des termes techniques. En effet, les premiers se sont appuyés sur une fonction seuil et les seconds ont utilisé l'*UMLS* comme ressource de connaissances pour détecter les termes qui posent plus de problèmes au lecteur non-spécialisé. En outre, [Deléger and Zweigenbaum, 2008a] et

[Deléger and Zweigenbaum, 2008b] ont aligné des segments de textes médicaux en vue de faciliter la compréhension des textes spécialisés par le lecteur. Leur méthode repose sur la constitution d'un corpus comparable (spécialisé/non-spécialisé) portant sur le domaine médical.

Tous ces travaux ont ainsi comme objectif d'effectuer le lien entre le langage des praticiens et celui des patients. Certains ont également comme but de constituer un vocabulaire médical contrôlé (*controlled consumer health vocabularies*) destiné au grand public. Ce dernier est défini comme un ensemble de concepts et expressions utilisé souvent par les patients lorsqu'il s'agit du domaine médical [Zeng and Tse, 2006]. Un tel vocabulaire paraît utile pour détecter les événements secondaires tels qu'ils sont exprimés par les patients dans les forums de santé et les réseaux sociaux.

CHAPITRE 2

MATÉRIEL ET MÉTHODE

Sommaire

Introduction

Comme mentionné dans la partie 1.2.1.3, un certain nombre de travaux présents dans l'état de l'art traitent le problème de détection d'événements secondaires publiés par les patients dans les forums de santé comme un problème de classification statistique résolu via les *champs aléatoires conditionnels (CRF)*. La performance de ces systèmes pour traiter un tel problème est assez satisfaisante (même en travaillant sur des données bruitées issues des médias sociaux [Ginn et al., 2014]), d'où l'intérêt de s'appuyer sur les *CRF* pour élaborer notre système de détection automatique d'événements secondaires.

La méthode appliquée en vue de mettre en place un tel système est détaillée dans cette partie. Cette dernière présente en effet le corpus étudié, illustre le processus d'annotation adopté afin de détecter les entités qui nous intéressent, notamment les événements secondaires, tout en mettant l'accent sur les étapes accomplies pour la constitution du modèle *CRF* et le choix de ses caractéristiques. Elle expose enfin les différentes configurations du modèle créé.

2.1 Constitution du corpus

2.1.1 Présentation du corpus

Dans le but de constituer un corpus pertinent pour notre étude, nous identifions tout d'abord les sources qui nous permet de sélectionner les forums de santé intéressants. Cette sélection a comme objectif d'aspirer les pages *HTML* qui répondent à nos besoins. Le choix des sources s'appuie en effet sur leur pertinence par rapport à notre sujet de recherche, ainsi que sur la facilité d'accéder à leurs contenus. Deux corpus sont ainsi extraits de la rubrique *forums* de *Doctissimo*, le premier comprend 13 847 fils de discussion portant sur les antidépresseurs et les anxiolytiques (corpus de base fournit en début de travail) et le second renferme uniquement 10 fils qui traitent de la migraine (corpus téléchargé pour évaluer certaines hypothèses de travail). Les deux corpus concernent des fils de discussion traités par les patients durant l'année 2014 et sont aspirés par la commande *wget*.

Tirées de la même source, les pages *HTML* qui constituent les deux corpus sont structurées d'une manière identique. Chaque page contient un nombre de messages publiés par les patients. Ces messages tournent autour d'un sujet précis (ex. : prise de poids lié au traitement *deroxat* ou les événements secondaires du *xanacs*), évoqué par la personne qui entame la discussion, et sont rattachés à un identifiant d'utilisateur (la personne qui a déposé le message) et son profil (ex. : patient, praticien, étudiant en médecine), ainsi qu'à une date de publication du message.

L'idée de se focaliser sur un tel forum se justifie par le fait qu'il représente une source d'informations intéressantes qui reflète les expériences personnelles des patients. De même, *Doctissimo* constitue un canal actif de communication entre plusieurs types d'utilisateurs, ce qui offre une diversité quant à la nature des informations publiées et limite les problèmes de mise à jour des forums. En effet, lors du traitement de données issues des médias sociaux, il semble crucial de prendre en compte les risques associés aux forums et réseaux sociaux, notamment lorsqu'il s'agit de la véracité et la pertinence des informations qui y sont publiées quotidiennement.

2.1.2 Prétraitement des données

L'aspiration des sources identifiées s'ensuit par une phase de prétraitement des données, durant laquelle une chaîne de prétraitements est implémentée. Cette dernière permet de faciliter l'exploitation des données afin d'extraire automatiquement les informations pertinentes à partir des pages Web aspirées. Parmi les différentes tâches effectuées dans ce prétraitement figure la transformation du corpus *HTML* en texte brut, le filtrage des données, le nettoyage des données récupérées, la production des fichiers tabulaires et l'évaluation de la sortie produite. La figure 2.1 illustre les différentes étapes concernant le prétraitement du corpus et l'annexe A.1.1 détaille le code élaboré pour mettre en œuvre cette chaîne.

- **Transformation du *HTML* au *txt* :** il s'agit de la transformation des pages aspirées au format texte à l'aide d'un parseur *HTML*, qui est dans notre cas le module *Perl Html::TableExtract*. Ce module permet d'extraire de chaque fichier du corpus le contenu des tableaux qui renferment les informations voulues.
- **Tri et filtrage des données brutes :** le filtrage des données est effectué à l'aide des expressions régulières qui permettent par exemple de supprimer des tableaux récupérés les lignes vides et ceux contenant des publicités. Elles aident

aussi à repérer le contenu des messages publiés, les identifiants des utilisateurs et leurs profils, ainsi que la date de publication du message. Plusieurs règles sont définies pour prendre en compte les diverses structures des fichiers *HTML*, soit les différents cas de figure tels que *profil supprimé*, *invité* ou *sans profil*.

- **Nettoyage des données récupérées :** les messages publiés par les utilisateurs sur les forums s'avèrent bruités et nécessitent d'être nettoyés et normalisés. Dans un premier temps, nous supprimons les citations (quand un utilisateur cite le message d'un autre avant de lui répondre) et certaines informations telles que « message édité par *nom utilisateur* » ou « message cité *nombre* fois ». Dans un deuxième temps, les messages sont normalisés et toutes marques typographiques liées aux forums sont supprimées.
- **Production des fichiers tabulaires :** la sortie finale de la chaîne de prétraitements correspond à un ensemble de fichiers tabulaires au format texte. Chaque fichier contient sept colonnes : le nom du fichier traité, un identifiant (compteur) pour chaque message publié, l'identifiant de l'utilisateur, son profil, la date de publication du message, son contenu textuel et le nom du médicament dont parlent les patients.
- **Évaluation de la sortie produite :** il s'agit de la vérification du code élaboré par le biais de commandes *Bash* afin de s'assurer de la pertinence des scripts réalisés et de la validité des résultats obtenus.

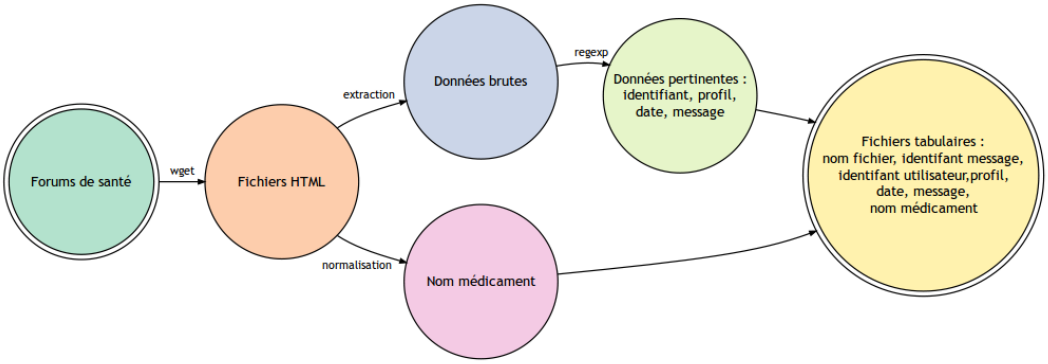


FIGURE 2.1 – Chaîne de prétraitements implementée

Le nom du médicament est en effet extrait du nom du fichier, soit directement, soit en calculant une distance d'édition (*Levenshtein*) entre l'ensemble des mots du nom de fichier et une liste de médicaments. Ce calcul semble nécessaire dans la mesure où les noms des fichiers téléchargés intègrent généralement le nom du médicament (61% des fichiers du corpus), mais ce dernier peut être écrit incorrectement (13.5%). À titre d'exemple, la normalisation du nom du médicament mal orthographié *xanacs* en *xanax* permet de rassembler les messages sous le nom correct du médicament.

Cette normalisation augmente la précision des informations extraites, ce qui semble utile puisque la suite du traitement repose sur les informations relatives aux noms des médicaments et aux événements secondaires. Le pourcentage de fichiers pour lesquels le nom du médicament n'est identifié ni directement, ni par la distance d'édition (38.8%) justifie le besoin d'un système de classification automatique qui attribue un nom de médicament à ces fichiers. Il est à noter que malgré cette normalisation, quelques noms de fichiers s'avèrent problématiques. C'est par exemple le cas du

fichier *ablyfy-remplacer-solian* pour lequel le traitement retenu est *Solian* (identifié directement, distance = 0) et non pas *Ablify* (identifié par la distance d'édition, distance = 1), étant donné que nous retenons le nom du médicament ayant la plus petite distance. L'annexe A.1.2 montre le script réalisé qui applique cette normalisation.

Un exemple de sortie produit par la chaîne de prétraitements est illustré par la figure 2.2. Il semble évident que le prétraitement des données facilite l'exploitation du corpus. Cette dernière s'appuie sur le contenu du message, d'où la nécessité de retenir des informations comme le nom du fichier qui permet de revenir plus facilement à la source de l'information (remonter dans le corpus), l'identifiant du message qui rattache le contenu du message avec des informations qui pourraient être utiles, l'identifiant de l'utilisateur qui s'avère utile pour toute étude statistique (étudier la fréquence d'intervention d'un utilisateur et la fiabilité de son message), le profil qui nous renseigne sur le statut de l'utilisateur et la date qui met l'accent sur une duré déterminée lors d'une étude épidémiologique.

(a) exemple de fichier d'entrée

```
lexomil-sujet_156653
1  taunio  null  26-10-2009  lexomil
bjr a tous. jai un test anti drogue à passer pour mon travail
et je crabure a lexomil depuis 1 semaine pour calmer mes
angoisses. jai tout d'un coup réalisé que lexomil contient
peut être de l'opiacée. c la catastrophe. pourriez vous me
confirmer si oui ou non ce medicament va faire réagir le test
urine que je vais devoir passer. si oui je perds mon travail,
c la catastrophe. merci de me répondre au plus vite car je dois
passer ce test ds exactement 3 jours ... y a t il de lopiacée
dans l'exomil ? tonio

lexomil-sujet_156653
2  pascou123  Doctinaute Hors Compétition  26-10-2009  lexomil
Bonjour Tonio. Je suis assistant-pharmacien. Je peux te rassurer
en te disant que le lexomil n'est pas un opiacé. Par définition,
un opiacé est un médicament qui contient de l'opium ou un de ses
dérivés et est utilisé pour contrer la douleur très forte. Par
exemple, les principaux opiacés utlisés sont la morphine,
l'hydromorphone, l'oxycodone et la méthodone. Dans ton cas, le
lexomil n'est qu'un anxiolytique dans la classe des benzodiazépines
et ne contient pas d'opiacés. Tu n'as rien à craindre. Bon courage
```

(b) exemple de fichier de sortie

FIGURE 2.2 – Résultat de la chaîne de prétraitements

2.2 Annotation du corpus

D'une manière générale, un corpus est divisé en trois parties, une pour l'apprentissage, une autre pour le développement et une dernière pour le test. Le corpus d'apprentissage sert à étudier un certain nombre de documents du corpus et à créer des règles ou un modèle statistique, celui de développement est utilisé pour évaluer les règles créées et le corpus de test permet d'évaluer les performances d'un système.

Dans cette perspective, quarante fichiers du corpus *antidépresseurs-anxiolytiques* et dix fichiers du corpus *migraine* ont été sélectionnés aléatoirement pour constituer nos corpus d'apprentissage. Quant aux corpus de test, ils concernent dix nouveaux fichiers pour chaque corpus. Il est à noter que nous n'avons pas travaillé sur l'intégralité du corpus *antidépresseurs-anxiolytiques* (mais sur un échantillon de 50 fichiers) afin de tester la pertinence de plusieurs hypothèses de travail (présentées dans la section 2.5), notamment en ce qui concerne la complexité de traiter un forum médical portant sur les antidépresseurs. Le corpus étudié s'avère ainsi de taille limitée, d'où l'intérêt de diviser le corpus uniquement en deux parties (entraînement et test).

Après l'annotation manuelle de dix fichiers parmi ceux du corpus d'apprentissage, nous avons tenté de faciliter le processus d'annotation, en l'automatisant dans un premier temps à l'aide d'un système à base de règles (*MEDINA* [Grouin, 2013]) et dans un second temps par un système à base d'apprentissage statistique (*Wapiti* [Lavergne et al., 2010]). Ces dix fichiers ont ainsi servi comme base pour les deux outils d'annotation utilisés.

2.2.1 Constitution du guide d'annotation

2.2.1.1 Objectifs

Le but de l'annotation du corpus est de constituer une référence humaine nécessaire lors des phases d'apprentissage et d'évaluation du modèle *CRF* créé. Autrement dit, l'objectif consiste à annoter manuellement les entités que nous souhaitons extraire de manière automatique, afin de créer une référence humaine essentielle pour entraîner le modèle et évaluer la performance du système développé. Cette annotation est en effet réalisée à l'aide de l'outil d'annotation *Brat*, en reposant sur des règles définies et détaillées dans le guide d'annotation (en annexe B) que nous avons élaboré en fonction des objectifs de recherche.

Selon ce dernier, une annotation du corpus permet dans un premier temps de repérer les entités sur lesquelles repose notre étude. Plus précisément, elle sert à détecter les traitements et les événements secondaires tels qu'ils sont exprimés par les patients, tout en distinguant les indications des traitements et les réactions provoquées par l'utilisation de certains médicaments. Elle permet aussi de spécifier la polarité des événements détectés (positifs, négatifs ou neutres), ainsi que de relever certaines informations utiles concernant la posologie des traitements comme la forme galénique du médicament et sa concentration. Dans un second temps, l'annotation met en évidence les liens entre traitement-événement secondaire et celles entre traitement-posologie par le biais de relations explicitement identifiées. La figure B.1 montre deux messages annotés manuellement en utilisant l'outil d'annotation *Brat*.

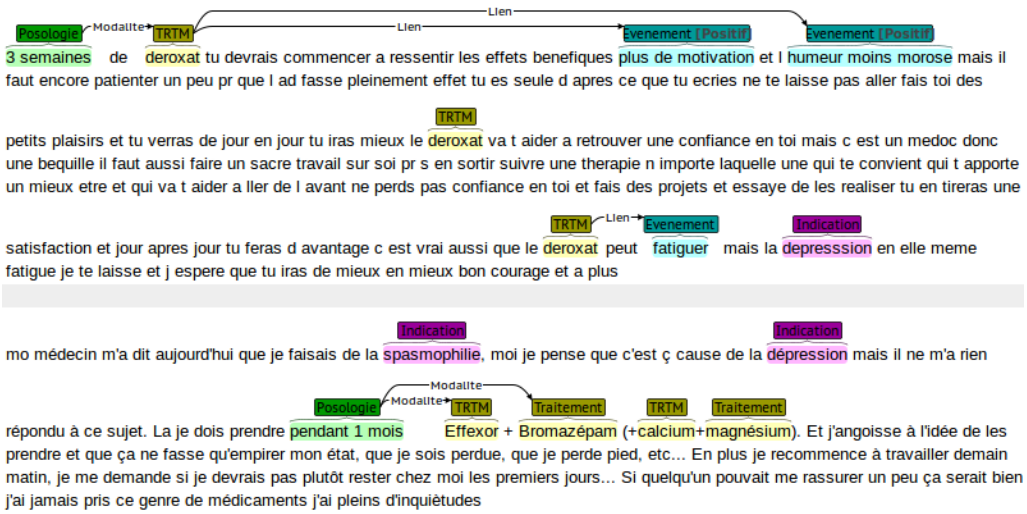


FIGURE 2.3 – Exemples de messages du corups *antidépresseurs-anxiolytiques* annotés manuellement sur *Brat*

2.2.1.2 Catégories et relations

Quatre catégories et un attribut de polarité sont pris en compte lors de l'annotation. Les catégories correspondent à *Événement*, *Traitement*, *Indication* et *Posologie*. Quant à l'attribut de polarité, il est ajouté à la catégorie *Événement* afin de préciser la positivité ou la négativité de l'événement secondaire selon la réaction provoquée par un traitement. Il est à noter que dans les cas où le contexte ne détermine pas la polarité de l'événement secondaire, celle-ci correspond à un attribut *neutre*.

- **Traitement (TRTM) :** les noms de médicaments dont parlent les patients dans leurs messages ex. : *Solian 100*, *Prozac*, *Effexor LP*.
- **Événement (EVT) :** les événements secondaires rapportés par les utilisateurs ex. : *grosse migraine*, *beaucoup de fatigue*, *régurgitation*.
- **Indication (IND) :** les raisons pour lesquelles le traitement a été prescrit. Plus précisemment, il s'agit des pathologies et symptômes que traite le médicament ex. : *dépression*, *crise d'angoisse*, *anxiété*.
- **Posologie (POSO) :** les informations associées au nom du médicament (concentration, dosage, fréquence, durée, forme galénique) ex. : *2 gélules*, *25mg*, *un comprimé par jour*.

Outre ces catégories, deux types de relations sont considérées lors de l'annotation. Le premier *Lien* met en rapport les traitements et les événements secondaires et le deuxième *Modalité* précise la modalité du traitement menant à l'événement secondaire. Il s'agit par exemple de la relation entre le médicament *Deroxat* et *vertige* pour la relation *Lien* et de celle entre le nombre 37.5 et le traitement *Effexor LP* dans le message « *bonjour, y-a-t-il accoutumance à* effexor LP *37,5 et au bout de combien de temps ? merci pour vos témoignages* » pour la relation *Modalité*.

2.2.1.3 Règles

Plusieurs règles ont été adoptées en annotant les messages déposés par les patients. Ces règles d'annotation sont appliquées au niveau de chaque message séparément en reposant sur certains principes. Parmi ces derniers figurent :

- **Redondance des entités :** l'annotation de toutes les occurrences d'une même entité liée à une catégorie étudiée au sein d'un message (à part quelques exceptions qui figurent dans le guide tels que les événements secondaires non-associés à un traitement particulier).
- **Portion textuelle la plus informative :** l'extraction de la portion de texte qui ramène le plus d'information. L'ajout de l'adjectif *constamment* dans *envie de dormir constamment* peut nuancer le sens de la portion annotée. Cette règle est en effet combinée avec d'autres qui délimitent les frontières des annotations pour prendre par exemple en compte les négations et ne pas annoter les prénons ou les déterminants (*l'Effexor* vs *Effexor* et *angoissé* vs *ni angoissé*).
- **Annotation au niveau des messages :** l'identification des relations qui existent uniquement au sein d'un même message et non pas celles entre les différents messages d'un fil de discussion. De même, les relations entre les traitements et les événements secondaires liés à l'arrêt du médicament ne sont pas extraites, étant donné qu'elles ne font pas partie des objectifs de notre étude. Il en va de même pour les relations traitement-indication (pathologie).

2.2.2 Annotation sans pré-annotation

Afin de vérifier la pertinence des règles mises en place dans le guide d'annotation, dix fichiers du corpus d'apprentissage sont annotés manuellement en double (parmi les annotateurs figure celui qui a rédigé le guide d'annotation). À l'issue de l'annotation, une phase d'adjudication des annotations est réalisée et s'ensuit d'une étape d'évaluation des annotations humaines. L'évaluation consiste en effet à calculer un accord inter-annotateurs entre les deux versions annotées pour ne pas biaiser les résultats. Ce calcul permet ainsi de mettre en évidence le taux d'accord entre les deux annotations par rapport aux accords obtenus par un simple tirage au sort.

La première phase d'annotation nous permet d'adapter le guide d'annotation élaboré, ainsi que de formuler certaines hypothèses quant au corpus traité. Nous constatons en effet que certains fils de discussion ne portent pas sur un sujet précis, mais traitent plutôt des sujets généraux. Ces fils de discussion ne comportent pas aucune mention d'événements secondaires, d'où l'intérêt de les écarter préalablement. Nous remarquons aussi que le thème de notre corpus de base portant sur les antidépresseurs et les anxiolytiques rend la tâche plus difficile. Les mêmes entités peuvent être liées soit aux événements secondaires, soit aux symptômes traités par un médicament. À titre d'exemple, le terme *crises d'angoisse* peut être considéré comme une indication d'un traitement ou un effet produit par un médicament selon le contexte dans lequel il figure. Cette ambigüité complexifie l'extraction des entités biomédicales, notamment quand il s'agit d'une automatisation de ce processus.

En effet, cette difficulté se confirme lors de l'adjudication et l'évaluation des résultats de l'annotation manuelle. Les catégories *Événement* et *Indication* enregistrent une F-mesure plutôt faible. Cette dernière varie respectivement entre 51.9 en appariement identique (*exact match*) et 58.8 en appariement relâchée (*inexact match*) pour *Événement* et entre 56.7 et 68.4 pour *Indication*. En revanche, les résultats montrent que les entités relatives à la catégorie *Traitement* sont les plus faciles à repérer. La F-mesure de cette catégorie atteint des scores élevés entre 93.5 en *exact match* et 96.1 en *inexact match*, ce qui semble logique étant donné que la catégorie *Traitement* est consacré aux noms de médicaments qui posent rarement des problèmes de frontières. De même, cette catégorie concerne les noms commerciaux des médicaments dont l'orthographe ne change pas.

Le tableau 2.1 illustre les résultats détaillés de la comparaison des deux versions annotées en matière de précision, rappel et F-mesure, ainsi qu'en indiquant le taux des vrais positifs, faux positifs et faux négatifs. Les résultats de l'appariement identique (première sous-colonne) s'avèrent plus élevés que ceux de l'appariement relâchée (deuxième sous-colonne). Cette hausse est due bien évidemment à la flexibilité de l'appariement par rapport aux frontières des entités détectées.

En outre, la comparaison des deux versions annotées manuellement nous renseigne sur les problèmes relatifs à l'identification des relations qui figurent au sein du corpus. Les résultats montrent que les liens entre *Posologie* et *Traitement* fournissent de meilleurs scores que ceux concernant *Traitement* et *Événement*. En d'autres termes, les relations traitement-posologie sont plus facile à extraire que celles entre traitement-événement secondaire. L'ambigüité des termes qui se rapportent à la catégorie *Événement* semble être à l'origine de cette difficulté. Certaines entités se trouvent étiquetées comme *Indication* au lieu d'avoir une étiquette *Événement*. Par conséquent, certaines relations *Traitement-Événement* ne sont pas détectées.

	VP		FP		FN		Précision		Rappel		F-mesure	
Événement	47	53	41	35	46	39	53.4	60.2	50.4	57.6	51.9	58.8
Indication	42	50	26	18	38	28	61.7	73.5	52.5	64.1	56.7	68.4
Posologie	22	33	24	13	16	3	47.8	71.7	57.8	91.6	52.3	80.4
Traitement	108	111	6	3	9	6	94.7	97.3	92.2	94.8	93.5	96.1
Total	219	247	97	69	109	76	69.3	78.1	66.7	76.4	68.0	77.3

TABLE 2.1 – Accords inter-annotateurs entre les deux versions annotées (dix fichiers annotés manuellement) : la sous-colonne de gauche correspond à une évaluation identique et celle de droite à une évaluation relâchée

De plus, la difficulté d'identifer des relations traitement-événement s'avère liée à la nature du corpus. Le corpus étudié est issu d'un forum de santé dans lequel certains messages ne précisent pas clairement ce type de lien. C'est par exemple le cas d'un utilisateur qui répond à un patient posant des questions à propos des événements secondaires du *Deroxat* par le message suivant : « vertiges oui, vomissement non ». Le tableau 2.2 montre les résultats globaux des accords inter-annotateurs concernant l'identification des relations qui figurent dans les dix fichiers annotés manuellement.

	VP	FP	FN	Précision	Rappel	F-mesure
Exact match	33	69	18	32.3	64.7	43.1
Inexact match	45	57	18	45.9	96.8	62.3

TABLE 2.2 – Accords inter-annotateurs des relations (dix fichiers annotés manuellement)

Il est à noter que la compréhension d'un guide d'annotation joue un rôle dans le processus d'annotation et a un impact sur le calcul de l'accord inter-annotateurs. Cette compréhension est en effet considérée comme un processus subjectif qui reflète la différence entre les interprétations des annotateurs concernant la longueur de la portion textuelle qui doit être annotée pour représenter un concept. Dans notre cas, les annotations qui ne sont pas identiques mettent en évidence les difficultés de déterminer si une personne témoigne de son expérience personnelle avec un médicament ou si elle fournit un commentaire sur le médicament dans son message. De même, elles illustrent une interprétation différente, de la part des annotateurs, quant à la traduction des métaphores et des expressions familières utilisées par les patients.

2.2.3 Annotation avec pré-annotation automatique

Partant du principe qu'une version pré-annotée automatiquement est plus rapide à traiter par des annotateurs humains et produit une version définitive d'une annotation de meilleure qualité, nous avons eu recours, en premier lieu, à l'outil *MEDINA*, un système à base de règles et une projection de lexiques conçu pour l'anonymisation des données personnelles dans les documents cliniques, mais qui permet également une annotation d'un corpus. Dans un second lieu, nous avons eu recours à l'outil *Wapiti* en vue de créer un modèle appris statistiquement. Il s'agit dans notre cas d'annoter automatiquement les entités biomédicales selon les mêmes catégories utilisées pour l'annotation manuelle et en reposant sur les règles du guide à l'aide des approches symboliques et statistiques.

2.2.3.1 Système à base de règles

Après une étude du corpus, plusieurs règles sont créées afin de trouver des déclencheurs pour les entités qui nous intéressent. Ces règles prennent la forme d'expressions régulières et concernent toutes les catégories définies à part la catégorie *Traitement*. Cette dernière est traitée grâce à un lexique constitué de 10 870 noms de médicaments, étant donné la difficulté de décrire une telle catégorie par des expressions régulières. Nous avons ensuite défini des variables pour alléger les règles créées sous forme d'expressions régulières, ainsi que pour représenter des déclencheurs et certaines entités.

Les règles et les variables définies durant cette phase d'annotation figurent dans le fichier grammaire de *MEDINA* en annexe C et certains exemples sont mentionnés dans le tableau 2.3. La première règle de ce tableau permet de capturer des entités relatives aux dosages des médicaments tels que *dosage faible* et *hautes dosages*. La deuxième règle concerne les événements secondaires mentionnés par un patient comme dans la phrase *j'avais eu des vertiges* et la troisième favorise la détection des indications exprimées après certaines verbes tels que *contrôler l'angoisse* ou *surmonter l'anxiété*.

Catégories	Variables et Règles
Posologie	$DOSE (hautes doses | dosage faible | petites doses | petite dose | faibles dosages) _DOSE
Événement	$ALPHA ([[:alpha :]]) $DET (d' | d | des | une | un) j(' |)(avais | ai) eu _DET (_ALPHA+ d'_ALPHA+ | _ALPHA+)
Indication	$INDVERB (calmer | surmonter | lutter contre | traiter | soigner | contrôler) _INDVERB l'(_ALPHA+)

TABLE 2.3 – Exemples de règles simplifiées appliquées sur le corpus à l'aide de l'outil *MEDINA*

Bien que la pré-annotation automatique permette de diminuer à moitié le temps d'annotation manuelle, elle nécessite la rédaction de règles de grammaire destinées à annoter un corpus avec le moins de bruit et silence possible (règles assez précises et générales en même temps). En outre, des listes exhaustives couvrant largement le domaine traité doivent être utilisées en vue que la projection du lexique soit efficace. Elles doivent aussi tenir compte de toutes les formes et variantes possibles de chaque entité. Un dernier point concernant cette approche se rapporte à la gestion des entités discontinues. L'outil utilisé dans cette phase ne propose pas la possibilité d'imbriquer les différentes étiquettes pour gérer de telles entités, ce qui diminue le rappel.

Le tableau 2.4 indique les résultats de l'évaluation de cette pré-annotation, en comparant la version annotée uniquement par *MEDINA* avec celle issue du consensus humain sur les dix fichiers annotés manuellement lors de la phase d'annotation sans pré-annotation. Ces résultats montrent que la pré-annotation automatique facilite la tâche d'annotation. Elle semble très utile, notamment pour les catégories *Traitement* et *Posologie*, mais elle l'est moins pour *Indication* et *Événement*. Ceci peut s'expliquer par la difficulté de créer des règles pour cette dernière catégorie, étant donné que les utilisateurs des forums ont tendance à s'exprimer en utilisant des structures de phrases différentes et difficiles à schématiser par des règles. Le langage des réseaux sociaux et forums possède des particularités qui le distingue syntaxiquement.

	VP		FP		FN		Précision		Rappel		F-mesure	
Événement	3	6	95	92	28	25	3.0	6.1	9.6	19.3	4.6	9.3
Indication	26	49	61	38	32	9	29.8	56.3	44.8	84.4	35.8	67.5
Posologie	21	41	27	7	33	12	43.7	85.4	38.8	77.3	41.1	81.1
Traitement	59	98	59	20	46	8	50.0	83.0	56.1	92.4	52.9	87.5
Total	109	194	242	157	139	54	31.0	55.2	43.9	78.2	36.3	64.7

TABLE 2.4 – Évaluation de la pré-annotation par *MEDINA* par rapport au consensus humain sur les dix fichiers annotés manuellement : la sous-colonne de gauche correspond à une évaluation identique et celle de droite à une évaluation relâchée

Par ailleurs, nous remarquons une légère amélioration des résultats de la phase avec pré-annotation qui figurent dans le tableau 2.5. Cette amélioration concerne la reconnaissance d'entités. Cependant, l'identification des relations demeure difficile à extraire en raison de l'ambigüité des entités et des contextes au sein des messages. L'évaluation des relations est illustrée dans le tableau 2.6. Il est à noter qu'a la fin de cette phase d'annotation, quelques modifications ont été apportées au guide d'annotation, À titre d'exemple, il nous a paru pertinent d'annoter uniquement les événements secondaires liés à la prise d'un traitement et d'écarter ceux relatifs à l'arrêt du traitement afin de se focaliser sur les objectifs du projet sur lequel nous travaillons. Le cas échéant, il semble crucial de spécifier le type de relations entre les traitements et les événements secondaires dans le but de différencier les événements *liés à la prise* de ceux *liés à l'arrêt*.

	VP		FP		FN		Précision		Rappel		F-mesure	
Événement	43	52	44	35	32	21	49.4	59.7	57.3	71.1	53.0	65.0
Indication	29	35	28	22	13	5	50.8	61.4	69.0	87.5	58.5	72.1
Posologie	42	52	32	22	20	11	56.7	70.2	67.7	82.5	61.7	75.9
Traitement	97	102	13	8	6	1	88.1	92.7	94.1	99.0	91.0	95.7
Total	211	241	117	87	71	38	64.3	73.4	74.8	86.3	69.1	79.4

TABLE 2.5 – Accords inter-annotateurs entre les deux versions annotées (dix fichiers pré-annotés par *MEDINA* puis corrigés manuellement) : la sous-colonne de gauche correspond à une évaluation identique et celle de droite à une évaluation relâchée

	VP	FP	FN	Précision	Rappel	F-mesure
Exact match	47	109	44	30.1	51.6	38.0
Inexact match	66	90	44	42.3	60.0	49.6

TABLE 2.6 – Accords inter-annotateurs des relations (dix fichiers pré-annotés par *MEDINA* puis corrigés manuellement)

2.2.3.2 Système à base d'apprentissage statistique

Sur la base de ces vingt fichiers annotés (dix fichiers annotés manuellement et dix autres pré-annotés par *MEDINA*), nous créons plusieurs modèles *CRF*, qui diffèrent selon la taille du corpus d'apprentissage, avec l'outil *Wapiti* en vue de faciliter l'annotation du corpus. La création de ces modèles prend en effet la forme d'un processus itératif qui consiste à rajouter à chaque itération dix fichiers sur les données d'apprentissage. Plus précisément, le modèle est appris dans un premier temps sur

les vingt fichiers annotés et corrigés manuellement pour être testé sur dix nouveaux fichiers non-annotés. Ces derniers sont ensuite corrigés manuellement (en double pour avoir un consensus), en vue d'enrichir les prédictions du *CRF* en lui donnant en entrée plus d'exemples annotés, et rajoutés au corpus d'apprentissage. Lors de la deuxième itération, le modèle est appris sur ces trente fichiers (20 fichiers de base + les 10 nouveaux fichiers) et dix nouveaux documents sont choisis pour constituer le corpus de test. Le schéma 2.4 illsutre le processus itératif d'annotation appliqué durant cette phase d'apprentissage statistique.

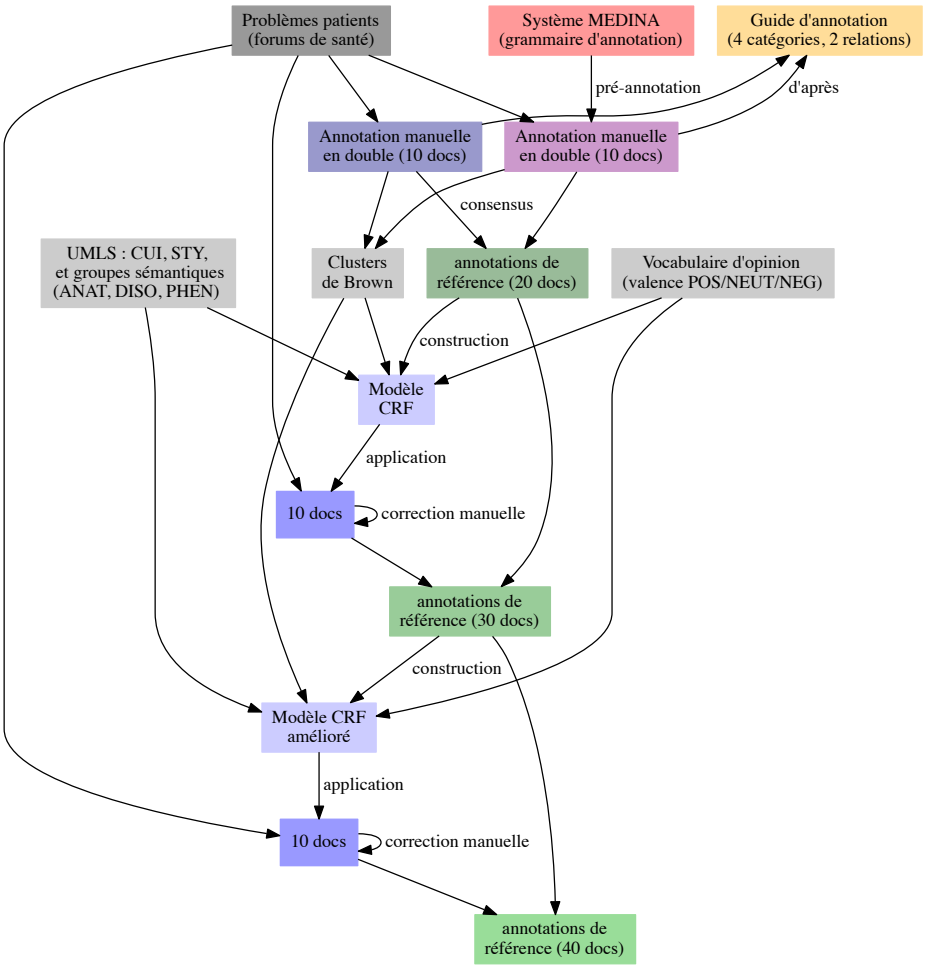


FIGURE 2.4 – Processus itératif de la constitution de la référence à l'aide d'un système à base d'apprentissage statistique

Le principe de base des *CRF* étant l'apprentissage par l'observation de contextes similaires, notre hypothèse de travail consiste à prévoir une annotation de meilleure qualité en augmentant le jeu de données d'apprentissage. D'une manière générale, nous pouvons nous attendre à avoir plus de prédictions correctes en fournissant plus d'exemples annotés en entrée au *CRF*. Une telle hypothèse induit ainsi une plus grande ressemblance entre les prédictions de *Wapiti* et les corrections humaines des annotations avec l'augmentation de la taillle des données sur lesquelles le modèle a

été appris. Les tableaux 2.7, 2.8 et 2.9 montre l'évolution des résultats d'annotation au fur et à mesure de l'augmentation du corpus d'apprentissage. Ils indiquent la différence entre les annotations de *Wapiti* avant et après la correction manuelle en *exact* et *inexact match*.

	VP		FP		FN		Précision		Rappel		F-mesure	
Événement	5	6	66	65	1	0	7.0	8.4	83.3	100	12.9	15.5
Indication	4	12	21	21	2	2	16.0	16.0	66.6	66.6	25.8	25.8
Posologie	9	12	54	51	4	1	14.2	19.0	69.2	92.3	23.6	31.5
Traitement	49	49	38	38	0	0	56.3	56.3	100	100	72.0	72.0
Total	67	71	179	175	7	3	27.2	28.8	90.5	95.9	41.8	44.3

TABLE 2.7 – Accords inter-annotateurs entre la version pré-annotée par *Wapiti* et celle corrigée manuellement après la pré-annotation - apprentissage sur 20 fichiers : la sous-colonne de gauche correspond à une évaluation identique est celle de droite à une évaluation relâchée

	VP		FP		FN		Précision		Rappel		F-mesure	
Événement	10	15	103	98	9	5	8.8	13.2	50.0	75.0	15.0	22.5
Indication	13	16	51	48	9	6	20.3	25.0	59.0	72.7	30.2	37.2
Posologie	12	15	50	47	6	3	19.3	24.1	66.6	83.3	30.0	37.5
Traitement	61	61	70	70	0	0	46.5	46.5	100	100	63.5	63.5
Total	96	107	274	263	25	14	25.9	28.9	79.3	88.4	39.1	43.5

TABLE 2.8 – Accords inter-annotateurs entre la version pré-annotée par *Wapiti* et celle corrigée manuellement après la pré-annotation - apprentissage sur 30 fichiers : la sous-colonne de gauche correspond à une évaluation identique est celle de droite à une évluation relâchée

	VP		FP		FN		Précision		Rappel		F-mesure	
Événement	9	14	104	99	11	6	7.9	12.3	45.0	70.0	13.5	21.0
Indication	14	17	50	47	8	5	21.8	26.5	63.6	77.2	32.5	39.5
Posologie	13	17	49	45	7	3	20.9	27.4	65.0	85.0	31.7	41.4
Traitement	64	64	67	67	0	0	48.8	48.8	100	100	65.6	65.6
Total	100	112	270	258	26	14	27.0	30.2	79.3	88.8	40.3	45.1

TABLE 2.9 – Accords inter-annotateurs entre la version pré-annotée par *Wapiti* et celle corrigée manuellement après la pré-annotation - apprentissage sur 40 fichiers : la sous-colonne de gauche correspond à une évaluation identique est celle de droite à une évluation relâchée

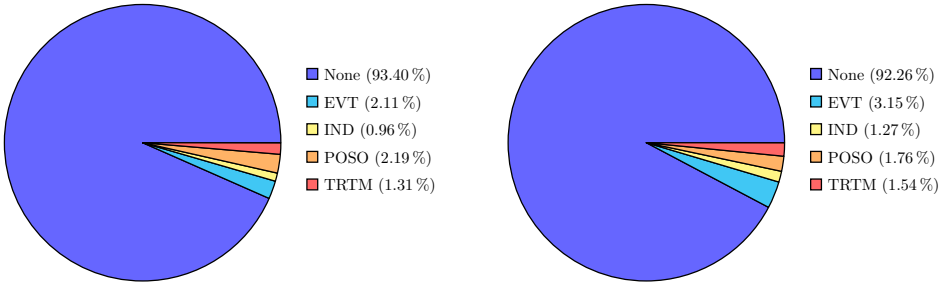
Ils montrent ainsi une légère amélioration des résultats avec l'augmentation de la taille des données d'apprentissage. Selon ces tableaux, les catégories les plus faciles à identifier automatiquement restent *Traitement* et *Posologie*, contrairement aux deux autres catégories. Les résultats restent pourtant modestes et favorisent le rappel au sujet de la précision.

Ces tableaux montrent également la complexité de l'extraction des événements secondaires de manière automatique. Cette difficulté est due aux différents types de relations qui existent entre un traitement et un événement. Une même entité peut représenter à la fois un événement positif, un événement négatif, voire même une indication de la pathologie. À titre d'exemple, *me fait dormir* peut être un effet lié

à l'efficacité d'un traitement (insomnifère), un effet bénéfique mais non désiré d'un traitement ou une réaction nocive du médicament. Il semble difficile pour un système de faire la distinction entre les entités qui se rapportent aux deux catégories *Événement* et *Indication* (ce qui explique les résultats de ces deux catégories), étant donné qu'elles reflètent des concepts identiques et sont différenciées à l'aide du contexte.

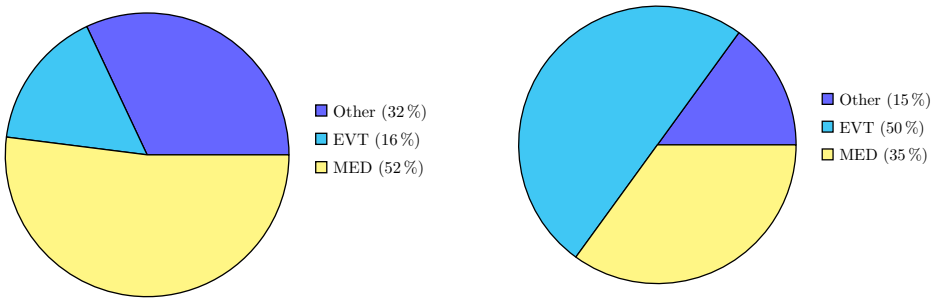
2.3 Statistiques sur le corpus annoté

À l'issue de l'annotation des deux corpus étudiés, nous présentons dans cette section quelques chiffres qui caractérisent nos corpus annotés (50 fils de discussion *antidépresseurs-anxiolytiques* et 20 fils de discussion *migraine*). À titre d'exemple, la figure 2.5 caractèrise les deux corpus en nombre d'entités, 2.6 présente le résultat d'un typage manuel des fichiers du corpus effectué selon le nom du fichier (le nom du fichier comprend un événement secondaire, le nom d'un traitement ou aucun des deux) et la figure 2.7 indique la proportion des formes différentes parmi les occurrences d'entités pour chaque catégorie traitée. Ces statistiques seront en effet utiles pour analyser les résultats de notre système, ainsi que pour comprendre le comprtement du système en réalisant ses prédictions.



(a) corpus *antidépresseurs-anxiolytiques* (b) corpus *migraine*

FIGURE 2.5 – Pourcentage des occurrences par catégorie dans les deux corpus



(a) corpus *antidépresseurs-anxiolytiques* (b) corpus *migraine*

FIGURE 2.6 – Pourcentage des fils de discussions qui portent sur un événement secondaire, un médicament et un sujet générique dans les deux corpus

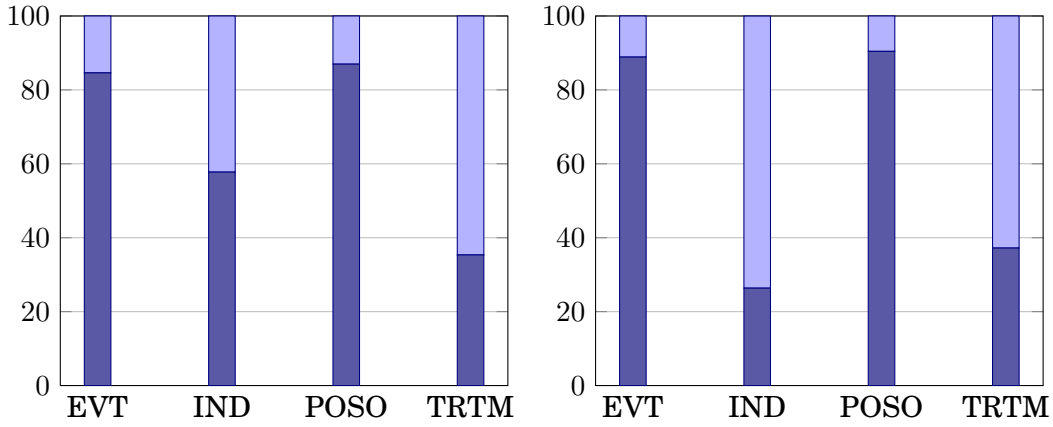


FIGURE 2.7 – Pourcentage des formes différentes parmi toutes les occurrences d'entités dans les deux corpus pour chaque catégorie : la figure de gauche concerne le corpus *antidépresseurs-anxiolytiques* et celle de droite le corpus *migraine*

	anitdépresseurs			migraine		
	formes	occ.	f/o	formes	occ.	f/o
Événement	330	390	**0.84**	496	558	**0.88**
Indication	145	251	0.57	114	432	0.26
Posologie	247	284	0.86	227	251	0.90
Traitement	181	512	0.35	249	669	0.37

TABLE 2.10 – Le nombre d'annotations par catégorie dans les deux corpus : la sous-colonne de droite indique le nombre d'occurrences d'entités, la sous-colonne de gauche présente le nombre de formes différentes et la troisième sous-colonne représente le rapport formes/occurrences

2.4 Création du modèle *CRF*

2.4.1 Algorithme et paramètre d'optimisation

Rappelons que notre objectif est la création d'un modèle de détection automatique d'événements secondaires par apprentissage supervisé et à l'aide des *CRF* tels qu'ils sont implémentés dans l'outil *Wapiti*. Les fichiers annotés comme expliqué dans la section 2.2 sont ainsi fournis à un tel outil comme un ensemble de données annotées (les classes sont connues) afin de l'aider à établir des critères de classification. Ces derniers sont réutilisés lors de la phase de test sur des données non-annotées dont *Wapiti* faudra retrouver la classe (catégorie).

À cette fin, plusieurs algorithmes sont implémentés dans *Wapiti* : *rprop*, *sgdl-1* et *l-bfgs*). L'efficacité de ces algorithmes dépend de la tâche à effectuer et du corpus traité. Dans notre cas, l'algorithme le plus adéquat s'avère être *rprop* en raison de sa rapidité et sa bonne performance. *sgd-l1* est en effet un algorithme stochastique (moins performant) utilisé d'une manière générale pour améliorer le temps de calcul et *l-bfgs* est connu par sa précision élevée, mais exige un espace mémoire important. *rprop* est aussi performant que ce dernier algorithme, mais nécessite un espace mémoire plus restreint (environ 30 itérations nécessaires lors de l'entraînement du modèle).

Outre le corpus d'apprentissage annoté et le choix de l'algorithme, nous fournissons à *Wapiti* un ensemble de caractérisitques sélectionnées qui reflètent des particularités du corpus étudié. Ces caractéristiques sont ensuite sélectionnées par le paramètre d'optimisation défini afin d'être utilisées lors de la prédiction des nouvelles étiquettes. En effet, en réglant le paramètre d'optimisation de *Wapiti* pour appliquer une pénalité *Laplacienne* 0.1 lors de la création du modèle, nous ne faisons pas de validation croisée, mais une régularisation *L1* [Tsuruoka et al., 2009] qui sélectionne les caractéristiques les plus intéressantes parmi celles qui sont définies. Cependant, en autorisant un tel terme de pénalité, le modèle créé risque d'être sur-appris. Il est à noter que plus la pénnalité est proche de zéro, moins le modèle est strict et moins d'entités sont annotées. Une pénalité proche de zéro augmente la précision et réduit ainsi le bruit des résultats (la plupart des annotations sont correctes).

2.4.2 Caractéristiques du modèle

Lors de la création du modèle, différents types de caractéristiques relatives à notre corpus sont définis. Elles varient entre des caractéristiques de surface, formes des tokens ainsi que des ressources externes et sont représentées selon un format prérequis par *Wapiti*. Ce dernier nécessite une tokénisation du corpus et peut avoir besoin d'autres traitements, tels que la lemmatisation, l'étiquetage morpho-syntaxique et un appariement de lexiques d'après les caractéristiques choisies. Il est à noter que certaines caractéristiques sont utilisées pour aider le *CRF* à attribuer la bonne étiquette à l'entité qui correspond, mais aussi pour ne pas annoter certains tokens (mettre une étiquette *out* si le système d'annotation applique la mèthode *Begin In Out (BIO)*).

Caractéristiques de surface Parmi les caractéristiques de surface considérées figurent la casse typographique du token (ex. : toutes les lettres du token sont en majuscule ou uniquement la première lettre) et sa taille (nombre de caractères), ainsi que la présence des chiffres et des signes de ponctuation au sein du token. Les caractéristiques de surface facilite le repérage des entités à annoter, en distinguant entre un nom de médicament commençant dans la plupart des cas par une majuscule et la concentration d'un médicament qui contient des chiffres.

Formes des tokens Il s'agit de la taille de la fenêtre de tokens examinée autour du token pivot (ex. : unigrammes, bigrammes ou trigrammes). Nous prenons en compte les unigrammes, les bigrammes et les trigrammes de tokens, afin de considérer les différentes séquences possibles composées de trois tokens.

Ressources externes Nous nous sommes appuyées sur divers types de ressources externes en constituant le modèle :

- **Lexiques :** plusieurs listes terminologiques sont prises en compte, parmi lesquelles il existe celles qui sont construites manuellement, celles qui sont constituées à partir des bases de données biomédicales et des lexiques affectifs disponibles en ligne, ou celles qui sont exportées de la version française de l'*UMLS*.
    1. Liste des médicaments : lexique de 10 870 noms de médicaments constitué à partir du site du *Vidal*, un site consacré aux informations sur les produits de santé et la sécurisation de la prescription des médicaments.

2. Liste des des évémenements secondaires : lexique extrait de la base de données en ligne des événements secondaires *MedEffet Canada*. Cette base renferme des informations concernant les événements secondaires soupçonnés associés à des produits de santé et rapportés par les consommateurs et les professionnels de la santé, ainsi que par les fabricants et les distributeurs. Les données brutes relatives à cette base concernent la période de 1965 à 2013 (31 décembre) et sont répartis sur 11 tables exportées en fichiers *ASCII* comprenant plusieurs champs. Nous extrayons automatiquement à l'aide d'une jointure entre ces fichiers les traitements et les événements secondaires correspondants (annexe A.2.1). 8 619 associations traitement-événement secondaire sont retenues. Nous retenons également la classe générique de chaque événement secondaire extrait en vue d'avoir une idée du type d'événement provoqué par chaque médicament. La reconnaissance de cette classe peut s'avérer utile en analysant les résultats et après avoir effectué la correpondance entre l'événement secondaire détecté et un traitement.
3. Lexique affectif : version préliminaire d'un lexique affectif français similaire à la liste *Affective Norms for English Words (ANEW)*. Ce lexique est constitué selon la méthode de [Pak and Paroubek, 2010]. Il renferme 879 formes extraites d'un corpus de *Tweets*. Chaque forme est attribuée un score qui correspond à sa valence. Cette dernière varie entre 1 (plus négatif) et 9 (plus positif). Nous utilisons ainsi cette liste pour déterminer la polarité de chaque mot du corpus et sa valence. Une telle tâche paraît utile dans la mesure où un événement secondaire peut figurer le plus souvent dans un contexte négatif.
4. UMLS : il nous semble pertinent d'utiliser les ressources conçues pour le domaine médical, telles que l'*UMLS* comme dans les travaux de [Leaman et al., 2010] et [Kang et al., 2014] afin de développer un système assez robuste. Ce metathesaurus est en effet conçu par la *U.S. National Library of Medicine* qui se charge du développement des ressources et des systèmes d'information biomédicaux. Il permet d'établir la correspondance entre les entités qui figurent dans les messages des patients et des classes plus génériques ou des groupes sémantiques comme *disorders phenomenon* et *anatomy* afin de regrouper plusieurs termes sous un même concept. Ces trois groupes sémantiques ont été choisi après une étude du corpus qui révèle que les portions textuelles relatives à la catégorie *Événement* comprennent des termes qui sont inclus dans ces groupes. De même, [Denecke, 2014] a établi une comparaison entre les documents cliniques et les réseaux sociaux en ce qui concerne les catégories sémantiques des mots utilisés. Selon cette comparaison, les catégories sématiques les plus utilisées dans le langage des médias sociaux ressemblent aux groupes sémantiques de l'*UMLS* utilisés.

- **Déclencheurs :** les déclencheurs sont des marqueurs qui peuvent nous renseigner sur la présence d'une entité précise au sein d'une phrase (ex. : les *titres* comme *M.* ou *Mme* pour repérer les entités de type *personne* en cas de reconnaissance d'entités nommées). Dans notre étude, il s'agit de la présence de certains mots liés aux événements secondaires tels que *mal*, *problème* ou *désordres* qui peuvent jouer un rôle dans le processus d'extraction. Partant du principe qu'un

événement secondaire figure le plus souvent dans un contexte négatif, nous avons également considéré les négations (*n'*, *ne*, *pas*, *plus*, etc.) comme des déclencheurs.

- **Morpho-syntaxe :** l'étiquetage en parties du discours devrait faciliter le repérage des entités (traitement ou événement secondaire) nécessaires pour notre étude en fonction de leurs catégories. De même, une lemmatisation permet de prendre en compte les différentes formes de surface d'un terme, ce qui peut améliorer la pertinence de la méthode adoptée. Destiné à de telles tâches, nous utilisons *Tree Tagger* pour étiqueter et lemmatiser chaque token du corpus. Les étiquettes morpho-syntaxiques et les lemmes sont ainsi utilisées comme caractéristiques du modèle.
- **Clusters :** les clusters de *Brown* [Brown et al., 1992] sont généralement utiles pour rassembler les termes qui partagent des contextes communs. Le *clustering* repose en effet sur des méthodes de classification automatique non-supervisée des données (méthode de regroupement hiérarchique ou de partitionnement de données) et est employé dans le domaine de l'analyse de données. Dans notre travail, les clusters sont employés en vue de regrouper les différents événements secondaires exprimés par les patients, étant donné qu'ils mettent en évidence les structures similaires entre les divers messages du corpus. Ils servent ainsi de caractéristiques pour aider l'outil *Wapiti* dans ses prédictions sur la base de contextes communs de la manière les patients expriment les réactions inattendues des médicaments dans leurs messages.

2.5 Configurations du modèle

La qualité des prédictions du système mis en œuvre nous permet de confirmer ou réfuter certaines hypothèses de travail. Nous mesurons ainsi la performance de notre système selon plusieurs configurations, dont certains sont *in domain* et d'autres sont *out domain*. Le premier type de configurations concerne en effet un corpus dont l'apprentissage et le test sont réalisés sur un même sous-domaine médical (ex. : apprentissage et test sur le corpus *antidépresseurs-anxiolytiques*), mais le second consiste à effectuer l'apprentissage du modèle et son test sur des corpus portant sur deux sous-domaines médicaux différents (ex. : apprentissage sur le corpus *antidépresseurs-anxiolytiques* et test sur le corpus *migraine*). Nous détaillons ici les différentes hypothèses de travail, ainsi que les configurations qui correspondent à ces hypothèses. Ces dernières sont récapitulées dans le tableau 2.11.

2.5.1 Configurations *in domain*

Parmi les configurations *in domain* prises en compte dans notre étude, il existe celles qui sont effectuées sur le corpus *antidépresseurs-anxiolytiques* en variant la taille du corpus d'apprentissage. Partant du principe que l'augmentation de la taille des données d'entraînement permet d'améliorer la performance d'un système appris statistiquement jusqu'à atteindre un certain point où cette amélioration stagne (hypothèse de travail 1), nous étudions l'impact de la variation de la taille du corpus *antidépressurs-anxiolytiques* sur les résultats. À cette fin, quatre configurations sont ainsi testées, en augmentant à chaque fois les données d'apprentissage de dix nouveaux fichiers. La taille du corpus varie d'une échelle entre dix et quarante fichiers,

tandis que les données de test restent toujours les mêmes et se limitent à dix fichiers annotés et corrigés.

Outre les configurations qui étudient l'impact de la variation de la taille du corpus, trois autres configurations *in domain* sont appliquées. Deux configurations testent la pertinence des catégories définies dans le guide d'annotation, notamment en ce qui concerne la distinction entre *Événement* et *Indication*, et la troisième permet de mettre l'accent sur la complexité de traiter un corpus portant sur les antidépresseurs et les anxiolytiques. Ayant ainsi comme hypothèse de travail qu'un système qui détecte uniquement les événements secondaires ou qui ne différencie pas les événements secondaires des indications des pathologies fournit de meilleurs résultats qu'un autre système qui traite plusieurs catégories (hypothèse de travail 2 et 3), deux nouvelles configurations sont testées. Elles concernent respectivement la constitution d'un modèle *CRF* antidépresseurs-anxiolytiques uniquement pour la catégorie *Événement* et d'un autre modèle où les deux catégories *Événement* et *Indication* sont fusionnées en une seule.

Enfin, la dernière configuration *in domain* nous permet de confirmer notre quatrième hypothèse qui indique qu'un forum portant sur les antidépresseurs-anxiolytiques complexifie l'automatisation de l'extraction des événements secondaires à partir des messages publiés par les patients sur les forums de santé (hypothèse de travail 4) puisque les mêmes événements peuvent être liés soit à un traitement, soit à la maladie elle-même. Dans cette configuration, l'apprentissage et le test concernent ainsi le deuxième corpus qui se rapporte à un sous-domaine médical différent. Les vingt fichiers du corpus *migraine* sont ainsi divisé en deux ensembles, les premiers dix sont gradés pour l'apprentissage et les dix autres sont consacrés pour le test.

2.5.2 Configurations *out domain*

Par ailleurs, trois configurations *out domain* sont également mises en œuvre. L'objectif principal de ces configurations consiste à évaluer la robustesse du système et à étudier son comportement vis-à-vis du sous-domaine du corpus étudié. La première d'entre elles consiste en effet à appliquer le modèle 40 du corpus *antidépresseurs-anxiolytiques* sur les dix fichiers test du corpus *migraine*. Elle permet ainsi d'évaluer l'hypothèse indiquant qu'un modèle *CRF* appris sur un sous-domaine déterminé et testé sur un autre sous-domaine fournit des résultats plus modestes que celui qui est appris et testé sur le même sous-domaine (hypothèse de travail 5).

Quant aux deux autres configurations, il s'agit de créer un nouveau modèle *CRF* (modèle 50) dont l'apprentissage est effectué sur les quarante fichiers d'entraînement du corpus *antidépresseurs-anxiolytiques* ainsi que les dix fichiers d'apprentissage du corpus *migraine*. Ce nouveau modèle est ensuite évalué une première fois sur le corpus test *migraine* et une deuxième fois sur le corpus test *antidépresseurs-anxiolytiques*. L'intérêt de ces deux dernières configurations est de vérifier si un système qui prend en compte plusieurs sous-domaines médicaux est moins performant que celui qui porte uniquement sur un seul sous-domaine (hypothèse de travail 6).

n°	Hypothèses	Configurations	Expériences
1	la performance d'un système s'améliore avec l'augmentation de la taille du corpus d'apprentissage	*in domain*	augmentation de la taille du corpus d'apprentissage (10, 20 30 et 40 fichiers)
2	un modèle *CRF* qui détecte uniquement les événements secondaires a une meilleure performance	*in domain*	détection uniquement des entités relatives à la catégorie *Événement*
3	le recouvrement entre la catégorie *Événement* et *Indication* a un impact sur les prédictions réalisées par le *CRF*	*in domain*	fusion des deux catégories *Événement* et *Indication*
4	l'étude d'un forum de santé différent (corpus *migraine*) que celui des antidépresseurs-anxiolytiques fournit de meilleurs résultats et simplifie la tâche de détection d'événements secondaires	*in domain*	création d'un modèle *CRF* portant sur les migraines et les maux de têtes (un thème différent que celui du corpus de base antidépresseurs-anxiolytiques)
5	un modèle *CRF* appris sur un sous-domaine médical et appliqué sur un autre sous-domaine différent est moins performant qu'un modèle appris et testé sur le même sous-domaine	*out domain*	application du modèle 40 du corpus *antidépresseurs-anxiolytiques* sur le corpus *migraine*
6	un système qui traite plusieurs sous-domaines biomédicaux confondus fournit des résultats plus modestes que celui qui traite uniquement un sous-domaine particulier	*out domain*	création du modèle 50 dont le corpus d'apprentissage englobe les 40 fichiers du corpus *antidépresseurs-anxiolytiques* et les 10 fichiers du corpus *migraine*

TABLE 2.11 – Synthèse des hypothèses de travail et des configurations

CHAPITRE 3

RÉSULTATS ET DISCUSSION

Sommaire

Introduction

Rappelons que l'objectif de notre travail consiste à détecter d'une manière automatique les événements secondaires rapportés par les patients dans les messages publiés sur les forums de santé. Afin de parvenir à un tel objectif, nous tentons de trouver la meilleure configuration du modèle de détection d'événements secondaires. Plusieurs configurations du modèle dévloppé sont ainsi testées.

En ce sens, cette partie met en lumière les différents tests effectués pour arriver à la configuration optimale qui fournit les meilleurs résultats possibles. Elle illustre ainsi les expériences réalisées, tout en présentant les résultats des tests menés et leurs analyses. Enfin, une discussion s'ensuit pour mettre l'accent sur les limites de notre méthode, ainsi que pour aborder quelques possibilités de son amélioration dans les travaux futurs.

3.1 Mesures d'évaluation

L'évaluation de la qualité des prédictions du modèle *CRF* nous renseigne sur la robustesse du système élaboré. De même, l'analyse des résultats obtenus nous permet d'avoir une idée sur les limites d'un système et de dégager certaines pistes d'amélioration potentielles. En vue d'avoir une idée sur la pertinence des choix des catégories définies lors de la phase d'annotation, ainsi que des caractéristiques utilisées lors de la constitution du modèle *CRF*, nous évaluons notre système en terme de précision, rappel et F-mesure. Ces derniers sont des métriques d'évaluation mono-classes utilisées fréquemment dans le domaine de la fouille de textes.

- **Précision :** mesure le nombre d'éléments correctement étiquetés par le système (vrais positifs) par rapport au nombre total d'éléments étiquetés par le système (vrais et faux positifs) $\frac{vrais\ positifs}{vrais\ positifs+faux\ positifs}$.
- **Rappel :** mesure le nombre d'éléments correctement étiquetés par le système par rapport au nombre d'éléments dans la référence $\frac{vrais\ positifs}{vrais\ positifs+faux\ négatifs}$.
- **F-mesure :** correspond à la moyenne harmonique pondérée du rappel et de la précision $\frac{(1+\beta^2)\times Précision\times Rappel}{\beta^2\times Précision+Rappel}$.

Les résultats sont ainsi décrits du point de vue du système (positif ou négatif) et en même temps du point de vue de la référence (vrai ou faux). Ils peuvent être ainsi définis en terme de vrais positifs, vrais négatifs, faux positifs et faux négatifs.

- ***Vrais positifs* (VP) :** nombre d'éléments étiquetés de la même manière dans l'hypothèse et la référence.
- ***Faux positifs* (FP) :** nombre d'éléments de l'hypothèse qui sont absents de la référence.
- ***Vrais négatifs* (VN) :** nombre d'éléments absents de l'hypothèse et de la référence.
- ***Faux négatifs* (FN) :** nombre d'éléments de la référence qui sont absents de l'hypothèse.

3.2 Résultats des expériences

3.2.1 Expériences relatives aux configurations *in domain*

3.2.1.1 Corpus antidépresseurs-anxiolytiques

La figure 3.1 illsutre l'évolution de la F-mesure obtenue par les différentes configurations appliquées sur le corpus *antidépresseurs-anxiolytiques* en faisant varier la taille du corpus d'apprentissage (hypothèse de travail 1). Les chiffres qui correspondent à la taille du corpus d'apprentissage sont présentés en abscisses et les valeurs de la F-mesure sont indiquées en ordonnées.

De même, le tableau 3.1 montre les résultats détaillés obtenus à l'aide de ces différentes configurations. Pour chaque modèle constitué, la précision (*P*), le rappel (*R*) et la F-mesure (*F*) sont calculés. Pour le modèle 40, nous indiquons également le nombre de *vrais positifs*. Ces scores correspondent en effet aux résultats des quatre premières configurations *in domain* détaillées dans la section 2.5.

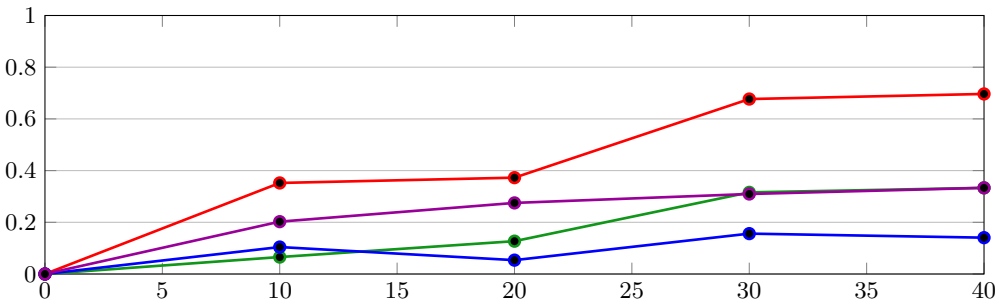


FIGURE 3.1 – Évolution de la F-mesure selon la taille du corpus d'apprentissage : *Traitement* en rouge, *Indication* en violet, *Posologie* en vert et *Événement* en bleu

	modèle 10			modèle 20			modèle 30			modèle 40			
	P	R	F	P	R	F	P	R	F	P	R	F	VP
EVT	85.7	5.5	10.4	75.0	2.7	5.3	50.0	9.2	15.6	45.5	8.3	**14.0**	**20**
IND	47.0	12.9	20.2	61.1	17.7	27.5	59.0	20.9	30.9	63.6	22.5	**33.3**	**22**
POSO	66.6	3.4	6.5	80.0	6.9	12.7	66.6	20.6	31.5	65.0	22.4	**33.3**	**20**
TRTM	100	21.3	35.2	100	22.9	37.2	100	51.1	67.7	100	53.4	**69.6**	**70**
Total	80.0	12.2	21.2	84.2	13.3	23.0	80.3	28.4	41.9	**80.3**	**29.5**	**43.1**	

TABLE 3.1 – Résultats des quatre premières configurations *in domain* appliquées sur le corpus *antidépresseurs-anxiolytiques* - Expérience # 1

Les deux tableaux 3.2 et 3.3 présentent respectivement les résultats fournis par la création d'un système qui détecte uniquement les événements secondaires (hypothèse de travail 2) et ceux fournis par le système qui ne considère pas la différence entre la catégorie *Événement* et *Indication* (hypothèse de travail 3). Ils dressent ainsi un bilan de la cinquième et la sixième configuration *in domain*, en calculant la précision, le rappel, la F-mesure et le nombre de *vrais positifs* pour chaque catégorie traitée.

	Précision	Rappel	F-mesure	VP
Événement	**57.1**	**7.4**	**13.1**	**14**

TABLE 3.2 – Résultats du système qui traite uniquement la catégorie *Événement* appliqué sur le corpus *antidépresseurs-anxiolytiques* - Expérience # 2

	Précision	Rappel	F-mesure	VP
Événement	**66.6**	**14.1**	**23.3**	**36**
Posologie	66.6	6.9	12.5	6
Traitement	100	32.8	49.3	43
Total	83.5	19.7	31.9	

TABLE 3.3 – Résultats de la fusion des deux catégories *Événement* et *Indication* testée sur le corpus *antidépresseurs-anxiolytiques* - Expérience # 3

3.2.1.2 Comparaison entre le corpus migraine et le corpus antidépresseurs-anxiolytiques

Quant aux expériences sur le corpus *migraine*, le tableau 3.4 montre les scores obtenus par le système appris sur les dix fichiers d'entraînement du corpus *migraine* et testé sur les dix fichiers test du même corpus (hypothèse de travail 4). Les résultats correspondent à ceux de la dernière configuration *in domain* et sont rapportés en précison, rappel, F-mesure et *vrais positifs*.

	Précision	Rappel	F-mesure	VP
Événement	**51.3**	**14.3**	**22.4**	**74**
Indication	**88.2**	**61.7**	**72.6**	**170**
Posologie	69.6	33.0	44.8	56
Traitement	97.2	64.9	77.8	250
Total	85.4	47.0	60.6	

TABLE 3.4 – Résultats du système appliqué sur le corpus *migraine* - Expérience # 4

3.2.2 Expériences relatives aux configurations *out domain*

Enfin, ces deux tableaux présentent les résultats des configurations *out domain*, soit les configurations qui reflètent les expériences croisées entre les deux corpus. Le premier tableau 3.5 illustre ainsi les scores obtenus par un système appris sur les quarante fichiers du corpus *antidépressurs-anxiolytiques* et évalué sur les dix fichiers test du corpus *migraine*, tandis que le second 3.6 concerne la création d'un nouveau modèle (modèle 50) qui englobe les données d'apprentissage des deux corpus (40 fichiers antidépresseurs-anxiolytiques et 10 fichiers migraine) pour être testé une première fois sur le corpus test du corpus *antidépresseurs-anxiolytiques* et une deuxième fois sur celui du corpus *migraine*.

	Précision	Rappel	F-mesure	VP
Événement	51.1	8.6	14.8	45
Indication	35.2	2.4	4.6	17
Posologie	50.8	25.4	33.9	59
Traitement	99.1	31.5	47.8	119
Total	73.7	17.7	28.5	

TABLE 3.5 – Résultats de la première configuration *out domain* : apprentissage sur le corpus *antidépresseurs-anxiolytiques* (modèle 40) et test sur le corpus *migraine* (modèle 10) - Expérience # 5

	test sur antidépresseurs				test sur migraine			
	P	R	F	VP	P	R	F	VP
Événement	53.5	13.8	**22.0**	**28**	62.8	48.7	**62.5**	**70**
Indication	56.5	20.9	30.5	23	85.4	53.0	65.4	151
Posologie	64.2	15.5	25.0	14	65.4	30.5	41.6	55
Traitement	98.9	74.8	**85.2**	**99**	98.5	74.3	**84.7**	**282**
Total	**82.3**	**37.6**	**51.6**		**87.2**	**48.7**	**62.5**	

TABLE 3.6 – Résultats de la deuxième configuration *out domain* : apprentissage fusionné sur les deux corpus et test sur le corpus *antidépresseurs-anxiolytiques* et sur le corpus *migraine* - Expérience # 6

3.3 Discussion

3.3.1 Résultats des configurations *in domain*

Selon les résultats présentés dans la section 3.2, nous remarquons que la précision, le rappel et la F-mesure varient selon les catégories traitées et les configurations appliquées. D'une manière générale, nous remarquons que les résultats globaux sont satisfaisants en matière de précision (entre 80% et 84%), mais restent modestes en terme de rappel (entre 12% et 29%). Dans un premier temps, le tableau 3.1 et la figure 3.1 valident notre première hypothèse de travail qui concerne l'efficacité du système avec l'augmentation de la taille du corpus d'apprentissage. Dans un second temps, ils nous montrent que parmi les quatre catégories définies dans le guide d'annotation, certaines semblent plus faciles à identifier que d'autres. Les entités qui correspondent à la catégorie *Traitement* s'avèrent les plus faciles à extraire automa-

tiquement, étant donné que les portions textuelles qui se rapportent à cette catégorie constituent des noms commerciaux généralement courts et ne témoignent pas d'une grande variation. Cela est en effet reflété par la précision, qui ne change pas dans toutes les configurations testées (100%) même avec un modèle appris sur un nombre limité de dix fichiers. Le rappel continue, à son tour, de s'améliorer au fur et à mesure que nous augmentons la taille du corpus d'apprentissage.

Nous constatons également que les résultats de la catégorie *Posologie* bénéficient de l'augmentation des données sur lesquelles le modèle a été appris. Le faible rappel (22.4%) de *Posologie* peut en effet être dû au regroupement de données de nature différente sous la même catégorie. En définissant la catégorie *Posologie* dans le guide d'annotation, quatre sous-catégories (dosage, concentration, forme galénique du médicament et mode d'administration) distinctes mais homogènes entre elles ont été regroupées ensemble. Contrairement à un être humain qui est capable de distinguer entre ces différentes formes (ex. : *par voie orale* comme mode d'administration d'un traitement, *sirop* comme forme galénique et *25mg* comme concentration), un *CRF* peut avoir plus de difficultés à effectuer cette différence. Il aurait été préférable de séparer les sous-catégories traitées, en redéfinissant la catégorie *Posologie* pour y garder uniquement les entités numériques et en créant une nouvelle catégorie qui regroupe les entités textuelles, telles que le mode d'administration et la forme galénique du médicament. Toutefois, cette distinction ne fait pas partie de notre sujet de recherche qui met plus l'accent sur la catégorie *Événement*. Il est à noter que les informations liées à la catégorie *Posologie* sont en effet importantes dans une étape postérieure afin d'établir les correspondances entre un événement secondaire et un traitement.

Par ailleurs, la tâche de détection des entités relatives aux deux catégories *Événement* et *Indication* demeure complexe, menant ainsi à des résultats moins satisfaisants. Cette complexité peut être expliquée par le recouvrement entre les deux catégories, le nombre limité d'annotation par rapport au corpus, la variété des entités liées à ces deux catégories et la complexité du corpus étudié. En effet, la ressemblance entre *Indication* et *Événement* est évalué par les expériences 2 et 3. Nous observons dans le tableau 3.2 une légère baisse du rappel du système mono-classe qui détecte les événements secondaires par rapport à un système multi-classes. Ceci ne correspond pas à notre deuxième hypothèse de travail, prouvant ainsi que le modèle *CRF* est enrichi par les annotations des autres catégories. Toutefois, les résultats du tableau 3.3 peut illustrer ce recouvrement avec l'augmentation du rappel de la catégorie *Événement* (passer de 8.3% à 14.1%). Ils montrent pourtant que le rappel de la catgorie *Événement* baisse par rapport à celui d'*Indication* (33% dans l'expérience 1), ce qui fait que la somme des *vrais positifs* des deux catégories *Événement* et *Indication* de l'expérience 1 reste plus élevée que celle de la catégorie *Événement* de l'expérience 3.

Le recouvrement entre les deux catégories en question peut être expliqué par l'absence de traits *temporels* dans les caractéristiques choisies lors de la création du modèle. La difficulté de différencier entre les catégories *Événement* et *Indication* repose essentiellement sur une notion de temporalité. Les indications sont des problèmes qui se manifestent avant la prise d'un traitement, tandis que les événements secondaires sont des problèmes provoqués par des médicaments et qui commencent à apparaitre après un certain temps de la prise du traitement. Un *CRF* aura ainsi plus de difficultés qu'un être humain à faire une telle inférence pour pouvoir distinguer entre les deux catégories sans le recours à des indices temporels. Cependant, cette fu-

sion dégrade les scores des autres catégories, d'où l'intérêt de traiter le recouvrement entre *Événement* et *Indication* à l'aide de deux relations différentes *motif*/*indication* et *cause*/*événement*, voire des attributs pour les catégories.

Quant à la variété des annotations par catégorie et leur pourcentage par rapport au reste du corpus, nous nous sommes rendues compte qu'elle joue un rôle dans la difficulté d'automatiser le processus d'extraction d'événements secondaires, en raison de l'absence de régularité comme montre le tableau 2.10 et la figure 2.7. Les deux catégories *Événement* et *Posologie* renferment des annotations très variées par comparaison à *Traitement* et *Indication* (la régularité d'*Indication* concerne notamment le corpus *migraine*). Cette variété est en effet produite par les particularités syntaxiques des messages des patients et la nature du corpus. Un corpus grand public tiré d'un forum de santé se caractérise par l'utilisation d'un registre familier, ainsi que par la péresence des fautes d'orthographe et des expressions imagées pour désigner les événements secondaires produits par les médicaments. À titre d'exemple, les patients ont tendance à exprimer leurs problèmes de manières différentes qui peuvent varier entre *voir des Schtroumpfs* (avoir des hallucinations), *avoir le QI d'une carotte* (avoir des problèmes de concentration), *être terrassé* (être fatigué). Ces expressions et constructions syntaxiques variées rendent la tâche plus complexe et s'avèrent difficile à être repérées automatiquement.

Outre l'absence de régularité, dans la figure 2.5, le nombre des tokens annotés dans les deux corpus s'avère limité par rapport à la taille du corpus, ce qui peut expliquer la nécessité d'annoter un corpus plus grand pour atteindre des résultats plus satisfaisants. Un grand corpus annoté s'avère également utile pour le repérage des événements secondaires malgré l'utilisation de ressources externes qui nous permet de faciliter la tâche et le recours aux lemmes lors de la projection du lexique pour gérer les différentes formes des tokens.

Enfin, en ce qui concerne le problème du forum *antidépresseurs-anxiolytiques*, les résultats du tableau 3.4 valide notre quatrième hypothèse de travail. Il existe dans ce forum des événements secondaires qui sont liés à la fois à un traitement et à la psychothérapie. De même, certains problèmes peuvent être étiquetés comme *Événement* ou *Indication* selon le contexte dans lequel ils figurent. C'est par exemple le cas de *fortes douleurs au dos et aux côtes* étiqueté comme *Événement* dans le message « *Voila je prends rivotril le soir pour dormir et là j'ai de fortes douleurs au dos et aux cotes, le hic c que j'ai pris 8 comprimés d'ixprim depuis quelques heures, est ce que prendre le rivotril tout de meme en plus ce n'est pas risqué ? je suis super énervée et tjs les douleurs donc je dormirai pas merci pour toute réponse* » et *douleurs intercostales* étiquetée comme *Indication* dans « *L ixprim c pour t douleur intercostales je suppose nan ? ya de la morphine dedans relie la notice...8 comprime c abuzer* ». Ces difficultés sont moins présentes dans le corpus *migraine*, ce qui mène à un faible pourcentage des formes différentes par rapport aux occurrences pour la catégorie *Indication*, comme le montre la figure 2.7, et une amélioration des résultats d'*Événement* et *Indication*. Le modèle 10 *migraine* se révèle plus performant que le modèle 40 *antidépresseurs-anxiolytiques*, ce qui s'explique d'une part par la présence de plus de régularités au niveau des entités qui se rapportent à la catégorie *Événement* et d'autre part, par la différence entre les problèmes liés aux symptômes d'une maladie et ceux produits par la prise d'un traitement. Le sujet sur lequel portent les fichiers du corpus a aussi un impact sur les résultats. La figure 2.6 montre que le corpus *migraine* englobe plus de fichiers pertinents par rapport à notre sujet de recherche.

3.3.2 Résultats des configurations *out domain*

Les tableaux 3.5 et 3.6 montrent les résultats des configurations *out domain* qui concernent la cinquième et la sixième hypothèse de travail. Nous remarquons dans le premier tableau que le modèle 40 *antidépresseurs-anxiolytiques* est plus efficace quand il est testé sur un corpus qui traite du même sous-domaine médical, validant ainsi notre cinquième hypothèse. Nous observons en effet une certaine dégradation des résultats globaux obtenus par l'application du modèle 40 *antidépresseurs-anxiolytiques* sur le corpus *migraine* par rapport aux résultats obtenus par l'application du même modèle sur le corpus *antidépresseurs-anxiolytiques*.

Cette dégradation peut être expliquée par la disparité entre les exemples qui figurent dans la référence et l'hypothèse, notamment pour les catégories *Indication* (passer de 33.3 à 4.6 de F-mesure) et *Traitement* (passer de 69.6 à 47.8 de F-mesure). Dans le corpus *antidépresseurs-anxiolytiques*, nous trouvons *dépression*, *crise d'angoisse* et *anxiété* comme les problèmes qui se rapportent à cette dernière catégorie, tandis que dans le corpus *migraine* ces problèmes concernent plus des entités telles que *maux de tête* et *céphalée*. Il en va de même pour *Traitement* où *Xanax* et *Seroplex* sont les médicaments utilisés en cas de dépression, mais *Imigrain* et *Nocertone* sont ceux prescrits en cas de crise de migraines.

Par ailleurs, un apprentissage fusionné sur les deux corpus traités améliore globalement les résultats, ce qui signifie que notre sixième hypothèse de travail n'est pas vérifiée. Après le rajout des dix fichiers *migraine* aux données d'apprentissage du modèle 40 *antidépresseurs-anxiolytiques*, nous observons que le modèle est enrichi par les données du corpus *migraine*. Le système devient plus performant et fournit de meilleurs résultats, notamment en effectuant le test sur le corpus *migraine*, malgré la difficulté de faire la différence entre les deux catégories *Indication* et *Événement*.

Les résultats de ces expériences sont mentionnés dans le tableau 3.6 qui montre une importante amélioration au niveau du rappel global (entre 8% et 18% environ selon le corpus de test). De même, si nous comparons ces résultats avec ceux obtenus par les modèles appris sur un seul sous-domaine (un seul corpus), nous remarquerons une augmentation de la F-mesure des catégories *Événement* et *Traitement*. Ces résultats confirment que le sous-domaine médical étudié a une incidence sur l'efficacité du système (dans notre cas, le sous-domaine *migraine* est plus facile à traiter que celui des *antidépresseurs-anxiolytiques*). Ils mettent aussi en lumière un problème relatif à la constitution de la référence du corpus de base auquel nous n'avons pas pensé : 50 fils de discussion ont été tirés aléatoirement du corpus, 40 parmi ces fichiers ont été annotés par ordre alphabétique des noms des fichiers selon le processus itératif expliqué dans la section 2.2. Les dix derniers fichiers (commençants par les lettres *J*, *K*, *L* et *P*) constituent ainsi le corpus de test pour le corpus de base. Ces derniers fichiers ne sont pas tous pertinents dans la mesure où certains portent sur des discussions générales et ne comportent pas des annotations intéressantes (ex. : *journal-de-bord* ou *poeme-dependances-sevrage*). Un tel problème apparait dans la figure 2.6 où nous pouvons observer que le pourcentage des fichiers portant sur un sujet général autre que les événements secondaires et les problèmes liés aux médicaments une partie non négligeable du corpus (32%).

3.3.3 Travaux futurs

D'après les résultats obtenus par les différentes expériences et leurs analyses, il existe plusieurs pistes potentielles pour améliorer la méthode implémentée. Parmi

ces possibilités d'amélioration, figure le typage des fils de discussion du corpus en s'inspirant du travail de [Yang et al., 2013]. Le corpus étudié est tiré d'un forum de santé qui contient des fils de discussion dans lesquels les patients parlent des événements secondaires (messages portant sur un médicament) et ceux qui ne le sont pas (discussion concernant un sujet général hors médicament). Il nous semble ainsi pertinent de rajouter à la phase de prétraitement du corpus un système permettant de réaliser un tel typage. Ce dernier peut prendre par exemple la forme d'un système à base de règles accompagné d'une fonction seuil définit pour calculer le nombre des noms de médicaments qui figurent au sein du fil de discussion. Dans ce cas, les fils contenant *N* nombre de traitements sont uniquement retenus, ce qui permet d'avoir un corpus homogène au niveau des annotations (par exemple, ne pas avoir des données de test qui renferment moins de discussions pertinentes pour notre étude que les données d'apprentissage ou inversement). Un système de typage peut également consister à l'application d'un algorithme de classification ou de *clustering* qui permet de regrouper les fils de discussion les plus proches selon certaines classes.

Un autre moyen d'amélioration consiste à ne pas traiter le problème de détection d'événements secondaires d'un point de vue strict syntaxiquement, mais plutôt de l'aborder d'une autre perspective plus relâchée. Cette dernière permet d'identifier uniquement les têtes de syntagmes qui symbolisent les événements secondaires et non pas d'extraire toute la portion textuelle exacte qui représente un événement. Il s'agit en effet de capturer des entités qui se rapportent à des classes sémantiques ou des concepts de l'*UMLS* ainsi que des déclencheurs qui côtoient souvent les termes indiquant la présence d'un problème de santé chez un patient. À titre d'exemple, l'événement *mal au ventre* serait détecté à l'aide de la présence du déclencheur *mal* qui marque un problème et du terme *ventre* qui correspond à un groupe sémantique de l'*UMLS* (partie anatomique). Une telle solution permet de simplifier le modèle, de le généraliser, ainsi que d'augmenter son rappel étant donné que les entités qui devront être annotées sont moins variées. Cette simplification va de même pour l'évaluation des résultats en adoptant une méthode d'évaluation moins strict concernant les frontières des entités détectées. La généralisation du modèle créé peut être aussi effectuée en utilisant les identifiants de l'*UMLS* (*CUI*) pour regrouper les annotations sous un même concept plus générique. Les entités telles que *prendre quelques kilos*, *pris 15 kilos* et *prise de poids* seront regroupés sous l'identifiant du concept général *prise de poids*. Une telle solution restreint considérablement les variations des portions textuelles à annoter.

L'analyse des sentiments qui figurent dans les messages à l'aide des pronoms [Ginn et al., 2014] et d'autres lexiques affectifs plus élaborés que celui que nous avons utilisé comme *SentiWordNet* [Sharif et al., 2014] est considérée comme une autre piste d'amélioration du modèle *CRF*. L'analyse des pronoms permet de filtrer les messages qui n'expriment pas une expérience personnelle d'un patient (ex. : une mère qui exprime son inquiétude pour son fils qui est sous traitement ou un utilisateur qui publie un commentaire concernant un médicament qu'il n'a pas testé) et le lexique permet de déterminer la positivité ou la négativité des mots. L'application de telle analyse nécessite de rajouter d'autres caractéristiques au modèle. Les pronoms et le vocabulaire de *SentiWordNet* seront ainsi pris en compte lors de la création de ce dernier.

Par ailleurs, nous pouvons effectuer une validation croisée (*k-fold cross-validation*) pour pallier les particularités d'un sous-corpus. Dans un tel cas, le corpus est divisé en *N* nombre de sous-parties dont *N-1* sont consacrées à l'apprentissage et

la *nième* partie est dédiée au test. Cette opération est répétée N fois et les résultats sont ensuite calculés en fonction de la moyenne des scores obtenus à chaque itération. Cette technique d'échantillonnage donne des résultats d'évaluation plus fiables, puisque le système n'est pas appris ou testé sur un sous-corpus qui englobe des cas de figures particulièrs. Le rajout d'autres groupes sémantiques de l'*UMLS*, pour rapporter plus d'information au *CRF* sur les entités à ne pas annoter, et la gestion des entités discontinues (dans ce travail, nous avons annoté les entités discontinues, mais nous ne les avons pas pris en compte lors de la création du modèle) constituent également des pistes à exploiter dans les travaux futurs.

Il serait également intéressant d'établir la correspondance entre le vocabulaire des patients et celui utilisé par les spécialistes. À cette fin, il semble crucial de traiter les problèmes des variantes linguistiques, des relations *partie de* et de l'emploi de différents registres et des expressions imagées. Ces problèmes peuvent être surmontés en ayant recours à des techniques de verbalisation et de nominalisation ou en utilisant des ontologies et des corpus parallèles et comparables (spécialisé/non-spécialisé comme dans [Deléger and Zweigenbaum, 2009] et [Deléger, 2009]. En effet, notre système ne gère pas les variantes linguistiques, même si la projection des listes des événements secondaires et celles extraites de l'*UMLS* est effectuée en se référant aux lemmes des entités. Nous prenons ainsi en compte les cas des apostrophes, les différentes casses, la présence/absence des lettres diacritiques et les différences entre pluriel et singulier tels que *sinusite* vs. *sinusites*, *seroplex* vs. *SEROPLEX*. La méthode implémentée ne gère pas non plus les cas comme *peur de mourir* vs. *peur de la mort*, *douleurs au dos* vs. *douleurs dorsales* et *mal au ventre* vs. *mal à l'estomac*. Quant à l'usage des différents registres, il s'agit de faire le lien entre *être claqué* et *être très fatigué* ou entre *couler l'eau comme une fontaine* et *transpirer*.

Un dernier point d'amélioration consiste à faire le lien entre les événements secondaires et les traitements en vue de déterminer le traitement qui a causé l'événement secondaire extrait du message. La gestion de ce rattachement concerne la distinction entre les événements liés à la prise d'un médicament et ceux qui sont dus aux effets conjugués avec d'autres problèmes. Les informations associées aux événements secondaires telles que le dosage, le mode d'adminstration et la durée du traitement sont ainsi utiles pour établir une telle correspondance.

Conclusion

La pharmacovigilance a comme but la surveillance des médicaments et la prévention des patients quant aux risques des événements secondaires en cas d'utilisation d'un médicament. Elle repose principalement sur le recueil d'informations à partir des notifications spontanées des praticiens et des patients. Néanmoins, les sites médicaux et les forums de santé semblent représenter aujourd'hui une source importante d'information qui peut être exploitée à des fins de pharmacovigilance. La quantité d'informations médicales disponibles sur les forums ainsi que sa croissance exponentielle permettent de mettre en œuvre de nouveaux mécanismes de recherche et d'extraction d'information, complémentaires aux moyens traditionnels, qui peuvent s'adresser à la question de détection d'événements secondaires.

Notre étude porte sur l'étude des messages publiés sur deux forums de *Doctissimo* qui se rapportent aux thèmes des antidépresseurs-anxiolytiques et de la migraine. Elle vise l'utilisation des *CRF* pour mettre en place une chaîne de traitements, capable de détecter automatiquement les événements secondaires provoqués par les médicaments. Nous veillons ainsi à la mise en application d'une méthode à base d'apprentissage supervisé conçu à cet égard. La chaîne de traitements implémentée n'est qu'une première approche possible qui peut être améliorée. Elle se compose de deux grandes étapes, l'annotation d'une partie du corpus après l'élaboration d'un guide et la définition d'une grammaire d'annotation pour la constitution de la référence, ainsi que la création du modèle *CRF* et le choix de ses caractéristiques et ses configurations.

Plusieurs expériences sont réalisées afin d'évaluer la performance du système et de tester certaines hypothèses de travail. Les premiers résultats de l'évaluation montrent que les trois phases qui se résument en l'étude des particularités du corpus traité, l'annotation de la référence et le choix des caractéristiques du modèle constituent des étapes importantes dans ce travail. En effet, les résultats de ces expériences et leurs analyses nous ont permis d'observer qu'un système à base d'apprentissage statistique est dépendant de la taille du corpus d'apprentissage. Plus les données d'entraînement sont de taille importante, plus le système est efficace dans la réalisation de ses prédictions.

Nous avons également pu constater qu'un modèle multi-calsses de détection d'événements secondaires est plus performant qu'un modèle mono-classe. Les informations concernant les catégories associées aux événements secondaires, telles que *Posologie* et *Traitement*, enrichissent le modèle et le rend plus robuste. Il en va de même pour les données sur lesquelles le modèle a été appris. Nous avons observé qu'un modèle est plus robuste en utilisant les données d'apprentissage portant sur les deux sous-domaines étudiés (antidépresseurs-anxiolytiques et migraine) au lieu de celles qui sont spécifiques à un seul sous-domaine médical. Toutefois, un modèle appliqué sur un autre sous-domaine que celui relatif aux données d'entraînement fournit de moins bons résultats.

Un dernier point remarqué concerne l'incidence d'un sous-domaine sur l'effica-

cité du modèle. Nous nous somme rendues compte qu'un sous-domaine médical peut s'avérer plus difficile à traiter qu'un autre en raison des particularités des entités à extraire. Ces dernières peuvent se rapporter à la fois à plusieurs catégories étudiées, ce qui complexifie la tâche d'extraction automatique des entités biomédicales.

Face aux résultats obtenus par les différentes expériences, plusieurs pistes d'amélioration sont ainsi envisageables. Certaines concernent la phase de prétraitement du corpus, telles que le typage des fils de discussion pour filtrer les messages non-pertinents, et d'autres sont relatives à l'approche adoptée pour mettre en application la méthode adoptée. Parmi ces dernières figurent l'extraction des têtes de syntagmes à la place des portions textuelles complètes qui représentent les événements secondaires, l'apprentissage par validation croisée pour pallier les particularités d'un sous-corpus, sans parler de l'ajout d'autres caractéristiques au modèle créé et la gestion des entités discontinues.

Outre ces modifications, il serait intéressant de considérer dans les travaux futurs la correspondance entre le vocabulaire patients et celui de spécialistes, ainsi que d'établir le lien entre traitement et événement secondaire. Dans le premier cas, les cas des variantes linguistiques et des différents registres utilisés sont traités, tandis que dans le second, la distinction entre un événement secondaire lié à un traitement et un autre relatif à un effet conjugué d'un autre problème est effectuée.

Même si plusieurs améliorations concernant la mise en place de notre méthode sont à prévoir et beaucoup d'étapes restent à affranchir pour atteindre de meilleurs performances, les apports de ce travail préliminaire sont prometteurs en matière de détection d'événements secondaires. Les meilleurs résultats de notre système, obtenus quand le modèle *CRF* est appris sur les deux corpus étudiés, ne sont pas loin de ceux des systèmes présents dans l'état de l'art. De plus, l'orientation applicative de ce travail nous a permis de mettre en œuvre des aspects techniques importants, ainsi que de mener une réflexion méthodique et pratique sur la question de la détection d'entités biomédicales.

BIBLIOGRAPHIE

[Agrawal et al., 1993] Agrawal, R., Imielinsk, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD Conference*, pages 207–216. – Cité page 19.

[Albert et al., 2003] Albert, S., Gaudan, S., Knigge, H., Raetsch, A., Delgado, A., Huhse, B., Kirsch, H., Albers, M., Rebholz-Schuhmann, D., and Koegl, M. (2003). Computer-assisted generation of a protein-interaction database for nuclear receptors. *Mol Endocrinol*, 17:135–143. – Cité page 16.

[Benton et al., 2011] Benton, A., Ungar, L., Hill, S., Hennessy, S., Mao, J., Chung, A., Leonard, C. E., and Holmes, J. H. (2011). Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. *J Biomed Inform*. – Cité page 21.

[Brown et al., 1992] Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Comput Linguist*, 18:467–479. – Cité page 41.

[Bundschus et al., 2008] Bundschus, M., Dejori, M., Stetter, M., Tresp, V., and Kriegel, H.-P. (2008). Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, 9. – Cité pages 16 et 20.

[Bunescu et al., 2006] Bunescu, R., Mooney, R., Ramani, A., and Marcotte, E. (2006). Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from Medline. In *BioNLP '06 Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, pages 49–56. – Cité page 16.

[Chang et al., 2004] Chang, J. T., Schütze, H., and Altman, R. B. (2004). GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics*, 20. – Cité page 16.

[Chee et al., 2011] Chee, B. W., Berlin, R., and Schartz, B. (2011). Predicting adverse drug events from personal health messages. In *AMIA Annual Symposium Proceedings*, pages 217–226. – Cité page 11.

[Chena et al., 2008] Chena, E. S., Hripcsakd, G., Xud, H., Markatoue, M., and Friedmand, C. (2008). Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. In *J Am Med Inform Assoc*, pages 87–98. – Cité page 17.

[Chun et al., 2006] Chun, H.-W., Tsuruoka, Y., Kim, J.-D., Shiba, R., Nagata, N., Hishiki, T., and Tsujii, J. (2006). Extraction of gene-disease relations from MEDLINE using domain dictionaries and machine learning. In *Proceedings of the Pacific Symposium on Biocomputing (PSB) 11*, pages 4–15. – Cité page 16.

[Cimino et al., 1997] Cimino, J. J., Elhanan, G., and Zeng, Q. (1997). Supporting infobuttons with terminological knowledge. In *AMIA Annu Symp Proc*, pages 528–532. – Cité page 22.

[Cohen and Hersh, 2005] Cohen, A. M. and Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Brief Bioinform*, 6:57–71. – Cité pages 15 et 16.

[Dai et al., 2010] Dai, H.-J., Chang, Y.-C., Tsai, R. T.-H., and Hsu, W.-L. (2010). New challenges for biological text-mining in the next decade. *Journal of Computer Science and Technology*, 25:7–18. – Cité page 16.

[de Quincey and Kostkova, 2010] de Quincey, E. and Kostkova, P. (2010). Early warning and outbreak detection using social networking websites: the potential of twitter. pages 21–24. – Cité page 12.

[Deléger, 2009] Deléger, L. (2009). *Exploitation de corpus parallèles et comparables pour la détection de correspondances lexicales : application au domaine médical*. PhD thesis. – Cité page 53.

[Deléger and Zweigenbaum, 2008a] Deléger, L. and Zweigenbaum, P. (2008a). Aligning lay and specialized passages in comparable medical corpora. In *Stud Health Technol Inform*, pages 89–94. – Cité page 22.

[Deléger and Zweigenbaum, 2008b] Deléger, L. and Zweigenbaum, P. (2008b). Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In *AMIA Annu Symp Proc*, pages 146–150. – Cité page 23.

[Deléger and Zweigenbaum, 2009] Deléger, L. and Zweigenbaum, P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *BUCC'09 Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*. – Cité page 53.

[Demaine et al., 2006] Demaine, J., Martin, J., Wei, L., and de Bruijn, B. (2006). LitMiner: integration of library services within a bio-informatics application. *Biomed Digit Libr*, 3:3–11. – Cité page 16.

[Denecke, 2014] Denecke, K. (2014). Extracting medical concepts from medical social media with clinical NLP tools : A qualitative study. In *Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*. – Cité page 40.

[Elhadad, 2006] Elhadad, N. (2006). Comprehending technical texts: Predicting and defining unfamiliar terms. In *AMIA Annu Symp Proc*, pages 239–243. – Cité page 22.

[Finkel et al., 2004] Finkel, J., Dingare, S., Nguyen, H., Nissim, M., Manning, C., and Sinclair, G. (2004). Exploiting context for biomedical entity recognition: from syntax to the web. In *NLPBA '04 Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 88–91. – Cité page 19.

[Fox and Fallows, 2003] Fox, S. and Fallows, D. (2003). Internet health ressources. Technical Report 42. – Cité page 21.

[Fukuda et al., 1998] Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T. (1998). Toward information extraction: identifying protein names from biological papers. In *Pac Symp Biocomput*, pages 707–718. – Cité page 16.

[Fundel et al., 2007] Fundel, K., K̈ffner, R., and Zimmer, R. (2007). Relex—relation extraction using dependency parse trees. *BMC Bioinformatics*, 23:365–371. – Cité page 16.

[Gaizauskas et al., 2003] Gaizauskas, R., G.Demetriou, P.J.Artymiuk, and Willett, P. (2003). Protein structures and information extraction from biological texts: The PASTA system. *BMC Bioinformatics*, 19:135–143. – Cité page 16.

[Ginn et al., 2014] Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., Connor, K., Sarker, A., Smith, K., and Gonzalez, G. (2014). Mining Twitter for adverse drug reaction mentions : A corpus and classification benchmark. In *Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*. – Cité pages 11, 25 et 52.

[Grouin, 2013] Grouin, C. (2013). *Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique*. PhD thesis. – Cité page 29.

[Gurulingappa et al., 2012] Gurulingappa, H., Rajput, A. M., and Toldo, L. (2012). Extraction of potential adverse drug events from medical case reports. *J Biomed Semantics*, 3. – Cité page 20.

[Hanisch et al., 2005] Hanisch, D., Fundel, K., Mevissen, H.-T., Zimmer, R., and Fluck, J. (2005). ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6. – Cité page 17.

[Harpaz et al., 2010] Harpaz, R., Haerian, K., Chase, H. S., and Friedman, C. (2010). Statistical mining of potential drug interaction adverse effects in FDA's spontaneous reporting system. In *AMIA Annu Symp Proc*, pages 281–285. – Cité page 19.

[jae Kim et al., 2006] jae Kim, J., Zhang, Z., Park, J. C., and Ng, S.-K. (2006). Biocontrasts: extracting and exploiting protein-protein contrastive relations from biomedical literature. *BMC Bioinformatics*, 22:597–605. – Cité page 16.

[Kandula et al., 2010] Kandula, S., Curtis, D., and Zeng-Treitler, Q. (2010). A semantic and syntactic text simplification tool for health content. In *AMIA Annu Symp Proc*, pages 366–370. – Cité page 22.

[Kang et al., 2014] Kang, N., Singh, B., Bui, C., Afzal, Z., van Mulligen, E. M., and Kors, J. A. (2014). Knowledge-based extraction of adverse drug events from biomedical text. *BMC Bioinformatics*, 15. – Cité pages 16, 18 et 40.

[Karamanis et al., 2007] Karamanis, N., Lewin, I., Seal, R., Drysdale, R., and Briscoe, E. (2007). Integrating natural language processing with FlyBase curation. In *Pac Symp Biocomput*, pages 245–256. – Cité page 17.

[Kazama et al., 2002] Kazama, J., Makino, T., Hitachi, Y. O., and Tsujii, J. (2002). Tuning support vector machines for biomedical named entity recognition. In *TBioMed '02 Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain*, pages 1–8. – Cité page 19.

[Keselman et al., 2008] Keselman, A., Smith, C. A., Divita, G., Kim, H., Browne, A. C., Leroy, G., and Zeng-Treitler, Q. (2008). Consumer health concepts that do not map to the UMLS: Where do they fit? *J Am Med Inform Assoc*, 15:496–505. – Cité page 22.

[Keselman et al., 2007] Keselman, A., Tse, T., Crowell, J., Browne, A., Ngo, L., and Zeng, Q. (2007). Assessing consumer health vocabulary familiarity: an exploratory study. *J Med Internet Res*, 9. – Cité page 22.

[Kongkaew et al., 2008] Kongkaew, C., Noyce, P. R., and Ashcroft, D. M. (2008). Hospital admissions associated with adverse drug reactions : a systematic review of prospective observational studies. *Ann Pharmacother*, 42:1017–1025. – Cité page 11.

[Lavergne et al., 2010] Lavergne, T., Capé, O., and Yvon, F. (2010). Practical very large scale CRFs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 504–513. – Cité page 29.

[Leaman et al., 2010] Leaman, R., Wojtulewicz, L., Sullivan, R., Yang, A. S. J., and Gonzalez, G. (2010). Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks. In *BioNLP '10 Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 117–125. – Cité pages 11, 21 et 40.

[Li et al., 2013] Li, Q., Deleger, L., Lingren, T., Zhai, H., Kaiser, M., Stoutenborough, L., Jegga, A. G., Cohen, K. B., and Solti, I. (2013). Mining FDA drug labels for medical conditions. *BMC Med Inform Decis Mak*. – Cité page 21.

[McDonald et al., 2004] McDonald, D. M., Chen, H., Su, H., and Marshall, B. B. (2004). Extracting gene pathway relations using a hybrid grammar : the Arizona relation parser. *Bioinformatics*, 20:3370–3378. – Cité page 16.

[McDonald and Pereira, 2005] McDonald, R. and Pereira, F. (2005). Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*. – Cité page 16.

[Minard et al., 2011] Minard, A.-L., Ligozat, A.-L., Abacha, A. B., Bernhard, D., Cartoni, B., Deléger, L., Grau, B., Rosset, S., Zweigenbaum, P., and Grouin, C. (2011). Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification. In *J Am Med Inform Assoc*, pages 588–593. – Cité page 20.

[Nikfarjam and Gonzalez, 2011] Nikfarjam, A. and Gonzalez, G. H. (2011). Pattern mining for extraction of mentions of adverse drug reactions from user comments. In *AMIA Annual Symposium Proceedings*, pages 119–126. – Cité page 22.

[OMS, 1969] OMS (1969). Pharmacovigilance internationale : rôle de l'hôpital, rapport d'une réunion de l'OMS. Technical Report 425. Réunion OMS sur le rôle de l'hôpital dans la pharmacovigilance internationale (1968: Genève, Switzerland) World Health Organization. – Cité page 11.

[Pak and Paroubek, 2010] Pak, A. and Paroubek, P. (2010). Construction d'un lexique affectif pour le français à partir de Twitter. In *TALN 2010 17ème Conférence sur le Traitement Automatique des Langues Naturelles*. – Cité page 40.

[Parker et al., 2013] Parker, J., Wei, Y., Yates, A., Frieder, O., and Goharian, N. (2013). A framework for detecting public health trends with Twitter. In *ASONAM '13 Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 556–563. – Cité page 12.

[Plovnick and Zeng, 2003] Plovnick, R. M. and Zeng, Q. T. (2003). Reformulation of consumer health queries with professional terminology: A pilot study. *J Med Internet Res*, 6. – Cité page 22.

[Rink et al., 2011] Rink, B., Harabagiu, S., and Roberts, K. (2011). Automatic extraction of relations between medical concepts in clinical texts. *J Am Med Inform Assoc*, 18:594–600. – Cité page 20.

[Roberts et al., 2008] Roberts, A., Gaizauskas, R., and Hepple, M. (2008). Extracting clinical relationships from patient narratives. In *BioNLP '08 Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 10–18. – Cité page 19.

[Rocha et al., 1993] Rocha, R. A., Rocha, B. H., and Huff, S. M. (1993). Automated translation between medical vocabularies using a frame-based interlingua. In *Proc Annu Symp Comput Appl Med Care*, pages 690–694. – Cité page 22.

[Sharif et al., 2014] Sharif, H., Zaffar, F., Abbasi, A., and Zimbra, D. (2014). Detecting adverse drug reactions using a sentiment classification framework. In *ASE International Conference on Social Computing*. – Cité page 52.

[Si et al., 2005] Si, L., Kanungo, T., and Huang, X. (2005). Boosting performance of bio-entity recognition by combining results from multiple systems. In *BIOKDD '05 Proceedings of the 5th international workshop on Bioinformatics*, pages 76–83. – Cité page 19.

[Smith et al., 2008] Smith, L., Tanabe, L. K., nee Ando, R. J., Kuo, C.-J., Chung, I.-F., Hsu, C.-N., Lin, Y.-S., Klinger, R., Friedrich, C. M., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C. A., Povinelli, R. J., Vlachos, A., Baumgartner, W. A., Hunter, L., Carpenter, B., Tsai, R. T.-H., Dai, H.-J., Liu, F., Chen, Y., Sun, C., Katrenko, S., Adriaans, P., Blaschke, C., Torres, R., Neves, M., Nakov, P., Divoli, A., na López, M. M., Mata, J., and Wilbur, W. J. (2008). Overview of biocreative II gene mention recognition. *Genome Biol*, 9. – Cité page 19.

[Sondhi et al., 2010] Sondhi, P., Gupta, M., Zhai, C. X., and Hockenmaier, J. (2010). Shallow information extraction from medical forum data. In *COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 158–166. – Cité page 21.

[Tsea and Soergela, 2003] Tsea, T. and Soergela, D. (2003). Exploring medical expressions used by consumers and the media: an emerging view of consumer health vocabularies. In *AMIA Annu Symp Proc*, pages 674–678. – Cité page 22.

[Tsuruoka et al., 2009] Tsuruoka, Y., Tsujii, J., and Ananiadou, S. (2009). Stochastic gradient descent training for L1-regularized log-linear models with cumulative penalty. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of AFNLP*, pages 477–485. – Cité page 39.

[Uzuner et al., 2011] Uzuner, O., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. In *J Am Med Inform Assoc*, pages 552–556. – Cité page 20.

[Wang et al., 2011] Wang, W., Haerian, K., Salmasian, H., Harpaz, R., Chase, H., and Friedman, C. (2011). A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from PubMed citations. In *AMIA Annu Symp Proc*, pages 164–170. – Cité page 19.

[Wang et al., 2009] Wang, X., Hripcsak, G., Markatoub, M., and Friedman, C. (2009). Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc*, 16:328–337. – Cité page 20.

[White et al., 2013] White, R. W., Tatonetti, N. P., Shah, N. H., Altman, R. B., and Horvitz, E. (2013). Web-scale pharmacovigilance: listening to signals from the crowd. pages 404–408. – Cité page 12.

[Yang et al., 2013] Yang, M., Wang, X., and Kiang, M. (2013). Identification of consumer adverse drug reaction messages on social media. In *PACIS 2013 Proceedings*. – Cité pages 21, 22 et 52.

[Yates and Goharian, 2013] Yates, A. and Goharian, N. (2013). ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In *ECIR'13 Proceedings of the 35th European conference on Advances in Information Retrieval*, pages 816–819. – Cité page 22.

[Yates et al., 2013] Yates, A., Goharian, N., and Frieder, O. (2013). Extracting adverse drug reactions from forum posts and linking them to drugs. In *ACM SIGIR Workshop on Health Search and Discovery: Helping Users and Advancing Medicine*. – Cité pages 12 et 22.

[Yeleswarapu et al., 2014] Yeleswarapu, S. J., Rao, A., Joseph, T., Saipradeep, V. G., and Srinivasan, R. (2014). A pipeline to extract drug-adverse event pairs from multiple data sources. *BMC Med Inform Decis Mak*, 14. – Cité page 21.

[Zeng and Cimino, 1996] Zeng, Q. and Cimino, J. J. (1996). Mapping medical vocabularies to the Unified Medical Language System. In *Proc AMIA Annu Fall Symp*, pages 105–109. – Cité page 22.

[Zeng and Tse, 2006] Zeng, Q. T. and Tse, T. (2006). Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc*, 13:24–29. – Cité pages 22 et 23.

[Zeng-Treitler et al., 2007] Zeng-Treitler, Q., Goryachev, S., Kim, H., Keselman, A., and Rosendale, D. (2007). Making texts in electronic health records comprehensible to consumers: a prototype translator. In *AMIA Annu Symp Proc*, pages 846–850. – Cité page 22.

[Zeng-Treitler et al., 2008] Zeng-Treitler, Q., Goryachev, S., Tse, T., Keselman, A., and Boxwala, A. (2008). Estimating consumer familiarity with health terminology: a context-based approach. *J Am Med Inform Assoc*, 15:349–356. – Cité page 22.

[Zhao, 2004] Zhao, S. (2004). Named entity recognition in biomedical texts using an HMM model. In *The International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 84–87. – Cité page 19.

[Zhou et al., 2004] Zhou, G., Zhang, J., Su, J., Shen, D., and Tan, C. (2004). Recognizing names in biomedical texts: A machine learning approach. *Bioinformatics*, 20:1178–1190. – Cité page 17.

[Zielstorff, 2003] Zielstorff, R. D. (2003). Controlled vocabularies for consumer health. *J Biomed Inform*. – Cité page 21.

[Zweigenbaum et al., 2007] Zweigenbaum, P., Demner-Fushman, D., Yu, H., and Cohen, K. B. (2007). Frontiers of biomedical text mining: current progress. *Brief Bioinform*, 8:358–375. – Cité pages 15 et 18.

INDEX

Annexes

SCRIPTS ÉLABORÉS

A.1 Constitution du corpus

A.1.1 Chaîne de prétraitements

```
#!/usr/bin/perl

<<DOC;
 MEGAHED Dalia
 Avril 2014

 Usage : perl extractTableHtmlCreateTab.pl repertoire-à-parcourir

 Le programme prend en entrée le répertoire contenant les fichiers html à
    traiter (forum Doctissimo.fr) et extrait de chaque fichier html (des
    tableaux ayant un attribut class="messagetable") le nom de l'utilisateur,
    son profil, son message et la date pour produire en sortie un répertoire
    contenant les fichiers txt correspondants
 Chaque fichier txt est un fichier tabulaire (7 colonnes : nomFichier\
    tidentifiantMsg\tuser\tprofil\tdate\tmsg\nomMed)

 Entrée : corpus-raw/sujets

 Sortie : sortieTab/*.txt
lexomil-sujet_151053_1 5 paxil123 Doctinaute de bronze 31-03-2014 Il me faisait
     la même chose aussi j'étais fatiguer en permanence lexomil

DOC

#-------------------------------------------------------------
#-------------------------------------------------------------

# Déclaration des modules nécessaires
use strict;
use warnings;
use utf8;
use Modern::Perl;
use HTML::Template;
use HTML::TableExtract;
use Data::Dumper;
use Storable;

# Définition globale des encodages d'entrée et de sortie du script à utf8
binmode STDIN, ':utf8';
binmode STDOUT, ':utf8';

```

```
36 # Vérification du nombre d'arguments de l'appel au script
37 my $usage="perl $0 repertoire-à-parcourir\n";
38 die $usage unless @ARGV==1;
39 die $usage unless -d $ARGV[0];
40
41 #-------------------------------------------------------------
42 #-------------------------------------------------------------
43
44 # Programme principal
45
46 my $dossier=$ARGV[0];
47 # Vérification que le nom du répertoire ne se termine pas par un "/"
48 $dossier=~s/[\/]$//;
49 my @corpus=<$dossier/*>;
50
51 # Créer un répertoire pour les fichiers de sortie
52 mkdir "sortieTab";
53 my $sortieDir="sortieTab";
54
55 # Parcours du répertoire fichier par fichier et appel des routines
56 foreach my $file(@corpus){
57     # next if $file =~ /^\.\.?$/;
58     my $content=&getFileContent($file);
59     &getTableContent($file, $content);
60 }
61
62 #-------------------------------------------------------------
63 #-------------------------------------------------------------
64
65 # Routines
66
67 sub getFileContent{
68
69     # Ouvre chaque fichier et concatène le contenu dans $fileContenu (slurp)
70     # Entrée : nom du fichier à concaténer
71     # Renvoie le contenu concaténé de chaque fichier
72
73     my $file=shift;
74     my $fileContent="";
75     open(IN, "<:encoding(UTF-8)", $file) || die $!;
76     while(my $ligne=<IN>){
77       chomp $ligne;
78       # $ligne=~s/\r//;
79       $ligne=~s/<span class=\"MoodStatus\">(.*?)>//g; # Supprimer les
    MoodStatus
80       $ligne=~s/<span class=\"signature\">(.*?)>//g; # Supprimer les signatures
81       $fileContent.=$ligne;
82     }
83     close IN;
84     return $fileContent;
85 }
86 #-------------------------------------------------------------
87
88 sub getTableContent{
89
90     # Parcourt chaque fichier html et cherche les blaises <table> qui ont des
    attributs class"=messagetable"
91     # Extrait le contenu des colonnes des tableaux repérées et produit un
    fichier txt tabulaire de la forme :
92     # nomFichier  identifiantMsg  user  profil  date  message
```

```
93      # Entrée : nom du fichier traité et la chaine du fichier concaténée
94      # Sortie : fichiers txt "tabulaires" pour chaque fichier html traité
95
96      my $file=shift;
97      my $content=shift;
98
99      # Récupérer le nom du fichier traité
100     $file=~s/.+\/([^\/]+)\.htm/$1/;
101     my $fileName=$1;
102
103     # Chercher le nom du médicament relatif au nom du fichier
104     # (correspodance faite par le hash produit par le script
        calculDistanceNomFichier.pl)
105     my $hashScore=retrieve("hashCorrespMedLev");
106
107     my $med=$hashScore->{$fileName};
108     if(not defined $med){ # si le nom du fichier n'a pas de correspodance dans
        le hash (nom du fichier ne contient pas un mot qui ne figure pas dans le
        dictionnaire)
109     $med="none";
110     }
111
112     # Spécifier les attributs des tableaux recherchées : tableaux ayant un
        attribut class="messagetable"
113     my $te=HTML::TableExtract->new(attribs => { class => 'messagetable' });
114     $te->parse($content);
115
116     # Créer un compteur pour les identifiants des messages
117     my $msgId=1;
118
119     # Reconstruire le chemin des fichiers de sortie
120     my $fileOut="$sortieDir/$file-tab.txt";
121     open(OUT, ">:encoding(UTF-8)", $fileOut) || die $!;
122
123     # Examiner tous les tableaux dans le fichier (dans $content)
124     foreach my $ts ($te->tables){
125     my @user=();
126     foreach my $row ($ts->rows){
127         # Supprimer les lignes vides et celles contenant des publicités
128         unless (not defined $$row[0] or $$row[0]=~/^\(Pub/){
129       my $tabLeft=$$row[0]; # nom de l'utilisateur et profil
130       my $tabRight=$$row[1]; # date et message
131       if($tabLeft=~/(.+?\n?)Profil : (.+)/){ # Cas général
132           my $user=$1;
133           chomp $user;
134           my $profil=$2;
135           print OUT $file."\t".$msgId."\t".$user."\t".$profil."\t";
136       }
137       elsif($tabLeft=~/^Profil/){ # Cas de "profil supprimé"
138           my $user=$tabLeft;
139           my $profil="none";
140           print OUT $file."\t".$msgId."\t".$user."\t".$profil."\t";
141       }
142       elsif($tabLeft=~/(.+)(Invit.$)/g){ # Cas de "invité"
143           my $user=$1;
144           my $profil=$2;
145           print OUT $file."\t".$msgId."\t".$user."\t".$profil."\t";
146       }
147       else{ # Cas sans profil
148           my $user=$tabLeft;
```

```
149           my $profil="none";
150           print OUT $file."\t".$msgId."\t".$user."\t".$profil."\t";
151       }
152       # Appel du routine de nettoyage
153       $tabRight=&cleanText($tabRight);
154       if($tabRight=~/Post. le (\d{2}-\d{2}-\d{4}).+:\d{2}?\s*(.+)/){
155           my $date=$1;
156           my $msg=$2;
157           print OUT $date."\t".$msg."\t".$med."\n\n";
158           $msgId++;
159       }
160     }
161   }
162   close OUT;
163   if(-z $fileOut) {unlink $fileOut}; # supprimer les fichiers vides
164 }
165 #-------------------------------------------------------------
166
167 sub cleanText{
168
169     # Nettoie $tabRight et le normalise
170     # Entrée : le message à nettoyer sous forme d'une chaine de caractères
171     # Renvoie le message nettoyé
172
173     my $text=shift;
174
175     $text=~s/\n+\r*/ /g; # Remplacer les sauts de lignes par des espaces
176     $text=~s/ +/ /g; # Normaliser les espaces
177     $text=~s/\t+/ /g; # Remplacer les tabulations par des espaces
178     $text=~s/\[citation.+?citation\]//g; # Supprimer les citations
179     $text=~s/\[img.+?img\]//g; # Supprimer les images
180     $text=~s/\[quotemsg.+?quotemsg\]//g; # Supprimer les quote messages
181     $text=~s/-{15}.+?//g;# Supprimer les signatures
182     $text=~s/\[g\]//g;
183     $text=~s/\[\/g\]//g;
184     $text=~s/_{2,}//g; # Supprimer les underscore (deux ou plus)
185     $text=~s/\.{4,}//g; # Supprimer les tirets (quatre ou plus)
186     $text=~s/Message .dit. par .+:\d{2}//g; # Supprimer Message édité par
        nomUtilisateur
187     $text=~s/Message cit. \d+ fois//g; # Supprimer Message cité \d fois
188
189     return $text;
190 }
```

A.1.2 Normalisation des noms des médicaments

```
#!/usr/bin/perl

<<DOC;
 MEGAHED Dalia
 Mai 2014

 Usage : perl calculDistanceNomFichier.pl repertoire-à-parcourir liste-médicaments dictionnaire-français

 Le programme prend en entrée le répertoire contenant les fichiers html à traiter (forum Doctissimo.fr) et calcule pour chaque nom de fichier
 la distance de Levenshtein entre les mots contenu dans le nom du fichier et une liste de médicaments (lst_med4) donnée en entrée au programme.
 Les mots traités sont uniquement ceux qui ne figurent pas dans un dictionnaire de formes fléchies du français (dictionnaire-fr.lxq) donné aussi en entrée au script.
 Le programme prend en compte les cas où le nom du médicament ne figure pas en premier dans le nom du fichier.
 La distance de Levenshtein est calculée uniquement entre un médicament figurant dans la liste et un des mots du nom de fichier commençant par la même lettre.

 Entrée : corpus-raw/sujets

 Sortie : hashCorrespMedLev : référence de la table de hachage contenant la correspondance entre le nom du médicament et le fichier ayant la plus petite distance

DOC

#-------------------------------------------------------------
#-------------------------------------------------------------

# Déclaration des modules nécessaires
use strict;
use warnings;
use utf8;
use Data::Dumper;
use Storable;

# Définition globale des encodages d'entrée et de sortie du script à utf8
binmode STDIN, ':utf8';
binmode STDOUT, ':utf8';

# Vérification du nombre d'arguments de l'appel au script
my $usage="perl $0 repertoire-a-parcourir liste-médicaments dictionnaire-français\n";
die $usage unless @ARGV==3;
die $usage unless -d $ARGV[0];

#-------------------------------------------------------------
#-------------------------------------------------------------

# Programme principal

my $dossier=$ARGV[0];
# Vérification que le nom du répertoire ne se termine pas par un "/"
$dossier=~s/[\/]$//;
```

```
48 my @corpus=<$dossier/*>;
49 my $i=0;
50
51 # Appel des routines
52 my $liste=&parcoursFichier($ARGV[1]);
53 my $dico=&parcoursFichier($ARGV[2]);
54 my $scoreLev=calculLevenshteinNomFichier(\@corpus, $liste, $dico);
55 &getMinScoreForMed($scoreLev);
56
57 #-------------------------------------------------------------
58 #-------------------------------------------------------------
59
60 # Routines
61
62 sub parcoursFichier{
63
64     # Parcours un fichier ligne à ligne (liste/dictionnaire) et stocke son
    contenu dans une table de hachage
65     # Entrée : liste des médicaments et dictionnaire des formes fléchies du
    français
66     # Renvoie une table de hachage contenant les mots du fichier
67
68     my $file=shift;
69     my %index=();
70
71     open(LIST, "<:encoding(UTF-8)", $file) || die $!;
72     while(my $ligne=<LIST>){
73       chomp $ligne;
74       $ligne=~s/\r//;
75       $index{lc($ligne)}++;
76     }
77     close LIST;
78     return \%index;
79 }
80
81 #-------------------------------------------------------------
82 sub calculLevenshteinNomFichier{
83
84     # Calcul la distance de Levenshtein entre le nom du médicament et le nom du
     fichier
85     # Appelle les deux routines levenshtein et min qui font le calcul
86     # Entrée : référence du dossier corpus, référence du hash contenant la
    liste et celle du hash contenant le dictionnaire
87     # Renvoie une table de hachage de la forme nomFichier=>médicament=>distance
     (uniquement pour les mots ayant une distance inférieure ou égale à 2 avec
    le médicament)
88
89     my $refCorpus=shift;
90     my @rep=@{$refCorpus};
91
92     my $refList=shift;
93     my %list=%{$refList};
94
95     my $refDico=shift;
96     my %dic=%{$refDico};
97
98     my %score=();
99
100    my $compteur=0; # compteur nombre de fichiers dans le corpus  qui intègrent
     le nom du traitement
```

```
101     my $cpt=0; # compteur nombre total de fichiers dans le corpus
102
103     foreach my $file(@rep){
104       $file=~/.+\/([^\/]+)\.htm$/;
105       my$fileName=$1;
106       my @splittedName=split(/[\W_]/, $fileName);
107       #print Dumper @splittedName;
108       if(exists $list{$splittedName[0]}){ # (traitement figure en premier dans
    le nom du fichier)
109           $score{$fileName}{$splittedName[0]}=0;
110           $compteur++;
111       }
112       else{
113           foreach my $mot(@splittedName){
114             if(($mot!~/\d+/) && (!(exists($dic{$mot})))){ # si le mot n'est pas
     un digit et ne figure pas dans le dictionnaire français
115               if(exists $list{$mot}){ # si le mot existe dans la liste des
    médicaments (traitement figure mais pas en premier dans le nom du fichier)
116               $score{$fileName}{$mot}=0;
117               $compteur++;
118               last; # retenir uniquement le premier traitement trouvé dans le
     nom du fichier
119             }
120             else{ # (traitement mal orthographié dans le nom du fichier)
121               foreach my $med(sort keys %list){
122                 if(substr($mot,0,1) eq substr($med,0,1)){ # si la première
    lettre du mot est la même que celle du médicament
123                   my $distance=levenshtein($med,$mot);
124                   $score{$fileName}{$med}=$distance if ($distance<=2); #
    seuil sur la distance
125                   #print "$fileName\t$med\t$distance\n";
126                 }
127               }
128             }
129           }
130         }
131       }
132       $cpt++;
133     }
134     # Afficher le pourcentage des fichiers qui intègrent le nom du traitement (
    fichiers pour lesquels pas besoin de calculer la distance)
135     $cpt=$cpt-30; # ne pas tenir compte des fichiers vides
136     my $prc=sprintf("%.1f",$compteur/$cpt);
137     warn "$prc ($compteur $cpt) = ",$prc*100,"%\n"; # pourcentage arrondi
138
139     #print Dumper \%score;
140     return \%score;
141 }
142 #-------------------------------------------------------------
143 sub levenshtein{
144
145     my @A=split //, lc shift;
146     my @B=split //, lc shift;
147     my @W=(0..@B);
148     my ($i, $j, $cur, $next);
149     for $i(0..$#A){
150     $cur=$i+1;
151     for $j(0..$#B){
152         $next=min($W[$j+1]+1,$cur+1,($A[$i] ne $B[$j])+$W[$j]);
153         $W[$j]=$cur;
```

```
154         $cur=$next;
155     }
156     $W[@B]=$next;
157     }
158     return $next;
159 }
160 
161 #-------------------------------------------------------------
162 sub min{
163 
164     if($_[0] < $_[2]){ pop @_; } else { shift @_; }
165     return $_[0] < $_[1]? $_[0]:$_[1];
166 }
167 
168 #-------------------------------------------------------------
169 sub getMinScoreForMed{
170 
171     # Cherche la plus petite distance calculée entre le nom du fichier et le
    médicament
172     # Entrée : référence de la table de hachage contenant les distances
    calculées (%score)
173     # Sauvegarde la table de hachage contenant la correspondance entre le nom
    du médicament et le nom du fichier ayant la plus petite distance
174 
175     my $refHash=shift;
176     my %hash=%{$refHash};
177 
178     my %levenshtein=();
179 
180     foreach my $nomFichier(keys %hash){
181       foreach my $medicament(sort {$hash{$nomFichier}{$a} <=> $hash{$nomFichier
    }{$b}} (keys %{$hash{$nomFichier}})){
182           $levenshtein{$nomFichier}=$medicament;
183           last if $i<1;
184       }
185     }
186     store \%levenshtein, "hashCorrespMedLev";
187     #print Dumper \%levenshtein;
188 }
```

A.2 Création du modèle *CRF*

A.2.1 Extraction des informations de *MedEffet Canada*

```
#!/usr/bin/perl

<<DOC;
 MEGAHED Dalia
 Avril 2014

 Usage : perl santeCanadaExtract.pl fichier-reactions.txt

 Le programme prend en entrée le fichier reactions.txt de la base de données en ligne des événements secondaires de Canada Vigilance téléchargée du site http://www.hc-sc.gc.ca/dhp-mps/medeff/databasdon/index-fra.php
 Le programme extrait les colonnes qui correspondent à l'événement secondaire, l'identifiant du rapport, l'identifiant du médicament susceptible de provoquer cet événement et la classe générale de l'événement secondaire

 Entrée : eurekaSante/reactions.txt
 "10902"$"109"$""$""$""$"10012735"$"Diarrhoea"$"Diarrhée"$"10017947"$"Gastrointestinal disorders"$"Affections gastro-intestinales"$"v.16.1"
 "10903"$"109"$""$""$""$"10047700"$"Vomiting"$"Vomissement"$"10017947"$"Gastrointestinal disorders"$"Affections gastro-intestinales"$"v.16.1"
 "11001"$"110"$""$""$""$"10012735"$"Diarrhoea"$"Diarrhée"$"10017947"$"Gastrointestinal disorders"$"Affections gastro-intestinales"$"v.16.1"

 Sortie : listeSanteCanada.txt
 Créatine phosphokinase sanguine augmentée  573376  10005470  Investigations
 Atrophie optique 900015369 10030910  Affections oculaires
 Vertiges labyrinthiques  346874  10047344  Affections du labyrinthe

DOC

#-----------------------------------------------------------------------
#-----------------------------------------------------------------------

# Déclaration des modules nécessaires
use strict;
use warnings;
use utf8;
use Data::Dumper;


# Définition globale des encodages d'entrée et de sortie du script à utf8
binmode STDIN, ':utf8';
binmode STDOUT, ':utf8';

# Vérification du nombre d'arguments de l'appel au script
my $usage= "Usage : $0 reactions.txt\n";
die "$ARGV[0] n'est pas un fichier\n$usage" unless -f $ARGV[0];
die $usage unless @ARGV==1;

#-----------------------------------------------------------------------
#-----------------------------------------------------------------------

# Programme principal

my $tabHash=&extract($ARGV[0]);
&report($tabHash);
```

```
50
51 #-------------------------------------------------------------------------
52 #-------------------------------------------------------------------------
53
54 # Routines
55
56 sub extract{
57
58   # lit le fichier ligne à ligne et splite les colonnes (délimiteur "$")
59   # recupère l'identifiant du rapport ($splited[1]), l'identifiant du
       médicament ($splited[5]),
60     # l'événement secondaire ($splited[7]) et la classe de l'événement
       indesirable ($splited[10])
61   # Entrée : nom du fichier (reactions.txt)
62   # Sortie : renvoie la référence de la table de hachage %hash: ADR=>{idRapport
       =>nombre, idMed=>nombre, class=>classe}
63
64   my $file=shift;
65
66   my %hash=();
67
68   open(IN, $file) || die $!;
69   while(my $line=<IN>){
70     chomp $line;
71     my @splitedTab=split(/\"\$\"/, $line);
72     my $idReport=$splitedTab[1];
73     my $id=$splitedTab[5];
74     my $adr=$splitedTab[7];
75     my $class=$splitedTab[10];
76     #$adr=~s/¿/'/g;
77     #$class=~s/¿/'/g;
78     $hash{$adr}{'idReport'}=$idReport;
79     $hash{$adr}{'idMed'}=$id;
80     $hash{$adr}{'class'}=$class;
81   }
82   close IN;
83   # print Dumper \%hash;
84   return \%hash;
85 }
86 #-------------------------------------------------------------------------
87
88 sub report{
89
90   # Imprime dans le fichier de sortie les colonnes extraites du fichier
       reactions.txt
91   # Entrée : référence de la table de hachage %tabHash
92   # Sortie : fichier tabulaire (listeSanteCanada.txt) avec quatre colonnes: ADR
       \tidRapport\tidMed\tclass
93
94   my $refHash=shift;
95   my %hash=%{$refHash};
96
97   open(OUT, ">:encoding(UTF-8)", "listeSanteCanada.txt") || die $!;
98   foreach my $adr(keys %hash){
99     print OUT $adr."\t".$hash{$adr}{'idReport'}."\t".$hash{$adr}{'idMed'}."\t".
       $hash{$adr}{'class'}."\n" if($hash{$adr}{'idMed'} ne "");
100  }
101  close OUT;
102 }
```

```
#!/usr/bin/perl

<<DOC;
 MEGAHED Dalia
 Avril 2014

 Usage : perl correspondanceSanteCanadaAdrMed.pl liste-SanteCanada.txt fichier-report_drug.txt

 Le programme prend en entrée le fichier drug_report.txt de la base de données en ligne des événements secondaires de Canada Vigilance téléchargée du site http://www.hc-sc.gc.ca/dhp-mps/medeff/databasdon/index-fra.php ainsi que la liste-SanteCanda (fichier tabulaire) créée par le script
 santeCanadaExtractListAdr.pl
 Le programme extrait les colonnes qui correspondent à l'événement secondaire et les médicaments qui peuvent provoquer cet événement

 Entrée : liste-SanteCanada.txt
 Créatine phosphokinase sanguine augmentée  573376  10005470  Investigations
 Atrophie optique 900015369 10030910  Affections oculaires

 cvponline_extract_20131231/drug_report.txt
"271"$"466"$"13406"$"ETHAMBUTOL"$"Concomitant"$"Concomitant"$"Oral"$"Oral"$""$""$""$""$""$""$""$""$""$""$"TABLET"$"COMPRIMÉS"
"273"$"466"$"10586"$"VITAMIN B6"$"Concomitant"$"Concomitant"$""$""$""$""$""$""$""$""$""$""$""$""$""$""$"NOT SPECIFIED"$"NON PRÉCISÉE"

 Sortie : correppondanceSanteCanada.txt
Créatine phosphokinase sanguine augmentée CUBICIN
Atteinte respiratoire par inhalation de gaz ou de vapeurs toxiques OXYGEN
Sous-dosage BUCKLEY'S COMPLETE BUCKLEY'S NIGHT TIME MIXTURE BUCKLEY'S COMPLETE PM - NIGHTTIME
Convulsions localisées L-ASPARAGINASE PREDNISONE

DOC

#-------------------------------------------------------------------------
#-------------------------------------------------------------------------

# Déclaration des modules nécessaires
use strict;
use warnings;
use utf8;
use Data::Dumper;


# Définition globale des encodages d'entrée et de sortie du script à utf8
binmode STDIN, ':utf8';
binmode STDOUT, ':utf8';

# Vérification du nombre d'arguments de l'appel au script
my $usage= "Usage : $0 liseSanteCanada.txt report_drug.txt\n";
die "$ARGV[0] n'est pas un fichier\n$usage" unless -f $ARGV[0];
die "$ARGV[1] n'est pas un fichier\n$usage" unless -f $ARGV[1];
die $usage unless @ARGV==2;

#-------------------------------------------------------------------------
#-------------------------------------------------------------------------

# Programme principal
```

```
53
54 my $tabHash=&extract($ARGV[1]);
55 &report($tabHash);
56
57 #-------------------------------------------------------------------------
58 #-------------------------------------------------------------------------
59
60 # Routines
61
62 sub extract{
63
64   # lit le fichier ligne à ligne et splite les colonnes (délimiteur "$")
65   # recupère l'identifiant du rapport ($splited[1]) et le médicament associé (
       $splited[3])
66   # Entrée : nom du fichier (report_drug.txt)
67   # Sortie : renvoie une référence d'une table de hachage %hash: idReport=>[
       drugName1, drugName2]
68
69   my $file=shift;
70
71   my %hash=();
72
73   open(FILE, $file) || die $!;
74   while(my $line=<FILE>){
75     chomp $line;
76     my @splitedTab=split(/\"\$\"/, $line);
77     my $idReport=$splitedTab[1];
78     my $drugName=$splitedTab[3];
79     push(@{$hash{$idReport}}, $drugName);
80   }
81   close FILE;
82   return \%hash;
83 }
84 #-------------------------------------------------------------------------
85
86 sub report{
87
88   # Imprime dans le fichier de sortie le ADR et les noms de médicaments
       associés
89   # Entrée : référence de la table de hachage %hash
90   # Sortie : fichier tabulaire (listeCorrespSanteCanada.txt) adr\tdrugName1\
       tdrugName2
91
92   my $refHash=shift;
93   my %hash=%{$refHash};
94
95   open(OUT, ">:encoding(UTF-8)", "listeCorrespSanteCanada.txt") || die $!;
96   open(LIST, "<:encoding(UTF-8)", "listeSanteCanada.txt") || die $!;
97   while(my $ligne=<LIST>){
98     chomp $ligne;
99     my @split=split(/\t/, $ligne);
100    my $adr=$split[0];
101    my $id=$split[1];
102      if(exists $hash{$id}){
103        print OUT $adr."\t".join("\t", @{$hash{$id}}),"\n";
104      }
105  }
106  close OUT;
107  close LIST;
108 }
```

ANNEXE B

GUIDE D'ANNOTATION

B.1 Guide d'annotation

B.1.1 Objectifs

Ce guide a pour objectif de constituer une annotation humaine nécessaire pour l'entraînement de *Wapiti* et la création d'un modèle *CRF*. Il renferme les principes d'annotation appliqués en annotant manuellement les éléments que nous souhaitons extraire de manière automatique à partir du corpus.

Cette annotation permet de détecter les événements secondaires tels qu'ils sont exprimés par les patients dans les messages déposés et d'autres entités biomédicales telles que les traitements et les indications des pathologies, ce qui facilite la distinction entre les indications des traitements et les événements provoqués par l'utilisation de certains médicaments. Elle aide également à identifier la nature des événements secondaires détectés, en mettant l'accent sur leur polarité (événements positifs, négatifs ou neutres). Outre les événements secondaires et les indications des médicaments, certaines informations relatives aux traitements tels que la forme galénique du médicament, sa concentration et son dosage sont aussi annotées. Enfin, cette annotation met en évidence le lien qui existe entre traitement-événement et traitement-posologie à l'aide des relations identifiées. La figure B.1 montre deux messages annotés manuellement en utilisant l'outil d'annotation *Brat*.

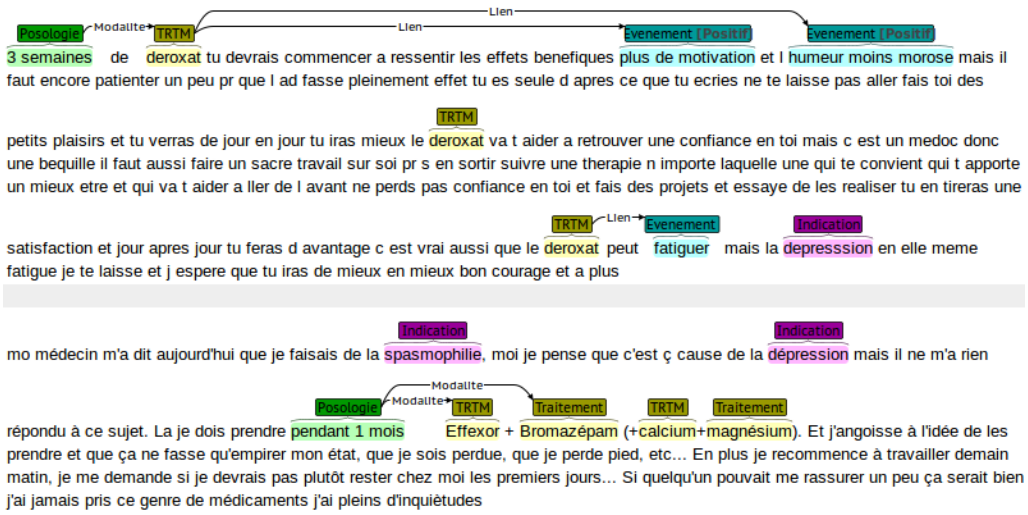


FIGURE B.1 – Exemples de messages annotés manuellement

B.1.2 Règles et exemples

B.1.2.1 Catégories

Les annotations s'articulent autour de quatre catégories, qui sont *Traitement*, *Événement*, *Indication* et *Posologie*.

- **Traitement (TRTM) :** les noms de médicaments dont parlent les patients dans leurs messages ex. : *Solian 100*, *Prozac*, *Effexor LP*.
- **Événement (EVT) :** les événements secondaires rapportés par les utilisateurs ex. : *grosse migraine*, *beaucoup de fatigue*, *régurgitation*.
- **Indication (IND) :** les raisons pour lesquelles le traitement a été prescrit. Plus précisemment, il s'agit des pathologies et symptômes que traite le médicament ex. : *dépression*, *crise d'angoisses*, *anxiété*.
- **Posologie (POSO) :** les informations associées au nom du médicament (concentration, dosage, fréquence, durée, forme galénique) ex. : *2 gélules*, *25mg*, *un comprimé par jour*.

Outre ces catégories, un attribut de polarité a été défini pour la catégorie *Événement*, en vue de déterminer la positivité ou la négativité de l'événement secondaire détecté. Dans la majorité des cas, la polarité attribuée aux événements est négative, étant donné qu'il s'agit d'une réaction inattendue provoquée par un médicament utilisé. Dans ces cas, aucun attribut est ajouté à la catégorie *Événement* lors de l'annotation.

En revanche, si un événement produit par un traitement se révèle bénéfique, il sera associé à un attribut *positif*. Il s'agit par exemple de la diffusion de bactéries de la flore intestinale dans la circulation sanguine et la lymphe provoqueée par le *cyclophosphamide* nécessaire en chimiothérapie anticancéreuse. Il s'est avéré que la réponse immunitaire contre ces bactéries aident le patient à mieux lutter contre la tumeur en stimulant de nouvelles défenses immunitaires.

Il est à noter que la détermination de la polarité d'un événement selon le contexte dans lequel il figure semble parfois difficile au sein de certains messages. Ces exemples difficiles à identifier portent ainsi un attribut *neutre*.

B.1.2.2 Relations

Deux types de relations sont considérés durant l'annotation. Le premier indique le lien qui existe entre les traitements et les événements secondaires. Il s'agit par exemple de la relation entre le médicament *Deroxat* et *vertige* ou de celle entre *Effexor* et *crispation*. Les relations de type *Lien* permettent ainsi d'établir les correspondances qui figurent entre traitement-événement secondaire. Il est à noter que les relations entre les traitements et les événements secondaires liés à l'arrêt du médicament ne sont pas extraites, étant donné qu'elles ne font pas partie des objectifs de notre étude. Il en va de même pour les relations traitements-indications

Quant au second type de relations, il concerne la modalité du traitement. Plus précisément, la relation *Modalité* nous renseigne sur la posologie du traitement utilisé qui a mené à l'événement secondaire. C'est par exemple la relation entre le nombre 37.5 et le traitement *Effexor LP* dans le message « *bonjour, y-a-t-il accoutumance à* effexor LP *37,5 et au bout de combien de temps ? merci pour vos témoignages* ».

B.1.2.3 Règles

Les règles d'annotation adoptées sont appliquées au niveau de chaque message séparément. Elles reposent sur plusieurs principes :

- Toutes les occurrences d'une même entité liée à une des catégories étudiées sont annotées. Seuls quelques cas font exception à cette règle.
  1. Les entités mal orthographiées qui figurent au sein d'un message contenant d'autres occurrences (de la même entité) écrites correctement. L'apport de l'annotation de l'entité mal orthographiée ne semble pas être significatif dans ce cas.
  2. Les informations qui ne sont pas relatives à un patient. À titre d'exemple, une personne qui exprime son angoisse car le patient qu'elle connait ne se porte pas bien. L'événement mentioné ne se rapporte pas dans ce cas au traitement.
  3. Les événements secondaires qui ne sont pas associés à un traitement particulier. C'est le cas par exemple d'un utilisateur qui mentionne un problème d'angoisse lié à la recherche d'un travail. Le problème d'angoisse est ici en relation avec la recherche d'emploi et non pas avec le traitement. Il en va de même pour les événements secondaires liés à l'arrêt de la prise du traitement.
- Les frontières des éléments à annoter sont définies de manière à retenir la portion de texte qui rapporte le plus d'information. L'ajout de l'adjectif *constamment* dans *envie de dormir constamment* peut nuancer le sens de la portion annotée.
  D'autres règles sont en effet définies pour bien limiter les frontières des annotations :
  1. Ne pas annoter les pronoms et les déterminants ex. : *l'Effexor* vs. *Effexor*.
  2. Gérer les entités discontinues ex. : *peur de mourir et que mon cœur s'arrête* vs. *peur de dormir* et *peur que mon cœur s'arrête*.
  3. Consérver le syntagme verbal uniquement si le verbe décrit une action constituant le signe ex. : *être en détresse* vs. *se lever la nuit sans se rendre compte*.
  4. Tenir compte de la négation ex. : *angoissé* vs. *ni angoissé*.
  5. Annoter le nom complet du médicament ex. *Effexor* vs. *Effexor LP*.
  6. Ne pas annoter les modalités relatives aux événements ex. : *fatiguer pendant un mois* vs. *fatiguer*.
- Annoter les termes génériques liés aux traitements ex. : *antidépresseur*, *neuroleptique* si aucun nom de médicament est présent dans le message et si plusieurs médicaments sont mentionnés mais un seul cause l'événement secondaire ex. : *je prends un neuroleptique et un AD mais l'AD me fatigue*.
- Identifier uniquement les relations entre les éléments les plus proches (et non pas toutes les relations possibles au sein d'un même message). Si un événement figure dans un message avec deux occurrences d'un nom de médicament, le nom présent dans la même phrase ou celui qui est le plus proche est choisi.
- Ne pas considérer les relations qui figurent entre les différents messages au sein d'un même fil de discussion.

- La relation traitement-événement est établie uniquement si le lien entre la pathologie et l'évenénemt secondaire est explicite au sein du message.

ANNEXE C

GRAMMAIRE D'ANNOTATION

```
//////////
// Grammaire utilisée par MEDINA
//////////


// Variables définies par l'utilisateur.
// - La déclaration se fait avec le symbole "$", l'utilisation avec le symbole "_" (meilleure visibilité).
// - Ne pas utiliser des noms de variables similaires à des mots du texte à traiter (qui existeraient en majuscules).
// - Vérifier qu'il y a une tabulation (et pas des espaces) entre le nom de variable et son contenu.

$CHIFFRE  \d
$DECI _CHIFFRE+,_CHIFFRE+|_CHIFFRE+\._CHIFFRE+
$ALPHA  ([a-zA-Z])
$MAJ  ([A-Z])
$DET  (d'|d|des|une|un)

$NUMBER (un|deux|trois|quatre|cinq|six|sept|huit|neuf|dix|un demi|une demi|une seule|1/2)
$POSO (mg|gélule|gélules|gelule|gelules|comprimé|comprimés|comprime|comprimes|fois|comp|CP entier|CP|cp| cp entier|entier|entiers|tablette|tablette et demi)
$TIME (le matin|le soir|par jour|par j|par jours|par soir|matin|soir)
$SLASH  (/jour|/j)
$PREP (pendant|durant|sur|il y a|Il y a|depuis)
$DURATION (jours|jour|mois|ans|an|année|annee|années|annees|semaines|semaine)
$DOSE (hautes doses|hautes dose|haute doses|haute dose|dosage faible|petites doses|petites dose|petite doses|petite dose|faibles dosages|faible dosage)
$PB (crises|crise|problèmes|problemes|problème|probleme|pb|prob|troubles|trouble)

$INDVERB  (calmer|surmonter|lutter contre|traiter|souffrir|soigner|contrôler|controler|affronter|aider)
$INDNOUN  (symptôme|symptômes|symptome|symptomes|traitement|souffrant|aide)
$INDPRO (les|la|le|des|de|tes|ta|ton|mes|ma|mon)


// Posologie

Posologie (_DOSE)
Posologie (.*pass.* à) !~_DECI+|_CHIFFRE+( _POSO|)~!
Posologie (.*pass.* a) !~_DECI+|_CHIFFRE+( _POSO|)~!
Posologie ((_DECI+|_CHIFFRE)(/jour|/j|)x(_DECI+|_CHIFFRE)(/jour|/j|))
```

```
37 Posologie ((_NUMBER|_DECI+|_CHIFFRE+)( |)_POSO( _TIME|_SLASH|))
38 Posologie ((_NUMBER|_DECI+|_CHIFFRE+) (le matin|le mat|le soir))
39 Posologie (.+ _PREP) !~(_CHIFFRE+|_NUMBER) _DURATION~!
40 Posologie (.+ _MAJ_ALPHA+)!~(_CHIFFRE+|_DECI+)~!
41
42 // Evenement
43
44 Evenement (.* ?j('| )(avais|ai) eu _DET) !~(_ALPHA+ et _ALPHA+|_ALPHA+ d'_ALPHA
       +|_ALPHA+)~!
45 Evenement (.* ?j('| )(avais|ai) _DET) !~(_ALPHA+ et _ALPHA+|_ALPHA+ d'_ALPHA+|
       _ALPHA+)~!
46 Evenement !~_PB( d'| des | de | d | )(_ALPHA+ et( d'| des | de | d | )_ALPHA+|
       _ALPHA+)~!
47 Evenement (.* ?je me sen. )!~(_ALPHA+ et _ALPHA+|_ALPHA+)~!
48 Evenement (impr.ssion de) !~_ALPHA+~!
49
50 // Indication
51
52 Indication  (.* ?_INDVERB _INDPRO) !~_ALPHA+~!
53 Indication  (.* ?_INDVERB l')!~_ALPHA+~!
54 Indication  (.* ?j('| )(avais fai.|ai fai.|) _DET) !~(_ALPHA+ et _ALPHA+|_ALPHA
       + d'_ALPHA+|_ALPHA+)~!
55 Indication  (.* ?je fai. _DET) !~(_ALPHA+ et _ALPHA+|_ALPHA+ d'_ALPHA+|_ALPHA+)
       ~!
56 Indication  (.* ?suite (à|a|aux|au) (une|un|)) !~_ALPHA+~!
```